Proceedings of the Fourteenth International AAAI Conference on Web and Social Media (ICWSM 2020)

Diffusion of Scientific Articles across Online Platforms

Igor Zakhlebin, Emőke-Ágnes Horvát
Northwestern University
2240 Campus Dr., Evanston, IL 60208
izakhlebin@u.northwestern.edu, a-horvat@northwestern.edu

Abstract

Online platforms have become the primary source of information about scientific advances for the wider public. As online dissemination of scientific findings increasingly influences personal decision-making and government action, there is a growing necessity and interest in studying how people disseminate research findings online *beyond* one individual platform. In this paper, we study the simultaneous diffusion of scientific articles across major online platforms based on 63 million mentions of about 7.2 million articles spanning a 7-year period. First, we find commonalities between people sharing science and other content such as news articles and memes. Specifically, we find recurring bursts in the coverage of individual articles with initial bursts co-occurring in time across platforms. This allows for a ranking of individual platforms based on the speed at which they pick up scientific information. Second, we explore specifics of sharing science. We reconstruct the likely underlying structure of information diffusion and investigate the transfer of information about scientific articles within and across different platforms. In particular, we *(i)* study the role of different users in the dissemination of information to better understand who are the prime sharers of knowledge, *(ii)* explore the propagation of articles between platforms, and *(iii)* analyze the structural virality of individual information cascades to place science sharing on the spectrum between pure broadcasting and peer-to-peer diffusion. Our work provides the broadest study to date about the sharing of science online and builds the basis for an informed model of the dynamics of research coverage across platforms.

Introduction

As scientific research increasingly shapes public discourse and impacts people's decision-making, researchers are more and more encouraged to communicate with the public (Peters et al. 2008). In recent years, social media and other online sources have come to dominate news consumption in general. In the U.S., more than 9 in 10 adults obtain news at least partially online (Pew Research Center 2018), with about two thirds getting news on social media (Shearer and Matsa 2018). Similarly, information about scientific advances is typically obtained from social media (Hargittai, Füchslin, and Schäfer 2018) and other online sources.

With social transmission, and particularly online sharing, becoming a key component of science communication, a suite of crucial questions emerge: *How is information about scientific advances shared by users of different online platforms? In what respects is the sharing of scientific findings similar to the diffusion of other types of content like news articles? What are the peculiarities of sharing scientific knowledge?* These questions are of key importance to both scientists and society at large. First, as the discourse about science is turning into an increasingly public topic, scientists rely on people's interest and taxpayer support to convey the value of and receive funding for their endeavors. Second, information overload resulting from the pace and scale of sharing and the use of multiple online dissemination channels make it especially hard for people to promote quality science.

This paper addresses important open questions about *(i)* who shares science, *(ii)* when, and *(iii)* on which online platforms. These basic questions have received little attention in the literature so far. Foundational work is based on surveys with scientists publishing in prime venues and describing their findings to nonscientists who then rate the likelihood of sharing them (Milkman and Berger 2014). This line of work has revealed the effect of content and linguistic style on shareability, and it has uncovered patterns in who is more likely to share scientific results. However, it remains unclear how people disseminate research findings in real settings given the complications introduced by the accelerating dynamics of collective attention (Lorenz-Spreen et al. 2019) and, most importantly, by how people navigate the ecosystem of online platforms.

Despite the ubiquity and importance of cross-platform sharing and consumption of information, they remain understudied even in the context of information diffusion more broadly. Existing literature has looked at diffusion within individual platforms like blogs (Adar and Adamic 2005), webpages (Ratkiewicz et al. 2010), Twitter (Lerman and Ghosh 2010; Vosoughi, Roy, and Aral 2018), Digg (Lerman and Ghosh 2010), Facebook (Cheng et al. 2014; 2016), and Wikipedia (Keegan, Gergle, and Contractor 2013). A few studies have looked into pairwise connections between the media, such as news outlets and blogs (Leskovec, Backstrom, and Kleinberg 2009), or news and Facebook (Tan, Friggeri, and Adamic 2016). To the best of our knowledge,

information diffusion between a larger number of platforms has not been considered in existing publications.

There are both experimental design and technical impediments to investigating cross-platform sharing activity. The bottleneck in terms of experimental design is to isolate a category of content that inherently spreads though a diverse set of online platforms. The technical challenge is then to track this cross-platform diffusion. Studying the online coverage of scientific articles successfully addresses both difficulties. On the one hand, focusing on scientific content assures a broad coverage, since science communication is prominent on most online platforms (Alt 2018) and users are, at least on social media, as likely to engage with scientific content as with "lighter" subjects such as entertainment (Hargittai, Füchslin, and Schäfer 2018). On the other hand, it is possible to unambiguously match posts about a scientific article through its publication URLs and Digital Object Identifier (DOI). It circumvents the need for using error-prone heuristics to connect the posts from different sources with widely varying representations, like grouping posts by potentially missing hyperlinks between them (Adar and Adamic 2005), hyperlinks in conjunction with text (Leskovec, Backstrom, and Kleinberg 2009), or their image and video content (Cheng et al. 2014; 2016).

We take two complementary perspectives to studying information diffusion between online platforms. First, we characterize dynamics of posting activity with respect to volume of posts over time, their organization in bursts (their typical size, recurrence and co-occurrence across platforms), as well as time differences between the platforms as they "catch up" on the stories. This exploration reveals several similarities between sharing scientific and other types of content. Second, we deduce the likely paths of information diffusion between the users of different platforms based on the time ordering of their posts (see Fig. 1). We analyze the structure of inferred networks to explore connections between platforms and the specific role that they play in the dissemination of scientific articles. The level of granularity of the structural analysis enables us to address questions that are specific to the dissemination of science within and across platforms: *(i)* Who are the main spreaders of scientific articles? *(ii)* What platforms are they using? *(iii)* At what rates does information propagate between platforms? and *(iv)* How does the structural virality of propagation trees vary by the platform sharing the article first?

Our paper makes three main contributions. First, we provide a first comprehensive temporal analysis of the diffusion of scientific articles online, which establishes similarities between the propagation patterns of science and more general news information. Second, our exploration of the likely structure of information diffusion characterizes the activity of users on various platforms, filling an important gap in existing literature with new knowledge about the extent to which different platforms share scientific articles. Third, our work is pioneering in cross-referencing the spread of information across different platforms, thereby informing substantively not only the domain of science communication, but also the broad study of information diffusion in an increasingly convoluted ecosystem of online platforms.

Related Work

Understanding the structure and dynamics of information diffusion has been a longstanding research problem in communication and media studies (Katz and Lazarsfeld 1955), marketing and management (Aral and Walker 2011), and information science (Szabo and Huberman 2010; Jamali and Rangwala 2009; Weng et al. 2012).

Mass communication. The structure of information diffusion has been in the purview of communication scholars for several decades. Based on initial observations of radio broadcasts and WWI propaganda, it was commonly believed that information spread by a mass medium directly reaches its audience's minds, akin to a "magic bullet" or a "hypodermic needle." Subsequent empirical research argued for a two-step flow of communication, where information originating in mass media is interpreted by opinion leaders that in turn influence their groups (Katz and Lazarsfeld 1955). Later research introduced additional steps into the model, defining a multi-step flow of communication (Weimann 1982).

