




The Landscape of Causal Inference: Perspective From Citation Network Analysis

Weihua An & Ying Ding

To cite this article: Weihua An & Ying Ding (2018) The Landscape of Causal Inference: Perspective From Citation Network Analysis, The American Statistician, 72:3, 265-277, DOI: [10.1080/00031305.2017.1360794](https://doi.org/10.1080/00031305.2017.1360794)

To link to this article: <https://doi.org/10.1080/00031305.2017.1360794>




View supplementary material 



Accepted author version posted online: 01 Aug 2017.
Published online: 08 Jun 2018.



Submit your article to this journal 





Article views: 756



View Crossmark data 



The Landscape of Causal Inference: Perspective From Citation Network Analysis

Weihua An ^a and Ying Ding ^b

^aDepartment of Sociology and Institute for Quantitative Theory and Methods, Emory University, Atlanta, GA; ^bSchool of Informatics and Computing, Indiana University, IN

ABSTRACT

Causal inference is a fast-growing multidisciplinary field that has drawn extensive interests from statistical sciences and health and social sciences. In this article, we gather comprehensive information on publications and citations in causal inference and provide a review of the field from the perspective of citation network analysis. We provide descriptive analyses by showing the most cited publications, the most prolific and the most cited authors, and structural properties of the citation network. Then, we examine the citation network through exponential random graph models (ERGMs). We show that both technical aspects of the publications (e.g., publication length, time and quality) and social processes such as homophily (the tendency to cite publications in the same field or with shared authors), cumulative advantage, and transitivity (the tendency to cite references' references), matter for citations. We also provide specific analysis of citations among the top authors in the field and present a ranking and clustering of the authors. Overall, our article reveals new insights into the landscape of the field of causal inference and may serve as a case study for analyzing citation networks in a multidisciplinary field and for fitting ERGMs on big networks. Supplementary materials for this article are available online.

ARTICLE HISTORY

Received February 2017
Revised July 2017

KEYWORDS

Big network; Causal inference; Citation network; ERGM

1. Introduction


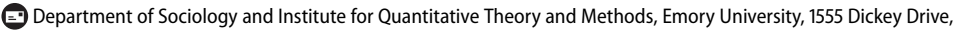
Causal inference is a multidisciplinary field that has grown rapidly in the past few decades (Morgan and Winship 2015). The specific methods include randomization test (Basu 1980; Rosenbaum 2002; Small, Ten Have, and Rosenbaum 2008; Ding, Feller, and Miratrix 2015), matching (Abadie et al. 2004; Abadie and Imbens 2006; Abadie and Imbens 2011), propensity score methods (Rosenbaum and Rubin 1983; Imbens 2000; Hirano, Imbens, and Ridder 2003; Imai and Dyk 2004; An 2010; Abadie and Imbens 2016), marginal structural models (Robins, Hernan, and Brumback 2000; Cole and Hernán 2008), instrumental variables (Imbens and Angrist 1994; Angrist, Imbens, and Rubin 1996; Abadie 2003; Hernán and Robins 2006; Small 2007; Small and Rosenbaum 2008; Okui et al. 2012; O'Malley et al. 2014; Baiocchi, Cheng, and Small 2014; Small et al. 2014; An 2015; An and Wang 2016; Ding, VanderWeele, and Robins 2017), directed acyclic graphs (Greenland, Pearl, and Robins 1999; Pearl 2000; Elwert 2013; Ogburn et al. 2014), etc. Comprehensive reviews can be found in Ho et al. (2007), Imai, King, and Stuart (2008), Angrist and Pischke (2009), Imbens and Wooldridge (2009), Rosenbaum (2010), Stuart (2010), Barringer, Eliason, and Leahy (2013), Imbens and Rubin (2015), Morgan and Winship (2015), Hu and Mustillo (2016), and Hernán and Robins (2018). In this article, we attempt to provide a review of the field of causal inference from the perspective of citation network analysis.

Prior citation network analyses have studied citations across journals (Stigler 1994; Peng 2015; Varin, Cattelan, and Firth 2016) or publications in a discipline (Ji and Jin 2016), but have


rarely considered citation networks in a multidisciplinary field that mixes researchers from different backgrounds and involves unique dynamics (Panofsky 2011; Wagner et al. 2011). In this article, we use key terminologies in causal inference to identify relevant publications and citations. Relatively speaking, our data are more comprehensive, covering significantly more journals (including all the journals indexed in *Web of Science* core collections) and a longer time period (1905–2014). More importantly, in this article we provide a rigorous study of citation formation mechanisms. From each publication record, we extract comprehensive information on author and publication characteristics and employ it in exponential random graph models (ERGMs; Hunter and Handcock 2006; Robins et al. 2007b; Robins, Pattison, and Wang 2009) to study citation formations. In our models we account for not only covariate effects, but also complex social processes such as homophily, transitivity, and preferential attachment. We also provide sociological interpretations of the results.

This article may also serve as a case study for how to estimate ERGMs on big networks. ERGMs are slow and difficult to fit on large networks, due to reliance on Monte Carlo Markov Chains (MCMCs) for estimation (Hunter and Handcock 2006). In this study, we employ several strategies to address the estimation problems, including multiple rounds of data-refinement to filter core publications and use of a local approximation to ERGMs.

This article proceeds as follows. In Section 2, we discuss past citation network studies and point out our contributions to the

CONTACT Weihua An  weihua.an@emory.edu 

Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/r/TAS.

 Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/TAS

© 2018 American Statistical Association

literature. In Section 3, we describe the data and the methods we used to analyze the data. In Section 4, we present the results and analyses. Finally, we conclude and discuss implications and limitations of this study.

2. Citation Network Analysis

Network methods have been used in the past to study scientific collaborations and structure of science (Newman 2001; Moody 2004; Shi, Foster, and Evans 2015). In particular, citation networks have been analyzed to study information diffusion (Yan et al. 2013) and scholarly impact (Stigler 1994; Walker et al. 2007; Ding 2011; Ji and Jin 2016; Varin, Cattelan, and Firth 2016). Most prior citation network studies are focused on providing descriptive analyses.

In this study, besides providing detailed descriptive analyses, we also employ ERGMs to investigate how technical features of the publications and social processes affect citation formations. First, we model the effects of a list of covariates (including publication and author characteristics) we extracted from publication records, whereas prior work mostly employed only a few covariates and often did not distinguish their effects on receiving and sending citations. We provide theoretical expectations on the effects of the covariates and present how publication and author characteristics can matter for citations (shown in the next section). Second, our analyses account for homophilous processes in citation formations (McPherson, Smith-Lovin, and Cook 2001). We expect that citations tend to form between publications that are in the same field and publications that share authors.

Third, we also examine multiple endogenous network formation processes in citations. Similar to prior work (Peng 2015), we examine transitivity in citations (i.e., if A cites B and B cites C, then A is more like to cite C). We argue that transitivity in citations can occur because researchers may use a snowball strategy to discover new studies by following the references of other studies. We also examine preferential attachment or cumulative advantage in citations (Merton 1968; Allison 1980), namely, publications that have been cited more tend to receive even more citations over time. We expect some publications will receive a lot more citations than others and so the variation in the number of received citations will be large. At the same time, the variation in the outgoing citations may be small. In addition, our study accounts for a special feature in citation: no forward referencing, namely, earlier publications cannot cite later ones.

Finally, prior work tends to study only citation networks among publications or journals. In this article, we also study the citation network among authors where a tie indicates the number of times one author has cited another. We use the author citation network mainly for two purposes. One is to rank authors based on their positions in the author citation network. The other is to cluster authors into groups, which helps delineate the different communities in causal inference research. In this study, we are particularly interested in the interactions (as indicated by citations) among top authors in the field, mostly because these interactions will likely have the largest impact on the development of the field.

Using ERGMs to study citation networks also poses challenges. As is known, ERGMs are difficult to fit on big networks.

First, to load data and carry out computations with big networks requires a big computer RAM (in our case, around 250 GB). We use supercomputers to solve the computer memory issue. Second, although we are able to fit a simple ERGM that include only covariate effects on the full citation network, we encounter problems in fitting ERGMs that can account for endogenous network formation processes. This is because such ERGMs rely on Monte Carlo Markov chains to simulate networks for estimations. For big networks involving thousands of nodes and edges such as the one studied in this article, the estimation process can be very slow (taking days or weeks). To address the computation speed issue, we fit the ERGMs that can account for endogenous network formation processes in addition to covariate effects only on the core citation network in which isolates (i.e., publications that do not cite or get cited by any other publications) are removed (Goodreau et al. 2008). The results based on the core network can be interpreted as capturing citation patterns among the core publications. One caveat is that some of the results may not be generalizable to the full citation network.

3. Data and Analytical Strategies

3.1. Data

We use the *Web of Science* core collections to obtain the publication records in causal inference. The *Web of Science*, maintained by Thomson Reuters, provides a comprehensive, multidisciplinary indexing service, with coverage from the early 20th century until the present, which therefore supports in-depth studies of scientific fields (Drake 2004). The *Web of Science* core collections include the following databases.

