双子の論文 カップリングによる引用の因果推論の簡単なフレームワーク

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

研究プロセスには、例えば、論文のタイトルをどうするか、どこで発表するかなど、多くの決定が含まれる。本論文では、そのような決定の効果を調査するための一般的なフレームワークを紹介する。効果を調査する際の主な困難は、現実には入手できない反実仮想の結果を知る必要があることである。我々のフレームワークの重要な洞察は、研究者が双子を反実仮想の単位とみなす、双子を用いた既存の反実仮想分析に触発されたものである。提案するフレームワークは、互いに引用し合う一組の論文を双子と見なすものである。このような論文は、類似のテーマで、類似のコミュニティで、並行した作品であることが多い。我々は、異なる決定を採用した双子の論文を調査し、これらの研究がもたらす研究インパクトの進展を観察し、これらの研究の影響の違いによって決定の効果を推定する。また、反実仮想研究に関するデータセットが少ないため、非常に有益であると考えられる我々のコードとデータを公開する。

CCS CONCEPTS

• **Information systems** → **Decision support systems**; **Data mining**.

KEYWORDS

因果推論、反実仮想データ、学術コミュニケーション

ACMリファレンスフォーマット 佐藤竜馬、山田誠、鹿島久志. 2022. Twin Papers: カップリングによる引用の因果推論のシンプルなフレームワーク. 第31回ACM国際情報・知識マネジメント会議(CIKM '22) 予稿集, 2022年10月17日～21日, アトランタ, GA, USA. ACM, New York, NY, USA, 5ページ https://doi.org/10.1145/3511808.3557716

1 INTRODUCTION

研究プロセスのどのような側面が被引用数に影響を与えるかは、長い間研究されてきた[6, 16, 18, 23]。すなわち、出版会場[18, 25, 29–31]、著者[12, 29, 31]、タイトル[3, 5, 14, 21, 22]、参考文献[27, 28]、トポロジカル特徴[7, 32]が引用の原因として考えられてきた。例えば、Yanら[31]は、著者の専門性と会場の影響はまた、Paiva ら[17]は、結果を説明する短いタイトルの論文がより多く引用されることを発見した。オープンアクセスの選択の影響を調査するためにランダム化比較試験を行ったDavisら[6]の顕著な例外を除いて、ほとんどの研究は観察研究に基づいている。これは主に、出版場所やタイトルをランダムに変更するなど、研究プロセスに介入することで、研究者のキャリアに悪影響を与える可能性があるためである。このため、既存の研究の多くは、相関関係のみを調査している。因果関係を見出そうとする研究もあるが[8, 18, 25]、それらは特定の統計モデルや共変量を仮定している。しかし、一般に共変量の選択は簡単ではなく、分析結果に決定的な影響を与える[26]。本論文では、交絡因子を調整することで、研究過程や引用における因果関係を見出すことを可能にするシンプルなフレームワークを提案する。また、本フレームワークは、手動分析の前に重要な因子をスクリーニングするために使用することができる。

再現性。再現性:コードと双子論文リストは https://github.com/joisino/twinpaper で公開されている。

2 OUR APPROACH

研究プロセスにおける二項対立の判断(例えば、タイトルにコロンを使うか、CIKMやSIGIRで論文を発表するか)を考えてみよう。ここでは、タイトルにコロンを使うかどうかを実行例として使用する。本研究の目的は、タイトルにコロンがあることで被引用数が増加するかどうか、また、増加した場合はそれも増加するかどうかを調査することである。因果関係を推定するための潜在的な結果の枠組みを考える。ここで、結果は、ある論文がある期間後に受け取った被引用数の(広いダイナミックレンジを持つため、生の値ではなく)ベース2の対数として定義される。論文のタイトルにコロンが使われた場合、論文xは治療を受けたと言う。論文が治療を受けた(あるいは受けなかった)場合の結果値である2つの可能な結果$Y_x(1)$と$Y_x(0)$が存在する。推定したい量は

$$\mathrm{ITE}_x \stackrel{\mathrm{def}}{=} Y_x(1) - Y_x(0), \tag{1}$$

$$\mathrm{ATE} \stackrel{\mathrm{def}}{=} \mathbb{E}_{x\sim p(x)}[\mathrm{ITE}_x] = \mathbb{E}_{x\sim p(x)}[Y_x(1) - Y_x(0)], \tag{2}$$

すなわち、治療が期待される結果がどれだけ増加するかということである。しかし、コロンがある場合とない場合で同じ論文を同時に出版できないため、重要な問題は、2つの結果のうち1つしか観察できないことである。Y_x^FとY_x^Cをそれぞれ事実上の結果と反実仮想の結果の値とする、すなわち、論文xがタイトルにコロンを付けて出版された場合、$Y_x^F = Y_x(1)$, $Y_x^C = Y_x(0)$, それ以外は$Y_x^F = Y_x(0)$, $Y_x^C = Y_x(1)$. Y_x^Cは観測できないので、ITE_xを得ることはできない。

