Drinking from the Firehose: Preprints, Chinese Scientists, and the Diffusion of Research on COVID-19

Caroline Fry*, Megan MacGarvie**

5/15/2020

Abstract

Platforms such as preprints websites have become an increasingly important way to rapidly disseminate new knowledge prior to peer review. While these platforms are accessible to scientists and audiences around the world, there is considerable uncertainty around the knowledge itself, particularly with articles produced by Chinese scientists. This study explores how readers allocate attention across preprints in a context of great urgency (the initial months of the COVID-19 pandemic), and the extent to which the community reduces uncertainty itself through endorsements on social media. We find that preprints with authors from Chinese institutions receive less attention than preprints with authors from the rest of the world. We also document that self-organizing screening mechanisms such as twitter endorsements drive attention, although these are less common and no more effective for Chinese authored preprints, replicating the original attention gap. The results suggest that platforms designed to, or used to, promote unfettered access to early research findings do not necessarily lead to democratization of science in a context of high urgency and uncertainty.

1. Introduction

The COVID-19 pandemic has led to an urgent need for scientific research related to the virus. The rapid spread and severity of the disease has forced researchers, clinicians and policy makers to quickly scale up efforts to combat the virus with limited time for evaluation and analysis. Although the peer review process, traditionally the mechanism through which scientific contributions are screened for accuracy and relevance, has accelerated, alternative platforms have emerged to help decision-makers in research and public health initiatives access knowledge as soon as it is produced, in the form of preprint articles. Preprint servers such as medRxiv.org and bioRxiv.org have hosted approximately 6,000 new COVID-19 articles as of the end of June, 2020. Much of this research has emerged from China, where the number of scientific publications has risen rapidly in the past decade (Xie and Freeman 2019), and was home to the earliest reported cases of COVID-19 and the first published articles on the virus.

The publication of research on preprint servers both facilitates the rapid diffusion of knowledge and requires researchers to evaluate articles before they have been fully vetted. Epidemiologist Marc Lipsitch has described the surge of research as a "firehose". Anthony Fauci, head of the US National Institute of Allergy and Infectious Diseases (NIAID) has said: "Eleven o'clock, 12 o'clock comes and you have 25 of these things to read…You can't ignore them…[but] it gets a little confusing what you can really believe."[1]

This paper studies the diffusion of a global public good -- scientific knowledge -- in the context of extreme urgency, where there is a large quantity of studies authored by scientists from around the globe. Although knowledge is a public good, scientific journals create artificial scarcity through copyright restrictions and high financial costs for authors and users alike (McCabe and Snyder 2018). The selection process, which screens articles for quality, may also biased in favor of certain types of articles or authors (Ross et al 2006; Lee et al 2013; Tomkins et al 2017). In order to bring articles to readers as soon as they written, preprint servers have recently emerged to publish work prior to peer review. Online preprint platforms remove many of these frictions for authors around the world, which is particularly useful in a crisis or a time of great urgency, allowing readers to access a complete picture of the latest research findings. Yet open access platforms have been found to have a small or negligible causal impact on citations to scientific papers (Gaule and Maystre 2011). And without the quality stamp of peer review, there is a great deal of uncertainty surrounding articles posted on these platforms. Readers may resort to using signals, particularly as they pertain to prior reputation of the authors, in allocating their attention. This could lead to less attention to preprints authored by Chinese scientists in particular, given their reputation arising from reports of fraud and misconduct (Hvistendahl 2013; Hvistendahl 2020) and concern that publishing incentives may have in some cases prioritized quantity over quality (Freeman and Huang 2015, Mallapaty 2020).

One way that the community of readers attempt to reduce the uncertainty associated with the quality of these preprints is to self-organize and screen preprints through endorsements on social networking platforms. Endorsements may influence the beliefs that readers have about the quality of different types of articles and subsequently drive attention. We examine whether endorsements

* University of Hawai'i at Manoa. ** Boston University and NBER.
[1] Lipsitch and Fauci quoted in Kupferschmidt (2020).

about preprints disproportionately benefit Chinese relative to non-Chinese authored preprints, thus decreasing the attention gap. Existing theories and evidence are ambiguous with respect to the effect of endorsements on different types of articles or authors. On the one hand, an endorsement could particularly benefit more disadvantaged groups of articles or authors as there is more uncertainty surrounding them (Long et al 1979; Stuart et al 1999), leading to a larger positive update in beliefs. On the other hand, endorsements could further penalize more disadvantaged groups as readers could discount new information (Hill and Stein 2020), or not change their beliefs enough to result in meeting the quality threshold for attention, giving further benefit to initially advantaged groups.

We base our empirical analysis on 4,447 preprint articles on the topic of COVID-19 posted in early 2020 on two preprint servers medRxiv.org and bioRxiv.org, which were the two most popular platforms for coronavirus research at the time of writing. These platforms played a large role in diffusing urgently needed science during the early months of the COVID-19 global pandemic, and we document that Chinese authors dramatically increase the rate of posting of preprints on the two study platforms during COVID-19. We ask how decision makers choose which articles to read when attention must be rationed and conventional signals of quality are not yet available.

We report three main findings. First, Chinese-authored preprints receive less attention (measured through downloads to the preprint) than those by authors from other countries, and particularly less than US-authored preprints. This relationship holds even after accounting for many of the characteristics that a reader observes and other quality measures, including the author's institution rank, the access Chinese authors have to early COVID-19 cases and the quality of the journal of ultimate publication of the preprint. That said, this gap in attention narrows throughout the lifetime of a preprint, suggesting that decision makers may take slightly longer to evaluate the merits of knowledge produced by scientists located outside of the historic centers for scientific excellence.

Second, the data indicate that endorsements drive attention: preprints mentioned by twitter users with more than 10,000 followers are more downloaded than those not mentioned by such a high-profile twitter user. This finding is consistent with Peoples et al (2016) who show that tweeted about papers receive more citations.

Third, we document an even greater gap in terms of twitter mentions and endorsements between Chinese and non-Chinese authored articles than we observe in the downloads. In addition, we find no evidence of a Chinese endorsement premium: the benefit from endorsements is no higher for Chinese authored preprints than for other preprints, and in fact it is significantly less in terms of the response of tweets to the endorsement. This finding suggests that attempts to resolve uncertainty of a given piece of knowledge by self-organizing networks, such as those that exist on twitter, may not result in a narrowing in the attention gap.

This paper makes two main contributions. First, the findings in this study contribute to the broad literature studying the determinants of knowledge diffusion, particularly under conditions of high urgency and uncertainty. Several studies have documented biases in the peer review process (Ross et al 2006; Lee et al 2013; Tomkins et al 2017) and issues associated with fee structures of journals (McCabe and Snyder 2018) that can restrict widespread diffusion of new knowledge. The response to which has been a rise in the use of alternatives to peer review (Ellison 2011; Fraser 2020). However, the extent to which platforms designed to remove entry barriers and costs exacerbate biases in attention, particularly under conditions of great urgency and uncertainty, is less well understood. Using novel data on attention to preprints during the COVID-19 pandemic, we document that preprint platforms do not resolve the gap in attention, and in fact that a mechanism by which communities self-organize to vet new knowledge through twitter endorsements replicates the bias found prior to the endorsement process. This result is consistent with a Bayesian explanation in which the community of decision makers infer the quality of a preprint given priors about the authors and update their priors according to new information and this in turn drives the allocation of attention.

Second, this paper also relates to the literature on the globalization of science, and particularly of the rising profile of Chinese science. As science becomes increasingly global, one major challenge is how decision makers can assess work produced by scientists at lesser known institutions. Our findings suggest that decision makers who use signals of reputation could be overlooking potentially important science emanating from China in particular. This has consequences for the diffusion of knowledge and the generation of new knowledge. To the extent that scientists are standing on the shoulders of familiar giants, the progress of global science and public health and economic advances could be limited.

2. Theoretical Background

Author institutional affiliation and knowledge diffusion. Scientific knowledge is complex, fast-changing and uncertain (Polanyi 1958; Zucker and Darby 1996; Jones 2009; Freedman et al 2015). How decision makers allocate finite attention and select which articles to read, therefore, is a core question in the economics of innovation. One way that decision makers can navigate the literature is through the use of informational cues or signals that are easily observable. A signal that is prominently featured on most scientific publications is authors' prior reputation or institutional affiliation, which translates into a quality signal for well-versed audiences.

Global science exhibits a core/periphery network structure with geographic affiliation in many ways determining position in the network (Wagner and Leydesdorff 2005; Leydesdorff and Wagner 2008; Zelnio 2012). The selective core of the network is associated with privilege, prestige and control (Clauset et al 2015), while peripheral actors suffer from lack of legitimacy and trust. In this study we consider preprints with a USA affiliation as emanating from the core, and those from China as emanating from the periphery. Until recently, the United States has been the global leader in scientific output and is responsible for the majority of the world's resources for science, and home to some of the world's most prestigious research institutions. In contrast, despite the notable rise in scientific output in China since the 1990s, reports of fraudulent science and misconduct are relatively common, lowering the global reputation of Chinese scientists (Hvistendahl 2013).[2] Hvistendahl confirms this bias against Chinese science and scientists, and describes that when China has achieved technological advances, "there's been a tendency to assume that they're products of theft, particularly by immigrant scientists" (Hvistendahl 2020).

Audiences may pay more attention to articles with authors from traditional locations of scientific excellence, and less to Chinese authored articles, due to the assumption that the work emanating from these institutions is of higher quality (Kim and King 2014), or for co-ordination purposes (Correll et al 2017). To the extent that platform users are mostly from traditional centers

[2] https://www.nytimes.com/2017/10/13/world/asia/china-science-fraud-scandals.html
https://qz.com/978037/china-publishes-more-science-research-with-fabricated-peer-review-than-everyone-else-put-together/
https://cen.acs.org/policy/research-funding/70-years-US-suspicion-toward/98/i11
https://wenr.wes.org/2018/04/the-economy-of-fraud-in-academic-publishing-in-china

of scientific excellence, there could also be a "home bias" of readers resulting in the same observational bias against Chinese research (Freeman and Xie 2020).

Prior research has documented that researchers rely on informational cues when allocating attention and assessing scientific information. Simcoe and Waguespack (2011) show that contributions to electronic engineering message boards by high-status authors are more likely to be mentioned in an online forum, especially when attention is scarce. Bikard (2018) uses simultaneous discoveries in academia and industry to document that inventors are more likely to reference work produced in academia in patent applications than identical work produced in industry. While this evidence provides a preliminary step in suggesting that signals do matter in the diffusion of knowledge, the question of the extent to which decision makers rely on signals to allocate scarce attention when using different platforms of communication, under different conditions, and as new information is revealed is increasingly important. Some observers have suggested that science communication is moving away from the model of publication in peer-reviewed journals and toward posting of analysis and code with debate on open platforms (see Sugimoto et al 2017 for a review of the literature on use of social media by researchers),[3] with social media platforms such as Twitter playing a more important role in communicating pre-publication research results.[4] With low costs to communicating research findings, and built in formats for public discussion and probing of new findings, these platforms enable the rapid communication of findings and new data, and in many ways democratize the communication of research findings. There is particular interest in new platforms as biases towards higher status researchers, particularly based on geographic affiliation, associated with the peer review process have been documented across a number of disciplines (Ceci and Peters 1982; Ross et al 2006; Smith et al 2014). However, the nature of interaction on these platforms, and the visibility of the author's name and affiliation, may have implications for which types of research receive the most attention, particularly in a time of great uncertainty and urgency.

