Altmetrics in the wild: An exploratory study of impact metrics based on social media

(presentation)

Jason Priem, Heather A. Piwowar, and Bradley H. Hemminger

Introduction

Though the true meaning of citations remains tendentious (Bornmann & Daniel, 2008), scientometricians continue to use them because they at least partly trace the *use* of scholarly products. Unfortunately, however, citations reflect only limited spectrum of uses, leaving others invisible.

Today, as growing numbers of scholars publicly read, bookmark, share, discuss, and rate using online tools, these "invisible impacts" are beginning to be seen. They are leaving traces on Web pages (Vaughan & Hysen, 2002; Thelwall, Vaughan, & Björneborn, 2005), in downloads (Brody, Harnad, & Carr, 2006; Bollen, Van de Sompel, Hagberg, & Chute, 2009), on blogs (Groth & Gurney, 2010), and on social media (Priem & Hemminger, 2010; Jiang, He, & Ni, 2011). Because measurements of these new traces may inform alternatives to traditional citation metrics, they been dubbed "altmetrics" (Priem, Taraborelli, Groth, & Neylon, 2010).

The sources cited above have claimed a number of advantages for altmetrics: they are faster than traditional citation measures, can measure different types of scholarly products, measure impact on non-scholarly audiences, and most importantly, they support a broader, more nuanced understanding of impact, a "bibliometric spectroscopy" (Cronin, 2001). However, despite a growing empirical base (Priem & Costello, 2010; Shema & Bar-Ilan, 2011; Uren & Dadzie, 2011) altmetrics based on social media are not yet well understood.

Questions

The goal of this study is to better understand the potential of altmetrics. We asked two main research questions:

1. How much and what kind of altmetrics data are out there?
    1.1. Which altmetrics sources generate enough data to be useful?
    1.2. Is the amount of altmetrics data growing or shrinking?
    1.3. How rich is the altmetrics data?
2. What do altmetrics measure?
    2.1. How do they correlate with one another and with citation metrics?
    2.2. Can we predict citation counts with altmetrics counts?

Methods

To answer these questions, we gathered altmetrics for a large sample of scholarly articles— all 24,334 articles published by the Public Library of Science (PLoS) before December 23, 2010. They came from seven journals representing a broad variety of scientific periodicals, ranging in Impact Factor from 4.4 to 12.9, and publishing between 623 and 14,102 articles apiece.

We used altmetrics and citation data supplied by PLoS, gathered from web APIs, and manually

downloaded as datasets to compile a list of altmetrics *events* (uses, bookmarks, etc.) for each PLoS article The resulting dataset contained approximately 1.8 million rows, recording events from these sources:

- citation in Web of Science, CrossRef, PubMed Central, and Scopus
- scholarly bookmarking/reference-management services Mendeley and CiteULike
- blogs as tracked by Postgenomic, Nature Blogs, and Research Blogging
- popular Web 2.0 services: Facebook, Twitter, and Delicious
- citations from Wikipedia
- comments from within the PLoS platform, as well as monthly PDF and html view counts.
- ratings by Faculty of 1000

To answer our questions, we had to separate a signal (people's usage of an article in variety of ways) from two kinds of noise: the trends in global adoption of different services, and the propensity for older articles to gather more citation/usage events of all types. Consequently, we turned to a signal processing technique—use of a Hamming window (Hamming, 1989)—to normalize data. For n events of type e on article published at time t, we first get the weighted mean count $n_{expected}$ for articles published within six months of t. This is number of events we'd expect for t, given its publication date. Then $n_{normalized}$ is simply $n_{actual} / n_{expected}$. In other words, we normalized a metric value for given article by dividing it by the average metric value of all articles published at about the same time, weighted such that articles published further away in time contributed less to the average. Since normalized event counts were highly skewed, we then log-transformed them; the resulting distributions of metric counts (excluding zeros) are summarized in Fig. 1.