CIKM '22, October 17–21, 2022, Atlanta, GA, USA

ACM ISBN 978-1-4503-9236-5/22/10…$15.00
https://doi.org/10.1145/3511808.3557716

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 data, 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 [3, 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/joisino/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:

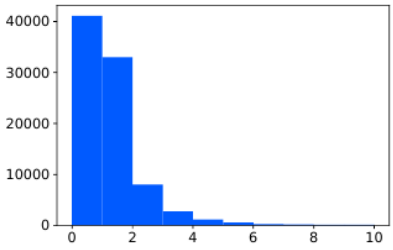
$$\mathrm{ITE}_x \stackrel{\mathrm{def}}{=} Y_x(1) - Y_x(0), \tag{1}$$

$$\mathrm{ATE} \stackrel{\mathrm{def}}{=} \mathbb{E}_{x\sim p(x)}[\mathrm{ITE}_x] = \mathbb{E}_{x\sim p(x)}[Y_x(1) - Y_x(0)], \tag{2}$$

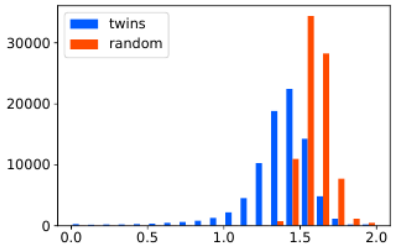
i.e., how much the treatment increases the outcome in expectation. However, the critical problem is that we can observe only one of the two outcomes because we cannot publish the same paper with and without a colon simultaneously. Let Y_x^F and Y_x^C be the factual and counterfactual outcome values, 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.

CIKM '22, October 17–21, 2022, Atlanta, GA, USA

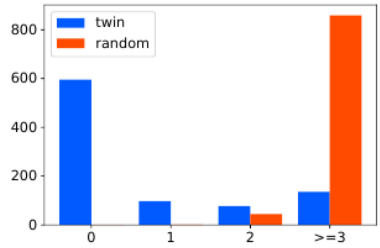
ACM ISBN 978-1-4503-9236-5/22/10…$15.00
https://doi.org/10.1145/3511808.3557716



（a）出版年の違い



(b) Distances of abstracts.



(c) Distances of twins on the collaboration network.

図1：(a)ほとんどの双子の論文は並行研究である。(b)ほとんどの双子は似たようなトピックを持っている。(c)ほとんどの双子は近いコミュニティに属している。これらの結果は、双子の論文の枠組みの前提を裏付けるものである。

3 METHOD

我々の提案するフレームワークは、医学と心理学の領域における双子に基づく因果推論フレームワーク[15]に触発されたものである。我々の提案するフレームワークの重要な洞察である双子論文は、互いに引用し合う一組の論文を反実仮想的な単位とおおよそみなすことができる、ということである。このような論文のペアを双子と呼ぶ。この定義の根拠は、双子論文は(1)並行研究、(2)類似したトピックに関する研究、(3)近いコミュニティで行われる傾向があり、これを実験で経験的に示すことにする。したがって、双子論文は、観測可能な交絡因子や観測不可能な交絡因子を含む、全てではないにしても、多くの交絡因子を調整することができる。もし、双子論文の被引用数が異なれば、何が違いを生んだかを調べることができる。例えば、タイトルにコロンが書かれていて、コロンなしで出版された双子論文yがあるとする。すると、ITEは次のように推定できる。

$$\widehat{\mathrm{ITE}} = Y_x^F - Y_y^F.$$

この値は、事実値のみから計算することができる。しかし、この推定値はノイズが多く、分散が大きい。そこで、平均効果、すなわちATEを考える。$\mathcal{D}^{colon/nocolon}$ = {(s, t) ∼ s, tは双子、sはコロン、tはコロン無し}とする。すると、ATEは次のように推定できる。

$$\widehat{\mathrm{ATE}}^{\text{colon/no colon}} = \frac{1}{|\mathcal{D}^{\text{colon/no colon}}|} \sum_{(s,t)\in\mathcal{D}^{\text{colon/no colon}}} Y_s^F - Y_t^F. \tag{3}$$

この値は、事実値のみから計算することができる。dblpデータセット[24]から双子論文を収集する。双子論文は全部で87, 396件あり、https://github.com/ joisino/twinpaperで公開されている。

4 ILLUSTRATIVE EXAMPLE

双子論文の利点を説明するために、Symposium on the Theory of Computing (STOC), Symposium on Foundations of Computer Science (FOCS), Neural Information Processing Systems (NeurIPS), International Conference on Machine Learning (ICML) に掲載された論文のみを含むデータセットのサブセットを作成する。STOCとFOCSは理論計算機科学の権威ある場であり、NeurIPSとICMLは機械学習の権威ある場である。例として、ある単語を追加する場合について考えてみる。

