Social Activity versus Academic Activity: A Case Study of Computer Scientists on Twitter

Subhash Chandra Pujari
Graz University of Technology
Graz, Austria
s.pujari@tugraz.at

Asmelash Teka Hadgu
L3S Research Center
Hannover, Germany
teka@l3s.de

Elisabeth Lex
Graz University of Technology
Graz, Austria
elisabeth.lex@tugraz.at

Robert Jäschke
L3S Research Center
Hannover, Germany
jaeschke@l3s.de

ABSTRACT

In this work, we study social and academic network activities of researchers from Computer Science. Using a recently proposed framework, we map the researchers to their Twitter accounts and link them to their publications. This enables us to create two types of networks: first, networks that reflect social activities on Twitter, namely the researchers' follow, retweet and mention networks and second, networks that reflect academic activities, that is the co-authorship and citation networks. Based on these datasets, we (i) compare the social activities of researchers with their academic activities, (ii) investigate the consistency and similarity of communities within the social and academic activity networks, and (iii) investigate the information flow between different areas of Computer Science in and between both types of networks. Our findings show that if co-authors interact on Twitter, their relationship is reciprocal, increasing with the numbers of papers they co-authored. In general, the social and the academic activities are not correlated. In terms of community analysis, we found that the three social activity networks are most consistent with each other, with the highest consistency between the retweet and mention network. A study of information flow revealed that in the follow network, researchers from Data Management, Human-Computer Interaction, and Artificial Intelligence act as a source of information for other areas in Computer Science.

CCS Concepts

•**Information systems** → *Social networks;*

Keywords

Twitter, Computer Science, Science 2.0

i-KNOW '15, October 21-23, 2015, Graz, Austria

DOI: http://dx.doi.org/10.1145/2809563.2809584

1. INTRODUCTION

Twitter has been and still is increasingly used by researchers to disseminate information [7, 12]. In parallel, a large body of research has grown analyzing this emerging communication and dissemination platform. E.g., there are works that study the usage of Twitter during Computer Science conferences to understand the motivation of using Twitter during conferences [12], others investigate the Twitter usage in scientific contexts as a mean for scientific communication [20].

All these studies describe Twitter as an efficient and fast way to communicate new findings to the research community across different areas of research. This is facilitated by sharing of comments, links to web pages, or multimedia content and through tagging them with hashtags. Unlike existing studies on Twitter usage during scientific conferences, we are interested in understanding general usage patterns within a specific discipline, i.e., Computer Science. In particular, we aim to compare traditional research activities and interactions like publishing and co-authoring papers with activities and interactions on Twitter like tweeting and following. We focus on the following research questions:

Researcher activity: How related are activities and interactions in research to those on Twitter?

Communities of researchers: How do researchers form communities and how consistent are those communities across different networks on Twitter and traditional academic networks?

Information flow between research areas: Which patterns of information flow can be found between different areas of Computer Science?

On Twitter, we study networks of users generated by actions like following, retweeting, and mentioning. The analysis of these interactions can help us better understand Twitter usage among researchers. For comparison with the researchers' 'real-life', we consider the co-author and citation networks induced by the publications they have written. We anticipate that the findings of such an analysis are a good starting point to develop methods to improve information sharing in the context of "Science 2.0". For instance, understanding the interaction of researchers on Twitter can help design algorithms for user recommendation. To the best of our knowledge, no previous works have compared the activity

and interaction of researchers from Computer Science on Twitter with traditional counterparts in research.

The main challenge in conducting such a study is the unavailability of a large-scale directory that contains information about researchers on Twitter. Thus, we exploit the approach presented in [8] for generating a list of researchers in Computer Science on Twitter. We further extend this method by mapping researchers to areas of Computer Science based on their publication venues. We also gather citation information and compare the corresponding network of researchers to the Twitter networks. Therefore, we compare two different clustering methods on the four networks (retweet, mention, follow, citation) using five different methods to evaluate the similarity of clusterings.

The main findings of this work are:

- In general, the activity of researchers on Twitter is weakly correlated to their activity in academia. However, there is a high correlation between their interaction in research and on Twitter. Specifically, it is more likely that co-authors interact with each other on Twitter than random pairs of researchers.
- The communities based on the Twitter networks are more similar to each other than to those based on the citation network. Furthermore, the community structure is more consistent between the retweet and mention network as compared to the other networks.
- Some areas of Computer Science (*Artificial Intelligence*, *Data Management*, *Human-Computer Interaction*) act as a main source of information for other areas. A high information flow can be found between related areas like *Software Engineering* and *Programming Languages*.

This paper is structured as follows: In Section 2 we review related work that is analyzing researchers on Twitter. In Section 3 we describe the collection of the data and the construction of the networks. The mapping of researchers to their research areas is then described in Section 4. Section 5 presents our experiments and their results. We conclude the paper and discuss future work in Section 6.

2. RELATED WORK

At present, we identify three main lines of research that are related to our work: Twitter usage of researchers, especially during conferences, group formation in scientific networks and information flow between areas of science.

