Co-Saved, Co-Tweeted, and Co-Cited Networks

Fereshteh Didegah
Danish Centre for Studies in Research & Research Policy, Department of Political Science & Government, Aarhus University, Aarhus, Denmark

Scholarly Communication Lab, Simon Fraser University, Vancouver, BCCanada. E-mail: fdidegah@sfu.ca

Mike Thelwall
Statistical Cybermetrics Research Group, University of Wolverhampton, Wulfruna Street, Wolverhampton WV1 1LY, UK. E-mail: m.thelwall@wlv.ac.uk

Counts of tweets and Mendeley user libraries have been proposed as altmetric alternatives to citation counts for the impact assessment of articles. Although both have been investigated to discover whether they correlate with article citations, it is not known whether users tend to tweet or save (in Mendeley) the same kinds of articles that they cite. In response, this article compares pairs of articles that are tweeted, saved to a Mendeley library, or cited by the same user, but possibly a different user for each source. The study analyzes 1,131,318 articles published in 2012, with minimum tweeted (10), saved to Mendeley (100), and cited (10) thresholds. The results show surprisingly minor overall overlaps between the three phenomena. The importance of journals for Twitter and the presence of many bots at different levels of activity suggest that this site has little value for impact altmetrics. The moderate differences between patterns of saving and citation suggest that Mendeley can be used for some types of impact assessments, but sensitivity is needed for underlying differences.

Introduction

Citation counts are widely used in research evaluation but have been criticized for being slow and reflecting only scientific impact, and not broader research impacts. Altmetrics are alternative indicators derived from social websites that may give quicker impact evidence and help assess the contributions of science to society beyond the scientific community (Priem, Taraborelli, Groth, & Neylon, 2011). Two prominent altmetrics for articles are counts of users in the social bookmarking site Mendeley and the number of tweets linking to them on Twitter. Previous altmetrics research has mainly investigated whether different altmetrics are related to citation counts, finding moderate correlations between citation counts and number of Mendeley users that had saved the article to their library (Mohammadi & Thelwall, 2014) and very weak correlations between citation counts and tweet counts (Haustein, Larivière, Thelwall, Amyot, & Peters, 2014; see also a recent review: Sugimoto, Work, Larivière, & Haustein, (2017)). Other studies have examined whether the factors that influence altmetrics differ from those that influence citations (Didegah, Bowman, & Holmberg, 2017; Haustein, Costas, & Lariviére, 2015) and why and how authors cite/save on the social web (for example, Mendeley: Mohammadi, Thelwall, & Kousha, 2016; Twitter: Priem, & Costello, 2010; Thelwall, Tsou, Weingart, Holmberg, & Haustein, 2013). Nevertheless, no study has compared citations with altmetrics at the level of networks of articles cited, tweeted, or saved by individual users to look for deeper insights into patterns of commonality and difference. This is a novel approach and not one of the standard analysis strategies for evaluating altmetrics (Sud & Thelwall, 2014). It is a useful additional technique because it can point to systematic network-level differences that may not be evident from examining individual articles. For example, fields or journals that are unusual in respect of being tweeted by disparate communities might be revealed by network metrics.

The current study explores the relationship between citations and altmetrics at the network level by comparing co-saved (to a Mendeley library), co-tweeted, and co-citation networks. A pair of documents is *co-cited* when another document simultaneously cites them both (for example, see: Boyack & Klavans, 2010; Small, 1973). In studies of

Received February 1, 2017; revised December 12, 2017; accepted February 4, 2018

© 2018 ASIS&T • Published online May 14, 2018 in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/asi.24028

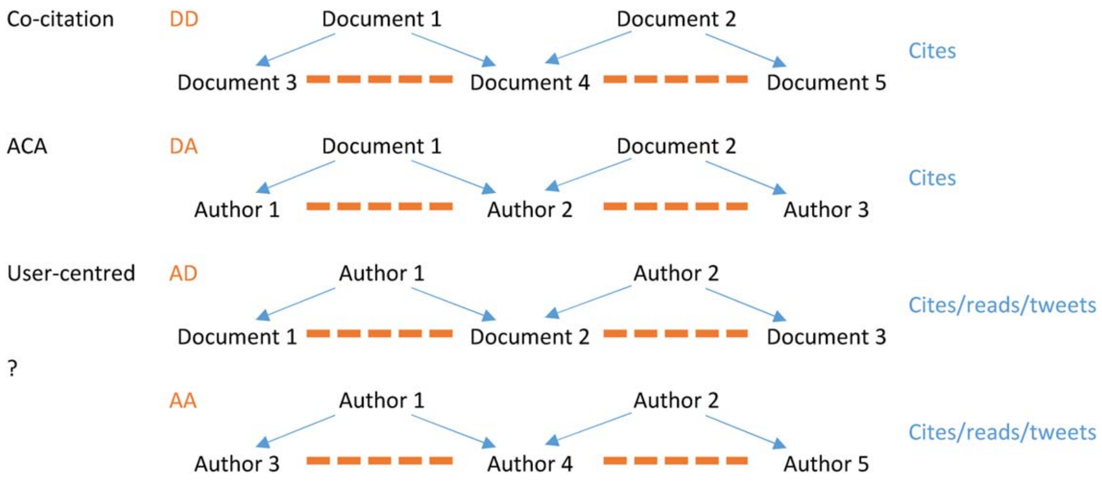

FIG. 1. DD: Traditional (document-document) co-citation network. A pair of documents is co-cited if a third document simultaneously cites both. The DD co-citation network generated by these data would have the two connections shown by the dotted lines: Document1: Document2 and Document2: Document3. DA: Traditional (document-author) author co-citation network. A pair of authors is co-cited if a document simultaneously cites both. The DA co-citation network generated by these data would have the two connections shown by the dotted lines: Author1:Author2 and Autor2:Author3. AD: (author-document) Author-based co-citation network. A pair of documents is co-cited if an author cites both (not necessarily in the same citing source document). The AD co-citation network generated by these data would have the two connections shown by the dotted lines: Document1: Document2 and Document2: Document3. AA: (author-author) Author-based author co-citation network. A pair of authors is co-cited if an author cites both (not necessarily in the same citing source document). The AA co-citation network generated by these data would have the two connections shown by the dotted lines: Author1: Author2 and Author2: Author3. [Color figure can be viewed at wileyonlinelibrary.com]

academic field structure and development (Small, 1981), co-citation and co-word connections have been widely used. In document co-citation analysis, articles that are cited together by a document are more likely to be semantically related than a random pair of documents, and the more often they are cited together, the stronger their relationship is likely to be (Small, 1973). Co-citations can therefore help to detect subject similarity (Small, 1973; Marshakova, 1973) and to investigate the intellectual structure of fields (White & McCain, 1998). Co-bookmarking of journals, counting the number of times that two journals are saved to a reference manager by the same user, has also been suggested for detecting similar journals (Haustein, 2012). Being co-saved is a weaker indicator of subject similarity than co-citation, however (Kraker et al., 2015). For Mendeley, the terms bookmarking, saving (to a library), and reading are all reasonable descriptions since most Mendeley users add articles to their library to read them, or after reading them (Mendeley: Mohammadi, Thelwall, & Kousha, 2016).

Author Co-citation Analysis (ACA) estimates the intellectual structure of a field based on co-citation relationships between the authors of the documents in that field (White & McCaine, 1998). ACA assumes that pairs of authors who are frequently cited by the same documents are more likely to produce semantically related research. No previous studies have analyzed author-document co-citation networks (that is, user-centered networks) as used here.

Co-citation, co-saved, and co-tweeted networks are all forms of user-centered document co-citation networks (Figure 1), with the actors being authors, Mendeley users, and tweeters. The difference with a standard co-citation network is that a pair of documents is (author/user) co-cited if the same author/user cites both documents. Thus, for co-citation networks, the citations could be from the same or different articles by one author. For Twitter, the co-citations could be from different tweets by the same tweeter. For Mendeley, the co-citations must be in the main library of a single user. Although co-citation networks have been previously studied for altmetrics, they have been traditional document co-citation networks rather than author/user-based co-citation networks (Jung, Lee, & Song, 2016). Author-based co-citation networks with citations are compared with altmetric equivalents to identify differences between them.

Mendeley co-saved networks connect pairs of articles when they are in the same user library. Co-saved networks of articles have previously been created to visualize the field of educational technology (Kraker, Schlögl, Jack, & Lindstaedt, 2015). Articles were included if they were in at least 16 user libraries associated with educational technology, giving 91 articles in total. A co-saved network has been used to identify the knowledge domain of an emerging field, Technology-Enhanced Learning (Kraker, Körner, Jack, & Granitzer, 2012). The co-saved networks gave useful insights into this field. A Mendeley co-saved network of disciplines with associated articles and reviews from 2012 has also been created to visualize all disciplines and subdisciplines of science (Bornmann & Haunschild, 2016). For this, each article included in the libraries of users associated with two different subdisciplines counted as a connection between the two subdisciplines. Both studies therefore attempted to visualize subject domains with Mendeley data, one using articles and the other using subdisciplines.

From the user perspective, a pair of articles in a user co-tweeted network is connected when they are both tweeted by the same tweeter. An author-document co-tweeted network has previously been created for the *Journal of the Association for Information Science and Technology* (JASIST) for articles published between 2001 and 2014, using tweets from 2007 to 2014 and compared with a (document-document) co-citation network (Jung, Lee, & Song, 2016), but this is not a like-for-like comparison because the networks are of different types and is also not optimal because Twitter is a real-time medium, so older articles are much less likely to be tweeted than newer ones.