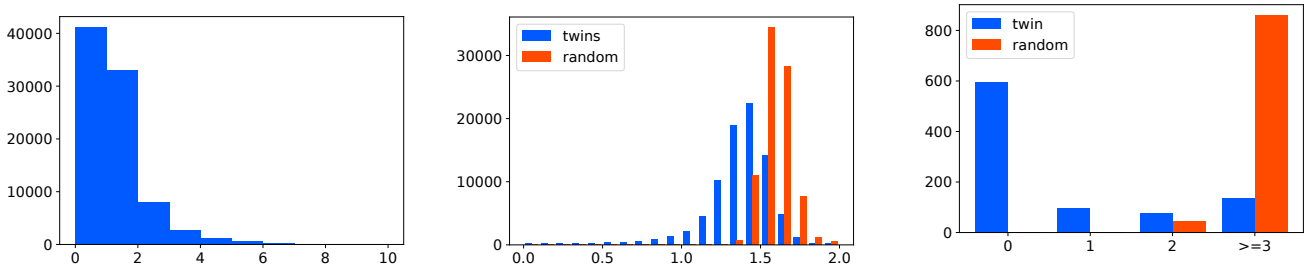
タイトルの "learning "は、インパクトにプラスの影響を与える。ここでは、扱うタイトルに "learning "があり、それなしには制御されている論文を考える。直感的には、この論文のタイトルを "Twin Papers: A Simple Learning Framework・・・"に変更しても、被引用回数はあまり変わらないだろう。したがって、その効果は小さいかゼロであると予想される。観測データで効果を推定する素朴な方法として

$$\widehat{ATE}_{\text{observational}} = \frac{1}{|\{i \text{ is treated}\}|} \sum_{i \text{ is treated}} Y_i^F - \frac{1}{|\{i \text{ is controlled}\}|} \sum_{i \text{ is controlled}} Y_i^F.$$

しかし、NeurIPSやICMLの論文はタイトルに「学習」する傾向があるため、選択バイアスがある。実際、AT_g $E_{observational}$ = 0.132となり、処理がプラスの効果を持つことがわかる。この結果は、NeurIPSやICMLの論文がSTOCやFOCSの論文よりも引用される傾向があることを反映しているに過ぎない。一方、双子論文と提案する推定量(すなわち(3))を用いた場合、AT_g E = -0.017となり、処理に効果がないことがわかる。

5 仮定の確認 5.1 双子は並列作品になる傾向がある

図 1(a)は、双子論文間の出版年の違いのヒストグラムである。この図から、双子のペアのうち84.8%が同じ年または翌年に出版されていることがわかる。しかし、双子の論文の中には、出版期間が異なるものもある。この現象の原因を調査する。出版年が5年以上異なる双子のペアをランダムに抽出し、表1に示す。最初の例では、20年分の差である。これは、1993年に出版された「非等間隔データに対する高速フーリエ変換」と同じタイトルの論文があるためであることがわかった。dblpデータセットでは、これらの論文がデータ処理の過程で混乱し、偽の双子が検出される可能性があります。他の例も同様の理由で発生した。全体として、並列でない双子はデータセットのノイズに起因する偽双子である。オプションとして、出版年の差を1つか2つずつ閾値処理するなど、前処理によってそのようなペアを削除することができます。以下の分析では、図1に示すように、このようなケースは稀であり、結果にあまり影響を与えないため、元のデータを使用する。



(a) Difference of publication years.

(b) Distances of abstracts.

(c) Distances of twins on the collaboration network.

Figure 1: (a) Most twin papers are parallel works. (b) Most twins have similar topics. (c) Most twins belong to close communities. These results corroborate the assumptions of the twin paper framework.

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

$$\widetilde{\mathrm{ITE}} = Y_x^F - Y_y^F.$$

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 $\mathcal{D}^{\text{colon/no colon}} = \{(s,t) \mid s \text{ and } t \text{ are twins, and } s \text{ has a colon, } t \text{ has no colons}\}$. Then, ATE can be estimated by

$$\widetilde{\mathrm{ATE}}^{\text{colon/no colon}} = \frac{1}{|\mathcal{D}^{\text{colon/no colon}}|} \sum_{(s,t)\in\mathcal{D}^{\text{colon/no colon}}} Y_s^F - Y_t^F. \tag{3}$$

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,

$$\widetilde{ATE}_{\text{observational}} = \frac{1}{|\{i \text{ is treated}\}|} \sum_{i \text{ is treated}} Y_i^F - \frac{1}{|\{i \text{ is controlled}\}|} \sum_{i \text{ is controlled}} Y_i^F.$$

However, there is a selection bias because papers in NeurIPS and ICML tend to have "learning" in the title. In fact, $\widetilde{ATE}_{\text{observational}} = 0.132$, 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)), $\widetilde{ATE} = -0.017$, 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.

