


**Please cite the Published Version**

Said, A, Bowman, TD, Abbasi, RA, Aljohani, NR, Hassan, SU and Nawaz, R  (2019) Mining network-level properties of Twitter altmetrics data. *Scientometrics*, 120 (1). pp. 217-235. ISSN 0138-9130

**DOI:** <https://doi.org/10.1007/s11192-019-03112-0>

**Publisher:** Springer Verlag

**Version:** Accepted Version

**Downloaded from:** <https://e-space.mmu.ac.uk/623508/>

**Additional Information:** This is a post-peer-review, pre-copyedit version of an article published in *Scientometrics*. The final authenticated version is available online at: <http://dx.doi.org/10.1007/s11192-019-03112-0>.

**Enquiries:**

If you have questions about this document, contact [openresearch@mmu.ac.uk](mailto:openresearch@mmu.ac.uk). Please include the URL of the record in e-space. If you believe that your, or a third party's rights have been compromised through this document please see our Take Down policy (available from <https://www.mmu.ac.uk/library/using-the-library/policies-and-guidelines>)

# Mining Network-Level Properties of Twitter Altmetrics Data

Anwar Said<sup>a</sup>, Timothy D. Bowman<sup>b</sup>, Rabeeh Ayaz Abbasi<sup>c</sup>, Naif Radi Aljohani<sup>d</sup>,  
Saeed-Ul Hassan<sup>a</sup>, Raheel Nawaz<sup>e</sup>

<sup>a</sup> Information Technology University, 346-B, Ferozepur Road, Lahore, Pakistan  
E-mail address: saeed-ul-hassan@itu.edu.pk, anwar.said@itu.edu.pk, Tel: + 92-322-228-9756

<sup>b</sup> School of Information Sciences, Wayne State University, Detroit, MI, United States  
E-mail address: timothy.d.bowman@wayne.edu

<sup>c</sup> Department of Computer Science, Quaid-i-Azam University, Islamabad, Pakistan.  
E-mail address: rabbasi@qau.edu.pk

<sup>d</sup> Faculty of Computing and Information Technology, King Abdulaziz University,  
Jeddah, Kingdom of Saudi Arabia  
E-mail address: nraljohani@kau.edu.sa

<sup>e</sup> Department of Operations, Technology, Events and Hospitality Management,  
Manchester Metropolitan University, Manchester, United Kingdom  
E-mail address: r.nawaz@mmu.ac.uk

## Abstract

Social networking sites play a significant role in altmetrics. While Twitter originates more than 90% of altmetric mentions, the known microscopic and macroscopic properties of Twitter altmetrics data are limited. In this study, we present a large-scale analysis of Twitter altmetrics data using social network analysis techniques on the ‘mention’ network of Twitter users. Exploiting the network-level properties of over 1.4 million tweets, corresponding to 77,757 scholarly articles, this study focuses on the following aspects of Twitter altmetrics data: a) the influence of organizational accounts; b) the formation of disciplinary communities; c) the cross-disciplinary interaction among Twitter users; d) the network motifs of influential Twitter users; and e) testing the small-world property. The results show that Twitter-based social media communities have unique characteristics, which may affect social media usage counts either directly or indirectly. Therefore, instead of treating altmetrics data as a black box, the underlying social media networks, which may either inflate or deflate social media usage counts, need further scrutiny.

**Keywords:** Altmetrics, Community Structure, Influential Users, Motifs, Overlapping Communities, Twitter

## 1 Background

Measuring the impact of scientific articles has been an active research area for the past decade, and various methods have been adopted. Among these, citation analysis is one

of the dominant means of research evaluation (Moed, 2010; Radicchi, Fortunato & Castellano, 2008). However, as citations focus solely on scientific impact and not on the broader societal impact of research, many funding organizations and scientific research councils have turned to altmetrics for evidence of the social usage of scholarly articles (Dinsmore, Allen & Dolby, 2014; Wilsdon, 2016).

The term altmetrics was first coined by Priem, Taraborelli, Groth, and Neylon (2010) in their study to evaluate traces of usage of scholarly documents in online contexts. Altmetrics refers to a system that tracks and measures the attention received by research objects, which can include articles, datasets, presentations, software and tools shared by scholars, scientific communities and the public in various Online Social Networks (OSNs), including Twitter, Reddit and Facebook (Haustein et al., 2016; Sugimoto, Work, Larivière & Haustein, 2017; Thelwall & Nevill, 2018; Yu, Xu, Xiao, Hemminger & Yang, 2017). These OSNs have demonstrated exponential growth over the past decade, and they connect large numbers of people (Priem & Hemminger, 2010; Wouters & Costas, 2012; Zahedi & Haustein, 2018) and facilitate the sharing of ideas and the ability to receive an immediate response from peers. Due to this rapid response capability, OSNs have attracted the attention of the scientific research community (Piwowar, 2013). According to Hassan et al. (2017), Twitter is among the most widely used OSN for information sharing and content dissemination, and more than 91% of altmetrics mentions stem from Twitter.

Twitter allows the sharing of public or private short messages known as tweets. Besides, Twitter affords users various options, including the functionality to retweet an original tweet, add hashtags to a tweet, create lists of relevant tweets and mention

another user in their tweet. This provides an opportunity for messages to be spread to a wide range of unknown audience members. A retweet allows Twitter users to share a chosen tweet with their followers, while a mention enables them to tag another user in a tweet or retweet. Various studies have shown the importance of Twitter in generating altmetrics data (Bornmann & Haunschild, 2018; Haustein et al., 2016; Haustein, Peters, Sugimoto, Thelwall & Larivière, 2014; Vainio & Holmberg, 2017), however very few have analysed altmetrics Twitter data by means of Social Network Analysis (SNA).

Alperin and Haustein (2017) explored altmetrics data using SNA by creating networks of the Twitter followers of seven highly tweeted articles and found that SNA can improve current altmetrics indicators. Imran et al. (2018) analysed Twitter social networks (retweet, mention) using altmetrics data to examine different network properties across academic fields. The authors highlight that the properties of these networks vary across the fields. Similarly, Didegah and Thelwall (2018) used SNA techniques to analyse researchers, who tend to save or tweet articles similar to those that they cite. While these studies represent initial attempts to study altmetrics Twitter data using SNA, they do not examine the underlying structure of the altmetrics Twitter social network.

In addition to examining Twitter, various studies have employed SNA on other sources of altmetrics data. Hoffmann et al. (2016) conducted a study investigating Researchgate<sup>1</sup> (R<sup>G</sup>), an academic, social networking site, to examine interactions among Swiss scholars in the field of management. Using the eigenvector centrality measure to rank the users, the study reported that high-profile scholars (professors) are

---

<sup>1</sup> <https://www.researchgate.net>

more central and dominant in the network than senior faculty members. In a survey of Academia.edu, Jordan (2014, 2017) reported that the interactions and relationships on the site functioned as an online business card, whereby users (mainly researchers) followed people without personally knowing them. Yan and Zhang (2018) conducted a large-scale analysis and collected the quantification of scientific reputation (termed as  $R^G$  score) of researchers at various levels from 61 US universities. The study reported that the scores closely and realistically reflected the institutions' research quality. Lutz and Hoffmann (2018) conducted a descriptive data analysis to explore various aspects of  $R^G$  followers/friends networks. They found that seniority is highly correlated to publication impact, which further leads to an increase in network centrality.