Research Questions

This study compares citations, Mendeley saves, and tweets for research articles at the network level to investigate network-level differences between them. As reviewed above, positive correlations have been found between citation counts and number of tweets and Mendeley saves, but how these differ from each other at the network level is unknown. For example, if two articles are cited by the same author, are they more likely to be found in the same Mendeley library than tweeted from the same account? Is the co-citation network of journals/fields more similar to the co-saved network of journals/fields than the co-tweeted network? The weaker correlations with citations for Twitter than for Mendeley suggests that there may be substantial differences between the two. This study investigates how authors, Mendeley users, and tweeters select articles to cite/read/tweet by assessing the extent to which pairs of articles that are cited by the same person are also read or tweeted by a single person (Figure 2), but not necessarily the same person. Any differences found can point to underlying patterns of behavior that are unique to one of these sources and can help when interpreting altmetrics. As well as focusing on the overlap between the three phenomena, this study investigates the role of journals, since these publish articles, and fields, since articles tend to contribute to individual fields. Twitter bots are also examined since these can influence the answers to the main questions (for arXiv preprints: Haustein, Bowman, Holmberg, Tsou, Sugimoto, & Larivière, 2016). The research questions focus on articles published in a single year because articles tend to be tweeted as soon as they are published, whereas reference lists can contain old articles, so huge differences between networks would be inevitable without this restriction.

- Q1: To what extent do the altmetric networks overlap with the citation network? More specifically, how many pairs of articles from the same year that are cited by the same person (that is, co-cited) are also in a single Mendeley user library (that is, co-saved) or have also been tweeted by the same person (that is, co-tweeted)?
- Q2: How does the importance of journals differ between networks? More specifically, how do the three different networks (that is, co-cited, co-saved, co-tweeted) differ in terms of the percentage of connected pairs of articles being from the same journal?

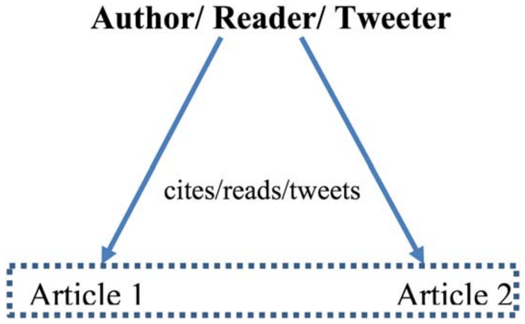

FIG. 2. Co-citation, co-saved and co-tweeted connections. [Color figure can be viewed at wileyonlinelibrary.com]

- Q3: How does the importance of narrow fields differ between networks? More specifically, how do the three different networks (that is, co-cited, co-saved, co-tweeted) differ in terms of the percentage of connected pairs of articles being from the same narrow field?
- Q4: To what extent are the results for Twitter affected by bots?

Questions 2 and 3 were investigated for all pairs of articles as well as for subsets consisting of highly connected pairs of articles to seek deeper patterns of use.

Methods

Data Collection

This study is based on all articles from the Web of Science (WoS) published in 2012. The year 2012 was selected to give enough time for articles to attract citations without being too old for the articles to attract many tweets and Mendeley users. Publication and citation data were obtained from the in-house WoS database of the Centre for Science and Technology Studies (CWTS), Leiden University in July 2016. Data were obtained from Mendeley in June 2016 and Twitter data were retrieved from Altmetric.com in June 2016. To get the save and tweet counts from Mendeley and Twitter, articles must have DOIs, and so articles without DOIs were discarded. Out of 1,320,205 articles published in 2012, 1,131,318 (85.69%) had a DOI. Even though DOIs are not necessary for the citation data, the sample of articles with DOIs was also used for the co-citation network for comparison purposes.

Out of 1,131,318 articles with a DOI, 994,738 (87.92%) had received at least one citation, with 5.88 citing authors (rather than citing publications) per article (5,849,858 citing authors in total). Corresponding records were identified in Mendeley for the 1,131,318 articles published in 2012 using a DOI search in the site. From this, 958,449 (84.71%) of the articles had at least one Mendeley user, with an average of 20.96 users each (20,098,335 article users in total).

Corresponding records were also identified in Twitter for the 1,131,318 articles published in 2012 using a DOI search in Altmetric.com. From this, 230,764 (20.39%) articles were tweeted at least once, with 4.56 tweeters per article (1,053,595 article tweeters in total). More than 11% of the articles had only one citation, more than 10% had only one tweeter, and less than 6% had only one user.

The three basic datasets were each refined separately to remove articles with few citations, tweets, or users. To find the appropriate cutoff thresholds to refine the datasets, a distribution-based approach was considered. The largest dataset, the distribution of saves per DOI, was plotted first. The distribution turning point suggested 23 saves per DOI but this was too large to be used as a cutoff, since it would generate huge co-occurrence matrices that were not technically manageable for further analyses. The cutoff threshold for each dataset was therefore determined instead by technical feasibility, as described below.

Pairs of Articles Cited by the Same Author (Co-Cited Dataset)

An author-based article co-citation dataset was created from all authors of the articles (as recorded in WoS) that had cited the articles in the 2012 dataset (Figure 2). The citing authors were identified using the CWTS list of disambiguated author names. The method has a high precision (95%) and recall (90%), but is less accurate for common (particularly Asian) names (for methodology, see: Caron and Van Eck, 2014). The co-citation dataset connects pairs of articles when they are cited by the same author, even if the citations are derived from different articles by that author. Articles with at least 10 citations were selected, resulting in 6,813,595 co-citation pairs. The requirement for 10 citations was a technical limitation to make it possible to process the entire dataset. This minimum threshold is the first cutoff used to construct the networks and measure the overlaps between the datasets.

Pairs of Articles Saved by the Same Mendeley User (Co-Saved Dataset)

A pair of articles is co-saved if they were both in the library of the same Mendeley user. Articles with at least 100 users were selected, which gave 19,167,468 unique co-saved pairs. Since Mendeley user counts tend to be higher than citation counts, it would not be reasonable to choose the same threshold for both.

Pairs of Articles Tweeted From the Same Account (Co-Tweeted Dataset)

A pair of articles is co-tweeted if both were tweeted from the same user account. Articles that were tweeted at least 10 times were selected, giving 3,156,129 unique co-tweeted pairs.

Field Classification

To investigate fields within the co-networks, all articles were classified into one of 817 narrow (also called meso-level) fields using CWTS article-level classifications. This scheme is derived from an algorithm that merges articles into groups based upon the citations between them and then labels the groups using terms that are common in the titles or abstracts of the articles (Waltman & Van Eck, 2012). Each narrow field is assigned up to five labels, so it is difficult to visualize them in the graphs. To create readable graphs, the 817 narrow fields were mapped into the following five broad fields proposed by CWTS. The co-network graphs based on the subject fields were built from narrow fields to see more detailed network interactions but the graph nodes are colored based on the five broad fields for greater readability:

1. Social Sciences and Humanities (SSH);
2. Biomedical and Health Sciences (BHS);
3. Physical Sciences and Engineering (PSE);
4. Life and Earth Sciences (LES);
5. Mathematics and Computer Science (MCS).

Since the CWTS fields are defined at the article level, they are more fine-grained than WoS subject categories or any other journal-level classifications.

Raw counts were used in all three analyses. This gives an advantage to articles in fields with long reference lists (citation network), a high proportion of Mendeley users (saving network), or a high proportion of Twitter users (Twitter network). These articles, and their containing journals, are more likely to obtain high values in the networks. Larger journals are also more likely to obtain high values. It would be possible to field normalize the data, for example by calculating the average co-citation/read/tweeted counts per field and dividing each score by the field average (for example, analogous to the Mean Normalized Citation Score: Waltman, van Eck, van Leeuwen, Visser, & van Raan, 2011). For journals, average scores per article could be used instead of total scores. These were not done for the current article because for a first analysis it is more transparent to focus on raw, untransformed data.

TABLE 1. Article pairs in different networks by journal, for WoS articles from 2012 with 10 citations, 100 readers and 10 tweets, respectively.