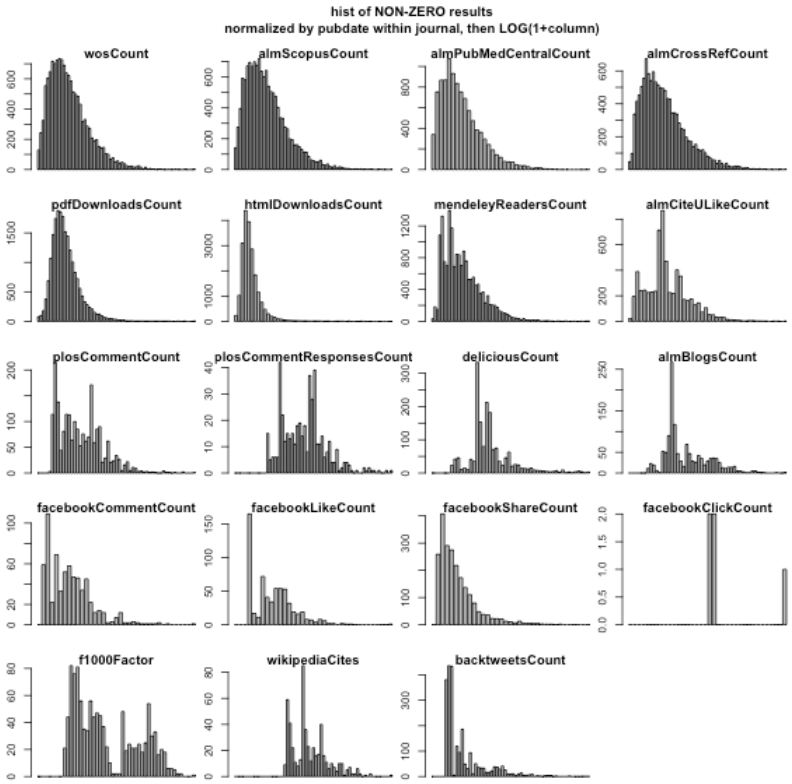

Figure 1

Results

RQ1.1: We found that some metrics were much more sparse than others (Fig. 2), but most event types showed activity on a reasonable percentage of articles.

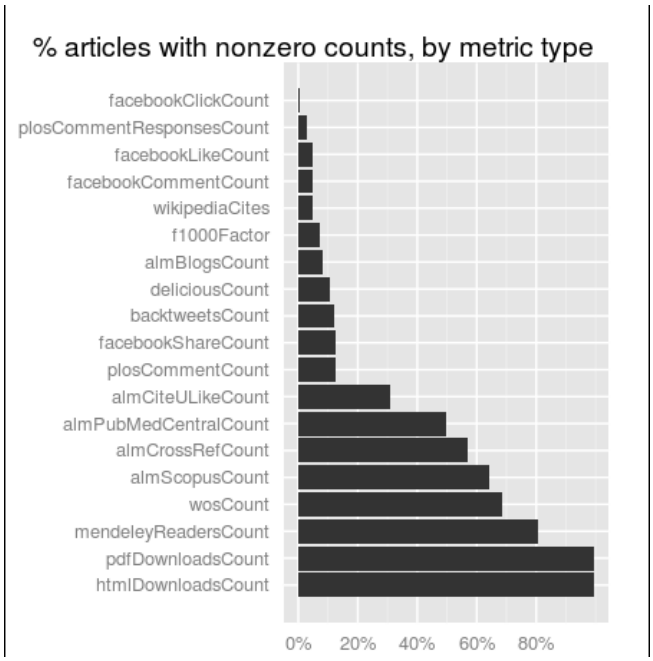

Figure 2

RQ1.2 The amount of use of a particular altmetrics source seemed to vary dramatically between communities and over time, likely reflecting differing community norms and the volatility of early-adopters' interest. Figures 3 and 4 give useful examples of this.

