Article

Patterns

Statistics in everyone's backyard: An impact study via citation network analysis

Highlights

- Citation data were collected for core statistics papers from the past two decades
- A comprehensive evaluation of their impact on other scientific fields was conducted
- The external impact of statistics has been increasing in volume and diversity
- The most influential statistics community for a given external topic was found

Authors

Lijia Wang, Xin Tong, Y.X. Rachel Wang

Correspondence

xint@marshall.usc.edu (X.T.), rachel.wang@sydney.edu.au (Y.X.R.W.)

In brief

To measure the impact of statistics on other scientific disciplines in the age of data revolution, we performed the first comprehensive analysis of citation data for core statistics theory and methodology papers from 1995 to 2018. We show that, overall, the external impact of statistics has been increasing in volume and diversity, while there also exist highly cited publications that have reached mostly one side of the audience. We propose a local clustering method for finding the most influential statistics community for any given external research topic.

Wang et al., 2022, Patterns *3*, 100532
August 12, 2022
https://doi.org/10.1016/j.patter.2022.100532

CellPress

Patterns

CellPress
OPEN ACCESS

Article

Statistics in everyone's backyard: An impact study via citation network analysis

Lijia Wang,[1] Xin Tong,[2,*] and Y.X. Rachel Wang[3,4,*]
[1]Department of Mathematics, University of Southern California, Los Angeles, CA 90007, USA
[2]Department of Data Sciences and Operations, University of Southern California, Los Angeles, CA 90007, USA
[3]School of Mathematics and Statistics, University of Sydney, Sydney, NSW 2006, Australia
[4]Lead contact
*Correspondence: xint@marshall.usc.edu (X.T.), rachel.wang@sydney.edu.au (Y.X.R.W.)
https://doi.org/10.1016/j.patter.2022.100532

THE BIGGER PICTURE How much impact has statistics made on other scientific fields in the era of big data? This work represents the first effort toward quantifying the external influence of statistical theory and method research through citation network analysis. We formulate the problem of finding the most relevant statistical research area for any external research topic as a local clustering problem, suggesting new applied and theoretical grounds for alternative community detection techniques. The results of our analysis confirm that statistics plays an active and expanding role in serving other disciplines. The data we have collected are rich in content and structure, lending themselves naturally to future modeling and analysis from different perspectives.

Development/Pre-production: Data science output has been rolled out/validated across multiple domains/problems

SUMMARY

Statistical methodologies are indispensable in data-driven scientific discoveries. In this paper, we make the first effort to understand the impact of recent statistical innovations on other scientific fields. By collecting comprehensive bibliometric data from the Web of Science database for selected statistical journals, we investigate the citation trends and compositions of citing fields over time, and we find increasing citation diversity. Furthermore, in a new setting, we apply a local clustering technique involving personalized PageRank with graph conductance for size selection to find the most relevant statistical innovation for a given external topic in other fields. Through a number of case studies, we show that the results from our citation data analysis align well with our knowledge and intuition about these external topics. Overall, we have found that the statistical theory and methods recently invented by the statistics community have made increasing impact on other scientific fields.

INTRODUCTION

The past decade has witnessed the success and impact of big data, whereby numerous areas in science, technology, and industry have been transformed by an ever-growing amount of data not only large in size, but also complex and rich in structure. As the discipline that focuses on the collection, analysis, and interpretation of data, statistics plays a central role in the data revolution. Over the years, fundamental concepts and tools have been developed in statistics to extract useful information from data.[1,2] On the other side, however, concerns have been expressed about the openness of the statistics community to addressing unstructured problems and the relevance of statistical research topics to the intended fields of application.[3,4] A central question in this debate revolves around understanding the outward facing impact of statistics.

In this paper, we consider measuring the impact of theoretical and methodological research in statistics on other scientific disciplines in recent decades. One direct way to measure the impact of academic works in general is through citation data. In the digital age, comprehensive bibliometric studies have been made possible by the existence of citation databases such as Web of Science and Scopus. From these databases, citations between papers can be extracted, represented as a

network, and studied using network analysis techniques. These citation networks have been used to track the movements of ideas and measure the distance between different scientific fields.[5,6] Rising scientific interdisciplinary knowledge flows have been documented[7–9] and shown to have a positive effect on the development of specific scientific topics.[10] Various measures of diversity in terms of cross-disciplinary citations have been proposed to evaluate papers or journals from an interdisciplinary aspect.[11–14] Many papers[15–17] have also analyzed how the diversity influences the (citation) impact of papers or journals. However, bibliometric studies focusing on publications in statistics have been relatively limited. Stigler[18] and Varin et al.[19] used the Bradley-Terry model to measure the import and export of knowledge between statistical journals. Ji and Jin[20] collected and analyzed citation and coauthorship networks for papers in four top statistical journals. In contrast to these papers focusing on the structure of citation patterns inside statistics, we provide the *first comprehensive study* analyzing the connections *between statistics and other fields*. Different from the aforementioned works, which study general interdisciplinary connections, our work focuses on the influence of statistical publications in the age of big data. Moreover, as we are primarily interested in evaluating how statistical tools have served other scientific fields, the flow of knowledge we focus on is one directional, i.e., from statistics to other fields, as reflected by external citations of statistics papers. We are also interested in analyzing the internal citations within statistics and comparing the internal and external impact.

We collect citation information for papers published in selected statistical journals from the Web of Science (WoS) Core Collection. These published papers are termed "source papers" for being the source of knowledge export; our complete dataset contains citations between source papers as well as their citations by papers (termed "citing papers") in other journals and fields. Using descriptive statistics, we characterize the trends in citation volumes and compositions of citing fields for the source papers over time, paying attention to fields external to statistics. We compare the internal and external citations for highly cited source papers and identify the corresponding statistical research areas highly ranked by both criteria. Citation trend analysis of these areas allows us to associate them with external fields on which they have made an intellectual impact.

Given a network, one of the most important structural features is the presence of communities, where subsets of nodes are more tightly connected with one another than with the rest of the network. On the citation network for source papers, *global* clustering techniques are usually used to partition the nodes into densely connected communities, as has been done in Ji and Jin,[20] offering an overall view of various research areas within statistics. However, in this paper, we are more interested in connecting these communities in statistics with research topics in other disciplines they have an influence on. That is, given an external research topic (e.g., COVID-19), we investigate the most relevant community in statistics, with relevance measured by the citation data. Thus, in contrast to common global clustering approaches, we formulate our community detection problem using a *local* clustering perspective.

Our local clustering procedure consists of two steps. First, given a small subset of seed nodes from a community of interest, all nodes are ranked in terms of their relevance to this target community using a local clustering algorithm. Many existing local clustering algorithms are based on seed expansion, which involves performing random walks starting from the seeds and ranking the nodes according to their landing probabilities from the walks, with well-known examples including personalized PageRank (PPR)[21–23] and heat kernels.[24,25] More recently, the theoretical properties of PPR have been studied under generative network models,[26,27] with the latter showing that PPR can include high-degree nodes outside the community of interest, while using the adjusted PPR (aPPR) algorithm[28] can correct the bias. The second step of our procedure uses conductance[28–30] to evaluate the quality of the community found along the sorted list of nodes, cutting it at an appropriate size to reveal the full community. Conductance measures the fraction of total edge volume that points outside the cluster, and a smaller conductance indicates the cluster is more separated from the rest of the network and more likely to be a community on its own. Combining aPPR for clustering and conductance for size selection, and adapting them to our data structure, we provide an integrative procedure for detecting the most relevant statistical research community for an external research topic of choice. We demonstrate its performance using several case studies from different scientific disciplines, where the results show that core statistical theory and method developments have stayed relevant and attuned to problems of high societal and scientific interest.

The main contributions of our paper include the following. (1) We provide the first comprehensive study analyzing the recent impact of statistics publications on other scientific disciplines and give a positive answer to the debate about the relevance and outward-facing impact of statistics as a discipline in the age of big data. (2) We apply a local clustering method utilizing aPPR and conductance measure to identify the most influential statistical community for an external topic, which requires combining internal and external citation information in a sensible way. The method is fully automated and can be applied to any external topic of choice, providing a different application for PPR and related techniques. In contrast, existing applications of PPR[31–33] and various modified versions of the algorithm[34–36] to citation networks are focused on analyzing the internal network of a field and ranking the papers in terms of their internal impact. (3) Under a commonly used network model, we provide the first theoretical justification for the combined use of aPPR and conductance to identify a target community, demonstrating that our approach is principled and generalizable.