Recent studies of Twitter provide support to models of both one-step, two-step, and multi-step flows of communication. In a 2011 study, 46% of content originated by "elite" Twitter users reached ordinary users through one or more intermediaries (Wu et al. 2011). A more nuanced analysis based on social movement data (Hilbert et al. 2017) demonstrated that aside of direct communication, there is also a two-step and multi-step flows present. In this paper, we assume that there is a network structure underlying the information diffusion, which allows us to capture all possible types of flow uniformly.

Information diffusion. Following the seminal paper about diffusion of memes between blogs by Adar and Adamic (Adar and Adamic 2005), research has focused on detecting, characterizing, and modeling information diffusion on various online platforms. The most prominent findings are: (*1*) the number of users individual messages reach follow fat-tailed distributions (Adar and Adamic 2005), (*2*) collective attention on various platforms is bursty in time (Lerman and Ghosh 2010; Ratkiewicz et al. 2010), and (*3*) that some bursts recur upon content resubmission (Lakkaraju, McAuley, and Leskovec 2013; Gilbert 2013; Cheng et al. 2016). These results have been demonstrated for particular individual platforms and it is not known whether they apply to a broad range of platforms.

Fewer papers have investigated pairwise connections between the platforms. In a study on meme tracking between news and blogs, authors found that online interest follows a "heartbeat"-like pattern and that news outlets tend to introduce the stories first, while blogs tend to catch them up later (Leskovec, Backstrom, and Kleinberg 2009). A study of news propagation through news outlets and Facebook has shown that posting activity on these two media unfolds in quick succession and that the entire news cycle typically lasts two days (Tan, Friggeri, and Adamic 2016). While providing important insights about information diffusion online, these studies have only considered a small subset of the biggest online platforms. This paper aims to cover a much broader portion of the Internet landscape.

Science communication online. While considerable re-

Table 1: Descriptive statistics of posts from different platforms covered by our data sample.

Platform	# Posts	# Users	Posts / User		Posts / Month
			50 pct.	99 pct.	
Twitter	52,678,741	4,708,238	1.0	149.0	
Facebook	4,054,074	768,330	1.0	64.0	
News	3,515,060	2,415	212.0	18,396.8	
Blogs	1,245,887	7,043	17.0	2,886.4	
Wikipedia	1,042,806	89,251	2.0	94.0	
Google+	750,217	160,323	1.0	52.0	
Reddit	138,723	30,299	1.0	40.0	
Total	63,425,508	5,765,899	1.0	138.0	

search has looked at how people use the Internet for sharing and engaging with various types of content from celebrity news to politics, very little of this work has considered how non-specialists interact online with science and research material (Bauer 2012; Brossard 2013; Hargittai, Füchslin, and Schäfer 2018). According to a recent report by the National Science Board, the Internet has become the most widely used source of science information among Americans (Beering and others 2014). Yet, literature on how science is presented online and how users interact with it is still budding; for a few exceptions, see (Brossard 2013; Scheufele 2013; Su et al. 2015). These studies highlight the importance of online platforms in accessing, interpreting, and discussing scientific material and call for more in-depth analyses of currently understudied areas of science communication.

Finally, there is a growing body of work that analyzes online communication about science in relation to particular events, e.g., conferences (Reinhardt et al. 2009) and focusing on, e.g., papers in specific journals (Robinson-Garcia et al. 2017). In this work, we attempt to reconstruct a maximally complete picture of information diffusion about scientific articles that is not confined to a particular platform or domain of knowledge.

Diffusion Modeling Framework

In this section, we present the used data and define notions to describe information diffusion in the rest of the paper.

Dataset Description

We have requested a recent data snapshot from Altmetric LLC[1], the largest service that tracks mentions of research outputs on the Internet. The particular advantages of using this data is its breadth, high granularity, and large time span (Alt 2018). It covers many types of research outputs such as journal articles, conference proceedings, book chapters, books, and reports. The wide range of Internet sources being tracked includes news outlets, blogs, Twitter, Facebook, Google+, Reddit, and Wikipedia. The data from these sources have been systematically collected since mid-2011.

We consider *users* of online media, i.e., individuals and organizations who make public *posts* talking about scientific advances. Due to the lack of a reliable way to determine the exact authorship of posts on news websites and blogs, we consider all news outlets and blogs individual "users" who represent their respective organizations. Twitter, Facebook, and Google+ users consist of a mix of individuals and organizations, while Reddit and Wikipedia only encompass individuals. In case of Wikipedia, the user of interest is the page editor who has first referenced the given scientific article.

We consider the users of different platforms as separate entities, despite the fact that, in reality, individuals and organizations might have accounts on multiple platforms. For example, journals like *Nature* own both a news outlet as well as profiles on Twitter, Facebook, and Google+. Matching reliably all such users across platforms would be difficult and would probably lead to loss of data. However, given that our method implicitly accounts for dependencies between users, as described in the "Inference of Diffusion Structure" section, if users consistently post the same content on different platforms in quick succession, we automatically connect them with edges in the information diffusion network.

When the user includes a link to the scientific article in their post, a *mention* is recorded. Our data includes only posts with such mentions. On platforms where users are allowed to repost publicly (Twitter, Facebook, and Google+), such reposts are also recorded. The data, however, does not include activity that happened "in response" to original posts, such as social media replies and reactions, unless such responses also include a direct link to the article. This selection criterion allows us to efficiently track how users mention the articles across different platforms.

We pre-processed the raw data dump from Altmetric and selected a sample that covers activity on 7 major online platforms over 7 years (June 10, 2011 – June 10, 2018). It included information about over 7.2 million research outputs that were mentioned at least once during that time period, about 63.5 million posts that mention them, and over 5.7 million users who wrote them. Of all research outputs, 75.1% were mentioned on a single platform (predominantly, Twitter), 17.6% on two platforms, and the remaining 7.3% on three different platforms or more. This means that we are able to study over half a million articles that were mentioned on three or more platforms.

Descriptive statistics of posts on different platforms and the users who wrote them are shown in Table 1. The largest number of posts across all platforms as well as the biggest number of unique users belong to Twitter, followed by Facebook. Next, according to the number of posts, are online news outlets and blogs. Throughout all platforms, we observe a large inequality in participation. As is the case with many Internet systems (Matei 2017), most users contribute just one post (except news and blogs, which are represented by organizations and not individuals), while a small fraction of them (on the order of 1%) contributes many dozens. With news outlets, there is a considerable reorganization in data provided by Altmetric, namely, around 2016 they have greatly expanded a list of news outlets they track.

Modeling Information Diffusion

Online activity around a scientific article typically unfolds in multiple stages. Typically, the article first appears online at its publisher's website. It can be automatically re-broadcast

[1] https://www.altmetric.com/research-access/

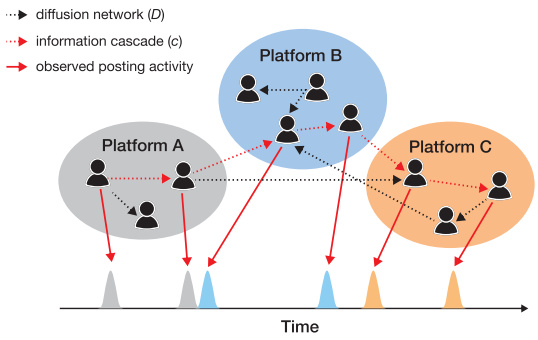

Figure 1: Information cascades propagate along the edges of unobserved diffusion network that includes users of different platforms. When information reaches individual users, they post on their respective platforms, generating sequences of posts that are ordered by publication time.