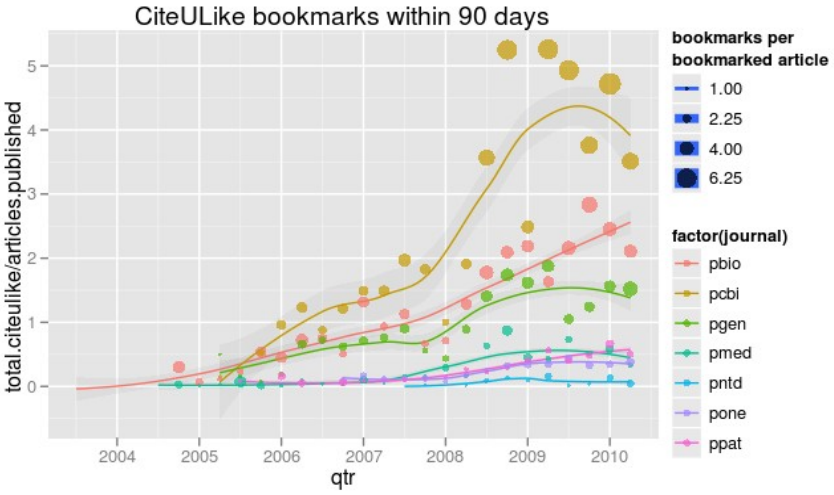

Figure 3

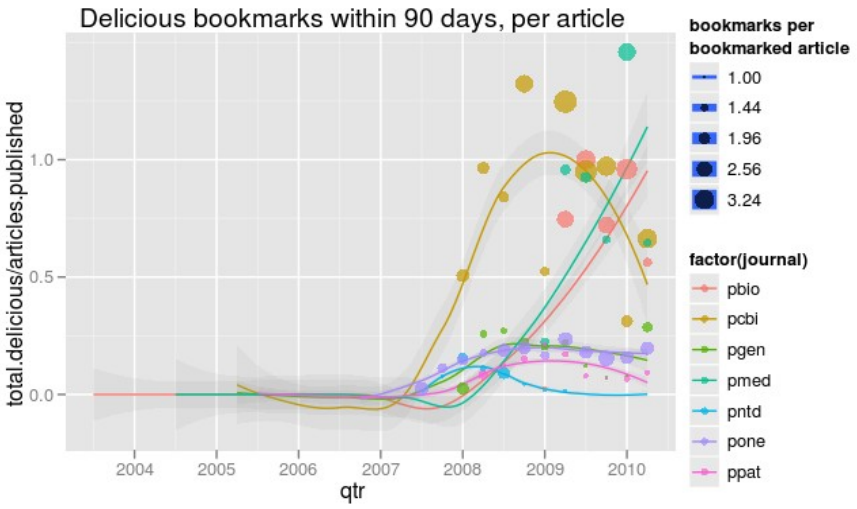

Figure 4

RQ1.3: The current study did not include analysis of supplemental altmetrics data like blog and tweet texts, bookmark tags, and user locations. These deeper analyses are likely to be fruitful: preliminary exploration into the contrasts between publisher- and user-supplied tags (Fig. 5) suggests that altmetrics can supply valuable additional data about articles.

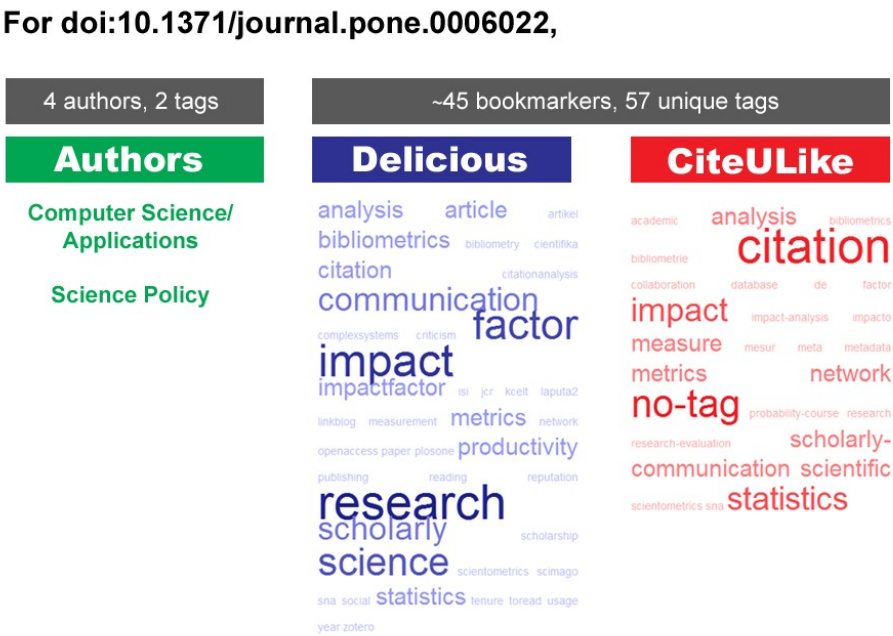

Figure 5

RQ2.1: Correlations between metric types are shown in Fig. 6; the overall matrix is similar to (Yan & Gerstein, 2011), which uses non-normalized counts. The scholarly bookmarking services Mendeley (r=.26) and CiteULike (r=.16) were somewhat correlated with citation, while general bookmarking services like Delicious were not. Fig. 7 shows the results of an exploratory factor analysis.