[3] See, for example, "The Scientific Paper is Obsolete," (James Somers, *The Atlantic,* April 5, 2018*)*,
[4] CRISPR pioneer Jennifer Doudna has written that "Preprints are not peer-reviewed or formally evaluated for scientific quality… preprints are quickly dissected on social media, enabling scientists to quickly replicate and build on findings. The rapid and open access to research will improve the communication of science and the involvement of non-scientists in the enterprise." https://www.economist.com/by-invitation/2020/06/05/jennifer-doudna-on-how-covid-19-is-spurring-science-to-accelerate

In parallel to the rise of democratizing platforms, science is becoming more global and while scientific outlets were traditionally dominated by a handful of prestigious institutions, we are seeing a rise in research produced by a more diverse community (NSF science and engineering indicators 2020; Holmgren and Schnitzer 2004). Thus, while it is documented that status signals can shape the diffusion of knowledge, their role in different platforms designed to diffuse knowledge, and the manner in which the community attempts to resolve uncertainty remains unknown.

Resolving uncertainty through endorsements. Traditional research on status benefits argues that signals of quality are most relevant in contexts where true quality cannot be observed (Podolny 1994; Podolny ad Phillips 1996). Podolny (2001) notes that "the rewards of status are contingent on the uncertainty that buyers face" (p. 36). Where there is greater uncertainty, decision makers will rely on prior expectations of quality to a greater extent. One way that audiences resolve uncertainty is to measure an actor's status or quality using the status of their affiliates, who provide implicit or explicit endorsements (Blau 1964; Long et al 1979; Stuart et al 1999). While endorsements are likely to drive attention as they alter the beliefs about quality of the focal actor or product, the theory and evidence on who benefits most from an endorsement is ambiguous.

First, endorsement comes with a reputational cost to the endorser. This may lead endorsers do be less willing to publicly declare the quality of an actor or product when the actor, or actor type, has a worse reputation. Second, even after endorsed, the endorsement itself could have a varying effect on different types. On the one hand, endorsements to more disadvantaged groups resolve considerable uncertainty and can result in a greater uptick in beliefs. This is documented in a variety of settings. Stuart et al (1999) find that endorsements from a higher status affiliate (who transfer some of their own status onto the endorsee and provide the community with some information about quality) for entrepreneurial firms are more effective for firms that are newer and smaller. Azoulay et al (2014) document that a quality signal from winning a prize has a greater effect on lower status and younger scientists, and Kovacs and Sharkey (2014) find that first time authors benefit from receiving a prize more than more seasoned authors.

However, to the extent that status is a co-ordination device (Correll et al 2017), or that prior beliefs are difficult to change, an endorsement could actually exacerbate the status gap. Some recent studies document a widening of the status gap in response to new information. Jin et al

(2019) explore the role of retractions of scientific information in credit allocation to scientists and find that a retraction is most harmful for less prominent scientists within a team. Hill and Stein (2020) find that high reputation teams that scoop lower reputation teams receive much more positive recognition than lower reputations teams that scoop a higher reputation team.[5] We examine the effect of an endorsement on attention to Chinese and non-Chinese authored articles using a model of beliefs and the allocation of attention.

Bayesian learning. Prior research on uncertainty about author quality has used a Bayesian framework to model beliefs and the effects of new information on those beliefs. (Azoulay et al. 2017, Jin et al. 2019, Reimers and Peukert 2020). Using a simple framework based on De Groot (1970) and similar to the model in Reimers and Peukert (2019), we consider a new preprint as having an uncertain quality W, with a prior distribution that is normal[6] with an unknown mean value μ and precision τ. A decision maker will download (D) a preprint if the expected quality $f(\mu,\tau)$ exceeds a certain threshold γ, which we assume constant across audiences. That is, D = 1 if $f(\mu,\tau)> \gamma$.

Suppose that there two types of preprints, H and L, and a decision maker is deciding where to allocate the most attention. H are those that have authors with a high reputation, known for producing high average quality knowledge, while L are those with authors with lower reputations, known for producing papers of lower average quality or average quality with low precision. Because the expectation of the value μ is greater for a H type than for a L type, at baseline, the decision maker will give more attention to the H type (proposition 1). That is, we assume that $f'(\mu)>0$ and $\mu_H>\mu_L$, therefore it follows that P(D = 1|H type)>P(D = 1|L type).

Proposition 1*. Preprints with high status authors are more likely to be downloaded.*

Consider now that decision makers receive new information about a given preprint, and therefore can learn and update their beliefs according to a Bayesian updating process. The decision-maker observes a random sample of n observations $(X_1,\ldots,X_n)$ from this posterior distribution, which has a mean μ' and precision $\tau+nr$.

$$\mu' = \frac{nr}{\tau+nr}\bar{x} + \frac{\tau}{\tau+nr}\mu \qquad [1]$$

[5] Where to "scoop" means to publish nearly identical results slightly before another team.

[6] We assume for simplicity that the expectation of quality of a preprint is normally distributed, but the logic of the model can be extended to alternative distributions.

Where $\bar{x}$ is the sample mean of the new data. Therefore, the posterior mean is a weighted average formed from new observations $\bar{x}$ and the mean of the prior distribution μ. Comparing the new mean with the old, we can see that

$$\mu' - \mu = \frac{\bar{x}-\mu}{\tau/nr+1} \quad [2]$$

This implies that there will be a change in beliefs when $\bar{x}$ is sufficiently different from μ. When n is low, more weight is placed on the mean of the prior distribution and the change in beliefs is small even if $\bar{x}$ is quite different from μ. However, as n increases (we obtain more pieces of information), more weight is placed on the mean of the observed data and beliefs change more. Therefore, as time passes and the community verifies the accuracy of a study, or more studies are released on a particular topic, more weight will be placed on the cumulative new information. In other words, the beliefs should reflect the true quality of the preprint more accurately as more information is revealed. Reimers and Peukert (2020) find that self-publishing increases the accuracy of publishers' predictions about a novel's market value by increasing the number of observations on similar novels. In our context, we expect status cues to be most important at the start of a preprint's lifetime, when n is low. As n increases with the emergence of more information, we expect attention, as measured by downloads, to be driven less by social cues (all else equal).

Proposition 2. *As time passes, downloads to a given preprint will be driven less by prior reputation of the authors*.

Now consider that decision makers learn about the quality of a piece of knowledge through endorsements from high status actors. In general, endorsements will increase downloads, as beliefs of quality are updated and raised. However, the effects of an endorsement are not likely to be uniform across types. Equation [1] implies that in the case where r is low relative to τ, and if $r < \tau$, more weight is placed on the prior and less on the new information. This suggests that H-type authors, who have precise priors (high τ), may see smaller changes in μ as a result of an endorsement, and conversely, L types who have low precision priors may see greater changes in μ as a result of an endorsement. However, an L type is less likely to meet the threshold γ, even after an endorsement, counteracting the potential larger effect of an endorsement. When $f(\mu', \tau + nr) < \gamma$, the preprint still will not be downloaded after an endorsement. We cannot separate these

two effects in the empirical exercise, but we can determine the overall change in D. With this, we arrive at the following proposition.

Proposition 3. *The impact of endorsements on downloads is lowest for preprints where precision of the update is small relative to the precision of the prior, and for L types that still don't meet the attention threshold ex-post.*

We can also think of r as capturing the amount of uncertainty about an endorsement. An endorsement from a high-status scientist can be viewed as a piece of new information with high precision (r) (as opposed to an endorsement from a lower-status scientist which may be viewed as having lower precision (r)). As shown in equation (2), when r is larger relative to τ, we can expect a bigger change in the estimated value of a preprint relative to priors. This leads to our final proposition.

Proposition 4. *Endorsements from higher-status scientists will result in greater changes in downloads than endorsements from lower-status scientists, all else equal.*

Finally, we consider an alternative response to endorsements, tweets, which could have a different threshold value γ. The threshold for tweeting is based on a different cost/benefit analysis of the audience member, who considers the reputational costs to themselves of tweeting, as well as the cost of the tweet itself. Given these reputational costs, we assume for now that the threshold for tweeting about a preprint is higher than the threshold for downloading a preprint. That said, we would expect the difference in attention to preprints by high- versus low-status authors to be much greater in tweets than in downloads due to the greater threshold required for action. And we would expect an endorsement to have an even lower impact on L types through tweets, as compared to downloads. This leads us to our final proposition:

Proposition 5. *Endorsements from higher-status scientists will result in a smaller change in tweets than in downloads for preprints with lower status authors, all else equal.*

3. Setting, Data, Measures and Descriptive Statistics

a. Setting and Data

We study these propositions using measures of attention to COVID-19 preprint articles. Preprint servers are designed to disseminate the latest knowledge, prior to peer review. They enable scientists to communicate their findings much more quickly than traditional journal publication

would, with a 48-hour process which screens for relevance but not scientific merit. While submission to preprint servers has exploded during the COVID-19 crisis, there is recognition in the scientific community of the challenges associated with the release of new knowledge before it has been fully vetted.[7] We explore the allocation of attention to preprints with a particular focus on Chinese authored preprints in the context of the COVID-19 crisis for three main reasons. First, the use of preprint platforms increased in the early months of the pandemic, and together with the acknowledged uncertainty that preprints present it provides a fascinating setting in which to understand the development of the dynamics of audience attention choices and self-organizing mechanisms. Second, Chinese scientists were responsible for many of the earliest findings in the COVID-19 pandemic, and posted many of them in English on preprint platforms, which renders our tests a conservative estimate of any bias given that the world was watching Chinese science. Third, the lack of travel and conferences during the early months of COVID-19 rules out the possibility that the bias is due to differential in-person exposure to research.

The sample of COVID-19 preprints[8] used in this study comprises 4,447 preprints posted on medRxiv.org and bioRxiv.org between 13th January and May 31st 2020. Figure 1 illustrates the explosion of preprints related to COVID-19 in the early months of 2020, which follows the trend in the increase in COVID-19 cases worldwide. We collect information on each preprint on the author affiliation and other preprint characteristics, as well as corresponding information on the number of times each article is downloaded each month and the tweets association with each preprint.