	All article pairs in the dataset			Highly co-occurring article pairs		
	Co-cited	Co-read	Co-tweeted	Co-cited	Co-read	Co-tweeted
Same journal	1.5%	4.6%	10.0%	18.6%	15.8%	22.0%
Different journals	98.5%	95.4%	90.0%	81.4%	84.2%	78.0%

Network Clustering

Networks were drawn in Gephi after aggregating all articles into their publishing journals and narrow fields. Different layouts were tested to draw the networks and Fruchterman Reingold was found to produce the clearest visualizations. The networks were clustered using a community

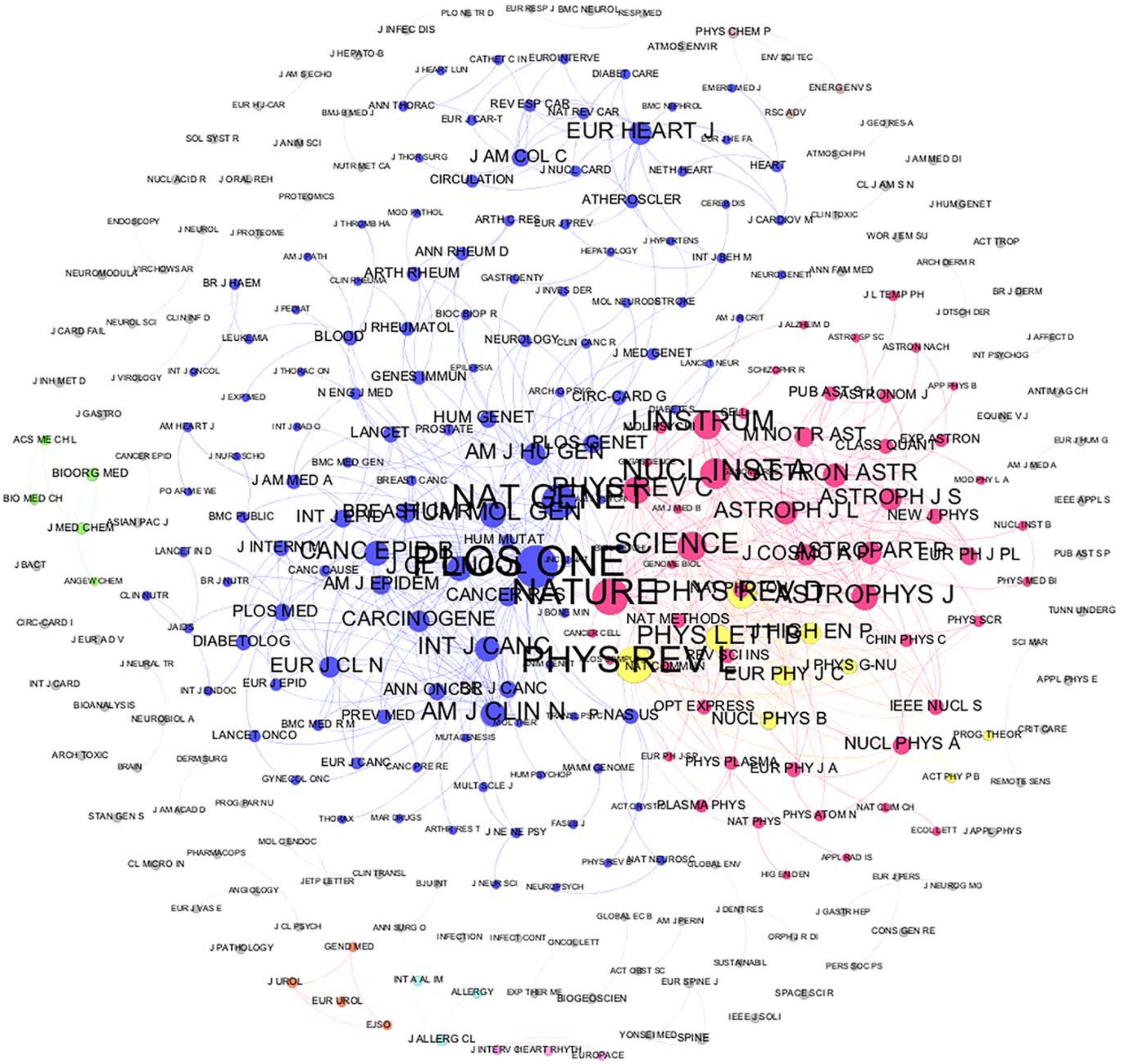

FIG. 3. Highly co-cited journals. Node size indicates the number of connections with other nodes in the network. Node color indicates cluster membership (60 different clusters were detected using the modularity algorithm but 90% of the nodes have been grouped into three main clusters, colored blue, red, and yellow; nodes in other clusters are too small to see colors). Lines connect nodes that are highly co-cited. Data: WoS articles from 2012 with 10 citations; co-citation relationships with strength at least 10. [Color figure can be viewed at wileyonlinelibrary.com]

TABLE 2. Some journals that are highly connected in each network.

Co-citations with articles in other journals		Co-read with articles in other journals		Co-tweeted with articles in other journals	
Journal	% of all co-citations	Journal	% of all co-reads	Journal	% of all co-tweets
Physics Letters B	139,902 (47%)	Science	20,358 (6%)	Nature	6,766 (4.4%)
Physical Review D	46,585 (15.6%)	Nature	38,385 (11.2%)	*PLoS One*	15,995 (10.4%)
Journal of High Energy Physics	43,012 (14.5%)	PLoS One	9,063 (2.6%)	PNAS	5,184 (3.3%)
		PNAS	11,640 (3.4%)		

detection technique, hierarchical optimization of modularity (Blondel, Guillaume, Lambiotte, & Lefebvre, 2008), which groups together nodes (journals/narrow fields) that tend to connect to each other but not to connect to others. Clusters were illustrated with different colors. The standard modularity network statistic was calculated for each network. This

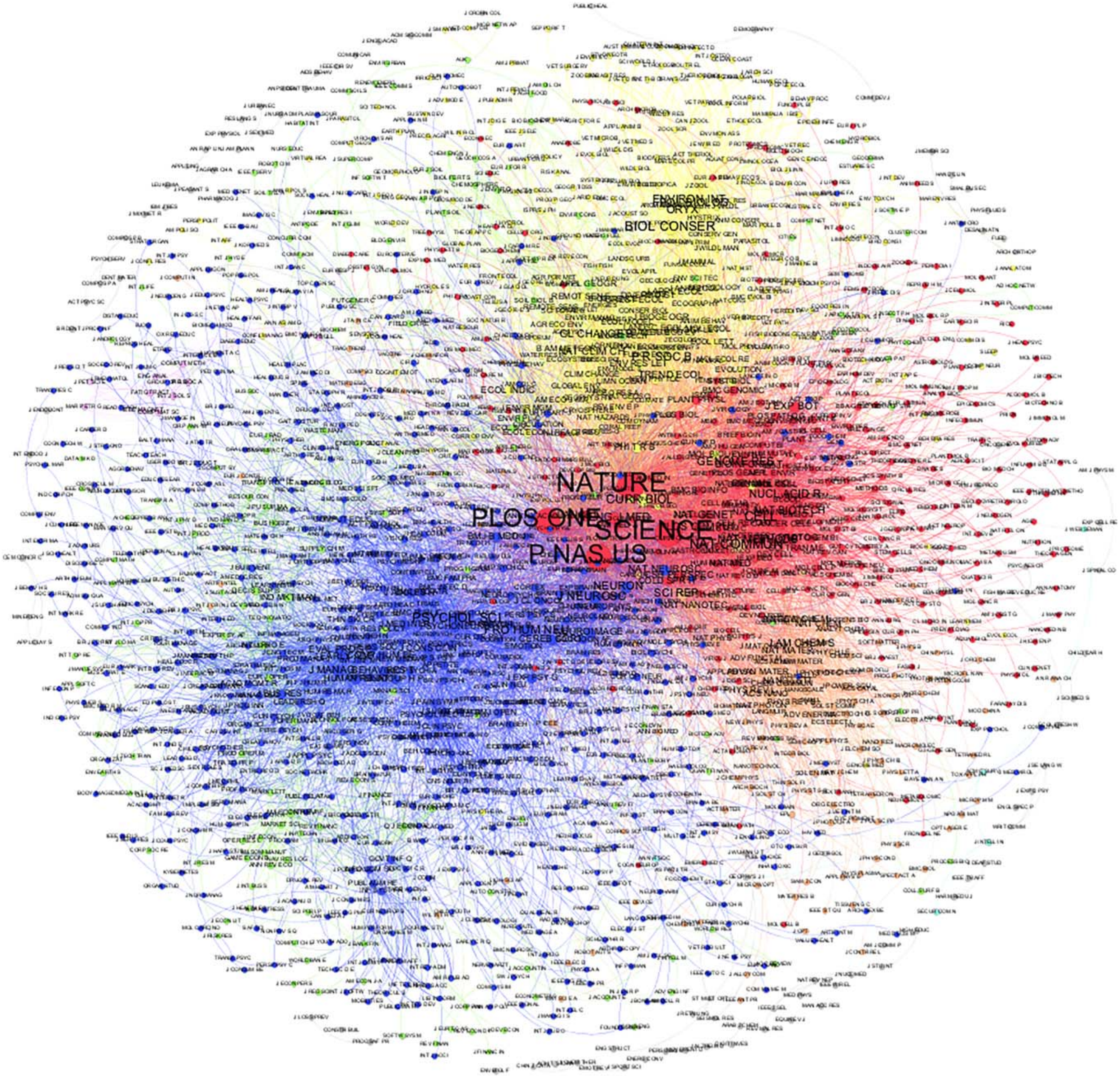

FIG. 4. Highly co-saved journals. (Using the modularity algorithm, 65 clusters were detected. The distribution of nodes across clusters is more balanced than the co-cited network. The top six clusters include 95% of the nodes). Data: WoS articles from 2012 with 100 Mendeley users; co-saved relationships with strength at least 100. [Color figure can be viewed at wileyonlinelibrary.com]

reports the proportion of connections within a cluster, minus the expected number of connections that would fall within the cluster if the edges were distributed at random (Newman & Girvan, 2004).

To create readable graphs, the samples had a second cutoff threshold: for high-co-occurrence, in addition to the existing citation (10), saving (100), and tweet (10) thresholds. The co-networks were limited to article pairs that were cited together at least 10 times (highly co-cited articles), read together at least 100 times (highly co-saved articles), or tweeted together at least 10 times (highly co-tweeted articles).