表1:並行論文でない双子論文。表1:並行論文でない双子論文:データセットのノイズによるもの。

Paper A	paper B
非等間隔格子の高速フーリエ変換に関する注意（1998）オフライン侵入検知システムの解析。多目的遺伝的アルゴリズムのケーススタディ（2005）組成CTLモデル検査における数式依存等価性（1994）	非等間隔データに対する高速フーリエ変換（2018）遺伝的アルゴリズムを用いたオフライン導入型命令検出（2016）組成CTLモデル検査における数式依存の等価性（2002）

表 2: コンテンツと論文の ATE。直感的には、この表は、1列目の処理をすると、3列目の量だけ論文のインパクトが大きくなることを読み取ることができる。2列目はデータセットからの双子の数、例えばコロンの場合は$\Delta D^{\text{¤¤¤}/nocolon}$ Δを示す。

Treatment	$\lvert\mathcal{D}\rvert$	$\widehat{\text{ATE}}$
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

5.2 双子は同じトピックになる傾向がある

双子論文のアブストラクトの L1 正規化 L1 bag-of-words 距離[19]を計算する。そして、ランダムなペアに対して同じ距離を計算する。図1(b)のヒストグラムは、双子論文は類似した抄録を持つ傾向があることを示している。これは、双子論文が同じか類似のトピックに属している傾向があることを示している。

5.3 '双子は同じコミュニティーにいる傾向がある

ノードを研究者とし、エッジをdblpデータセットを用いて、2人の研究者が協働したことを示すコラボレーションネットワークを構築する。コラボレーションネットワークで親しい研究者は同じ研究コミュニティにいると考えられる。論文Aの著者と論文Bの著者がコラボレーションネットワークにおいて持つ最小距離として、2つの論文A、Bの距離を計算する。図1(c)より、コラボレーションネットワークでは、双子論文の著者が接近する傾向があることがわかる。

6 ANALYSIS WITH TWINS

目次とスタイル 情報量論的文献[3, 5, 14, 21, 22]で議論されているように、論文の内容やスタイルが論文の被引用数に影響を与える可能性がある。我々は6つの処理を定量的に調査している。まず、研究者によっては、自分の手法に名前を付け、コロンを付けて論文の冒頭に置くこともある。例えば、本論文では、"Twin Papers: A Simple."から始まっている。このような論文のタイトルはキャッチーであり、クリックの可能性をより多く提供することができる。表2の2行目から、タイトルにコロンを含めると、被引用数が若干向上することがわかる。この結果は、タイトルにコロンを追加することで被引用数にプラスの効果があると報告したButer and van Raan [5]の知見と一致する。なお、コロンの影響については議論があり[11, 14, 17]、ドメインに依存する可能性がある[5]。我々の分析はコンピュータサイエンスの論文(すなわち、dblp論文)で行われ、結論は他のドメインには適切でない可能性がある。しかし、我々のフレームワークは一般的であり、他のデータセットで使用すれば、他のドメインにも適用できることを強調する。第二に、論文のタイトルの長さは論文によって異なります。短いタイトルは理解しやすく、印象も強い。一方、長いタイトルは目に留まりやすく、検索エンジンにリストアップされる可能性が高い。どちらが優れているかを定量的に調査する。双子のペアそれぞれについて、タイトルの短い方を処理し、もう一方を制御していると考える。タイトルの長さが同じペアは削除している。表2の最初の行は、タイトルを短くする方が若干良いことを示しており、これはAyres and Vars [4]と一致するが、その効果は小さい。タイトルを長くする効果は、informetrics領域では議論のある話題であり[10, 13, 14, 20, 22]、時には逆の効果も確認されている[10, 13]。この分析で観察された小さな効果は、文献と一致している。参考文献が多い論文ほど、参考文献の逆引きの可能性が高いかもしれない。表2の3行目では、参考文献を長くすることで、引用回数に中程度の正の効果があることが示されている。この結果は、Haslamら[12]やOnodera and Yoshikane[16]の知見と一致する。次に、抄録の長さについて考察する。抄録が長い論文は、検索される確率が高い。表 2 の 4 行目は、抄録が長いほど若干プラスの効果があることを示している。次に、論文の長さについて調査する。論文が長いほど、内容や証拠が多いと考えられる。また、論文が長いほど、検索クエリに引っかかる可能性が高くなる。一方、読者は長すぎる論文を読むことに消極的である可能性がある。表2の5行目は、論文を長くすることが中程度の正の効果を持つことを示している。この結果は、Falagasら[8]の知見と一致する。最後に、自己引用は被引用数を増やすための一般的な戦略である[2, 9]。自己引用は、被引用数を直接的に増やすだけでなく、露出度を向上させる。さらに、Google ScholarやSemantic Scholarなどの学術検索エンジンでは、検索結果に引用数を提供しており、引用数の増加によりクリックの可能性が高まる。ここでは、少なくとも1人の共通著者がいる論文が被引用数を多くした場合、その論文を処理すると考える。表2の6行目は、自己引用が被引用数に対して強い正の効果を持つことを示している。これは、Fowler and Aksnes [9]の知見と一致する。

優先順位 双子論文は並行論文であるが、出版時期が若干異なる。本分析では、先に出版されたものを対象としている。意外なことに、推定されたATEは-0.187であり、これは初期の出版物の引用数がやや少ないことを意味する。これは、後期出版物の質が高いためと推測される。少なくとも、この結果は、出版を急いでも、長期的には利益が出ないことを示している。

会場。学会やジャーナルに論文を発表することで、その論文はコミュニティで知られるようになり、コミュニティでの論文の影響に影響を与える。会場によって読者や参加者が異なるため、会場によって効果が異なる場合があります。本節では、会場の選択の影響について検討する。まず、会場の各ペア(a,b)について、$D^{a/b}$ = {(s,t) ∼ s, t は双子、s は a, t は b で出版} と構成する。

Table 1: Twin papers that are not parallel works. They are due to noise of the dataset.