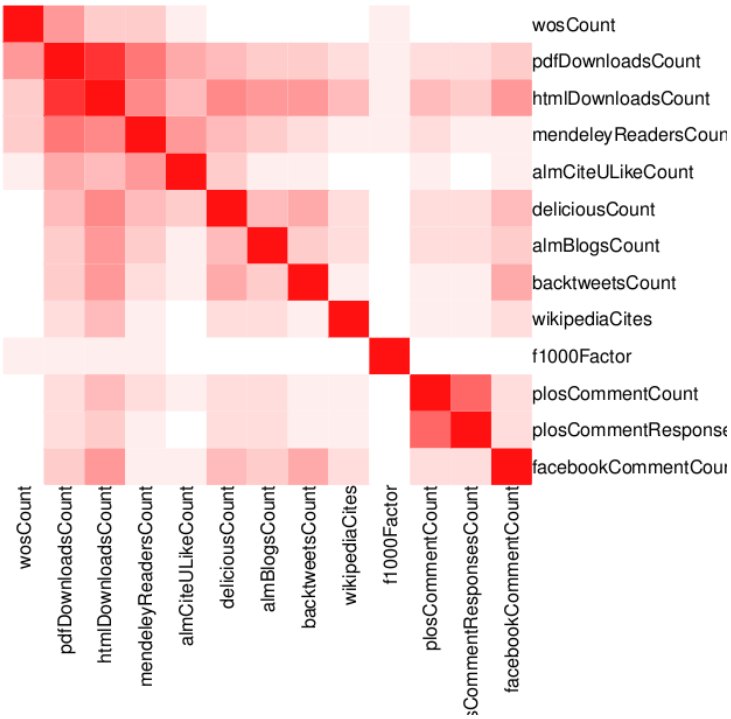

Figure 6

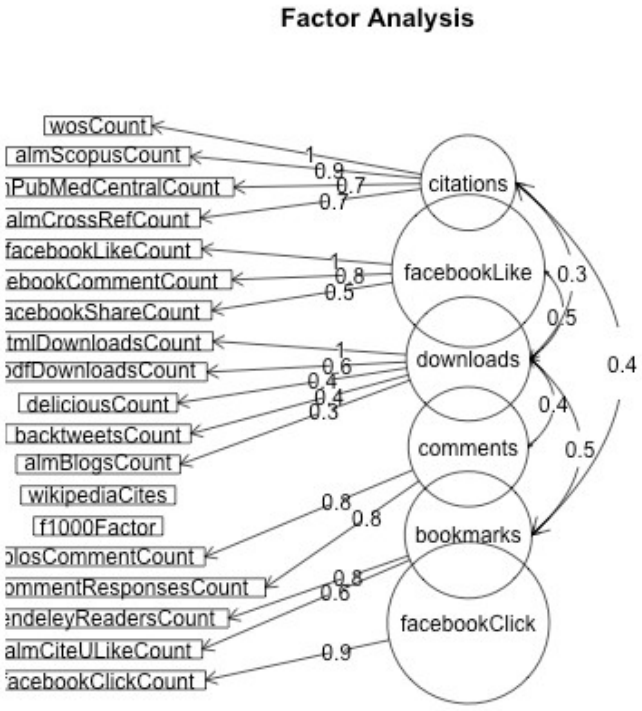

Figure 7

RQ2.2: We used linear regression to understand whether altmetric counts contributed new information to the prediction of Web of Science citation counts. Altmetrics were found to improve the model fit, as shown in Table 1 (differences are significant at the 0.001 level).

Table 1: Linear regression models

Model	R^2
base (article age + journal + authors count)	.07
base +usage (html, pdf downloads)	.42
base + usage + altmetrics	.46
base + altmetrics	.19

It would be useful if we could use altmetrics to predict *future* citation. Additional data and analyses are needed to understand the feasibility of this. Fig. 8 shows that in at least some cases, certain event types seem to presage others in a way that might inform future automated prediction algorithms.

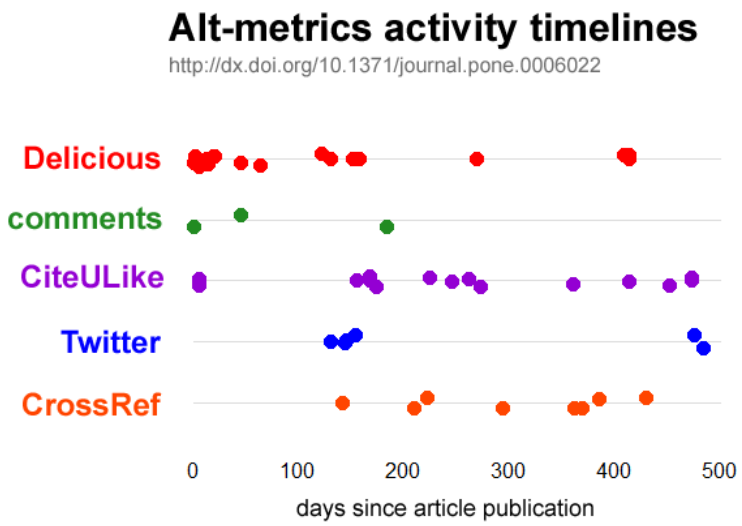

Figure 8

Discussion

Our results suggest that there are enough altmetrics events to support robust research into usable impact measures. We also find that social media altmetrics are a significant contributor to citation prediction, which to our best knowledge has never before been demonstrated.