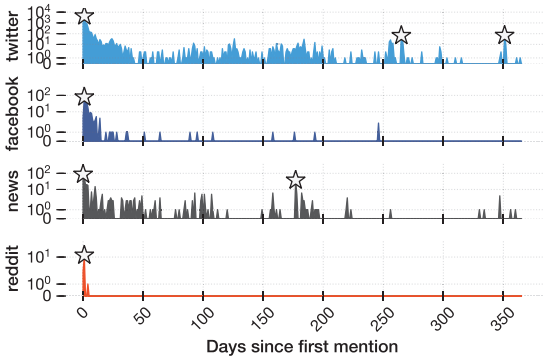

Figure 2: Mentions of the paper "United States Health Care Reform: Progress to Date and Next Steps" by Barack Obama on different online platforms one year since its publication date. Detected peaks of ten or more posts per day are marked with stars (☆).

by other websites that systematize papers from specific domains and provide a more convenient access to them (e.g., PubMed and F1000 Prime). These websites together provide a set of seed URLs that could be used to reference the article. Online users who have learned about the article directly can then use those URLs to link to the article in their posts. Over time, other users see these posts and learn about the article. Some of them would want to share it with their followers and they will post about it themselves. Yet others will see both old and new posts and repeat the process (cf. Fig. 1).

This process is commonly known as formation of an *information cascade* (Easley and Kleinberg 2010). In our case, public posts of some user can cause other users to post, which, in cascading fashion, can cause yet others to post. Although we don't know the exact structure of how the information spreads within such cascades, we can build heuristics: if user A posted before user B, it is likely that A has learned about the article before B and might have influenced B to post, while the contrary is unlikely. Thus, online posting activity about scientific articles can be used as a proxy for studying underlying information cascades.

Our model of information diffusion is based on the independent cascade model (Goldenberg, Libai, and Muller 2001) and network inference methods (Leskovec, Backstrom, and Kleinberg 2009). It makes the following choices. First, we assume that there is a static underlying *diffusion network* between the set of users V. This network spans all platforms and can be modeled as a directed graph $D = (V, E)$ that represents who exerts influence on whom, or, conversely, who seeks information from whom. Second, we assume that information cascades can only propagate along the edges of this diffusion network $c \in C : V_c \subset V, E_c \subset E$. When user u posts about an article, each user v influenced by u in D can post as well with some probability that depends only on the properties of u and v, and is independent of other nodes. Information cascades are thus formally forests, whose individual trees do not interfere with each other. Additionally, individual information cascades do not interact and propagate regardless of each other. Our third assumption is that when users propagate information, they link directly to the seed URLs. This modeling decision allows us to resolve the problem with some users linking directly to the scientific source (e.g., Wikipedia editors), while others use the built-in reposting capabilities of their platform (e.g., Twitter users). Reposts are thus considered independent new posts containing a link to the article.

This model is illustrated on Fig. 1. Users of various platforms propagate information about scientific articles amongst each other, forming information cascades related to individual articles. Upon receiving the new information, each user posts on their corresponding platform, together producing the time-ordered sequences of posts spanning different platforms. Our goal for the rest of the paper is to address the posed research questions based on this information.

Dynamics of Diffusion

To systematically characterize how scientific articles are talked about online, we first analyze when and on which platforms are they mentioned. Are they associated with similar dynamics of posting activity to other types of content like news articles, photos, or videos?

To illustrate our set-up throughout this investigation, we order posts about an article by time to form a histogram of how much it was mentioned on different platforms over time (see Fig. 2). This histogram demonstrates three common traits of information diffusion w.r.t. time: (*1*) posting activity tends to peak early on all media, (*2*) it is organized in sequences of bursts, and (*3*) initial bursts have a tendency to co-occur on different media within a couple of hours. These observations motivate the analyses to follow.

Temporal Variation

To understand how posting activity varies across time on different platforms, we look at the first year of activity on each platform for individual articles, counting from the time of their first mention on any platform. We estimate the proba-

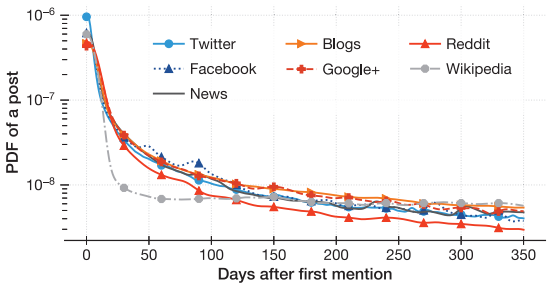

Figure 3: Probability of a post on each platform during the year after the first post about the article.

bility distribution of a post occurring during that time window with a Gaussian kernel density estimator with bandwidth computed by Scott's method (Scott 1992) as shown in Fig. 3. Across all platforms, most of the posting activity happens during the first three days after the first post, then it decreases until it reaches a low plateau. Although there may still be local peaks in posing activity for individual articles, they appear at a much lower rate giving birth to the said plateau. KL-divergence values between these probability distributions reveal an inherent organization of platforms into groups. The first group can be characterized as "breaking news" platforms, such as news outlets, Twitter, and Facebook (intra-group $D_{\mathrm{KL}} \leq 0.01$). The second group consists of blogging platforms, such as blogs, Google+, and Reddit ($D_{\mathrm{KL}} \leq 0.05$). Finally, Wikipedia forms the third group that is considerably different from the rest.

Burstiness

It has been shown before that posting of web links (URLs), news, photos, and videos on certain platforms is organized in bursts (Adar and Adamic 2005; Lerman and Ghosh 2010; Cheng et al. 2014; 2016). Here, we demonstrate that bursty behavior is typical for dissemination of scientific articles as well, across all considered platforms.

Temporal activity is considered to be bursty if inter-event times follow a fat-tailed distribution (Barabási 2005). To test for burstiness, we computed time intervals between posts for a year after the first mention of each article. We fitted the distribution of those intervals against a range of possibilities such as exponential, power-law, lognormal, and truncated power-law distributions (Clauset, Shalizi, and Newman 2009). In all cases, exponential and power-law distributions were rejected in favor of truncated power-law or lognormal distributions. For each platform, except Wikipedia, the truncated power-law represents a significantly better fit ($p < 10^{-4}$). The power-law exponent ranges from 1 for Facebook and blogs to 1.159 for Google+, 1.183 for Twitter, 1.206 for Reddit, and 1.34 for news sites. These values are consistent with previous measurements of other bursty online systems (Barabási 2005). In case of Wikipedia, the lognormal distribution represents a significantly better fit ($p < 10^{-6}$). When time differences are measured in seconds, parameter values for Wikipedia are $\mu = 16.584$, and $\sigma = 0.423$. These values are similar to the ones obtained for Digg website (Doerr, Blenn, and Van Mieghem 2013; Blenn and Van Mieghem 2016).

Recurrence of Bursts

To detect individual bursts, we use the method and parameter choices from (Cheng et al. 2016). Since the shapes of bursts can vary widely, the simplest way to identify them is by their *peaks*. We define each burst as a spike in the number of posts in a given day compared to the the days that surround it. Following (Cheng et al. 2016), we detect a burst whenever the daily number of posts is at least $h_0 = 10$, no less than twice the average number of daily posts, and is a local maximum within a window of $\pm w = 7$ days. Additionally, the number of posts per day between the two adjacent peaks should drop below the half of smaller peak's height.

