STI 2018 Conference Proceedings

Proceedings of the 23rd International Conference on Science and Technology Indicators

All papers published in this conference proceedings have been peer reviewed through a peer review process administered by the proceedings Editors. Reviews were conducted by expert referees to the professional and scientific standards expected of a conference proceedings.

Chair of the Conference

Paul Wouters

Scientific Editors

Rodrigo Costas
Thomas Franssen
Alfredo Yegros-Yegros

Layout

Andrea Reyes Elizondo
Suze van der Luijt-Jansen

The articles of this collection can be accessed at https://hdl.handle.net/1887/64521

ISBN: 978-90-9031204-0

© of the text: the authors
© 2018 Centre for Science and Technology Studies (CWTS), Leiden University, The Netherlands

This ARTICLE is licensed under a Creative Commons Atribution-NonCommercial-NonDetivates 4.0 International Licensed

23rd International Conference on Science and Technology Indicators (STI 2018)
"Science, Technology and Innovation indicators in transition"
12 - 14 September 2018 | Leiden, The Netherlands
#STI18LDN

Exploiting Social Networks of Twitter in Altmetrics Big Data

Mubashir Imran*, Aqsa Akhtar*, Anwar Said*, Iqra Safder*, Saeed-Ul Hassan* and Naif Radi Aljohani**

*mubashir.imran@itu.edu.pk; mscs15059@itu.edu.pk; anwar.said@itu.edu.pk; iqra.safder@itu.edu.pk; saeed-ul-hassan@itu.edu.pk
Information Technology University, Lahore (Pakistan)

** nraljohani@kau.edu.sa
King Abdulaziz University, Jeddah (Saudi Arabia)

Background

Social media platforms have enabled online users to disseminate information across the web. This massive diffusion of information by online users has resulted in the term "user-generated content" (Lee, 2011). This content is aligned to the opinions and interests of the social media networks' various communities, which presumably have similar interests. More recently, scientific communities have started actively to adopt social media platforms to emulate the impact and influence of scholarly literature (Priem & Bradely, 2010).

Altmetrics captures the opinions on scholarly literature shared by these online scientific communities (Priem et al., 2010). It has allowed the timely capture and measurement of scholarly communication over the web and provided a web trace of the social media activities undertaken by the scholarly community. The activities of over a dozen social media platforms are captured by Altmetric.com, including Facebook, Twitter, Google+ and Wikipedia. Altmetric.com started by capturing data from various online platforms, and it is one of the largest Altmetrics data aggregators. According to Hassan et al. (2017), Twitter has the highest coverage, having 91% of the social activity trace that is monitored by altmetric.com.

In recent years, several attempts have been made to explore aspects of tweets in Altmetrics data. Liu and Fang (2017) presented a methodology for scoring tweets on the basis of the sentiments contained. The proposed approach extracts tweets associated with just the top 100 articles as scored by altmetric.com. Each tweet is assigned a weight on the basis of the sentiments exhibited, such as neutral, praise, agreement, interest, surprise, recommendation or expansion. The designed system deliberately overlooked negative or controversial tweets. By applying sentiment analysis tools such as SentiStrength (Thelwall et al., 2010) and Sentiment140 (Go, 2016) to these scholarly tweets, Konkiel (2017) revealed the limitations of these tools. Haustein et al. (2016) manually identified bots in the Twitter dataset, and argued that automated Twitter accounts that publish scholarly tweets behave differently from general Twitter bot accounts. Their study also showed that over 9% of the total tweets in 2012 arXiv submissions were generated by bot accounts. Similarly, Costas et al. (2017) presented an approach to identify active scholars on Twitter. They claim that 2% of all scholars on the Web of Science (WoS) are active on Twitter, with the largest populations being of scholars in the fields of social sciences and humanities.

Given the recognized needs of the scientometrics community and its recent interest in the advancement of social media platforms to complement traditional, bibliometric-based scientific assessments, we[1] explored the behaviour and properties of scholarly communities on Twitter. In this study, we examined the chief commonalities and differences of Twitter-based social media activity by users across 17 broader disciplines by employing a dataset of over 800k tweets.