1. Science Citation Index Expanded (1955-present)
2. Social Sciences Citation Index (1900-present)
3. Arts & Humanities Citation Index (1975-present)
4. Book Citation Index Science (2005-present)
5. Book Citation Index Social Sciences & Humanities (2005-present)

We use a series of key terms in causal inference to identify records of interest. If any of the key terms appears in the title, abstract, author, or keywords of a record, the record is identified as relevant. The key terms include: (1) causal inference, (2) propensity score, (3) potential outcome(s), (4) causality, (5) counterfactual(s), (6) causation, and (7) causal effect(s). The search identifies 158,663 raw records. Including only standard publications like articles, reviews, and proceeding papers reduces the number of records to 152,419.

Our browsing of the remaining records suggests that many of the records are not tightly related to causal inference methods and their applications. Also to obtain a manageable size of records that is central to the field, we further refine the dataset by restricting it to a number of research areas that are known for their interests and activities in causal inference, including (1) biomedical social sciences; (2) social work; (3) family studies; (4) public environmental occupational health; (5) business economics; (6) social sciences other topics; (7) health care sciences services; (8) mathematical computational biology; (9) operations research and management science; (10) government and

law; (11) information science and library science; (12) medical laboratory technology; (13) medical informatics; (14) mathematics; (15) behavioral sciences; (16) public administration; (17) mathematical methods in social sciences; (18) education and educational research; (19) social issues; and (20) sociology. Such refinement reduces the number of records to 33,306, which can be viewed as representing the core publications in causal inference. These publications cite 1,491,743 references in total. A significant portion of the references are outside of the field of causal inference. Since we are mostly interested in the citation patterns within the field of causal inference, we keep only those references that include the key search terms. In the end, we identify 33,306 core publications and 8442 citations/references.

For each publication, we are able to extract and construct a series of covariates. These covariates are used to model a series of possible mechanisms for citation formations. For example, researchers may cite others' work because the cited research has a higher quality (being original and innovative) and is more visible (easier to access and published earlier). Researchers also may use citations to legitimize their own research, for example, by citing research from prestigious institutions. As shown below, each of the mechanisms may be measured by several different covariates, and each of the covariates may measure different aspects of the mechanisms. Thus, it is difficult to cleanly separate the mechanisms and the contributions of the covariates to the mechanisms. To facilitate analyses and interpretations of the results, we divide the covariates into author characteristics and publication attributes. Below we briefly outline our expectations for their contributions to citation formations. The author characteristics include the following variables.

1. Whether the publication receives any funding (binary variable, yes = 1; no = 0). We expect publications receiving funding support are more likely to be cited. This may be because the authors of these publications have more resources to refine their work or because the funding has served as a screening mechanism to select quality research. Similarly, authors with funding support may be more likely to cite others because they have more resources to engage with the academic community.
2. Whether the corresponding author is from the U.S. (binary variable, yes = 1; no = 0). As many of the pioneer studies in causal inference have been done in the U.S. and publications by U.S.-based authors may have higher visibility than their counterparts, we expect publications whose main author is from the U.S. are more likely to be cited. At the same time, because U.S. authors may be more familiar with the field, we also expect them to cite more works.
3. Whether the first author has a short last name (i.e., with no more than four letters) (binary variable, yes = 1; no = 0). Counting such occurrence only in the first authors is shown to be less problematic when it is used as the independent variable (Long, McGinnis, and Allison 1980). We use this variable to approximate the race of the first author of the publication, because presumably East Asian (particularly Chinese and Korean) researchers tend to have shorter last names. Since it is very difficult to identify the race of the authors, this approach, although imperfect, offers a practical solution to studying the

citation patterns associated with race.¹ We expect that the publications whose first author has a short last name are less likely to be cited or to cite others, possibly because this group includes relatively more graduate students and also because the work of this group (especially those abroad) may be less integrated with the mainstream literature.

4. Whether any of the authors is from the top 100 universities (binary variable, yes = 1; no = 0).² We use this variable to approximate the prestige of the authors' affiliations. We expect publications with authors from prestigious universities are more likely to be cited. Also, researchers at more prestigious universities may be more informed of the field and so are more likely to cite others' research.
5. Number of authors (count variable). In general, we expect publications with more authors are more likely to cite and to be cited by others, simply because more authors may help produce more innovative work and increase the visibility of the publications. At the same time, because of the uniqueness of statistical research, methodological breakthroughs tend to be obtained by the most talented and prepared, so significant publications tend to have fewer authors. Therefore, we may also expect that publications with many authors tend to be less innovative and consequentially, to be cited less. These two mechanisms are probably both at play. So, it is an empirical question to see which mechanism dominates.

The publication characteristics include the following variables.

1. Publication format (binary variable, article = 1; others = 0). As original research tends to be presented in the article format, we expect that papers receive more citations than other types of publications. Also, we expect papers to cite less than other types of publications (e.g., reviews).
2. Publication page count (continuous variable). We expect that longer publications naturally tend to cite more. As longer publications also tend to be reviews or contain original research, we also expect that they will be cited more by others.
3. Number of keywords (count variable). Publications with more keywords are more likely to engage with various topics in causal inference. Thus, we expect them to both receive and send more citations.
4. Whether the publication is in a method journal (binary variable, yes = 1; no = 0). We expect method papers to be cited more but not necessarily to cite more, because methodological innovations may be used in many applied studies while the innovations do not have to rely a lot on prior work.

¹ To assess the quality of this name measure, we extract a random sample of 300 records. We find 34 common east Asian names, among which 24 (71%) are correctly coded as ones while the other 10 are incorrectly coded as zeros. Among the 266 non-East Asian names, 246 (92.4%) are correctly coded as zeros. In general, the measurement error tends to bias the estimated effect of this variable toward zero.

² We use the Times Higher Education World University Ranking, because of its comprehensive coverage of universities in the world. The data are available at <http://www.timeshighereducation.co.uk/world-university-rankings/2014-15/world-ranking>.

5. Whether the publication is in the top 30 journals with the most publications in causal inference (binary variable, yes = 1; no = 0). We use this variable to approximate the prestige and the visibility of the publication. We expect that publications in top outlets will be cited more and may also cite more (partly because the publications need to entertain a wider audience).
6. Publication year. As research published recently has less time to be recognized, we expect it to receive fewer citations. Also, as research published recently has more prior work to cite from, we expect it to cite more.

For computational convenience and better interpretability of the results, we binarize selected variables (i.e., number of authors, publication length, number of keywords, and publication year). As the distributions of these variables are right-skewed, we binarize the variables at their mean instead of median so that only an appropriately small fraction of publications are identified as possessing certain salient features (i.e., long publications).

To represent assortative mixing mechanisms and forward referencing in citation formations, we create three dyadic variables. (1) Whether two publications are in the same academic field (yes = 1; no = 0). (2) Whether two publications share any authors (yes = 1; no = 0). (3) Whether a publication is published earlier than another one (yes = 1; no = 0). This variable is used to represent forward referencing.

3.2. Analytical Strategies

We adopt three strategies to analyze the data. First, we provide descriptive analyses. We show the number of publications over years, the top journals, the top authors, the most cited publications, etc. We also present the basic features of the citation network, including density (the proportion of all possible citations that are realized in the observed citation network), transitivity (the tendency for a publication to cite its references' references), centralization (the tendency for citations to disproportionately concentrate on a few publications), distributions of indegree (the number of received citations), outdegree (the number of sent citations), betweenness (the number of times a publication is on the shortest path that connects two other publications), and components. Indegree in this context may be viewed as reflecting a publication's influence in the field, outdegree as an indicator for a publication's engagement with the field, and betweenness as reflecting a publication's brokerage power (i.e., the role of connecting diverse academic areas). We list these statistics for both the full and the core citation networks.

Second, we use ERGMs to examine citation patterns. An ERGM assumes that the observed network comes from a family of random networks (Wasserman and Pattison 1996; Handcock et al. 2003; Robins et al. 2007a; Robins et al. 2007b; Robins, Pattison, and Wang 2009).

$$\text{Prob}(W = w|X) = \exp(\theta'g(w, X))/C, \quad (1)$$

where w represents the observed network and $g(w, X)$ the included model terms. θ is the vector of model parameters. C is a normalizing factor that ensures the probability will sum to one.

The model is equivalent to a conditional logit model (Hunter et al. 2008).

$$\text{logit}[P(W_{ij} = 1|w^r, X)] = \theta' \delta^{ij}(w, X), \quad (2)$$

where the log odds of publication i citing publication j (i.e., $w_{ij} = 1$), conditional on their attributes X and the rest of the network w^r , are dependent on the change statistics $\delta^{ij}(w, X)$ (i.e., the changes in the model terms when w_{ij} is flipped from 0 to 1). Hence, the coefficients in ERGMs can be interpreted as conditional log odds ratios.

