RIETI Discussion Paper Series 20-E-058

Incentive or Disincentive for Disclosure of Research Data? A Large-Scale Empirical Analysis and Implications for Open Science Policy

KWON, Seokbeom
University of Tokyo

MOTOHASHI, Kazuyuki
RIETI

The Research Institute of Economy, Trade and Industry
https://www.rieti.go.jp/en/

RIETI Discussion Paper Series 20-E-058
June 2020

Incentive or Disincentive for Disclosure of Research Data? A Large-Scale Empirical Analysis and Implications for Open Science Policy[i]

KWON Seokbeom (University of Tokyo), MOTOHASHI Kazuyuki (RIETI, University of Tokyo)

Abstract

The incentive for scientists to disclose their research data hinges on the extent to which data disclosure brings academic credit (the credit effect) compared to the dissipation of academic credit through intensified scientific competition (the competition effect). In this study, we examine the net effect on the academic credit received by research publications of data-providing researchers publicly disclosing research data. To accomplish this, we compared the citation impact of scientific journal articles that disclosed original data with those that did not.
An analysis of metadata of over 310,000 Web of Science (WoS)-indexed journal articles published in 2010 shows that in the early period after publication, more citations accrued to articles that disclosed original data than to those that did not. However, this difference faded over time and the pattern was later reversed. Additional analysis shows that the credit effect dominates for data-disclosing research published in journals with higher scholarly reputations, whereas the competition effect dominates for research published in journals with lower scholarly reputations. This study contributes to on-going policy discussion concerning the need for institutional measures to promote open science and the disclosure of research data by scientists.

Keywords: Data Sharing; Open Science; Data Disclosure; Digital Economy Policy
JEL codes: I23; O33

[i] This study was conducted as part of the "Digitalization and Innovation Ecosystem: Holistic Approach" project undertaken at the Research Institute of Economy, Trade and Industry (RIETI). The authors would like to thank Professor Nagaoka and RIETI discussion paper seminar participants for their helpful comments.

1. Introduction

Disclosing and sharing research data promotes follow-on research by reducing the cost to subsequent researchers in accessing necessary information. In addition, it helps to prevent duplicating research (Hilgartner and Brandt-Rauf, 1994; Shibayama et al., 2012). Once research data are shared, scientists no longer need to expend effort obtaining the same data for different research. As a result, resources can be more efficiently allocated to novel research projects (e.g., Arrow, 1972). Active disclosure of original research data, therefore, contributes to knowledge diffusion and encouraging future researchers to access existing knowledge for follow-on research (Rosenberg, 1996), all of which is essential for the cumulative progress of science (Furman and Stern, 2011; Mokyr, 2002).

Opening access to data in one research field can foster research in other fields. As a variety of data become available, researchers have more opportunities to explore solutions to new problems by combining the disclosed data (Fecher et al., 2015; Whitlock, 2011). The story of an international research team at the University of California at San Diego, Stanford University, and the Korean Institute of Science and Technology Information, seeking a cancer cure through drug repositioning, is an example. This research team combined four disclosed datasets on gene expression profiles, compound activity measurements, cancer cell line molecular profiles, and cancer patient samples to explore which existing drugs are effective in cancer treatment. This allowed them to demonstrate that pyrvinium pamoate (i.e., anthelmintic) is effective in the treatment of liver cancer (Chen et al., 2017).

Sharing research data facilitates the development of scientific theory by enabling the replication and validation of research (Anderson et al., 2008; Christensen and Miguel, 2018; Eisenberg, 2006; Mueller-Langer et al., 2019) while helping to detect research misconduct (Campbell, 2009; Levelt et al., 2012). With access to disclosed data, researchers increase their chances of finding anomalies in existing scientific theory, which is the essential driver of scientific paradigm shifts (Kuhn, 1962).

Individual researchers may also benefit from disclosing their research data. Data disclosure can enhance the visibility of research publications, thus bringing them greater academic credit. It may also improve the chances for career development and future funding opportunities, as the credit from data disclosure may help to build individual academic reputations (McKiernan et al., 2016).

Due to its significance, there have been a number of institutional efforts to promote and support the disclosure of research data. For example, the US National Institute of Health (NIH) requires that all NIH-

funded research projects with funding of more than $500,000 to include a data-sharing plan.[1] Another example is the establishment of archives for storing biomaterials and related information to facilitate new research (Furman and Stern, 2011). Public money is invested in building the infrastructure for storing and sharing research data. In 2003, the University of Rochester launched a digital archive to preserve and share comprehensive academic data.

However, there is an ongoing concern that researchers continue to not disclose their research data (Cohen and Taubes, 1995; Nelson, 2009; Thursby et al., 2009) because of a lack of explicit incentives for individual researchers to do so in the first place (Mueller-Langer and Andreoli-Versbach, 2018). This is often framed as a social dilemma (Linek et al., 2017a; Scheliga and Friesike, 2014). The academic originality of research publications is often dependent on the originality of the data used by researchers. When disclosing data, researchers potentially lose competitive advantage in their future research while inviting more competition into their research field (Haeussler, 2011; Haeussler et al., 2014; Thursby et al., 2009). As a result, research data disclosure may encourage new research that will quickly replace the scholarship in the data-disclosing research. Intensified scientific competition and the resulting expedited knowledge replacement has the potential to dissipate the academic credit accrued to the original research, thus, disincentivizing researchers from disclosing their data. The disincentive to disclose data could be particularly acute for early-career researchers whose job security is often dependent on academic performance (Andreoli-Versbach and Mueller-Langer, 2014; Haeussler et al., 2014). The professional careers of such researchers could be jeopardized by data disclosure (Marshall, 2002).

Sharing research data may also impose an economic cost on individual researchers. The storing and reorganizing of data for disclosure require researchers to spend resources in the absence of explicit compensation. Indeed, studies emphasize that it is crucial that researchers be compensated for the personal cost of data disclosure and be given proper credit for sharing research data, as a means to incentivize disclosure (Mukherjee and Stern, 2009).

In sum, disclosing research data may allow individual researchers to gain academic credit for their research publications, and doing so is socially desirable. Yet, it also has the potential to dissipate the academic credit for that work. The existence of these two opposite effects indicates that the extent to which individual researchers are willing to disclose their research data hinges on whether the data disclosure brings benefits greater than the cost to the scientist's scholarship.

This study examines the net effect of disclosing research data on the academic credit of those

[1] See https://grants.nih.gov/grants/guide/notice-files/NOT-OD-03-032.html

researchers who disclose their data. We undertook an empirical analysis of metadata of WoS-indexed journal articles published in 2010, in conjunction with the WoS data citation index database (known as *DataCite*). *DataCite* connects WoS-indexed articles with published data sources on the Internet (Mongeon et al., 2017). As of 2019, *DataCite* indexed 7.8 million datasets and 1.1 million data studies in more than 380 data repositories.[2] Considering the citations accrued to a journal paper as a proxy for the academic credit awarded to the underlying research (Merton, 1973), we estimated the effect of data disclosure on academic credit to research publications by comparing the citation count of journal articles that disclosed the original data and those that did not.

A series of multivariate regression analyses show that more citations were accrued to data-disclosing papers than non-data disclosing papers in the early period after publication. However, the pattern was reversed later— on average, papers that disclosed the original data received fewer citations than those that did not as time went by. These findings support the idea that data disclosure enhances the visibility of associated research publications and boosts academic credit (***credit effect***) in the early period. However, data disclosure intensifies scientific competition and eventually expedites the knowledge replacement process (***competition effect***). The credit and competition effects are canceled out in the middle period, and the competition effect prevails later on.

Additional analysis reveals that the interplay between these two effects is shaped by the scholarly reputation of the journal where the data-disclosing research was published. For data-disclosing papers published in highly reputable journals, the credit effect was prominent. In contrast, the competition effect was strong for data-disclosing research published in journals with a lower reputation.

This study complements previous research that examined scientists' behavior and incentive issues in one-on-one scientific information exchange (Blumenthal et al., 1997; Campbell et al., 2000; Hilgartner and Brandt-Rauf, 1994; McKiernan et al., 2016; Shibayama et al., 2012; Tenopir et al., 2011; Walsh and Hong, 2003) by shedding empirical light on the public disclosure of research data. Along with these previous studies, the findings of this research add empirical clues to our understanding of the sustainability of open science practice and whether individual scientists are adequately incentivized to comply with open science policy for research data sharing.

The remainder of this paper is structured as follows: In section 2, we review the literature to obtain theoretical insight into the net effect of data disclosure on the academic credit earned by data-providing research. Two strands of literature are reconciled: 1) the academic benefits and costs of sharing research

[2] See https://clarivate.com/webofsciencegroup/solutions/webofscience-data-citation-index/

data, and 2) the factors affecting researchers' willingness to disclose their data. Section 3 illustrates the empirical research design, and section 4 presents the analysis results. In sections 5, we attempt to disentangle the credit and competition effects. In sections 6, we analyze how the scholarly reputation of journals moderates the interplay between the credit and competition effects. Sections 7 and 8 present the implications of the findings and concluding remarks.

2. Credit vs. Competition

How does the disclosure of research data affect the academic credit accrued to the original research? The existing literature suggests that two opposite effects may be at play. These are the credit effect and the competition effect.

On the one hand, researchers may receive greater academic credit by disclosing their data. McKiernan et al. (2016) argue that the credit effect is realized in five ways. First, by disclosing data, the visibility of associated research publications to subsequent researchers is enhanced. This can boost the citation impact of those publications, thus enhancing the academic reputation of the data disclosing researcher. Second, by disclosing their data, researchers have more opportunities to signal the credibility of their work to other scientists, thus benefiting their career development and providing opportunities to collaborate with other scientists who are now aware of their work. Third, the resulting expanded opportunities for future research may provide more opportunities to receive research funding. The argument for the "benefits" to individual researchers of disclosing their data assumes that the academic community will give credit to the original data disclosure by recognizing the scientific work with which it is associated. Survey research conducted by Scheliga and Friesike (2014) and Fecher et al. (2017) found that this assumption appears to be valid, as most of the scientists surveyed recognized the importance of sharing scientific information and its benefits to the advancement of individual research.