RESULTS

Overview of the citation data

We conducted our study on all the papers published from 1995 to 2018 in five influential statistics journals: *Annals of Applied Statistics* (AOAS), *Annals of Statistics* (AOS), *Biometrika*, *Journal of the American Statistical Association* (JASA), and *Journal of the Royal Statistical Society: Series B* (JRSSB; JRSSB used two publication names during 1995–1997; we included both in our study). Among our selected journals, AOS, *Biometrika*, JASA, and JRSSB are considered by many researchers in the statistics community as top outlets for theory and method papers. We

A

Number of citations	Proportion
0	1.92%
1 ~ 500	97.02%
501 ~ 5000	1.00%
> 5000	0.06%

B

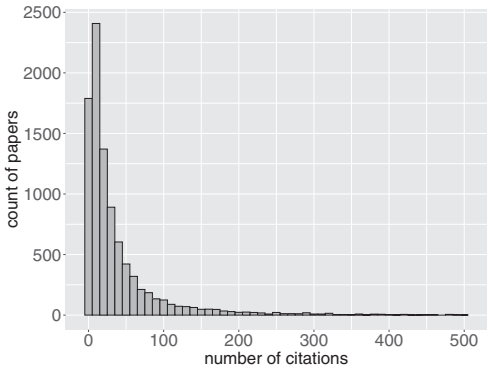

Figure 1. The distribution of citation counts for the source papers

(A) The proportions of papers falling into four citation brackets.

(B) Histogram of papers with $\leq$500 citations.

have also included AOAS as a representative journal with a broad applied focus. A total of 9,338 papers published in these journals in the time span were considered. We call these publications "source papers," since they act as a source of knowledge for papers citing them. In addition to the citations between the source papers, 264,356 papers cited these source papers in our dataset (collected by December 2020); these papers are called "citing papers." Our citation network consists of 273,694 nodes, including all the source and citing papers, and edges, representing citations between the source papers and from the citing papers to the source papers.

Citation trends

As shown in Figure 1A, the majority of source papers had fewer than 500 citations, with 1.92% of the source papers receiving zero citations. Figure 1B further plots the distribution of the citation counts for source papers with citations from 0 to 500. We observe that removing the zero-citation papers would lead to a power-law distribution of the citation counts. Notably, 0.06% of the papers (six papers) received more than 5,000 citations.

Looking at the trends over the years, the total number of citations for each journal grows consistently (Figure S1), and the growth is not due to the journals expanding their volumes of publications. In fact, there was no significant increase in the annual number of publications in each journal (Figure S2), except for AOAS. AOAS was established in 2007 and subsequently went through a fast growth period before stabilizing. To account for the effect of publication numbers, for each year T, we normalized the annual citation count for each journal by the total number of published papers from 1995 to T in that journal, since any citing paper published in year T is free to cite source papers in the period 1995–T. Figure 2A shows that the normalized citations still increase consistently over the years for all the journals, among which JRSSB enjoys substantially more citations per article after 2002. AOAS's normalized citations have been growing quickly, as a relatively new journal. Using a different way of normalization, for each journal, Table S1 computes the average counts of citations received by the journal's papers in the first few years of their publication, and Figure S3 records the change in average citations in the first 5 years over time. Both show that papers in JRSSB have more citations.

It is clear that citation counts are not distributed equally across all the papers, and one possible way to measure citation inequality is through the Lorenz curves[6,20] in Figure 2B. Curves closer to the bottom right corner indicate greater extent of inequality. Most journals have highly similar curves, while JRSSB appears to have the most significant inequality. This can be explained by the fact that there are four papers that each received more than 5,000 citations, accounting for 49.5% of the total citations in JRSSB in this period. After these four papers are removed, the normalized citation counts for JRSSB become much closer to the other journals, but it remain the highest of all the journals (Figure S4).

Diversity of citing fields over time

We divided the dataset into 83,503 internal and 190,191 external papers. The internal papers include statistics papers labeled as "STATS" and mathematics papers (excluding STATS papers) labeled as "MATH." Also, we classify the external papers into five broad categories: arts and humanities ("ART"), life sciences and biomedicine ("BIO"), physical sciences ("PHY"), social sciences ("SOC"), and technology ("TECH"). Figure 3A shows the research area breakdown for all the citations over the years; Figure S5 plots the proportions of these areas. As expected, in the earlier years of our period of study, most of the citations are from within statistics. However, the proportion of external citations soon begins to increase at a fast pace and finally exceeds half. Among the external citations, BIO and TECH have heavy weights. The proportion of external citations also increases over time for all the journals, with AOAS and JRSSB having larger proportions than the others.

One way to summarize the distribution of proportions and put the diversity measure for each journal on the same scale is through the Gini concentration (Herfindahl) index,[18] where we compute the scaled sum of squared proportions of the internal category and other external categories. A value close to 100 would indicate that most of the citations come from the internal category, whereas lower values suggest the journals have more diverse citation profiles from external categories. Figure 3B plots the change in the Gini concentration index for each journal over the years. Overall, the trends agree with the patterns of increasing diversity in citation proportions at each journal level (Figures S6 and S7) and at the overall level (Figures 3A and S5). All the journals have demonstrated increasing connections with external fields, with AOAS, JASA, and JRSSB showing more external connections than the others.

Internal and external impact of most highly cited papers

Now we turn to examine the internal and external impact of some specific source papers selected based on their high citation counts. Do highly cited papers always have high impact, both internally and externally? To this end, we first ranked the source papers according to their internal and external citation counts separately. Focusing on papers in the top 20 list by either internal

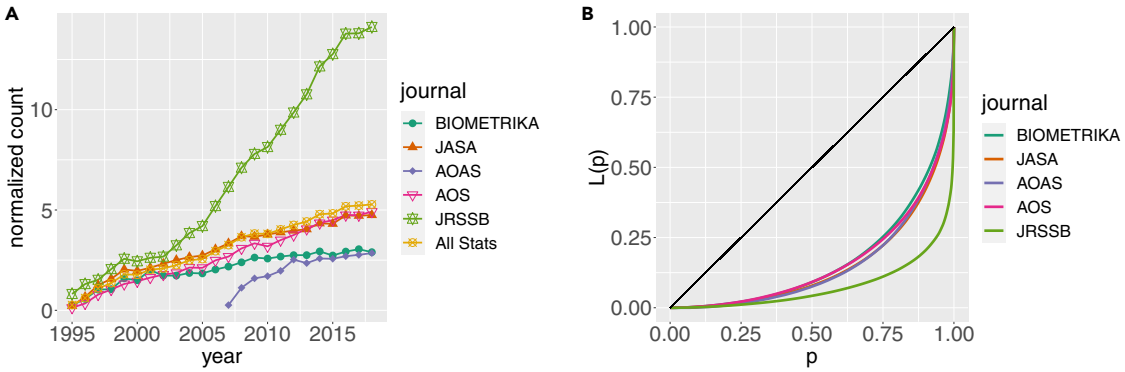

Figure 2. Citation trends and distributions for each journal

(A) The normalized number of citations over the years. "All Stats" refers to all the source papers.

(B) The Lorenz curve for each journal.

or external counts, Figure 4 shows their respective ranks internally and externally. One can see that most of these papers are ranked high under both criteria, except for a few outliers. We focus on the most obvious two (boxed in red) and provide their information in Table 1 and further analysis below.

The first paper[37] in Table 1 ranks in the top 20 based on the internal citation counts, but its external rank is relatively lower in comparison. Since the paper is about distribution theory, we find, unsurprisingly, that most of the citations come from fields closely related to statistics. Table S1 provides the top 10 WoS categories and their number of occurrences among the citations, with "Statistics & Probability" appearing most often. Also, most of these categories contain the keyword "math," which explains the higher internal rank. The other categories (e.g., "Computer Science, Interdisciplinary Applications") are still closely related to statistics or mathematics. Upon removal of the internal papers, the occurrences of these categories, other than statistics and mathematics, decrease significantly (Table S3), suggesting that many of the previous counts are contributed by internal papers with multiple category labels. Overall, the paper has reached a larger audience within statistics and mathematics, most likely due to its technical nature.

The second paper[38] in Table 1 demonstrates the opposite pattern, with a high external rank but a low internal rank. This paper proposes a practical method of evaluating and adjusting for the possibility of publication bias (e.g., a preference for positive results), a well-known phenomenon in published academic research, especially in meta-analysis, which thus has attracted wide scientific interest. Table S4 lists the top 10 most frequent WoS categories among all the citations. One can see that the list is dominated by psychiatry and psychology, while statistics- or mathematics-related categories are not present. This list remains almost unchanged after removing all the internal papers from the citations (Table S5). We have additionally searched for keywords related to publication bias in the title and author keywords of the internal papers. The search returns only 59 papers, confirming that the topic is less explored internally and could be a potential area for further theoretical and methodological development in statistics. We note that Figure 4 has another paper[39] with a low internal rank (469) and a high external rank (12). The paper has a category profile similar to that of Duval and Tweedie[38] (Table S6), and hence detailed discussion is omitted.