Twitter Usage of Researchers.. In many academic disciplines, conferences serve as a platform for interaction and to foster collaboration between researchers. DeVocht et al. [5] visualized collaborations and online social interactions at scientific conferences in the light of scholarly networking. Their aim is to illustrate the various ways of scientific interaction by aligning co-authorship, citation and social media based networks in one visualization. This helps investigate conferences and researchers from multiple perspectives, based on their collaborations and online interactions. Works studying researchers from the Semantic Web community during a conference and their usage of different media for information dissemination found that Twitter is one of the top three services used by that community to spread information [11, 12]. Unlike traditional media of information dissemination, researchers set up accounts on Twitter to reach a wider audience outside the realm of their own area of research. Different motivations of researchers using Twitter were studied in [7] by monitoring their activity during the ED-MEDIA 2009 conference. They found that researchers use Twitter in many ways including commenting, sharing, and arranging of online and offline collaborations, thus Twitter can contribute to strengthen a scientific community. Mazarakis et al. [14] have investigated Twitter activity and tweet content of researchers tweeting during the Science 2.0 conference 2014. They found that the researcher-specific tweeting behavior follows a power law, i.e., only a few users tweeted often during the conference, while the majority of users tweeted only occasionally. The authors categorized the tweets according to their purpose and they found that over 80% of the tweets either reported concrete contents of the conference such as information about a presentation, or they were used to share resources by posting links to, e.g., Slideshare or Figshare.

Group Formation in Scientific Networks. Another line of research that is related to our work is linked to analyzing the interaction between groups of researchers in scientific networks. In [2], they studied citation networks with the goal to identify different fields of science, their structure, size and interconnectedness. Jung et al. [10] studied the formation of communities in citation networks around research fields and researchers and how these communities change over time. They found that citation networks can be used to successfully predict scientific communities in citation networks up to five years into the future using link prediction and community detection methods. In our work, we also aim to understand interaction patterns between groups of researchers, as well as communities.

Information Flow Between Areas of Science. Also related to our work is the analysis of the information flow between different fields of science. In [18] the information flow between different fields of Computer Science was studied using publication venues as a source for identifying research areas. A finding that is useful to and further confirmed by our study in the case of the Twitter network is that more theoretical research areas like *Algorithms and Theory* have less information flow whereas applied areas like *Data Management* have more information flow from and to other areas of Computer Science. We are not aware of similar studies that map the flow of information between different areas of (Computer) Science on Twitter.

3. DATASETS AND NETWORKS

In this section, we describe how we collected our datasets and how we constructed the networks we used in our experiments. The overall approach and the experiments we carried out in this work are depicted in Figure 1.

Dataset Collection. Our work is based on the dataset that was created and published in [8] which is available on GitHub.[1] Between 11/2013 and 01/2014 Twitter accounts of users who followed the Twitter handles of popular Computer Science conferences were extracted and mapped to author profiles in DBLP. We here briefly recap how this dataset was created. Based on a list from Wikipedia[2] the official Twitter accounts of the conferences, e.g., *@www2015*, were identifed by a) performing a *Web search* to find the official Web page of the conference and extract its Twit-

[1] https://github.com/l3s/twitter-researcher

[2] http://en.wikipedia.org/wiki/List_of_computer_science_conferences

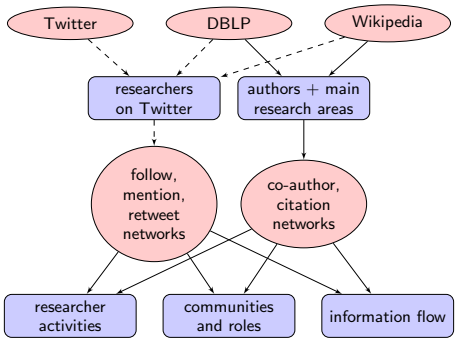


Figure 1: **Our approach and datasets**. Red circles denote the datasets that were used and the networks that were generated based on them. Blue rectangles represent the outcomes of our experiments. The dashed arrows correspond to task that we completed in previous work [8], the solid arrows to the experiments that this work particularly focuses on.

ter account from the page, and b) building potential screen names from acronyms of conferences with years appended, such as, e.g., *www2015, iknow2015*, and performing a *Twitter search* to check whether these accounts exist. Based on the identified 170 conference Twitter accounts (of 98 conferences), *candidate users* were collected, i.e., the Twitter accounts which follow or retweet any of the conference accounts. This resulted in a dataset that contained 52 678 Twitter users. These candidates were then mapped to their author profile in DBLP,[3] if possible. To avoid ambiguities, researchers with a homonym on Twitter or in DBLP were removed from the dataset. In the end, 9 191 of the candidates could be linked to their corresponding author profile. A manual evaluation of a random sample of 150 of the 9 191 users showed an accuracy of 73% with many of the remaining users also being researchers in Computer Science.

