Twin Papers: A Simple Framework of Causal Inference for Citations via Coupling

Ryoma Sato
r.sato@ml.ist.i.kyoto-u.ac.jp
Kyoto University / RIKEN AIP
Kyoto, Japan

Makoto Yamada
myamada@i.kyoto-u.ac.jp
Kyoto University / RIKEN AIP
Kyoto, Japan

Hisashi Kashima
kashima@i.kyoto-u.ac.jp
Kyoto University / RIKEN AIP
Kyoto, Japan

ABSTRACT

The research process includes many decisions, e.g., how to entitle and where to publish the paper. In this paper, we introduce a general framework for investigating the effects of such decisions. The main difficulty in investigating the effects is that we need to know counterfactual results, which are not available in reality. The key insight of our framework is inspired by the existing counterfactual analysis using twins, where the researchers regard twins as counterfactual units. The proposed framework regards a pair of papers that cite each other as twins. Such papers tend to be parallel works, on similar topics, and in similar communities. We investigate twin papers that adopted different decisions, observe the progress of the research impact brought by these studies, and estimate the effect of decisions by the difference in the impacts of these studies. We release our code and data, which we believe are highly beneficial owing to the scarcity of the dataset on counterfactual studies.

CCS CONCEPTS
• Information systems → Decision support systems; Data mining.

KEYWORDS
causal inference, counterfactual studies, scholarly communication

ACM Reference Format:
Ryoma Sato, Makoto Yamada, and Hisashi Kashima. 2022. Twin Papers: A Simple Framework of Causal Inference for Citations via Coupling. In Proceedings of the 31st ACM International Conference on Information and Knowledge Management (CIKM ’22), October 17–21, 2022, Atlanta, GA, USA. ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3511808.3557716

1 INTRODUCTION

It has been studied for a long time what aspects of research processes affect the number of citations [6, 16, 18, 23]. Namely the publication venues [18, 25, 29–31], authors [12, 29, 31], titles [5, 5, 14, 21, 22], references [27, 28], and topological features [7, 32] have been considered as the cause of citations. For example, Yan et al. [31] argue that the authors’ expertise and venue impact are important factors, and Paiva et al. [17] found that articles with short titles describing the results were cited more often.

Except for a notable exception of Davis et al. [6], who conducted a randomized control trial for investigating the impact of the choice of open access, most studies are based on observational studies. This is primarily because intervening research processes, e.g., by randomly changing publication venues or titles, may cause adverse impacts on the researchers’ careers. For this reason, most of the existing studies investigate only correlations. Although some studies [8, 18, 25] tried to find causal relations, they assumed specific statistical models and covariates. However, in general, the choice of covariates is not straightforward and crucially affects the results of analysis [26]. In this paper, we propose a simple framework for adjusting confounders and thereby enabling us to find causal relationships in research processes and citations. Our framework can also be used for screening important factors before manual analysis.

Reproducibility: Our code and the list of twin papers are available at https://github.com/joisin/o/twinpaper.

2 OUR APPROACH

Let us consider a binary decision in the research process (e.g., whether to use a colon in the title, or publishing the paper in CIKM or SIGIR). We use whether to use a colon in the title as a running example. The goal of this study is to investigate whether a colon in the title increases the number of citations, and if any, how many citations. We consider a potential outcome framework for causal estimation, where the outcome is defined as the base-2 logarithm (instead of the raw value because of its broad dynamic range) of the number of citations a paper receives after a certain period. We say a paper x receives a treatment if a colon is used in the title of x. There are two possible outcomes $Y_x(1)$ and $Y_x(0)$, the outcome value if the paper receives (resp. does not receive) the treatment.