Table 2 lists all the papers that are ranked in the top 20 both internally and externally. We classify these papers roughly into five topics: Markov chain Monte Carlo (MCMC), causal inference (causal), penalized regression, false discovery rate (FDR), and Bayesian model selection. To investigate the influence of these papers on other fields, we considered the aggregated citations by the five topics and broke down the citations by category labels, similar to Figure 3A. Note that we refined the categories by adding two category labels, "BE" for the research area business and economics and "CS" for the research area computer science, since we noticed a considerable number of citations from these two areas, especially for causal inference and penalized regression. Figure 5 (and Figure S8) shows that the influence on other fields differs by statistical research topic. FDR and Bayesian model selection have always attracted a substantial proportion of citations from BIO, even from the earlier years. MCMC and penalized regression have more citations from CS than the others. On the other hand, causal inference has the largest proportion of citations from SOC and BE among the five topics.

Local clustering reveals the most relevant statistical research areas for external topics

We applied local clustering to our citation data to find the most relevant statistical research areas for given external topics. Our method involves first finding seed papers for a given topic using citation data, followed by searching for related source papers using aPPR and community size selection using conductance. The details and the main algorithm (Algorithm 1) can be found in the experimental procedures, where we also provide theoretical justifications of the method under a commonly used network model. We choose three external topics (single-cell transcriptomics, labor economics, and flu) of high general interest, spanning biology, economics, and epidemiology. The size of the community found for each topic is listed in Table 3. We can see that these subnetworks indeed have significantly denser connections (and in some cases, higher clustering coefficients)

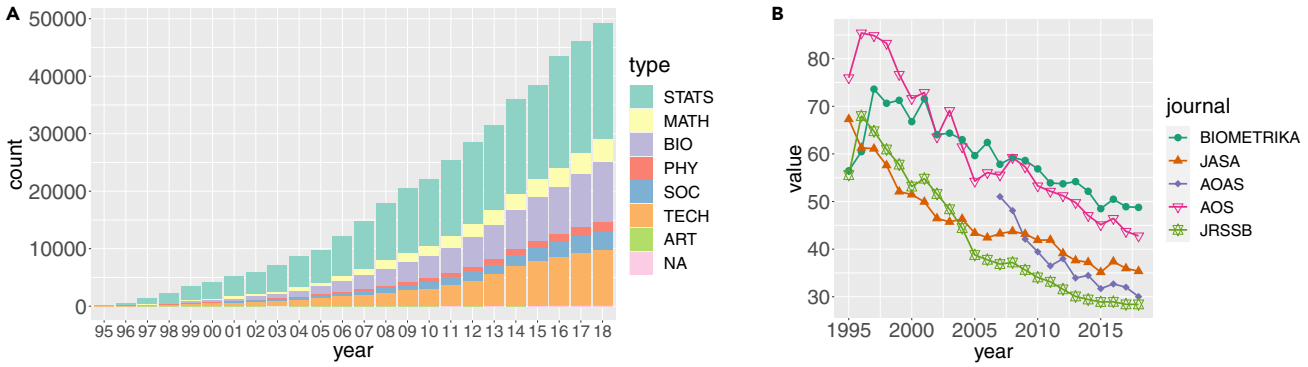

Figure 3. Diversity of citing fields in statistics journals

(A) The annual counts of internal and external citations for all the source papers.

(B) The yearly Gini concentration index for each journal.

than the whole network. The subnetworks and the word clouds generated from the keywords of the subnetwork papers can be found in Figure 6. Furthermore, for the topics of labor economics and flu, we investigated how these influential communities evolved over time by breaking the study period into five roughly even windows and applying local clustering to each (Note S2.1). As single-cell transcriptomics is a relatively recent topic (the earliest publication in our dataset being from 2013), instead of temporal changes, we examined the robustness of our results with respect to the tuning parameters. We discuss these results in more detail below, interpreting the results with our understanding of the topics.

Single-cell transcriptomics

Rapid advances in single-cell sequencing technologies in the past decade have enabled researchers to profile different aspects of an individual cell, in particular, its transcriptome. After appropriate preprocessing, single-cell transcriptomic data usually take the form of a large, sparse matrix, with tens of thousands of rows representing genes and columns representing cells. The sparse, noisy, and heterogeneous nature of such data has proved a fertile ground for the development of statistical and computational methods (see, e.g., Kharchenko[50] for a review). Inspecting the subnetwork and word cloud in Figure 6A, perhaps unsurprisingly, we find that a significant fraction of the papers selected are concerned with multiple testing and connected to the hub node 79.[48] As an example, multiple testing is routinely performed in the analysis of single-cell RNA-sequencing (scRNA-seq) data for identifying differentially expressed genes, which involves applying a statistical test to a large number of genes to determine if their expression levels are significantly different between two sets of cells. The word cloud also suggests clustering as another main keyword; in the subnetwork, clustering is a topic shared by the set of papers tightly knit around nodes 35[51] and 78.[52] In the analysis pipeline of scRNA-seq data, clustering is applied to a dimension-reduced scRNA-seq matrix to identify distinct subpopulations of cells, which can correspond to different cell types or states. The related feature selection and model selection problems are highly relevant in this context, as they help researchers determine genes (features) that distinguish these subpopulations and the total number of subpopulations observed. Finally, in Tables S7 and S8, we examine the stability of this cluster found in Figure 6A with respect to the two main tuning parameters in our method (the threshold parameter in the construction of preference vector and the teleportation constant, see the experimental procedures). The Jaccard index values are reasonably close to 1 for most of the parameter ranges, indicating that the cluster found in Figure 6A is stable.

Labor economics

Labor economics aims to understand the functioning and dynamics of the markets for wage labor. Many fundamental questions in this subject—How does education affect income? How does health care affect income?—are of a causal nature. Economists and governments would like to design policies that might achieve certain economic and social welfare goals based on causal analysis. Randomized controlled trials (RCTs) are usually not available for labor economics problems. Therefore, it makes sense to see that an overwhelming majority of the statistics papers selected in the subnetwork and word cloud in Figure 6B are in the realm of causal inference. We note here that causal inference itself is a rapidly growing interdisciplinary field spanning statistics, econometrics, psychology, computer science, and many other disciplines; thus, contributions to its development do not only come from statistics. Our results intend to mostly reflect the influence from the statistical perspective.

Concretely, in the word cloud, the frequently appearing keywords (minus "test") are all technical terms in causal inference—"propensity score," "instrumental variable," "structure model," "matched sampling," "treatment effect," "matching," and "observational study." Notably, node 152 (circled in red),[41] a hub in the subnetwork, links the structural equations framework in econometrics and the potential outcomes framework in statistics. The paper provides conditions for a causal interpretation of the instrumental variable (IV) estimand and quantifies the bias of violations of the critical assumptions.

Investigating how the most influential community has changed over time, Figure S9 reveals that node 152 (circled in red) has attracted wide attention and become the hub of a cluster of causal

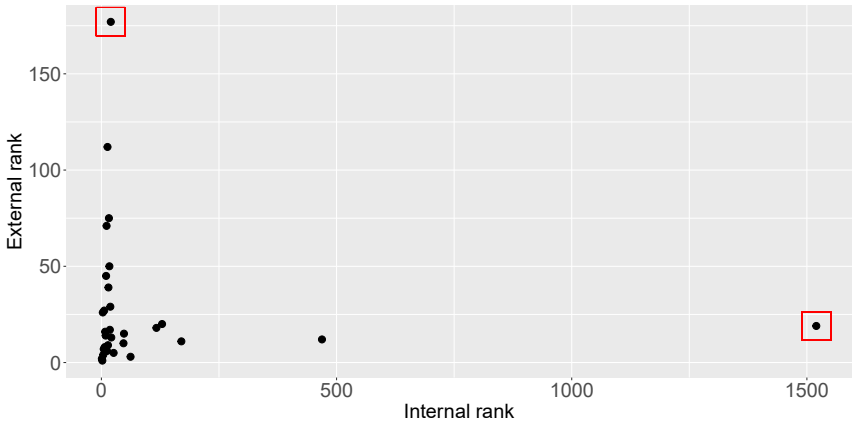

Figure 4. The comparison of external and internal ranks for highly cited papers

inference papers since 2000. Meanwhile, there are clues that other statistical papers started influencing labor economics after 2011, forming small disconnected clusters (Figure S9D). For example, the word cloud (Figure S11) for the cluster around node 85[53] (boxed in blue, Figure S9D) shows that one new contribution was from (social) network analysis. Later, this cluster formed a connection with the causal inference community through node 163[54] (Figures 6B and S9E), which applied the causal effect estimation method to social network data. In addition, we note that many papers appearing in the selected community (for both the overall period in Figure 6B and the more recent period in Figure S9E), such as node 34[55] and node 36,[56] are rather recent. This coincides with the recent surge in the study of causal inference in the statistical community in the past few years and offers some evidence that the new developments quickly penetrate into other research fields.