Constructing Networks and Representation. For our experiments, we constructed five different networks: the (i) follow, (ii) retweet, and (iii) mention networks of the 9 191 identified researchers on Twitter, as well as their (iv) citation, and (v) co-authorship networks. For (i)–(iii), we crawled the required information, i.e., followers, followees, and the researchers' tweets using the Twitter API. For (iv)–(v), we acquired the authorship and citation information from the DBLP dataset[4] that is provided by ArnetMiner [19]. The DBLP dataset is updated every month and consist of of 1 287 395 publications.

The networks are represented as weighted and directed graphs (except for the co-author network) $G_\alpha = (V, E)$ such that V is a set of vertices, E a set of edges, and $\alpha \in \{\text{follow, retweet, mention, co-author, citation}\}$. In the case of a retweet network, a researcher $u \in V$ retweeting another researcher $v \in V$ results in an edge $(u, v) \in E$ of G_{retweet}. The edges for the other three directed networks, i.e., the Twitter follow and mention networks as well as the citation networks, are created analogously.

In the co-authorship network, $G_{\text{co-author}}$, there is an edge between u and v, if there exists a publication in DBLP where u and v are co-authors with an edge weight w that equals to the number of papers u and v have co-authored together. The weight function $w : E \to \mathbb{N}$ is defined according to Table 1 which also shows the number of vertices (users) and edges in all our five networks as well as the density $(2|E|/(|V|^3 - |V|))$. In the weighted networks, a higher edge weight denotes a higher strength of the link. Consequently, the *mention* and *retweet* networks can better represent the dynamics of the underlying social network as they capture the strength of the relationships between the researchers.

[3]http://dblp.uni-trier.de/

[4]https://arnetminer.org/lab-datasets/citation/DBLP_citation_2014_May.zip

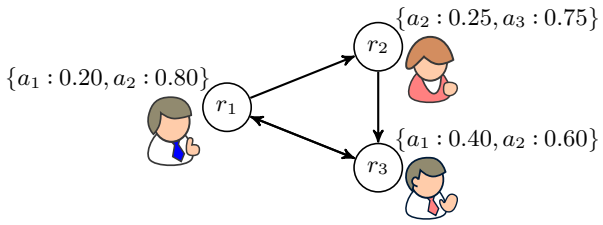


Figure 2: An exemplary network of researchers showing the soft distribution of area scores.

4. MAPPING RESEARCHERS TO AREAS

To get a more complete picture of our researchers, we mapped them to their respective research areas following a two-step procedure: First, we identified the researchers' publication venues and second, we mapped these venues to their respective research area. The final result is a soft assignment of area membership for each researcher as shown in Figure 2. In this example, researcher r_1 is working with share of 0.2 in area a_1 and with a share of 0.8 in area a_2. In Section 5.3 we use this assignment of areas to researchers to analyze the information flow between areas of Computer Science. We here present the details of our approach.

Mapping Researchers to Conferences. First, we identified at which conferences the researchers have published. For this, we queried DBLP to obtain the publication records. Note that for this study, we only considered *inproceedings* entries. In DBLP, a publication record is represented as a tuple ($title$, $year$, $conference$, ($author_1$, $author_2$, ..., $author_n$)) whereas conferences are represented by their acronym. Some of the conferences in the dataset also have a suffix attached to the acronym, e.g., *ICPP Workshop*, *KDD Workshop on Data Mining using Matrices and Tensors*, which we removed in order to get our final *author-conferences* mappings.

Mapping Conferences to Research Areas. To map the conferences of the authors to the respective areas, we again used the *List of Computer Science Conferences* from Wikipedia since it also assigns the conferences to one of the 14 areas given in Table 2. In the end, the list consisted of 268 conferences from 14 different areas after removing duplicate conference acronyms. Note that we favored this list over other potential sources such as the ACM classification to generate these mappings not only because of the mapping of conferences to areas but also because the list has been created and is curated in a consensus-driven community effort.

Finally, we joined the *author-conferences* and *conference-research area* mappings to create a final *author-research area* mapping. We also considered how often a researcher has had published in a research area and divided this number by her total number of publications to quantify the importance of a conference for a researcher. This is denoted by an *area*

Table 1: Overview on the analyzed networks and the meaning of the edge weight function $w : E \to \mathbb{N}, (u, v) \mapsto w(u, v)$.

network α	$\lvert V\rvert$	$\lvert E\rvert$	density	$w(u, v)$ means
follow	7 969	135 282	0.43%	$w(u, v) = 1$ if u follows v, else 0
mention	6 030	73 357	0.40%	u mentioned v $w(u, v)$ times in a tweet
retweet	5 050	48 592	0.38%	u retweeted $w(u, v)$ tweets of v
citation	5 163	105 004	0.79%	u cites $w(u, v)$ papers of v (authors mapped to Twitter)
co-author	1 313 098	5 124 388	0.00%	u wrote $w(u, v)$ papers together with v (all authors from DBLP)

Table 2: Areas of Computer Science and their acronyms.

acronym	area of Computer Science
AI	Artificial Intelligence
ATH	Algorithms & Theory
CA	Computer Architecture
CB	Computational Biology
CDP	Concurrent, Distributed & Parallel Computing
CG	Computer Graphics
CN	Computer Networking and Networked Systems
DM	Data Management
ED	Education
HCI	Human-Computer Interaction
OS	Operating Systems
PL	Programming Languages
SE	Software Engineering
SNP	Security & Privacy

score that can have a value between 0 and 1. This "soft" assignment of areas provides a more accurate picture of researchers' work since many researchers do in fact work in more than one area.