The quantity we want to estimate is:

$$ITE_x \overset{\text{def}}{=} Y_x(1) - Y_x(0),$$

$$ATE \overset{\text{def}}{=} \mathbb{E}_{x \sim p(x)}[ITE_x] = \mathbb{E}_{x \sim p(x)}[Y_x(1) - Y_x(0)],$$

i.e., how much the treatment increases the outcome in expectation. However, the critical problem is that we can observe only one of the two outcomes because we cannot publish the same paper with and without a colon simultaneously. Let $Y_x^F$ and $Y_x^C$ be the factual and counterfactual outcome values, respectively, i.e., if paper x is published with a colon in the title, $Y_x^F = Y_x(1), Y_x^C = Y_x(0)$, and otherwise, $Y_x^F = Y_x(0), Y_x^C = Y_x(1)$. One cannot obtain $ITE_x$ because $Y_x^C$ is not observable.
3 METHOD

Our proposed framework is inspired by the causal inference framework based on twins [15] in the medical and psychological domains. The key insight of our proposed framework, twin papers, is that we can roughly regard a pair of papers that cite each other as counterfactual units. We call such a pair of papers twins. The rationale behind this definition is that twin papers tend to be (1) parallel works, (2) on similar topics, and (3) in close communities, which we will empirically show in the experiments. Therefore, twin papers can adjust many, if not all, confounders, including observable and unobservable ones. If the numbers of citations the twin papers receive are different, we can investigate what made the difference. Suppose a paper \( x \) was published with a colon in the title and has a twin paper \( y \) which was published without a colon. Then, we can estimate ITE by

\[
\text{ITE} = Y_F - Y_I.
\]

This value can be computed solely from factual values. However, this estimate is noisy and has a high variance. Therefore, we consider the average effect, i.e., ATE. Let \( D_{\text{colon/no colon}} = \{(s, t) \mid s \text{ and } t \text{ are twins, and } s \text{ has a colon, } t \text{ has no colon}\} \). Then, ATE can be estimated by

\[
\text{ATE}_{\text{colon/no colon}} = \frac{1}{|D_{\text{colon/no colon}}|} \sum_{(s, t) \in D_{\text{colon/no colon}}} Y_F^s - Y_I^t.
\]

This value can be computed solely from factual values.

We gather twin papers from the dblp dataset [24]. There are 87,396 twins in total, which are available in https://github.com/joisino/twinpaper.

4 ILLUSTRATIVE EXAMPLE

To illustrate the benefit of twin papers, we create a subset of the dataset that contains only papers published in Symposium on the Theory of Computing (STOC), Symposium on Foundations of Computer Science (FOCS), Neural Information Processing Systems (NeurIPS), and International Conference on Machine Learning (ICML). STOC and FOCS are prestigious venues in theoretical computer science, and NeurIPS and ICML are prestigious venues in machine learning. As an example, we consider if adding a word “learning” in the title has a positive effect on the impact. We consider a paper with “learning” in the title to be treated and that without it is controlled. Intuitively, just changing the title of this paper to “Twin Papers: A Simple Learning Framework...” would not change the number of citations much. Therefore, we expect the effect is small or zero. A naive approach to estimating the effect with observational data is,

\[
\bar{ATE}_{\text{observational}} = \frac{1}{|\{i \text{ is treated}\}|} \sum_{i \text{ is treated}} Y_I^i - \frac{1}{|\{i \text{ is controlled}\}|} \sum_{i \text{ is controlled}} Y_I^i.
\]

However, there is a selection bias because papers in NeurIPS and ICML tend to have “learning” in the title. In fact, \( \bar{ATE}_{\text{observational}} = 0.132 \), which indicates that the treatment has a positive effect. This result just reflects the fact that NeurIPS and ICML papers tend to receive more citations than STOC and FOCS papers. By contrast, if we use twin papers and the proposed estimator (i.e., (3)), \( \bar{ATE} = -0.017 \), which indicates the treatment has no effects.

5 CONFIRMING ASSUMPTIONS

5.1 Twins Tend to Be Parallel Works

Figure 1 (a) shows the histogram of the differences of publication years between twin papers. This indicates that 84.8 percent of twin pairs are published in the same or the next year. However, some twin papers are published in different periods. We investigate the cause of this phenomenon. We draw random twin pairs whose publication years are different by more than five years and show them in Table 1. The difference in the first example is as many as twenty years. We found out that this is because there is a paper with the same title as “Fast Fourier transforms for nonequispaced data” published in 1993. The dblp dataset confused these papers, maybe in the data processing process, and spurious twins are detected. Other examples were caused due to similar reasons. Overall, twins that are not parallel works are spurious twins caused by noise in the dataset. Optionally, we can remove such pairs by preprocessing, e.g., thresholding the difference of publication years by one or two. We use the original data in the following analysis because such cases are rare, as shown in Figure 1, and do not affect the results much.
We build a collaboration network, where a node is a researcher, and we compute L1-normalized L1 bag-of-words distances [19] of the abstracts of twin papers. Twin papers are parallel works, their publication dates are slightly different. We consider the one published earlier as the treatment and the other as the control. We remove the pairs with the same title length. The first row of Table 2 shows that shortening the title is slightly better, which is consistent with Ayres and Vars [4], but the effect is small. The effect of longer titles has been a controversial topic in the informetrics domain [10, 13, 14, 20, 22], and sometimes the opposite effects have been confirmed [10, 13]. The small effect observed in this analysis is consistent with the literature.