Many studies have shown that SNA techniques can be used to mine complex user interactions by employing graph-based, spectral and probabilistic approaches (Abbasi, Altmann & Hossain, 2011; Otte & Rousseau, 2002). Notably, users with similar interests usually have similar sub-network structures and patterns, when these are represented in the form of a social network (Fortunato, 2010). Similarly, users with unique patterns of interactions across the network can be identified from their position and local neighbourhood within the network. SNA provides practical methods to discover patterns in complex networks (Barabási, 2016; Fortunato, 2010). Keeping the complex dynamics of altmetrics data in context, the authors leveraged SNA approaches to investigate both the microscopic and macroscopic properties.

The objectives of this study are to address the following research questions:

- What type of user accounts are influential in the altmetrics Twitter network?

- How do Twitter network communities form and what do these communities represent in the altmetrics Twitter network?
- To what extent do Twitter users interact with scientific publications across fields?
- What are the common means of communication (network motifs) in the altmetrics Twitter network?
- Does the altmetrics Twitter network satisfy the small-world property?

The rest of the article is structured as follows. Section 2 provides a brief overview of the dataset and social network formation. Section 3 investigates both the microscopic and macroscopic properties of the altmetrics Twitter network, using SNA approaches including centrality measures, community detection, and recurring patterns. Finally, Section 4 concludes the article and discusses future directions.

## **2 Twitter Social Mention Network**

This study uses the dataset released by Altmetric.com on June 14, 2016 (version dataset-jun-4-2016.tar.gz). This version of the Altmetric.com dataset contains approximately 4.5 million JSON files, each containing information on a single publication. Altmetric.com captures mentioning of scholarly publications in various online contexts by tracking Digital Object Identifier (DOI) use and certain domain names. Because Altmetric.com does not include citation counts for publications, the count for each scholarly object in this dataset was obtained using the Scopus API – for more details about the dataset, see Hassan et al. (2017). From this, the authors obtained a subset of all scholarly articles published in 2015 that had at least one citation (through February 2017) and at least one tweet with at least one user mention, as captured by Altmetric.com. The authors had chosen to limit the dataset to articles with at least one

citation in order to have a manageable dataset on which to apply network analyses. The final dataset for this study is over 1.4 million tweets, corresponding to 77,757 scholarly articles. Note that these selected publications (with at least one citation each) comprise 15.2% of all Scopus publications indexed in Altmetric.com in 2015.

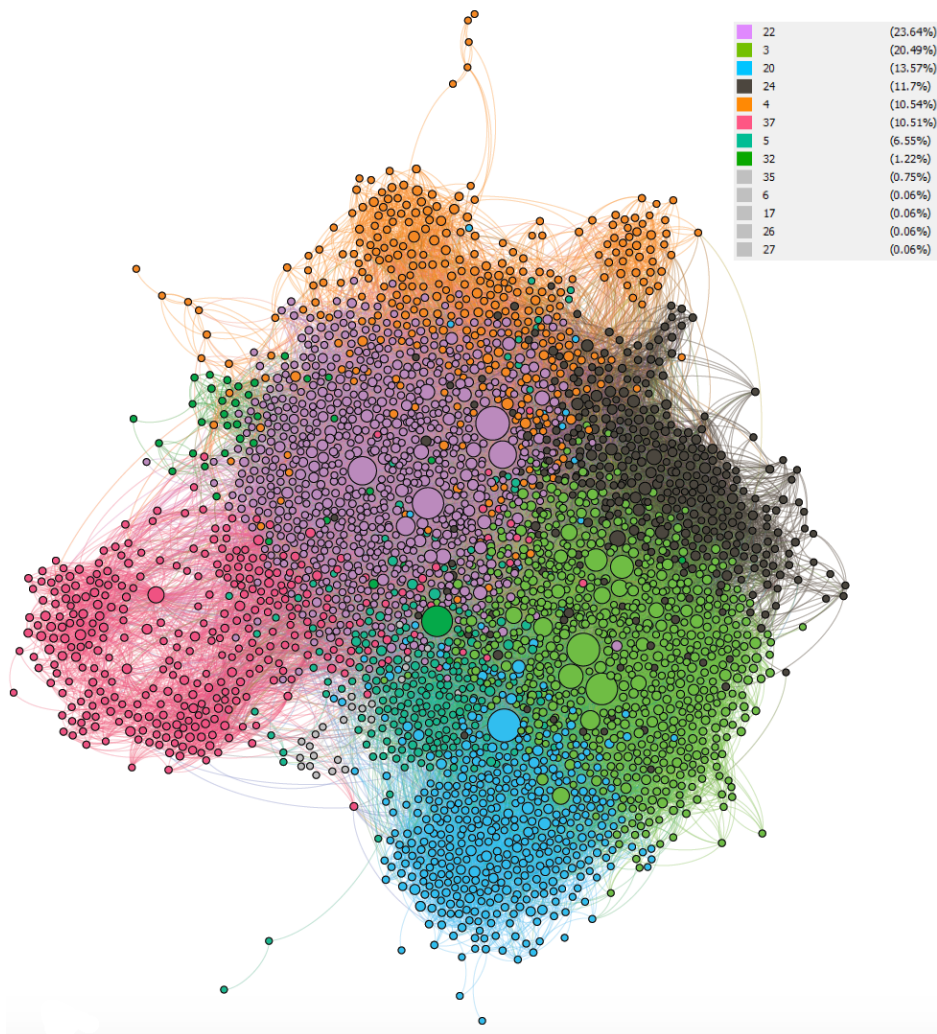
The selected data consist of the following attributes: (i) tweet ID; (ii) Altmetric.com ID; (iii) screen-name; (iii) screen-name-mention; (iv) retweet; (v) mention; (vi) subject field; and (vii) date of publication. Here, the tweet ID and Altmetric.com ID are unique IDs representing the tweets and articles, respectively. However, there can be a one-to-many association between Altmetric ID and tweet ID, as a single article can be tweeted many times. The screen-name attribute represents the user who tweets, while the screen-name-mention attribute represents the user who is mentioned in the tweet. Both the screen-name and the screen-name-mention attributes were used to create the social network, which has been termed as Altmetrics Twitter Social Network (ATSN).