5. EXPERIMENTS AND RESULTS

In this section we describe the setup of our experiments and the results of our analysis. We start with a comparison of academic activities of researchers in Computer Science and their activities on Twitter. We then tackle the challenge of identifying communities of researchers and compare communities based on different networks with each other. We extend this by an analysis of the information flow between different areas of Computer Science on Twitter.

5.1 Activity of Researchers

In this section we analyze and compare the activity and interaction of researchers on Twitter with that in academia to tackle our first research question, namely *How related are activities and interactions in research to those on Twitter?*

Individual Success in Research and on Twitter. It seems obvious that Twitter users who have more followers are also more frequently mentioned in tweets or get more retweets.[5] However, it is not clear whether any of these demonstrations of interest are correlated to the academic success of the researchers. On the one hand, one could assume that successful researchers are also 'famous' on Twitter and hence have, e.g., many followers. On the other hand, one could argue that success is the result of hard work which does not leave time for much Twitter activity, resulting in fewer followers. We therefore try to answer the question, whether success on Twitter is correlated to success in academia.

[5]Indeed, the average number of followers, retweets, and mentions all exhibit a high pairwise correlation (Spearman's $\rho > 0.690, p < 0.001$).

Table 3: The Spearman correlation coefficients between average research activity and average Twitter activity per year.

research activity	followers	mentions	retweets
publications	-0.027	-0.064	-0.044
citations	0.038	-0.029	-0.020
co-authors	-0.011	-0.037	-0.028

To measure the success on Twitter, we consider the average number of followers, retweets, and mentions per year. We use these three activities instead of the number of tweets of a user, since a user cannot easily manipulate them. Similarly as peer reviewers decide upon the acceptance of publications and thereby affect the number of publications of an author, the peers on Twitter decide about the number of followers, retweets, and mentions a user has. The average number of articles a researcher has published per year can be regarded as a simple indicator for research productivity, and the average number of citations he or she received as an indicator for success in research. Additionally, we consider the average number of co-authors per year.[6] For all measures, we take the average per year to avoid a bias towards the age of the researcher or the Twitter account. For the Twitter data, we consider for each user only the time between the registration (returned by the Twitter API in the field `user.created_at`) and the date of the crawl. The users' tweets were collected using the `statuses/user_timeline` call of the Twitter API. For the average number of publications and co-authors the timespan between the first and the last publication in the DBLP dataset is considered. For the average number of citations per year only citation counts from years following (and including) the first year where a citation can be found are considered.

Table 3 shows the Spearman correlation coefficients between the average number of publications, citations, and co-authors and the average number of followers, retweets, and mentions, respectively. As we can see, the research activities are not correlated to the Twitter activities. (Incidentally, the outcome is the same, when we consider the absolute values instead of averages, though we omitted the details.) This suggests that researchers use Twitter in many diverse ways irrespective of their activities and success in academia. In particular, we could not find that a higher research productivity implies more followers on Twitter. This first analysis thus could not provide any evidence of a relationship between real-world and Twitter activities.

Collaboration in Research and Interaction on Twitter. Since science is increasingly becoming a collaborative

[6]Which is correlated to the average number of publications per year ($\rho = 0.607, p < 0.001$).

Table 4: The absolute and relative numbers of pairs of co-authors that (not) interacted with each other on Twitter.

interaction	following	mentioning	retweeting
none	3 121 (47%)	4 342 (66%)	4 835 (73%)
unilateral	1 009 (15%)	892 (14%)	1 020 (15%)
reciprocal	2 453 (37%)	1 349 (20%)	728 (11%)

endeavor, we further investigate the cooperation between researchers. Co-authorship is a typical joint activity and evidence of collaboration. We want to know whether and how researchers that co-authored a publication together also interacted with each other on Twitter. The co-author network connects users that have written at least one joint publication where the edge weight indicates how many publications they have co-authored (cf. Section 3). We consider pairs of co-authors to find out whether researchers that are strongly connected in the co-author network also closely interact on Twitter. In contrast to co-authorship, the relationships on Twitter are unilateral, i.e., a user can follow a user without the other user following back. Therefore, we distinguish between unilateral and reciprocal relationships and count for how many pairs of co-authors (i) neither of the two authors follows the other (*none*), (ii) only one author follows the other (*unilateral*), and (iii) both authors follow each other (*reciprocal*). We do the same for the mentioning and retweeting relationships.