Paper A	paper B
A note on fast Fourier transforms for nonequispaced grids (1998)	Fast Fourier transforms for nonequispaced data (2018)
Analysis of an Off-Line Intrusion Detection System: A Case Study in Multi-Objective Genetic Algorithms (2005)	IMPROVED OFF-LINE INTRUSION DETECTION USING A GENETIC ALGORITHM (2016)
Formula-Dependent Equivalence for Compositional CTL Model Checking (1994)	Formula-Dependent Equivalence for Compositional CTL Model Checking (2002)

Table 2: ATEs for the contents and papers. Intuitively, this table reads that if we make the treatment in the first column, the paper has more impact by the amount in the third column. The second column shows the number of twins from the dataset, e.g., $|\mathcal{D}^{\text{colon/no colon}}|$ in the colon case.

Treatment	$\lvert\mathcal{D}\rvert$	$\widetilde{\text{ATE}}$
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

5.2 Twins Tend to Be on the Same Topic

We compute L1-normalized L1 bag-of-words distances [19] of the abstracts of twin papers. We then compute the same distances for random pairs. The histograms in Figure 1 (b) show that twin papers tend to have similar abstracts. This indicates that twins tend to be on the same or similar topics.

5.3 Twins Tend to Be in the Same Community

We build a collaboration network, where a node is a researcher, and an edge indicates that two researchers have collaborated, using the dblp dataset. Researchers close in the collaboration networks are considered to be in the same research community. We compute the distance of two papers A and B as the minimum distance between the authors of paper A and the authors of paper B in the collaboration network. Figure 1 (c) shows that the authors of twin papers tend to be close in the collaboration network.

6 ANALYSIS WITH TWINS

Contents and Styles. As discussed in the informetrics literature [3, 5, 14, 21, 22], the contents and style of the paper may affect the number of citations of the paper. We investigate six treatments quantitatively.

First, some researchers name their method and put it at the beginning of the paper with a colon. For example, this paper starts with "Twin Papers: A Simple..." Such paper titles are catchy and may provide more chances of clicks. The second row of Table 2 shows that including a colon in the title slightly improves the number of citations. This finding is consistent with the findings of Buter and van Raan [5], who reported that adding a colon in the title had a positive effect on the number of citations. We note that the impact of a colon has been controversial [11, 14, 17] and may depend on domains [5]. Our analysis is done with computer science papers (i.e., dblp papers), and the conclusion may not be appropriate for other domains. However, we emphasize that our framework is general and can be applied to other domains if used with other datasets.

Second, the lengths of paper titles vary from paper to paper. Short titles are easy to understand and provide strong impressions, whereas long titles have more chances to be caught in the eye and to be listed in search engines. We investigate which is better quantitatively. For each pair of twins, we consider that the one with the shorter title is treated and the other is controlled. We remove the pairs with the same title lengths. 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 Haslam et al. [12] and Onodera and Yoshikane [16].

Then, we consider the length of the abstract. A paper with a longer abstract has more chance of being searched. The fourth row of Table 2 shows that longer abstract has a slightly positive effect.

Next, we investigate the length of the paper. A longer paper is considered to have more content and evidence. Besides, the longer the paper is, the more chances it has to be caught by search queries. On the other hand, readers may be reluctant to read too long papers. The fifth row of Table 2 shows that lengthening the paper has a moderately positive effect. 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 a 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 $\mathcal{D}^{a/b} = \{(s, t) \mid s \text{ and } t \text{ are twin, } s \text{ is published in } a, t \text{ is published in } b\}$.

表3:出版会場のATE。ATEが正となるように処理を選んでいる。したがって、直感的には、この表は、1列目の会場が2列目の会場より4列目の金額だけ優れていると読み取れる。

Treatment (a)	Control (b)	$\lvert\mathcal{D}^{a/b}\rvert$	$\widetilde{\text{ATE}}$
Journal of Cognitive Neuroscience	NeuroImage	817	0.539
IEEE Transactions on Information Theory	International Symposium on Information Theory	459	1.93
Neural Computation	IEEE Transactions on Neural Networks	216	0.632
Neural Computation	Neural Networks	199	0.76
Symposium on the Theory of Computing	Foundations of Computer Science	182	0.252
IEEE Transactions on Signal Processing	International Conference on Acoustics, Speech, and Signal Processing	178	2.78
Neural Computation	Neurocomputing	153	2.41
Journal of Economic Theory	Games and Economic Behavior	125	0.668
IEEE ACM Transactions on Networking	International Conference on Computer Communications	110	1.29
Symposium on the Theory of Computing	Symposium on Discrete Algorithms	105	0.693