Formally, the ATSN is an un-weighted directed graph  $G = (V, E)$ , where  $V$  represents a set of vertices  $\{v_1, v_2, \dots, v_n\}$  and  $E$  is the edge set  $(u, v)$ . The authors constructed a directed edge between  $u \rightarrow v$ , which indicates that  $u$  mentions  $v$ , and an adjacency matrix  $A$  where  $A_{uv} = 1$ , where there is a directed edge between node  $u$  and  $v$  and 0 otherwise. The total numbers of nodes and edges in the network are 149,830 and 374,822, respectively. Further statistics, including average degree, average path length, component ratio and weakly connected components, are shown in Table 1. It can be observed that more than 98% of the nodes form a giant component within the network. The visualization of the network is presented in Figure 1; for a better presentation of the visualization, only the 3,206 layout nodes that have a total degree  $\geq 30$  are

displayed. Different coloured nodes represent the various communities in the network, while the size of the nodes depicts their eigenvectors' centrality values. Among the eight communities that individually comprise over 1% of the network size, two communities, (i.e. #22 and #3) appear to be the largest, with more than a 20% share of the whole network.

**Table 1:** Statistics of the Altmetrics Twitter Social Network (ATSN).

Nodes	Edges	Avg. Clustering Coefficient	Avg. Degree	Avg. Path Length	Component Ratio	Weakly Connected Components
149830	374822	0.084	2.5	8.45	0.004	681



**Figure 1:** Visualization of the altmetrics Twitter 'mention' network. The network demonstrates a color-wise community structure; nodes of the same colors belong to the same community. For better visualization, nodes of total degree <30 are filtered from the network. This visualization is performed in Gephi using OpenOrd and ForceAtlas layouts on a setting to prevent overlap. The colors were assigned automatically using a community-detection algorithm (modularity) and the size of nodes is determined by the eigenvector centrality measure, with min size = 10 and max size = 30. Note that 2.14% of the nodes remain after filtering the network with total degree < 30.



### **3 Investigating the Microscopic and Macroscopic Properties of ATSN**

In this section, the micro- and macro-level properties of the ATSN are presented. First, the micro-level properties are investigated to reveal the network's influential users; then the macro-level properties are investigated to reveal the community structure of the network. Note that Gephi (Bastian et al, 2009) was used to conduct the required analysis.

#### **3.1 Twitter influential users**

To find the influential users in the social network, the notion of centrality was used, which is a well-known concept in SNA (Borgatti, 2005). The centrality measure computes the global and local influence of the nodes by mining their connectivity within a network. Over the years, a large number of centrality measures have been proposed; among them, eigenvector centrality is one of the most widely used for finding central nodes in social networks (Bonacich, 2007; Carrington, Scott & Wasserman, 2005).

In the context of the ATSN, the important nodes (users) are those mentioned by other nodes (users). Note that the importance of a node can be computed using a simple SNA measure, such as In-degree Centrality. However, this fails to capture the importance of a node (user) with fewer mentions, and a node with a low In-degree may yet be relevant if another important node mentions it. Centrality measures such as Eigenvector Centrality or PageRank can identify these important nodes. Other centrality measures such as Closeness Centrality or Betweenness Centrality were considered unsuitable for the ATSN, as these also do not take into consideration the relationships of a node with other vital node.

**Table 2:** Top 20 IUs (influential Twitter users) with their Ev-centrality (eigenvector) and PR (PageRank) centrality values. IUs are shown in order, with respect to EV-centrality; however, PR values also follow a similar pattern. The top 20 IUs concerning PR value also remain the same, with a minor change in their order.

IUs	Type	Field	EV-centrality	PR
PLOSONE	Journal	Science and Medicine	1	0.0058
Science magazine	Journal	General	0.844	0.0056
JAMA current	Journal	Health Sciences	0.828	0.0047
nature	Journal	General	0.81	0.0054
TheLancet	Journal	Medicine	0.777	0.0066
Bmj latest	Journal	Medical	0.71	0.0048
PNASNews	Journal	Science	0.61	0.0027
NEJM	Journal	Medicine	0.596	0.0036
CellCellPress	Journal	Biology	0.457	0.0019
NatureNews	Journal	General	0.42	0.004
Nature	Journal	General	0.413	0.0036
PLOSbiology	Journal	Biology	0.359	0.0014
WHO	Organization	Health Sciences	0.312	0.0014
NatureBiotech	Journal	Biotechnology	0.304	0.0012
CurrentBiology	Journal	Biology	0.298	0.0012
NatureMedicine	Journal	Medicine	0.293	0.0013
BJSM BMJ	Journal	Medicine	0.271	0.0016
PLOS	Journal	Medicine	0.26	0.0011
NatureGenet	Journal	Biology/Genetic	0.251	0.0007
AnnalsofIM	Journal	Medicine	0.232	0.0011

With the above considerations, the Eigenvector Centrality measure was chosen to identify the important nodes in the network, since it is based on a network spectrum (eigenvalues), which capture a global view of the whole network through an adjacency matrix. The measure computes the centrality of nodes with respect to the centrality value of its neighbors. Its mathematical formulation is shown in Eq. 1:

$$Ax = \lambda x \quad (1)$$

where  $\lambda$  is a normalization constant,  $A$  is the adjacency matrix and  $x$  is the vector of the eigenvalue scores.

Using the eigenvector centrality measure, the top 20 influential users are presented (see Table 2). To provide a comparison, PageRank values are displayed alongside the eigenvector centrality values in Table 2. Apart from the World Health Organization (WHO), all the top 20 influential users are organizational accounts associated with highly reputable journals, which demonstrates their dominance in the ATSN. Table 2

shows that organizational accounts associated with journals play a significant role in altmetrics and that most of these journals are in medical fields.

Among the influential users found, PLOS One has the highest eigenvalue and the highest degree value in the whole network. This account has a high co-mention relationship with other medical accounts associated with medical journals, including PLOS Biology, Nature Neuro and PLOS Medicines. Science Magazine is the second highest organizational account, with both high centrality and degree values. Several other accounts, such as PNAS, Nature and Scientific Reports, demonstrate a high co-mention relationship with Science. Our analysis reveals that the ATSN can be affected by the following primary social media biases: a) type of Twitter handler (e.g. organizational accounts in the current scenario have higher centrality scores than other accounts); b) reputation of the Twitter account (e.g. highly reputable journals have more followers and hence a better chance of obtaining wide social coverage); and c) the relation of a user to influential users (i.e. this relationship can lead to the spread of a tweet to a wide audience).