One advantage of ERGMs is that they can model incoming and outgoing citations simultaneously. In addition, ERGMs can account for both covariate effects and endogenous network formation processes. In this study, we specify two ERGMs. The first ERGM includes only covariate effects, namely, receiver effects, sender effects, and homophily effects. The receiver effects indicate whether publications with certain characteristics are more likely to be cited than those without these characteristics. The sender effects indicate whether publications with certain characteristics are more likely to cite others than those without these characteristics. The homophily effects model assortative mixing mechanisms and indicate whether publications with the same attributes (e.g., in the same field or sharing authors) are more likely to cite one another than those with different attributes. This model is fit on the full citation network.

In the second ERGM, we additionally include several network endogenous formation processes and a variable indicating forward referencing. We include the geometrically weighted edgewise shared partners (GWESP) and the geometrically weighted dyadwise shared partners (GWDEP) to account for transitivity in the citations. GWESP indicates the tendency for citations to be triangular (i.e., if A cites B and B cites C, then A is more likely to cite C). GWDEP indicates the tendency for citations to run across two paths but not close a triangle (i.e., A cites B and B cites C, but A does not cite C). Usually, a positive coefficient for GWESP together with a negative coefficient for GWDEP provides strong evidence for transitivity in the citations (Hunter 2007; Papachristos, Hureau, and Braga 2013). We include the geometrically weighted in-degree distribution (GWDEGREE), and out-degree distribution (GWODEGREE) to represent possible dispersions in the number of received or sent citations and account for cumulative advantage in citations (Merton 1968; Allison 1980). A negative coefficient for each term indicates preferential attachment in the received or sent citations (Hunter 2007) and that the network exhibits a core-periphery structure (Lusher, Koskinen, and Robins 2013). Last, we include a dyadic variable that indicates whether a publication is published earlier than the other one. As forward referencing is rare, we expect its coefficient to be negative and large in size, meaning earlier publications are extremely unlikely to cite newer ones.

As mentioned before, ERGMs that can account for endogenous network formation processes are difficult to fit on big networks. To facilitate the estimation process, we fit the second ERGM only on the core network where isolates (publications with no citations) are removed. Excluding the isolates helps

focus the analyses on interactions among the central publications in the field.³ Furthermore, we employ a pseudo-maximum likelihood estimation (PMLE) (a local approximation to ERGM) to facilitate the estimation.

$$\text{logit} \left[P(w_{ij} = 1 | w_{ij}^L) \right] = \theta' \delta^{ij}(w, X) \quad (3)$$

PMLE assumes ties are independent conditioning on local network structures w_{ij}^L .⁴ We estimate the ERGMs by employing the “statnet” package in R (Handcock et al. 2003).

Our third main strategy to analyze the data is to provide more detailed analyses of the interactions among the top authors (i.e., those with the most publications in causal inference). We choose to focus on these authors because knowing how these elite authors interact helps us better understand how the field is structured at the top and how it may evolve in the future. Specifically, based on the publication citation network, we construct an author citation network in which a link represents the number of times one author cites another one. We use this author citation network for two major purposes. First, we rank the authors according to their network centralities. We consider three network centralities. Indegree is the number of times an author is cited by others, which measures an author’s general intellectual influence in the field. Outdegree is the number of times an author cites others, which measures an author’s engagement (or familiarity) with others’ work in the field. Betweenness is the number of times an author is on the shortest path connecting two other authors in the network, which indicates an author’s connectivity with different schools of thoughts in the field. Second, we use the network to cluster the authors into groups. We present clustering results based on the K-Means method (Hartigan and Wong 1979) that attempts to minimize within-group variance in citation patterns. Additional results based on hierarchical clustering algorithm (Butts 2008b) are available upon request.

4. Results

4.1. Descriptive Results

Figure 1 (and Table A1 in the online supplementary material) shows the number of publications in causal inference over time between 1905 and 2014. Note that the number of publications in 2014 is counted only until middle May in 2014. Before 1980, there were only 261 publications in total. Since the 1980s, the number of publications has quickly increased. The number of yearly publications increased from 38 in 1983 to 3844 in 2013. There is no doubt causal inference as a field has grown rapidly. If the development of causal inference follows the typical S-shaped diffusion curve (Rogers 2010), then it seems like the increase will continue as no leveling-off has been discernible.

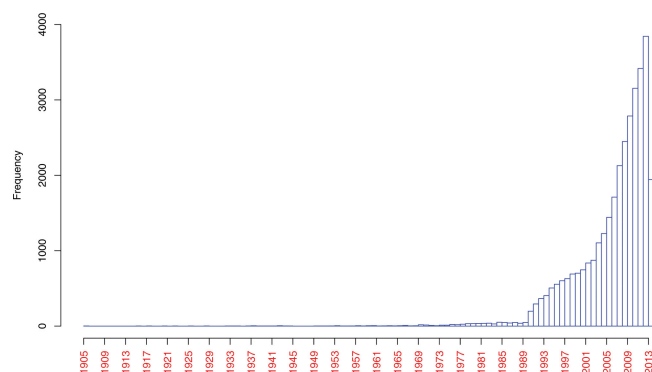


Figure 1. Number of publications in causal inference over time.

Table 1. Journals with the Most Causal Inference Publications

Rank	Journal	Publications
1	Social Science and Medicine	446
2	Statistics in Medicine	363
3	BMC Public Health	347
4	Applied Economics	326
5	American Journal of Epidemiology	291
6	Environmental Health Perspectives	287
7	Medical Care	239
8	Applied Economics Letters	205
9	Drug Safety	205
10	Journal of Clinical Epidemiology	195
11	International Journal of Epidemiology	191
12	Journal of Epidemiology and Community Health	190
13	Epidemiology	186
14	BMC Health Services Research	175
15	Journal of General Internal Medicine	171
16	American Journal of Preventive Medicine	167
17	Energy Economics	161
18	Accident Analysis and Prevention	150
19	Biometrics	148
20	Economic Modelling	143
21	Economics Letters	138
22	Cancer Epidemiology Biomarkers and Prevention	135
23	American Journal of Public Health	133
24	American Journal of Managed Care	125
25	Journal of the American Statistical Association	125
26	Journal of Econometrics	123
27	Health Services Research	119
28	Pharmacoeconomics	118
29	Health Economics	108
30	Expert Systems with Applications	106

Table 1 shows the top 30 journals with the most publications in causal inference. Overall, it appears that journals in health, economics, and statistics are among the most important outlets for causal inference research. If we group the journals by academic disciplines, indeed, health, economics, and statistics are the top three areas which have produced the most research in causal inference. Other active disciplines in the field include management, sociology, biology, and education. See Table 2 for more detail.

Table 3 lists the 30 institutions with the most publications in causal inference. We identify institutions based on authors’ email addresses. It appears that universities with a strong medical and/or public health school produce more research in causal inference. The field also seems to have evolved into a polycentric structure with a few strongholds like Harvard University, University of Michigan, Columbia University, Johns Hopkins University, Yale University, and Centers for Disease Control and Prevention.

³ To provide sensitivity analysis, we also fit the first ERGM on the core network. That the results do not significantly differ from those based on the full network indicates that the results from the second ERGM based on the core network may be (reasonably) generalizable to the full network.

⁴ When there are no endogenous network formation processes in the model, like in the first ERGM, PMLE is equivalent to MCMLE. Note that if MCMLE describes the true model, PMLE may under-estimate endogenous network formation processes and provide more conservative inferences (i.e., wider confidence intervals) on covariate effects (van Duijn, Gile, and Handcock 2009).

Table 2. Disciplines with the most causal inference publications

Rank	Field	Publications
1	Health	10503
2	Economics	4005
3	Statistics	2151
4	Management	2038
5	Sociology	1533
6	Biology	1296
7	Education	1229
8	Public Policy and Management	1223
9	Psychology	1091
10	Political Science	781
11	Law	726
12	Computer Science and Informatics	266

Table 4 lists the top authors in the field. Panel A shows the 50 most prolific authors while Panel B the 50 most cited authors. There are some overlaps in the two lists.

Table 5 shows the 20 publications receiving the most citations from other publications in the field of causal inference. Many well-known, influential papers are well represented in the list. Indeed, if a researcher wants to study the most influential research in causal inference, Table 5 is a good starting point. Additionally, Table A2 lists the top 20 publications that cite the most references within the field of causal inference. Table A3 lists the 20 publications that have the highest betweenness centrality.

We present the summary statistics of the full citation network in panel A of Table 6. The network is sparse, as indicated by the low density. The centralization score is small, around 0.03,

reflecting that the citations are not concentrated on a few publications. The degree of reciprocity is very low—only about 1% of the citations are mutual, probably as a result of no forward referencing. There is a sizable degree of transitivity—about 33% of the publications cite their references' references. The network includes 28,802 components with a mean size around 1 and a standard deviation in the size around 23.38. The largest component includes 3968 publications. On average, each publication receives about 0.25 citations whereas one publication receives the maximal 2028 citations (Figure A1). The majority of the publications cite none in the field, whereas one publication cites the maximal 18 references (Figure A2). Indegree and betweenness are highly correlated, suggesting that publications that receive many citations are also likely to be cited across different disciplines. Indegree and outdegree are weakly correlated, indicating that highly cited publications do not necessarily cite more.