Furman and Stern (2011) provide compelling evidence for the existence of the "credit effect." Their study was based on the establishment of the Biological Resource Center (BRC) tasked with collecting and distributing information regarding existing biological organisms. By using the event of transferring biomaterials to the BRC as the external shock, they estimated the causal effect of the BRC on the citation count accrued to the transferred biomaterials-associated research articles. They found that the enhanced visibility of those publications due to the transfer of biomaterials to BRC increased their citation impact. Likewise, disclosing research data can boost academic credit to the associated original research.

However, disclosing original data may have the opposite effect. Disclosing data has the potential to

draw more competitors into the data-disclosing researcher's field of scholarship, and those competitors may publish their research findings using those data before the data-providing researcher does (Thursby et al., 2009). More importantly, data disclosure may negatively impact the extent to which the data-disclosing research earns academic credit because of the expedited the knowledge replacement process. Once data are disclosed, subsequent researchers could well generate enhanced and superseding findings using the disclosed data to outdate the original data-disclosing research. Meanwhile, follow-on research may use the disclosed data to attempt to invalidate or highlight the critical limitations of the original research, which could negatively affect long-term academic credit that the data-disclosing research might otherwise receive. Existing theoretical and empirical research suggests this may be the case.

Mukherjee and Stern (2009) conducted theoretical research exploring when researchers choose "disclosure" of their research materials rather than "secrecy." Their model explicitly demonstrated an academic credit tradeoff between disclosing data and keeping data secret. Researchers may get more academic credit from data disclosure so long as the subsequent research using the disclosed data gets more credit, while keeping data secret may help researchers to appropriate their research. In other words, researchers choose disclosure only if it is anticipated to bring more academic credit to their original research publications. Otherwise, secrecy is a better strategy. Considering that stronger academic competitive intensity reduces the expected amount of academic credit, scientific competition makes scientists reluctant to share their research data.

Research by Walsh et al. (2007) and Cohen and Walsh (2007) suggests that scientific competition is the crucial determinant of researchers' material-sharing decisions. They report that researchers use "secrecy" more often than formal intellectual property rights (e.g., patent) to maintain exclusive access to their research materials. Using the survey data of 507 biomedical researchers, they showed that one of the main factors driving restricting access to research materials in the field of biomedicine is the stiff scientific competition faced by researchers.

Haeussler et al. (2014) have found that "competition" is the key factor driving researchers' data disclosure decisions. Using a game-theoretic model of researchers' data disclosure decisions, they showed that, as academic competition increases, researchers become more conservative in sharing their data. An empirical test of this hypothesis, using the survey data from 1,173 bio-scientists in German and UK research organizations, supported the model prediction. The survey further revealed that the tendency was moderated by the career stage of the researcher. Researchers who had not received tenure were significantly less likely to share their data, while no such tendency was found for tenured researchers. Anderson et al. (2008) reached similar conclusions following a survey of 488 researchers in

the fields of management and economics. One of their findings indicated that the willingness to share research data increased for researchers who had received tenure.

The probable existence of credit and competition effects of data disclosure suggests that the prevalence of one over the other is dependent on the extent to which data disclosure accrues net academic credit to the researcher. When the credit effect is greater than the competition effect, researchers can expect boosted academic credit to their research publications and are incentivized to disclose their data. In contrast, when the competition effect prevails, researchers will be unwilling to disclose their data.

To sum up this literature review, the credit effect originates from the potential for subsequent opportunities for the data-disclosing researcher, while the competition effect stems from the subsequent diminishment of academic credit to the data-disclosing research due to the process of knowledge-replacement. This implies a difference in the timing of the two effects: the credit effect arises first, as data disclosure enables subsequent research, while the competition effect emerges later, as resulting research may overtake the data-disclosing research and diminish the credit accruing to subsequent research conducted by the date-disclosing researcher. Therefore, we expect that:

Data disclosing research gains more academic credit than non-data disclosing research in the early period after publication. However, this difference effect gradually disappears and is reversed as the competition effect emerges.

3. Method

3.1. Overview of Research Design

We consider a journal article as the body of novel scientific knowledge and the number of citations accrued to the article as a proxy for academic credit given to the research (Merton, 1973). According to this notion, our analysis focused on comparing the citation counts of journal articles where original data were disclosed and those where original data were not disclosed.

If data disclosure brought more academic credit to the associated research work, a journal article from which the original data had been disclosed would receive more citations than its non-disclosed counterpart. To examine associated time dynamics, an analysis was conducted of the annual citation count difference for the nine years[3] after a journal article was published.

[3] The annual citation count for 2019 was incomplete at the time of data retrieval.

3.2. Data

Our data consisted of Clarivate's WoS-indexed journal articles that disclosed their original data and articles published in the same journals in the same year in the absence of data disclosure.

We began by retrieving information from the WoS core collection database on journal articles published in 2010 where original data had been disclosed. To this end, we first searched for journal articles that had "associated data". Although WoS records articles with disclosed original data as articles with "associated data", this label is also applied to articles that simply cite existing data. To identify which articles disclosed the original data, we obtained detailed information on the associated data from *DataCite* and checked whether there was an overlap between the authors of the associated data and the authors of the journal article. If the associated data and the article in question were published by the same author(s), we considered that the article is one disclosed the original research data. Using this author-matching algorithm, we identified 15,271 journal articles disclosing the original data. An example of such a journal article is Gu and Rice's (2010) "Three conformational snapshots of the hepatitis C virus NS3 helicase reveal a ratchet translocation mechanism", published in the Proceedings of the National Academy of Science. The study reveals the x-ray crystal structure for a set of NS3h complexes which are known as one of the causes of liver disease. The authors published their data in the World Protein Bank and the data were indexed in the WoS.[4] A further example is Riffle and Davis's (2010) "The Yeast Resource Center Public Image Repository: A large database of fluorescence microscopy images", published in BMC Bioinformatics. This article provides detailed information about a data repository that stores microscopy images of yeast and that was established by one of the authors in 2010.

For each article disclosing original data, we retrieved the metadata for articles published in the same year and in the same journal as the focal article, but for which original data were not disclosed. This group of articles became the comparison group in the analysis. Using a journal-based search, we obtained 295,634 articles in the comparison group. Some articles had missing metadata. These were excluded from the sample. In total, our dataset contained the detailed metadata of 310,900 articles published in 1,240 WoS-indexed journals in 2010.

3.3. Variables

Dependent Variable. The dependent variable was the annual citation count accrued to each article from 2010 to 2018 (from ***FWD2010*** to ***FWD2018***). The annual citation count had a non-negative value without

[4] See https://www.wwpdb.org/pdb?id=pdb_00003kqu

an upper limit.

Independent Variable. The independent variable was a binary variable that had a value of 1 for articles disclosing original data and a value of 0 for articles that did not disclose data (***Original Data Disclosed***).

Control Variables. We introduced several control variables that may be associated with both the dependent and independent variables. First, we controlled for the journal fixed effect (***Journal FE***). Average article citation counts have a high-level heterogeneity depending on the journal, while some journals explicitly encourage authors to disclose their data. To take into account probable journal-level heterogeneity, we introduced a set of dummy variables corresponding to each journal that appeared in the data.

Second, journal articles about popular topics, such as artificial intelligence, machine learning, etc., draw more academic attention than other, less popular, topics, resulting in a higher accumulation of annual citation counts. At the same time, those articles might be less likely to disclose data due to highly intensive competition in those fields (Cohen and Walsh, 2007; Haeussler et al., 2014; Walsh et al., 2007). To even out this probable compounding effect, we used keyword information to control for the popularity of the research topic in each journal article (***Popularity***). In WoS, two types of keywords are assigned to journal articles: author's keywords and 'keywords plus', which is curated by Clarivate. Both sets of keywords information were used. Topic popularity was measured based on how many articles in the dataset had the same keywords as the focal article. More specifically, we calculated popularity as follows:

- We constructed the list of keywords appearing in the author's keywords and keyword plus fields for all articles in the dataset.
- We used lemmatization and the elimination of stop words/punctuation to standardize the keywords.
- We calculated the frequency of each keyword in the list. This frequency list became U.
- For each article, we listed the frequency of each keyword by referring to U. The list of keywords and their frequency is F.

The popularity measure of article i was calculated as the natural log of the median value of F. The greater the value of this variable, the higher its topical popularity in the sample.

Third, we controlled for whether articles originated from funded or non-funded research projects (***Funding***). Increasingly, the funding bodies of some countries, such as the NSF and NIH in the US, have introduced policy initiatives to promote research data disclosure. Therefore, the likelihood of a journal

article having disclosed data could positively correlate with whether the associated research project was funded. In addition, it is widely known that articles arising from funded research receive more citations than articles from unfunded research. To take into account the funding-associated compounding effect, we introduced a binary variable with a value of 1 if the article in question acknowledged research funding support, and 0 otherwise.

Fourth, we controlled for whether a research article was published by international collaboration. Wagner et al. (2019) argue that international collaborative research suffers from greater transaction costs of exchanging research ideas and information among collaborators. Such transaction costs may suppress the exploration of novel ideas and, thus, may be negatively associated with the academic credit earned by the resulting research publications. In addition, the involvement of researchers in different countries may result in delays and additional costs in deciding on data disclosure. To take this into account, we introduced a binary variable with a value of 1 if more than one country appeared in the author's country information (***Int Collabo***).

Fifth, to account for the extent to which the research article of interest built upon pre-existing research, we controlled for the number of cited references and took the natural log-transformation, adding 1 to the number of cited references to consider its skewed distribution (***ln(nRef+1)***).