### **3.2 Mining community structure**

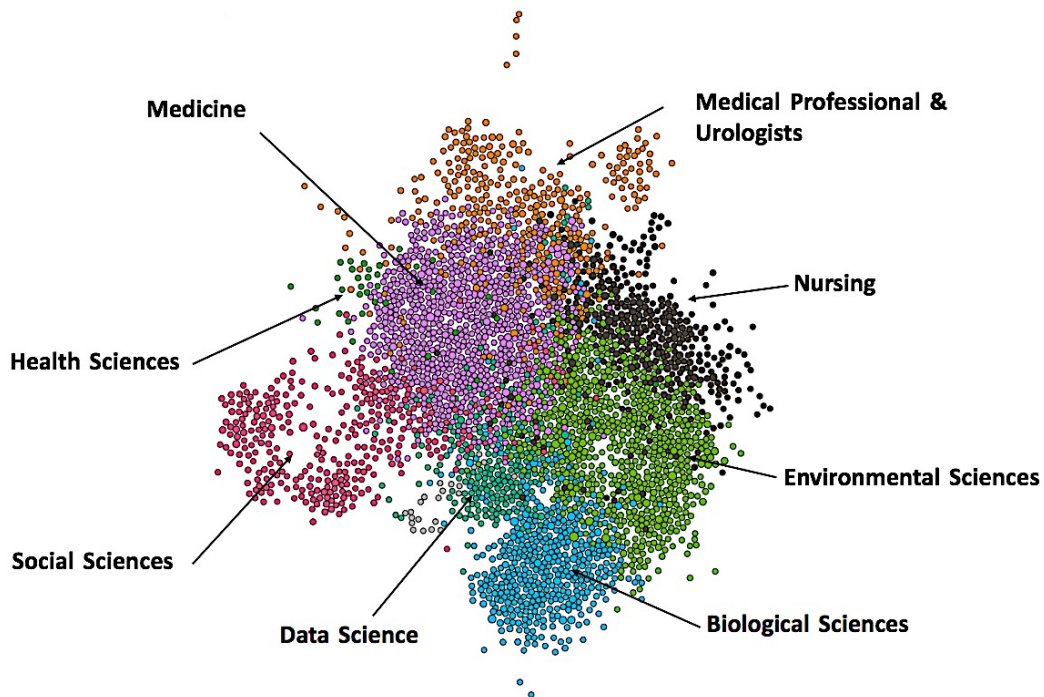
Communities are subgroups of nodes that are densely connected to their members and sparsely connected to the rest of the network (Said, Abbasi, Maqbool, Daud & Aljohani, 2018; Yang, McAuley & Leskovec, 2013). Complex real-world networks have abundant hidden information that is not easily detected by simple observation. However, most of that information can be extracted by analyzing the community structure of the networks.

In order to explore the community structure of the ATSN, a state-of-the-art community-detection algorithm invented by Blondel et al. (2008) was deployed. This algorithm uses a heuristic modularity optimization approach that works hierarchically to provide competitive results concerning modularity and time. Note that the algorithm is based on a greedy approach in order to optimize modularity that has time complexity of  $O(n \log n)$ . The modularity is scaled between -1 and 1 and evaluates a given community by measuring its inter- and intra-linkages.

The Blondel algorithm utilizes heuristics, since finding all combinations of the nodes to form the community is a far from trivial task. First, the algorithm optimizes the modularity of all nodes by generating small communities; second, these small communities are coupled to form relatively larger groups. These steps are repeated until the modularity has converged. Using the Blondel community-detection algorithm in Gephi, eight major communities were identified in the network. Next, the subject fields of these selected communities were identified by matching the Twitter profiles of the top five influential users in each community. A list of the selected users, along with their subject fields and account types (scholar/organization), is attached in Appendix-A, Table A-1.

Among the selected eight communities, Community #22 was found to be the most significant, covering 23% of the users in the network (see Fig. 1). The top five influential users of this community included JAMA, TheLancet, BMJ and NEJM, all organizational accounts associated with medical journals (see Appendix Table A-1). It was discovered that all the important user accounts in Community #22 are in the field of medicine, so it was labeled as such (see Fig. 2). In addition to medicine, certain fields

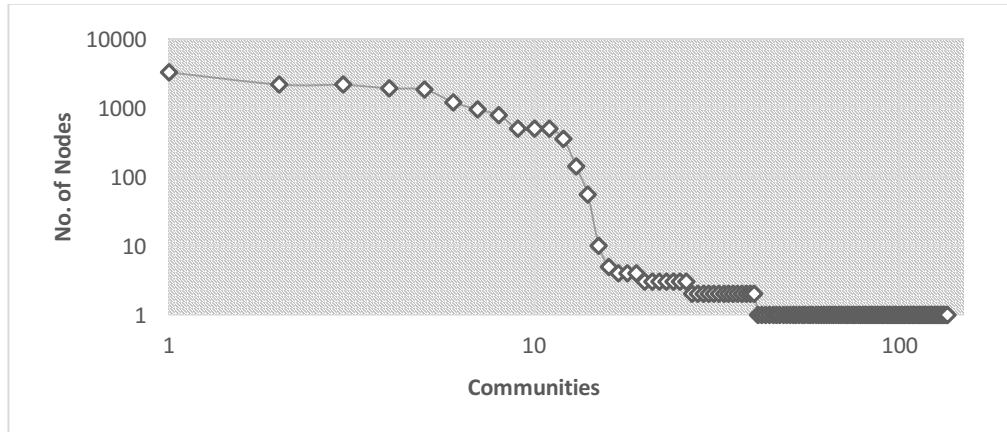
are prominent, such as biological sciences, environmental sciences, social sciences and the emerging field of data science. The emerging community of data science (i.e. Community #5) covers over 5% of the network and includes well-known data scientists including Albert Barabási and Shannon McGregor. Note that Figure 2 highlights only the major communities that are commonly found in altmetrics data; a number of other fields or subfields, such as computer science, physics and mathematics, are not discernible, which may be due to the exclusion filter applied to the node's degree (i.e.  $< 30$ ).



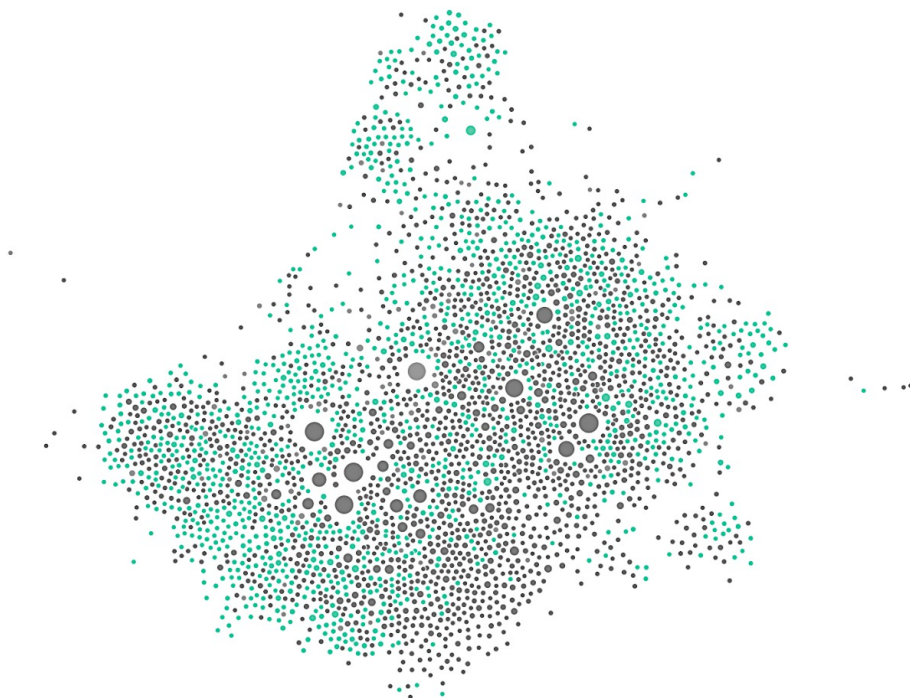
**Figure 2:** Community structure of the ATSN shown in Fig 1. Each community is labeled after matching it to its respective Twitter profile to better understand the network structure.