Panel B of Table 6 shows the summary statistics of the core citation network. Density and reciprocity are still very low. There is still a reasonable degree of transitivity. Indegree still strongly correlates with betweenness, while the correlation between outdegree and betweenness is weak. However, the publications in the core network (by construction) have at least one or more citations. There are also fewer but usually larger components. The distributions of indegree and betweenness are relatively less skewed than their counterparts in the full network. Also, the correlation between indegree and outdegree becomes negative. The publications in the two networks also differ significantly in covariates distributions (Table A4). In the next section, we show that although the publication attributes and the network

Table 3. Institutions with the most causal inference publications

Rank	A. Ranked by Overall Publications		B. Ranked by Publications Since 2010	
	Institution	Publications	Institution	Publications
1	Harvard University	501	Harvard University	278
2	University of Michigan	274	University of Michigan	140
3	Johns Hopkins University	167	Columbia University	96
4	Columbia University	166	Johns Hopkins University	87
5	Yale University	134	Yale University	79
6	Centers for Disease Control and Prevention (CDC)	130	Centers for Disease Control and Prevention (CDC)	76
7	University of Washington	125	University of California, Berkeley	75
8	University of Minnesota	113	Penn State University	73
9	Penn State University	110	Duke University	70
10	University of California, Berkeley	109	Stanford University	66
11	Duke University	104	University of Minnesota	66
12	University College London, UK	104	McGill University, Canada	65
13	University of North Carolina at Chapel Hill	102	University of Illinois	63
14	McGill University, Canada	101	University of Melbourne, Australia	63
15	University of Melbourne, Australia	101	New York University	61
16	Stanford University	94	University College London, UK	61
17	University of New South Wales, Australia	91	University of Washington	58
18	University of Chicago	87	University of North Carolina at Chapel Hill	58
19	University of California, Los Angeles	86	University of New South Wales, Australia	55
20	University of Toronto, Canada	86	London School of Economics and Political Science, UK	53
21	London School of Economics and Political Science, UK	85	Bristol University, UK	52
22	Arizona State University	83	Monash University, Australia	52
23	Bristol University, UK	83	London School of Hygiene and Tropical Medicine, UK	51
24	London School of Hygiene and Tropical Medicine, UK	83	Imperial College London, UK	50
25	Imperial College London, UK	82	Massachusetts Institute of Technology	50
26	University of Manchester, UK	82	Partners Healthcare	50
27	World Bank	80	University of Washington	48
28	New York University	79	University of Manchester, UK	47
29	Vanderbilt University	75	University of Chicago	47
30	Massachusetts Institute of Technology	73	University of Toronto, Canada	47

Table 4. Top authors in causal inference

A. The 50 Most Prolific Authors						B. The 50 Most Cited Authors					
Rank	Author	Pubs	Rank	Author	Pubs	Rank	Author	Cites	Rank	Author	Cites
1	Rubin, DB	66	26	Li, Y	23	1	Rubin, DB	2682	26	Hudgens, MG	69
2	Robins, J	63	27	Stuart, EA	23	2	Rosenbaum, PR	2096	27	Mark, SD	69
3	Vanderweele, TJ	61	28	Imbens, GW	22	3	Robins, J	1373	28	Newey, WK	69
4	Van Der Laan, MJ	51	29	Small, DS	22	4	Imbens, GW	732	29	Petersen, ML	68
5	Rosenbaum, PR	50	30	Brookhart, MA	21	5	Hernan, MA	708	30	Sinisi, SE	68
6	Smith, GD	48	31	Gilbert, PB	21	6	Brumback, B	498	31	Ashley, R	61
7	Kawachi, I	39	32	Narayan, PK	21	7	Pearl, J	373	32	Granger, CWJ	61
8	Lee, CC	39	33	Olsen, J	21	8	Frangakis, CE	345	33	Schmalensee, R	61
9	Austin, PC	38	34	Heckman, J	20	9	Thomas, N	278	34	Vansteelandt, S	60
10	Greenland, S	37	35	Imai, K	20	10	Vanderweele, TJ	278	35	Lin, DY	59
11	Schneeweiss, S	35	36	Lawlor, DA	20	11	Greenland, S	274	36	Angrist, JD	57
12	Vansteelandt, S	35	37	Lee, J	20	12	Abadie, A	223	37	Kronmal, RA	57
13	Miller, RR	34	38	Schisterman, EF	20	13	Ridder, G	212	38	Psaty, BM	57
14	Hernan, MA	33	39	Zhang, J	20	14	Blevins, D	211	39	Davidian, M	51
15	Cole, SR	31	40	Geng, Z	19	15	Hirano, K	211	40	Halloran, ME	49
16	Lechner, M	30	41	Glasgow, RE	19	16	Ritter, G	211	41	Rotnitzky, A	49
17	Lee, S	28	42	Goetghebeur, E	19	17	Wulfsohn, M	211	42	Schneeweiss, S	49
18	Kim, J	27	43	Joffe, MM	19	18	Phillips, PCB	166	43	Florens, JP	48
19	Glynn, RJ	25	44	Kim, S	19	19	Toda, HY	145	44	Card, D	47
20	Kivimaki, M	25	45	Millimet, DL	19	20	Hahn, JY	121	45	Bosch, RJ	45
21	Platt, RW	25	46	Moodie, EEM	19	21	Hernandez-Diaz, S	108	46	Joffe, MM	44
22	Payne, JE	24	47	Tang, CF	19	22	Bang, H	107	47	Rassen, JA	44
23	Savitz, DA	24	48	Wang, Y	19	23	Gilbert, PB	105	48	Avorn, J	42
24	Shahbaz, M	24	49	King, G	18	24	Van Der Laan, MJ	95	49	Glynn, RJ	42
25	Zhang, Y	24	50	Pearl, J	18	25	Tsiatis, AA	93	50	Struchiner, CJ	42

features differ somewhat between the core and the full citation networks, the mechanisms for citation formations are quite similar.

4.2. ERGM Results for Predicting Citation Patterns

Table 7 shows the ERGM results. Model 1 contains only covariate effects and is fitted on the full citation network. The term “edges” act like an intercept in regressions. The negative coefficient here indicates the baseline connectivity of the citation network is low. Among the covariate effects, we find that all else

equal, publications with funding support, a U.S. corresponding author, first author having a longer last name, fewer authors, and authors from prestigious universities, longer publications, publications with more keywords, earlier publications, and publications in method journals and top journals are significantly more likely than their counterparts to receive citations. At the same time, publications with funding support, a U.S. corresponding author, first author having a longer last name, fewer authors, and authors from prestigious universities, non-article publications, longer publications, publications with more keywords, more recent publications, and publications in method journals

Table 5. Top 20 Publications Receiving the Most Citations Within the Field of Causal Inference

Rank	Author	Article title	Citations
1	Rosenbaum, PR; Rubin, DB	The Central Role of the Propensity Score in Observational Studies for Causal Effects	2028
2	Robins, J; Hernan, MA; Brumback, B	Marginal Structural Models and Causal Inference in Epidemiology	346
3	Frangakis, CE; Rubin, DB	Principal Stratification in Causal Inference	223
4	Robins, J; Blevins, D; Ritter, G; Wulfsohn, M	G-Estimation of the Effect of Prophylaxis Therapy for Pneumocystis-Carini Pneumonia on the Survival of Aids Patients	221
5	Hirano, K; Imbens, GW; Ridder, G	Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score	211
6	Imbens, GW	The Role of the Propensity Score in Estimating Dose-Response Functions	208
7	Pearl, J	Causal Diagrams for Empirical Research	181
8	Rubin, DB; Thomas, N	Matching Using Estimated Propensity Scores: Relating Theory to Practice	161
9	Greenland, S; Pearl, J; Robins, J	Causal Diagrams for Epidemiologic Research	155
10	Abadie, A; Imbens, GW	Large Sample Properties of Matching Estimators for Average Treatment Effects	155
11	Hernan, MA; Brumback, B; Robins, J	Marginal Structural Models to Estimate the Causal Effect of Zidovudine on the Survival of Hiv-Positive Men	149
12	Toda, HY; Phillips, PCB	Vector Autoregressions and Causality	145
13	Hahn, JY	On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects	121
14	Hernan, MA; Hernandez-Diaz, S; Robins, J	A Structural Approach to Selection Bias	108
15	Bang, H	Doubly Robust Estimation in Missing Data and Causal Inference Models	107
16	Vanderweele, TJ	On the Distinction Between Interaction and Effect Modification	105
17	Frangakis, CE; Rubin, DB	Addressing Complications of intention-to-Treat Analysis in the Combined Presence of All-Or-None Treatment-Noncompliance and Subsequent Missing Outcomes	76
18	Hernan, MA; Robins, J	Instruments for Causal Inference: An Epidemiologist'S Dream?	72
19	Greenland, S	Quantifying Biases in Causal Models: Classical Confounding vs. Collider-Stratification Bias	72
20	Robins, J; Mark, SD; Newey, WK	Estimating Exposure Effects By Modeling the Expectation of Exposure Conditional on Confounders	69