To construct the suitable control set we collect the full sample of preprints in the same subject areas as the COVID-19 preprints, posted between July 1 2019 and January 30th 2020 on the two preprint servers mentioned above, and generate similar preprint level variables as described for the COVID-19 preprints. This results in a sample of 10,637 control preprints which we trace attention measures for six months after posting consistent with the COVID-19 preprint group.

b. Measures

Dependent Variables. First, we collect information on each preprint on the number of times each article is downloaded (both PDF downloads, and abstract downloads) each month for the five

[7] See "Coronavirus Tests Science's Need for Speed" (Wudan Yan, *New York Times,* April 14, 2020).
[8] Those preprints classified by the preprint server staff prior to posting as related to COVID-19 research.

months after its initial posting using the Altmetrics measures linked to each preprint given in the preprint servers. Second, we measure the number of times a preprint was mentioned on Twitter by day (aggregated at the monthly level for most empirical models), as well as by type of tweeter (scientist or health professional, and whether the tweeter is in the top 95th percentile in terms of number of followers). The former is achieved by running key word searches that would indicate a tweeter is a scientist or health professional on a given tweeters' bio text.[9]

For dates between June 1 2020 and August 20 2020 we also gather daily downloads and tweets associated with each COVID-19 preprint which we use in the event study analysis.

Author location. We explore the role of geographic affiliation in knowledge diffusion. Specifically – we examine the extent to which knowledge diffusion and changes in uncertainty are moderated by a Chinese, or a USA affiliation. We generate the Chinese or USA affiliation dummies for each preprint using address information from the preprint authors' affiliations.

Control variables. One main concern with using geographic location as a status signal in the context of a global pandemic is that some locations have better access to crucial inputs into the scientific process, in this case – proximity to COVID-19 cases. Access to inputs could influence attention through either a signaling mechanism or improving the actual quality of the work. If this access is correlated with measures of status based on geographic location, this could confound our results. In the context our study, the use of a Chinese affiliation is particularly problematic, as this was also the home to the earliest cases. Therefore, we control for preprint authors' access to cases in some of the specifications through measuring their proximity to the outbreak at the time of doing the research.

In order to measure proximity to the outbreak, after extracting the city and country of each author of each preprint, we match each author on every preprint to the cumulative number of COVID-19 cases by country, and for the United States, Canada, Australia and China by region (e.g. state or province) 6 days prior to the posting of the preprint (to account for a lag in research time) using data from the repository for the 2019 Novel Coronavirus Visual Dashboard operated by the Johns Hopkins University Center for Systems Science and Engineering (JHU CSSE).[10] If

[9] Due to a limitation with the data availability at the granular tweet level the number of tweets per preprints is capped at 10,000 (but the number of preprints with more than 10,000 tweets represents just 0.2% of the COVID-19 preprint sample).

[10] Supported by ESRI Living Atlas Team and the Johns Hopkins University Applied Physics Lab (JHU APL).

there are authors from multiple countries or regions on a given preprint we take the maximum cases across preprint authors 6 days prior to posting as the number of COVID-19 cases in author location, and calculate the percentage of all global cases on that day in the author's location. Finally, we generate a dummy variable that takes the value of 1 if a preprint author is affiliated with an institution in Wuhan, the first city in the world with documented COVID-19 cases, becoming a household name in early 2020.

In order to account for the quality of the institution of the preprint author(s), each author in every preprint article is matched to institution rankings using the 2019 Scimago Institutions Research Rankings[11] database as well as the Nature Publishing Index[12]. Where authors do not receive a ranking, this is generally because they do not mention their institution in the preprint author details or they are affiliated with an unranked institution (the ranking requires a company or institution has at least 100 publications overall to be included). We create the highest-ranking institution of any author associated with a given preprint, as well as the ranking for the first and last author of a given preprint. In the majority of specifications we use a dummy variable representing status that takes the value of 1 if an author in the preprint is affiliated with an institution that is in the top 50 globally ranked institutions, according to the Scimago Institutions Research Rankings (details of institutions in this top 50 globally ranked institution list are provided in Appendix A). Additional measures of status include a dummy variable that takes the value of 1 if an author of the preprint has been awarded a prestigious scientific prize (the Nobel Prize, a Lasker Prize, a Breakthrough Prize, or a Wolf Prize in medicine or biological sciences) since the year 2000, a dummy variable that takes the value of 1 if an author is affiliated with an institution in the top 100 globally ranked institutions according to Scimago, a dummy variable that takes the value of 1 is an author is affiliated with an institution in the top 50 ranked institutions in the Nature Publishing Index, and a raw measure of the highest Scimago rank of the authors in a given preprint.

In addition, for each COVID-19 and control preprint article we construct preprint-level variables such as the number of authors associated with a preprint, whether the team is international, scientific discipline, whether the authors provide the data associated with the preprint

[11] The Scimago Institutions Research Ranking incorporates a variety of research output measures in an index at the institutional level and ranks just under 6,500 global institutions across academic, private and government sectors. We use the Research Ranking in 2019 in this study.

[12] The Nature Index tracks contributions to research articles published in 82 high-quality natural science journals, and provides absolute and fractional counts of article publication at the institutional level.

(or whether it is publicly available data), and the preprint project funding source (private/public/philanthropic).

Endorsements. Endorsements from highly followed tweeters who post about a given preprint, or scientific blogs that reference a given preprint are identified in order to assess the impact of reduced uncertainty about said preprint on daily downloads. For tweets we identify tweets about preprints from tweeters who have more than 10,000 followers (the 95th percentile of all tweeters). 296 preprints are identified as being tweeted about by tweeters with more than 10,000 followers between June 5 and July 5 2020. We read through these tweets aside from one tweet we deemed the tone to be positive or neutral in nature. For the tweet 'events', we take the first day between June 5 and July 5 as the event day and analyse changes in daily downloads in the 3 days before the event to that in the 10 days after the event, as compared to changes in daily downloads of comparable preprints that are not tweeted about by a high profile tweeter in the time period.

Publication outcomes. Finally, we match COVID-19 and control preprints to their ultimate publication outcome. First, we collect the full set of 2020, and the second half of 2019, journal articles across biology and medical fields from the Elsevier Scopus database, and all 2020 COVID-19 related articles from Elsevier Scopus database, PubMed Central depository and the Web of Science. Second, we assign preprints to publications using a method that matches the title and authors of the articles. Then, for preprints that do have a matched publication, we identify the 2018 source (journal) normalized impact factor per paper (SNIP), a measure of the frequency with which the average article in a journal has been cited in a particular year corrected for field differences, developed by Elsevier Scopus. This allows for a measure of the post peer review quality, and a proxy for the impact, of the preprint to which we can compare with attention measures. We choose the impact factor, or SNIP, of the publication as opposed to the citations to get a more objective measure of quality that is independent of networks and attention.

c. Descriptive Statistics

Figure 1a illustrates the growth in the rate of posting of scientific articles through the preprint servers in this study in the first six months of 2020. Authors of the preprints are from a number of countries, although particularly in the very early months, preprint authors mostly emanate from the United States and China. That said, the number of preprints from other countries in the world, particularly Germany, the United Kingdom and India increases in the later months in the study

period. By the end of May 2020, authors from 94 countries in the world had posted preprints on the two servers, with a noticeable absence of lower-income countries in this list.

The Chinese province with the first known cases, Hubei province, produced 261 preprints by May 2020, mostly in February and March 2020. The rate of preprint production is correlated with the national COVID-19 caseload (Figure 1b). Fraser et al (2020) document that the date of the first preprint posting from authors from a given country on the bioRxiv.org and medRxiv.org sites is correlated with the date of the first confirmed case in a country. In the two major contributing countries to preprints, China and the United States, we observe some interesting trends. As cases grow exponentially in the United States, as does the number of preprints from United States based authors, whereas in China, the number of cases flattens in the second quarter of 2020, and the number of preprints emanating from China actually declines (Figure 1c). This dominance of China on the preprint platforms in the early months of 2020 is in stark contrast to their relative absence just prior to the crisis (Table 1).

Table 2 provides some descriptive statistics about the COVID-19 preprints. Preprints have an average of eight authors, with a maximum of one hundred and seventy-eight. Around thirty percent of preprints have an international team and thirty four percent of preprints have authors from the United States. Figure 2 illustrates the lifecycle of downloads and tweets following posting on the preprint server. In general, attention measures tend to be greatest in the month of posting, rapidly declining following the first month. The rate of tweets to a given preprint declines even more rapidly than that of downloads throughout a preprint's lifetime.

4. Results

a. Empirical strategy

As a preliminary step, we analyze whether a preprints' authors' institutional affiliation influences the rate of attention to the preprint. Specifically, we measure whether there is a relationship between preprint downloads and tweets and the location of the authors' institution. We use the following general empirical framework to assess this relationship (equation 3).

$$Y_{ijt} = \beta_0 + \beta_1 China_i + \beta_2 USA_i + X_{it}\partial + \varepsilon_{ijt} \quad [3]$$

Where Y_{ijt} is the number of downloads, tweets, or tweets from a highly followed user per paper i published in country j in time t, and $China_i$ is a dummy that takes the value of 1 if preprint i has any author affiliated with a Chinese institution, and USA_i is a dummy that takes the value of 1 if preprint I has any author affiliated with a USA institution. We control for the maximum research ranking of paper i author(s)' institutions prior to the crisis, and $X_{it,}$ a set of preprint-specific control variables reflecting the life cycle of preprints and the pandemic as well and variables representing the cumulative number of COVID-19 cases in the preprint authors' location at the time of research, scientific field, geographic origin and other features of the preprint.

We measure changes in attention to a focal preprint following twitter endorsements described above, allowing us to account for any underlying heterogeneity in the preprints themselves (equation 4).

$$Y_{ijt} = \beta_0 + \beta_1 AfterEvent_{it} * Treated_i + \beta_2 AfterEvent_{it} * Rank_i + \partial_\tau + \gamma_i + \varepsilon_{ijt} \quad [4]$$

Where $Treated_i$ = variable that takes the value of 1 if a preprint is associated with a tweet from a highly followed user, and $AfterEvent_{it}$ = variable that takes the value of 1 if the focal measurement time is after an event. We include calendar day and individual preprint fixed effects consistent with our approach to analyze changes in the attention rates following events.

The majority of the dependent variables of interest are skewed and non-negative (Figure 3 illustrates the distribution of pdf downloads across all COVID-19 preprints). Due to the large skew in outcomes, and following tradition in the study of scientific and technical change, we present estimates based on ordinary least squares models with inverse hyperbolic sine transformation of the dependent variables. Standard errors are clustered at the preprint level (Abadie et al 2017).

b. Findings

In Table 3 we explore the role of geographic affiliation of preprint authors in attention. Specifically, we ask whether a USA, or China affiliation results in more or less attention than average. We find that in general attention is greater for preprints with USA authors, and lower for preprints with China-based authors (column 1). In terms of the magnitude of these differences, column 1 implies that Chinese authored preprints receive 12% fewer downloads per month as compared to other preprints. This corresponds to over 100 fewer downloads per month, or 500 in the 5 months following posting of the preprint.

In the event that attention is driven by access to early cases in the first few months of the pandemic, we include a control for whether the authors of the preprint are affiliated with an institution in Wuhan, China (column 2). Although the positive coefficient on Wuhan suggests that signaling access to early cases affects attention, the negative coefficient on Chinese authors remains negative, and actually becomes more negative, suggesting that outside of Wuhan, Chinese authors receive substantially less attention to their COVID-19 preprints than do authors within Wuhan. Interestingly, we see no association between Wuhan authors and downloads where 'Wuhan' is not specifically named in the affiliation field of the article (column 3). This provides further support of our hypothesis that labels of institution location on an articles act as a signal to readers.