Last, studies have repeatedly found that the novelty of the research idea is likely to be associated with the citation impact and its timing of recognition by subsequent research (e.g., Wang et al., 2017). Meanwhile, willingness to disclose research data may be associated with its novelty, as novelty in research may be associated with the intensity of scientific competition in the relevant research field. To control for this compounding effect, we introduced the operationalization of ***Novelty*** employed by Lee et al. (2015) by adapting the method formulated by Uzzi et al. (2013). Based on the presumption that the cited references are proxy for the prior knowledge that enabled the focal research and the atypical knowledge combination is the source of the scientific novelty, this variable quantifies how atypical it is for the cited journals in a publication to jointly appear in the corpus of associated publications.

For regression analysis, we fit our data to the generalized negative binomial (GNB) model because the dependent variable was a count variable that had left-skewed distribution (i.e., overdispersion problem).

If the credit effect was prevalent, ***original data disclosed*** had a statistically significant positive value. In contrast, if the competition effect was dominant, the independent variable was anticipated to have a statistically significant negative value.

4. Results

4.1. Descriptive Analysis

Table 1 presents the summary statistics of the key variables and their pairwise correlations. All the absolute values of correlations are below 0.3, indicating no critical multicollinearity. Note that 305,245 of the 310,900 articles had valid popularity values. This was due to the lack of keywords in 5,655 articles (1.8%). With regard to the *Novelty* measure, 310,150 articles had valid values. The other 750 had no cited journal information (0.75%).

[Insert Table 1 about here]

Table 2 compares the summary statistics of the key variables. From 2010 to 2018, the mean value of the annual citation count for the data-disclosing articles was consistently greater than that of the comparison group. Virtually no differences in ***Popularity***, ***ln(nRef+1)***, and ***Int Collabo*** were found. However, a greater share of data-disclosing articles acknowledged funding and contained novel research than those that did not disclose data.

[Insert Table 2 about here]

Figure 1 displays the distribution of WoS subject categories (WoS SCs) of the data-disclosing articles. For visualization purposes, only the top 20 WoS SCs are presented. Multidisciplinary science took the largest share, followed by Life Science related fields.

[Insert Figure 1 about here]

4.2. Regression Analysis

4.2.1. Main Analysis

Table 3 presents the GNB regression results. The coefficient of ***Original Data Disclosed*** in the first column was positive but statistically insignificant at the 0.1 significance level. From 2011 to 2013, the coefficients were positive and statistically significant at the 0.05 significance level. During those three years, journal articles that disclosed data received more citations than counterparts that did not disclose data. Note that the size of the coefficients decreased over time. This indicates that the difference between the two groups of articles gradually shrank. Note also that the coefficients of the independent variable from the fifth to the last columns had both negative values and size increases. In the last column, the coefficient of ***Original Data Disclosed*** turned negative and statistically significant at the 0.1 significance level.

[Insert Table 3 about here]

The regression results provide evidence to support our expectation that, in the early period, a journal article disclosing original data receives more academic credit than an article that does not disclose original data, but that this pattern gradually disappears and is reversed over time.

4.2.2. Selection Bias

One may raise a concern that our findings could be driven by multiple selection bias. For example, data-oriented research becomes the subject of data disclosure while theory-driven studies have no data to disclose. If these findings originated from such an inherent difference between data-oriented and theoretical research, the analysis results could not be interpreted as the effect of data disclosure. Besides, it is also feasible that a journal article presenting a highly impactful scientific discovery is more likely to disclose associated research data and, therefore, would receive more citations than its counterpart in the early period. Although these self-selection bias issues do not explain another finding, that data disclosing articles received fewer citations than their counterparts later, selection bias remains a challenge to the validity of the findings.

To test if selection bias was the primary driver of these findings, we conducted two additional analyses. First, we analyzed the articles that disclosed research data only. In this analysis, we exploited variations in the timing of data disclosure. Among the data disclosing articles, some disclosed the original data in the same year as the article was published (early data-disclosed articles) while others released the data later (late data-disclosed articles). In this setting, the citation count that the late data-disclosed articles received become the proxy for the counterfactual of the early data-disclosed papers for the period between the data disclosure year of early and late data disclosing papers.

Of the 15,271 data-disclosed articles, 56% were early data-disclosed articles. We then constructed the article-year panel data using the annual citation count from 2010 to 2018 as the dependent variable. We ran the conditional fixed effect (at the article level) negative binomial regression. Table 4 presents the panel regression results.

[Insert Table 4 about here]

The first column shows the regression result using data-disclosed articles in 2010 and 2011. The coefficient of ***Original Data Disclosed*** was positive and statistically significant at the 0.01 significance level. The second column reports the regression result using data-disclosed articles in 2010 and 2012.

The coefficients of the independent variable remained positive and statistically significant at the 0.05 significance level, but the size decreased. We ran the same regression again, replacing the data with (2010 vs. 2013) to (2010 vs. 2018). The coefficients of the independent variable remained positive and statistically significant up to 2010 vs. 2014 column. From the 2010 vs. 2015 column, the statistical significance of the coefficient disappeared, and the signs turned to negative later. This finding is consistent with our cross-sectional analysis which found that a data-disclosing article received more citations for the early period, but this pattern disappeared over time, with the data-disclosing article receiving fewer citations later.

The second analysis exploited the fact that scientifically impactful research projects are often funded by research grants (King, 1987; Shapira and Wang, 2010).

Research funding decisions are based on the peer review of research proposals by disciplinary experts. This serves as an institutional device to select scientifically impactful and feasible research projects. By using the information of whether a paper acknowledged research grants as a rough indicator (***Funding***) of the quality of the research outcome, we examined whether the association between the independent and dependent variables was moderated by ***Funding***. For the operationalization of this test, we generated an interaction term between ***Funding*** and ***Disclosed Original Data***; ***FundingxOriginal Data***. If the selection bias in question was a crucial factor behind the main regression result, we expected a negative correlation of the interaction term. Table 5 shows that the coefficients of the interaction term were statistically insignificant at the 0.1 significance level. There was no evidence to support the argument that self-selection bias was the major driver of our findings.

[Insert Table 5 about here]

4.2.3. Matthew Effect

Our previous analysis showed that in the early period, data-disclosing articles received more citations than articles that did not disclose data. Yet, this finding could have originated from the Matthew effect, whereby scholars with higher reputations are more likely to disclose their data because they are likely to have established careers (i.e., tenured) and are, therefore, less sensitive to the consequences of data disclosure. Meanwhile, articles authored by leading scholars are likely to be cited more, simply because of the scholar's reputation (Merton, 1968). Although the Matthew effect does not explain the finding that in the later period— data-disclosing articles received fewer citations than articles that did not have disclosed data, the effect may still threaten the validity of our findings regarding the presence of the credit effect. To check whether the Matthew effect was the primary driver of our first

finding, we conducted an additional analysis by matching data-disclosed articles with non-data disclosed articles on authors and journals.

For each article in the data-disclosing group, we matched one article that did not disclose the original data but was published by at least two of the same authors[5] in the same journal and the same year as the data-disclosing article in question. This matching process substantially reduced the sample size because only a handful of authors published two or more articles in the same journal in the same year. Table 6 reports the generalized negative binomial regression results.

[Insert Table 6 about here]

The regression analysis shows that a data disclosing article received more citations than its matched article and this difference was statistically significant at the 0.1 significance level in 2011. From 2012, the size of this citation difference decreased and was reversed and enlarged from 2016. Although the statisti cal difference was insignificant after 2012, this should not discount the meanings of the overall pattern b ecause the best estimator of the citation count difference followed the same pattern as that of the main analysis. Note that the sample in this analysis was selective for articles authored by those who published at least two articles in the same journal in 2010. Accordingly, our matching procedure might have subst antially removed the statistical power in estimating the difference in the citation rate between matched articles.

5. Disentangling the Credit and Competition Effects

In the previous analysis, we showed that the credit and competition effects exist but take effect at different times. How can we decouple the credit from the competition effect? We borrowed the citation function model (Caballero and Jaffe, 1993; Jaffe and Trajtenberg, 1996, 2002) to address this question.

The citation function model explains the probability of a scientific work receiving a citation from subsequent research into two exponential processes: the knowledge diffusion process (i.e., the development of follow-on research enabled by the published information in the focal research) and the effect by which the knowledge is replaced by follow-on knowledge. We used Arora et al.'s (2013) notations to describe the citation function model. The probability that a journal article s was cited by another article S ($P(s,S)$) was expressed as:

$$P(s,S) = \alpha(s,S) \times \exp(-\beta_1 \Delta T) \times (1 - \exp\ (-\beta_2 \Delta T))$$

[5] Accordingly, we selected articles authored by two or more authors from the beginning.

Where $\alpha(s,S)$ is the function of the attributes of articles *s* and *S*, ΔT is the publication time difference between articles *s* and *S*, $\exp(-\beta_1\Delta T)$ is the knowledge replacement effect and $(1-\exp\ (-\beta_2\Delta T))$ corresponds to the knowledge diffusion effect. Then, the expected citations that article s receives (E_s) was represented as:

$$E_s = V \times \exp(-\beta_1\Delta T) \times (1-\exp\ (-\beta_2\Delta T))$$

Where $V=\sum_i \alpha(a,i)$ and *i* are the index of all articles that potentially cite the article at ΔT. The expected citation count was calculated by combining the number of articles that potentially cited *s*, their attributes shared with those of article *s*, and the two parameters in the exponential processes (β_1,β_2) with the time difference between article *s* and the citing article (ΔT).

For simplicity, we assumed that *V* was constant over time.[6] By definition, the credit effect by data disclosure expedites the knowledge diffusion process, which was modeled as an increased level of β_2. Likewise, the competition effect was modeled into an increased level of β_1.

We estimated the two effects by fitting our data to the citation function model. Because the estimated coefficient in regression corresponded to the difference in the annual citation count between a journal article disclosing data and a comparable article that did not disclose data, we fit the empirical data to the following functional form using the non-linear function least square method.

$$\Delta E_s = V \times [\exp(-(\beta_1+\delta_1)\Delta T) \times \left(1-\exp\ (-(\beta_2+\delta_2)\Delta T)\right) - \exp(-\beta_1\Delta T) \times \left(1-\exp\ (-\beta_2\Delta T)\right)]$$

Where δ_1 represents the competition effect and δ_2 represents the credit effect.