Table 4 shows the absolute and the relative values of pairs of co-authors that interacted with each other on Twitter by following, mentioning, or retweeting. Surprisingly, many co-authors do neither follow each other (47%) nor mention or retweet each other (66% and 73%, respectively). However, when one co-author follows the other, then it is more likely a mutual than a unilateral relationship: 37% of the co-author pairs follow each other and only for 15% of them only one author follows the other. Although this is also true for mentioning, retweeting is more often unilateral than reciprocal. Interestingly, the fraction of unilateral relationships is very similar (around 15%) for all three types of interaction. Compared to the set of all researchers (not only co-authors), the interaction between co-authors is much more intense: for more than 99.5% of the pairs of researchers we cannot observe any kind of follow, mention, or retweet interaction on Twitter[7] which is far more than the highest value of 73% for the retweeting interaction among co-authors. In contrast to individual activity, these results indicate that for some aspects there exists a strong relationship between research activities and activities on Twitter.

Strength of Interaction. As we have seen, existing real-life relationships between researchers increase the likelihood of interaction on Twitter. We now want to investigate whether closer scientific collaboration induces stronger interaction on Twitter. Therefore, we analyze the fraction of co-authors that (mutually) follow, retweet, or mention each other as a dependency of the number of co-authored publications (cf. Figure 3). We can see that, as the number of co-authored publications increases, the reciprocal follow relationship increases as well. The Spearman correlation between those two values is with $\rho = 0.782$ quite high ($p < 0.001$). For mentioning and retweeting, however, the increase is less strong ($\rho = 0.345, p = 0.126$, and $\rho = 0.628, p = 0.003$, respectively). Unilateral relationships are constantly low for all three types of interaction. We can also observe a high correlation ($\rho = 0.810, p < 0.001$) between the number of co-authored articles and the fraction of those co-authors that had any kind of reciprocal interaction on Twitter (not shown as a plot). The results confirm that real-life interactions between researchers increase the likelihood for joint interactions on Twitter, where the strength of real-life activity has a positive influence on the Twitter interaction.

[7]See also Table 1 which shows the density of the individual networks.

Table 5: The number of communities in the four different networks as computed by the Louvian and CNM methods.

method	citation	follow	mention	retweet
Louvian	9	9	16	14
CNM	9	8	9	8

5.2 Communities and Networks

We now perform a detailed community and network analysis to find differences and commonalities with respect to the community forming behavior of researchers. This helps us to tackle our second research question, namely *How consistent are communities within different networks on Twitter and across traditional academic networks?*

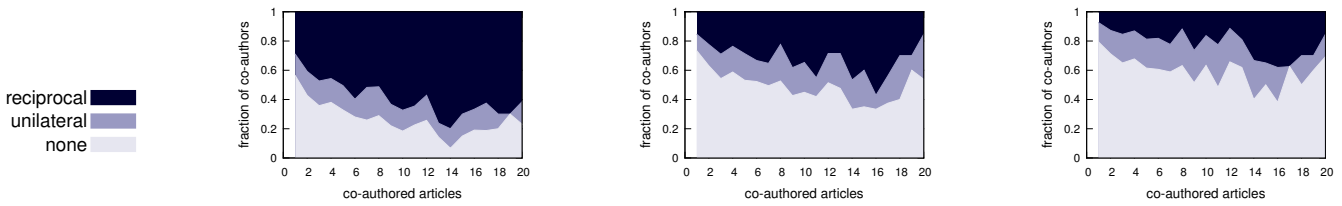
Community Detection. Community detection is a graph clustering technique to partition a graph into a modular structure such that nodes within a community have more links among each other than with the rest of the network. Such communities are a basic property of real-world networks in which some underlying rule governs the formation of such structures [16]. This governing rule can be a common relation between nodes in the network, like people studying at the same university or living in the same region. For finding communities in the four different networks, we use the Louvian [1] and Clauset-Newman-Moore (CNM) [3] algorithms as two different modularity-based methods for community detection. We used the implementations in Gephi[8] for the Louvian and in SNAP[9] for the CNM method. For the Louvian method we set the resolution parameter to 1.0 which allowed us to get communities with a larger size. The number of communities with more than 10 nodes which we could find with the two methods are shown in Table 5.

Quantitative Comparison. To analyze the consistency of the detected community over the different networks and to determine which of the networks from academia and Twitter are more similar to each other with respect to their community structure, we use five different measures for comparison. The community consistency measures we used are

- the *Rand index* (rand) [17],
- the *adjusted Rand index* (adjusted-rand) [9],
- *normalized mutual information* (nmi) [4],
- the *split-join score* (split-join) [6], and
- *variation of information* (vi) [15].

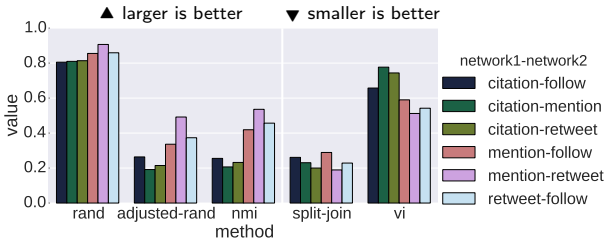
We used the *igraph*[10] network analysis library to compute these different measures. For each pair of networks

[8]https://gephi.org/
[9]http://snap.stanford.edu/
[10]http://igraph.org/

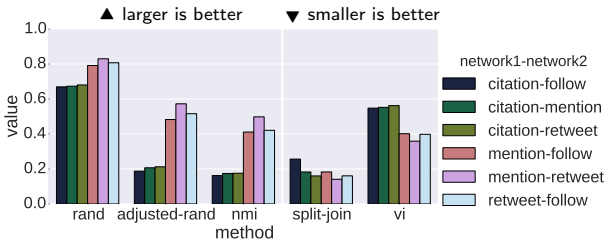


(a) following (b) mentioning (c) retweeting

Figure 3: How does closer scientific collaboration affect the interaction on Twitter? The plots show the fractions of co-authors that follow, mention, or retweet each other depending on the number of co-authored publications.