Similar to (Cheng et al. 2016), we observe that bursts within each platform routinely occur more than once, i.e., they *recur*. The probability of recurrence given that the first burst has been observed is different for each platform: 14.3% for Twitter, 11.6% for Wikipedia, 5.9% for Facebook, 5.6% for news, 4.1% for Google+, and 0.88% for blogs. The overall number of bursts in different platforms are shown on Fig. 4A. Accordingly, the number of articles with k bursts decreases with k. Burst counts for the most active platforms (Twitter, news, Facebook) decrease sublinearly under log-transformation. This indicates that the probability of a $k+1$-th burst, once k bursts have been observed, decreases with k. This confirms the findings previously obtained on Facebook alone (Cheng et al. 2016) and shows that they hold up on other platforms.

Recurring bursts become smaller and less frequent: their size (Fig. 4B) typically decreases over time, and they recur about one month apart, or longer for subsequent bursts (Fig. 4C). Twitter, Facebook, and Google+ all exhibit this pattern. It is likely driven by collective attention that decays over time with novelty of the content, so it attracts less references (Wu and Huberman 2007). Interestingly, news outlets and Wikipedia follow a different trend than other platforms. For them, burst sizes and distances lack clear dependence on recurrence. This suggests that the intensity of posting activity on these two platforms does not strongly depend on the novelty of information. This poses contrast to the previous findings from (Cheng et al. 2016) and shows that they do not apply to two out of five platforms considered here.

Bursts Linked across Platforms

Next, we investigate the succession of bursts on different platforms to explore their characteristic "catch-up times". Specifically, do bursts happen at the same time across multiple platforms or are there any delays that could make one platform preferable over another for potential science marketing endeavors?

To answer this question, we order the time of occurrence for all the bursts associated with an article on different platforms. Then we identify adjacent bursts as linked if they are no more than $\Delta w = 3$ days apart. Replication with values of Δw in the range of $1, \ldots, 5$ produces the same results

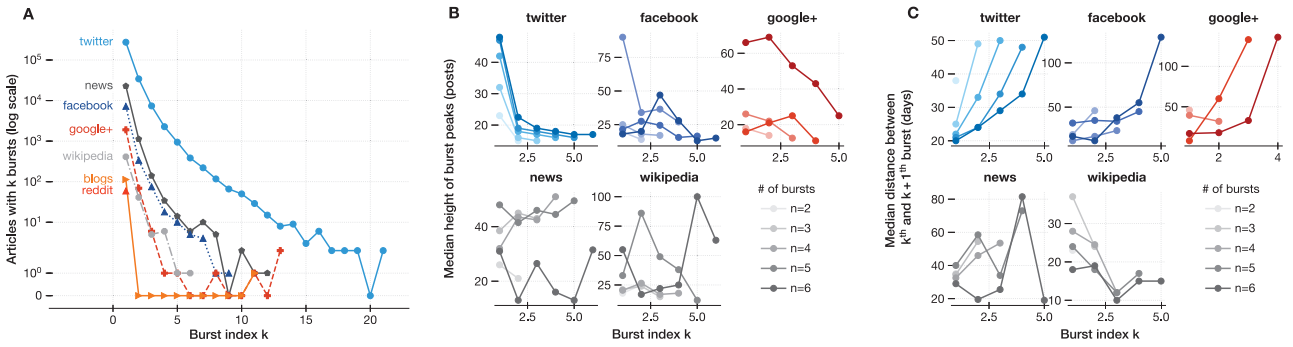

Figure 4: Histograms of: (**A**) numbers of bursts on each platform during one year since the first mention, (**B**) median number of posts in kth burst's peak for each platform, and (**C**) median distance in days between subsequent bursts.

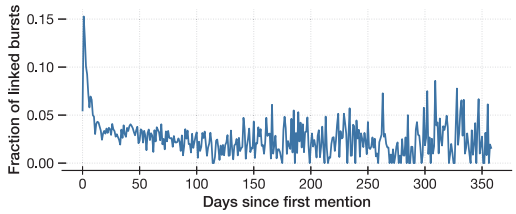

Figure 5: Fraction of linked bursts on different days.

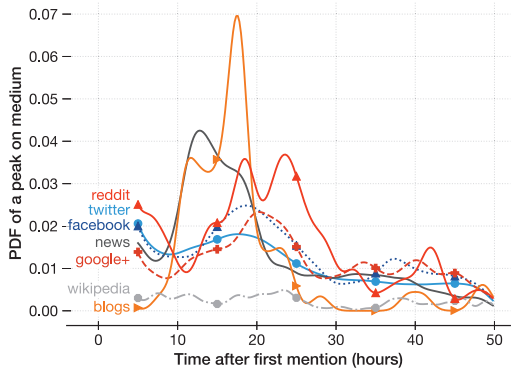

Figure 6: Probability distribution of bursts occurring on each platform during first 50 hours after the first mention.

as described below. Since we use $\Delta w < w$ for burst detection, it is guaranteed that linked bursts are from different platforms. Then we find the fraction of linked bursts in the overall number of bursts happened on that day (Fig. 5).

In general, there is a low probability that burst on one platform will be accompanied by a burst on other platforms (around 3% for any period after 15 first days). However, one day after the initial mention this probability jumps to 15% from the initial 6%. Given that only 17.6% of articles are ever mentioned on more than one platform, we argue that approximately one day after the first mention there is a strong tendency for bursts to be linked across platforms. This is a novel finding that establishes the time frame for which it is meaningful to look for linked bursts.

Platform Catch-up Times

Finally, we want to determine the speed at which the information about scientific papers reaches the different platforms and causes bursts of posting activity. For that, we analyze the relative positions of linked bursts during the three days since the initial post about the paper on any platform. In this time frame we have enough data to study bursts on an hourly scale. Switching to his granularity allows us to determine which platforms start mentioning research outputs earlier, and which of them — later.

We re-run the burst detection algorithm with a new time window of one hour. To adjust for smaller numbers of posts and prevent "bunching" of bursts together, we select new values of $h_0 = 5$ and $w = 10$. For every detected burst, we record the hour of its peak. Then we aggregate these values by calculating the frequency-based empirical probability of a burst happening in each platform within hourly time frames (see Fig. 6).

News outlets are the fastest to reach the peak in approximately 12.75 hours after the first mention, followed by blogs (17.5h), Twitter (17.75h), Facebook (18.75h), Google+ (20.25h), Reddit and Wikipedia (23.5h). These timings suggest a pattern for diffusion of information between the platforms: after a burst is observed on one platform, it can propagate to other platforms with longer catch-up times. Although we used a substantially different method, our results are similar to previous observations or information hand-off from the literature. Leskovec et al. found that the hand-off of information from news to blogs happened with 2.5h gap for a broad range of memes (Leskovec, Backstrom, and Kleinberg 2009). We extend estimation of catch-up times to all 7 media and allow for direct comparisons between all of them.

Inference of Diffusion Structure

To explore the propagation of scientific articles between different platforms and to better understand the peculiarities of sharing science, we reconstruct the network of information

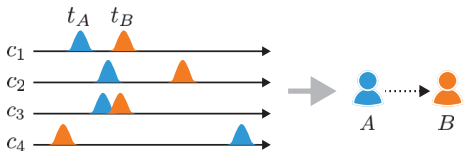

Figure 7: Underlying principle of NETINF: the more cascades c_i with user A posting shortly before user B, the more likely a diffusion link $A \rightarrow B$ is.

diffusion based on the observed sequences of posts on different platforms. We evaluate the resulting network on a set of ground-truth data and check robustness of our results.