Bot Identification

Twitter bots are computer programs that automatically tweet content following a predefined set of rules. They are sometimes difficult to identify, particularly when content is posted from accounts by both humans and bots. The “Bot or Not?”[1] application reports the probability of a tweeter being a bot or human, but is not reliable for accounts that tweet research (Haustein et al., 2016). Manual checking was therefore needed but was only possible

[1] http://truthy.indiana.edu/botornot/

DOI: 10.1002/asi

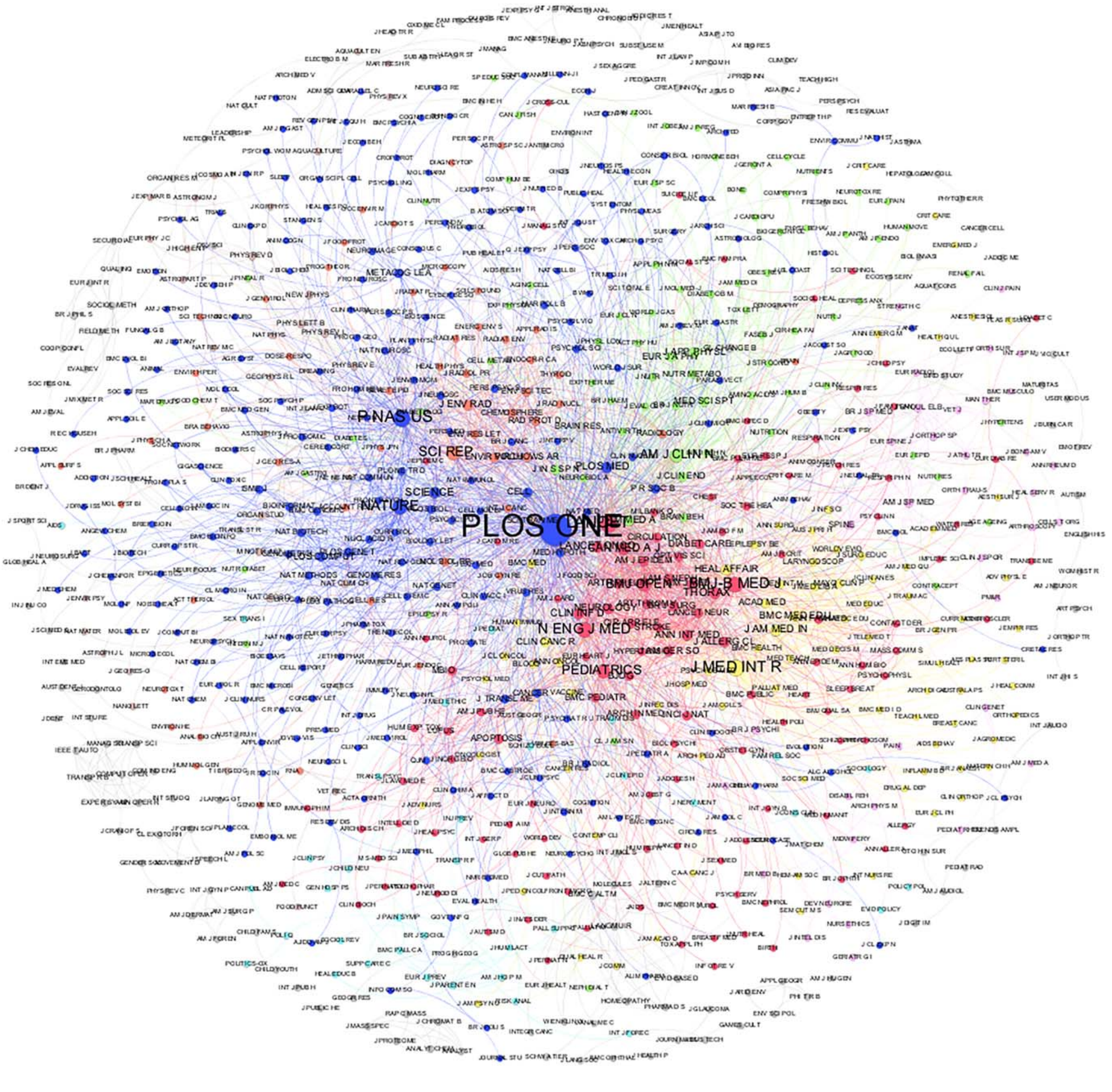

FIG. 5. Highly co-tweeted journals. (The modularity algorithm detected 75 clusters with 12 clusters having more than 1% of the nodes. The nodes are more fairly distributed across the clusters in this network. The top five clusters include more than 65% of the nodes in the network). Data: WoS articles from 2012 with 10 tweeters; co-tweeted relationships with strength at least 10. [Color figure can be viewed at wileyonlinelibrary.com]

for a small subset of the accounts. To ensure maximum power, and because bots can be prolific, the 800 accounts with the most tweets were checked (tweeting 42 to 1,236 distinct DOIs). A second sample of 100 accounts that had tweeted between 10 to 34 DOIs (moderate tweeters) and a sample of 100 accounts that had tweeted between 2 to 7 DOIs (occasional tweeters) was selected to assess the prevalence of bots in other account strata. To identify the bot accounts, the method of Haustein et al. (2016) was applied with additional criteria: checking the frequency of tweets posted, the photo of tweeter, other photos taken and posted, types and content of interactions (replies or comments) with others on her/his own tweets and retweets.

Results

RQ1: Overlaps Between Co-Citation, Co-Saving, and Co-Tweeting

After restricting the three raw datasets to articles with 10 citations, 100 Mendeley users, and 10 tweets, as discussed above, co-cited pairs of articles were compared with

TABLE 3. Article pairs in different networks by narrow field.

	All article pairs in the dataset			Highly co-occurring article pairs		
	Co-cited	Co-read	Co-tweeted	Co-cited	Co-read	Co-tweeted
Same narrow field	4.4%	7.6%	6.9%	72.3%	43.5%	21.3%
Different narrow fields	95.6%	92.4%	93.1%	27.7%	56.5%	78.7%

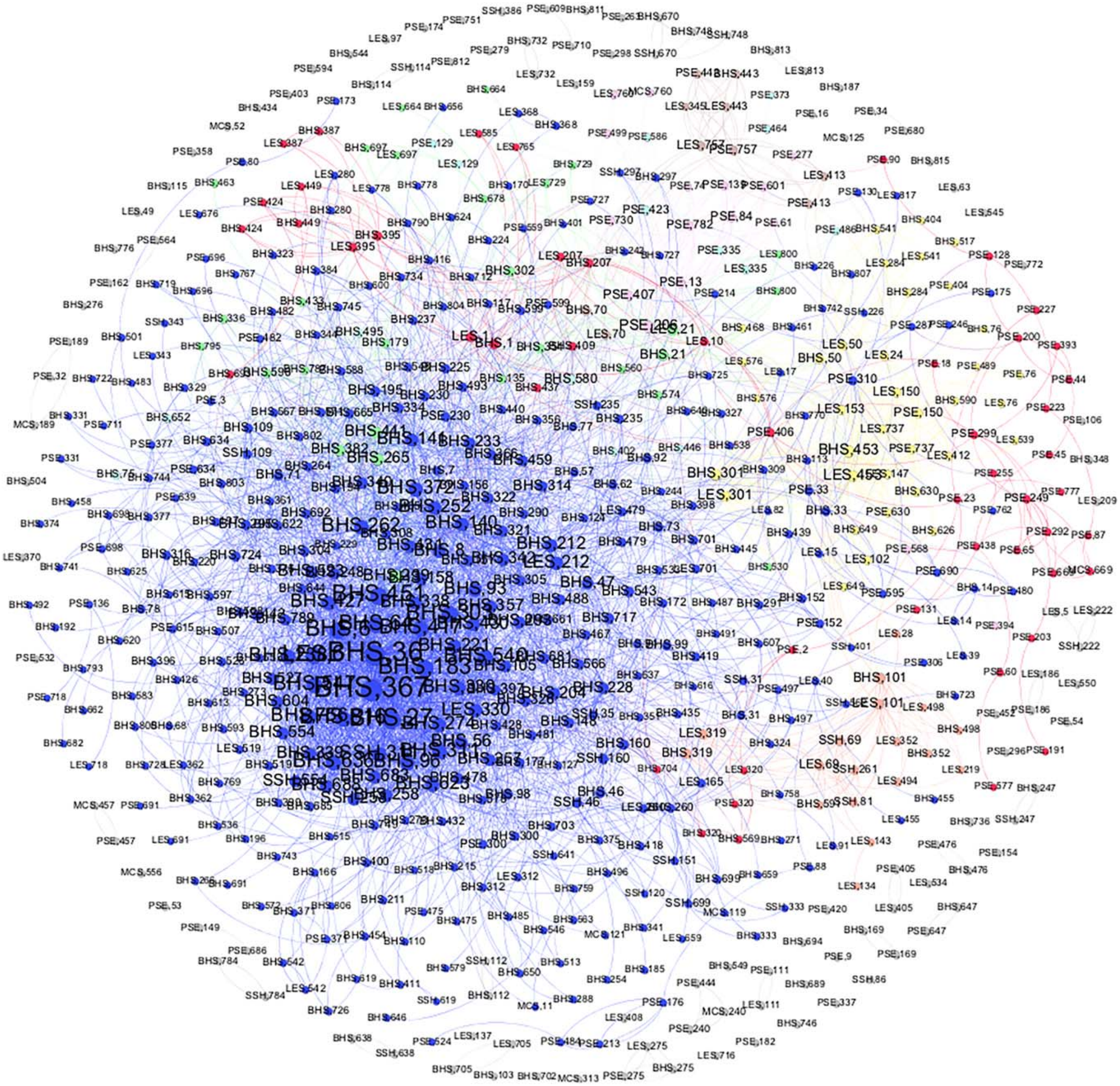

FIG. 6. Co-cited narrow field network. Nodes are labeled with broad field abbreviation and narrow field code (broad, narrow). Node size indicates the number of connections with other nodes in the network. Node color indicates cluster membership (The modularity algorithm detected 80 clusters with the top one covering 54% of the nodes in the network, illustrated with blue circles). Lines connect nodes that are highly co-cited. Data: WoS articles from 2012 with 10 citations; co-citation relationships with strength at least 10. [Color figure can be viewed at wileyonlinelibrary.com]

co-saved and co-tweeted pairs to assess whether articles that are co-cited are also co-saved and co-tweeted. Because these three thresholds are different, the most meaningful overlap threshold is the percentage of the smaller dataset because this is less influenced by the differences in sizes between the two datasets.