Data and Method

In this paper, we used a dataset of 4.5 million JSON files, originally obtained from Altmetric.com (version jun-4-2016.tar.gz). Each article in the Altmetrics database is associated with a unique identifier, *altmetric_id*. Each file in the dataset (identified by *altmetric_id*) contains one or more pieces of social interaction information. For this study, we used only Altmetrics data mapped to 2015. We extracted 884,048 tweets' text using tweet_IDs associated with *altmetric_id*, from Twitter.com via Twitter API. Further, we processed the tweets to separate out the mentioned and the retweeted users' names from the tweets. Note that tweets that contained no mentioned user (identified by '@') and those that were retweets (identified by 'RT') were deleted from the dataset. Moreover, all rows in which the original users had mentioned or retweeted their own tweets were removed. Next, we extracted edge lists representing a weighted graph of associations between tweeter, retweeter and mentioned user. Note that our Twitter-based network consists of 2,71,582 unique Twitter users (nodes), along with 27,07,684 links (edges) between the nodes. Lastly, the graph was further divided into 17 disciplines, using the All Subject Journal Classification (ASJC) mapping employed by Scopus.

In order to analyse the various aspects of graphs formed by the network of tweeters for each respective discipline and their behaviour, we employed a state-of-the-art graph visualization open-source software *Gephi* (Bastian et al., 2009). The graph's properties, along with abbreviations and references to the used algorithm, are provided below.

- Average Degree (AD): In a network, the Average Degree of a graph is defined as the average number of connections that a node has with other nodes (Lancichinetti & Fortunato, 2009).
- Weighted Average Degree (WAD): Similar to the Average Degree, the Weighted Average Degree is calculated from the average number of connections that a node has with another node in a network, where the weight is commonly defined as the total number of edges for a particular node (Lancichinetti & Fortunato, 2009).
- Graph Density (GD): A graph is referred to as a complete graph if each node is connected to every other node via an edge. Graph Density takes this property into consideration and counts the number of edges in a graph to compare how close a graph is to a complete graph.
- Modularity (Mod.): Modularity in a graph is a measure of strength for nodes forming modules or clusters in the network. A group of nodes that are densely connected generates a high modularity value; however, there is a possibility that these groups of nodes have sparsely interlinked connections (Blondel et al., 2008).

[1] This work was partially supported by faculty research & development funds at the Information Technology University. The authors are grateful to Digital Science & Research Solutions Inc., which provided Altmetrics data for this research free of charge.

- Connected Component (CComp): The Connected Component is a sub-graph of an undirected graph, having a path that connects every other node in that sub-graph (Robert, 1972).
- Clustering Coefficient (CCoef.): The Clustering Coefficient is the measure of the degree to which all the nodes in the graph form clusters (Latapy, 2008).
- Eigenvector Centrality (EC): Eigenvector Centrality is the amount (measure) of influence that a specific node has in a network, based on its connections (Bonacich, 2007).
- Average Path Length (APL): Path length is defined as the longest path between two nodes in a network. Note that the Average Path Length provides an impression of information dissemination in a network (Brandes, 2001).

In addition to above quantitative measures, we visualized the interaction among Twitter users across the disciplines. For this, we formulated clusters (sub-communities) within each discipline by computing the modularity of the respective network. Each cluster, identified by a different colour, represents a sub-community within a discipline.

Results and Discussion

Overall, we analysed our data in two ways: a) we presented an array of quantitative measures to study social communities' interactions regarding scholarly literature across disciplines; b) we visualized networks of selected disciplines to study the community structures and interactions across disciplines.

Table 1: Network properties of communities across disciplines

Disciplines	**Node**	**Edges**	**AD**	**Mod.**	**CComp.**	**CCoef.**
Other Life & Health Sciences	230211	709677	5.331	0.703	5035	0.38
Medicine	169260	483523	5.713	0.699	3854	0.389
Biochemistry, Genetics & Molecular Biology	75695	192289	5.081	0.676	2524	0.406
General (*Science*, *Nature*, *PNAS*)	57981	107099	1.847	0.664	2063	0.406
Agricultural, Biological Sciences & Veterinary	56419	133855	4.745	0.69	2154	0.409
Social Science	49209	82420	3.35	0.85	2538	0.424
Health Professions & Nursing	29617	60902	4.113	0.767	1196	0.396
Environmental Sciences	19870	40373	2.032	0.773	981	0.156
Engineering	15600	26439	1.695	0.787	949	0.146
Earth & Planetary Sciences	11718	19980	3.41	0.798	528	0.43
Chemistry	10432	17612	3.377	0.755	609	0.441
Physics & Astronomy	8078	11863	2.937	0.839	623	0.486
Economics, Business & Decision Sciences	7697	9963	2.589	0.938	726	0.471
Computer Science	6300	8582	2.724	0.868	623	0.423
Mathematics	5612	9488	3.381	0.782	385	0.45
Materials Sciences	5192	9619	3.254	0.77	420	0.522
Arts and Humanities	3917	4957	2.531	0.948	449	0.511