Overall, the analysis reveals that the top influential users of each community are either organizational accounts associated with well-known journals or leading scholars in a particular field. Large communities such as Communities #22 (Medicine), #3 (Environmental Sciences) and #20 (Biological Sciences) are dominated by well-known

journals that can influence a giant component of the network. By contrast, the relatively small communities such as Communities #5 and #32 are dominated by top scholars in to their respective fields, who may have an influence on only a small sub-network. It was also found that the community size distribution closely follows a power-law distribution (see Fig. 3).



**Figure 3:** Community size distribution: A few communities (x-axis) have a large number of Twitter users (y-axis), while a large number of communities have fewer users, which closely follows a power-law distribution.



**Figure 4:** Altmetrics scholars (green) and journal/organizational accounts (gray) visualization. Here the size of the nodes represents their importance, based on eigenvector centrality measure. Overall, the green nodes (scholars) are of less importance than the gray nodes, which include organizational accounts associated with journals and organizations.

In order to investigate the presence of scholar and journal/organizational accounts, the method proposed by Costas et al. (2017) was employed – see Figure 4, in which green nodes represent scholars while gray nodes depict non-scholars. Note that the size of a node highlights its importance (its eigenvector centrality value) in the network. One can see that the green nodes (scholars) are rarely influential, which again emphasizes the significant role of journal/organizational accounts in the ATSN.

### 3.3 Analysis of overlapping communities across fields

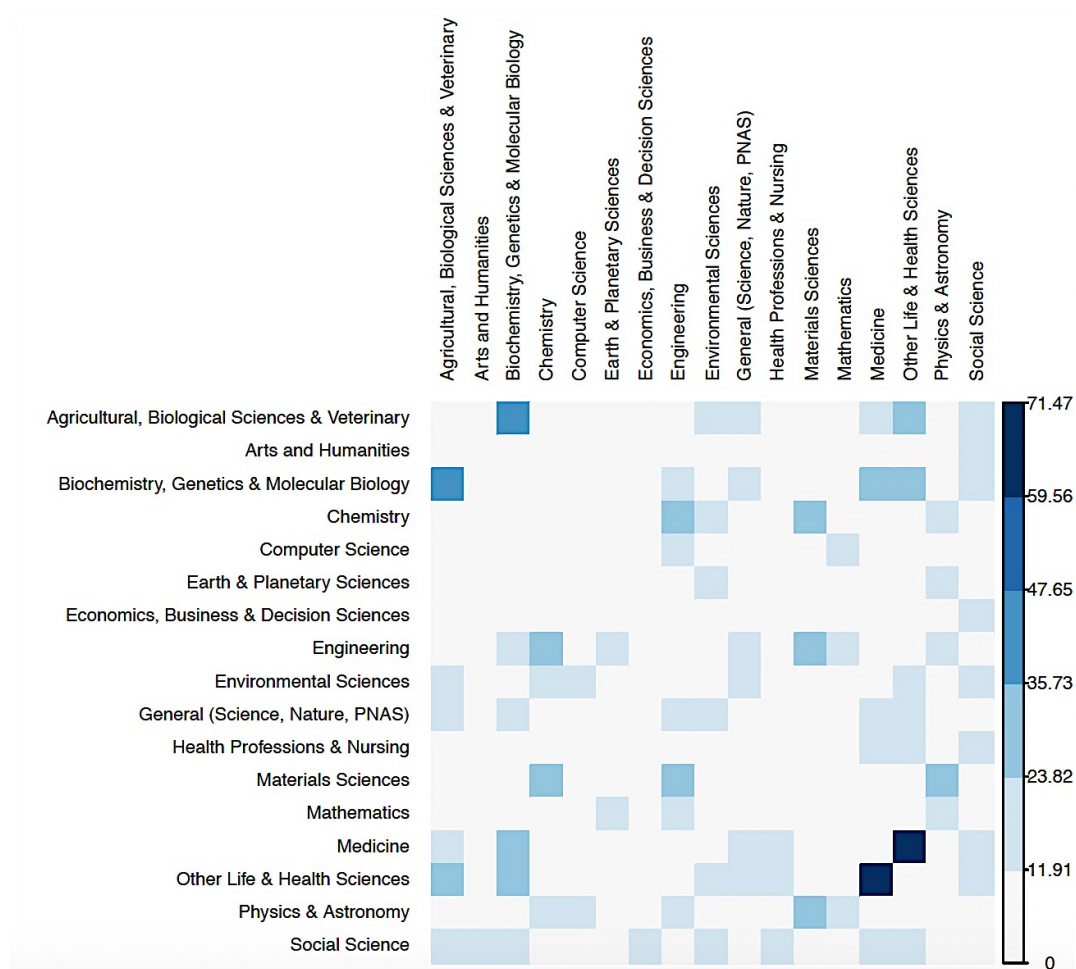
Community overlap is a significant feature of many real-world networks. It is known that people in a social network naturally have multiple community membership. For example, in a social network a person may interact with several social groups, such as family, colleagues and friends. Similarly, in the ATSN a scholar may be active across various subject fields. In this subsection, the Twitter users who mention scientific articles across subject fields are identified.

Following the work of Haddawy et al. (2016), 77,757 selected scientific articles were mapped to 17 broad subject fields using the All Science Journal Classification embedded in Scopus (<https://www.scopus.com>). To find the overlap across the fields, a well known Jaccard (1901) similarity measure was used, which may be computed as shown in Eq. 1:

$$J_{U_i, U_j} = \frac{|U_i \cap U_j|}{|U_i \cup U_j|} \quad (1)$$

Here,  $U_i$  and  $U_j$  are sets of users who tweet on scientific articles in two different fields  $i$  and  $j$ .

Figure 5 shows a strong overlap (around 70%) between the fields of Medicine and Other Life & Health Sciences. For example, over 70% of the users who tweet a scientific article from the field of Medicine also tweeted a scientific article from the field of Other Life & Health Sciences. The Medicine field has an approximate 35% overlap with Biochemistry, Genetics and Molecular Biology. The second highest overlap (up to 60%) is found in the fields of Biochemistry, Genetics and Molecular Biology and Agriculture, Biological Sciences & Veterinary Science. Furthermore, the field of Material Sciences has a similar overlap (i.e. over 10%) with each of the three fields of Chemistry, Engineering and Physics & Astronomy.



**Figure 5:** Twitter users mentioning scientific articles overlap across 17 different fields, as identified by the All Science Journal Classification



### 3.4 Twitter-user communication via motif identification

Most real-world networks contain recurring patterns known as network motifs, building blocks of networks that occur at numbers higher than those in random networks (Alon, 2007; Milo et al., 2002). The underlying structure of the various natural networks varies; likewise, these may have a distinct network motif. For example, the motifs shared by the Word Wide Web network are unlike the motifs shared by protein-protein interaction networks, nonetheless their identification exposes various kinds of interactions found in real-world networks.