A paper with more references may have more chances of reverse lookups of references. The third row of Table 2 shows that lengthening the reference has a moderately positive effect on the number of citations. This result is consistent with the findings of Falagas et al. [8].

Finally, self citation is a common strategy to increase the number of citations [2, 9]. Self citations do not only increase the number of citations directly but also improve the exposure. Furthermore, many scholarly search engines such as Google Scholar and Semantic Scholar provide citation numbers in the search results, and the increase of citation numbers will increase the chances of clicks. We consider a paper is treated if the paper is cited by a paper that has at least one common author. The sixth row of Table 2 shows that self citation has a strong positive effect on the number of citations. This is consistent with the findings of Fowler and Aksnes [9].

Priority. Although twin papers are parallel works, their publication dates are slightly different. We consider the one published earlier is treated in this analysis. Surprisingly, the estimated ATE was −0.187, which means that earlier publications receive slightly fewer citations. We hypothesize that this is because the quality of later publications is better. At least, this result indicates that hurrying to publish does not benefit in the long run.

Venue. Publishing a paper in a conference or journal makes the paper known in the community and has an effect on the impact of the paper in the community. As each venue has different readers and participants, different venues may have different effects. We investigate the impact of the choice of venue in this section. First, for each pair (a, b) of venues, we construct \( D_{a/b} = \{ (s, t) \mid s \text{ and } t \text{ are twin, } s \text{ is published in } a, t \text{ is published in } b \} \).

| Treatment | \(|D|\) | ATE |
|-----------|-----|----|
| Including a Colon in the Title | 21080 | 0.356 |
| Lengthening the Title | 84970 | -0.126 |
| Lengthening the reference | 81857 | 0.710 |
| Lengthening the abstract | 82917 | 0.248 |
| Lengthening the paper | 65730 | 0.630 |
| Self citation | 10582 | 1.30 |
We confirmed that twin papers adjusted three conditions in Section 5. We argue that much more conditions are adjusted by twin papers. For example, the research problems they tackle are considered to be the same or similar. Besides, we hypothesize that the qualities of papers would also be adjusted to some extent, if not totally, because too low-quality papers are unlikely to be cited. Importantly, the quality of a paper is difficult to quantify, and thus we cannot numerically validate this hypothesis. We argue that the ability to control such unobservable/quantifiable confounders is the strength of twin papers because other methods such as multivariate analysis cannot handle them.

### 7.2 Limitations
First, twin papers do not necessarily control all confounding factors. For example, if authors decide the venue, and after that, they decide to add a colon in the title following the custom of the venue, then, the choice of the venue becomes a confounding factor. In this case, one needs to adjust the confounding factor using auxiliary features. We stress that our framework is general and can be combined with other adjustment methods such as multivariate analysis and stratified analysis, and importantly, the strength of twin papers is that it can adjust many, if not all, factors with a simple procedure.

Second, strictly speaking, twin papers are not true counterfactual results. In reality, if two papers have similar topics and cite each other, the research impacts of these papers affect one another. This limitation is common with the original study on twins. However, the twin paper framework is much less sensitive to biases than previous studies using random samples from observational data. Combining our framework with manual analysis, e.g., multivariate analysis and stratified analysis, will further mitigate this problem.

The third limitation is that twins are rare in some domains. We found that the data mining domain had few twins. We hypothesize that this is because many data mining conferences prohibit submitting papers to arXiv during submission, and it hinders authors from finding concurrent papers on the same topic.