表4:処理の組み合わせのATE。直感的には、この表は、1列目と2列目の処理をした場合、3列目の量だけ、論文のインパクトが大きくなることを読み取ることができる。

Treatment A	Treatment B	$\lvert\mathcal{D}\rvert$	$\widetilde{\text{ATE}}$
Lengthen the reference	Lengthen the paper	41546	1.04
Lengthen the reference	Self citation	6473	1.75
Lengthen the paper	Self citation	5018	1.66

式(3)に基づき、a での出版が b での出版を上回る ATE を推定する。表3は、$D^{a/b}$が10大きい会場の結果である。IEEE Transaction on Signal Processingでの出版は、International Conference on Acoustics, Speech, and Signal Processing (ICASSP)での出版と比較して、多くの引用を誘発していることが分かる。また、Symposium on the Theory of Computing (STOC)での出版は、Foundations of Computer Science (FOCS)での出版と同等であるが、STOCの方が若干優れている。一方、STOCはSymposium on Discrete Algorithms (SODA)よりも明らかに優れている。

6.1 Are the effects additive?

処理Aが引用回数を2倍にし、処理Bが引用回数を2倍にしたとする。すると、処理Aと処理Bの両方を採用した場合、引用回数は4倍になるのでしょうか?なお、効果を対数領域で測定すると、効果が加法的であれば、被引用数は乗法的である。表4から、効果は亜加法的であることがわかる。例えば、表2によれば、参考文献と論文を長くした場合の効果はそれぞれ0.710と0.630であり、両処理の効果は1.04 < 0.710 + 0.630 である。しかし、複数の肯定的な処理を組み合わせると、正の効果があり、単一の処理よりも優れている。なお、先行研究[1, 8, 10]で採用された線形モデルでは、このような非線形性を扱うことができない。

7 DISCUSSION

7.1 Other Confounders

第 5 節では、双子論文が 3 つの条件を調整していることを確認した。我々は、双子論文によって調整される条件が非常に多いことを主張する。例えば、双子論文が取り組む研究課題は、同一または類似であると考えられる。また、論文の質も、完全ではないにせよ、ある程度調整されると仮定している。なぜなら、あまりに質の低い論文は引用されにくいからである。重要なのは、論文の質を定量化することが困難であるため、この仮説を数値的に検証することができないことである。多変量解析のような他の方法では対応できないため、このような観測不能/定量化不能な交絡因子を制御できることが双子論文の強みであると主張する。

7.2 Limitations

第一に、双子論文は必ずしもすべての交絡因子を制御しているわけではない。例えば、会場を決めてから、会場の習慣に従ってタイトルにコロンを追加するとすると、会場の選択が交絡因子となる。この場合、補助的な特徴を用いて交絡因子を調整する必要がある。我々の枠組みは一般的であり、多変量解析や層別解析などの他の調整方法と組み合わせることができることを強調する。また、重要なことは、双子論文の強みは、すべてではないにしても、多くの要因を簡単な手順で調整できることである。第二に、厳密に言えば、双子論文は真の反実仮想の結果ではない。現実には、2つの論文が類似したテーマを持ち、互いに引用し合えば、これらの論文の研究インパクトは互いに影響し合う。この限界は、双子に関する元の研究と共通するものである。しかし、双子論文の枠組みは、観察データから無作為に抽出したサンプルを用いた先行研究よりも、バイアスの影響を受けにくい。我々のフレームワークと手動分析、例えば多変量解析や層別解析を組み合わせることで、この問題をさらに軽減することができる。第三の限界は、双子論文が領域によっては稀であることである。データマイニングの領域では、双子が少ないことがわかった。これは、多くのデータマイニングカンファレンスが投稿時にarXivへの投稿を禁止しており、同じトピックに関する同時投稿論文を見つける妨げになっているためと推測される。

8 CONCLUSION

本論文では、研究過程における意思決定の効果を調査するための簡単なフレームワークを提案した。双子論文が同様の条件で出版されていることを実証的に確認し、論文内容や出版会場への影響についていくつかのケーススタディを実施した。

ACKNOWLEDGMENTS

この研究は、日本学術振興会科研費21J22490およびJST CREST科研費JPMJCR21D1の支援を受けて行われた。

Table 3: ATEs for the publication venues. We choose treatments so that ATE is positive. Therefore, intuitively, this table reads that the venue in the first column is better than that in the second column by the amount in the fourth column.

Treatment (a)	Control (b)	$\lvert\mathcal{D}^{a/b}\rvert$	$\widetilde{\text{ATE}}$
Journal of Cognitive Neuroscience	NeuroImage	817	0.539
IEEE Transactions on Information Theory	International Symposium on Information Theory	459	1.93
Neural Computation	IEEE Transactions on Neural Networks	216	0.632
Neural Computation	Neural Networks	199	0.76
Symposium on the Theory of Computing	Foundations of Computer Science	182	0.252
IEEE Transactions on Signal Processing	International Conference on Acoustics, Speech, and Signal Processing	178	2.78
Neural Computation	Neurocomputing	153	2.41
Journal of Economic Theory	Games and Economic Behavior	125	0.668
IEEE ACM Transactions on Networking	International Conference on Computer Communications	110	1.29
Symposium on the Theory of Computing	Symposium on Discrete Algorithms	105	0.693

Table 4: ATEs for the combination of treatments. Intuitively, this table reads that if we make the treatments in the first and second columns, the paper has more impact by the amount in the third column.