[Insert Figure 2 about here]

Figure 2 presents the results. The red dashed line draws the estimated credit effect (i.e., forcing δ_1=0), while the blue dashed line represents the competition effect (i.e., forcing δ_2=0). The black solid line is a fitted value, assuming that $\delta_1>0$ and $\delta_2>0$ using the non-linear least square method. The gray dashed line is the actual empirical observation obtained from the main regression.

Our estimation using the citation function model vividly shows that not only the presence of the competition and credit effects but also the difference in the timing of the emergence of the two effects— credit effects come into place first and the competition effect emerges later.[7]

[6] Relaxation of this assumption does not change the shape of the citation function model.

[7] As a complementary analysis, we conducted simulations of citation pattern in the four cases— neither effect exists, only the

6. Scholarly Reputation of Venue for Publication as the Moderating Factor

Because the academic incentive for data disclosure is dependent on which effect is dominant, exploring the moderating factors could help to extend our understanding of how the academic incentive for data disclosure is configured.

We argue that the scholarly reputation of the journal where the data-disclosing research is published is one moderating factor.

The academic reputation of the venue where research is published is likely to be positively associated with the size of the credit effect. The higher the academic reputation of the place where the research is published, the greater the perceived scholarly credibility of the research. When the data from research with high scholarly credibility is disclosed, subsequent research may be accelerated and, as a consequence, the credit effect of the data disclosing research boosts.

Meanwhile, the greater the scientific credibility of the research, the greater the difficulty and time required to generate new scientific knowledge to replace the findings in the focal research. Therefore, the competition effect of data disclosure is likely to be weakened when data-disclosing research is published in a place with a high scholarly reputation.

Given that scholarly journals are the venues for publishing original research works, we expected that the relative prominence of the credit effect over the competition effect would be positively associated with the scholarly reputation of the journal where the data-disclosing research was published.

For an empirical test of this expectation, we operationalized the scholarly reputation of the publishing journal with a discipline-normalized Journal Impact Factor (JIF) for 2010.[8] Some limitations to the use of JIF for operationalization should be noted. JIF can be manipulated by journal stakeholders, such as editors, reviewers, and publishers, it is dependent on how long the journal has been in business, and it is imprecise in measuring the scientific impact of an individual's research (Greenwood, 2007) or the novelty of that research (Wang et al., 2017).

Nevertheless, it is undeniable that the JIF is a widely accepted and useful bibliometric indicator of the scholarly reputation of academic journals that serve as venues for the publication of original scientific research. Given that individual scientists, research funders, and research institutes consider papers published in highly reputable journals to be high-quality research (Guimera et al., 2005; Smith,

---

credit effect exists, only the competition effect exists, and both effects exist. The simulation result is provided in the Appendix.

[8] JIF information was obtained from the Clarivate's Citation Report (https://jcr.clarivate.com/JCRLandingPageAction.action)

2006), JIF can still be a useful proxy for assessing the scholarly reputation of the place where the research has been published (Garfield, 1972).

We divided the articles in the sample into five scientific disciplines using WoS subject categories (WoS SCs). WoS assigns one or more of almost 230 subject categories to journal articles (c.f., assign the WoS SCs for journals). Referring to the work of Leydesdorff et al. (2013) and Rafols et al. (2010), we aggregated the WoS SCs into the five categories[9] of Biology, Medicine, Engineering & Mathematics, Physics & Chemistry, and Social Sciences & Psychology. We generated a binary variable with a value of 1 for articles published in journals with a JIF higher than the median JIF within the field (***JIF above Median***). Then, we generated an interaction term between ***JIF above Median*** and ***original data disclosed*** (***JIF above Median x Data Disclosed***)**.** Note that the size of the sample was reduced to 304,906 because some journals in the sample did not have valid JIF in 2010. Table 7 reports the regression results.

[Insert Table 7 about here]

The coefficients of ***Original Data Disclosed*** were positive and statistically significant at the 0.05 significance level in 2010. Yet, the coefficient signs turned to negative and statistically significant from 2012. This finding indicates that for articles published in journals with a JIF lower than the median value within the same discipline, the dominance of the credit effect was short-lived and the competition effect became prevalent earlier than observed in the main regression analysis.

Meanwhile, the coefficient of ***JIF above Median x Data Disclosed*** was negative and statistically significant at the 0.01 significance level in 2010. This implies that in the early period, the net credit effect for an article published in journals with a higher JIF than the median value of within-field JIFs was smaller than that of articles published in journals with below-median JIF. Yet, this pattern was reversed from 2011. The marginal effect of data disclosure was substantially greater for papers published in highly-reputable journals than for those published in less-reputable journals. Note that the sign and size of the coefficients of the interaction term persisted positive and increased until 2018. This finding implies that the relative prominence of the credit effect over the competition effect is positively associated with the scholarly reputation of the journal where the data-disclosing research was published.

In sum, our additional analyses support the expectation that the scholarly reputation of the journal where the research is published moderates the interplay between the credit and competition effects of

[9] They clustered the WoS SCs into five groups based on the co-citation pattern at journal level. The details and relevant software are provided in http://leydesdorff.net/wc15/index.htm. We combine Biology and Medicine into Life Science.

data disclosure. These findings suggest that the academic incentive to disclose data is positively moderated by the academic reputation of the place where the associated research work is published.

7. Discussion

In this study, we examined the net effect of research data disclosure on the academic credit earned by the associated research. The reconciliation of previous studies suggested that data disclosure enhances the visibility of the associated research (credit effect) while inviting greater competition into the associated research field (competition effect), thus accelerating the generation of new knowledge that replaces and outdates the findings of the data-disclosing research. We anticipated that research that discloses original data receives more academic credit than its counterpart soon after article publication, but this pattern is reversed by the emergence of the competition effect later. We tested this hypothesis by analyzing WoS-indexed journal articles published in 2010.

Our regression analysis found evidence to support this. Controlling for article-level characteristics, it was found that articles disclosing original data earned more citations than their counterparts for the first three years after publication. However, this pattern disappeared in the middle period, and, by eight years after publication, the pattern had reversed. We decomposed the two effects and elucidated the differential timing of their emergence by fitting our data into the citation function model.

Follow-on analyses revealed that the scholarly reputation of the journal in which data-disclosing research is published moderates the interplay between the credit and competition effects. The credit effect is more prominent for research published in journals with a higher scholarly reputation. In contrast, the competition effect is stronger for research published in journals with lower scholarly reputations. Our finding implies that the perceived scholarly credibility of the data-disclosing research may boost the credit effect while weakening the competition effect by slowing down the knowledge replacement process.

Given that the academic incentive for data disclosure is defined by the net consequences of the credi t and competition effects, our findings suggest that the incentive for sharing research data is associated with at least two factors: time and the perceived scholarly credibility of the research.

First, the academic incentive for research data disclosure may differ according to the time horizon. Because the credit effect emerges earlier than the competition effect, data disclosure may help researchers to gain academic credit for their work early on after its publication. However, the competition effect emerges later, and gradually comes to predominate. Therefore, in the long term, researchers face a net disincentive to disclose their data. This implies that, from a long-term perspective,

research data may not be shared publicly as much as is socially desirable.

Second, with all other factors remaining constant, the extent to which scientists are willing to disclose their data may partly depend on their expectations of the scholarly reputation of the journal where their work will be published. Our findings suggest that scientists who expect to publish in prestigious scholarly journals are relatively active in disclosing their data because of the dominance of the credit effect. In contrast, there may be fewer incentives to disclose data when publishing research in a journal with a relatively lower scholarly reputation. This suggests that more nuanced policies must be developed to effectively promote research data disclosure, as scientists will be less willing to voluntarily disclose their original data if the data-disclosing research is likely to be published in journals with lower academic reputations.

What policy options should be considered to mitigate the disincentive to disclose data and what undesirable effects should be taken into account in policy design? Although there could be various policy measures, we discuss two options that are opposite ends to provide an outlook of the ranges of the policies that may need to be in consideration.

The first is to institutionalize the legal protection of research data ownership. This would allow scientists to control access to their data post-disclosure. This gives researchers the option of controlling the competition effect ex-post. One way to implement this is to use the licensing scheme, whereby data-providing researchers disclose data while retaining control of the terms of its use (Eisenberg, 2006). However, this option may harm the concept of open science that has been essential to scientific advance. The propertiziation of research data could result in complicated legal and social welfare issues, such as the tragedy of the anticommons caused by the increased transaction costs incurred in the negotiation of data use (Eisenberg, 2006), as in patent licensing (Heller and Eisenberg, 1998), which arguably impedes scientific progress.

The second is to mandate that all scientists disclose their research data. Enforcing scientists to disclose their data if they are in receipt of public research grants could be one way to implement this policy measure. Recently, in the US, it has become a requirement of NIH and NSF funding that recipients submit their plans for data disclosure/sharing. Such actions can be understood as an early step. However, it should be noted that the extent to which this policy measure is effective in promoting data disclosure is questionable given that it may bring unintended consequences, such as driving researchers to put sub-optimal effort into building and managing their research data from the start (Mueller-Langer and Andreoli-Versbach, 2018).

8. Conclusion

This study contributes to advancing the discussion around the incentive for scientists to disclose their data. So far, scholars and policymakers have focused on the theoretical aspects of the costs and benefits of data disclosure to the individual researcher. Despite its importance, the question of whether researchers are incentivized to voluntarily release their research data to the public has insufficiently been addressed. Our study provides the first large-scale empirical study that seeks to advance this ongoing discussion. We hope our findings are useful for evidence-based policymaking toward promoting research data-sharing among scientists.