Network Inference Algorithm

To reconstruct diffusion network, we use NETINF, an efficient algorithm for network inference (Gomez-Rodriguez, Leskovec, and Krause 2012). NETINF models posting activity for each article through information cascades, one per article. The structure of cascade c shows the likely path of information transfer. The posting activity associated with c is described by a vector of timestamps at which the users have posted $(t_1, \ldots, t_{|V|})$. The task of network inference is then to determine the most likely diffusion structures for all cascades given the sequences of post timestamps. As its main heuristic, NETINF uses (*i*) the number of cascades where pairs of users have participated together combined with (*ii*) differences in their posting times (see Fig. 7 for illustration).

Each information cascade c is considered to be a subgraph of a static diffusion network $D = (V, E)$. Each node in the cascade, i.e., each user who has posted, can only have one "parent" node, from which it received information about the article before posting itself. Cascade c thus belongs to $\mathcal{T}_c(D)$, the set of all subgraphs of D that are forests and contain the nodes of c.

The probability of information transmission between users A and B depends only on the time difference between their posts $\Delta t_{A,B} = t_B - t_A$. Specifically, $P_c(A, B) \propto \frac{1}{\Delta t_{A,B}^{\alpha}}$, since power-law distribution agrees the most with our data. Additionally, there is a small probability $\varepsilon = 10^{-8}$ that a transmission of information would happen between any pair of users at random.

For each cascade, only the most likely propagation subgraph $T \in \mathcal{T}_c(D)$ is considered, i.e., $P(c|D) \approx \max_{T \in \mathcal{T}_c(D)} P(c|T)$, where $P(c|T) \propto \prod_{(u,v) \in T} P_c(u, v)$. Since cascades are assumed to be independent of each other, the probability of all cascades given diffusion network can be found by $P(C|D) = \prod_{c \in C} P(c|D)$. The optimization problem that NETINF solves can be formulated as finding the network D that maximizes this $P(C|D)$. Since this probability monotonically increases with addition of new edges to D, an additional constraint is added that the diffusion network should not contain more than k edges:

$$\hat{D} = \text{argmax}_{|D| \leq k} P(C|D).$$

The algorithm greedily optimizes for $P(C|D)$: at each iteration, it considers the probability gains from adding each possible edge $e = (u, v)$ such that $\exists c : t_u \leq t_v$, and adds an edge with the largest gain to D. Gains are computed as the sum of differences in log-likelihoods of individual cascades $\sum_{c \in C} \log P(c|T \cup e) - \log P(c|T)$.

Although the ground truth diffusion network may be not static, NETINF deals with temporality implicitly. If for a sufficiently long period of time, two users consistently post one after another, an edge will likely be inferred between them, even if this pattern of activity stops afterwards. Given that we evaluate the results of network inference based on the entire time period, we should be able to recover a large portion of the most likely influences.

Evaluation

To run NETINF, we selected the top 1,000 posters from each platform, for a total of 7,000 users. To help exclude social media bots and spammers, we determined top users by the number of unique scientific articles they mentioned. We only considered cascades with at least two top users participating in them, which yielded 1,784,540 cascades, by the number of articles. Using the timestamp vectors of selected cascades as an input, we inferred 100,000 most likely diffusion edges between the top users.

To obtain baselines for the quality of the inference, we also generated a set of benchmark networks with the same number of nodes and edges and compared them with results of the NETINF algorithm:

- RANDOM — an Erdős-Rényi random network with all edges having the same probability of existence. It shows the results of guessing the edges at random.
- NAIVE — a network with edges (A, B) that could have participated in the largest total number of cascades ($\Delta t_{A,B} > 0$). It demonstrates a simple approach based on the knowledge of who usually posts after whom.

To validate the results of network inference, we compare them with ground truth data about known connections between users. We collected all posts belonging to our data sample that were authored by selected top users of Twitter and Facebook. If one of the users reposted or linked to posts of another user within the same platform at least once, we considered the two to have a diffusion link between them in the ground truth network:

- TWITTER — our data contains over 700K pairs of posts retweeting, quoting, or linking to one another that identify over 19,5K unique connections between Twitter users;
- FACEBOOK — more than 101K posts are reposts or contain links to other posts, which allows us to detect 1,544 unique connections between Facebook users.

Although this way of determining ground truth connections is not perfect, it is the best available option given that both Twitter and Facebook severely limit access to information about their users' followers and friends. Our procedure for building the ground truth networks worked better in the case of Facebook than Twitter, because Twitter automatically links reposts to the original post, regardless of the actual trajectory of the tweets. We find that stricter definitions of connections between the users of these platforms

Table 2: Evaluation results of generated networks.

Dataset	Method	$F1^{max}$	k^{max}
TWITTER	RANDOM	0.002	99,443
	NAIVE	0.100	99,845
	NETINF	0.361	78,491
FACEBOOK	RANDOM	0.000	—
	NAIVE	0.045	86,267
	NETINF	0.495	71,977

(e.g., two or more posts linking to each other) lead to similar results.

Results

The ground truth networks contain only a subset of nodes of the inferred network. Precision and recall are thus measured based on connections between the nodes that belong to ground-truth networks. Since the NETINF algorithm can infer arbitrary many edges, its precision and recall may vary greatly with the number of inferred edges k. With addition of the first edges, the precision is high and recall is low. Adding more edges makes precision decrease and recall to increase. To balance these changes, we use F1-score. With addition of edges, the score increases until it reaches its maximum, and then starts to decrease. Table 2 reports the maximum F1-scores and numbers of edges k^{max} at which they were achieved.

In all cases, NETINF performs much better than both benchmarks according to their maximum F1 values. It also achieves its maximum performance with smaller number of edges k^{max}. Even though NETINF performs substantially below a perfect score of 1.0, the results indicate a successful reconstruction of the underlying diffusion structures for both ground truth networks. Network inference is a hard problem, where a random guessing approach and a simple heuristic perform extremely poorly in comparison. We also expect that the differences in the accuracy on the two data sets are tied to the differences in completeness of the ground truth networks obtained from Facebook and Twitter, respectively.

Additional Tests and Robustness Checks

As a robustness check, we studied how the inference results depends on the time frame covered by the selected data. We limited the posting activity to first two years (2011–2013) and last two years (2016–2018) of data and inferred 75,000 most likely diffusion edges based on each subsample. We chose the threshold based on the k^{max} values reported in Table 2. The evaluation against TWITTER and FACEBOOK ground truth networks produced the following $F1^{max}$ values: 0.189 and 0.074 for the first two years, and 0.301 and 0.483 for the last two years. Accordingly, using the first two years of the data does not provide results in agreement with the 7-year sample. During 2011-2013, Altmetrics was setting up their system and we expect thus that this time frame is the least reliable part of our data. When we use the last two years, however, the results are qualitatively similar to the 7-year sample. As we discuss in data description, the only notable difference between the subsample and the entire data is the fact that in 2015 (see trends shown in Table 1), Altmetrics expanded their coverage of news. Hence, we see an increased prevalence of news users in the cascades, which also means that news become the source of information more often than before and have a higher coreness in the network (see "Structure of Diffusion" for more details).