Studying motifs in the ATSN can enable identification of recurring patterns. To do so, the RAND-ESU algorithm developed by Wernicke (2006) was implemented in the FANMOD<sup>2</sup> tool for the analysis. The full enumeration option was chosen during the setup of the tool, and the process generated 1,000 random networks. Only those motifs whose  $Z$  values were greater than 2 were chosen. The value of  $Z$  can be computed as shown in Eq. 2:

$$Z_x = \frac{X - \mu_x}{\sigma_x} \quad (2)$$

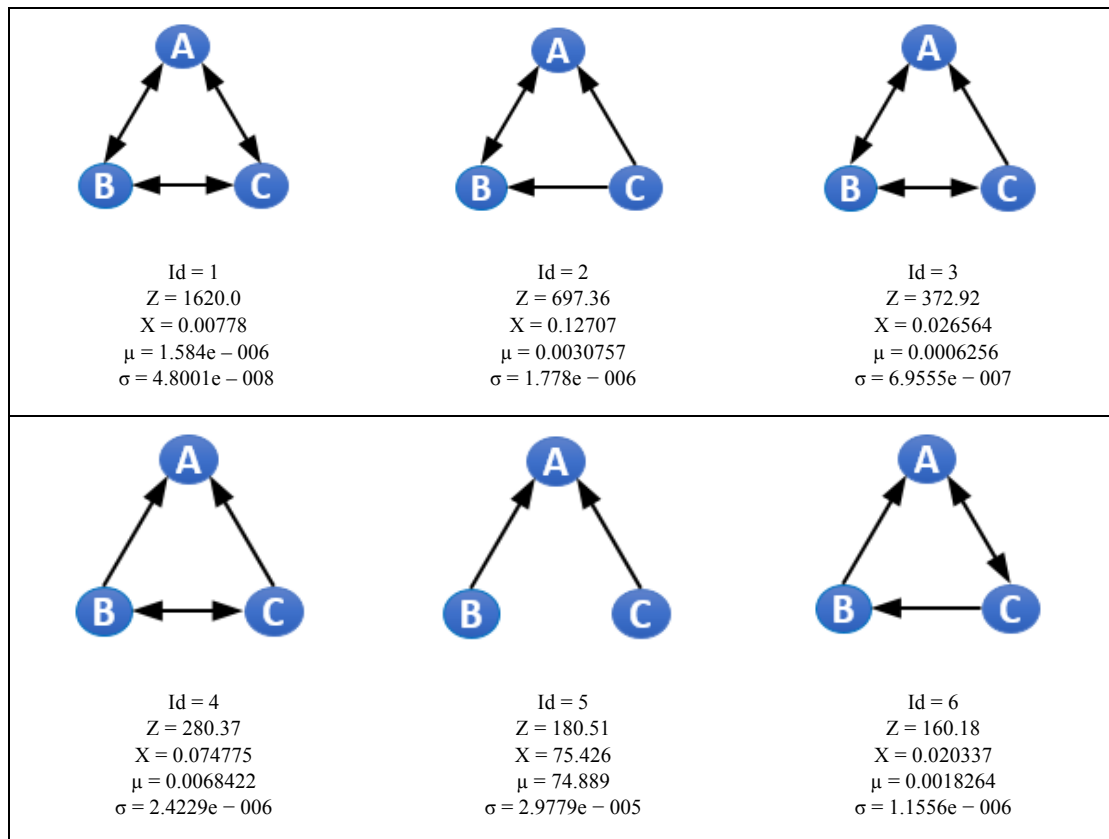
where  $X$  is the number of times motif  $x$  appears in the network,  $\mu$  is the mean number of times it appears in the random graph and  $\sigma$  is the standard deviation of the motif appearing in the random graph, when  $Z < 0$  motifs occur less often than in random networks (meaning that they are not a property of the network being studied), and for  $Z > 0$  the motifs occur more often than in random graphs (meaning that they are a potential feature of the network). A value of  $|Z| > 2$  is significant, as its probability of occurring by chance is  $< 5\%$  (Milo et al., 2002).

---

<sup>2</sup> <http://theinf1.informatik.uni-jena.de/motifs/>

Using the aforementioned FANMOD tool, triplets were detected and six different motifs with  $Z > 2$  were found, as shown in Figure 6. The first motif has the highest  $Z$  value, meaning that it is (bi-directionally) fully connected, which implies that most of the time the users mention each other in a closed circle of the network. For example, the first motif indicates that User A mentions B and C; User B mentions A and C; and User C mentions B and A.

**Figure 6:** Motifs and their corresponding scores in the ATSN.



This type of pattern is quite common in online social networks, due to strong ties and because the strong triadic closure property is satisfied (Easley et al., 2010). The second most common motif demonstrates that less-connected users usually mention the network's influential users; and influential users are more likely to mention each other. In simple terms, popular users always receive more endorsements from their followers

than less popular users. This motif suggests that User C mentions A and B, and Users A and B mention each other, but they do not mention C. This is because Users A and B are influential and known to each other, while User C knows only A and B, and Users A and B do not know C. Motifs 3, 4 and 6 are variants of Motifs 1 and 2. Motif 5 is of interest, because it demonstrates that Users B and C mention A, yet Users B and C do not mention each other. The non-reciprocal nature of Motif 5 indicates that the communication may represent a broadcast rather than a conversation.

When considering all of the motifs presented above in Figure 6, it becomes clear that the commonality between them is the existence of well-known users (IUs). For example, User A is endorsed in every motif created. Such types of nodes, also referred to as hubs, are commonplace in real-world networks. Hubs have significant impact on network topology and serve to distinguish real-world networks from random networks. Furthermore, hubs both play a significant role in information diffusion and influence propagation in social networks.

The second prominent aspect of the ATSN is the strong connectivity among subgroups of nodes, which reflects frequent interaction among users, leading to the community structure of the network. The high  $z$  value of Motif 1 clearly demonstrates the existence of strongly connected components in the ATSN. There may also be more than one hub in a single community, which further helps to increase the interaction ratio among the members of that community. Motifs 2 and 3 can appear in large communities such as Medicine and Environmental Sciences (as shown in Fig. 1).

### 3.5 Small-world property test

Real-world networks differ from random networks in that they have different unique properties. Random networks have a binomial distribution, which can be approximated by a Poisson distribution in the  $k \ll N$  limit (Barabási, 2016). In contrast, the degree of distribution in real-world networks is quite different: most of the real-world networks appear to be scale-free networks and follow a power-law distribution (Barabási, 2009).