By shedding empirical and theoretical light on the public disclosure of research data, our study complements the study of one-on-one information exchange between scientists. In addition to individual scientist's strategic behavior in the one-on-one exchange of information for science (Hilgartner, 1996; Kim and Adler, 2015; Linek et al., 2017b; Murray, 2010; Walsh and Hong, 2003), our study reveals the factors and dynamics at play when scientists decide whether or not to publicly disclose their research data. In particular, our study elaborated on the incentives to disclose data by empirically disentangling the competition and credit effects of data disclosure as well as exploring a moderating factor that has not previously been examined.

Policymakers can benefit from our findings. Thus far, the policy discussion concerning fostering of research data disclosure has focused on establishing the infrastructure required for data storage and sharing. Examples of such efforts include building a library for biomaterials/data storage and creating a large-scale digital archive for researchers to store/share data. Although these efforts may facilitate the disclosure of data by individual researchers, they are of limited effectiveness in addressing the inherent disincentive for data disclosure, as shown in the present study. Nelson's (2009) account of an empty archive shows that building the infrastructure cannot alone address the issue. Based on our findings, policymakers can extend their effort to devise institutional measures to offset the disincentives to disclose data by taking into account the presence and differential timing of two opposing effects of data disclosure along with the factor moderating these two effects.

This study has some limitations that we hope to address in future research. First, the unit of analysis in this study was the journal article. However, much of the discussion concerns whether the future careers of "researchers" benefit or are hampered by the disclosure of their data. Our study does not provide a direct answer to this question. Future research can examine whether a difference exists in the "future research performance" of data-disclosing researchers and researchers who do not disclose their data. Such an examination would include the odds of getting tenure or individual publication

performance.

Second, it is plausible that researchers make strategic decisions regarding data disclosure by taking into account credit and competition effects. There have been numerous studies elucidating the determinants of the data disclosure decision (Fecher et al., 2017; Haeussler et al., 2014; Kim and Adler, 2015; Linek et al., 2017a; Zenk-Möltgen et al., 2018). With insight from those studies, together with the conclusions of this study, we believe that exploring more moderating factors for the credit and competition effects at the researcher level will provide deeper insight into the dynamics of the data disclosure.

REFERENCE

1. Anderson, R.G., Greene, W.H., McCullough, B.D., Vinod, H.D., 2008. The role of data/code archives in the future of economic research. Journal of Economic Methodology 15, 99-119.
2. Andreoli-Versbach, P., Mueller-Langer, F., 2014. Open access to data: An ideal professed but not practised. Research Policy 43, 1621-1633.
3. Arora, A., Branstetter, L.G., Drev, M., 2013. Going soft: How the rise of software-based innovation led to the decline of Japan's IT industry and the resurgence of Silicon Valley. Review of Economics and Statistics 95, 757-775.
4. Arrow, K.J., 1972. Economic welfare and the allocation of resources for invention, Readings in industrial economics. Springer, pp. 219-236.
5. Blumenthal, D., Campbell, E.G., Anderson, M.S., Causino, N., Louis, K.S., 1997. Withholding research results in academic life science: evidence from a national survey of faculty. Jama 277, 1224-1228.
6. Caballero, R.J., Jaffe, A.B., 1993. How high are the giants' shoulders: An empirical assessment of knowledge spillovers and creative destruction in a model of economic growth. NBER macroeconomics annual 8, 15-74.
7. Campbell, E.G., Weissman, J.S., Causino, N., Blumenthal, D., 2000. Data withholding in academic medicine: characteristics of faculty denied access to research results and biomaterials. Research Policy 29, 303-312.
8. Campbell, P., 2009. Data's shameful neglect. Nature 461, 145.
9. Chen, B., Ma, L., Paik, H., Sirota, M., Wei, W., Chua, M.-S., So, S., Butte, A.J., 2017. Reversal of cancer gene expression correlates with drug efficacy and reveals therapeutic targets. Nature communications 8, 16022.
10. Christensen, G., Miguel, E., 2018. Transparency, reproducibility, and the credibility of economics research. Journal of Economic Literature 56, 920-980.
11. Cohen, J., Taubes, G., 1995. Share and share alike isn't always the rule in science. Science 268, 1715.
12. Cohen, W.M., Walsh, J.P., 2007. Real impediments to academic biomedical research. Innovation policy and the economy 8, 1-30.
13. Eisenberg, R.S., 2006. Patents and data-sharing in public science. Industrial and Corporate Change 15, 1013-1031.
14. Fecher, B., Friesike, S., Hebing, M., 2015. What drives academic data sharing? PloS one 10.
15. Fecher, B., Friesike, S., Hebing, M., Linek, S., 2017. A reputation economy: how individual reward considerations trump systemic arguments for open access to data. Palgrave Communications 3, 17051.
16. Furman, J.L., Stern, S., 2011. Climbing atop the shoulders of giants: The impact of institutions on cumulative research. American Economic Review 101, 1933-1963.
17. Garfield, E., 1972. Citation analysis as a tool in journal evaluation. Science 178, 471-479.
18. Greenwood, D.C., 2007. Reliability of journal impact factor rankings. BMC medical research methodology 7, 48.
19. Gu, M., Rice, C.M., 2010. Three conformational snapshots of the hepatitis C virus NS3 helicase reveal a ratchet translocation mechanism. Proceedings of the National Academy of Sciences 107, 521-528.
20. Guimera, R., Uzzi, B., Spiro, J., Amaral, L.A.N., 2005. Team assembly mechanisms determine collaboration network structure and team performance. Science 308, 697-702.
21. Haeussler, C., 2011. Information-sharing in academia and the industry: A comparative study. Research Policy 40, 105-122.
22. Haeussler, C., Jiang, L., Thursby, J., Thursby, M., 2014. Specific and general information sharing among competing academic researchers. Research Policy 43, 465-475.
23. Heller, M.A., Eisenberg, R.S., 1998. Can patents deter innovation? The anticommons in biomedical research. Science 280, 698-701.

24. Hilgartner, S., 1996. Access to data and intellectual property: scientific exchange in genome research. Intellectual property rights and research tools in molecular biology, 28-39.
25. Hilgartner, S., Brandt-Rauf, S.I., 1994. Data access, ownership, and control: Toward empirical studies of access practices. Knowledge 15, 355-372.
26. Jaffe, A.B., Trajtenberg, M., 1996. Flows of knowledge from universities and federal laboratories: Modeling the flow of patent citations over time and across institutional and geographic boundaries. Proceedings of the National Academy of Sciences 93, 12671-12677.
27. Jaffe, A.B., Trajtenberg, M., 2002. Patents, citations, and innovations: A window on the knowledge economy. MIT press.
28. Kim, Y., Adler, M., 2015. Social scientists' data sharing behaviors: Investigating the roles of individual motivations, institutional pressures, and data repositories. International Journal of Information Management 35, 408-418.
29. King, J., 1987. A review of bibliometric and other science indicators and their role in research evaluation. Journal of information science 13, 261-276.
30. Kuhn, T.S., 1962. The structure of scientific revolutions. Chicago and London.
31. Lee, Y.-N., Walsh, J.P., Wang, J., 2015. Creativity in scientific teams: Unpacking novelty and impact. Research Policy 44, 684-697.
32. Levelt, W.J., Drenth, P., Noort, E., 2012. Flawed science: The fraudulent research practices of social psychologist Diederik Stapel.
33. Leydesdorff, L., Carley, S., Rafols, I., 2013. Global maps of science based on the new Web-of-Science categories. Scientometrics 94, 589-593.
34. Linek, S.B., Fecher, B., Friesike, S., Hebing, M., 2017a. Data sharing as social dilemma: Influence of the researcher's personality. PloS one 12, e0183216.
35. Linek, S.B., Fecher, B., Friesike, S., Hebing, M., 2017b. Data sharing as social dilemma: Influence of the researcher's personality. PloS one 12.
36. Marshall, E., 2002. DNA sequencer protests being scooped with his own data. Science 295, 1206-1207.
37. McKiernan, E.C., Bourne, P.E., Brown, C.T., Buck, S., Kenall, A., Lin, J., McDougall, D., Nosek, B.A., Ram, K., Soderberg, C.K., 2016. Point of view: How open science helps researchers succeed. Elife 5, e16800.
38. Merton, R.K., 1968. The Matthew effect in science: The reward and communication systems of science are considered. Science 159, 56-63.
39. Merton, R.K., 1973. The sociology of science: Theoretical and empirical investigations. University of Chicago press.
40. Mokyr, J., 2002. The gifts of Athena: Historical origins of the knowledge economy. Princeton University Press.
41. Mongeon, P., Robinson-Garcia, N., Jeng, W., Costas, R., 2017. Incorporating data sharing to the reward system of science: Linking DataCite records to authors in the Web of Science. Aslib Journal of Information Management 69, 545-556.
42. Mueller-Langer, F., Andreoli-Versbach, P., 2018. Open access to research data: Strategic delay and the ambiguous welfare effects of mandatory data disclosure. Information Economics and Policy 42, 20-34.
43. Mueller-Langer, F., Fecher, B., Harhoff, D., Wagner, G.G., 2019. Replication studies in economics—How many and which papers are chosen for replication, and why? Research Policy 48, 62-83.
44. Mukherjee, A., Stern, S., 2009. Disclosure or secrecy? The dynamics of open science. International Journal of Industrial Organization 27, 449-462.
45. Murray, F., 2010. The oncomouse that roared: Hybrid exchange strategies as a source of distinction at the boundary of overlapping institutions. American Journal of sociology 116, 341-388.
46. Nelson, B., 2009. Data sharing: Empty archives. Nature News 461, 160-163.

