Dissemination of scholarly literature in social media

Pablo Moriano, Emilio Ferrara, Alessandro Flammini, and Filippo Menczer
Center for Complex Networks and Systems Research
School of Informatics and Computing
Indiana University, Bloomington, USA

ABSTRACT

Social media data have been increasingly used to assess the impact of scholarly research. Such data provide complementary metrics (often called *altmetrics*) to traditional impact indicators. This paper provides a summary on the diffusion of scholarly content in social media, based on a collection of tweets citing papers from a set of 27 academic publishers within various fields between 2011 and 2013. We first show that there has been an increasing adoption of Twitter as a channel to disseminate scholarly literature. In particular, between 2012 and 2013, the number of scholarly tweets and the fraction of tweets (over the entire corpus) have increased by 91.2% and 42.6% respectively. We then analyze the structure of the information diffusion network. We show that the distributions of the numbers of times a specific paper is tweeted, retweeted, and the number of connected components in the diffusion network are scale-free. These preliminary results suggest that, as for other kinds of information, there are underlying mechanisms that lead some scholars and their products to become viral.

1. INTRODUCTION

Social media have been adopted as widely used sources of data in the study of complex social dynamics. In recent years, data from social media platforms have attracted the attention of researchers who study the diffusion and impact of scholarly content [8, 2]. Social media data, specifically data gathered from platforms such as LinkedIn, Facebook and Twitter, offer complementary ways (called *altmetrics*) to measure scholarly impact against traditional citation indicators [6]. Here, we are particularly interested in quantifying how scientific knowledge is disseminated in these environments.

While the reliability and validity of altmetrics is still under investigation [3], the use of traditional citation metrics as measures of impact of scholarly publications is being reevaluated. On the one hand, people argue that traditional metric indicators are exclusively based on the transfer of knowledge in close research communities [9]. On the other hand, altmetrics are supposed to account for a broader audience, since they are based on data from online sources used by different types of populations [7]. Focusing on Twitter, we provide evidence of the increasing use of social media in the dissemination of scholarly content, we characterize the distribution of popularity of scholarly products, and illustrate statistical patterns in the aggregate network of scholarly information diffusion. We focus on a multidisciplinary set of 27 well-known academic publishers. Our contribution is threefold. First, we analyze to what extent scientific publications are discussed in Twitter. In doing so, we measure the increase in the use of Twitter as a dissemination tool for academic content. Second, we quantify the distribution of the numbers of times a specific paper is tweeted and retweeted. Third, we characterize the structure of the diffusion network for scholarly products by characterizing the distribution of the number of connected components for the diffusion network. Our results suggest that, similar to a variety of physical, biological, and man-made phenomena [5], Twitter communication dynamics about scientific literature is also characterized by scale-free distributions.

Table 1: Publisher domains used in the study.

1	acm.org	10	ieee.org	19	plosone.org
2	acs.org	11	jbc.org	20	pnas.org
3	ams.org	12	jstor.org	21	sciencedirect.com
4	aps.org	13	mdpi.com	22	sciencemag.org
5	arxiv.org	14	metapress.com	23	springer.com
6	biomedcentral.com	15	nature.com	24	ssrn.com
7	cell.com	16	nejm.org	25	thelancet.com
8	doi.org	17	oxfordjournals.org	26	wiley.com
9	elsevier.com	18	plos.org	27	worldscientific.com

2. DATA COLLECTION

Our collection of tweets is based on a 10% sample of the Twitter stream, and spans three years of data (2011–2013). In particular, we use tweets collected by the Truthy project [4]. In our dataset we retain only tweets and retweets with an explicit or shortened link to a paper within a pre-selected set of publishers (Table 1.)

3. RESULTS AND DISCUSSION

Fig. 1(A) show the total number of tweets about scholarly papers for every year during the observation period. We note a growing amount of scholarly communication. In particular, there is an increase of 165.1% between 2011 and 2012, and one of 91.2% between 2012 and 2013. Fig. 1(B) illustrates the variation in the percentage of scholarly tweets. For instance, there is an increase of 29.5% between 2011 and 2012, and 42.6% between 2012 and 2013. This suggests that social media platforms are increasingly perceived as tools to promote academic literature, and that scholarly content is capturing a growing share of social media attention.

Fig. 2 shows the number of tweets about scholarly papers during the observation period by month and by publisher. The most

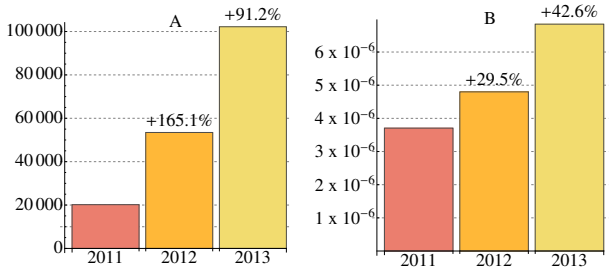

Figure 1: (A) Scholarly tweets per year. (B) Percentage of scholarly tweets over time.