Treatment A	Treatment B	$\lvert\mathcal{D}\rvert$	$\widetilde{\text{ATE}}$
Lengthen the reference	Lengthen the paper	41546	1.04
Lengthen the reference	Self citation	6473	1.75
Lengthen the paper	Self citation	5018	1.66

We estimate ATE of publishing in a over publishing in b based on Eq. (3). Table 3 shows the results for the venues with the 10 largest $|\mathcal{D}^{a/b}|$. We can observe that publishing in IEEE Transaction on Signal Processing provokes many citations compared to publishing in International Conference on Acoustics, Speech, and Signal Processing (ICASSP). In addition, publishing in Symposium on the Theory of Computing (STOC) is comparable to publishing in Foundations of Computer Science (FOCS), although STOC is slightly better. By contrast, STOC is clearly better than Symposium on Discrete Algorithms (SODA).

6.1 Are the effects additive?

Suppose treatment A doubles the number of citations and treatment B doubles the number of citations. Then, if we adopt both treatments A and B, will the number of citations quadruple? Note that as we measure the effect in a log domain, if the effect is additive, the number of citations is multiplicative. Table 4 shows that the effect is sub-additive. For instance, the effects of lengthening the reference and paper are 0.710 and 0.630, respectively, according to Table 2, and the effect of both treatments is $1.04 < 0.710 + 0.630$. However, combining several positive treatments does have positive effects and is better than a single treatment. It should be noted that the linear models adopted in previous research [1, 8, 10] cannot handle this kind of nonlinearity.

7 DISCUSSION

7.1 Other Confounders

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/unquantifiable 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 GrantNumber 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, Hasitha 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 Fredrick E Vars. Determinants of citations to articles in elite law reviews. *Journal of Legal Studies*, 29(S1):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 JL 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 23rd ACM International Conference on Information and Knowledge Management, CIKM*, pages 491–498. ACM, 2014.

[8] Matthew E Falagas, Angeliki Zarkali, Drosos E Karageorgopoulos, Vangelis Bardakas, 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 Fowler and Dag Aksnes. Does self-citation pay? *Scientometrics*, 72(3): 427–437, 2007.

[10] Farrokh Habibzadeh and Mahboobeh Yadollahie. Are shorter article titles more attractive for citations? crosssectional study of 22 scientific journals. *Croatian medical journal*, 51(2):165–170, 2010.

[11] James Hartley. Planning that title: Practices and preferences for titles with colons in academic articles. *Library & Information Science Research*, 29(4):553–568, 2007.

[12] Nick Haslam, Lauren Ban, Leah Kaufmann, Stephen Loughnan, Kim Peters, Jennifer Whelan, and Sam Wilson. What makes an article influential? predicting impact in social and personality psychology. *Scientometrics*, 76(1):169–185, 2008.

[13] Thomas S Jacques and Neil J Sebire. The impact of article titles on citation hits: an analysis of general and specialist medical journals. *JRSM short reports*, 1(1): 1–5, 2010.

[14] Hamid R Jamali and Mahsa Nikzad. Article title type and its relation with the number of downloads and citations. *Scientometrics*, 88(2):653–661, 2011.

[15] Matt McGue, Merete Osler, and Kaare Christensen. Causal inference and observational research: The utility of twins. *Perspectives on psychological science*, 5(5): 546–556, 2010.

[16] Natsuo Onodera and Fuyuki Yoshikane. Factors affecting citation rates of research articles. *Journal of the Association for Information Science and Technology*, 66(4): 739–764, 2015.

[17] Carlos Eduardo Paiva, João Paulo da Silveira Nogueira Lima, and Bianca Sakamoto Ribeiro Paiva. Articles with short titles describing the results are cited more often. *Clinics*, 67(5):509–513, 2012.

[18] Ryoma Sato, Makoto Yamada, and Hisashi Kashima. Poincare: Recommending publication venues via treatment effect estimation. *Journal of Informetrics*, 16(2): 101283, 2022.

[19] Ryoma Sato, Makoto Yamada, and Hisashi Kashima. Re-evaluating word mover's distance. In *Proceedings of the 39th International Conference on Machine Learning, ICML*, 2022.

[20] Stefan Stremersch, Isabel Verniers, and Peter C Verhoef. The quest for citations: Drivers of article impact. *Journal of Marketing*, 71(3):171–193, 2007.

[21] Stefan Stremersch, Nuno Camacho, Sofie Vanneste, and Isabel Verniers. Unraveling scientific impact: Citation types in marketing journals. *International Journal of Research in Marketing*, 32(1):64–77, 2015.

[22] Sinisa Subotic and Bhaskar Mukherjee. Short and amusing: The relationship between title characteristics, downloads, and citations in psychology articles. *Journal of information science*, 40(1):115–124, 2014.