47. Rafols, I., Porter, A.L., Leydesdorff, L., 2010. Science overlay maps: A new tool for research policy and library management. Journal of the American Society for information Science and Technology 61, 1871-1887.
48. Riffle, M., Davis, T.N., 2010. The yeast resource center public image repository: a large database of fluorescence microscopy images. BMC bioinformatics 11, 263.
49. Rosenberg, S.A., 1996. Secrecy in medical research. Mass Medical Soc.
50. Scheliga, K., Friesike, S., 2014. Putting open science into practice: A social dilemma? First Monday 19.
51. Shapira, P., Wang, J., 2010. Follow the money. Nature 468, 627-628.
52. Shibayama, S., Walsh, J.P., Baba, Y., 2012. Academic entrepreneurship and exchange of scientific resources: Material transfer in life and materials sciences in Japanese universities. American Sociological Review 77, 804-830.
53. Smith, R., 2006. Commentary: The power of the unrelenting impact factor—Is it a force for good or harm? International Journal of Epidemiology 35, 1129-1130.
54. Tenopir, C., Allard, S., Douglass, K., Aydinoglu, A.U., Wu, L., Read, E., Manoff, M., Frame, M., 2011. Data sharing by scientists: practices and perceptions. PloS one 6, e21101.
55. Thursby, M., Thursby, J., Haeussler, C., Jiang, L., 2009. Do academic scientists share information with their colleagues? Not necessarily.
56. Uzzi, B., Mukherjee, S., Stringer, M., Jones, B., 2013. Atypical combinations and scientific impact. Science 342, 468-472.
57. Wagner, C.S., Whetsell, T.A., Mukherjee, S., 2019. International research collaboration: Novelty, conventionality, and atypicality in knowledge recombination. Research Policy 48, 1260-1270.
58. Walsh, J.P., Cohen, W.M., Cho, C., 2007. Where excludability matters: Material versus intellectual property in academic biomedical research. Research Policy 36, 1184-1203.
59. Walsh, J.P., Hong, W., 2003. Secrecy is increasing in step with competition. Nature 422, 801-802.
60. Wang, J., Veugelers, R., Stephan, P., 2017. Bias against novelty in science: A cautionary tale for users of bibliometric indicators. Research Policy 46, 1416-1436.
61. Whitlock, M.C., 2011. Data archiving in ecology and evolution: best practices. Trends in Ecology & Evolution 26, 61-65.
62. Zenk-Möltgen, W., Akdeniz, E., Katsanidou, A., Naßhoven, V., Balaban, E., 2018. Factors influencing the data sharing behavior of researchers in sociology and political science. Journal of Documentation 74, 1053-1073.

TABLES

Table 1. Correlation and Summary Statistics

	Original Data Disclosed	Popularity	Funding	ln(nRef+1)	Int Collabo	Novelty
Original Data Disclosed	1.00					
Popularity	0.06	1.00				
Funding	0.08	0.07	1.00			
ln(nRef+1)	0.08	-0.12	0.18	1.00		
Int Collabo	0.04	-0.02	0.10	0.08	1.00	
Novelty	0.07	0.23	0.22	0.22	0.01	1.00
Obs	310,900	305,245	310,900	310,900	310,900	310,150
Mean	0.05	-5.90	0.77	3.57	0.28	0.26
Std.Dev	0.22	1.64	0.42	0.54	0.45	1.28
Min	0	-10.41	0	0	0	-17.00
Max	1	-0.61	1	6.93	1	4.71

Table 2. Descriptive Comparison

	Data Disclosed (Obs.15,271)				No Data Disclosed (Obs.295,629)			
Variable	Mean	Std. Dev.	Min	Max	Mean	Std. Dev.	Min	Max
FWD2010	1.16	3.14	0	108	0.90	2.34	0	292
FWD2011	5.72	9.96	0	246	3.89	7.12	0	712
FWD2012	7.37	12.80	0	346	5.04	9.37	0	932
FWD2013	7.34	13.66	0	515	5.11	10.52	0	1525
FWD2014	7.11	14.57	0	728	5.03	11.70	0	2052
FWD2015	6.64	14.74	0	803	4.79	12.26	0	1677
FWD2016	6.18	15.40	0	918	4.55	12.80	0	1534
FWD2017	5.80	16.19	0	1083	4.30	13.39	0	1694
FWD2018	5.30	16.25	0	1131	4.03	13.73	0	1958
Popularity	-5.45	1.60	-10.41	-0.61	-5.92	1.64	-10.41	-0.61
Funding	0.91	0.28	0	1	0.77	0.42	0	1
ln(nRef+1)	3.76	0.51	0	6.70	3.56	0.54	0	6.93
Int Collabo	0.35	0.48	0	1	0.27	0.44	0	1
Novelty	0.67	0.82	-10.88	4.51	0.24	1.29	-17	4.71

Table 3. Generalized Negative Binomial Regression

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
VARIABLES	FWD2010	FWD2011	FWD2012	FWD2013	FWD2014	FWD2015	FWD2016	FWD2017	FWD2018
Original Data Disclosed	0.0302	0.0396***	0.0252**	0.0258**	-0.00211	-0.00565	-0.00748	-0.0264	-0.0310*
	(0.0218)	(0.0118)	(0.0113)	(0.0121)	(0.0133)	(0.0146)	(0.0159)	(0.0172)	(0.0183)
Popularity	0.0195***	0.0350***	0.0409***	0.0450***	0.0468***	0.0466***	0.0471***	0.0471***	0.0463***
	(0.00244)	(0.00149)	(0.00152)	(0.00168)	(0.00184)	(0.00194)	(0.00204)	(0.00215)	(0.00226)
Funding	-0.0335***	0.0380***	0.0496***	0.0421***	0.0390***	0.0312***	0.0230***	0.0144	0.00924
	(0.00982)	(0.00616)	(0.00633)	(0.00718)	(0.00788)	(0.00850)	(0.00888)	(0.00926)	(0.00969)
ln(nRef+1)	0.310***	0.382***	0.405***	0.416***	0.421***	0.430***	0.430***	0.435***	0.443***
	(0.00962)	(0.00590)	(0.00579)	(0.00630)	(0.00676)	(0.00729)	(0.00766)	(0.00814)	(0.00856)
Int Collabo	0.153***	0.119***	0.110***	0.107***	0.107***	0.105***	0.101***	0.107***	0.104***
	(0.00776)	(0.00481)	(0.00489)	(0.00540)	(0.00583)	(0.00628)	(0.00659)	(0.00698)	(0.00733)
Novelty	0.00194	0.0120***	0.0126***	0.00897***	0.00613**	0.00452	0.00387	0.00513	0.00411
	(0.00429)	(0.00270)	(0.00265)	(0.00288)	(0.00306)	(0.00321)	(0.00334)	(0.00349)	(0.00365)
Constant	-2.079***	-0.632***	-0.0890	-0.108	-0.0952	-0.00552	0.131	-0.0768	-0.0547
	(0.250)	(0.144)	(0.162)	(0.137)	(0.204)	(0.244)	(0.277)	(0.378)	(0.384)
lnalpha	0.519***	-0.464***	-0.478***	-0.354***	-0.242***	-0.149***	-0.0570***	0.0167***	0.0948***
	(0.00687)	(0.00558)	(0.00551)	(0.00578)	(0.00600)	(0.00613)	(0.00617)	(0.00637)	(0.00653)
Observations	304,906	304,906	304,906	304,906	304,906	304,906	304,906	304,906	304,906
Journal FE	YES	YES	YES	YES	YES	YES	YES	YES	YES

Robust standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Table 4. Panel Regression with the Data-Disclosed Articles only

	2010 vs. 2011	2010 vs. 2012	2010 vs. 2013	2010 vs. 2014	2010 vs. 2015	2010 vs. 2016	2010 vs. 2017	2010 vs. 2018
VARIABLES	FWD	FWD	FWD	FWD	FWD	FWD	FWD	FWD
Original Data Disclosed	1.884***	0.500***	0.349***	0.138***	-0.0482	-0.226***	-0.0917	0.0453
	(0.0675)	(0.0642)	(0.0126)	(0.0161)	(0.0646)	(0.0164)	(0.182)	(0.121)
Constant	-0.451***	0.903***	1.059***	1.252***	1.447***	1.615***	1.491***	1.353***
	(0.0679)	(0.0650)	(0.0158)	(0.0187)	(0.0653)	(0.0190)	(0.182)	(0.121)
Observations	86,967	82,215	100,431	90,297	81,864	92,673	81,468	81,630
Number of Articles	9,663	9,135	11,159	10,033	9,096	10,297	9,052	9,070
Article FE	YES	YES	YES	YES	YES	YES	YES	YES

Conditional Fixed Effect Negative Binomial Model employed, Standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Table 5. Regression with the Interaction term between Funding and Original Data Disclosed