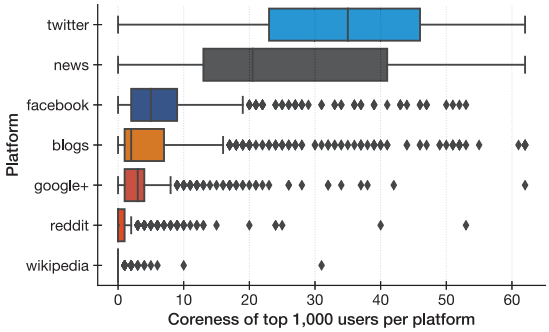

Figure 8: Distributions of coreness values in the inferred diffusion network for top 1,000 users of each platform.

We have also considered extensions of NETINF algorithm that encounter additional factors such as full text content of posts and hyperlinks used in them by extending the formulation of diffusion probability $P_c(A, B)$. However, such extensions improved the accuracy of results only marginally ($\leq 1\%$), which is incommensurate with their complexity. The inference was dominated by the number of cascades that each edge could potentially participate in, similar to the way the NAIVE benchmark operates.

Structure of Diffusion

The inferred diffusion network has the level of granularity to enable tackling questions that are specific to the dissemination of science within and across platforms about who are the main spreaders of scientific articles and what platforms are they using. Additionally, we can also quantify the frequency of transmissions between platforms as well as study the interplay between broadcasting that results from, e.g., press releases, and actual peer-to-peer spreading of scientific articles.

Positions of Platform Users

First, we investigate the most prominent users in the inferred network. Table 3 shows users with the highest out-degrees for each platform, i.e., the number of outgoing diffusion links for users within individual platforms. These findings show that some of the most active users are large professional outlets like *Nature*, *EurekAlert!*, and *PsyPost*, as well as some individuals. Different areas of knowledge are well represented, notably, medicine, psychology, and physics, among others.

We further inquire whether users of different platforms systematically differ in terms of their positions in the network. First, we compute the coreness value of each user

Table 3: Top users by out-degree in the inferred network.

Rank	News	Blogs	Twitter	Facebook	Google+	Reddit
1	EurekAlert!	I F***ing Love Science	@NatureNews	Nature	Alex Psi	officialcitral
2	MedicalXpress	Neuroscience RSS Feeds	@uranus_2	NatureNews	Tim Cannon	burtzev
3	Phys.org	ZME Science	@animesh1977	JAMAJournal	Nature News & Comment	starspawn0
4	Yahoo! News	PsyPost	@EricTopol	NatureReviews	Jay Cross	anutensil
5	Medical News Today	Sci-News.com	@JAMA_current	TheConversationUS	Nature	mvea

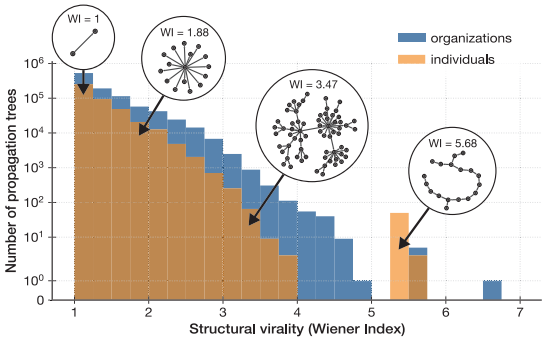

Figure 9: Structural virality of individual propagation trees. Histogram color denotes whether the tree has been started by organization or individual. Examples show the different tree structures with different values of the Wiener index.

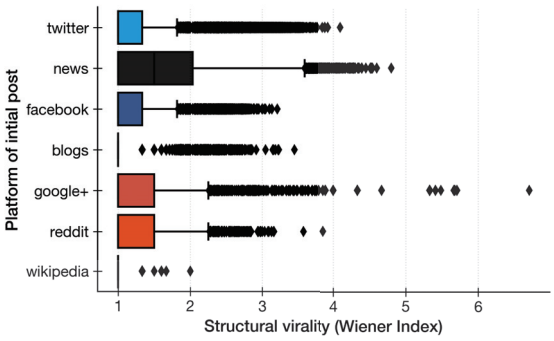

Figure 10: Structural virality of individual propagation trees, grouped by the platform on which initial post was made.

through a repeated pruning of the diffusion network based on degree, such that in the k-core of the network, every user has at least k connections. Then, we aggregate coreness values by platform. We choose coreness to describe the position of users in the network, because it is a robust measure that can counteract local noise in the inferred network. Unlike centrality indices and indicators depending on paths, coreness is more stable to changes in individual connections.

As Fig. 8 shows, there is a split in the typical coreness of users of different platforms: Twitter users and news outlets have a considerably higher coreness than most users of other platforms. It means that the diffusion network has a core-periphery structure where news organizations and Twitter users belong to the tightly connected core of the network, while the rest lay closer to the periphery. Within this structure, information can propagate sequentially, akin to a multi-step flow of communication, where users from different platforms participate at different steps of the flow. The closer one seeds information to the core, the more opportunities information has to propagate to peripheral nodes. Accordingly, we expect on average Twitter users and news outlets to be more influential than users of other platforms.

However, we notice that there are outliers on all platforms with high coreness values (≥ 30). These users can compete with Twitter and news in terms of influence. It also means that users of most platforms are represented across all coreness levels and, consequently, steps of communication flow.

Structural Virality

Previous analyses have equated posting with information diffusion. To test whether the initial stages of a cascade reflect less actual diffusion across platforms and rather efforts by publishers, we use the concept of structural virality to place inferred cascades on the spectrum between pure broadcast and person-to-person spreading (Goel et al. 2015). Then, we compare the typical structures of cascades initiated by individuals and organizations.

Each cascade consists of one or more trees. The root of each tree is the user who started the information diffusion and the edges show how information has been propagating between users. Such trees have a wide range of possible structures, with structural virality being an established way to characterize them. This measure is calculated effectively through the Wiener index originating from mathematical chemistry and is defined as the average distance between all pairs of nodes in a diffusion tree. Denoting the distance between a pair of nodes i, j in a tree T with n nodes as d_{ij}, the index is defined as:

$$\mathrm{WI}(T) = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} d_{ij} \qquad (1)$$

Smaller values of WI correspond to conditions where information was spread in a single, possibly large, broadcast, while bigger values stand for a viral diffusion, where information was propagated by different users in multiple steps, with each user responsible for only a fraction of all sharing. For instance, WI < 2 characterizes trees with star structures, where all nodes are directly connected to the root. With larger values of WI, tree branching becomes more complex. Finally, large values of WI denote propagation trees that are chain-like, where information has been passed on sequentially in multiple steps.

The structural virality of all propagation trees is shown on

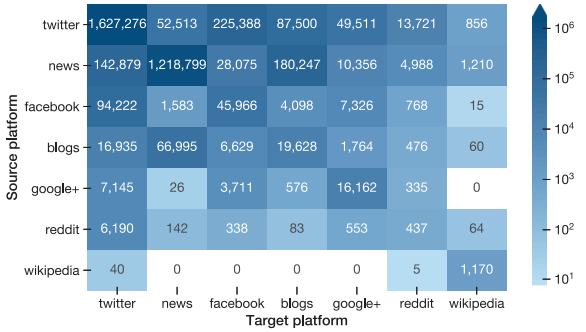

Figure 11: Number of times the information has propagated from each source platform to target platform based on the inferred propagation trees.