(a) Louvian



(b) Clauset-Newman-Moore

Figure 4: A quantitative comparison of communities using different similarity and distance measures.

Table 6: The overlap between the four different networks. The upper (lower) triangular matrix shows the number of overlapping nodes (edges).

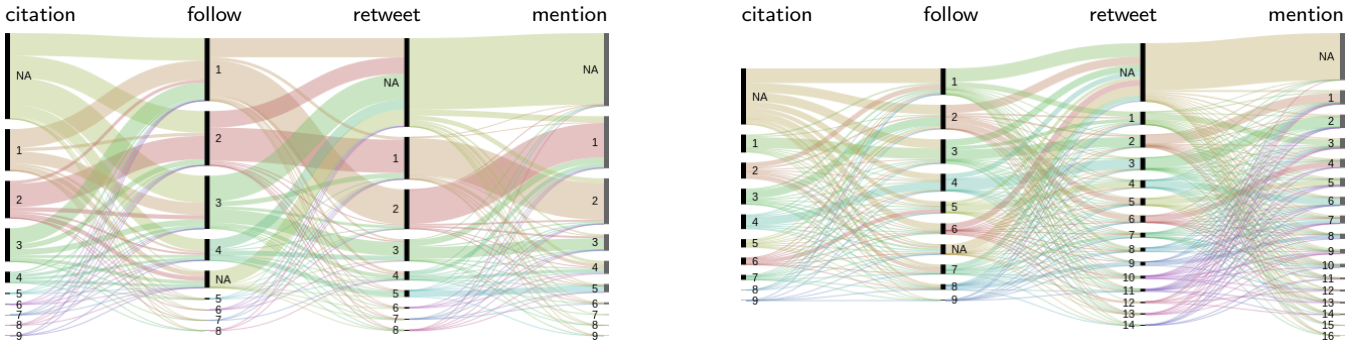
	citation	follow	mention	retweet
citation		4617	3269	3109
follow	8285		5906	5021
mention	4166	45 379		5043
retweet	2905	30 777	48 209	

we computed the common users. This is necessary, since not every user is contained in all networks due to lack of retweeting activity, for instance. Then for these users, we compute the community consistency score based on their community membership. The value of *adjusted-rand*, *rand* and *nmi* varies from 0 to 1, where 1 implies a complete match of communities. In contrast, *vi* and *split-join* are distance measures between network clusters, in which case a higher value signifies less similarity. For completely similar clusters, the *vi* and *split-join* score is 0. The maximum score in case of *split-join* is $2 \times$ (number of nodes) and in case of *vi* it is $2 \times \log$(number of clusters).

The overlap of the nodes and edges between the Twitter networks is larger as compared to the academic network (cf. Table 6). The results of the community consistency comparisons are shown in Figure 4 for the two community detection algorithms normalized by the maximum value of a score among all the networks in either of the two community detection algorithms. The general tendency is that for the first three measures, the community consistency score is higher between the Twitter networks compared to the citation network and lower for the other two measures. Comparing the results with the first three measures, the Twitter communities are more similar to each other than to the communities in the citation network. This can be attributed to the fact that the Twitter communities evolve over a common underlying network. The picture is not so clear using the split-join score and the variation of information measure, though, where the distances of the communities in the mention and retweet networks are lower than the distances between some of the Twitter networks and the citation network. As we can see community consistency score is lowest for *variation of information* and *split-join* for these retweet and mention network. The communities of the retweet and mention networks are more similar to each other than to the follow network. The similarities between the citation network and the Twitter networks are generally lower.

Consistency Over Different Networks. By analyzing the overlap between communities in the different networks we can find how similar these networks are based on how well the communities from one network can be mapped to communities in the other network. The networks with higher overlap between communities suggest that link formation or interaction between groups of individuals is more similar in them. To find the overlap and visualize it we take the union of all the nodes in all four networks in our dataset. The visualization of the overlap between communities in the Twitter networks as computed by the CNM and Louvian methods is shown in Figure 5. The communities are ordered by their number of nodes. The sets denoted by *NA* ('not available') represent the subset of researchers which are not present in the given network or cannot be mapped to any community that have more then 10 nodes.