We next attempt to account for underlying quality of the work. Readers may be less likely to download a Chinese authored preprint than another preprint because they are lower quality, rather than because of their location or reputation. We present a large amount of the information available to readers, and other variables that may affect the actual quality of a preprint, in order to reduce concerns of omitted-variables. First, we control for author location COVID-19 cases in column 4, which we argue is a shifter in the cost of doing high quality research through enabling access to patients and samples. Similar to the Wuhan dummy, this is a generalized version. We find that an author's location percentage of global COVID-19 cases is significantly associated with pdf downloads of an article, but that the negative coefficient on Chinese author, after accounting for whether the author is in Wuhan or not, remains relatively unaffected. This suggests that any increased attention to Chinese preprints resulting from earlier access to cases is concentrated in attention to Wuhan preprints. Or in other words – it is only the very early cases that matter for attention for Chinese authored preprints.

Second, we include controls in the regression framework for the preprint author institution rank, if we consider that authors at higher ranked institutions are either more able to produce high quality research, or have more access to resources. We find that downloads are strongly correlated with the rank of the authors (column 5). And the negative association between Chinese author and downloads becomes significantly more negative after accounting for the rank of the institution. This suggests that there are relatively more higher ranked institutions in preprints authored by Chinese scientists as compared to the rest of the world, and that on average preprints coming from these higher ranked institutions receive more attention.

Third, we control for whether the data is made publicly available in column 6. We consider the former measure a signal of transparency, reducing uncertainty about the quality of the product.

Fourth, we include a control variable for whether the preprint has been published in a peer reviewed journal at the time of writing, and if so, the source normalized journal impact factor of the ultimate publication outlet. Although there is variation within journals on quality, this provides a coarse quality measure of the article. As revealed in column 7, downloads are highly correlated with the ultimate publication outcome, but the coefficient on Chinese authors remains negative to this control for underlying quality.

We restrict the sample in column 8 to just preprints that appear in peer-reviewed journals at the time of writing, and control for the impact factor of the journal again, reporting that for this subset of preprints, there is no observable negative relationship between downloads and Chinese authors. This implies that the negative relationship between Chinese authors and downloads of a preprint is driven by the set of preprints that did not make it into peer reviewed journals at the time of writing, either because they are lower quality or because for these preprints the publication process is longer.

Finally, we measure the relationship between abstract downloads of the preprint and author characteristics in columns 9 and 10. In order to download the abstract on the preprint platforms, readers see the title and the names of the authors of the preprint on the main page in a list of preprints, but the author affiliation is not available. The relationship between Chinese authors and abstract downloads is less negative than that for PDF downloads, suggesting that while author names and title alone does drive differences in attention, the author affiliations as well as the abstract text has an additional impact on the rate of downloads above the author names and title. Interestingly, there is little difference between the coefficients on Chinese author in columns 9 and 10, which implies that the author institution rank has a limited effect on the rate of abstract downloads. This suggests that the effect that author institution rank above operates via either a signalling mechanism, or via the quality of the abstract, driving audiences to download the full text.

Passage of time. As more information is made available throughout a preprint's lifetime, such as comments, new information and data, and news reports, decision makers can update their beliefs about the quality of work. This kind of updating is likely to be heterogeneous by preprint type.

Figure 1d presents the difference in downloads over time between preprints with different categories of authors, holding preprint field and month of posting constant for the subset of preprints produced in January through March 2020. These preprints are likely to have the most ex-ante uncertainty surrounding them as they were posted at the start of the epidemic when there was very little known about the virus and transmission itself. When compared to the role of the signal in the first month of posting, institutional rank and USA authors appear to drive attention less over time, while the relationship between preprints having Chinese (and Wuhan) based authors and attention becomes more positive over time. This results in a narrowing of the gap in attention between high status and low status authored preprints. In contrast, the relationship between author location COVID-19 cases and attention remains constant throughout a preprint's lifetime. This suggests that audiences take longer to assess the quality of the Chinese preprints, but that attention to these articles eventually catches up. One American-based doctor confirms this updating in the perception of Chinese based scientists during the pandemic, saying: "*We had a talk from a doctor in Wuhan through zoom at the start of the pandemic… it changed my impression of Chinese doctors and research*".

Endorsements. We investigate the role of twitter mentions as a self-organizing screening mechanism used to help communities reduce uncertainty about preprint quality. We consider twitter mentions of a preprint as 'endorsements', the magnitude of which varies with the number of followers of the tweeter. First we explore which preprints are more likely to be endorsed in Table 4. Similar to the pattern for downloads or attention, USA preprints are more likely to be tweeted about and Chinese preprints less so, although the magnitude of the difference is greater in tweets than in downloads. This lower rate of tweets to Chinese authored preprints could be due to lower use of twitter by Chinese researchers (Sugimoto et al 2017), who are more likely to be within the same social network as each other, or a hesitation to publicly endorse Chinese preprints. Similarly, we examine which preprints are tweeted by users with more than 10,000 followers (top 95th percentile of tweeters in our sample). Compared to the non-endorsed preprints, 'endorsed' preprints are most likely to be authored by USA based, and less likely to be authored by Chinese based authors.

We report the results of the difference-in-differences specification in Table 5. We compare any change in the rate of downloads and tweets of a focal preprint following a tweet mention by a tweeter with more than 10,000 followers of that preprint to that of comparable preprints that were

not mentioned in a tweet by a tweeter with more than 10,000 followers. We identify similar preprints before the tweet event using a coarsened exact matching procedure that ensures that the rate of daily tweets (which also results in comparable rates of daily downloads) leading up the tweet event is comparable. In Tables 5, column 1, the coefficient estimate implies that following a tweet of a preprint by a tweeter with more than 10,000 followers, the rate of daily downloads increases by 23% relative to the trajectory of daily downloads of preprints not tweeted by a high-profile tweeter. Similarly, the rate of daily tweets increases by around 65% relative to the trajectory of daily tweets of preprints not tweeted by a high-profile tweeter.

Figure 4a shows the dynamic version of the model, for all preprints and for separate categories of authors. We interact the treatment variable with indicator variables for the number of days before and after the preprints earliest tweet between June 5 and July 5. We graph the estimates along with 95% confidence intervals. We don't see any evidence that preprints receive more downloads in the period running up to a high-profile tweet (Figure 4a), but afterwards we see an initial jump in downloads followed by a leveling off suggesting that the impact is persistent.

We explore any heterogeneity in the impact of a tweet event by preprint author location. We find that the impact of a tweet on daily downloads is statistically similar for preprints Chinese authors as the average impact, but smaller for those with USA authors (although this smaller effect disappears when any differential effects according to the rank of the authors institution is accounted for). In contrast, the impact on daily tweets is much less for Chinese authors than for other preprints. These findings are consistent with the Bayesian framework described in Section 2. When a signal is very informative, it has a large impact on attention unless the levels of precision of the prior are greater than the prior beliefs about the quality of a preprint, or in instances when the expected quality even after an endorsement don't meet a threshold for action. We assume that readers have a strong prior about USA authored preprints which explains the smaller effect for those preprints on downloads, and for Chinese authored preprints there is a tension in terms of the possible opposing effects of an endorsement. On the one hand, the priors about Chinese authored preprints are greater and so should result in a larger impact of an endorsement, but on the other hand, they still may not reach the threshold for attention resulting in a smaller impact. These opposing effects cancel each other out to result in Chinese authored preprints having a similar sized effect from endorsements in downloads as other preprints.

In terms of the impact of an endorsement on tweets, the threshold for tweeting a Chinese preprint is greater than for downloading, consistent with the finding that it is this negative effect of Chinese authors that dominates the positive effect in the tweet outcomes, resulting in a smaller overall effect of an endorsement on tweets for Chinese authored preprints. In contrast, the threshold for tweeting USA authored preprints can be considered lower, explaining the relatively large effect of an endorsement on daily tweets for USA authored preprints. This difference between the rate of increase of tweets versus downloads for Chinese authored preprints is consistent with the idea that a co-ordination mechanism is one mechanism driving the baseline penalty for Chinese preprints. Tweets are visible to the community, and as Keynes (1936:158) put it, "*Worldly wisdom teaches us that it is better to fail conventionally than succeed unconventionally.*"

Reassuringly, the baseline effect of an endorsement is not found for tweets by tweeters with fewer than 1000 followers (Table 6 column 1). This is consistent with the predictions of the theoretical model that the status of the endorser also matters. However, one finding worth pointing out is that a tweet from a tweeter with less than 500 followers actually does have a positive impact on downloads for preprints with authors from top 50 ranked institutions, USA institutions, and on subsequent rate of tweets. It is helpful to put these findings into the context of the theoretical model, which suggests that the perceived quality of preprints with authors from USA institutions are closer to the threshold for downloading then other preprints. Therefore, even a small endorsement could push an audience member to the point that the threshold is met for action (downloading the PDF).

Magnitude of the Bias. While the previous analysis identified the role of author location in knowledge diffusion, whether some types of preprints are receiving more attention than they 'should' given underlying quality is an empirical question. In order to estimate whether preprints receive more or less attention than their underlying quality would predict, we compare the downloads and tweets to the ultimate publication outcome, using the journal impact factor as a proxy for quality of the preprint.

In order to assess whether preprints authored by different types of researchers are over/under downloaded or tweeted relative to expectations given their publication outlet, we generate a predicted line of attention measures of a preprint from the journal impact factor of the publication outlet as well as broad field of the preprint and age of the preprint using the pre-COVID-19 preprints. We then regress the residual of this line for each COVID-19 preprint found in a peer-

reviewed outlet on measures of status of the authors of the preprints and present the results in Table 7. It is important to note that this is a subset of preprints that have appeared in peer reviewed journals at the time of writing, and interestingly this subset of Chinese preprints experience no download penalty. This could be for a few reasons: the very best Chinese preprints (coming out of the best institutions) could also be the ones pushed forward into peer reviewed journals, and there could be less uncertainty about these ex ante, or the earliest preprints (that are most likely to be published at time of writing) could have received less download penalty as they were the very first pieces of knowledge on the virus. Either way, despite no download penalty there is a twitter penalty and we explore the extent to which these preprints are downloaded and tweeted about more or less than would be expected, given their ultimate publication outcome.

The statistically significant coefficients on the institutional rank variable suggest that a preprint's author institutional rank leads to a higher-than-expected-rate of download and tweets (albeit slightly less so), as compared to what would be expected given eventual publication outlet in the pre-COVID period. Preprints with Wuhan authors are downloaded much more than expected, while preprints with Chinese authors are tweeted much less than expected. The relationship between the residual and downloads is not particularly informative if there is a relationship between downloads and publication outcome (for example, if peer reviewers look more favorably upon articles that they are already familiar with), but the variation in the relationship between the residual and attention measures by preprint author status is slightly concerning. If preprint platforms and social networks are increasingly prevalent platforms of knowledge distribution, these results suggest that decision makers could increasingly be unaware of knowledge coming out of less distinguished institutions, regardless of quality.