Flu

The global pandemic of COVID-19 has further ignited wide research interests in the modeling and prediction of the spread of an epidemic. We choose flu as an example of epidemics due to its long history of study and frequent appearance in the literature of epidemiology. (The results from using COVID-19 as the topic are presented in Note S2.2, which includes Table S9 and Figures S15 and S16.) Many of the keywords in Figure 6C are related to stochastic processes and state-space modeling. The word MCMC appears the most often, being a commonly used technique for parameter estimation in these epidemic models. Looking more closely at the subnetwork, many of the papers focus on refining the susceptible-infectious-recovered (SIR) model for infectious diseases, including flu and SARS.

Figure 6C has two hub nodes, 141[57] and 35.[58] Node 141 (circled in red), which is concerned with the parameter estimation problem for different types of observed data, started as a branch of the MCMC community and became the center of the inferential community over time (Figure S10). Then in 2011–2015, another cluster brought insight from dynamic systems (the most frequent keyword in Figure S12) to the classic SIR model, as node 35 (boxed in blue) extends the SIR model by incorporating incubation stage and time dynamics to track the spread of flu. This innovation started a new front for the studies of this epidemic and became another center of the most influential community during 2016–2011 (Figure S10).

DISCUSSION

In this paper, we studied the citation network arising from selected statistical papers in the past two decades, a period coinciding with the rise of big data and statistics being perceived as playing increasingly important roles in many scientific disciplines. Unlike previous studies on statistics citation networks, we focused on the connections between statistics and other disciplines and used citation data to investigate the external influence of various statistical works.

First performing descriptive analysis, we showed that both the overall volume of citations and the diversity of citing fields have been increasing over time for all the journals considered. Even typical theoretical journals such as AOS have been attracting a significant proportion of external citations in recent years, which is quite encouraging. Next by distinguishing between internal and external citations, we identified research areas in statistics that have high impact under both criteria. The most highly cited papers were ranked high both internally and externally. In contrast, papers with a large number of external citations but relatively fewer internal citations can point to areas where future development in relevant theory and methods may be rewarded by immediate visibility outside statistics. Last, using the technique of local clustering, we identified the statistical research communities most relevant to various external topics of interest. Presenting a number of case studies using external topics of high general interest, we showed that the communities selected align well with our intuition and understanding of the topics.

Our study takes the first step toward understanding the influence of statistical works on other disciplines that use tools and methods from statistics to aid their discoveries. The data we have collected can be of independent interest, opening opportunities for further modeling and analysis from different perspectives. We also note that some of the limitations in our current study can be addressed by expanding the scope of the data. For example, in analyzing the trend of diversity of citing fields, it would be ideal to collect information about the number of published papers in each citing field and include it as a normalization factor. The data could also be expanded to include more journals and other types of source publications, such as conferences and books, over a longer period of time to allow for a more comprehensive historical view and richer analysis. We leave the collection and analysis of these more extensive data as future work.

Compared with global clustering, the theoretical properties of local clustering techniques are less well characterized under generative network models. We have performed theoretical analysis of our local clustering method under the degree-corrected stochastic block model (DC-SBM). We note that although the DC-SBM does not explicitly capture the acyclic structure typically present in citation networks, it is well known that such models are locally tree-like in the sparse case. Pursuing a model

Table 1. Papers with significantly different internal and external ranks

Title and reference number	Rank (internal)	Rank (external)	No. of citations
The multivariate skew-normal distribution[37]	20	177	749
A nonparametric "trim and fill" method of accounting for publication bias in meta-analysis[38]	1,520	19	1,362

Table 2. Papers whose internal and external citations both rank in the top 20

Title and reference number	Area (statistics)	Rank (internal)	Rank (external)	No. of citations
Reversible jump Markov chain Monte Carlo computation and Bayesian model determination[40]	MCMC	8	16	2,868
Identification of causal effects using instrumental variables[41]	Causal	18	17	2,125
Least angle regression[42]	Penalized regression	7	8	4,252
The control of the false discovery rate in multiple testing under dependency[43]	FDR	12	6	5,062
Model selection and estimation in regression with grouped variables[44]	Penalized regression	9	14	2,935
Regularization and variable selection via the elastic net[45]	Penalized regression	5	7	5,790
A direct approach to false discovery rates[46]	FDR	14	9	3,186
Bayesian measures of model complexity and fit[47]	Bayesian model selection	4	4	6,743
Controlling the false discovery rate: a practical and powerful approach to multiple testing[48]	FDR	2	1	46,899
Regression shrinkage and selection via the lasso[49]	Penalized regression	1	2	16,905

that directly incorporates acyclic features would be an interesting direction for future work. Our application and theoretical results of local clustering can also be extended to incorporate mixed membership modeling and temporal changes in the evolution of communities. We have currently used textual data (e.g., keywords) as a way to validate the target communities found; it would be more interesting to include such data as covariates in the network model subject to clustering analysis.

We end the discussion by acknowledging the limitations of citation itself as a form of data measuring intellectual influence, some of which have already been pointed out in previous studies.[18,19] Not all citations carry the same weight: a paper could be mentioned just in the literature review or serve as the foundation that inspired the paper citing it; arguably the latter type of citation is more important. Citations are not always attributed to the correct source, and the modern-day style of research relying on search engines such as Google is likely to bias toward papers already with high citation counts. Many data scientists and practitioners in industry do not necessarily publish their works but can still make use of ideas and tools in statistical papers, resulting in missing citations. Nevertheless, despite these limitations, citation data provide a useful and necessary first passage into investigating the intellectual influence of scientific works.

EXPERIMENTAL PROCEDURES

Resource availability

Lead contact

The lead contact for this paper is Y.X. Rachel Wang (rachel.wang@sydney.edu.au).

Materials availability

The citation data were downloaded from the WoS database.

Data and code availability

The citation data used in this article and the code for performing local clustering can be found at Zenodo: https://doi.org/10.5281/zenodo.6565329

Data collection

Using a Python script, we crawled the bibliographic database the WoS Core Collection to collect the source papers. We included only publications whose document types are listed as "article" in WoS. For each source paper, the WoS database provides a list of papers citing it and the corresponding publication information. We finished extracting these citation lists before December 2020; all the papers citing the source papers, excluding the source papers themselves, form our citing papers. Note that the citing papers are from journals other than the selected five statistics journals, or from these five journals but published in 2019 and 2020 (since papers published there before 2019 are already included in our source papers; note that the accessibility of citing papers depends on the university library VPN used to access the WoS database). Rather than limiting to "article" as we did for the source papers, the citing papers could be of any document type. Based on the lists of citations, we built the citation network. It can be represented by a binary adjacency matrix $\mathbf{A} \in \{0,1\}^{273694\times 9338}$, in which

$$A_{ij} = \begin{cases} 1, & i \text{ cites } j; \\ 0, & \text{otherwise.} \end{cases} \quad \text{(Equation 1)}$$

In this matrix, we assign each source paper to an index in $\mathcal{I}_s = \{1,\ldots,9338\}$ and each citing paper to an index in $\mathcal{I}_c = \{9339,\ldots,273694\}$. Our current study did not contain citations from the source papers to the citing papers, since we were primarily interested in the impact of source papers on other scientific works.

We obtained the publication information for both source and citing papers from the WoS database. In particular, the following variables were central to our analysis: (1) article title, (2) publication source title (e.g., journal or conference names), (3) publication year, (4) author keywords, (5) abstract, (6) WoS categories (e.g., "Statistics & Probability" and "Mathematical & Computational Biology"), and (7) research areas (e.g., "Mathematics").