The partition into communities for the retweet and mention network is more consistent with fewer users that could not be assigned. As can be seen in Figure 5b, for the Louvian method we have a more consistent mapping of communities, as communities 2, 3, 4, 5, 6, 8, and 9 in the retweet network do almost completely match to communities 1, 2, 6, 5, 4, 7, and 8 in the mention network. Therefore, the communities between the retweet and the mention network are more consistent as compared to those in the follow network. This can be attributed to the fact that retweet and mention are more active ways of interacting with other users, and users on Twitter usually interact with a small set of users out of



(a) Clauset-Newman-Moore

(b) Louvian

Figure 5: Community overlap in different Twitter networks.

the users they are following. Also, in Figure 5a we can observe a similar pattern as the communities 1, 2, 3 and 5 in the retweet network have a large number of common nodes to communities 2, 1, 3. 5 in mention network respectively.

Another interesting observation that shows that the communities between the retweet and mention network are more consistent is that communities from the follow network are split in the retweet network and then mapped consistently to communities in the mention network for the Louvian method. For instance, community 2 in the follow network splits into the communities 5 and 6 in the retweet network which are mapped to the communities 5 and 4 in the mention network. These results are consistent with the outcome of the community consistency score for different networks as all scores have a higher similarity between the mention and the retweet network.

5.3 Information Flow Between Areas

We define the information flow similar to how it is defined between research publications based on the citation links between them [18]. However, instead of a hard assignment based on researchers' publication venues, we use a soft assignment with fractional values as defined in Section 4. In the citation network a paper citing another paper induces an information flow from the paper that is cited to the one that is citing it. Similarly, in case of Twitter the information flows from a user that is followed to one following him or her, since the tweets of users being followed are visible to their followers. Based on the areas assigned to researchers and the direction of the follow links we can analyze the information flow between areas of Computer Science. The flow between areas is induced by the flow from users from one area that are followed to the users from another area that are following them. Since the area assignment is fractional, the flow score between the areas is computed as the product of the area scores assigned to the followed user and the follower.

Taking our example from Figure 2, when user r_1 with an area assignment of $\{a_1 : 0.20, a_2 : 0.80\}$ is following user r_2 with an area assignment of $\{a_2 : 0.25, a_3 : 0.75\}$ this will result in a flow of information from user r_2 to user r_1. In terms of the flow between areas of research we multiply the corresponding fractional values of area assignments from the followed user to the follower. E.g., for the flow from area a_2 to area a_1 we multiply the score 0.25 for area a_2 from user r_2 with the score 0.20 of area a_1 from user r_1, such that this follow relationship contributes with 0.05 to the flow of information between area a_2 and area a_1. The complete area

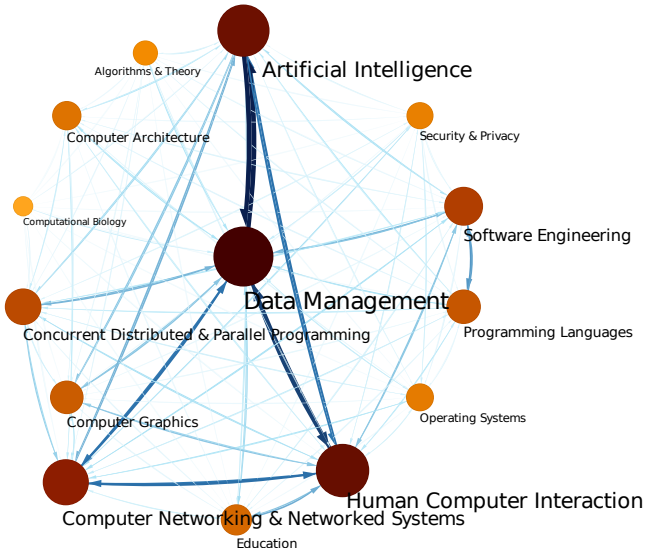


Figure 6: The information flow between different areas of Computer Science based on the follow network.

flow from user r_1 to user r_2 is then $\{(a_2, a_1) : 0.05, (a_2, a_2) : 0.20, (a_3, a_1) : 0.15, (a_3, a_2) : 0.60\}$.

The final weight of a directed edge between areas is then equal to the sum of all such area scores between pairs of researchers belonging fully or partially to these areas and having a directed link in the follow network.

Figure 6 shows the flow of information between the different areas in the follow network. The edge width corresponds to the sum of all area scores and therefore represents the flow of information between research areas. The size of the nodes denotes the amount of information flowing through the nodes which corresponds to the sum of the outflow and inflow. The main nodes are the areas *DM*, *HCI*, *AI*, and *CN* which are the main sources of information for many other nodes in the network. The areas which are related in research have a higher uni-directional or bidirectional flow between them, as we can see between *DM* and *AI*, and *DM* and *HCI*. The flow between these two pairs is higher then between any other pair of nodes. As *SE* and *PL* are related areas, they have a relatively high flow between them as compared to and from other areas. Similarly, the areas *CN*, *OS*, and *SNP* have a higher information flow between them, apart from the areas which are the main source of informa-

tion. Overall, the information flow between areas based on the following behavior of researchers on Twitter seems to be consistent to real-life interaction and collaboration among researchers working in different areas of Computer Science.