Table 1 shows the network properties of Twitter users across disciplines. Disciplines such as Other Life & Health Sciences (*Nodes*: 169,260, *Edges*: 483,523) and Medicine (*Nodes*: 230,211, *Edges*: 709,677) form the largest and denser networks, having the greatest number of nodes and edges in their graph, with an average degree of greater than 5. The research disciplines such as Arts & Humanities (*Modularity*: 0.948) and Economics, Business & Decision Sciences (*Modularity*: 0.948) have the highest modularity and form more distinct sub-communities within their network than other disciplines, which are more interleaved. These disciplines also possess a high clustering coefficient, clearly showing that the sub-communities within each discipline are highly connected to each other.

Figure 1 is a graphical demonstration of some tightly coupled disciplines with low interconnectivity across the communities: Arts & Humanities; and Economics, Business & Decision Sciences. We found that the social media communities for both these disciplines are highly interconnected, resulting in fewer interactions across the clusters with other disciplines.

Figure 1: Disciplines with low interconnected communities

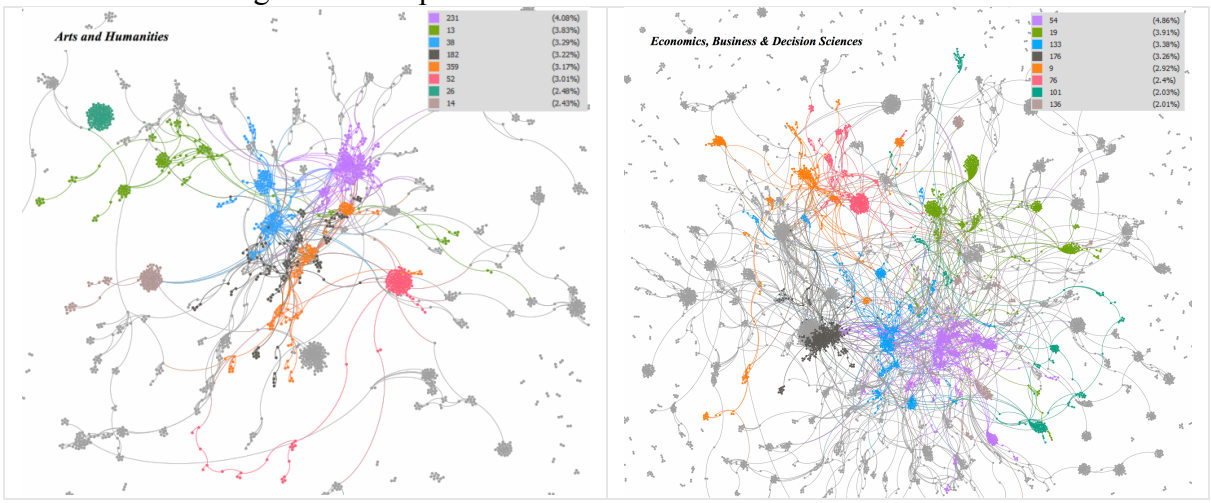
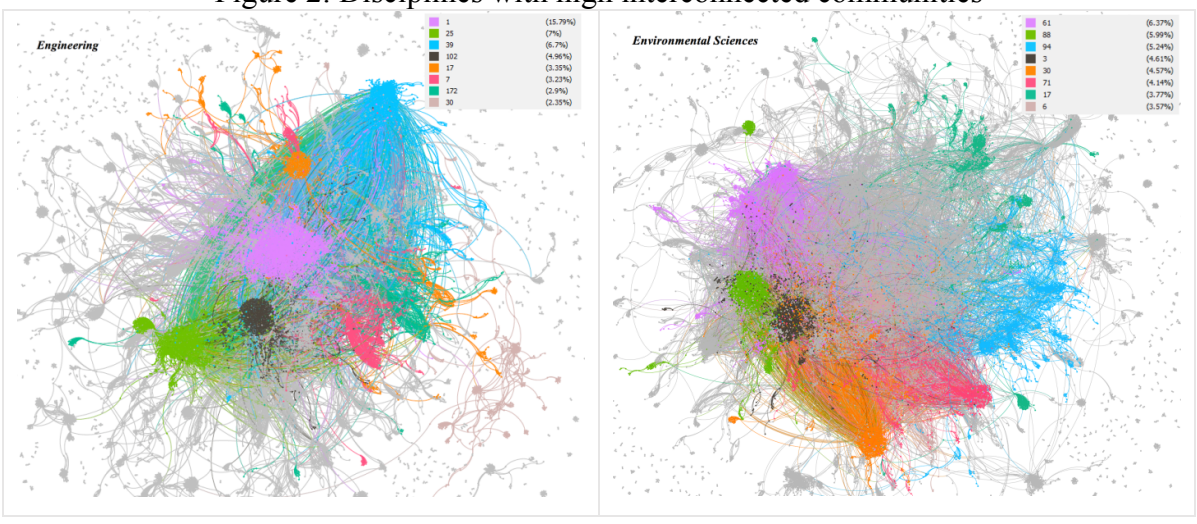