5. Discussion

The rising use of platforms designed to disseminate early research findings raises the question of how decision makers allocate attention. Without conventional quality stamps that the peer review process provides decision makers the community undertakes self-organizing screening mechanisms. This study explores the relationships between author location and the diffusion of new knowledge on new platforms, and following self-organizing screening mechanisms. Measuring rates of attention (downloads) to preprints in the context of COVID-19 pandemic preprints we find that the geographic affiliation of preprints authors is a determinant of attention.

Specifically, we find that preprints with Chinese based authors tend to receive lower attention than authors from the rest of the world, even once accounting for proximity of Chinese scientists to early COVID-19 cases, which drive attention. In fact, there is a noticeable 'Wuhan effect' (the earliest region located in China with COVID-19 cases) which drives attention, but even after accounting for the positive attention that Wuhan authored preprints receive, we still notice a bias against Chinese authored preprints.

We measure any changes in attention to a given preprint following endorsements on twitter by a highly followed individual. We find that Chinese based authors are less likely to be tweeted about than their peers around the world, even after accounting for quality of the preprint, and find that endorsements do drive attention, albeit no more positively for Chinese authored preprints than other preprints. Given that Chinese authored preprints are subject to lower attention in the first place, this suggests that endorsements via twitter can replicate the attention gap. Our results are consistent with Jin et al (2019) in supporting Merton's (1968) proposition that the "rich" have an advantage over the relatively "poor" in light of new information and that this can lead to persistent cumulative advantage. The type of uncertainty or new information seems important in driving wedges in the status distribution, and future research should explore how features of the endorser or endorsement create heterogeneity in the effect.

We consider the findings to also have practical implications for decision makers in a time of extreme uncertainty. As Carley et al 2010 state: "The urgency and severity of the COVID-19 pandemic contains threats and opportunities to clinicians wishing to practice EBM (Evidence Based Medicine)", decision makers struggle to allocate attention and assess the evidence in a timely manner. Given the observation that much of the attention in the early months of the COVID-19 pandemic is directed at the work of higher status authors, a cautionary tale is offered that preprint platforms and social networks such as twitter have a limited effect in levelling the playing field for global scientists.

As we enter into different paradigms of uncertainty (future global pandemics, consequences of climate change) whereby the latest findings from scientists are critical, it is important to understand how science is communicated, and which kinds of information decision makers give their attention to. Decision makers cannot act on information they have not seen, and ultimately the drivers of attention to new scientific knowledge have consequences for the rate and direction of innovation, health and prosperity outcomes, and long-run economic growth.

References

Abadie A, et al (2017) When Should You Adjust Standard Errors for Clustering? NBER Working Paper No. w24003.

Azoulay, P., Stuart, T., Wang, Y (2014) "Matthew: Effect or Fable?," *Management Science* 60(1): 92-109.

Azoulay, P., et al. (2015) "Retractions." *Review of Economics and Statistics* 97.5: 1118-1136.

Azoulay, Pierre, Alessandro Bonatti, and Joshua L. Krieger. "The career effects of scandal: Evidence from scientific retractions." *Research Policy* 46.9 (2017): 1552-1569.

Bikard, Michaël. "Made in academia: The effect of institutional origin on inventors' attention to science." *Organization Science* 29.5 (2018): 818-836.

Blau (1964) Exchange and Power in Social Life
New York: Wiley

Ceci, Stephen J., and Douglas P. Peters. "Peer review: A study of reliability." *Change: The Magazine of Higher Learning* 14.6 (1982): 44-48.

Clauset A, Arbesman S, Larremore DB (2015) "Systematic Inequality and Hierarchy in Faculty Hiring Networks" Science Advances 1(1): e1400005.

Correll, Shelley J., et al. "It's the conventional thought that counts: How third-order inference produces status advantage." *American Sociological Review* 82.2 (2017): 297-327.

Gaule, P, Maystre N. "Getting cited: Does open access help?." *Research Policy* 40.10 (2011): 1332-1338.

De Groot 1970 Optimal statistical decisions New York, McGraw-Hill
Hicks, John R. "Keynes' theory of employment." *The Economic Journal* 46.182 (1936): 238-253.

Fraser et al (2020) "Preprinting a pandemic: the role of preprints in the COVID-19 pandemic". bioRxiv preprint doi https://doi.org/10.1101/2020.05.22.111294

Freedman LP, Cockburn IM, Simcoe TS (2015) The economics of reproducibility in preclinical research. PLoS Biol. 13(6):e1002165

Freeman, Richard B. and Wei Huang, (2015) "China's 'Great Leap Forward' in Science and Engineering." In *Global Mobility of Research Scientists*, edited by Aldo Geuna,, Academic Press, San Diego, Pages 155-175, https://doi.org/10.1016/B978-0-12-801396-0.00006-5.

Fry, CV et al (2020) "Consolidation in a Crisis: Patterns of International Collaboration in COVID-19 Research" PLoS ONE 15(7): e0236307. https://doi.org/10.1371/journal.pone.0236307

Fry CV (2020) “Viral Privilege: Evidence from the Ebola Epidemic” Working paper

Hill, R and Stein, C (2020) Scooped! Estimating Rewards for Priority in Science. Working paper

Holmgren M, Schnitzer SA (2004) Science on the Rise in Developing Countries. PLoS Biol 2(1): e1. https://doi.org/10.1371/journal.pbio.0020001

Hvistendahl, M (2013) China’s Publication Bazaar. Science 342(6162): 1035-1039

Hvistendahl, M (2020) The Scientist and the Spy: A true story of China, the FBI, and Industrial Espionage. Riverhead

Jin, Ginger Zhe, Benjamin Jones, Susan Feng Lu, and Brian Uzzi (2019), “The Reverse Matthew Effect: Consequences of Retraction in Scientific Teams,” *The Review of Economics and Statistics* 2019 101:3, 492-506

Jones B (2009) The burden of knowledge and the ‘death of the Renaissance man’: Is innovation getting harder? Rev. Econom. Stud.76(1):283–317.

Kim, Jerry W., and Brayden G. King. "Seeing stars: Matthew effects and status bias in major league baseball umpiring." *Management Science* 60.11 (2014): 2619-2644.

Kovács, Balázs, and Amanda J. Sharkey. "The paradox of publicity: How awards can negatively affect the evaluation of quality." *Administrative science quarterly* 59.1 (2014): 1-33.

Kupferschmidt, K., “Preprints bring ‘firehose’ of outbreak data,” *Science*, February 28, 2020

Lee, Carole J., et al. "Bias in peer review." *Journal of the American Society for Information Science and Technology* 64.1 (2013): 2-17.

Leydesdorff L, Wagner CS (2008) “International Collaboration in Science and the Formation of a Core Group” Journal of Informetrics 2(4): 317-325.

Long, J. Scott, Paul D. Allison, and Robert McGinnis. "Entrance into the academic career." *American sociological review* (1979): 816-830.

Mallapaty, Smriti (2020). “China bans cash rewards for publishing papers.” *Nature* 579, 18. doi: 10.1038/d41586-020-00574-8

McCabe, Mark J., and Christopher M. Snyder. "Open Access as a Crude Solution to a Hold-Up Problem in the Two-Sided Market for Academic Journals." *The Journal of Industrial Economics* 66.2 (2018): 301-349.

Merton, Robert K. (1968). "The Matthew Effect in Science". Science. 159 (3810): 56–63.

NSF science and engineering indicators 2020, last accessed on 8.15.20 https://ncses.nsf.gov/pubs/nsb20201/global-science-and-technology-capabilities

Peoples BK, Midway SR, Sackett D, Lynch A, Cooney PB (2016) Twitter Predicts Citation Rates of Ecological Research. PLoS ONE 11(11): e0166570. https://doi.org/10.1371/journal.pone.0166570

Peukert, Christian and Reimers, Imke, (2019) "Digital Disintermediation and Efficiency in the Market for Ideas". Available at SSRN: https://ssrn.com/abstract=3110105 or http://dx.doi.org/10.2139/ssrn.3110105

Podolny, J. 1994. Market uncertainty and the social character of eco-nomic exchange. Admin. Sci. Quart.39(3) 458–483

Podolny, Joel M., and Damon J. Phillips. "The dynamics of organizational status." *Industrial and Corporate Change* 5.2 (1996): 453-471.

Podolny, Joel M. "Networks as the pipes and prisms of the market." *American journal of sociology* 107.1 (2001): 33-60.

Polanyi M (1958) Personal Knowledge (Routledge, London),http://www.press.uchicago.edu/ucp/books/book/chicago/P/bo19722848.html.

Ross, Joseph S., et al. (2006) "Effect of blinded peer review on abstract acceptance." *Jama* 295.14: 1675-1680.

Simcoe, T. Waguespack, D.M. (2011) "Status, Quality and Attention: What's in a (Missing) Name?" *Management Science*, 57(2): 274-290.

Smith MJ, Weinberger C, Bruna EM, Allesina S (2014) The Scientific Impact of Nations: Journal Placement and Citation Performance. PLoS ONE 9(10): e109195. https://doi.org/10.1371/journal.pone.0109195

Stuart, Toby E., Ha Hoang, and Ralph C. Hybels. "Interorganizational endorsements and the performance of entrepreneurial ventures." *Administrative science quarterly* 44.2 (1999): 315-349.

Sugimoto, Cassidy R., et al. (2017) "Scholarly use of social media and altmetrics: A review of the literature." *Journal of the Association for Information Science and Technology* 68.9: 2037-2062.

Tomkins, Andrew, Min Zhang, and William D. Heavlin. "Reviewer bias in single-versus double-blind peer review." *Proceedings of the National Academy of Sciences* 114.48 (2017): 12708-12713.

Wagner CS, Leydesdorff L (2005) “Network Structure, Self-organization, and the Growth of International Collaboration in Science” Research Policy 34(10): 1608-1618.

Waldfogel, J (2017) “How Digitization Has Created a Golden Age of Music, Movies, Books, and Television,” *Journal of Economic Perspectives*, VOL. 31, NO. 3, SUMMER 2017 (pp. 195-214)

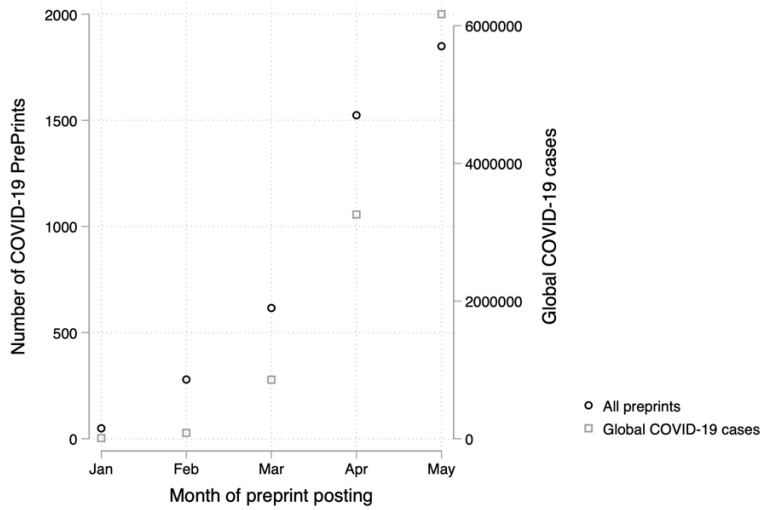
Xie Q, Freeman RS (2019) “Bigger Than You Thought: China’s contribution to scientific publications and its impact on the global economy *China & World Economy* 27(1) 1-27

Zelnio R, (2012) “Identifying the Global Core-Periphery Structure of Science” Scientometrics 91(2): 601-615.