6. CONCLUSION & FUTURE WORK

In this paper, we studied social activity networks in comparison to academic activity networks of researchers from Computer Science. We found that, in general, there is no correlation between the social activity networks of researchers on Twitter, i.e. the follow, retweet and mention networks and the academic activity networks, i.e. the co-authorship and citation networks. Nevertheless, co-authors who also interact with each other on Twitter typically have a reciprocal relationship – with an increasing tendency when they have written more papers together. Comparing the communities from our three social activity networks and the two academic activity networks, we found that the social activity networks are most consistent to each other, with the highest consistency between the retweet and mention network. Also, our study showed that the follow network is most similar to the citation network. We also investigated the information flow in our networks and we found that researchers from the Computer Science areas *Data Mining (DM)*, *Human Computer Interaction (HCI)* and *Artificial Intelligence (AI)* act as a source of information for other Computer Science areas. Naturally, there is a higher chance that information flows between areas that deal with related topics since it is very likely that researchers from related but different disciplines attend the same conferences, tweet at the conference and start following each other on Twitter. For future work, we plan to repeat this study also for other disciplines such as economics as e.g. Mahrt et al. in [13] report that Twitter usage may differ between disciplines. Finally, we plan to build a Web application that features experts from and across different areas and disciplines together with methods to recommend researchers to follow.

Acknowledgements

This work was performed in the context of the Leibniz Research Alliance 'Science 2.0'. The work was supported by the Know-Center Graz and the EU-funded project 'Learning Layers' (Grant Agreement 318209). The Know-Center is funded within the Austrian COMET Program – 'Competence Centers for Excellent Technologies' – under the auspices of the Austrian Ministry of Transport, Innovation and Technology, the Austrian Ministry of Economics and Labor and by the State of Styria. COMET is managed by the Austrian Research Promotion Agency (FFG).

7. REFERENCES

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

[2] K. W. Boyack, R. Klavans, and K. Börner. Mapping the backbone of science. *Scientometrics*, 64:351–374, 2005.

[3] A. Clauset, M. E. J. Newman, and C. Moore. Finding community structure in very large networks. *Physical Review E*, 70:066111, 2004.

[4] L. Danon, A. Díaz-Guilera, J. Duch, and A. Arenas. Comparing community structure identification. *J. Stat. Mech.*, 9:8, 2005.

[5] L. De Vocht, S. Softic, A. Dimou, R. Verborgh, E. Mannens, M. Ebner, and R. Van de Walle. Visualizing collaborations and online social interactions at scientific conferences for scholarly networking. In *Proc. WWW*, pages 1053–1054, 2015.

[6] S. Dongen. Performance criteria for graph clustering and markov cluster experiments. Technical report, CWI (Centre for Mathematics and Computer Science), Amsterdam, The Netherlands, 2000.

[7] M. Ebner and W. Reinhardt. Social networking in scientific conferences – Twitter as tool for strengthen a scientific community. In *Proc. EC-TEL*, Berlin/Heidelberg, Oct. 2009. Springer.

[8] A. T. Hadgu and R. Jäschke. Identifying and analyzing researchers on Twitter. In *Proc. Web Science*, pages 23–32. ACM, 2014.

[9] L. Hubert and P. Arabie. Comparing partitions. *Journal of Classification*, 2(1):193–218, 1985.

[10] S. Jung and A. Segev. Analyzing future communities in growing citation networks. *Knowledge-Based Systems*, 69(0):34 – 44, 2014.

[11] J. Letierce, A. Passant, J. Breslin, and S. Decker. Understanding how Twitter is used to widely spread scientific messages. In *Proc. Web Science*, 2010.

[12] J. Letierce, A. Passant, J. G. Breslin, and S. Decker. Using Twitter during an academic conference: The #iswc2009 use-case. In W. W. Cohen and S. Gosling, editors, *ICWSM*. The AAAI Press, 2010.

[13] M. Mahrt, K. Weller, and I. Peters. Twitter in scholarly communication. In *Twitter and Society*, pages 399–410. Peter Lang, New York, 2014.

[14] A. Mazarakis and I. Peters. Tweets and scientific conferences: The use case of the science 2.0 conference. In *Proceedings of the 2nd European Conference on Social Media 2015 (ECSM 2015)*, 2015.

[15] M. Meilă. Comparing clusterings by the variation of information. In *Learning Theory and Kernel Machines*, pages 173–187. Springer, 2003.

[16] M. E. J. Newman. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, 103(23):8577–8582, 2006.

[17] W. Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66(336):846–850, 1971.

[18] X. Shi, B. L. Tseng, and L. A. Adamic. Information diffusion in computer science citation networks. *CoRR*, abs/0905.2, 2009.

[19] J. Tang, J. Zhang, L. Yao, J. Li, L. Z. 0007, and Z. Su. Arnetminer: extraction and mining of academic social networks. In *Proc. KDD*, pages 990–998. ACM, 2008.

[20] K. Weller, E. Dröge, and C. Puschmann. Citation analysis in twitter: Approaches for defining and measuring information flows within tweets during scientific conferences. In M. Rowe, M. Stankovic, A.-S. Dadzie, and M. Hardey, editors, *Making Sense of Microposts (#MSM2011)*, pages 1–12, May 2011.