Research areas for each paper

Even though the WoS categories and research areas can help us identify the research field each paper belongs to, we still had to make a decision about whether a citation should be considered inside (internal) or outside (external)

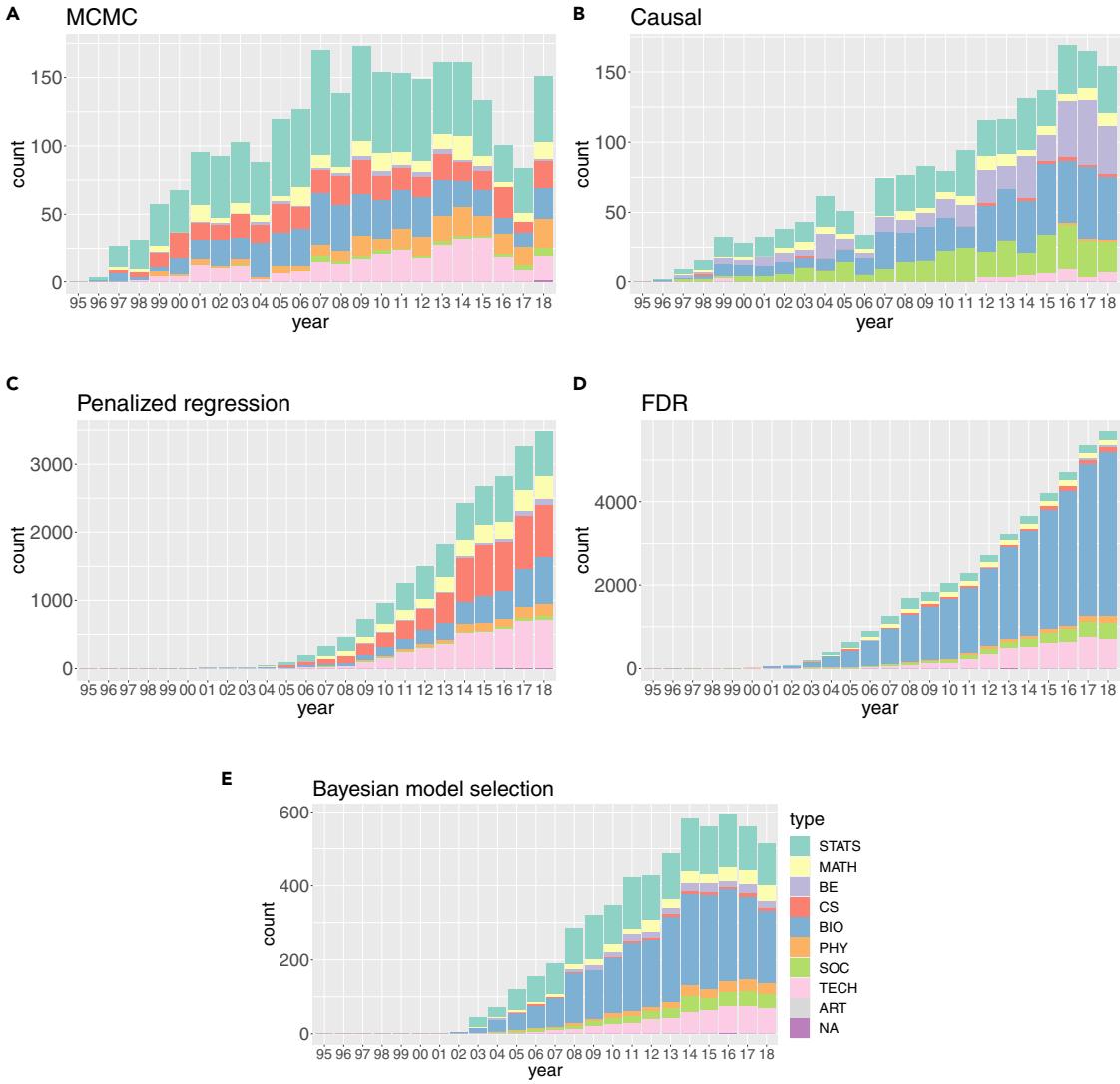

Figure 5. Breakdown of citations for the papers, whose internal and external citations both rank in the top 20, aggregated by the five statistical topics

(A) MCMC.
(B) Causal.
(C) Penalized regression.
(D) FDR.
(E) Bayesian model selection.

of statistics. This is a subjective decision in some sense, given the interdisciplinary nature of many research topics in statistics and the overlap of statistics with fields such as mathematics, computational biology, and econometrics. We took the following approach, which perhaps can be viewed as conservative in estimating external impact. We considered two types of internal papers. The first type included papers containing the tag "Statistics & Probability" in their WoS categories, which applied to all the papers published in common statistics and/or probability journals. These papers are labeled as STATS in our subsequent plots. The second type includes papers whose WoS categories contain the keyword "math" (e.g., "Mathematics" and "Mathematical & Computational Biology") while excluding STATS papers. Additional papers selected by this step were published in journals such as *Journal of Econometrics* or *BMC Bioinformatics* and thus from fields reasonably close to statistics. In what follows, these papers are labeled as MATH and counted as internal citations. The rest of the papers were considered as external. Then, we used the papers' research areas (https://images.webofknowledge.com/images/help/WOS/hp_research_areas_easca.html) provided by WoS to classify the external papers into five broad categories: arts and humanities (ART), life sciences and biomedicine (BIO), physical sciences (PHY), social sciences (SOC), and technology (TECH). In our dataset, only 98 of the papers did not have any specified categories (nor research areas), thus we labeled their categories (and research areas) as "NA." If an external paper listed multiple research areas, each area was weighted equally and contributed a fractional count to the total in Figure 3. Figure 5 considers a finer classification, including two extra categories, business and economics (BE) and computer science (CS), since we noticed a that considerable number of citations were from these two areas. To avoid double counting, papers with the BE (or CS) label were not counted in SOC (or TECH), which is the broad category BE (or CS) belongs to in WoS. Similar to before, multiple labels for one paper were weighted equally.

Algorithm 1. Local clustering

Input: adjacency matrix A, preference vector π, and teleportation constant α.

1. Compute the aPPR vector p^* in Equation 4 based on (A, π, α).
2. Construct the sequence of clusters $\{\mathcal{C}_n\}_{n=1}^{N}$ according to Equation 5 and p^*.
3. Calculate conductance values $\{\varphi(\mathcal{C}_n)\}_{n=1}^{N}$ by Equation 6.
4. Find the first local minimum $\varphi(\mathcal{C}_{n^*})$ in $\{\varphi(\mathcal{C}_n)\}_{n=1}^{N}$.

Output: local cluster $\mathcal{C}_{n^*}$.

Lorenz curve

We measured citation inequality through the Lorenz curves. For journal j, define

$$L(p) = \frac{\sum_{i=1}^{\lfloor p \times N_j \rfloor} d_{(i)}}{\sum_{i=1}^{N_j} d_{(i)}},$$

where N_j is the number of publications, p is the percentage, and $d_{(1)}, d_{(2)}, \ldots, d_{(N_j)}$ are the citation numbers in a non-decreasing order of papers in journal j published in 1995–2018. $L(p)$ calculates the percentage of citations shared by the least-cited p percent of papers as a measure of inequality. A Lorenz curve is the graph of $L(p)$.

Gini concentration index (Herfindahl index)

We measured the diversity for each journal by Gini concentration index following Stigler.[18] Let

$$\text{Gini Concentration index} = 100 \times \sum_i s_i^2,$$

where s_i is the proportion of citations from research category i; and we considered the same categories as shown in Figure 3A except that we combined STATS and MATH into one internal category. The index attains a maximum of 100 when there is only one category with proportion equaling 1 (all citations were from the internal category in our case) and decreases as the proportions become more spread out across different categories, which in turn indicates increased diversity.

Local clustering

In the following sections, we first describe the local clustering procedure in a general network setting before presenting details on how it was applied to our citation data. We also present the theoretical properties of the local clustering procedure under the DC-SBM (Karrer and Newman[59]).

DC-SBM

To analyze the behavior of local clustering, we adopted the popular DC-SBM,[59] which captures both node heterogeneity and community structure, as the underlying network model. While such a model may not capture all the features of our citation network, the presence of node heterogeneity is reflected by the uneven distribution of citation counts, and it is plausible to assume the underlying communities correspond to different research topics. For convenience of notation, we will describe the DC-SBM and local clustering procedure using a general symmetric adjacency matrix A and a general set of nodes I.

In the original SBM,[60] N nodes are assigned to K blocks or communities, and the probability of an edge between two nodes depends only on their community memberships. To abbreviate notations, write the set $\{1, \ldots, n\}$ as $[n]$ for any integer n. The set of nodes $\mathcal{I} = [N]$ is partitioned into K blocks by the function $g : [N] \to [K]$. Let n_k denote the size of block k and $\mathcal{I}_k$ denote the set of nodes in block k for $k \in [K]$. The proportion of members in block k is $\tau_k = n_k/N$. We consider the case that the number of blocks K is fixed, and τ_k is bounded below by a constant for all the $k \in [K]$. The probability of an edge between nodes i and j is:

$$A_{ij} | g \overset{\text{ind.}}{\sim} \text{Bernoulli}\left(B_{g(i)g(j)}\right), \ \forall i, j \in \mathcal{I}, \ i \neq j,$$

where $B \in [0,1]^{K \times K}$ is the connectivity matrix.

DC-SBM introduces node heterogeneity by adding a degree parameter θ_i for each node i, so that the probability of an edge between i and j becomes:

$$A_{ij} | g, \ \theta \overset{\text{ind.}}{\sim} \text{Bernoulli}\left(\theta_i \theta_j B_{g(i)g(j)}\right), \ \forall i, j \in \mathcal{I}, \ i \neq j. \qquad \text{(Equation 2)}$$

Some constraint is needed on θ_i for identifiability, and we adopt the constraint $\sum_{i \in \mathcal{I}_k} \theta_i = n_k$ for all $k \in [K]$ following Karrer and Newman.[59] The degree of node i is defined as $d_i = \sum_{j \in \mathcal{I}} A_{ij}$.