[23] Iman Tahamtan and Lutz Bornmann. Core elements in the process of citing publications: Conceptual overview of the literature. *Journal of Informetrics*, 12 (1):203–216, 2018.

[24] Jie Tang, Jing Zhang, Limin Yao, Juanzi Li, Li Zhang, and Zhong Su. Arnetminer: extraction and mining of academic social networks. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD*, pages 990–998. ACM, 2008.

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

[26] Tyler J VanderWeele. Principles of confounder selection. *European journal of epidemiology*, 34(3):211–219, 2019.

[27] Elizabeth S Vieira and José ANF Gomes. Citations to scientific articles: Its distribution and dependence on the article features. *Journal of Informetrics*, 4(1): 1–13, 2010.

[28] 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.

[29] Shuai Xiao, Junchi Yan, Changsheng Li, Bo Jin, Xiangfeng Wang, 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.

[30] Yuxin 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.

[31] Rui Yan, Jie Tang, Xiaobing 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.

[32] 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.

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, Hasitha 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 Fredrick E Vars. Determinants of citations to articles in elite law reviews. *Journal of Legal Studies*, 29(S1):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 JL 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 23rd ACM International Conference on Information and Knowledge Management, CIKM*, pages 491–498. ACM, 2014.

[8] Matthew E Falagas, Angeliki Zarkali, Drosos E Karageorgopoulos, Vangelis Bardakas, 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 Fowler and Dag Aksnes. Does self-citation pay? *Scientometrics*, 72(3): 427–437, 2007.

[10] Farrokh Habibzadeh and Mahboobeh Yadollahie. Are shorter article titles more attractive for citations? crosssectional study of 22 scientific journals. *Croatian medical journal*, 51(2):165–170, 2010.

[11] James Hartley. Planning that title: Practices and preferences for titles with colons in academic articles. *Library & Information Science Research*, 29(4):553–568, 2007.

[12] Nick Haslam, Lauren Ban, Leah Kaufmann, Stephen Loughnan, Kim Peters, Jennifer Whelan, and Sam Wilson. What makes an article influential? predicting impact in social and personality psychology. *Scientometrics*, 76(1):169–185, 2008.

[13] Thomas S Jacques and Neil J Sebire. The impact of article titles on citation hits: an analysis of general and specialist medical journals. *JRSM short reports*, 1(1): 1–5, 2010.

[14] Hamid R Jamali and Mahsa Nikzad. Article title type and its relation with the number of downloads and citations. *Scientometrics*, 88(2):653–661, 2011.

[15] Matt McGue, Merete Osler, and Kaare Christensen. Causal inference and observational research: The utility of twins. *Perspectives on psychological science*, 5(5): 546–556, 2010.

[16] Natsuo Onodera and Fuyuki Yoshikane. Factors affecting citation rates of research articles. *Journal of the Association for Information Science and Technology*, 66(4): 739–764, 2015.

[17] Carlos Eduardo Paiva, João Paulo da Silveira Nogueira Lima, and Bianca Sakamoto Ribeiro Paiva. Articles with short titles describing the results are cited more often. *Clinics*, 67(5):509–513, 2012.

[18] Ryoma Sato, Makoto Yamada, and Hisashi Kashima. Poincare: Recommending publication venues via treatment effect estimation. *Journal of Informetrics*, 16(2): 101283, 2022.

[19] Ryoma Sato, Makoto Yamada, and Hisashi Kashima. Re-evaluating word mover's distance. In *Proceedings of the 39th International Conference on Machine Learning, ICML*, 2022.

[20] Stefan Stremersch, Isabel Verniers, and Peter C Verhoef. The quest for citations: Drivers of article impact. *Journal of Marketing*, 71(3):171–193, 2007.

[21] Stefan Stremersch, Nuno Camacho, Sofie Vanneste, and Isabel Verniers. Unraveling scientific impact: Citation types in marketing journals. *International Journal of Research in Marketing*, 32(1):64–77, 2015.

[22] Sinisa Subotic and Bhaskar Mukherjee. Short and amusing: The relationship between title characteristics, downloads, and citations in psychology articles. *Journal of information science*, 40(1):115–124, 2014.

[23] Iman Tahamtan and Lutz Bornmann. Core elements in the process of citing publications: Conceptual overview of the literature. *Journal of Informetrics*, 12 (1):203–216, 2018.

[24] Jie Tang, Jing Zhang, Limin Yao, Juanzi Li, Li Zhang, and Zhong Su. Arnetminer: extraction and mining of academic social networks. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD*, pages 990–998. ACM, 2008.

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

[26] Tyler J VanderWeele. Principles of confounder selection. *European journal of epidemiology*, 34(3):211–219, 2019.

[27] Elizabeth S Vieira and José ANF Gomes. Citations to scientific articles: Its distribution and dependence on the article features. *Journal of Informetrics*, 4(1): 1–13, 2010.

[28] 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.

[29] Shuai Xiao, Junchi Yan, Changsheng Li, Bo Jin, Xiangfeng Wang, 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.

[30] Yuxin 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.

[31] Rui Yan, Jie Tang, Xiaobing 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.

[32] 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.