Watts and Strogatz (1998) studied the dynamics of small-world networks and presented a model to generate such networks. A small-world network satisfies two properties. First, it has a high average clustering coefficient. This quantifies the connectivity of nodes in their neighborhood and is defined as the fraction of the number of existing edges between neighbors of node  $i$  among all possible edges between these neighbors (Said et al., 2018). The average clustering coefficient is shown in Eq. 3:

$$C = \frac{1}{N} \sum_{i=1}^n C_i \quad (3)$$

where  $C_i, C \in [0,1]$

Second, the short Average Path Length (APL) =  $\log(N)$ , which represents the global-scale property of the network, is defined as the average path length between all possible pairs of the network nodes, as shown in Eq. 4:

$$APL = \frac{2}{N^2} \sum_{u,v \in P} l_{uv} \quad (4)$$

where  $l_{uv}$  represents the shortest path length between node  $u$  and  $v$ ,  $P = \{ (u,v) \mid l_{ij} < \infty ; u,v = 1, \dots, N \}$  and  $N_1 = |P|$ .

As there is no direct method to test a network for small-world properties (Bialonski, Horstmann & Lehnertz, 2010), the following steps were undertaken:

- Computing Average Path Length (APL) and average Clustering Coefficient (CC) of a real network
- Creating an appropriate ensemble of the null model using any random model; a Fast Random Networks Model (Batagelj & Brandes, 2005) was used for this purpose
- Computing  $APL_r$  and  $CC_r$  of the null model
- Computing the normalized shortest path  $\lambda = \frac{APL}{APL_r}$  and  $\gamma = \frac{CC}{CC_r}$
- Checking for the criteria  $\lambda \approx 1$  and  $\gamma > 1$ .

Following the steps mentioned above, the APL of the network was initially determined by taking the Large Connected Component (LCC) into consideration, as the network was disconnected. Note that the LCC covers more than 98% of the nodes of the network. To generate an appropriate ensemble, a Fast Random Networks Model was used with the same number of nodes, and the probability was set to 0.00048. There were 100 random networks generated; the APL and average LCC were computed in order to compare these to the original graph. The experiment resulted in a 0.913 value for  $\lambda$  and 213.55 for  $\gamma$ , which demonstrates that the ATSN satisfies small-world characteristics and has a structure unlike that of a random network.

## 4 Concluding Remarks

This study explored multiple aspects of the ATSN using an SNA approach. It was found that organizational accounts associated with highly reputable journals, such as *PLOS One*, *Nature* and *Science*, play an essential role. The authors showed that, due to these

accounts' influence, the network forms a giant component covering more than 98% of nodes. Moreover, the community structure of the network was examined and found to have field-wise high intra-connectivity, resulting in a field-wise community structure. Large communities are dominated by organizational accounts associated with journals, while small communities are dominated by experts in the field.

As expected, substantial overlap was found between relevant fields, for example, the field of Medicine had a 70% overlap with Health Sciences, while the field of Biochemistry, Genetics & Molecular Biology had a 60% overlap with Agriculture, Biological Sciences & Veterinary Science. It was discovered that users in Medicine-related fields demonstrate a much greater overlap than users in the field of Engineering. Finally, users who were active in fields such as Social Sciences, Earth & Planetary Sciences and Economics, Business & Decision Sciences showed no significant overlap with other fields.

Overall, this work demonstrated a novel approach to examining the ATSN. We showed that Twitter-based social media communities have different characteristics. While some communities are highly interconnected, others are highly coupled yet have low interconnectivity. Such characteristics may affect social media usage counts, either directly or indirectly. Instead of regarding altmetrics as a black box, researchers and consumers of altmetrics should consider the underlying social media networks that may be either inflating or deflating the measures of social usage. Therefore, a more comprehensive examination is advised before adopting these very promising altmetrics indices.

In future, the authors plan to normalize the effect of influential users that may give rise to bias in generating social usage data in Twitter-based altmetrics. Future work will include a comparison of various community-detection algorithms using the ATSN. Finally, future studies should be conducted on even larger data, irrespective of citation count, as this is one of the potential limitations of existing studies.

## Acknowledgements

The research has been supported by the National Research Programme for Universities grant No. 6857/Punjab/NRPU/R&D/HEC/2016, funded by the Higher Education Commission of Pakistan, with Dr. Saeed Ul Hassan as Principal Investigator.

## References

- Abbasi, A., Altmann, J., & Hossain, L. (2011). Identifying the effects of co-authorship networks on the performance of scholars: A correlation and regression analysis of performance measures and social network analysis measures. *Journal of Informetrics*, 5(4), 594–607.
- Alon, U. (2007). Network motifs: theory and experimental approaches. *Nature Reviews Genetics*, 8(6), 450.
- Alperin, J. P., & Haustein, S. (2017). Applying social network analysis to explore Twitter diffusion patterns. In Altmetrics17 Workshop.
- Barabási, A. L. (2009). Scale-free networks: a decade and beyond. *Science*, 325(5939), 412-413.
- Barabási, A.-L. (2016). Network Science. Cambridge University Press.
- Bastian, M., Heymann, S., & Jacomy, M. (2009, March). Gephi: an open source software for exploring and manipulating networks. In Third international AAAI conference on weblogs and social media.
- Batagelj, V., & Brandes, U. (2005). Efficient generation of large random networks. *Physical Review E*, 71(3), 036113.
- Bialonski, S., Horstmann, M.-T., & Lehnertz, K. (2010). From brain to earth and climate systems: Small-world interaction networks or not? *Chaos: An*

- Interdisciplinary Journal of Nonlinear Science*, 20(1), 013134.
- 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.
- Bonacich, P. (2007). Some unique properties of eigenvector centrality. *Social Networks*, 29(4), 555–564.
- Borgatti, S. P. (2005). Centrality and network flow. *Social Networks*, 27(1), 55–71.
- Bornmann, L., & Haunschild, R. (2018). Allegation of scientific misconduct increases Twitter attention. *Scientometrics*, 115(2), 1097–1100.
- Carrington, P. J., Scott, J., & Wasserman, S. (2005). *Models and methods in social network analysis* (Vol. 28). Cambridge university press.
- Costas, R., van Honk, J., & Franssen, T. (2017). Scholars on Twitter: who and how many are they?. arXiv preprint arXiv:1712.05667.
- Didegah, F., & Thelwall, M. (2018). Co-saved, co-tweeted, and co-cited networks. *Journal of the Association for Information Science and Technology*.
- Dinsmore, A., Allen, L., & Dolby, K. (2014). Alternative perspectives on impact: the potential of ALMs and altmetrics to inform funders about research impact. *PLoS Biology*, 12(11), e1002003.
- Easley, D., & Kleinberg, J. (2010). *Networks, crowds, and markets: Reasoning about a highly connected world*. Cambridge University Press.
- Fortunato, S. (2010). Community detection in graphs. *Physics Reports*, 486(3–5), 75–174.
- Haddawy, P., Hassan, S. U., Asghar, A., & Amin, S. (2016). A comprehensive examination of the relation of three citation-based journal metrics to expert judgment of journal quality. *Journal of Informetrics*, 10(1), 162-173.
- Hassan, S.-U., Imran, M., Gillani, U., Aljohani, N. R., Bowman, T. D., & Didegah, F. (2017). Measuring social media activity of scientific literature: an exhaustive comparison of Scopus and novel altmetrics big data. *Scientometrics*, 113(2), 1037–1057.
- Haustein, S., Bowman, T. D., Holmberg, K., Tsou, A., Sugimoto, C. R., & Larivière, V. (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.
- Hoffmann, C. P., Lutz, C., & Meckel, M. (2016). A relational altmetric? Network