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
VARIABLES	FWD2010	FWD2011	FWD2012	FWD2013	FWD2014	FWD2015	FWD2016	FWD2017	FWD2018
FundingxOriginal Data	-0.0206	-0.0227	-0.00170	-0.0271	-0.0177	0.00643	-0.0345	-0.0450	-0.0244
	(0.0562)	(0.0327)	(0.0321)	(0.0341)	(0.0385)	(0.0385)	(0.0417)	(0.0455)	(0.0485)
Original Data Disclosed	0.0488	0.0602*	0.0268	0.0503	0.0139	-0.0115	0.0237	0.0143	-0.00889
	(0.0542)	(0.0317)	(0.0312)	(0.0332)	(0.0376)	(0.0373)	(0.0404)	(0.0438)	(0.0463)
Popularity	0.0195***	0.0350***	0.0409***	0.0450***	0.0468***	0.0466***	0.0471***	0.0471***	0.0463***
	(0.00244)	(0.00149)	(0.00152)	(0.00168)	(0.00184)	(0.00195)	(0.00204)	(0.00215)	(0.00226)
Funding	-0.0329***	0.0386***	0.0497***	0.0428***	0.0394***	0.0310***	0.0239***	0.0155*	0.00985
	(0.00995)	(0.00626)	(0.00643)	(0.00731)	(0.00801)	(0.00865)	(0.00903)	(0.00941)	(0.00984)
Ln(nRef+1)	0.310***	0.382***	0.405***	0.416***	0.421***	0.430***	0.430***	0.435***	0.443***
	(0.00962)	(0.00590)	(0.00579)	(0.00630)	(0.00676)	(0.00729)	(0.00766)	(0.00814)	(0.00856)
Int Collaboration	0.153***	0.119***	0.110***	0.107***	0.107***	0.105***	0.101***	0.107***	0.104***
	(0.00776)	(0.00481)	(0.00489)	(0.00540)	(0.00583)	(0.00628)	(0.00659)	(0.00698)	(0.00733)
Novelty	0.00192	0.0119***	0.0126***	0.00896***	0.00612**	0.00452	0.00385	0.00511	0.00410
	(0.00429)	(0.00270)	(0.00265)	(0.00288)	(0.00306)	(0.00321)	(0.00334)	(0.00349)	(0.00365)
Constant	-2.079***	-0.632***	-0.0890	-0.109	-0.0955	-0.00542	0.131	-0.0777	-0.0552
	(0.250)	(0.144)	(0.162)	(0.137)	(0.204)	(0.244)	(0.277)	(0.378)	(0.384)
lnalpha	0.519***	-0.464***	-0.478***	-0.354***	-0.242***	-0.149***	-0.0570***	0.0167***	0.0948***
	(0.00687)	(0.00558)	(0.00551)	(0.00578)	(0.00600)	(0.00613)	(0.00617)	(0.00637)	(0.00653)
Observations	304,906	304,906	304,906	304,906	304,906	304,906	304,906	304,906	304,906
Journal FE	YES	YES	YES	YES	YES	YES	YES	YES	YES

Robust standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Table 6. Test on Mattew Effect

VARIABLES	(1) FWD 2010	(2) FWD 2011	(3) FWD 2012	(4) FWD 2013	(5) FWD 2014	(6) FWD 2015	(7) FWD 2016	(8) FWD 2017	(9) FWD 2018
Original Data Disclosed	-0.139**	0.0505	0.0803*	0.0644	0.0419	0.0177	-0.00705	-0.0262	-0.0402
	(0.0619)	(0.0380)	(0.0477)	(0.0489)	(0.0750)	(0.0882)	(0.104)	(0.126)	(0.139)
Popularity	0.112***	0.145***	0.175***	0.182***	0.181***	0.181***	0.159***	0.165***	0.155***
	(0.0414)	(0.0285)	(0.0289)	(0.0286)	(0.0331)	(0.0345)	(0.0373)	(0.0401)	(0.0455)
Funding	0.0177	0.338**	0.383**	0.383**	0.433**	0.452**	0.417**	0.480**	0.435*
	(0.134)	(0.156)	(0.158)	(0.157)	(0.197)	(0.192)	(0.197)	(0.208)	(0.236)
Ln(nRef+1)	0.172	0.190	0.220*	0.201	0.176	0.133	0.0734	0.0618	0.0354
	(0.156)	(0.119)	(0.127)	(0.134)	(0.153)	(0.156)	(0.173)	(0.183)	(0.209)
Int Collabo	0.452**	0.375**	0.393**	0.352**	0.323**	0.310*	0.316*	0.291*	0.276
	(0.179)	(0.159)	(0.158)	(0.160)	(0.162)	(0.164)	(0.179)	(0.170)	(0.176)
Novelty	0.0761	0.0898	0.146*	0.165*	0.170*	0.145	0.147	0.163	0.158
	(0.0827)	(0.0823)	(0.0854)	(0.0863)	(0.0946)	(0.0899)	(0.0936)	(0.0992)	(0.106)
Constant	0.434	1.553***	1.681***	1.774***	1.791**	1.910***	1.972**	1.927**	1.915**
	(0.605)	(0.574)	(0.614)	(0.626)	(0.725)	(0.727)	(0.766)	(0.809)	(0.900)
lnalpha	0.940***	0.127	0.108	0.173*	0.278***	0.322***	0.401***	0.501***	0.596***
	(0.0961)	(0.0990)	(0.0943)	(0.0918)	(0.0995)	(0.102)	(0.105)	(0.119)	(0.123)
Observations	3,478	3,478	3,478	3,478	3,478	3,478	3,478	3,478	3,478
Journal FE	Clustered	Clustered	Clustered	Clustered	Clustered	Clustered	Clustered	Clustered	Clustered

Robust standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Table 7. Moderating Effect by Scholarly Reputation of Journals

VARIABLES	(1) FWD2010	(2) FWD2011	(3) FWD2012	(4) FWD2013	(5) FWD2014	(6) FWD2015	(7) FWD2016	(8) FWD2017	(9) FWD2018
JIF above MedianxData Disclosed	**-0.184*** **	**0.0845*** **	**0.173*** **	**0.174*** **	**0.230*** **	**0.206*** **	**0.188*** **	**0.260*** **	**0.241*** **
	(0.0558)	**(0.0306)**	**(0.0300)**	**(0.0315)**	**(0.0342)**	**(0.0366)**	**(0.0398)**	**(0.0436)**	**(0.0461)**
JIF above Median	**0.727*** **	**0.701*** **	**0.662*** **	**0.673*** **	**0.697*** **	**0.705*** **	**0.721*** **	**0.726*** **	**0.730*** **
	(0.00852)	**(0.00553)**	**(0.00582)**	**(0.00666)**	**(0.00758)**	**(0.00867)**	**(0.00968)**	**(0.0108)**	**(0.0119)**
Original Data Disclosed	**0.153*** **	**0.0115**	**-0.0737*** **	**-0.0908*** **	**-0.163*** **	**-0.156*** **	**-0.154*** **	**-0.228*** **	**-0.231*** **
	(0.0504)	**(0.0265)**	**(0.0259)**	**(0.0270)**	**(0.0290)**	**(0.0308)**	**(0.0330)**	**(0.0357)**	**(0.0367)**
Popularity	0.0665***	0.0801***	0.0854***	0.0873***	0.0896***	0.0883***	0.0850***	0.0835***	0.0807***
	(0.00281)	(0.00193)	(0.00197)	(0.00220)	(0.00249)	(0.00261)	(0.00276)	(0.00295)	(0.00321)
Funding	0.101***	0.162***	0.132***	0.108***	0.101***	0.0890***	0.0703***	0.0549***	0.0462***
	(0.0112)	(0.00777)	(0.00801)	(0.00905)	(0.0104)	(0.0113)	(0.0126)	(0.0140)	(0.0156)
lnref	0.322***	0.358***	0.373***	0.387***	0.396***	0.401***	0.395***	0.399***	0.407***
	(0.0101)	(0.00694)	(0.00713)	(0.00823)	(0.00943)	(0.0111)	(0.0128)	(0.0147)	(0.0166)
Int Colalbo	0.307***	0.230***	0.218***	0.213***	0.211***	0.211***	0.216***	0.222***	0.220***
	(0.00942)	(0.00655)	(0.00681)	(0.00763)	(0.00860)	(0.00929)	(0.0100)	(0.0107)	(0.0116)
Novelty	0.0390***	0.0735***	0.0672***	0.0557***	0.0442***	0.0401***	0.0307***	0.0273***	0.0191***
	(0.00380)	(0.00255)	(0.00249)	(0.00267)	(0.00291)	(0.00314)	(0.00337)	(0.00361)	(0.00387)
Constant	-1.577***	-0.185***	0.113***	0.106***	0.0633*	-0.0100	-0.0530	-0.124**	-0.231***
	(0.0359)	(0.0247)	(0.0256)	(0.0295)	(0.0346)	(0.0412)	(0.0489)	(0.0576)	(0.0659)
lnalpha	0.848***	-0.0669***	-0.111***	-0.0138**	0.0855***	0.170***	0.253***	0.321***	0.396***
	(0.00707)	(0.00573)	(0.00595)	(0.00631)	(0.00673)	(0.00723)	(0.00766)	(0.00821)	(0.00869)
Observations	304,906	304,906	304,906	304,906	304,906	304,906	304,906	304,906	304,906
Journal FE	NO	NO	NO	NO	NO	NO	NO	NO	NO