There was almost no overlap between the three networks at the level of individual articles. Only 4,598 article pairs matched between the co-citation and co-tweeted sets, accounting for 0.06% of the co-citation set and 0.14% of the co-tweeted set. Between the co-citation and co-saved networks only 11,389 pairs matched, accounting for 0.16% of the co-citation set and 0.05% of the co-saved set. Thus, pairs of articles that were frequently cited by the same author were rarely also tweeted by the same tweeter or registered in Mendeley by the same user (even though it is not necessarily the same person for each network). There was a much larger overlap of 428,774 article pairs matching between the co-saved and co-tweeted networks. This accounts for 13.58% of the co-tweeted set and 2.23% of the co-saved set.

RQ2: The Importance of Journals for the Three Datasets

In the sample of 6,813,595 co-cited pairs of articles, only 1.52% were from the same journal. Similarly, only 4.6% of the co-saved article pairs are from the same journal and 10% of co-tweeted article pairs were from the same journal. After limiting the co-cited set to highly co-cited pairs, 18.6% of the highly co-cited articles were from the same journal. Similarly, 15.8% of highly co-saved pairs were from the same journal and 22% of highly co-tweeted pairs were from the same journal (Table 1).

The co-citation network of journals (Figure 3) with 60 clusters of journals has a modularity of 0.34, suggesting that the connections between the journals within a cluster are moderately denser than the connections with journals outside of the cluster. Near the center is a large set of high-impact physics journals (for example, *Physics Letters B, Physical Review D, Journal of High Energy Physics*, and *Physical Review Letters*), and multidisciplinary journals (for example, *Science, PLoS One*, and *Nature*). Physics journals including *Physics Letters B, Physical Review D*, and *Journal of High Energy Physics* have extremely dense connections with some other journals; their connectivity in the network differs from that of multidisciplinary journals because the former are densely connected to fewer journals but the latter are much less densely connected to more journals. For example, *Physics Letters B* is connected with 36 other journals but its articles are co-cited 139,902 times with articles in the other journals; *Physical Review D* is connected with 36 other journals but its articles are co-cited 46,585 times with articles in those journals; *Journal of High Energy Physics* has connections with 43,012 articles in 30 journals. For multidisciplinary journals, *PLoS One* has connections with 99 other journals but its articles are co-cited 1,902 times with articles in these other journals; *Nature* articles are co-cited with only 905 articles from 79 other journals and *Science* is connected with 65 other journals (2,011 times) (Table 2). The prominence of physics is probably due to its large journals, which have a size advantage in the network. *PLoS One* also has a size advantage.

TABLE 4. Co-cited, co-read, and co-tweeted counts for the broad fields.

Main field	Co-cited	Co-read	Co-tweeted
BHS	74%	50%	59%
LES	11.5%	22%	17%
SSH	4%	11%	14%
PSE	10%	14%	9%
MCS	0.5%	3%	1%

The co-saved network of journals (Figure 4) with 65 clusters of journals has a higher modularity (0.57) than the co-citation network and is therefore more clustered. It also includes considerably more journals. Some journals also have particularly strong intracluster connections. The high-impact general journals: *Nature, Science, PLoS One*, and *Proceedings of the National Academy of Sciences of the USA* (PNAS) belong to one cluster. *Nature* is highly co-saved with 387 other journals (38,385 times), including *Cell, Genome Research, Science*, and *Bioinformatics. Science* is highly co-saved with 382 other journals (6,502 times), including *Nature, Cell*, PNAS, and *Genome Research. PLoS One* and PNAS both have connections with 357 other journals. The journals *Cell, Bioinformatics, Genome Research, Nano Lett*, and *ACS Nano* also have strong connections with the highly connected journals above (Table 2).

TABLE 5. The five most co-cited, co-read, and co-tweeted narrow fields.

Narrow field	Co-cited	Narrow field	Co-read	Narrow field	Co-tweeted
Coffee consumption; caffeine; green tea catechin	2.5%	Microarray experiments; gene expression data	3%	Publication bias; research utilization; project paths	1.6%
Aldehyde dehydrogenase; hepatitis c virus infection; alcohol consumption	1.9%	Propanediol; butanol production	2%	Gut microbiota; probiotic property	1.4%
Virgin olive oil; fish consumption; cardiolipin	1.5%	Transposable element; codon usage	1.7%	Health information and literacy; internet addiction; violent video game	1.3%
Apelin; fatty acid synthase; uncoupling protein	1.4%	Gut microbiota; probiotic property	1.6%	Late life depression; borderline personality disorder; mindfulness meditation	1.3%
FTO gene; quantitative trait; population stratification	1.3%	Electron transfer; water oxidation	1.6%	Microarray experiments; gene expression data	1.2%

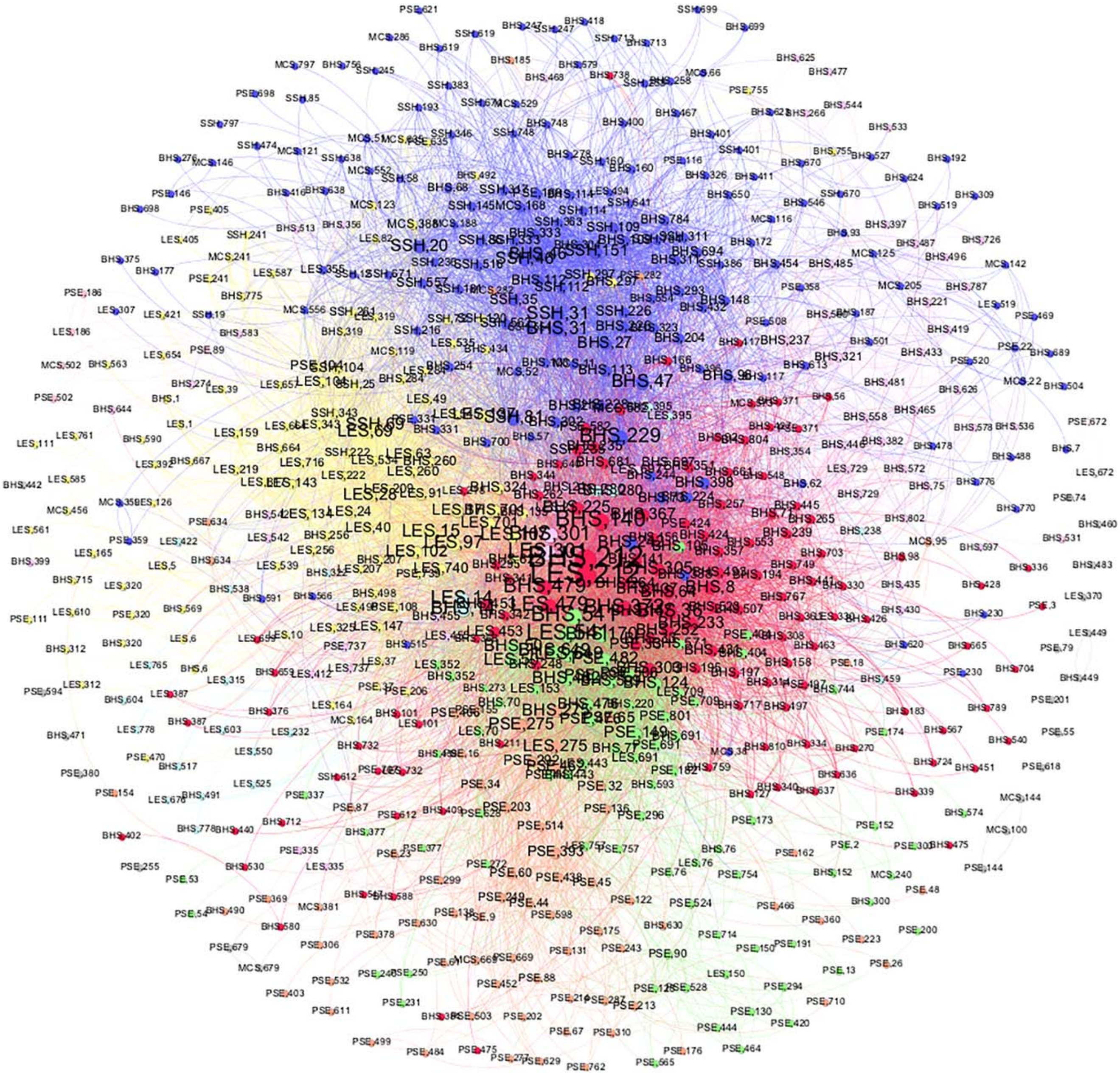

FIG. 7. Co-saved narrow field network (26 clusters were identified through the modularity algorithm). Data: WoS articles from 2012 with 100 Mendeley users; co-saved relationships with strength at least 100. [Color figure can be viewed at wileyonlinelibrary.com]