Table 6. Summary Statistics of the Citation Networks

A. The full citation network				B. The core citation network			
I. Basic network statistics				I. Basic network statistics			
Density	8.E-06	Components	28802	Density	4.E-04	Components	265
Centralization	0.03	Min	1	Centralization	0.21	Min	2
Reciprocity	0.01	Mean	1.16	Reciprocity	0.00	Mean	18
Transitivity	0.33	Max	3968	Transitivity	0.33	Max	3968
Isolates	28537	SD	23.38	Isolates	0	SD	243.58
II. Summary information of the centrality measures				II. Summary information of the centrality measures			
	Indegree	Outdegree	Betweenness		Indegree	Outdegree	Betweenness
Min.	0	0	0	Min.	0	0	0
Mean	0.25	0.25	0.52	Mean	1.75	1.75	3.62
Max.	2028	18	1867	Max.	2028	18	1871
SD	11.89	0.87	19.99	SD	31.37	1.61	52.70
Skewness	150.81	5.75	63.37	Skewness	57.26	2.59	23.98
III. Spearman rank correlations of the centrality measures				III. Spearman rank correlations of the centrality measures			
	Indegree	Outdegree	Betweenness		Indegree	Outdegree	Betweenness
Indegree	1			Indegree	1		
Outdegree	0.13	1		Outdegree	− 0.32	1	
Betweenness	0.61	0.24	1	Betweenness	0.61	0.14	1

Table 7. ERGM Results for Predicting the Citation Networks

	Model 1 full network			Model 2 core network			Model 3 core network		
	Est	SE		Est	SE		Est	SE	
Edges	− 13.82	0.15	***	− 7.53	0.15	***	− 6.82	0.32	***
Receiver effects									
Funding	1.25	0.05	***	1.20	0.05	***	0.63	0.12	***
U.S. Author	0.38	0.02	***	− 0.02	0.02		− 0.10	0.05	
Short Last Name	− 0.54	0.04	***	− 0.43	0.04	***	− 0.23	0.08	**
More Authors	− 1.25	0.04	***	− 1.14	0.04	***	− 0.91	0.08	***
Prestigious University	0.59	0.03	***	0.20	0.03	***	0.45	0.08	***
Article	− 0.11	0.12		− 0.52	0.12	***	− 0.41	0.24	
Longer Publications	1.32	0.05	***	1.21	0.05	***	0.70	0.11	***
More Keywords	0.15	0.02	***	− 0.01	0.02		− 0.19	0.05	***
Recent Publications	− 2.55	0.04	***	− 3.03	0.04	***	− 1.16	0.09	***
Method Journal	0.29	0.03	***	− 0.93	0.03	***	− 0.07	0.08	
Top Journal	1.14	0.02	***	0.25	0.02	***	− 0.14	0.06	*
Sender effects									
Funding	0.36	0.03	***	0.16	0.03	***	0.09	0.06	
U.S. Author	0.50	0.02	***	0.20	0.02	***	0.27	0.05	***
Short Last Name	− 0.07	0.03	*	− 0.04	0.03		0.01	0.06	
More Authors	− 0.34	0.03	***	− 0.18	0.03	***	− 0.38	0.06	***
Prestigious University	0.42	0.03	***	0.17	0.03	***	0.16	0.06	**
Article	− 0.34	0.08	***	− 0.37	0.08	***	− 0.48	0.17	**
Longer Publications	0.13	0.03	***	0.04	0.03		0.27	0.07	***
More Keywords	0.17	0.02	***	0.13	0.02	***	0.33	0.05	***
Recent Publications	0.67	0.03	***	0.31	0.03	***	0.13	0.07	
Method Journal	1.60	0.02	***	0.52	0.02	***	1.01	0.06	***
Top Journal	0.66	0.02	***	0.09	0.02	***	0.04	0.05	
Homophily									
Same Field	0.83	0.02	***	0.82	0.02	***	0.94	0.06	***
Shared Authors	5.98	0.08	***	4.94	0.09	***	4.69	0.28	***
Network structures									
GWESP (Transitivity)							3.47	0.05	***
GWDSP (Two-Path)							− 0.49	0.02	***
GWIDEGREE							− 5.77	0.16	***
GWODEGREE							3.86	0.00	***
Forward Referencing							− 5.E+15	4.E+04	***
Number of Nodes	33,306			4,769			4,769		
Number of Edges	8,367			8,367			8,367		

Note: Estimates are conditional odds ratio. Significance code: *, $P < 0.05$; **, $P < 0.01$; ***, $P < 0.001$.

and top journals are significantly more likely than their counterparts to cite others. Thus, both author and publication characteristics matter for citations and their effects are similar for both receiving and sending citations.

We also find significant homophily in the citations. Publications in the same academic field and with shared authors are much more likely to cite one another than otherwise. In particular, the odds of publications with shared authors citing each other is about 400 ($e^{5.98} = 395$) times the odds of publications without shared authors citing each other.

In model 2, we fit the same ERGM as in model 1 on the core citation network to partly check consistency of the results. The results are broadly similar to those in model 1, with a few notable differences. For example, US authors are no longer more likely to receive citations. Article publications start to receive fewer citations. The coefficients for the length of the first author's last name, number of keywords, and publication length become statistically insignificant, probably because of the reduced network size. Hence, on one hand, the estimates from the core network seem to approximate the ones from the full network quite well and we hope this will still be the case in the more complex ERGMs. On the other hand, the discrepancies caution us not to overly project the results from the core network to the full network.

Model 3 additionally controls for several endogenous network formation processes and is fitted on the core network.⁵ The magnitude or significance of many previous estimates change somewhat. Most notably, the receiver effects of funding support and short last name reduce almost by half, while the receiver effect of authors from prestigious universities doubles. The receiver effect of publishing in method journals reduces so much in size that it loses statistical significance. The receiver effect of publishing in top journals switches the sign (from positive to negative). Overall, the results show the importance of controlling for endogenous network formation processes to more accurately estimate the covariate effects. At the same time, many of the covariate effects are relatively robust and consistent across the three models. Publications with funding support, first author having a longer last name, fewer authors, and authors from prestigious universities, that is longer, and published earlier are significantly more likely than their counterparts to receive citations. Publications with a U.S. corresponding author, fewer authors, authors from prestigious universities, a non-article format, and more keywords and published in top journals are significantly more likely than their counterparts to cite others. Homophilious effects based on academic field and shared authorship are also consistent across the models.

Model 3 shows that endogenous network formation processes also play an important role in citation formations. The coefficient for "GWESP" indicates that a citation toward a reference's reference is over 30 ($e^{3.47} = 32$, $P < 0.001$) times more likely to occur than not. The negative coefficient for "GWDSP" indicates that citations that do not close a triangle are less likely to occur. The results on "GWDSP" and "GWESP" together provide strong evidence that citations tend to be transitive, which

may be driven by the fact that researchers tend to snowball-sample the literature to learn about studies in the field. The negative coefficient for "GWIDEGREE" indicates there is preferential attachment in indegree (a few publications are highly cited by others) and the citation network exhibits a core-periphery structure (Hunter 2007; Lusher, Koskinen, and Robins 2013). In contrast, the positive coefficient for "GWODEGREE" indicates that there is less variation in the number of outgoing citations. The extremely large and negative coefficient for "forward referencing" indicates that it is very rare to cite future publications, as one would expect.

4.3. Analyses of the Citation Network Among Top Authors

Figure 2 shows the citation network among the 50 most prolific authors in causal inference. Each node represents an author and each edge the citation counts between authors. The node size is proportional to the received citations. The graph indicates that the network exhibits a core-periphery structure. In the center are a few well-known experts such as J. Robins, D. B. Rubin, M. A. Hernán, P. R. Rosenbaum, J. Pearl, S. Greenland, and G. W. Imbens. Citations also tend to be asymmetric, namely, some authors cite others a lot while not being cited equivalently. This is different from friendship networks, where mutual relations are more prevalent. The asymmetry may reflect differential in intellectual status and a strong consensus on who the influential researchers are (Cole, Cole, and Dietrich 1978). The network also appears to be a small-world network in which any two authors can reach each other by only a few steps (Milgram 1967; Watts and Strogatz 1998; Watts 1999).