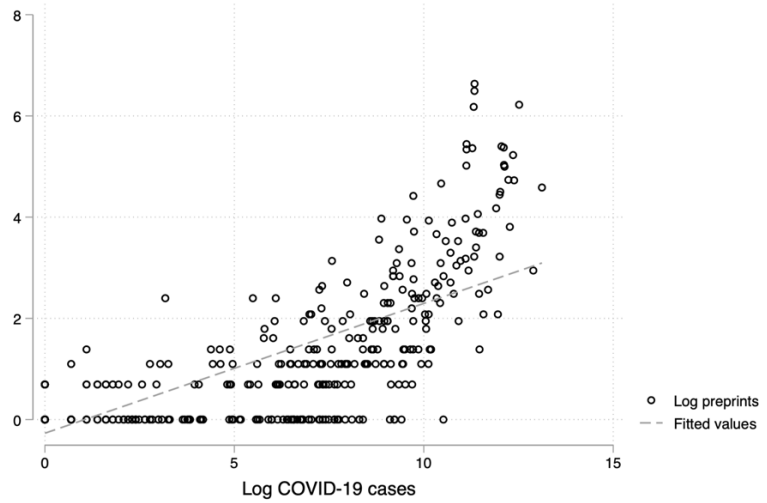
Zucker LG, Darby MR (1996) Star scientists and institutional trans-formation: Patterns of invention and innovation in the formation of the biotechnology industry. Proc. Natl. Acad. Sci. USA93(23):12709–12716

Figures & Tables

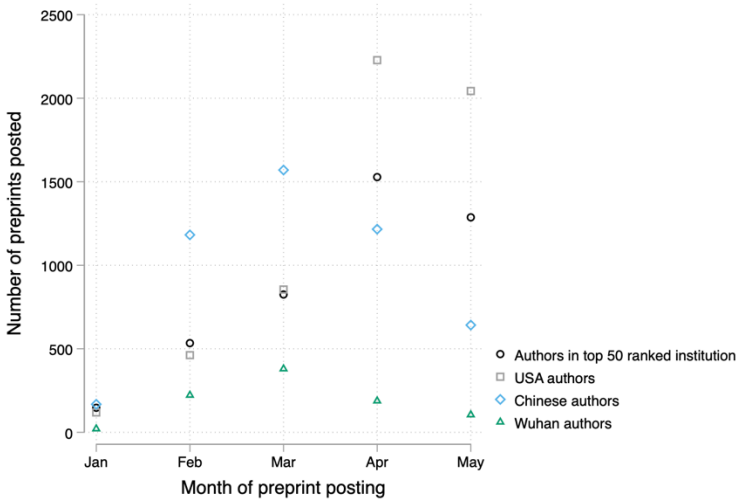
Figure 1. COVID-19 preprints posted on bioRxiv.org and medRxiv.org



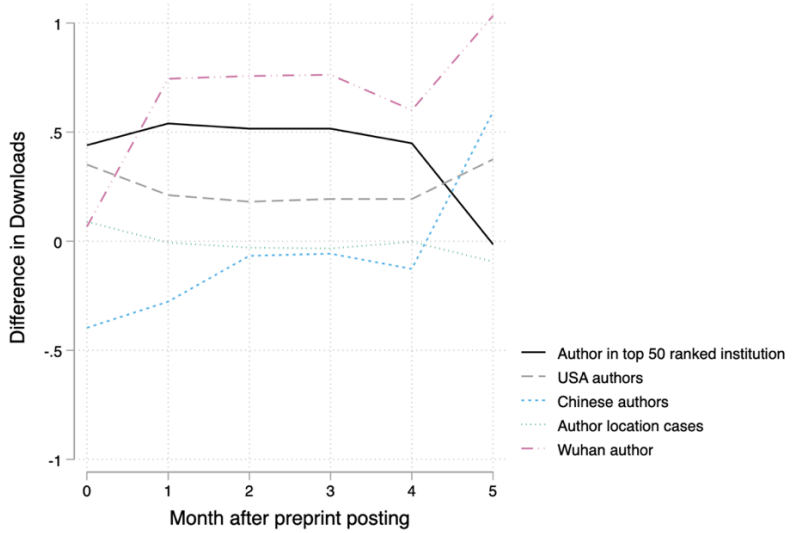
Panel A. COVID-19 preprints and global COVID-19 cases



Panel B. COVID-19 preprints and author country COVID-19 cases



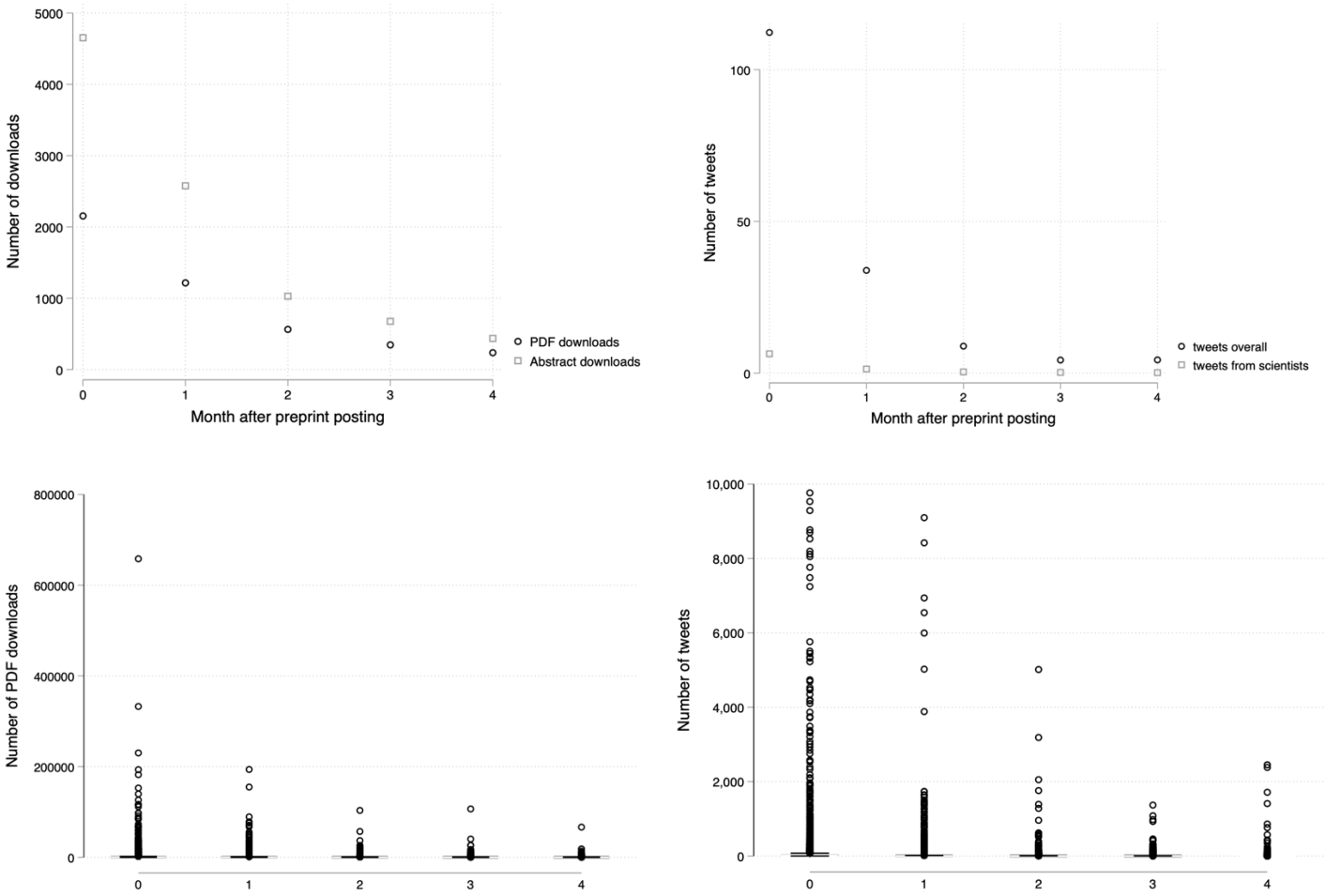
Panel C. COVID-19 preprints by type of author



Panel D. Changes in attention to COVID-19 preprint by author type

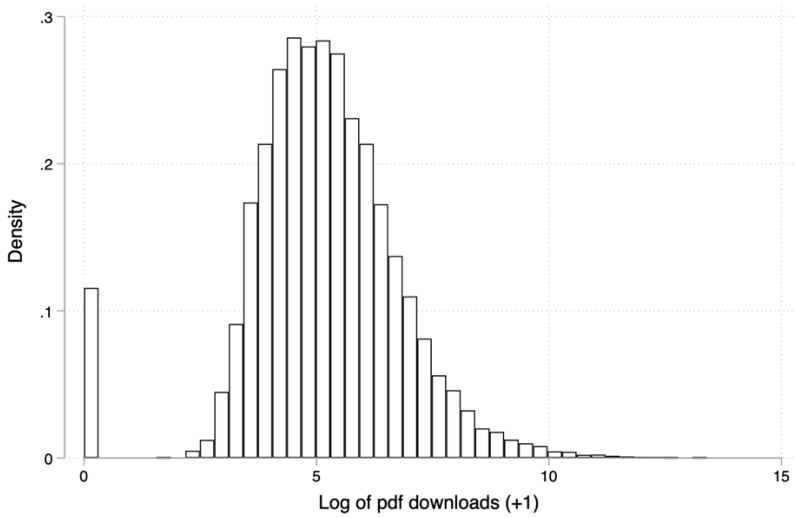
Note:
Panel A. We compute the total number of COVID-19 preprints published in each month in 2020 on the left-hand y-axis, and the global cumulative cases of COVID-19 at the end of each month on the right-hand y-axis.
Panel B. We compute the log of the cumulative number of COVID-19 related preprints at the country level produced by authors affiliated with the country (including just countries with any preprint in the time period) by the last day of each month in January 13-May 31 2020, and plot against the log of the cumulative number of COVID-19 cases in the country on the last day of the month.
Panel C. We plot the number of preprints posted by different 'types' of authors in the early months of 2020.
Panel D. We plot the relative difference in pdf downloads in a given month after posting between preprints posted in Jan – Mar 2020 inclusive with different 'types' of authors, and that of baseline preprints (preprints published in the same month in the same field, but without high ranking, USA, or Chinese authors).