The co-tweeted journal network (Figure 5) with 75 clusters has a high degree of modularity (0.65) and so is more clustered than the co-saved network (that is, denser connections within clusters and/or weaker connections between clusters). Some medical and biomedical journals have strong co-tweeted connections with each other in the network. *PLoS One* has co-tweeted connections with 275 other journals (15,995 times) and is in the center of graph. This journal has been highly co-tweeted with PNAS, *Nature*, and *Scientific Reports*. Several other PLOS publications (*PLoS Medicine, PLoS Computational Biology, PLoS Pathogens, PLoS Genetics, PLoS Biology*) also have strong co-tweeted connections with *PLoS One*. PNAS is highly co-tweeted with 110 journals (5,184 times) and *Nature* with 88 journals (6,766 times). *Science* is linked with 49 journals (2,907 times). *Lancet* is connected with 100 other journals but its articles are co-cited only 2,012 times with other journals (Table 2).

A feature of journal networks is open accessibility, which could be important for the communities, such as Twitter, where not all users are scholars and some may have restricted access to scientific outputs behind paywalls. Open access journals are more visible in the co-tweeted network than in the other two networks (15% of the journals in the co-tweet network are listed in the Directory of Open Access Journals [DOAJ], in contrast to 8% of the journals in the co-

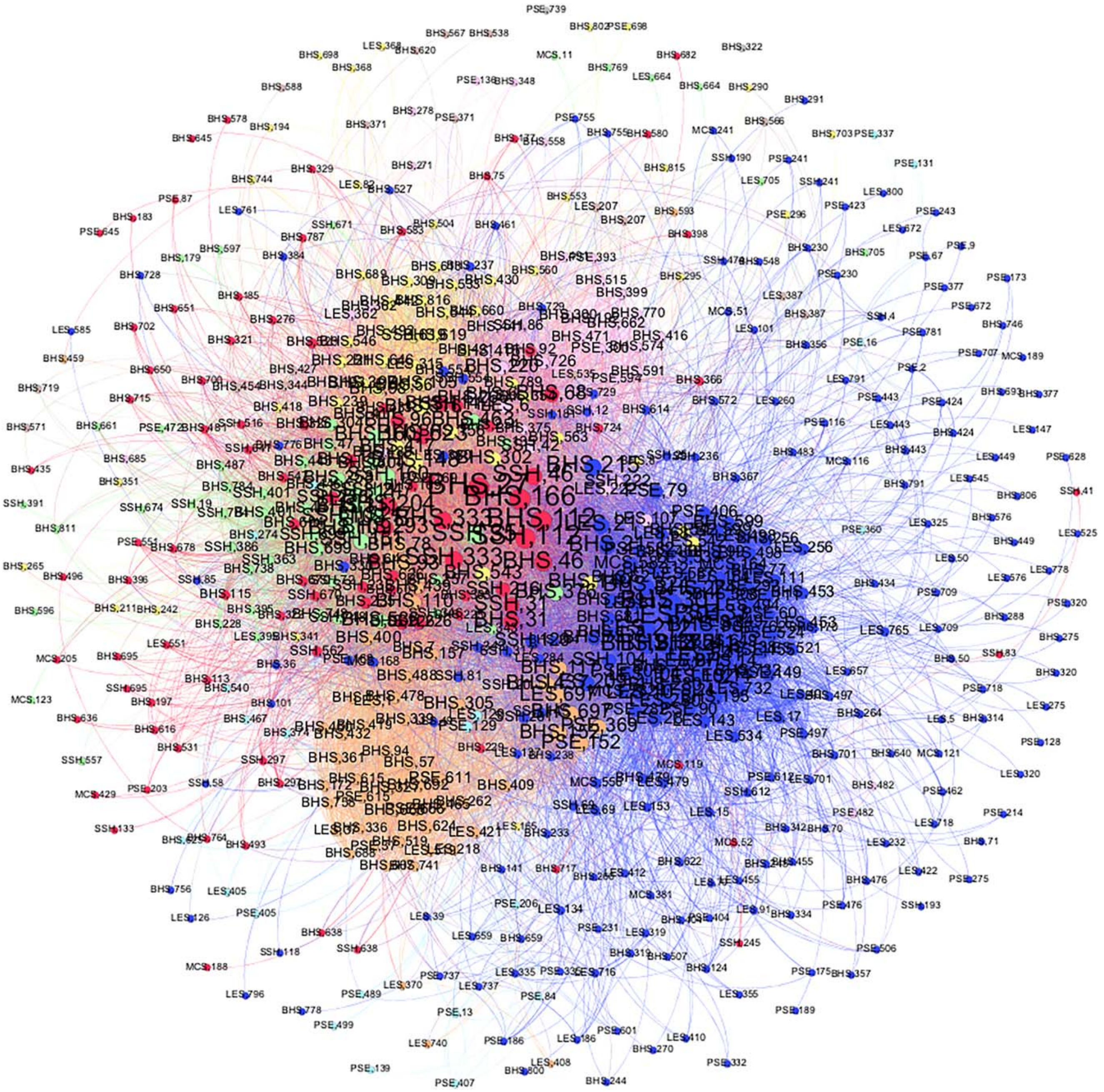

FIG. 8. Co-tweeted narrow field network. (Using the modularity algorithm, 10 clusters were detected with the top one covering 50% of the nodes in the network). Data: WoS articles from 2012 with 10 tweeters; co-tweet relationships with strength at least 10. [Color figure can be viewed at wileyonlinelibrary.com]

citation network and 9.5% of the journals in the co-saved network).

RQ3: The Importance of Narrow Fields for the Three Datasets

In the sample with the first cutoff, 4.4% of the co-cited article pairs, 7.6% of the co-saved article pairs and 6.9% of the co-tweeted article pairs are from the same narrow (meso-level) fields. As mentioned in the Methods section, to have more readable graphs the samples were limited to highly co-cited (co-cited at least 10 times), highly co-saved (co-saved at least 100 times), and highly co-tweeted (co-tweeted at least 10 times) pairs of articles. For this restricted sample, article pairs were much more likely to be from the same narrow fields: 72.3% of highly co-cited article pairs, 43.5% of highly co-saved article pairs, and 21.3% of highly co-tweeted article pairs are from the same narrow fields (Table 3).

The highly co-cited network with 65 clusters has a modularity of 0.49, indicating moderately stronger connections between the nodes within clusters than between clusters (Figure 6). The network is dominated by the Biomedical and Health Sciences (BHS) broad area, which accounts for 74%