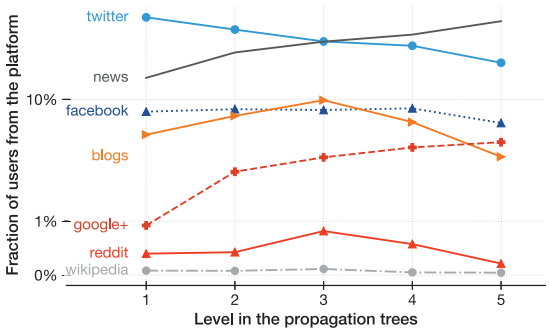

Figure 12: Percentage of users from different platforms at different levels of inferred propagation trees (on a log scale).

Fig. 9. Accordingly, viral dissemination of research articles is orders of magnitude less likely than a direct broadcast.

To better describe the role of individuals and organizations in the propagation of scientific articles, we automatically classify users based on the following heuristic: We split the screen name into tokens using spaces and capitalization of words, and check them against a comprehensive list of first names from different countries[2]. If one of the tokens matches a known name, the user is considered an individual, and otherwise — as organization.

We find that diffusion started by organizations tends to be more structurally viral than the one started by individuals and that they both follow the same general pattern. This is a surprising finding, since organizations have mostly been associated with broadcasting, while users are an integral part of viral sharing. Yet, organizations involved in sharing science have the necessary platform infrastructure and public acceptance that attracts the attention of online users.

After grouping the users who made the initial posts by the platform they used (Fig. 10), we see that the propagation trees started by news outlets have by far the highest median virality. They are followed by cascades started in Google+, Reddit, Twitter, and Facebook. Notably, there are outliers in most platforms, indicating that the diffusion of science occasionally becomes viral, regardless of the platform where it was seeded.

Diffusion across Platforms

To better understand how information spreads between the platforms, we analyze information cascades that traverse them. For each pair of platforms, we count the number of times the information has spread from a user of one platform to a user of another platform according to the inferred propagation trees. This gives us the counts presented in Fig. 11. These numbers correspond to the number of pieces of information transferred between the platforms. They don't correspond to individual papers, as information about a single paper might get transferred across the same pair of platforms or within the same platform multiple times.

Most of the time, the source of information is Twitter users (over 2 million times) and news outlets (over 1.5 million). These two platforms exert the largest influence on the others, similar to conclusions of our analysis of user coreness. The data also shows that Wikipedia appears to be a "sink" for information from other media. Yet, its full influence is likely to be underestimated in our data.

In the case of Twitter, news, Google+, and Wikipedia, scientific articles are typically shared within the same platform. Other platforms, however, indicate interesting transmissions. For instance, articles propagate more commonly from Facebook and Reddit to Twitter, and these platforms are in turn most influenced by Twitter. Similarly, blogs influence news outlets more often than other blogs, and they are in turn most influenced by news. These asymmetries are caused in part by the differences in posting frequencies across platforms: Twitter users are more active than Facebook and Reddit users, and news publish more regularly than blogs. Note that this result provides conservative estimates of the asymmetries between sharing within and across platforms, because we based it on the activity of 1,000 most active users from each platform. Extrapolating it on more users per platform will likely reveal larger heterogeneities.

Finally, we ask which users start sharing articles, i.e., represent the root nodes of cascades, and who keep the propagation going. We denote the users who start the information diffusion as Level 1, the ones who received information directly from them as Level 2, and so on. In Fig. 12, we plot the proportion of users from different platforms at each level of the propagation trees. For example, out of the users at Level 1, 70% are from Twitter, 16.7% from news outlets, 7.5% from Facebook, 4.3% from blogs, and less than 2% from other platforms combined.

Taken together, Twitter and news account for more than 80% of users at each level, but their percentages change. The proportion of Twitter users drops to 23.9% by Level 5, while the share of news outlets grows to about 64%. Thus, original posts about scientific papers tend to appear on Twitter, but due to the lower virality of the cascades originating

[2]Name lists taken from https://github.com/alt-code/Research/tree/master/SimpleGenderComputer/namelists .

from this platform, they stop before reaching further levels. On the other hand, news tend to generate the most viral cascades. Combined with the tendency of news outlets to primarily influence other news outlets, this increases their share in further levels. Other platforms have much lower proportions of users across all levels, and their dynamics is largely determined by the activity within them and their interactions with Twitter and news outlets.

Discussion

The dissemination of scientific findings impacts public opinion and policy. Yet, it is unclear what role users on online platforms play in generating and spreading information about scientific advances. Moreover, our knowledge about how any type of information diffuses across various online platforms is still rudimentary. To fill this gap, we studied here posts about 7.2 million scientific articles from a selection of seven online platforms. We make a number of contributions that corroborate and build upon discoveries from existing literature.

In a series of papers exploring dynamics of posting activity within individual platforms, it has been shown that such activity is organized in bursts that occasionally recur (Gilbert 2013; Cheng et al. 2014; 2016). These results replicate in the dissemination of scientific articles and hold in case of most platforms we consider. However, we also highlight important exceptions. Notably, news outlets and Wikipedia do not conform to the common tendencies of cascade sizes and inter-cascade time gaps. Furthermore, when we link bursts *across* platforms, we discover that bursts on different platforms tend to happen with specific time offsets after the initial mention of a scientific article.

Additionally, in a study about meme tracking across news websites and blogs (Leskovec, Backstrom, and Kleinberg 2009), it was found that online news outlets introduce stories first and blogs catch up later. In this paper, we reproduce this pattern and extend it to other platforms, determining the relative timing of bursts in every one of them.

In relation to different theories about the flow of communication, we uncover that both individuals and organizations participate in dissemination of scientific information online. We find that posts about scientific articles can originate on various platforms, but their structural virality depends on the source platform and on whether the user is an individual or an organization. In general, the content originating on news outlets is associated with higher virality than the one created by other platforms; and organizations tend to produce more viral content than individuals. Furthermore, there are systematic trends in the rates of information transfer between different platforms, which also determine where scientific articles are shared at different stages of the diffusion.

In the domain of disseminating science online, we move beyond the commonly studied question of popularity of scientific articles (Robinson-Garcia et al. 2017; MacLaughlin, Wihbey, and Smith 2018). With the aim to uncover how online platforms and individual users communicate with each other, we provide a framework for studying information diffusion across multiple platforms, which opens up avenues for a holistic investigation of online phenomena in a broad ecosystem of platforms.

The directions for future work are twofold. So far, we have examined temporal and structural aspect of information diffusion across multiple platforms. In future work, we plan to focus on the content of posts to see how it changes thoughout this diffusion process and how diverse is the overall coverage of scientific advances. On the other hand, we plan to further study whether the online posting activity contains collective intelligence signals that would be useful for evaluation of scientific articles themselves. This could potentially be used to create more robust mechanisms for collective evaluations of research outputs, in the spirit of research on altmetrics.

Acknowledgments. Authors would like to thank Stacy Konkiel of Altmetric LLC for providing data from their platform, Nick Diakoupulos, Darren Gergle, Nick Vincent, and the anonymous reviewers for their valuable feedback and suggestions. This work has been partially funded by NSF CRII award (IIS-1755873).

References

Adar, E., and Adamic, L. A. 2005. Tracking information epidemics in blogspace. In *Web Intelligence, 2005. Proceedings. The 2005 IEEE/WIC/ACM International Conference On*, 207–214. IEEE.

2018. What outputs and sources does Altmetric track? https://help.altmetric.com/support/solutions/articles/6000060968-what-outputs-and-sources-does-altmetric-track.