Adjusted personalized PageRank

Given a symmetric adjacency matrix A for N nodes, define the diagonal matrix $D = \text{diag}(d_1, \ldots, d_N)$, where d_i is the degree of node i, and the graph transition matrix $P = D^{-1}A$. The PPR vector $p \in [0,1]^N$ is the stationary distribution of the process:

$$p^\top = \alpha \pi^\top + (1 - \alpha) p^\top P,$$

where $\alpha \in (0, 1]$ is the teleportation constant and $\pi \in [0,1]^N$ is a probability vector called the "preference vector" encoding one or multiple seed nodes. For example, if there is one seed node $v_0 = 1$, $\pi = (1, 0, \ldots, 0)^\top$ (assuming that without loss of generality it belongs to block 1, the target block). The goal is to recover all the nodes with the same community membership as v_0 by ranking the elements in the PPR vector p.

In our setting, we chose source papers that had high citation counts by a set of topic papers as the seed nodes. For a source paper $j \in \mathcal{I}_s$ and a set of topic papers $\mathcal{I}_t$, its citation count was $a_j = \sum_{i \in \mathcal{I}_t} A_{ij}$, where $\mathbf{A}$ is the citation network defined in Equation 1. The preference vector $\pi \in [0,1]^{9338}$ was calculated as:

$$\pi_k = \frac{a\prime_k}{\sum_{j \in \mathcal{I}_s} a\prime_j} \text{ where } a\prime_j = \begin{cases} a_j, & a_j \geq t; \\ 0, & a_j < t. \end{cases} \qquad \text{(Equation 3)}$$

Here t is a chosen threshold constant. We extended the setting of a single seed node in Kloumann et al.[26] and Chen et al.[27] to multiple seed nodes, but still made the assumption that they all belong to the same community. While it is unlikely that all papers cited by a specific topic come from the same community, the threshold t helps us prune the vector π and makes the assumption more reasonable.

Related to PPR, the aPPR vector is defined as:

$$p_i^* = \frac{p_i}{d_i} \text{ for } i = 1, \ldots, N, \qquad \text{(Equation 4)}$$

Table 3. Summary statistics for the most relevant statistical communities (subnetworks) for external topics compared with the whole network for all source papers

Topic	Size	Graph density	Average clustering coefficient
Single-cell transcriptomics	79	0.031	0.608
Labor economics	108	0.039	0.402
Flu	30	0.73	0.232
All source papers	9,338	0.001	0.252

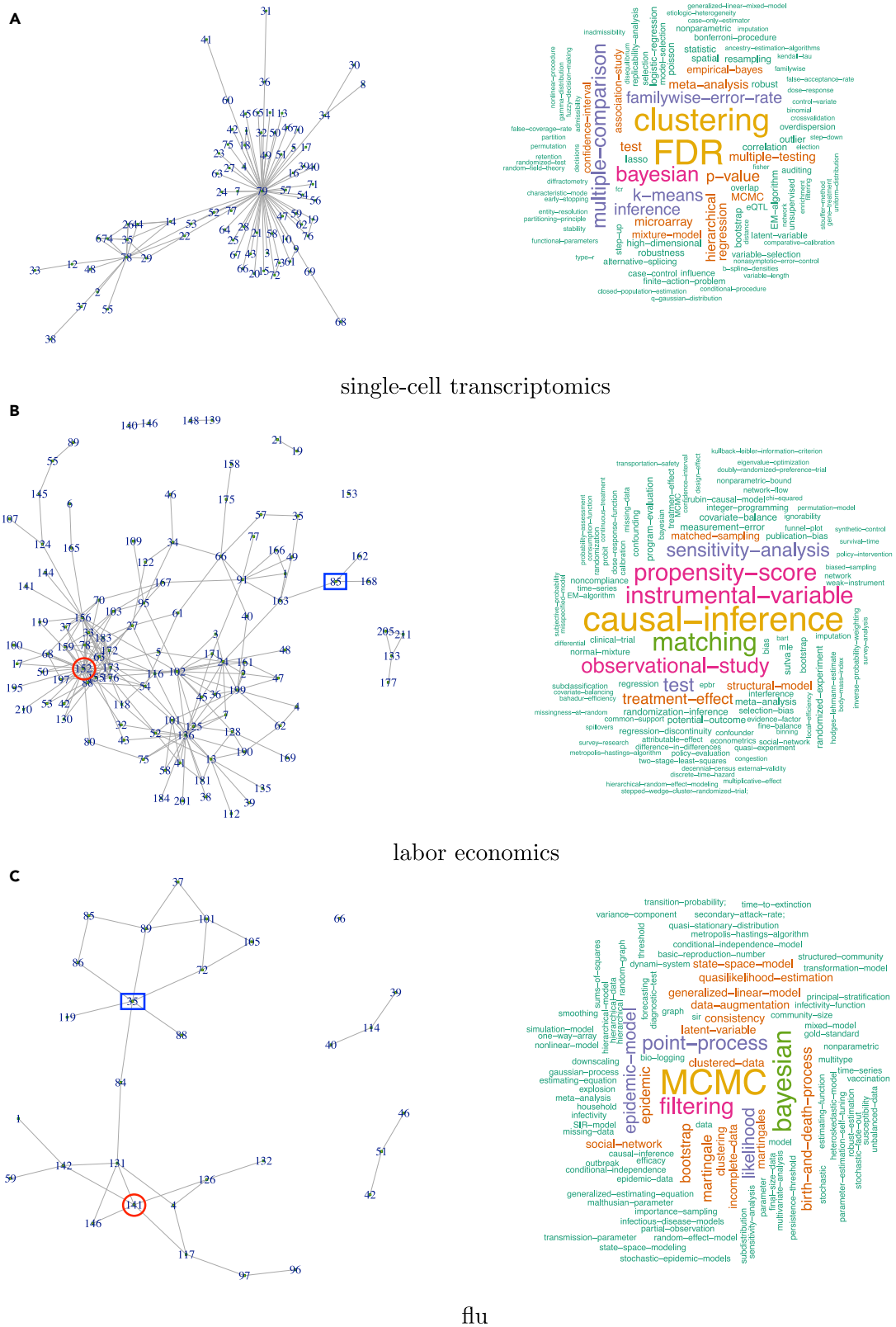

Figure 6. Networks and word clouds generated from the source papers found by local clustering for three topics

(A) Single-cell transcriptomics.

(B) Labor economics.

(C) Flu.

where p_i is the ith entry in the PPR vector. Formally, let n be a community size cutoff. Then n nodes with the largest p_i^* values are selected as members in the target community, that is,

$$\mathcal{C}_n = \left\{ i \middle| p_i^* \geq p_{(n)}^* \right\}, \quad \text{(Equation 5)}$$

where $p_{(1)}^*, \ldots, p_{(N)}^*$ is the sorted list of p^* in a non-increasing order. We show that the aPPR vector sorted the nodes in terms of their relevance to the target community with high probability under the DC-SBM in Note S3.1.

Conductance

It remains to choose the correct size n for $\mathcal{C}_n$ to fully recover the target community. To achieve this, an objective function is needed to evaluate the quality of the clusters found. Conductance is a popular objective function to be optimized, either globally or locally,[29,30] and is often used in conjunction with a local clustering algorithm like PPR.[21,61] It tends to favor small clusters weakly connected to the rest of the graph, and one would expect such an assortative structure in citation networks with communities defined by research topics.

For a set of nodes $\mathcal{I}' \subseteq \mathcal{I}$, where $\mathcal{I}$ denotes all the nodes in A, we define its conductance φ as

$$\varphi(\mathcal{I}') = \frac{\sum_{i \in \mathcal{I}'} \sum_{j \notin \mathcal{I}'} A_{ij}}{\sum_{i \in \mathcal{I}'} A_{i\cdot}}, \quad \text{(Equation 6)}$$

where $A_{i\cdot} = \sum_{j \in \mathcal{I}} A_{ij}$.

Using aPPR to sort the nodes in terms of their relevance to the target community, the sorted list of nodes leads to a sequence of clusters $\{\mathcal{C}_n\}_{n=1}^N$ (by Equation 5) and their conductance values $\{\varphi(\mathcal{C}_n)\}_{n=1}^N$. Our next theorem establishes that the correct choice of n occurs at a local optimum along this sequence, justifying the practice of choosing the community size cutoff by inspecting the conductance plot.