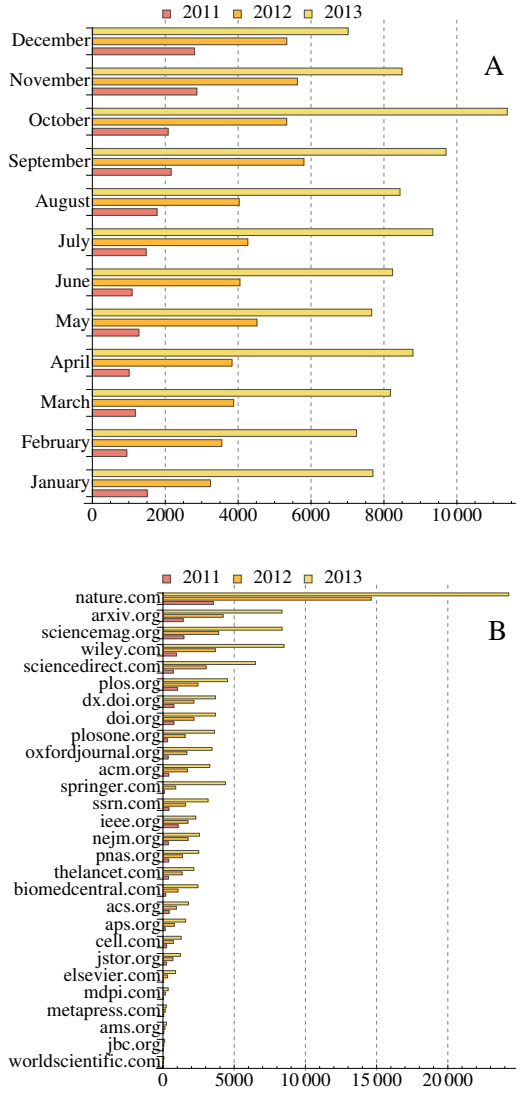

Figure 2: Scholarly tweets per (A) month, (B) publisher.

tweeted papers are published by Nature, arXiv, and Science.

Fig. 3(A) shows the *complementary cumulative distribution function* (CCDF) of the number of times a url to a specific paper is contained in a tweet. The distributions follow a power law with estimated scaling exponents [1] of 3.1, 2.7, and 2.6, respectively. Fig. 3(B) shows the CCDF of the number of retweets that reference a specific paper. Corresponding distributions are also approx-

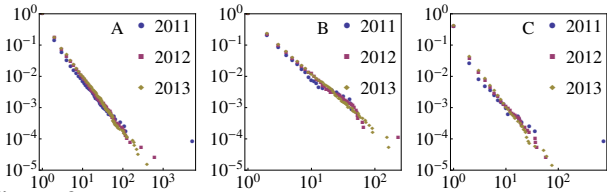

Figure 3: CCDF of the numbers of times a specific paper is (A) tweeted, (B) retweeted, and (C) connected components in the paper diffusion network.

imately power laws, with exponents of 2.8, 2.7, and 2.7. To represent the dissemination of academic literature within the collected set of tweets, for each paper we construct a directed graph in which nodes represent users. An edge from user X to Y represents a mention of Y by X or a retweet by Y of a message from X. Fig. 3(C) shows the CCDF of the number of connected components for this diffusion network (exponents of 2.5, 3.0, and 3.0). This analysis leads us to hypothesize that, similarly to other information diffusion phenomena, scholarly paper dissemination may follow a preferential attachment mechanism. This results in a few widely popular papers accounting for a great fraction of mentions, retweets and community penetration.

4. CONCLUSIONS

We analyzed the use of Twitter in scholarly-related discussion. We observed an increasing use of social media as channels for scholarly content dissemination. We also studied the structure of these diffusion networks, finding that most conversations tend to spread within small numbers of communities. However, there are some papers that become very popular and go viral spreading in multiple communities. The present analysis is at the level of publisher domains; tracking URLs of individual papers will allow us to investigate whether the number of distinct papers and the fraction of papers represented in Twitter are growing. Further investigation will also aim to understand under what conditions scholarly-related conversations become viral in social media. Comparison with spreading of information in other contexts (e.g., politics, social movements, or the news) will shed light on the underlying forces that shape academic content dissemination. This work was supported in part by NSF grant CCF-1101743.

5. REFERENCES

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

[2] S. Haustein, I. Peters, C. R. Sugimoto, M. Thelwall, and V. Larivière. Tweeting biomedicine: An analysis of tweets and citations in the biomedical literature. *J. Assoc. Inf. Sci. Technol.*, 65(4):656–669, 2014.

[3] R. Kwok. Research impact: altmetrics make their mark. *Nature*, 500(7463):491–493, 2013.

[4] K. McKelvey and F. Menczer. Truthy: Enabling the study of online social networks. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work Companion*, pages 23–27, 2013.

[5] M. E. J. Newman. The structure and function of complex networks. *SIAM Rev.*, 45(2):167–256, 2003.

[6] J. Priem and K. L. Costello. How and why scholars cite on twitter. In *Proceedings of the American Society for Information Science and Technology*, pages 1–4, 2010.

[7] J. Priem, D. Taraborelli, P. Groth, and C. Neylon. altmetrics: a manifesto. `http://altmetrics.org/manifesto/`, 2011.

[8] X. Shuai, A. Pepe, and J. Bollen. How the scientific community reacts to newly submitted preprints: Article downloads, twitter mentions, and citations. *PLoS One*, 7(11):e47523, 2012.

[9] C. H. Yeong and B. J. J. Abdullah. Altmetrics: the right step forward. *Biomed. Imag. Interv. J.*, 8(3):e15, 2012.