Figure 2: Disciplines with high interconnected communities



By contrast, in terms of their Twitter-based social media networks, Engineering and Environmental Sciences appear to be two highly interconnected disciplines. Figure 2 shows the significant connections between clusters 1 and 39, and clusters 25 and 102 in Engineering. We observed that more than 15% of all Twitter users are grouped under a single cluster; that is, cluster # 1, in Engineering. Similarly, a strong interconnectivity is seen in the Environmental Sciences discipline.

Our results indicate that, among the social network communities that interact concerning scholarly literature, Arts, Humanities, Economics, Business & Decision Sciences and Computer Science tend to form smaller sub-communities. These communities share and communicate their opinions regarding scholarly work within only a small, selective group. They do not interact across the network. By contrast, the social media communities that interact in disciplines such as Engineering, Environmental Sciences and Medicine appear to be highly coupled. In addition, the various communities in these disciplines actively interact with each other.

Concluding Remarks

We present a novel way to examine Twitter-based social media networks. We show that Twitter-based social media communities have dissimilar characteristics. While some communities are highly interconnected, such as Engineering and Environmental Sciences, others are highly coupled yet have low interconnectivity, such as Arts & Humanities and Economics, Business & Decision Sciences. We believe that such characteristics may affect social media usage counts, either directly or indirectly. We argue that, instead of regarding Altmetrics as a black box, we need to scrutinize the underlying social media networks that may be inflating or deflating social usage. Thus, a more comprehensive examination is advised before the adoption of these very promising Altmetrics data.

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.

Bonacich, P. (2007). Some unique properties of eigenvector centrality. *Social networks*, *29*(4), 555-564.

Brandes, U. (2001). A faster algorithm for betweenness centrality. *Journal of mathematical sociology*, *25*(2), 163-177.

Costas, R., van Honk, J., & Franssen, T. (2017). Scholars on Twitter: who and how many are they?. *arXiv preprint arXiv:1712.05667*.

Go, A., Bhayani, R., & Huang, L. (2016). Sentiment140. *Site Functionality, 2013c. URL http://help. sentiment140. com/site-functionality. Abruf am*, *20*.

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 T witter. *Journal of the Association for Information Science and Technology*, *67*(1), 232-238.

Konkiel, S. (2017). Adapting sentiment analysis for tweets linking to scientific papers. *The Idealis*.

Lancichinetti, A., & Fortunato, S. (2009). Benchmarks for testing community detection algorithms on directed and weighted graphs with overlapping communities. *Physical Review E*, *80*(1), 016118.

Latapy, M. (2008). Main-memory triangle computations for very large (sparse (power-law)) graphs. *Theoretical computer science*, *407*(1-3), 458-473.

Lee, E. (2011). Facilitating student-generated content using Web 2.0 technologies. *Educational Technology*, 36-40.

Liu, X. Z., & Fang, H. (2017). What we can learn from tweets linking to research papers. *Scientometrics*, *111*(1), 349-369.

Bastian, M., Heymann, S., & Jacomy, M. (2009). Gephi: an open source software for exploring and manipulating networks. *Icwsm*, *8*(2009), 361-362.

Priem, J., & Hemminger, B. H. (2010). Scientometrics 2.0: New metrics of scholarly impact on the social Web. *First Monday*, *15*(7).

Priem, J., Taraborelli, D., Groth, P., & Neylon, C. (2010). Altmetrics: A manifesto.

Tarjan, R. (1972). Depth-first search and linear graph algorithms. *SIAM journal on computing*, *1*(2), 146-160.

Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., & Kappas, A. (2010). Sentiment strength detection in short informal text. *Journal of the American Society for Information Science and Technology*, *61*(12), 2544-2558.