Theorem 1

Under the DC-SBM and appropriate assumptions, for sufficiently large N, there exits n' with $n' - n_1 = \Omega(N)$ such that

$$\varphi(\mathcal{C}_{n_1}) - \varphi(\mathcal{C}_n) \leq -\frac{1}{N}\Omega_P(|n - n_1|) \quad \text{(Equation 7)}$$

uniformly for $n \in [n']$, where n_1 is the size of the target block.

The details of Theorem 1 (assumptions and proof) are presented in Note S3.2. Here the Ω notation indicates that $n' - n_1$ is bounded below by a constant order of N; Ω_P indicates with high probability the difference between $\varphi(\mathcal{C}_{n_1})$ (the local optimum value) and any surrounding value $\varphi(\mathcal{C}_n)$ is bounded from below by a constant order of $|n - n_1|/N$. The bound in Equation 7 and the lower bound on $n' - n_1$ guarantee that the optimum at n_1 is well separated from its neighborhood, and this neighborhood is wide enough to be observed in a conductance plot. In Note S3.3 (Tables S10–S11 and Figures S17–S20), we demonstrate the performance of Algorithm 1 in recovering the target community using simulated data and examine its robustness with respect to the number of seeds and the teleportation constant.

Our local clustering procedure is summarized in Algorithm 1.

Applying Algorithm 1 to citation data

For all the case studies, we first used a keyword search to select a set of topic papers $\mathcal{I}_t$, for an external topic of interest. The seed nodes were constructed using citation information between the source papers $\mathcal{I}_s$ and the topic papers in $\mathcal{I}_t$ as described in Equation 3, and Algorithm 1 was performed on $\mathcal{I}_s$ and their network $\mathbf{A}^s$. For clustering purpose, we considered two papers as related in content if a citation existed between them; the direction of this citation was less important if we thought of it as a form of association. For this reason, we treated $\mathbf{A}^s$ as an undirected network in this section. That is,

$$A_{ij}^s = \begin{cases} 1, & \text{there is a citation between } i \text{ and } j; \\ 0, & \text{otherwise} \end{cases}, \quad \text{(Equation 8)}$$

for $i, j \in \mathcal{I}_s = \{1, \ldots, 9338\}$.

We set the teleportation constant $\alpha = 0.15$ following Chen et al.[27]. In general, we found our procedure to be robust to the choice of α, and detailed simulation studies can be found in Note S3.3.

It remains to describe the construction of the preference vector π, which relied on the selection of topic papers $\mathcal{I}_t$. For each external topic, papers in $\mathcal{I}_t$ were chosen by keyword searches among all the papers. More concretely, for the topics of single-cell transcriptomics and labor economics, we found all the papers that contain the relevant keywords ("Single-cell" [or "single cell"] and "RNA-seq" for the topic of single-cell; "labor" for the topic of labor economics) in their abstracts. For a more accurate search result, we further restricted the labor economics papers to the category SOC using the labels. The single-cell papers could come from a more diverse set of categories, and as shown in Figure S13A, most of our selected papers were from BIO. For the topic of flu, we noted that many papers may use flu datasets as examples of their analytic methods instead of focusing on the topic itself. To select papers with a sharper focus on the topic, we searched for papers with "flu" or "influenza" in their title instead of their abstract. The proportions of category labels for the flu papers are illustrated in Figure S13B, which indicates that most of them were from BIO. Having constructed $\mathcal{I}_t$, we chose the seed nodes in π from the source papers with high citation counts by $\mathcal{I}_t$. For each topic, we constructed the preference vector by Equation 3. We chose the threshold t based on the citation counts from the topic papers to the source papers. For the topics "single-cell" and "labor economics," the top papers received more than 90 citations; we set $t = 10$. For the topic "flu," the highest citation count was less than 90, and we set $t = 5$. The conductance plot for each topic is shown in Figure S14. In most cases, there was an obvious local minimum leading to a reasonable community size. In Figure S14B, we chose the first minimum occurring after $n \geq 10$ for a more plausible subnetwork size and clearer interpretation of the result.

For the temporal studies of labor economics and flu topics, we broke our study period into five windows: 1995–2000, 2001–2005, 2006–2010, 2011–2015, and 2016–2021. For each window, the construction of seed nodes used only papers from that time period; the adjacency matrix A, which is the input in Algorithm 1, contained citation information among papers only within or before that period. Then, we plotted the network of the influential community for each time window in Figures S9 and S10. For selected subnetworks in these figures, we present their word clouds in Figures S11 and S12.

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.patter.2022.100532.

ACKNOWLEDGMENTS

The authors would like to thank Dr. Tung-Yu Wu for help with the data collection process and Prof. Peter J. Bickel and Prof. Jingyi Jessica Li for many fruitful discussions. Y.X.R.W. gratefully acknowledges funding from the Australian Research Council DECRA Fellowship (DE180101252).

AUTHOR CONTRIBUTIONS

X.T. and Y.X.R.W. conceived the study and designed the data collection procedures and methods. L.W. carried out the methods, wrote the code, and performed empirical and theoretical analysis. X.T. and Y.X.R.W. supervised the execution. All authors contributed to the writing of the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: November 25, 2021
Revised: April 25, 2022
Accepted: May 25, 2022
Published: June 16, 2022