- centrality on Research Gate as an indicator of scientific impact. *Journal of the Association for Information Science and Technology*, 67(4), 765–775.
- Imran, M., Akhtar, A., Said, A., Safder, I., Hassan, S.-U., & Aljohani, N. R. (2018). Exploiting Social Networks of Twitter in Altmetrics Big Data. In *23rd International Conference on Science and Technology Indicators (STI 2018), September 12-14, 2018, Leiden, The Netherlands*. Centre for Science and Technology Studies (CWTS).
- Jaccard, P. (1901). Distribution de la flore alpine dans le bassin des Dranses et dans quelques régions voisines. *Bull Soc Vaudoise Sci Nat*, 37, 241-272.
- Jordan, K. (2014). Academics and their online networks: Exploring the role of academic social networking sites. *First Monday*, 19(11).
- Jordan, K. (2017). *Understanding the structure and role of academics' ego-networks on social networking sites* (PhD Thesis). The Open University.
- Lutz, C., & Hoffmann, C. P. (2018). Making academic social capital visible: Relating SNS-based, alternative and traditional metrics of scientific impact. *Social Science Computer Review*, 36(5), 632–643.
- Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., & Alon, U. (2002). Network motifs: simple building blocks of complex networks. *Science*, 298(5594), 824–827.
- Moed, H. F. (2010). Measuring contextual citation impact of scientific journals. *Journal of Informetrics*, 4(3), 265–277.
- Otte, E., & Rousseau, R. (2002). Social network analysis: a powerful strategy, also for the information sciences. *Journal of Information Science*, 28(6), 441–453.
- Piwowar, H. (2013). Altmetrics: Value all research products. *Nature*, 493(7431), 159.
- Priem, J., & Hemminger, B. H. (2010). Scientometrics 2.0: New metrics of scholarly impact on the social Web. *First Monday*, 15(7).
- Radicchi, F., Fortunato, S., & Castellano, C. (2008). Universality of citation distributions: Toward an objective measure of scientific impact. *Proceedings of the National Academy of Sciences*, 105(45), 17268–17272.
- Said, A., Abbasi, R. A., Maqbool, O., Daud, A., & Aljohani, N. R. (2018). CC-GA: A clustering coefficient based genetic algorithm for detecting communities in social networks. *Applied Soft Computing*, 63, 59–70.
- Thelwall, M., & Nevill, T. (2018). Could scientists use Altmetric. com scores to predict longer-term citation counts? *Journal of Informetrics*, 12(1), 237–248.

- Vainio, J., & Holmberg, K. (2017). Highly tweeted science articles: who tweets them? An analysis of Twitter user profile descriptions. *Scientometrics*, 112(1), 345–366.
- Wasserman, S., & Faust, K. (1994). *Social network analysis: Methods and applications* (Vol. 8). Cambridge university press.
- Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of ‘small-world’ networks. *Nature*, 393(6684), 440.
- Wernicke, S. (2006). Efficient detection of network motifs. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 3(4), 347–359.
- Wilsdon, J. (2016). *The metric tide: Independent review of the role of metrics in research assessment and management*. Sage.
- Wouters, P., & Costas, R. (2012). *Users, narcissism and control: tracking the impact of scholarly publications in the 21st century*. SURFfoundation Utrecht.
- Yan, W., & Zhang, Y. (2018). Research universities on the ResearchGate social networking site: An examination of institutional differences, research activity level, and social networks formed. *Journal of Informetrics*, 12(1), 385–400.
- Yang, J., McAuley, J., & Leskovec, J. (2013). Community detection in networks with node attributes. In *Data Mining (ICDM), 2013 IEEE 13th international conference on* (pp. 1151–1156). IEEE.
- Yu, H., Xu, S., Xiao, T., Hemminger, B. M., & Yang, S. (2017). Global science discussed in local altmetrics: Weibo and its comparison with Twitter. *Journal of Informetrics*, 11(2), 466–482.
- Zahedi, Z., & Haustein, S. (2018). On the relationships between bibliographic characteristics of scientific documents and citation and Mendeley readership counts: A large-scale analysis of Web of Science publications. *Journal of Informetrics*, 12(1), 191–202.

## Appendix A

This section presents the top 5 influential users from each community with their fields and types. Field attribute represents the subject-field or domain of the user, which helps to understand the community structure of the network. Note that subject-field and/or domain is manually extracted from the Twitter users' profiles to properly label the influential users. There are eight major communities which cover more than 98% of users of the whole network.

**Table A-1:** Influential users with their fields and types. Community # is the community identifier assigned by Gephi as shown in Figure 1. Nodes represent screen-names of Twitter users. Some abbreviations include HPN (Health Profession & Nursing), ORG (Organization), Sci. (Science), Med. (medicine), Pro. (Professional), Res. (Research), Env. (Environmental), and BNC (BraveNewClimate).

Community #	Nodes	Type	Field	Community #	Nodes	Type	Field
22	Jama current	Journal	Medicine	4	Annals Oncology	Journal	Oncology
22	TheLancet	Journal	Medicine	4	EUplatinum	Journal	Urology
22	TheBMJ	Journal	Medicine	4	BldCancerDoc	Researcher/Pro.	Medicine
22	NEJM	Journal	Medicine	4	UroWeb	ORG	Urology
22	WHO	Health org	Medicine	4	LNelsonMD	Med Pro.	Health Sci.
3	FabianWadsWorth	Res.	Env-Sci.	37	resiapretorius	Researcher	Social Sci.
3	DrHelenMcGregor	Researcher	Env-Sci.	37	DDPSCmaker	Res. group	Social Sci.
3	EnviroTaff	Researcher	Env-Sci.	37	EricTopol	Researcher	Psychology
3	PdeMenocal	Researcher	Env-Sci.	37	mwilsonsayres	Researcher	Social Sci.
3	DrHelenMc- Gregor	Researcher	Env-Sci.	37	Graham Coop	Researcher	Social Sci.
20	PLOSONE	Journal	Sci. & Med.	5	riotta	Researcher	Data Sci.
20	ESAFrontiers	Journal	Biological Sci.	5	prdeville	Researcher	Data Sci.
20	NYCuratrix	Researcher	Biological Sci.	5	barabasi	Researcher	Data Sci.
20	ehekkala	Researcher	Biological Sci.	5	shannimcg	Researcher	Data Sci.
20	PestSmartCRC	ORG	Biological Sci.	5	bonstewart	Researcher	Social Sci.
24	Liver4Kids	Researcher	HPN	32	JNeurophysiol	Journal	Neuro
24	CincyChildrens	Health org	HPN	32	adrianhaith	Researcher	Health Sci.
24	UBC	University	HPN	32	blamlab	LAB	Health
24	Gretaknits	Researcher	HPN	32	spornslab	LAB	Health
24	jclinicalinvest	Journal	HPN	32	introspection	LAB	Health