Robust standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

FIGURES

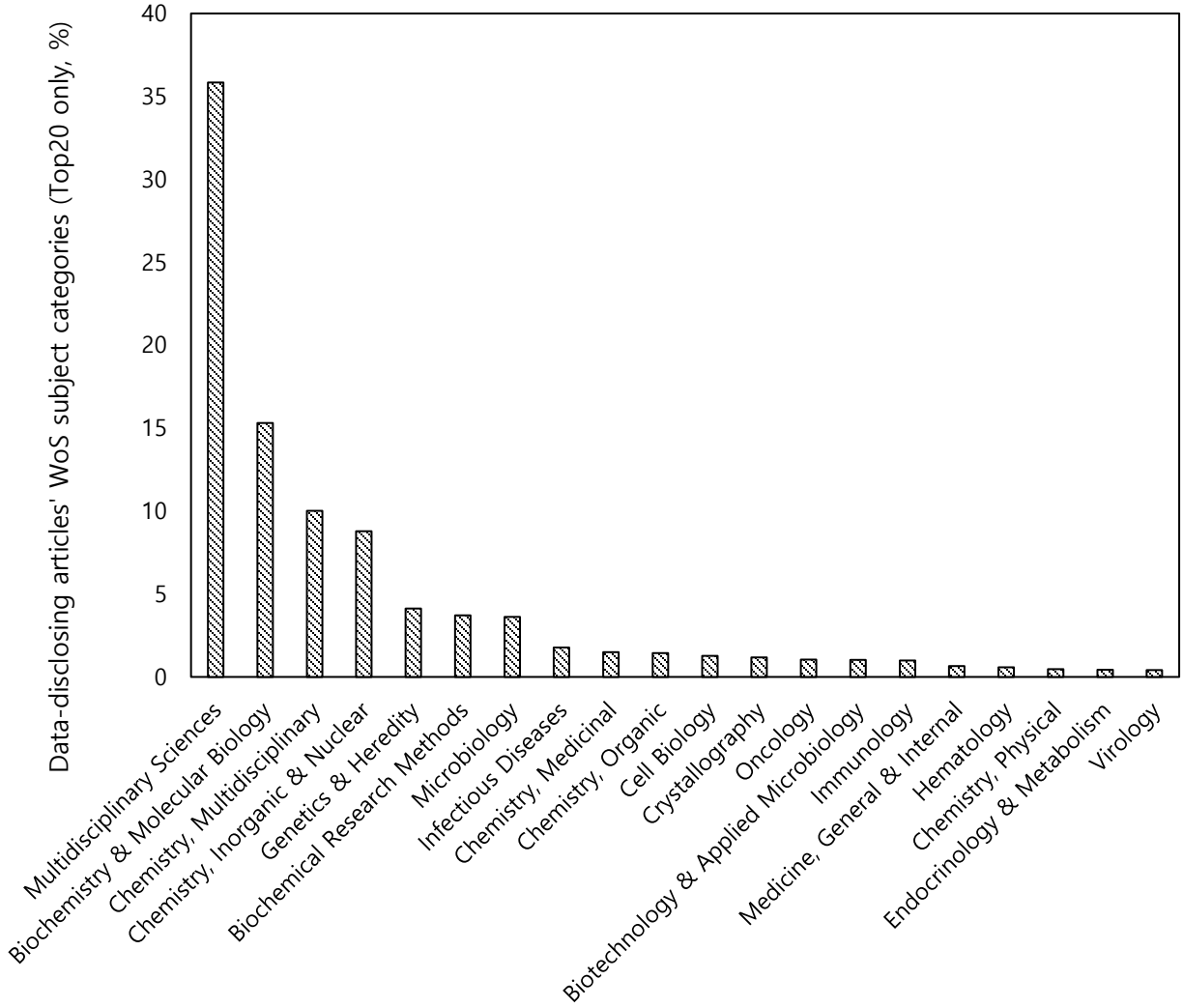

Figure 1. Distribution of WoS Subject Categories of data-disclosing articles (Top 20 only)

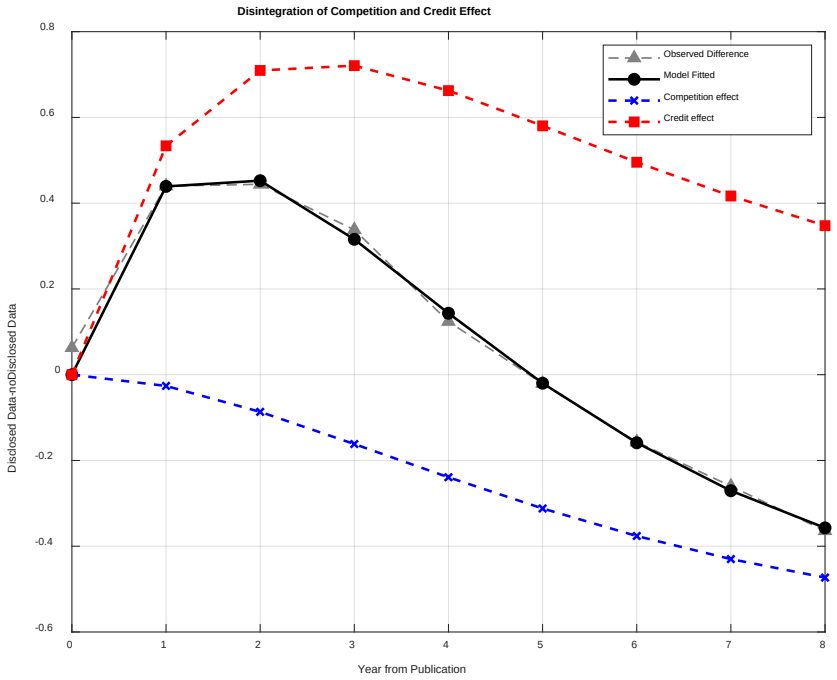

Figure 2. Decoupling the Credit and Competition Effect

Note: Observed difference was fitted to the OLS regression model

Appendix. Simulation with the Citation Function Model

Using the citation function model, we draw the graph of expected citations count of an article received from its publication year to nine years after. We do this for each of the following four scenarios: (1) No credit effect and no competition effect exist (no change in β_1 nor β_2), (2) Only credit effect exists (increase in β_2 but no change in β_1), (3) only competition effect exists (increase in β_1 but no change in β_2), (4) both credit and competition effects exist (increase in β_1 and β_2). As the purpose of this simulation is to check the difference in the pattern of citation by the scenarios, we set the parameter to arbitrary numbers. Figure AP presents the result.

[Insert Figure AP about here]

The expected citation count by scenario (1) is provided on the left top. Because no credit effect and no competition effects are presumed in this scenario, there is virtually no difference (blue solid line) in the expected citation count between when data was disclosed (black solid line) and when data is not disclosed (gray dashed line). The figure on the right top presents the expected citation count by scenario (2). In this "credit effect only" scenario, the difference in the expected citation count between data-disclosed and data-not disclosed article (blue solid link) increases for a certain period but decreases later. Note that the blue solid line never goes below zero. This indicates that if the credit effect exists only, the data disclosing article will receive more citations always than a comparable article that does not disclose the data. On the left bottom, the expected citation count in scenario (3) is presented. By the competition effect, the expected citation count for the data-disclosing article is fewer citations than that of a comparable article that does not disclose the data. Hence, the difference in the citation count becomes below zero always. Finally, the right bottom figure presents the expected citation count in scenario (4). For the early period, the data-disclosing article gest more citations than the article that does not disclose the data. Later, this pattern is reversed; the data-disclosed article receives fewer citations than its counterpart. Note that, in this scenario, the blue solid line goes below zero in the late period.

In our empirical analysis, we showed that the articles with disclosed data received more citations than the counterparts in the early period. Yet, this citation difference gradually disappeared and was reversed in the late period. The observed pattern in our empirical analysis is in line with the scenario (4). From this analysis, we argue that our empirical findings were originated from the interplay between the credit effect and the competition effect.

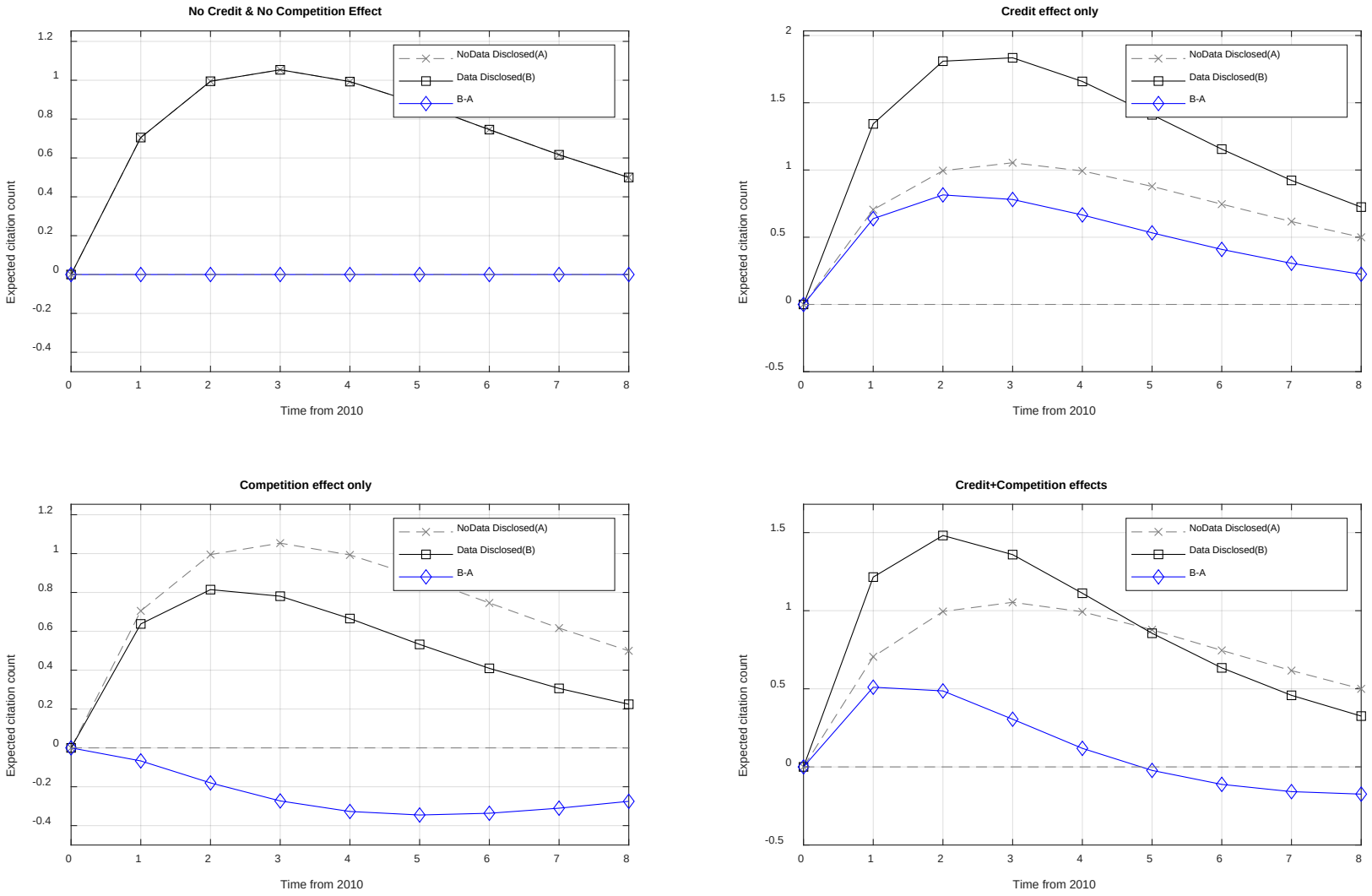

Figure AP. Simulation Result using Citation Function