We list the centralities of these authors in Table A5. Indegree reflects the number of times that an author has been cited by others and outdegree the number of times of citing others. Betweenness reflects the degree to which an author has been cited by other authors from diverse disciplines. There is a strong correlation between indegree and betweenness centralities ($\rho = 0.84$), indicating the most cited authors are also likely to be cited by others from diverse disciplines. There is also a sizable correlation between outdegree and either of the indegree and betweenness centralities ($\rho \approx 0.5$). But there is also some degree of discrepancy across these centrality measures. For example, T. J. Vanderweele is (currently) not the most cited author in this group, but he has cited others the most and has been cited by many authors from diverse disciplines (as indicated by his high betweenness centrality). As these centrality measures reflect different aspects of intellectual influence, they should be used jointly to understand a researcher's influence in the field.

We also use the K-means method to cluster the authors (Hartigan and Wong 1979). Figure 3 shows the clusters with different colors and shapes. More detailed results can be found in Table A6. One cluster (marked as black circle) is mainly composed of authors whose research is relevant to biostatistics and health sciences (e.g., J. Robins, M. A. Hernán, and D. S. Small) and/or who use graphic approaches (e.g., J. Pearl and S. Greenland). Another cluster (marked as green square) includes authors mostly from statistics and social sciences. The final cluster (marked as red triangle) consists of authors whose research spans statistics/social sciences and biostatistics/health sciences (e.g., T. J. Vanderweele).

⁵ Figure A3 provides some assessment of the goodness of fit of Model 3 (Goodreau et al. 2008). The model seems to have captured structural features of the network well, as there are only a few places at which the observed statistics lie out of the 95% confidence intervals of the statistics in simulated networks.

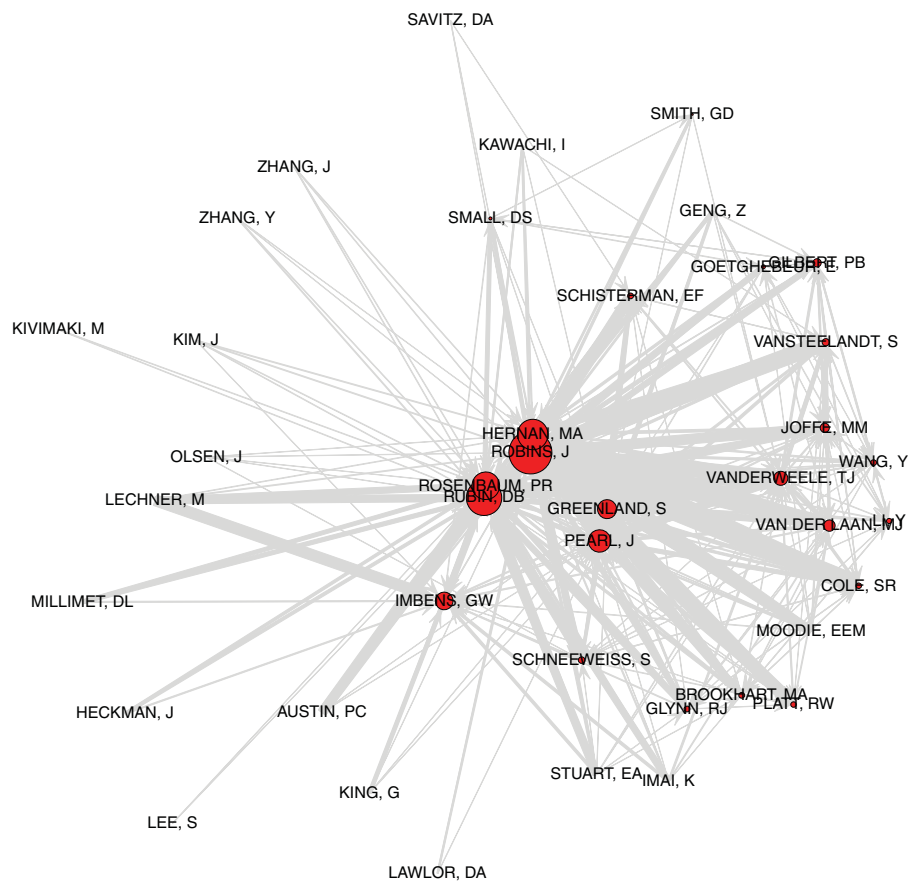


Figure 2. The citation network among the 50 most prolific authors.

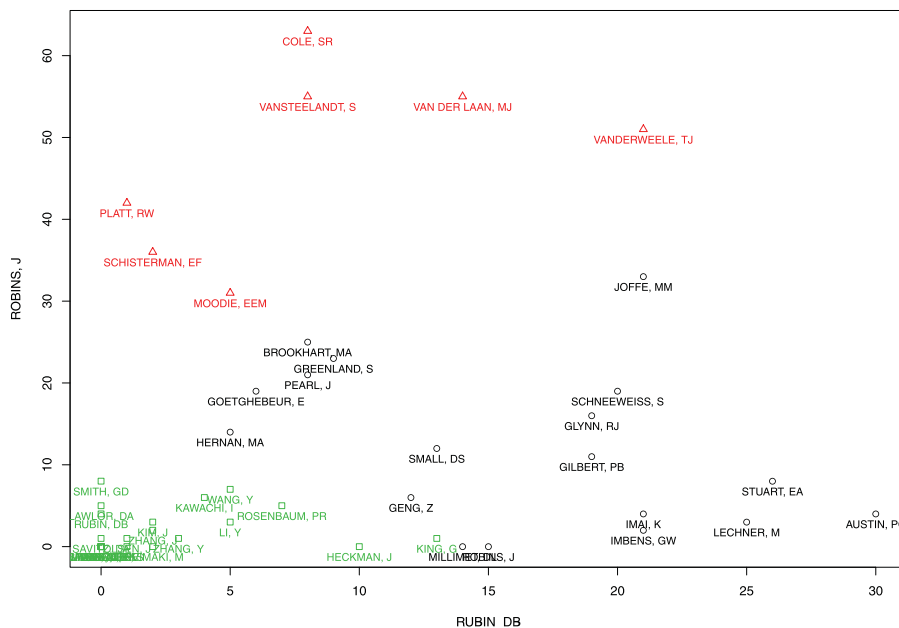


Figure 3. Clustering the top Authors by the K-Means Method.

5. Conclusion and Discussion

In this study, we review the field of causal inference through the lens of citation networks. Like previous studies, we present detailed descriptive analyses by showing the top authors, most cited research, and the properties of the citation networks. But

more importantly, we employ ERGMs to provide detailed analyses of citation formation mechanisms in the field. We show that both publication and author characteristics matter for citations. Citations can occur because of not only technical concerns (e.g., through citing papers in method journals), but also social processes, such as legitimization (e.g., through citing publications

by authors from prestigious institutions), homophily (the tendency to cite publications in the same academic field or with shared authors), preferential attachment (the tendency for a few publications to receive many citations), and transitivity (the tendency to cite references' references). In addition, we also provide specific analyses of the citations among the top authors in the field. We show that authors might play different roles in the field according to their network positions. The ones receiving more citations tend to be the founding figures in the field. The ones citing others a lot are knowledge distributors. And the ones who are cited by many authors from different fields are connectors who help connect research from otherwise disconnected communities. We also show that these authors might be divided into three large groups representing researchers from statistics/social sciences, biostatistics/health sciences, and those whose research spans the two broad areas. It appears that more synthesis across these different disciplines is needed in future development of the field. Overall, our article helps reveal insights into the social structure and processes in the field of causal inference. It may serve as a case study for citation network analysis and for fitting ERGMs on big networks.

With that said, the current study is still exploratory. It opens a door for a series of questions that are worth further investigations. A popular approach in sociology of science conceives a scientific field as a hierarchically structured network composed of various relations (e.g., opposition or collaboration and domination or subversion) among researchers with different technical capital and social positions (Bourdieu and Wacquant 1992). Within the field, there are constant competitions and power struggles over culture (e.g., preference for research priority and research style and recognition of scientific competency), capital (e.g., distribution of material goods like laboratory space, funding, and job positions), and boundaries of the field (Bourdieu 1975; Bourdieu 1991; Bourdieu 2004; Gieryn 1983). On one hand, our analyses show that there is significant inequality in the number of papers authored and the number of citations received by researchers in the field. There is also significant asymmetry in the citations between authors. These are signs of hierarchy. On the other hand, the field may be less hierarchical than expected. There are several institutions and multiple disciplines that all contributed significantly to the field. Most authors are in a small world who can reach one another via a few steps in the citation network. Hence, the field may be more appropriately characterized by a polycentric structure with a high level of cohesion within groups (i.e., disciplines and institutions) and a moderate level of cohesion across groups. Thus, to avoid fractionalization and increase cohesion in the field, more intergroup interactions seem to be needed (Moody 2004).

Questions about diffusion of innovations may also be pursued. Our study shows that many influential publications in causal inference have only one or two authors, which confirms that opinion leaders are often minority (Rogers 2010). Future work may study how personal factors and social processes (e.g., cumulative advantage) help propel the success of the leaders and whether resources are increasingly concentrated toward the leaders and how that will affect the field (Merton 1968; Cole and Cole 1973; Merton 1973; Allison and Stewart 1974; Allison 1980; Allison, Long, and Krauze 1982; Allison and Long 1990). At the field level, the potential outcomes framework is

paradigm-shifting (Kuhn 1970). Future work needs to show how this new paradigm spreads across disciplines (Cohen 2015) and whether it has resolved the "liability of newness" (Freeman, Carroll, and Hannan 1983) and will become THE legitimate framework for causal inference.