Differences in citation correlation between scholarly and general services suggest altmetrics can gather impact on varied audiences, while the low correlation with citation suggests that altmetrics captures a sort of impact mostly orthogonal to that reflected in citations.

Does this kind of impact matter? Future research must examine scholars' use of social media tools to gain a better picture of what a Twitter citation or a CiteULike bookmark, for instance, really means. We should also extend this research to a more diverse sample (since PLoS, as open-access publisher, may be exceptional), and build on the potential of altmetrics for prediction and recommendation.

Full datasets, event crawler code, and analysis code are available at https://github.com/jasonpriem/plos_altmetrics_study

References

Bollen, J., Van de Sompel, H., Hagberg, A., & Chute, R. (2009). A principal component analysis of 39 scientific impact measures. *PloS one*, *4*(6), e6022.

Bornmann, L., & Daniel, H. D. (2008). What do citation counts measure? A review of studies on citing behavior. *Journal of Documentation*, *64*(1), 45.

Brody, T., Harnad, S., & Carr, L. (2006). Earlier Web usage statistics as predictors of later citation impact. *Journal of the American Society for Information Science and Technology*, *57*(8), 1060-1072. doi:10.1002/asi.20373

Cronin, B. (2001). Bibliometrics and beyond: some thoughts on web-based citation analysis. *Journal of Information Science*, *27*(1), 1. doi:10.1177/016555150102700101

Groth, P., & Gurney, T. (2010). Studying Scientific Discourse on the Web using Bibliometrics: A Chemistry Blogging Case Study. *Proceedings of WebSci10: Extending the Frontiers of Society*, Raleigh, NC: US. Retrieved from http://journal.webscience.org/308/.

Hamming, R. W. (1989). *Digital filters*. Englewood Cliffs, NJ: Prentice-Hall.

Jiang, J., He, D., & Ni, C. (2011). Social reference: aggregating online usage of scientific literature in CiteULike for clustering academic resources. *Proceeding of the 11th annual international ACM/IEEE joint conference on Digital libraries*, JCDL '11 (pp. 401–402). Ottawa, Ontario, Canada: ACM. doi:10.1145/1998076.1998155

Priem, J., & Costello, K. L. (2010). How and why scholars cite on Twitter. *Proceedings of the 73rd ASIS&T Annual Meeting*. Presented at the American Society for Information Science & Technology Annual Meeting, Pittsburgh PA, USA. doi:10.1002/meet.14504701201

Priem, J., & Hemminger, B. H. (2010). Scientometrics 2.0: Toward new metrics of scholarly impact on the social Web. *First Monday*, *15*(7). Retrieved from http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/2874/2570

Priem, J., Taraborelli, D., Groth, P., & Neylon, C. (2010). alt-metrics: a manifesto. Retrieved August 15, 2011, from http://altmetrics.org/manifesto/

Shema, H., & Bar-Ilan, J. (2011). Characteristics of Researchblogging.org science Blogs and Bloggers. Presented at the altmetrics11: Tracking scholarly impact on the social Web (An ACM Web Science Conference 2011 workshop), Koblenz, Germany. Retrieved from http://altmetrics.org/workshop2011/shema-v0/

Thelwall, M., Vaughan, L., & Björneborn, L. (2005). Webometrics. *Annual Review of Information Science and Technology*, *39*(1).

Uren, V., & Dadzie, A.-S. (2011). Relative Trends in Scientific Terms on Twitter. Presented at the altmetrics11: Tracking scholarly impact on the social Web (An ACM Web Science Conference 2011 workshop), Koblenz, Germany. Retrieved from http://altmetrics.org/workshop2011/uren-v0/

Vaughan, L., & Hysen, K. (2002). Relationship between links to journal Web sites and impact factors. *Aslib Proceedings*, *54*(6), 356-361.

Yan, K.-K., & Gerstein, M. (2011). The Spread of Scientific Information: Insights from the Web Usage Statistics in PLoS Article-Level Metrics. *PLoS ONE*, *6*(5), e19917. doi:10.1371/journal.pone.0019917