REFERENCES

1. Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., Hung Byers, A., et al. (2011). Big Data: The Next Frontier for Innovation, Competition, and Productivity (McKinsey Global Institute).
2. Provost, F., and Fawcett, T. (2013). Data science and its relationship to big data and data-driven decision making. Big Data *1*, 51–59. https://doi.org/10.1089/big.2013.1508.
3. Jordan, J.M., and Lin, D.K. (2014). Statistics for big data: are statisticians ready for big data? International Chinese Statistical Association Bulletin *52*, 133–149.
4. Malley, J.D., and Moore, J.H. (2013). The Disconnect between Classical Biostatistics and the Biological Data Mining Community.
5. Shi, F., Foster, J.G., and Evans, J.A. (2015). Weaving the fabric of science: dynamic network models of science's unfolding structure. Soc. Network. *43*, 73–85. https://doi.org/10.1016/j.socnet.2015.02.006.
6. Varga, A. (2019). Shorter distances between papers over time are due to more cross-field references and increased citation rate to higher-impact papers. Proc. Natl. Acad. Sci. Unit. States Am. *116*, 22094–22099. https://doi.org/10.1073/pnas.1905819116.
7. Rinia, E.J., Van Leeuwen, T.N., Bruins, E.E., Van Vuren, H.G., and Van Raan, A.F. (2001). Citation delay in interdisciplinary knowledge exchange. Scientometrics *51*, 293–309. https://doi.org/10.1023/a:1010589300829.
8. Van Leeuwen, T., and Tijssen, R. (2000). Interdisciplinary dynamics of modern science: analysis of cross-disciplinary citation flows. Res. Eval. *9*, 183–187. https://doi.org/10.3152/147154400781777241.
9. Van Noorden, R. (2015). Interdisciplinary research by the numbers. Nature *525*, 306–307. https://doi.org/10.1038/525306a.
10. Steele, T.W., and Stier, J.C. (2000). The impact of interdisciplinary research in the environmental sciences: a forestry case study. J. Am. Soc. Inf. Sci. *51*, 476–484. https://doi.org/10.1002/(sici)1097-4571(2000)51:5<476::aid-asi8>3.0.co;2-g.
11. Mansilla, V.B., Feller, I., and Gardner, H. (2006). Quality assessment in interdisciplinary research and education. Res. Eval. *15*, 69–74. https://doi.org/10.3152/147154406781776057.
12. Stirling, A. (2007). A general framework for analysing diversity in science, technology and society. J. R. Soc. Interface *4*, 707–719. https://doi.org/10.1098/rsif.2007.0213.
13. Kajikawa, Y., and Takeda, Y. (2008). Structure of research on biomass and bio-fuels: a citation-based approach. Technol. Forecast. Soc. Change *75*, 1349–1359. https://doi.org/10.1016/j.techfore.2008.04.007.
14. Dias, L., Gerlach, M., Scharloth, J., and Altmann, E.G. (2018). Using text analysis to quantify the similarity and evolution of scientific disciplines. R. Soc. Open Sci. *5*, 171545. https://doi.org/10.1098/rsos.171545.
15. Levitt, J.M., and Thelwall, M. (2008). Is multidisciplinary research more highly cited? a macrolevel study. J. Am. Soc. Inf. Sci. Technol. *59*, 1973–1984. https://doi.org/10.1002/asi.20914.
16. Larivière, V., and Gingras, Y. (2010). On the relationship between interdisciplinarity and scientific impact. J. Am. Soc. Inf. Sci. Technol. *61*, 126–131. https://doi.org/10.1002/asi.21226.
17. Yegros-Yegros, A., Rafols, I., and D'este, P. (2015). Does interdisciplinary research lead to higher citation impact? the different effect of proximal and distal interdisciplinarity. PLoS One *10*, e0135095. https://doi.org/10.1371/journal.pone.0135095.
18. Stigler, S.M. (1994). Citation patterns in the journals of statistics and probability. Stat. Sci. *9*, 94–108. https://doi.org/10.1214/ss/1177010655.
19. Varin, C., Cattelan, M., and Firth, D. (2016). Statistical modelling of citation exchange between statistics journals. J. Roy. Stat. Soc. *179*, 1–63. https://doi.org/10.1111/rssa.12124.
20. Ji, P., and Jin, J. (2016). Coauthorship and citation networks for statisticians. Ann. Appl. Stat. *10*, 1779–1812. https://doi.org/10.1214/15-aoas896.
21. Andersen, R., and Lang, K.J. (2006). Communities from seed sets. In Proceedings of the 15th international conference on World Wide Web, pp. 223–232.
22. Whang, J.J., Gleich, D.F., and Dhillon, I.S. (2013). Overlapping community detection using seed set expansion. In Proceedings of the 22nd ACM international conference on information & knowledge management, pp. 2099–2108.
23. Kloumann, I.M., and Kleinberg, J.M. (2014). Community membership identification from small seed sets. In Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 1366–1375.
24. Chung, F. (2009). A local graph partitioning algorithm using heat kernel pagerank. Internet Math. *6*, 315–330. https://doi.org/10.1080/15427951.2009.10390643.
25. Kloster, K., and Gleich, D.F. (2014). Heat kernel based community detection. In Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 1386–1395.
26. Kloumann, I.M., Ugander, J., and Kleinberg, J. (2017). Block models and personalized pagerank. Proc. Natl. Acad. Sci. Unit. States Am. *114*, 33–38. https://doi.org/10.1073/pnas.1611275114.
27. Chen, F., Zhang, Y., and Rohe, K. (2020). Targeted sampling from massive block model graphs with personalized pagerank. J. Roy. Stat. Soc. B *82*, 99–126. https://doi.org/10.1111/rssb.12349.
28. Andersen, R., Chung, F., and Lang, K. (2006). Local graph partitioning using pagerank vectors. In 2006 47th Annual IEEE Symposium on Foundations of Computer Science (FOCS'06) (IEEE), pp. 475–486.
29. Yang, J., and Leskovec, J. (2015). Defining and evaluating network communities based on ground-truth. Knowl. Inf. Syst. *42*, 181–213. https://doi.org/10.1007/s10115-013-0693-z.
30. Van Laarhoven, T., and Marchiori, E. (2016). Local network community detection with continuous optimization of conductance and weighted kernel k-means. J. Mach. Learn. Res. *17*, 5148–5175.
31. Chen, P., Xie, H., Maslov, S., and Redner, S. (2007). Finding scientific gems with google's pagerank algorithm. Journal of Informetrics *1*, 8–15. https://doi.org/10.1016/j.joi.2006.06.001.
32. Ma, N., Guan, J., and Zhao, Y. (2008). Bringing pagerank to the citation analysis. Inf. Process. Manag. *44*, 800–810. https://doi.org/10.1016/j.ipm.2007.06.006.
33. Zeng, A., Shen, Z., Zhou, J., Wu, J., Fan, Y., Wang, Y., and Stanley, H.E. (2017). The science of science: from the perspective of complex systems. Phys. Rep. *714-715*, 1–73. https://doi.org/10.1016/j.physrep.2017.10.001.
34. Walker, D., Xie, H., Yan, K.-K., and Maslov, S. (2007). Ranking scientific publications using a model of network traffic. J. Stat. Mech. Theor. Exp. *2007*, P06010. https://doi.org/10.1088/1742-5468/2007/06/p06010.
35. Su, C., Pan, Y., Zhen, Y., Ma, Z., Yuan, J., Guo, H., Yu, Z., Ma, C., and Wu, Y. (2011). Prestigerank: a new evaluation method for papers and journals. Journal of Informetrics *5*, 1–13. https://doi.org/10.1016/j.joi.2010.03.011.
36. Zhou, J., Zeng, A., Fan, Y., and Di, Z. (2016). Ranking scientific publications with similarity-preferential mechanism. Scientometrics *106*, 805–816. https://doi.org/10.1007/s11192-015-1805-1.
37. Azzalini, A., Valle, A.D., and AzzAlini, A. (1996). The multivariate skew-normal distribution. Biometrika *83*, 715–726. https://doi.org/10.1093/biomet/83.4.715.
38. Duval, S., and Tweedie, R. (2000). A nonparametric "trim and fill" method of accounting for publication bias in meta-analysis. J. Am. Stat. Assoc. *95*, 89. https://doi.org/10.2307/2669529.
39. Lo, Y., Mendell, N.R., and Rubin, D.B. (2001). Testing the number of components in a normal mixture. Biometrika *88*, 767–778. https://doi.org/10.1093/biomet/88.3.767.
40. Green, P.J. (1995). Reversible jump Markov chain Monte Carlo computation and bayesian model determination. Biometrika *82*, 711–732. https://doi.org/10.1093/biomet/82.4.711.

41. Angrist, J.D., Imbens, G.W., and Rubin, D.B. (1996). Identification of causal effects using instrumental variables. J. Am. Stat. Assoc. *91*, 444–455. https://doi.org/10.1080/01621459.1996.10476902.

42. Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. Ann. Stat. *32*, 407–499. https://doi.org/10.1214/009053604000000067.

43. Benjamini, Y., and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. Ann. Stat. 1165–1188. https://doi.org/10.1214/aos/1013699998.

44. Yuan, M., and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. J. Roy. Stat. Soc. B *68*, 49–67. https://doi.org/10.1111/j.1467-9868.2005.00532.x.

45. Zou, H., and Hastie, T. (2005). Regularization and variable selection via the elastic net. J. Roy. Stat. Soc. B *67*, 301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x.

46. Storey, J.D. (2002). A direct approach to false discovery rates. J. Roy. Stat. Soc. B *64*, 479–498. https://doi.org/10.1111/1467-9868.00346.

47. Spiegelhalter, D.J., Best, N.G., Carlin, B.P., and Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. J. Roy. Stat. Soc. B *64*, 583–639. https://doi.org/10.1111/1467-9868.00353.

48. Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. Roy. Stat. Soc. B *57*, 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x.

49. Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. J. Roy. Stat. Soc. B *58*, 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x.

50. Kharchenko, P.V. (2021). The triumphs and limitations of computational methods for scrna-seq. Nat. Methods *18*, 723–732. https://doi.org/10.1038/s41592-021-01171-x.

51. Sugar, C.A., and James, G.M. (2003). Finding the number of clusters in a dataset: an information-theoretic approach. J. Am. Stat. Assoc. *98*, 750–763. https://doi.org/10.1198/016214503000000666.

52. Tibshirani, R., Walther, G., and Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. J. Roy. Stat. Soc. B *63*, 411–423. https://doi.org/10.1111/1467-9868.00293.

53. Hunter, D.R., Goodreau, S.M., and Handcock, M.S. (2008). Goodness of fit of social network models. J. Am. Stat. Assoc. *103*, 248–258. https://doi.org/10.1198/016214507000000446.

54. Aronow, P.M., and Samii, C. (2017). Estimating average causal effects under general interference, with application to a social network experiment. Ann. Appl. Stat. *11*, 1912–1947. https://doi.org/10.1214/16-aoas1005.

55. Ganong, P., and Jäger, S. (2018). A permutation test for the regression kink design. J. Am. Stat. Assoc. *113*, 494–504. https://doi.org/10.1080/01621459.2017.1328356.

56. Li, F., Morgan, K.L., and Zaslavsky, A.M. (2018). Balancing covariates via propensity score weighting. J. Am. Stat. Assoc. *113*, 390–400. https://doi.org/10.1080/01621459.2016.1260466.

57. Britton, T. (1998). Estimation in multitype epidemics. J. Roy. Stat. Soc. B *60*, 663–679. https://doi.org/10.1111/1467-9868.00147.

58. Dukic, V., Lopes, H.F., and Polson, N.G. (2012). Tracking epidemics with google flu trends data and a state-space seir model. J. Am. Stat. Assoc. *107*, 1410–1426. https://doi.org/10.1080/01621459.2012.713876.

59. Karrer, B., and Newman, M.E.J. (2011). Stochastic blockmodels and community structure in networks. Physical review E *83*, 016107. https://doi.org/10.1103/physreve.83.016107.

60. Holland, P.W., Laskey, K.B., and Leinhardt, S. (1983). Stochastic blockmodels: first steps. Soc. Network. *5*, 109–137. https://doi.org/10.1016/0378-8733(83)90021-7.

61. Wu, X.-M., Li, Z., So, A.M.-C., Wright, J., and Chang, S.-F. (2012). Learning with partially absorbing random walks. NIPS (News Physiol. Sci.) *25*, 3077–3085.