We have used multiple text mining methods to extract information from publication records. But given the complexity of the data some measurement error is inevitable. Future work may improve our research by collecting more detailed data, which will be useful for investigating citation patterns related to race, gender, and country of origin. Future work may also apply our methods to studying other scientific fields and compare the results across fields. Finally, the relational event model (Butts 2008a) may be used to model time to occurrences of citations and study temporal changes in citation networks.

Supplementary Material

The supplementary materials contain additional tables and figures.

Acknowledgments

The authors would like to thank Dr. Jian Xu for the help with data collection. The authors are grateful to Professor Nicole Lazar and anonymous reviewers for the valuable comments on earlier drafts of this article.

ORCID

Weihua An  <http://orcid.org/0000-0003-0334-2930>
Ying Ding  <http://orcid.org/0000-0003-2567-2009>

References

- Abadie, A. (2003), "Semiparametric Instrumental Variable Estimation of Treatment Response Models," *Journal of Econometrics*, 113, 231–263. [265]
- Abadie, A., Drukker, D., Herr, J. L., and Imbens, G. (2004), "Implementing Matching Estimators for Average Treatment Effects in Stata," *The Stata Journal*, 1, 1–18. [265]
- Abadie, A., and Imbens, G. (2006), "Large Sample Properties of Matching Estimators for Average Treatment Effects," *Econometrica*, 74, 235–267. [265]
- (2016), "Matching on the Estimated Propensity Score," *Econometrica*, 84, 781–807. [265]
- (2011), "Bias-Corrected Matching Estimators for Average Treatment Effects," *Journal of Business & Economic Statistics*, 29, 1–11. [265]
- Allison, P. D. (1980), "Inequality and Scientific Productivity," *Social Studies of Science*, 10, 163–179. [266,268,275]
- Allison, P. D., and Long, J. S. (1990), "Departmental Effects on Scientific Productivity," *American Sociological Review*, 55, 469–478. [275]
- Allison, P. D., Long, J. S., and Krauze, T. K. (1982), "Cumulative Advantage and Inequality in Science," *American Sociological Review*, 47, 615–625. [275]
- Allison, P. D., and Stewart, J. A. (1974), "Productivity Differences among Scientists: Evidence for Accumulative Advantage," *American Sociological Review*, 39, 596–606. [275]
- An, W. (2010), "Bayesian Propensity Score Estimators: Incorporating Uncertainties in Propensity Scores into Causal Inference," *Sociological Methodology*, 40, 151–189. [265]
- (2015), "Instrumental Variables Estimates of Peer Effects in Social Networks," *Social Science Research*, 50, 382–394. [265]
- An, W., and Wang, X. (2016), "LARE: Instrumental Variable Estimation of Causal Effects through Local Average Response Functions," *Journal of Statistical Software*, 71, 1–13. [265]

- Angrist, J., Imbens, G., and Rubin, D. (1996), "Identification of Causal Effects Using Instrumental Variables," *Journal of the American Statistical Association*, 91, 444–455. [265]
- Angrist, J. D., and Pischke, J.-S. (2009), *Mostly Harmless Econometrics: An Empiricist's Companion*, New Jersey: Princeton University Press. [265]
- Baiocchi, M., Cheng, J., and Small, D. S. (2014), "Instrumental Variable Methods for Causal Inference," *Statistics in Medicine*, 33, 2297–2340. [265]
- Barringer, S., Eliason, N. S. R., and Leahey, E. (2013), "A History of Causal Analysis in the Social Sciences," in *Handbook of Causal Analysis for Social Research*, ed. Morgan, S. L., Berlin: Springer. [265]
- Basu, D. (1980), "Randomization Analysis of Experimental Data: The Fisher Randomization Test," *Journal of the American Statistical Association*, 75, 575–582. [265]
- Bourdieu, P. (1975), "The Specificity of the Scientific Field and the Social Conditions of the Progress of Reason," *Information (International Social Science Council)*, 14, 19–47. [275]
- (1991), "The Peculiar History of Scientific Reason," *Sociological Forum*, 6, 3–26. [275]
- (2004), *Science of Science and Reflexivity*, Chicago, IL: University of Chicago Press. [275]
- Bourdieu, P., and Wacquant, L. J. (1992), *An Invitation to Reflexive Sociology*, Chicago, IL: University of Chicago press. [275]
- Butts, C. T. (2008a), "A Relational Event Framework for Social Action," *Sociological Methodology*, 38, 155–200. [275]
- (2008b), "Social Network Analysis with sna," *Journal of Statistical Software*, 24, available at <http://www.jstatsoft.org/v24/i06/>. [269]
- Cohen, M. (2015), *Paradigm Shift: How Expert Opinions Keep Changing on Life, the Universe, and Everything*, Andrews UK Limited. [275]
- Cole, J., and Cole, S. (1973), *Social Stratification in Science*, Chicago, IL: University of Chicago Press. [275]
- Cole, S., Cole, J. R., and Dietrich, L. (1978), "Measuring the Cognitive State of Scientific Disciplines," in ed. Y. Elkana, *Toward a Metric of Science*, New York: Wiley, pp. 209–251. [273]
- Cole, S. R., and Hernán, M. A. (2008), "Constructing Inverse Probability Weights for Marginal Structural Models," *American Journal of Epidemiology*, 168, 656–664. [265]
- Ding, P., Feller, A., and Miratrix, L. (2015), "Randomization Inference for Treatment Effect Variation," *Journal of the Royal Statistical Society, Series B*, 78, 655–671. [265]
- Ding, P., VanderWeele, T. J., and Robins, J. (2017), "Instrumental Variables as Bias Amplifiers With General Outcome and Confounding," *Biometrika*, 104, 291–302. [265]
- Ding, Y. (2011), "Applying Weighted PageRank to Author Citation Networks," *Journal of the American Society for Information Science and Technology*, 62, 236–245. [266]
- Drake, M. A. (2004), *Encyclopedia of Library and Information Science*, New York: Marcel Dekker. [266]
- Elwert, F. (2013), "Graphical Causal Models," in ed. S. Morgan, *Handbook of Causal Analysis for Social Research*, New York: Springer, pp. 245–273. [265]
- Freeman, J., Carroll, G. R., and Hannan, M. T. (1983), "The Liability of Newness: Age Dependence in Organizational Death Rates," *American Sociological Review*, 48, 692–710. [275]
- Gieryn, T. F. (1983), "Boundary-work and the Demarcation of Science from Non-science: Strains and Interests in Professional Ideologies of Scientists," *American Sociological Review*, 48, 781–795. [275]
- Goodreau, S. M., Handcock, M. S., Hunter, D. R., Butts, C. T., and Morris, M. (2008), "A Statnet Tutorial," *Journal of Statistical Software*, 24, available at <http://www.jstatsoft.org/v24/i09/>. [266]
- Greenland, S., Pearl, J., and Robins, J. M. (1999), "Causal Diagrams for Epidemiologic Research," *Epidemiology*, 10, 37–48. [265]
- Handcock, M. S., Hunter, D. R., Butts, C. T., Goodreau, S. M., and Morris, M. (2003), "statnet: Software tools for the Statistical Modeling of Network Data," available at <http://statnetproject.org>. [268,269]
- Hartigan, J. A., and Wong, M. A. (1979), "Algorithm AS 136: A K-Means Clustering Algorithm," *Journal of the Royal Statistical Society, Series C*, 28, 100–108. [269,273]
- Hernán, M. A., and Robins, J. M. (2006), "Instruments for Causal Inference: An Epidemiologist's Dream?" *Epidemiology*, 17, 360–372. [265]
- (2018), *Causal Inference*, Boca Raton, FL: CRC press. [265]
- Hirano, K., Imbens, G. W., and Ridder, G. (2003), "Efficient Estimation of Average Treatment Effects using the Estimated Propensity Score," *Econometrica*, 71, 1161–1189. [265]
- Ho, D. E., Imai, K., King, G., and Stuart, E. A. (2007), "Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference," *Political Analysis*, 15, 199–236. [265]
- Hu, A., and Mustillo, S. (2016), "Recent Developments of Propensity Score Methods in Observational Studies: Multi-categorical Treatment, Causal Mediation, and Heterogeneity," *Current Sociology*, 64, 60–82. [265]
- Hunter, D. R. (2007), "Curved Exponential Family Models for Social Networks," *Social Networks*, 29, 216–230. [268,273]
- Hunter, D., Handcock, M., Butts, C., Goodreau, S., and Morris, M. (2008), "ERGM: A Package to Fit, Simulate and Diagnose Exponential-family Models for Networks," *Journal of Statistical Software*, 24, available at <http://www.jstatsoft.org/v24/i03/paper>. [268]
- Hunter, D. R., and Handcock, M. S. (2006), "Inference in Curved Exponential Family Models for Networks," *Journal of Computational and Graphical Statistics*, 15, 565–583. [265]
- Imai, K., and Dyk, D. A. V. (2004), "Causal Inference with General Treatment Regimes: Generalizing the Propensity Score," *Journal of the American Statistical Association*, 99, 854–866. [265]
- Imai, K., King, G., and Stuart, E. A. (2008), "Misunderstandings between Experimentalists and Observationalists about Causal Inference," *Journal of the Royal Statistical Society, Series A*, 171, 481–502. [265]
- Imbens, G. (2000), "The Role of the Propensity Score in Estimating Dose-Response Functions," *Biometrika*, 87, 706–710. [265]
- Imbens, G. W., and Angrist, J. D. (1994), "Identification and Estimation of Local Average Treatment Effect," *Econometrica*, 62, 467–476. [265]
- Imbens, G. W., and Rubin, D. B. (2015), *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*, New York: Cambridge University Press. [265]
- Imbens, G. W., and Wooldridge, J. M. (2009), "Recent Developments in the Econometrics of Program Evaluation," *Journal of Economic Literature*, 47, 5–86. [265]
- Ji, P., and Jin, J. (2016), "Coauthorship and Citation Networks for Statisticians," *The Annals of Applied Statistics*, 10, 1779–1812. [265,266]
- Kuhn, T. S. (1970), *The Structure of Scientific Revolutions* (2nd ed.), University of Chicago Press. [275]
- Long, J. S., McGinnis, R., and Allison, P. D. (1980), "The Problem of Junior-Authored Papers in Constructing Citation Counts," *Social Studies of Science*, 10, 127–143. [267]
- Lusher, D., Koskinen, J., and Robins, G. (2013), *Exponential Random Graph Models for Social Networks: Theory, Methods, and Applications*, New York: Cambridge University Press. [268,273]
- McPherson, M., Smith-Lovin, L., and Cook, J. (2001), "Birds of a Feather: Homophily in Social Networks," *Annual Review of Sociology*, 27, 415–444. [266]
- Merton, R. K. (1968), "The Matthew Effect in Science: The Reward and Communication Systems of Science are Considered," *Science*, 159, 56–63. [266,268,275]
- (1973), *The Sociology of Science: Theoretical and Empirical Investigations*, University of Chicago Press. [275]
- Milgram, S. (1967), "The Small World Problem," *Psychology Today*, 1, 61–67. [273]
- Moody, J. (2004), "The Structure of a Social Science Collaboration Network," *American Sociological Review*, 69, 213–238. [266,275]
- Morgan, S. L., and Winship, C. (2015), *Counterfactuals and Causal Inference: Methods and Principles for Social Research* (2nd ed.), New York: Cambridge University Press. [265]
- Newman, M. E. (2001), "The Structure of Scientific Collaboration Networks," *Proceedings of the National Academy of Sciences*, 98, 404–409. [266]
- Ogburn, E. L., VanderWeele, T. J., et al. (2014), "Causal Diagrams for Interference," *Statistical Science*, 29, 559–578. [265]
- Okui, R., Small, D. S., Tan, Z., and Robins, J. M. (2012), "Doubly Robust Instrumental Variable Regression," *Statistica Sinica*, 173–205. [265]