### 8 CONCLUSION
In this paper, we proposed a simple framework for investigating the effect of the decisions in research processes. We empirically confirm that twin papers are published under similar conditions, and conduct several case studies on the effects on the contents of the paper and publication venues.

### ACKNOWLEDGMENTS
This work was supported by JSPS KAKENHI Grant Number 21J22490 and JST CREST Grant Number JPMJCR21D1.
REFERENCES

[1] Giovanni Abramo, Ciriaco Andrea D’Angelo, and Giovanni Felici. Predicting publication long-term impact through a combination of early citations and journal impact factor. Journal of Informetrics, 13(1):32–49, 2019.

[2] Dag W Aksnes. A macro study of self-citation. Scientometrics, 56(2):235–246, 2003.

[3] Anupama Annalingam, Hastha Damayanthi, Ranil Jayawardena, and Priyanga Ranasinghe. Determinants of the citation rate of medical research publications from a developing country. Springerplus, 3(1):1–6, 2014.

[4] Ian Ayres and Frederick E Vars. Determinants of citations to articles in elite law reviews. Journal of Legal Studies, 29(5):427–450, 2000.

[5] Reindert K Buter and Anthony FJ van Raan. Non-alphanumeric characters in titles of scientific publications: An analysis of their occurrence and correlation with citation impact. Journal of Informetrics, 5(4):608–617, 2011.

[6] Philip M Davis, Bruce V Lewenstein, Daniel H Simon, James G Booth, and Mathew J Connolly. Open access publishing, article downloads, and citations: randomised controlled trial. BMJ 337, 2008.

[7] Feruz Davletov, Ali Selman Aydin, and Ali Cakmak. High impact academic paper prediction using temporal and topological features. In Proceedings of the 22nd ACM International Conference on Information and Knowledge Management, CIKM, pages 491–498. ACM, 2014.

[8] Matthew E Falagas, Angeliki Zarkali, Drosos E Karageorgopoulos, Vangelis Barkas, and Michael N Mavros. The impact of article length on the number of future citations: a bibliometric analysis of general medicine journals. PLoS One, 8(2):1–8, 2013.

[9] James Hartley. Planning that title: Practices and preferences for titles with colons. Perspectives on psychological science, 5(3):348–362, 2009.

[10] Thomas S Jacques and Neil J Sebire. The impact of article titles on citation hits: studies of non-alphanumeric characters in titles of scientific publications. Evolutionary Psychology, 7(3):348–362, 2009.

[11] Thomas S Jacques and Neil J Sebire. The impact of article titles on citation hits: studies of non-alphanumeric characters in titles of scientific publications. Evolutionary Psychology, 7(3):348–362, 2009.

[12] Yuxin Xiao, Adit Krishnan, and Hari Sundaram. Discovering strategic behaviors for collaborative content-production in social networks. In Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD, pages 990–998. ACM, 2008.

[13] Vincent A Traag. Inferring the causal effect of journals on citations. Quantitative Science Studies, 2(2):496–504, 2021.

[14] Gregory D Webster, Peter K Jonason, and Tatiana Orozco Schember. Hot topics and popular papers in evolutionary psychology: Analyses of title words and citation counts in evolution and human behavior, 1979–2008. Evolutionary Psychology, 7(3):348–362, 2009.

[15] Shuai Xiao, Junchi Yan, Changsheng Li, Bo Jin, Xiaokang Yang, Stephen M. Chu, and Hongyuan Zha. On modeling and predicting individual paper citation count over time. In Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI, pages 2676–2682, 2016.

[16] Yuexuan Xiao, Adit Krishnan, and Hari Sundaram. Discovering strategic behaviors for collaborative content-production in social networks. In The Web Conference, WWW, pages 2078–2088, 2020.

[17] Rui Yan, Jie Tang, Xiaoming Liu, Dongdong Shan, and Xiaoming Li. Citation count prediction: learning to estimate future citations for literature. In Proceedings of the 20th ACM International Conference on Information and Knowledge Management, CIKM, pages 1247–1252, 2011.

[18] Xiao Yu, Quanquan Gu, Mianwei Zhou, and Jiawei Han. Citation prediction in heterogeneous bibliographic networks. In Proceedings of the Twelfth SIAM International Conference on Data Mining, SDM, pages 1119–1130, 2012.