DOI: 10.1002/asi

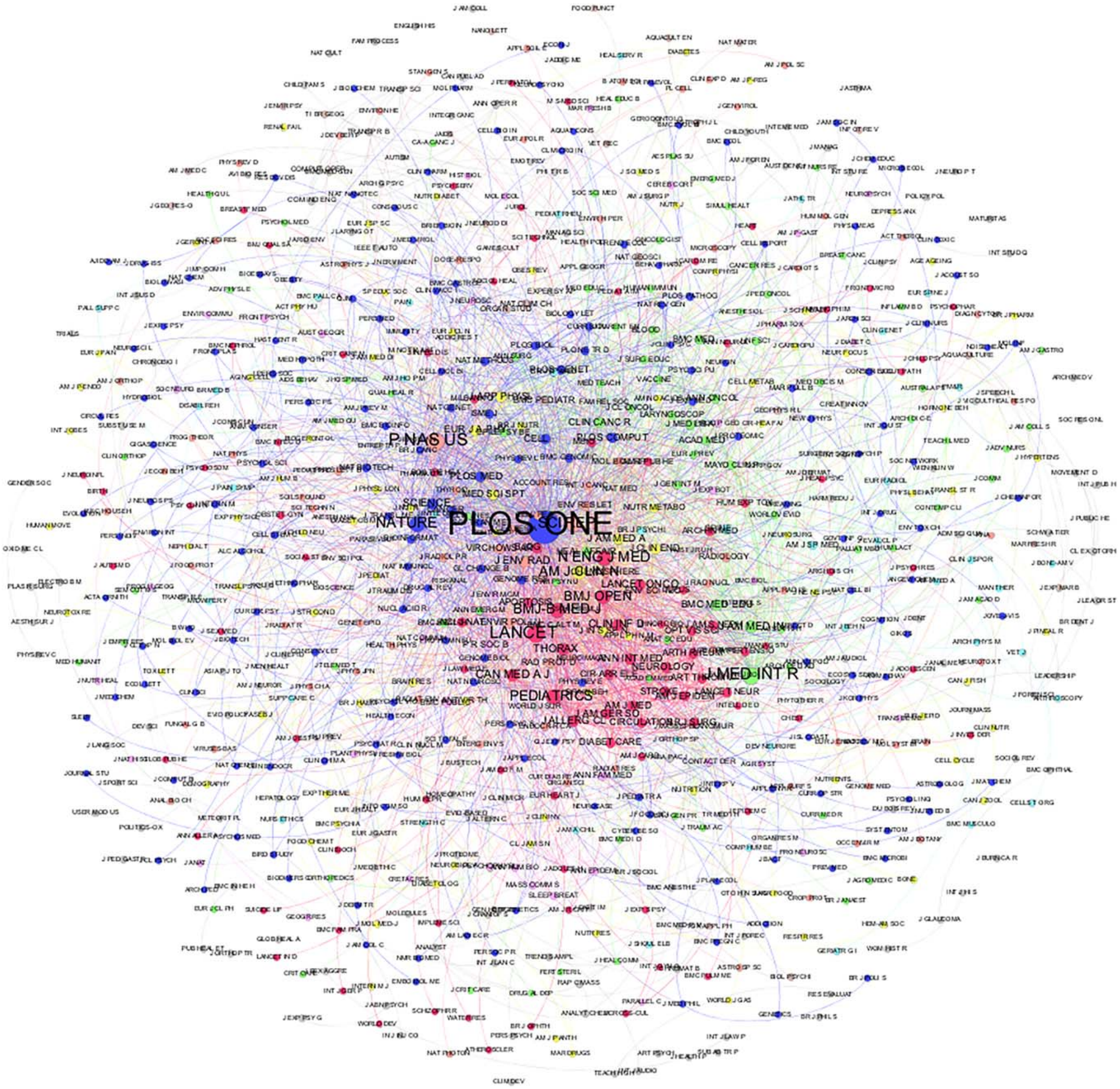

FIG. 9. Highly co-tweeted journals after removing manually identified bots, mainly from the 800 most prolific accounts. (58 clusters were identified through the modularity algorithm. The top four clusters cover around 60% of the nodes in the network). [Color figure can be viewed at wileyonlinelibrary.com]

of the highly co-cited articles. BHS articles are mainly (84%) co-cited with other BHS articles. Less than 4% of the nodes in the network belong to SSH and less than 0.5% to MCS (Table 4). The most highly connected narrow field in the network is a biomedical-medical field relating to coffee consumption and caffeine (Table 5).

The co-saved network with 26 clusters has a higher modularity (0.60) than the co-citation network and so is even more densely clustered (Figure 7). BHS articles again dominate the network (50%), supported by LES (22%), PSE (14%), SSH (11%), and MCS (3%) (Table 4). The most connected narrow field in the network is a BHS-LES field about microarray experiments and gene expression data. This narrow field is also among the highly co-tweeted fields. Gut microbiota and a probiotic property is also the fourth top co-saved narrow field and second top co-tweeted narrow field (Table 5).

The co-tweeted network with 10 clusters has a similar modularity (0.49) to the co-cited network (Figure 8). Most co-tweeted articles are in BHS (59%), supported by LES (17%), SSH (14%), PSE (9%), and MCS (1%) (Table 4). The top connected narrow field in this network is within SSH and BHS and is about publication bias, research utilization, and project paths (Table 5).

RQ4: How Important Are Twitter Bots for the Results?

Since bots can be prolific, Twitter networks may reflect bot rules rather than aggregate human actions. Twitter bots were therefore filtered from the dataset to create a new co-tweeted network with the cleaned data. The first author's assessment of the sample of 1,000 accounts was as follows.

- Prolific tweeters: 46% bot; 37.7% human; 9% mixed; 7.3% other languages, deactivated, and protected accounts.
- Moderate tweeters: 21% bot; 50% human; 19% mixed; 10% deactivated and protected accounts.
- Occasional tweeters: 11% bot; 75% human; 4% mixed; 10% deactivated and protected accounts.

The results confirm that accounts are more likely to be bots if they are more active, at least in terms of the number of DOIs tweeted. All bots identified were removed from the dataset and a new co-tweeted network of journals was created (Figure 9, a filtered version of Figure 5). The new network has 734 fewer nodes and 6,199 fewer connections, despite the changes being (mainly) due to filtering of the top 800 accounts. As can be seen from the graphs, there is a huge decrease in the number of peripheral nodes while the core journals in the original network remain core in the revised network: *PLoS One*, PNAS, *Nature*, and some biomedical and medical journals.

Discussion and Conclusions

This research is limited by analyzing articles from a single year, and with a single classification scheme. The source of evidence about Mendeley users and tweets is also incomplete, which can affect the results. The impossibility of checking all Twitter accounts for bots may systematically affect the results, especially by making tweeters seem to frequently tweet articles from the same journal. The high co-occurring networks are affected by the relatively arbitrary choice of cutoff threshold, which affects comparisons between networks.

Intersections between networks: There were surprisingly low overlaps between the networks: only 4,008 (0.06%) co-cited article pairs were also co-tweeted and only 10,689 (0.16%) co-cited article pairs were also co-saved. There was a much larger overlap between the altmetric networks since 428,774 (13.58%) of the co-tweeted article pairs were also co-saved. It is therefore very rare that two articles that are cited by the same author in the same year are also in the same Mendeley user library or tweeted from the same Twitter account. Many factors can help to explain this result.

- Authors that publish or co-author in different fields may generate many co-cited pairs of articles from publications with little in common.
- Authors' citations are influenced by external factors such as the peer-review process, self-citation inclinations, rewarding peers, which may be less (or more) important for saving and tweeting.
- The nature of co-relationships means that prolific individual accounts can have undue influence on the results because a list of n entities produces $n(n-1)/2$ co-relationships. For example, a prolific author publishing many articles (since 2012) with a total of 100 unique citations to articles published in 2012 would generate 4,950 different co-citation pairings but if s/he only tweeted five articles published in 2012 then this would produce just 10 co-tweeted pairings. If s/he entered all her/his references into Mendeley and no other articles, then this would generate 4,950 co-saved pairings. It seems from this example that it is much easier and more natural for a scientist to produce many co-citation and co-saved pairings than many co-tweeted pairings. Moreover, the most successful researchers seem to rarely use Twitter and Mendeley (Mas-Bleda, Thelwall, Kousha, & Aguillo, 2014), exacerbating the differences with co-citations.
- Older articles within 2012 are less likely to be tweeted (Thelwall, Haustein, Larivière, & Sugimoto, 2013) but age within a year probably has little effect on Mendeley libraries for 2012 (Thelwall & Sud, 2016).
- Twitter users may tend to self-publicize, tweet selectively, or with a topical focus, whereas authors may cite, and Mendeley users may include, a variety of fields and methods that have informed their work(s).
- Both Twitter users and Mendeley users include some nonacademics (for example, students, librarians, practitioners) that have different relationships with academic research. Nevertheless, academics are presumably the main authors of scientific articles and they are also the main users of Mendeley (Mohammadi, Thelwall, Haustein, & Larivière, 2015). Although Mendeley is a tool for managing references and making bibliographies, it can also be used to store and share documents. Mendeley users with wide reading interests may therefore generate eclectic libraries in a way that referees may not permit for the reference lists of published documents.
- The presence of bots in Twitter generates spurious matches, even though some were removed.
- The full-text accessibility of an article may differently affect the likelihood of it being tweeted, cited, or added to Mendeley.
- The social origins of Twitter cause some different patterns of use (Neylons, 2014; Didegah, Bowman, Bowman, & Hartley, 2016), such as articles with funny titles and common social topics being tweeted more often. Although Twitter is mainly used by the public (Haustein, Tsou, Minik, Brinson, Hayes, Costas, & Sugimoto, 2016), tweets about research (and hence the articles in the dataset) may primarily originate from scholars and serve to publicize, rather than discuss, recent research (Thelwall, Tsou, Weingart, Holmberg, & Haustein, 2013).
- The author name disambiguation used has high precision but is imperfect and therefore some of the co-cited article pairs are not co-cited.

Network features: Despite the very low numbers of matches between pairs of articles in the three networks, they have some common patterns at the level of journals and subject fields. In the main samples, a very low percentage of articles in all three networks are from the same journals and the same narrow fields. However, while only 1.5% of the co-cited article pairs are from the same journals, 10% of the co-tweeted article pairs are from the same journals. This suggests that some editors, authors, or publishers regularly tweet recent articles from a journal. This may also be influenced by some journals (for example, *Nature*) hosting a one-click button that tweets a precomposed link to an article. When the networks are limited to highly co-occurring article pairs, the patterns differ. In the sample of highly co-cited articles, 72% of article pairs are from the same narrow fields, showing an extreme field concentration in the highly co-cited set (perhaps unsurprisingly, given that narrow fields were defined by citation relationships). The author-centered co-citation method used here gives an advantage to disciplinary journals since prolific authors may cite huge numbers of articles from one main discipline, whereas those that author in general journals like *Nature* are likely to publish mainly in disciplinary journals. This may also be part of the reason why the overlap between datasets is small. Around 44% of article pairs in the highly co-saved set but only 21% of the highly co-tweeted article pairs are from the same narrow fields. The highly co-tweeted network connections might cover more diverse fields due to the high rate of systematic tweeting of general journals, such as *Science, Nature*, and *PLoS One*, that publish articles from many different specialities. It is also possible that tweeters and Mendeley users have wider multidisciplinary interests than do citers. This would agree with a finding that educational technology articles in Mendeley user libraries are more diverse than authors' cited references (Kraker et al., 2015) and a finding that topics discussed relating to JASIST on Twitter are more diverse than the journal citation network (Jung, Lee & Song, 2016).