Figure 2. Average rate of attention per preprint, following posting of preprint



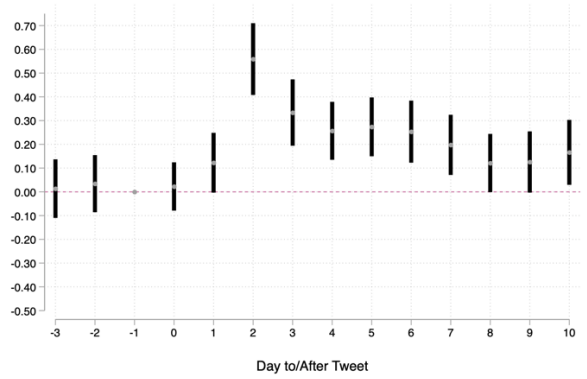
Note: We compute the average number of pdf and abstract downloads for each COVID-19 preprint each month after posting on the preprint server on the lefthand side, and the average number of total tweets, and tweets from scientists for each preprint each month after posting on the righthand side. The sample in the later months after posting is smaller due to the real time data collection.

Figure 3. Monthly downloads of preprints

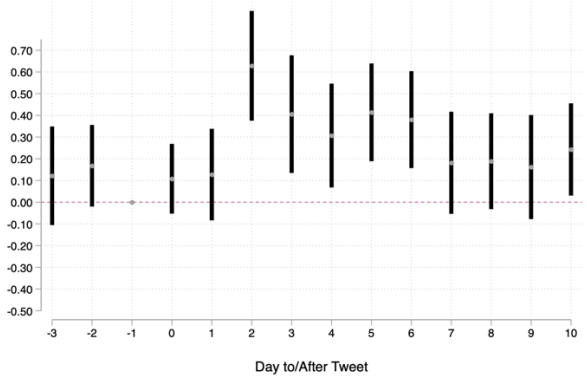


Note: We compute the log number of pdf downloads each month for each COVID-19 preprint.

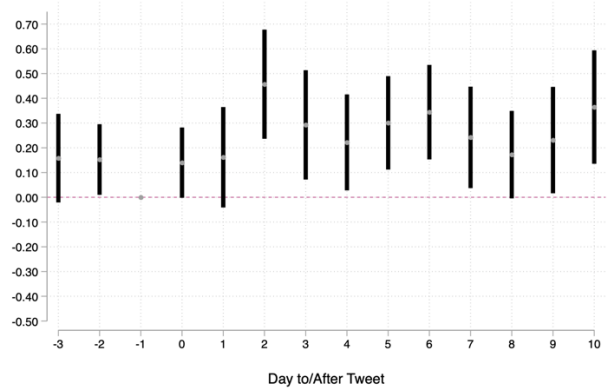
Figure 4. Event study diagrams of impact of tweets from highly followed tweeter on daily downloads



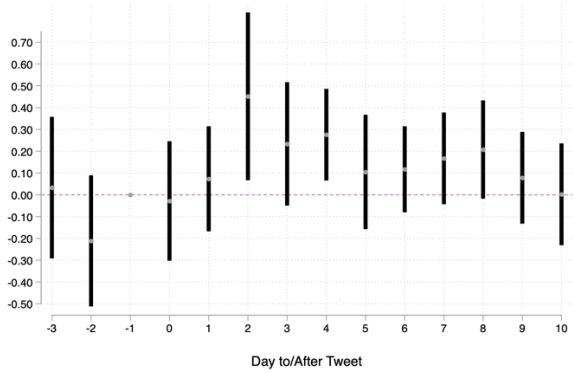
Panel A. All preprints



Panel B. Preprints with authors from top 50 ranked institutions



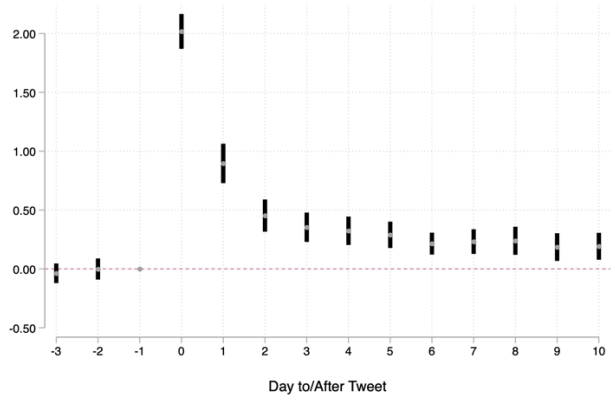
Panel C. USA authored preprints



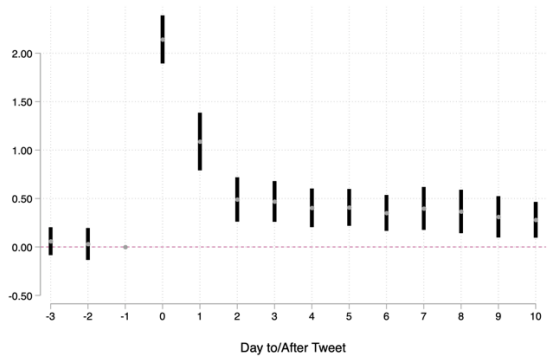
Panel D. Chinese authored preprints

Note. We plot the coefficient estimates stemming for conditional fixed effects ordinary least squares specifications in which inverse hyperbolic sine daily pdf downloads are regressed onto day fixed effects, individual preprint fixed effects, as well as interaction terms between treatment status and the number of days before/after the endorsement (the indicator variable for treatment status interacted with the day before the endorsement is omitted).

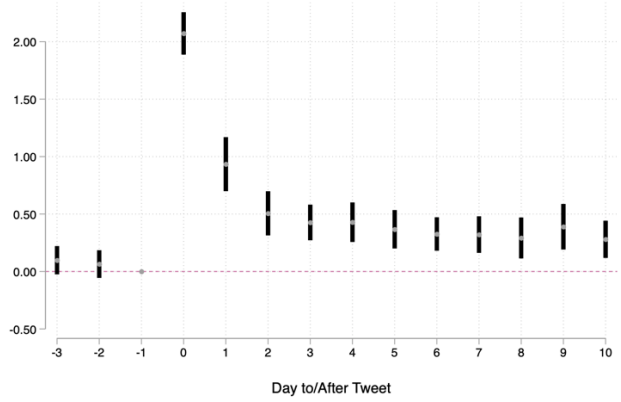
Figure 5. Event study diagrams of impact of tweets from highly followed tweeter on daily tweets



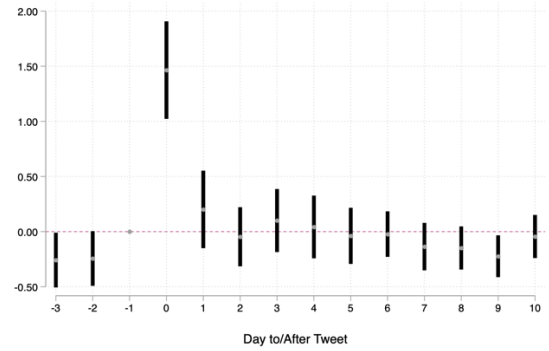
Panel A. All preprints



Panel B. Preprints with authors from top 50 ranked institutions



Panel C. USA authored preprints



Panel D. Chinese authored preprints

Note. We plot the coefficient estimates stemming for conditional fixed effects ordinary least squares specifications in which inverse hyperbolic sine daily tweets are regressed onto day fixed effects, individual preprint fixed effects, as well as interaction terms between treatment status and the number of days before/after the endorsement (the indicator variable for treatment status interacted with the day before the endorsement is omitted).

Table 1. COVID-19 preprint characteristics as compared to pre-COVID-19 preprint characteristics

	COVID-19 preprints (N=4,447)		Preprints prior to COVID-19 (N=10,637)	
Variable	**Mean**	**Std Dev**	**Mean**	**Std Dev**
Number of authors	8.566***	9.108	7.757	9.297
Chinese authors	0.237***	0.425	0.0807	0.272
USA authors	0.338	0.473	0.516***	0.500
International team	0.308	0.462	0.386***	0.487
Authors in top 50 ranked institutions	0.244	0.429	0.349***	0.477
Data made publicly available	0.609***	0.488	0.195	0.396
Biology	0.164	0.370	0.737***	0.439
Medicine	0.420***	0.494	0.125	0.330
Number pdf downloads per month	902***	7,146	35	1,196
Number abstract downloads per month	1,874 ***	12,657	202	4,333
Number pdf downloads in first month	2,153***	14,474	50	2,245
Number abstract downloads in first month	4,653***	21,842	652	9,582
Number tweets per month	32***	301	3	10
Number tweets from scientists per month	2***	11	2	7
Number tweets in first month	122***	666	12	18
Number tweets from scientists in first month	6***	22	8	12

Note: difference of means test compares mean values across COVID-19 preprints and preprints posted prior to COVID-19. *,**, *** represent significance at the 0.1, 0.05 and 0.01 level respectively.

Table 2. Descriptive statistics of 4,447 COVID-19 preprints

Variable	Mean	Median	Std Dev	Min	Max
Month posted	4.13	4	0.95	1	5
Day of month posted	17.39	18	8.49	1	31
Number of authors	8.57	6	9.11	1	178
Chinese authors	0.24	0	0.43	0	1
Hubei province authors	0.059	0	0.24	0	1
Wuhan city authors	0.045	0	0.21	0	1
USA authors	0.34	0	0.47	0	1
Italy authors	0.052	0	0.22	0	1
International team	0.31	0	0.46	0	1
Highest ranked institution of authors	492	212	904	1	6156
Author in top 50 ranked institutions	0.24	0	0.43	0	1
No author in ranked institution	0.12	0	0.33	0	1
Ranking of last author institution	768	369	1151	4	6156
Ranking of first author institution	779	368	1184	4	6156
Data made publicly available	0.61	1	0.49	0	1
Public funding[13]	0.25	0	0.43	0	1
Private funding	0.0036	0	0.060	0	1
Philanthropic funding	0.15	0	0.36	0	1
Biology	0.16	0	0.37	0	1
Medicine	0.42	0	0.49	0	1
Number pdf downloads per month	902	165	7,146	0	658,207
Number abstract downloads per month	1,874	367	12,657	0	743,364
Number pdf downloads in first month	2,153	421	14,474	0	658,207
Number abstract downloads in first month	4,653	1,213	21,734	0	574,400
Number pdf downloads per day (June-Aug 2020)	13	4	67	0	11,527
Number tweets per month	33	1	301	0	9,763
Number tweets from scientists per month	2	0	11	0	513
Number tweets in first month	122	11	666	0	10,000
Number tweets from scientists in first month	6	1	22	0	468
Number of followers of preprint tweeters	3903	509	36,427	0	9,645,715

Note: The full set of 4,447 preprints on COVID-19 related topics posted prior to May 31 2020 on the preprint servers bioRxiv.org and medRxiv.org is downloaded alongside relevant information on the preprint and the number of downloads and tweets per month.