- O'Malley, A. J., Elwert, F., Rosenquist, J. N., Zaslavsky, A. M., and Christakis, N. A. (2014), "Estimating Peer Effects in Longitudinal Dyadic Data Using Instrumental Variables," *Biometrics*, 70, 506–515. [265]
- Panofsky, A. L. (2011), "Field Analysis and Interdisciplinary Science: Scientific Capital Exchange in Behavior Genetics," *Minerva*, 49, 295–316. [265]
- Papachristos, A. V., Hureau, D., and Braga, A. A. (2013), "The Corner and the Crew: The Influence of Geography and Social Networks on Gang Violence," *American Sociological Review*, 78, 417–447. [268]
- Pearl, J. (2000), *Causality: Models, Reasoning, and Inference*, New York: Cambridge University Press. [265]
- Peng, T.-Q. (2015), "Assortative Mixing, Preferential Attachment, and Triadic Closure: A Longitudinal Study of Tie-Generative Mechanisms in Journal Citation Networks," *Journal of Informetrics*, 9, 250–262. [265,266]
- Robins, G., Pattison, P., Kalish, Y., and Lusher, D. (2007a), "An Introduction to Exponential Random Graph (p^*) Models for Social Networks," *Social Networks*, 29, 173–191. [268]
- Robins, G., Pattison, P., and Wang, P. (2009), "Closure, Connectivity and Degrees: New Specifications for Exponential Random Graph (p^*) Models for Directed Social Networks," *Social Networks*, 31, 105–117. [265,268]
- Robins, G., Snijders, T., Wang, P., Handcock, M., and Pattison, P. (2007b), "Recent Developments in Exponential Random Graph (p^*) Models for Social Networks," *Social Networks*, 29, 192–215. [265,268]
- Robins, J. M., Hernan, M. A., and Brumback, B. (2000), "Marginal Structural Models and Causal Inference in Epidemiology," *Epidemiology*, 11, 550–560. [265]
- Rogers, E. M. (2010), *Diffusion of Innovations*, Simon and Schuster. [269,275]
- Rosenbaum, P., and Rubin, D. (1983), "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika*, 70, 41–55. [265]
- Rosenbaum, P. R. (2002), *Observational Studies*, New York: Springer. [265]
- (2010), *Observational Studies* (2nd ed.), New York: Springer. [265]
- Shi, F., Foster, J., and Evans, J. (2015), "Weaving the Fabric of Science: Dynamic Network Models of Science's Unfolding Structure," *Social Networks*, 43, 73–85. [266]
- Small, D., Ten Have, T., and Rosenbaum, P. R. (2008), "Randomization Inference in a Group-Randomized Trial of Treatments for Depression: Covariate Adjustment, Noncompliance and Quantile Effects," *Journal of the American Statistical Association*, 103, 271–279. [265]
- Small, D. S. (2007), "Sensitivity Analysis for Instrumental Variables Regression with Overidentifying Restrictions," *Journal of the American Statistical Association*, 102, 1049–1058. [265]
- Small, D. S., and Rosenbaum, P. R. (2008), "War and Wages: The Strength of Instrumental Variables and Their Sensitivity to Unobserved Biases," *Journal of the American Statistical Association*, 103, 924–933. [265]
- Small, D. S., Tan, Z., Lorch, S. A., and Brookhart, M. A. (2014), "Instrumental Variable Estimation When Compliance is not Deterministic: The Stochastic Monotonicity Assumption," unpublished manuscript, available at <http://xxx.tau.ac.il/abs/1407.7308>. [265]
- Stigler, S. M. (1994), "Citation Patterns in the Journals of Statistics and Probability," *Statistical Science*, 9, 94–108. [265,266]
- Stuart, E. A. (2010), "Matching Methods for Causal Inference: A Review and a Look Forward," *Statistical Science*, 25, 1. [265]
- van Duijn, M. A., Gile, K. J., and Handcock, M. S. (2009), "A Framework for the Comparison of Maximum Pseudo-likelihood and Maximum Likelihood Estimation of Exponential Family Random Graph Models," *Social Networks*, 31, 52–62. [269]
- Varin, C., Cattelan, M., and Firth, D. (2016), "Statistical Modelling of Citation Exchange between Statistics Journals," *Journal of Royal Statistical Society, Series A*, 179, 1–33. [265,266]
- Wagner, C. S., Roessner, J. D., Bobb, K., Klein, J. T., Boyack, K. W., Keyton, J., Rafols, I., and Börner, K. (2011), "Approaches to Understanding and Measuring Interdisciplinary Scientific Research (IDR): A Review of the Literature," *Journal of Informetrics*, 5, 14–26. [265]
- Walker, D., Xie, H., Yan, K., and Maslov, S. (2007), "Ranking Scientific Publications using a Model of Network Traffic," *Journal of Statistical Mechanics*, P06010. [266]
- Wasserman, S., and Pattison, P. E. (1996), "Logit Models and Logistic Regressions for Social Networks: I. An Introduction to Markov Graphs and p^* ," *Psychometrika*, 61, 401–425. [268]
- Watts, D. J. (1999), "Networks, Dynamics, and the Small-World Phenomenon," *American Journal of Sociology*, 105, 493–527. [273]
- Watts, D. J., and Strogatz, S. H. (1998), "Collective Dynamics of 'Small-World' Networks," *Nature*, 393, 440–442. [273]
- Yan, E., Ding, Y., Cronin, B., and Leydesdorff, L. (2013), "A Bird's-eye View of Scientific Trading: Dependency Relations Among Fields of Science," *Journal of Informetrics*, 7, 249–264. [266]