The top physics journals and high-impact well-known journals, *Science*, *PLoS One*, and *Nature* are central in the network of highly co-cited articles. This network is affected not only by author reference choices but also by peer-review and editorial policies (for example, restrictions to the number of references; the types of article that articles in a journal are expected to cite). These large journals also have an advantage in the network because it is based on the total number of co-citing pairs and is not normalized for journal size.

The network of highly co-saved articles has a different topology from the co-citation network. The network has a few clusters with strong intercluster connections. There is a big group containing high-impact multidisciplinary journals (*Nature*, *PLoS One*, and PNAS) in the middle, which is strongly connected to some other big groups of medical and biomedical, engineering, and social sciences journals. *PLoS One* (open access) and PNAS (delayed open access) have strong connections in the network, perhaps suggesting the importance of open access journals for Mendeley saving. The network structure is less affected by publishing restrictions and may better reflect users' interests, including perhaps for articles that they do not intend to cite.

The co-tweeted network is different from the co-citation and co-saved networks. *PLoS One* is in the center of the network and has strong connections with several other journals, including some from PLoS. Since 15% of the journals in the co-tweeted network are open access in comparison to 8% for the co-cited network, free full-text access seems to be important for tweeting, since any tweet recipient could immediately read the article. This percentage only includes full open access journals listed in DOAJ, and some tweeted articles in other journals will also be open access.

The co-citation network is dominated by the biomedical and health fields, with only 4% of the network containing social sciences and humanities. Biomedical and medical sciences are also the largest parts of the co-tweeted and co-saved networks but with more social sciences and humanities articles (14% co-tweeted; 11%, co-saved).

Twitter bots: Automated tweeting from bots is a major issue for those hoping to use Twitter to assess the wider impacts of science (Haustein et al., 2016). Based on the manual checks in the current study, 46% of prolific article tweeters are bots, while 21% of moderate and 11% of occasional article tweeters are bots. Thus, accounts that tweet 50 to 1,000 articles from the same year have an approximately even chance of being bots and there are also many bots that tweet only a few articles, even though these are a low proportion of their strata. Thus, those using Twitter citations must accept the likelihood of a substantial bot presence even if they undertake a substantial manual or automatic data cleaning operation. Removing bots from the highly co-tweeted sample caused a drop in the number of nodes and links but did not affect the relative position of some journals, such as *PLoS One*.

In conclusion, the results cast grave doubt on the use of Twitter in altmetrics. Not only are there bots at all levels of activity in terms of tweeting articles (prolific, moderate, occasional) but the relatively high focus on journals by tweeters suggests that articles are selected to be tweeted in a very different way in which articles are selected by academics to be referenced. Although the Mendeley results were much closer to those for citations, there were also substantial differences, showing that this source needs to be used cautiously as an alternative to citations, especially from the perspective of creating networks. Future research is needed to explore in more detail why these differences exist.

Acknowledgment

This research is an extended version of an extended abstract accepted for presentation at the 3am Altmetrics Conference, Bucharest, Romania (September, 2016).

References

Blondel, V.D., Guillaume, J.L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment, 2008(10), P10008.

Bornmann, L., & Haunschild, R. (2016). Overlay maps based on Mendeley data: The use of altmetrics for readership networks. Journal of the Association for Information Science and Technology, 67(12), 3064–3072.

Boyack, K.W., & Klavans, R. (2010). Cocitation analysis, bibliographic coupling, and direct citation: Which citation approach represents the research front most accurately? Journal of the American Society for Information Science and Technology, 61(12), 2389–2404.

Caron, E., & Van Eck, N.J. (2014). Large scale author name disambiguation using rule-based scoring and clustering. *Proceedings of the 19th International Conference on Science and Technology Indicators*, 79–86.

Didegah, F., Bowman, T.D., & Holmberg, K. (2017, in press). On the differences between citations and altmetrics: An investigation of factors driving altmetrics vs. citations for Finnish articles. Journal of the Association for Information Science and Technology.

Didegah, F., Bowman, T., Bowman, S., & Hartley, J. (2016). Comparing the characteristics of highly cited titles and highly alted titles. *International Conference of Science and Technology Indicators*, Valencia, Spain. September 14–16. http://ocs.editorial.upv.es/index.php/STI2016/STI2016/paper/viewFile/4543/2327

Haustein, S. (2012). Multidimensional Journal Evaluation: Analyzing Scientific Periodicals beyond the Impact Factor. Berlin: De Gruyter Saur: 143–147.

Haustein, S., et al. (2016). Tweets as impact indicators: Examining the implications of automated “bot” accounts on Twitter. Journal of the Association for Information Science and Technology, 67(1), 232–238.

Haustein, S., Costas, R., & Lariviére, V. (2015). Characterizing social media metrics of scholarly articles: the effect of document properties and collaboration patterns. PLoS One, 10(3). https://doi.org/10.1371/journal.pone.0120495

Haustein, S., Larivière, V., Thelwall, M., Amyot, D., & Peters, I. (2014). Tweets vs. Mendeley readers: How do these two social media metrics differ? IT-Information Technology, 56(5), 207–215.

Haustein, S., et al. (2016). “Identifying Twitter user communities”. *3am conference*. Bucharest, Romania, September 28–29.

Jung, H., Lee, K., & Song, M. (2016). Examining characteristics of traditional and Twitter citation. Frontiers in Research Metrics and Analytics, 1, 6.

Kraker, P., Körner, C., Jack, K., & Granitzer, M. (2012). Harnessing user library statistics for research evaluation and knowledge domain visualization. In *Proceedings of the 21st International Conference on World Wide Web (pp. 1017–1024)*. Lyon, France, 16–20 April.

Kraker, P., Schlögl, C., Jack, K., & Lindstaedt, S. (2015). Visualization of co-savedership patterns from an online reference management system. Journal of Informetrics, 9(1), 169–182.

Marshakova, I. (1973). System of document connections based on references. Nauchno-Tekhnicheskaya Informatsiya Seriya 2-Informatsionnye Protsessy I Sistemy, (6), 3–8.

Mas-Bleda, A., Thelwall, M., Kousha, K. & Aguillo, I.F. (2014). Do highly cited researchers successfully use the social web? Scientometrics, 101(1), 337–356.

Mohammadi, E., & Thelwall, M. (2014). Mendeley readership altmetrics for the social sciences and humanities: Research evaluation and knowledge flows. Journal of the Association for Information Science and Technology, 65(8), 1627–1638.

Mohammadi, E., Thelwall, M. & Kousha, K. (2016). Can Mendeley bookmarks reflect readership? A survey of user motivations. Journal of the Association for Information Science and Technology, 67(5), 1198–1209.

Mohammadi, E., Thelwall, M., Haustein, S., & Larivière, V. (2015). Who reads research articles? An altmetrics analysis of Mendeley user categories. Journal of the Association for Information Science and Technology, 66(9), 1832–1846.

Newman, M.E., & Girvan, M. (2004). Finding and evaluating community structure in networks. Physical Review E, 69(2), 026113.

Neylon, C. (2014). Altmetrics: What are they good for. PLOS Opens, http://blogs.plos.org/opens/2014/10/03/altmetrics-what-are-they-good-for/.

Priem, J., & Costello, K.L. (2010). How and why scholars cite on Twitter. Proceedings of the American Society for Information Science and Technology (ASIST 2010) (1–4) doi:10.1002/meet.14504701201

Priem, J., Taraborelli, D., Groth, P., & Neylon, C. (2011). Altmetrics: a manifesto. http://altmetrics.org/manifesto/

Small, H. (1973). Co-citation in the scientific literature: A new measure of the relationship between two documents. Journal of the American Society for Information Science, 24(4), 265–269.

Small, H. (1981). The relationship of information science to the social sciences: A co-citation analysis. Information processing & management, 17(1), 39–50.

Small, H. (1999). Visualizing science by citation mapping. Journal of the Association for Information Science and Technology, 50(9), 799.

Sud, P. & Thelwall, M. (2014). Evaluating altmetrics. Scientometrics, 98(2),1131–1143.

Sugimoto, C.R., Work, S., Larivière, V., & Haustein, S. (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.

Thelwall, M., Haustein, S., Larivière, V., & Sugimoto, C.R. (2013). Do altmetrics work? Twitter and ten other social web services. PLoS One, 8(5), e64841. http://doi.org/10.1371/journal.pone.0064841

Thelwall, M. & Sud, P. (2016). Mendeley readership counts: An investigation of temporal and disciplinary differences. Journal of the Association for Information Science and Technology, 57(6), 3036–3050.

Thelwall, M. Tsou, A., Weingart, S., Holmberg, K., & Haustein, S. (2013). Tweeting links to academic articles. Cybermetrics, 17(1), http://cybermetrics.cindoc.csic.es/articles/v17i1p1.html.

Waltman, L., van Eck, N.J., van Leeuwen, T.N., Visser, M.S., & van Raan, A.F. (2011). Towards a new crown indicator: An empirical analysis. Scientometrics, 87(3), 467–481.

Waltman, L., & Van Eck, N.J. (2012). A new methodology for constructing a publication-level classification system of science. Journal of the American Society for Information Science and Technology, 63(12), 2378–2392.

White, H.D., & McCain, K.W. (1998). Visualizing a discipline: An author co-citation analysis of information science, 1972–1995. Journal of the American society for information science, 49(4), 327–355.

DOI: 10.1002/asi