[13] Funding data only available for the 3,589 articles posted on medrxiv

Table 3. Relationship between author location and PDF downloads of COVID-19 preprints

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Dependent variable	PDF downloads								Abstract downloads	
	Full sample							Preprints in peer-reviewed journals	Full sample	
USA author	0.30***	0.30***	0.30***	0.30***	0.15***	0.15***	0.16***	0.093	0.29***	0.14***
	(0.046)	(0.046)	(0.046)	(0.047)	(0.051)	(0.051)	(0.052)	(0.15)	(0.047)	(0.052)
Chinese author	-0.12*	-0.18***	-0.16***	-0.17**	-0.23***	-0.23***	-0.22***	0.023	-0.14**	-0.15**
	(0.063)	(0.067)	(0.066)	(0.067)	(0.067)	(0.067)	(0.068)	(0.18)	(0.069)	(0.069)
Wuhan author		0.35***	0.42***	0.18	0.23	0.23	0.22	0.46	0.49***	0.24
		(0.13)	(0.13)	(0.15)	(0.15)	(0.15)	(0.15)	(0.38)	(0.14)	(0.16)
Wuhan author without 'Wuhan' named in preprint affiliation details			-0.46*	-0.57**	-0.53**	-0.53**	-0.54*	0.16	-0.56*	-0.65**
			(0.26)	(0.26)	(0.26)	(0.26)	(0.26)	(0.45)	(0.29)	(0.28)
Author location % of global COVID-19 cases				0.82***	0.81***	0.81***	0.83***	-0.24		0.87***
				(0.28)	(0.29)	(0.29)	(0.29)	(0.67)		(0.30)
Author in top 50 ranked institution					0.44***	0.44***	0.45***	0.34**		0.38***
					(0.058)	(0.058)	(0.058)	(0.17)		(0.059)
Author in top 50-100 ranked institution					0.15*	0.15*	0.16*	0.16		0.16*
					(0.085)	(0.085)	(0.086)	(0.20)		(0.088)
Data publicly available						0.014	0.016	-0.084		0.041
						(0.048)	(0.049)	(0.14)		(0.049)
Published in peer reviewed journal							-0.12			-0.078
							(0.088)			(0.090)
Source Normalized Impact Factor (SNIP) of publication outcome							0.15***	0.15***		0.13***
							(0.028)	(0.029)		(0.029)
Mean of the dependent variable	902.88							1612.07	1874.22	
Nb preprint-month observations	22,235	22,235	22,235	22,235	22,235	22,235	22,235	3,450	22,235	22,235
Nb preprints	4,447	4,447	4,447	4,447	4,447	4,447	4,447	690	4,447	4,447

Note: Estimates stem from ordinary least squares models with outcome variables inverse hyperbolic sine transformed. All models include a full set of calendar month, and preprint age (in months) fixed effects, as well as a control for the source of the preprint (medRxiv.org/bioRxiv.org) and the scientific field of the preprint. Standard errors are clustered at the level of the preprint.

Table 4. Relationship between author location and tweets of COVID-19 preprints

Dependent variable	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	Number of tweets					Dummy if tweeted by user with >10,000 followers				
	Full sample				Preprints in peer-reviewed journals	Full sample				Preprints in peer-reviewed journals
USA author	0.27***	0.28***	0.13***	0.13***	0.083	0.027***	0.028***	0.010	0.0099	-0.021
	(0.034)	(0.034)	(0.037)	(0.036)	(0.095)	(0.0059)	(0.0059)	(0.0064)	(0.0064)	(0.016)
Chinese author	-0.16***	-0.20***	-0.28***	-0.25***	-0.19*	-0.025***	-0.033***	-0.042***	-0.039***	-0.022
	(0.039)	(0.041)	(0.040)	(0.040)	(0.11)	(0.0068)	(0.0070)	(0.0071)	(0.0071)	(0.020)
Wuhan author		0.25***	0.10	0.065	-0.33		0.045***	0.028	0.024	-0.013
		(0.093)	(0.098)	(0.097)	(0.23)		(0.016)	(0.018)	(0.017)	(0.038)
Author location % of global COVID-19 cases			0.59***	0.58***	0.76**			0.070**	0.068**	0.10*
			(0.19)	(0.18)	(0.35)			(0.034)	(0.032)	(0.061)
Author in top 50 ranked institution			0.36***	0.32***	0.19*			0.049***	0.044***	0.032*
			(0.044)	(0.044)	(0.11)			(0.0077)	(0.0076)	(0.019)
Author in top 50-100 ranked institution			0.10	0.090	-0.081			0.011	0.0097	-0.031
			(0.061)	(0.059)	(0.12)			(0.011)	(0.011)	(0.021)
Published in peer reviewed journal				-0.11**					-0.0058	
				(0.057)					(0.0098)	
Source Normalized Impact Factor (SNIP) of publication outcome				0.16***	0.16***				0.018***	0.019***
				(0.023)	(0.024)				(0.0033)	(0.0034)
Mean of dependent variable	32.74				60.23	0.22				0.28
Nb preprint-month observations	22,235	22,235	22,235	22,235	3,450	22,235	22,235	22,235	22,235	3,450
Nb preprints	4,447	4,447	4,447	4,447	690	4,447	4,447	4,447	4,447	690

Note: Estimates stem from ordinary least squares models with outcome variables inverse hyperbolic sine transformed in columns 1-5, and dummy outcomes in columns 6-10. All models include a full set of calendar month, and preprint age (in months) fixed effects, as

well as a control for the source of the preprint (medRxiv.org/bioRxiv.org) and the scientific field of the preprint. Standard errors are clustered at the level of the preprint.

Table 5. Effect of Highly Followed Tweet on Preprint Attention

Dependent variable	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	PDF downloads				Number of tweets			
After endorsement X endorsed	0.21***	0.27***	0.27***	0.25***	0.50***	0.54***	0.54***	0.50***
	(0.037)	(0.090)	(0.090)	(0.087)	(0.045)	(0.076)	(0.076)	(0.071)
After endorsement X endorsed X USA author		-0.13	-0.13	-0.17		0.0027	0.000015	-0.060
		(0.10)	(0.10)	(0.13)		(0.095)	(0.095)	(0.11)
After endorsement X endorsed X Chinese author		-0.035	0.026	-0.0061		-0.25**	-0.23**	-0.25**
		(0.15)	(0.18)	(0.18)		(0.098)	(0.10)	(0.11)
After endorsement X endorsed X Wuhan author			-0.19	0.083			-0.085	0.086
			(0.25)	(0.41)			(0.15)	(0.24)
After endorsement X endorsed X author location % of global COVID-19 cases				-0.54				-0.25
				(0.57)				(0.33)
After endorsement X endorsed X author in top 50 ranked institution				0.13				0.17*
				(0.13)				(0.10)
Nb preprint-day observations	1,749,650	1,749,650	1,749,650	1,749,650	1,749,650	1,749,650	1,749,650	1,749,650
Nb preprints	4,379	4,379	4,379	4,379	4,379	4,379	4,379	4,379

Note: The events studied are the first tweet between June 5-July 5 for a given preprint from a tweeter with more than 10,000 followers, giving 228 treated preprints, and 4,151 matched control preprints, used with replacement and weighted. Estimates stem from fixed effects ordinary least square specifications with dependent variables being inverse hyperbolic sine of counts of downloads or tweets per preprint per day for 3 days before and 10 days after the event (or counterfactual). All models incorporate a full suite of preprint age, calendar day and preprint fixed effects. Standard errors are clustered at the preprint level.

Table 6. Effect of Less Followed Tweet on Preprint Daily Downloads

Dependent variable	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	PDF downloads				Number of tweets			
After `mini' endorsement X endorsed	-0.0088	-0.054***	-0.054***	-0.064***	0.15***	0.16***	0.16***	0.16***
	(0.016)	(0.019)	(0.019)	(0.020)	(0.0049)	(0.0062)	(0.0062)	(0.0060)
After `mini' endorsement X endorsed X USA author		0.088***	0.088***	0.086***		-0.0041	-0.0041	-0.011
		(0.026)	(0.026)	(0.029)		(0.0077)	(0.0077)	(0.0090)
After `mini' endorsement X endorsed X Chinese author		0.074**	0.066*	0.054		-0.034***	-0.034***	-0.039***
		(0.032)	(0.035)	(0.034)		(0.0077)	(0.0080)	(0.0086)
After `mini' endorsement X endorsed X Wuhan author			0.049	0.020			0.00039	0.016
			(0.070)	(0.15)			(0.018)	(0.016)
After `mini' endorsement X endorsed X author location % of global COVID-19 cases				0.20				-0.064
				(0.15)				(0.045)
After `mini' endorsement X endorsed X author in top 50 ranked institution				0.012				0.022*
				(0.031)				(0.012)
Nb preprint-day observations	511,994	511,994	511,994	511,994	511,994	511,994	511,994	511,994
Nb preprints	3,286	3,286	3,286	3,286	3,286	3,286	3,286	3,286

Note: The events studied are the first tweet between June 5- July 5 for a given preprint from a user with less than 1000 followers (and no tweet from a user with more than 1000 followers in the same day, giving 2,168 treated preprints, and 1,149 matched control preprints, used with replacement and weighted. Estimates stem from fixed effects ordinary least square specifications with dependent variables being inverse hyperbolic sine of counts of downloads or tweets per preprint per day for 3 days before and 10 days after the event (or counterfactual). All models incorporate a full suite of preprint age, calendar day and preprint fixed effects. Standard errors are clustered at the preprint level.

Table 7. The relationship between publication outcomes and preprint attention as a function of author location

	(1)	(2)	(3)	(4)	(5)	(6)
Dependent variable	Residual of established pre-COVID relationship between SNIP of publication outcome and PDF downloads			Residual of established pre-COVID relationship between SNIP of publication outcome and number of tweets		
Author in top 50 ranked institutions	0.36***	0.32***	0.33***	0.33***	0.31***	0.30***
	(0.071)	(0.077)	(0.077)	(0.055)	(0.059)	(0.059)
USA authors		0.10	0.11		0.085	0.085
		(0.073)	(0.073)		(0.056)	(0.071)
Chinese authors		0.099	0.022		-0.21***	-0.22***
		(0.077)	(0.082)		(0.059)	(0.063)
Wuhan author			0.33**			0.20
			(0.13)			(0.10)
Mean of the dependent variable		1.30			0.69	
Nb preprint-month observations	3,450	3,450	3,450	3,450	3,450	3,450
Nb preprints	690	690	690	690	690	690

Note: Estimates stem from ordinary least squares models with outcome variables being the residual of the predicted line of the relationship between pdf downloads (columns 1-4) and tweets (columns 5-8) and source normalized impact per paper (SNIP) of the preprint ultimate publication. The predicted line is established by regressing the SNIP of the publication outcome of 633 pre-COVID control preprints that were identified in peer reviewed journals and their PDF downloads, including a full set of controls including preprint age, as well as a control for the source of the preprint (medRxiv.org/bioRxiv.org) and the scientific field of the preprint. The sample of preprints used is just those COVID-19 preprints that have appeared in a peer reviewed journal at the time of data collection.