Aral, S., and Walker, D. 2011. Creating social contagion through viral product design: A randomized trial of peer influence in networks. 34.

Barabási, A.-L. 2005. The origin of bursts and heavy tails in human dynamics. *Nature* 435:207.

Bauer, M. W. 2012. Public attention to science 1820–2010 – A 'longue durée' picture. In Rödder, S.; Franzen, M.; and Weingart, P., eds., *The Sciences' Media Connection –Public Communication and Its Repercussions*, volume 28. Dordrecht: Springer Netherlands. 35–57.

Beering, S., and others. 2014. Science and technology public attitudes and understanding.

Blenn, N., and Van Mieghem, P. 2016. Are human interactivity times lognormal? *arXiv preprint arXiv:1607.02952*.

Brossard, D. 2013. New media landscapes and the science information consumer. *Proceedings of the National Academy of Sciences* 110(Supplement_3):14096–14101.

Cheng, J.; Adamic, L.; Dow, P. A.; Kleinberg, J. M.; and Leskovec, J. 2014. Can cascades be predicted? 925–936. ACM Press.

Cheng, J.; Adamic, L. A.; Kleinberg, J. M.; and Leskovec, J. 2016. Do cascades recur? In *Proceedings of the 25th International Conference on World Wide Web*, 671–681. International World Wide Web Conferences Steering Committee.

Clauset, A.; Shalizi, C. R.; and Newman, M. E. J. 2009. Power-law distributions in empirical data. *SIAM Review* 51(4):661–703.

Doerr, C.; Blenn, N.; and Van Mieghem, P. 2013. Lognormal infection times of online information spread. *PloS one* 8(5):e64349.

Easley, D., and Kleinberg, J. 2010. Information cascades. In *Networks, Crowds, and Markets: Reasoning about a Highly Connected World. Cambridge University Press*.

Gilbert, E. 2013. Widespread underprovision on Reddit. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work*, 803–808. ACM.

Goel, S.; Anderson, A.; Hofman, J.; and Watts, D. J. 2015. The structural virality of online diffusion. *Management Science* 150722112809007.

Goldenberg, J.; Libai, B.; and Muller, E. 2001. Talk of the network: A complex systems look at the underlying process of word-of-mouth. *Marketing Letters* 12(3):211–223.

Gomez-Rodriguez, M.; Leskovec, J.; and Krause, A. 2012. Inferring networks of diffusion and influence. *ACM Transactions on Knowledge Discovery from Data* 5(4):1–37.

Hargittai, E.; Füchslin, T.; and Schäfer, M. S. 2018. How do young adults engage with science and research on social media? Some preliminary findings and an agenda for future research. *Social Media + Society* 4(3):2056305118797720.

Hilbert, M.; Vásquez, J.; Halpern, D.; Valenzuela, S.; and Arriagada, E. 2017. One step, two step, network step? Complementary perspectives on communication flows in twittered citizen protests. *Social Science Computer Review* 35(4):444–461.

Jamali, S., and Rangwala, H. 2009. Digging Digg: Comment mining, popularity prediction, and social network analysis. In *2009 International Conference on Web Information Systems and Mining*, 32–38. Shanghai, China: IEEE.

Katz, E., and Lazarsfeld, P. 1955. Personal influence.

Keegan, B.; Gergle, D.; and Contractor, N. 2013. Hot off the wiki: Structures and dynamics of Wikipedia's coverage of breaking news events. *American Behavioral Scientist* 57(5):595–622.

Lakkaraju, H.; McAuley, J.; and Leskovec, J. 2013. What's in a name? understanding the interplay between titles, content, and communities in social media. In *Seventh International AAAI Conference on Weblogs and Social Media*.

Lerman, K., and Ghosh, R. 2010. Information contagion: An empirical study of the spread of news on Digg and Twitter social networks. *ICWSM* 10:90–97.

Leskovec, J.; Backstrom, L.; and Kleinberg, J. 2009. Meme-tracking and the dynamics of the news cycle. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 497–506. ACM.

Lorenz-Spreen, P.; Monsted, B. M.; Hövel, P.; and Lehmann, S. 2019. Accelerating dynamics of collective attention. *Nature Communications* 10(1759).

MacLaughlin, A.; Wihbey, J.; and Smith, D. A. 2018. Predicting news coverage of scientific articles. *Proceedings of ICWSM* 10.

Matei, S. A. 2017. *Structural Differentiation in Social Media: Adhocracy, Entropy, and the 1% Effect*. New York, NY: Springer Science+Business Media.

Milkman, K. L., and Berger, J. 2014. The science of sharing and the sharing of science. *Proceedings of the National Academy of Sciences* 111(Supplement 4):13642–13649.

Peters, H. P.; Brossard, D.; de Cheveigné, S.; Dunwoody, S.; Kallfass, M.; Miller, S.; and Tsuchida, S. 2008. Interactions with the mass media. *Science* 321(5886):204–205.

Pew Research Center. 2018. Digital news fact sheet. http://www.journalism.org/fact-sheet/digital-news/.

Ratkiewicz, J.; Fortunato, S.; Flammini, A.; Menczer, F.; and Vespignani, A. 2010. Characterizing and modeling the dynamics of online popularity. *Physical Review Letters* 105(15).

Reinhardt, W.; Ebner, M.; Beham, G.; and Costa, C. 2009. How people are using Twitter during conferences. *Creativity and Innovation Competencies on the Web. Proceedings of the 5th EduMedia* 145–156.

Robinson-Garcia, N.; Costas, R.; Isett, K.; Melkers, J.; and Hicks, D. 2017. The unbearable emptiness of tweeting—About journal articles. *PLOS ONE* 12(8):e0183551.

Scheufele, D. A. 2013. Communicating science in social settings. *Proceedings of the National Academy of Sciences* 110(Supplement_3):14040–14047.

Scott, D. W. 1992. *Multivariate Density Estimation: Theory, Practice, and Visualization*. John Wiley & Sons.

Shearer, E., and Matsa, K. E. 2018. News use across social media platforms 2018. Technical report, Pew Research Center.

Su, L. Y.-F.; Akin, H.; Brossard, D.; Scheufele, D. A.; and Xenos, M. A. 2015. Science news consumption patterns and their implications for public understanding of science. *Journalism & Mass Communication Quarterly* 92(3):597–616.

Szabo, G., and Huberman, B. A. 2010. Predicting the popularity of online content. *Communications of the ACM* 53(8):80.

Tan, C.; Friggeri, A.; and Adamic, L. 2016. Lost in propagation? Unfolding news cycles from the source. *ICWSM* 10.

Vosoughi, S.; Roy, D.; and Aral, S. 2018. The spread of true and false news online. *Science* 359(6380):1146–1151.

Weimann, G. 1982. On the importance of marginality: One more step into the two-step flow of communication. *American Sociological Review* 764–773.

Weng, L.; Flammini, A.; Vespignani, A.; and Menczer, F. 2012. Competition among memes in a world with limited attention. *Scientific Reports* 2(1).

Wu, F., and Huberman, B. A. 2007. Novelty and collective attention. *Proceedings of the National Academy of Sciences* 104(45):17599–17601.

Wu, S.; Hofman, J. M.; Mason, W. A.; and Watts, D. J. 2011. Who says what to whom on Twitter. In *Proceedings of the 20th International Conference on World Wide Web - WWW '11*, 705. Hyderabad, India: ACM Press.