Using Network Analysis to Characterize Participation and Interaction in a Citizen Science Online Community

Ishari Amarasinghe (✉)[1 [0000−0003−2960−480]], Sven Manske[2 [0000−0002−5098−1682]], H. Ulrich Hoppe[2 [0000−0003−3240−5785]], Patricia Santos[1 [0000−0002−7337−2388]], and Davinia Hernández-Leo[1 [0000−0003−0548−7455]]

ICT Department, Universitat Pompeu Fabra, Barcelona, Spain[1]
{ishari.amarasinghe,patricia.santos,davinia.hernandez-leo}@upf.edu
RIAS Institute, Duisburg, Germany[2]
{sm,uh}@rias-institute.eu

Abstract. Citizen Science (CS) projects provide a space for collaboration among scientists and the general public as a basis for making joint scientific discoveries. Analysis of existing datasets from CS projects can broaden our understanding of how different stakeholder groups interact and contribute to the joint achievements. To this end, we have collected publicly available forum data from the "Chimp&See" project hosted on the Zooniverse platform via crawling its *Talk* pages. The collected data were then analysed using Social Network Analysis (SNA) and Epistemic Network Analysis (ENA) techniques. The results obtained shed light on the participation and collaboration patterns of different stakeholder groups within discussion forums of the "Chimp&See" project.

Keywords: Citizen Science · Discussion forums · Social Network Analysis · Epistemic Network Analysis.

1 Introduction

"Citizen Science" (CS) is a growing trend that builds on the participation of persons not considered as professional scientists in scientific activities and achievements. Although in some areas of science we may still see individual contributions, CS is essentially collaborative and in many cases also interdisciplinary. Often the collaborative activities take place in online communities and make use of different types of collaboration technologies. Accordingly, we see CS as a very interesting subject for research in the area of collaboration technologies.

As summarised and discussed by Haklay et al. [10], there is a huge variety of different and competing definitions of CS. In this spectrum, we find a specific contrast between approaches using citizen science as a means to enhance and support "official" science through crowd-sourcing activities as opposed to grassroots initiatives driven by the volunteers, i.e. the citizen scientists, themselves. We accept that all these manifestations belong to the reality of CS. Accordingly,

the quality of scientific output in terms of research results and publications as well as the personal growth, learning and enrichment on the part of the volunteers should be considered as equally valid goals.

Scientometric or bibliometric methods can be applied to CS projects to measure scientific quality based on publications. Kullenberg and Kasperowski [15] have conducted such an analysis with a focus on biology and environmental studies as prominent areas of CS activities. Their analysis also shows the overlaps between the concepts of CS, crowdsourcing and public engagement.

Our focus is on the participation and collaboration patterns that are characteristic for the contributions of volunteers in online CS projects. Emerging technologies are influencing the scientific research process by streamlining data collection in CS, improving data management, expediting communication, etc. [18] In particularly, social networking is enhancing the dialog between scientists and citizen scientists via virtual forums and communities, increasing the collective capital [7]. As indicated by Newman et al. [18] the success of such approaches is dependent on diverse stakeholders' contributions. In order to understand better the phenomenology of CS forums, important questions to be explored are: Which kinds of roles and interactions can we observe and ascertain? Are the volunteers in a way instrumentalised as a kind of "mechanical turks" or do they engage in "legitimate peripheral participation" [16] as a source of learning and personal growth? Aristeidou et al. [1] have studied profiles and levels of engagement on the part of volunteers in CS projects relying on behavioural data from the underlying collaboration platforms. The indicators and metrics used in that study are based on duration measures and counts of certain activity types. In contrast, Huang at el. [14] have used discourse-analytic techniques to study the dynamics of interactions in two CS projects dealing with local environmental problems.

The work reported here has been conducted in the context of the European research project CS Track [1]. One of the cornerstones of CS Track's approach to monitoring and analysing CS projects and activities is the computational analysis of web-based sources. In the case study described here, we use publicly available forum data from the Zooniverse platform [2]. Zooniverse is one of the world's largest CS platforms that invites the public to participate in scientific data analysis and engage in discussions with professional scientists [23]. In our analysis of these data we go beyond a counting/aggregating approach by applying network analysis techniques of two different types: Social interactions and communication in the forum are converted into actor-actor networks using a social network analysis approach [25]. Social network analysis reveals relational structures and possibly role patterns in these communities [13]. In our analysis, we also consider the dynamics of network measures (namely eigenvector centrality) over time, which can indicate if an actor moves from a peripheral to a more central position in the network. In addition to these social relations, we use the approach of "Epistemic Network Analysis" [20] to characterise forum interac-

[1] CS Track project: https://cstrack.eu. Retrieved: 2021-04-26.
[2] Zooniverse: https://www.zooniverse.org. Retrieved: 2021-04-26.

tions in terms of certain types of knowledge building and exchange. Overall, this study is of explorative nature. The guiding aim is to identify (or not) forms of participation that would indicate the taking over of initiative and responsibility by the volunteers in the example project.

2 Analysis Methods

2.1 Social Network Analysis

Originating from the idea of using network or graph models to model social relationships, Social Network Analysis (SNA) provides a rich set of mathematically based analysis techniques, which have been applied in social science and economy studies [5]. SNA methods are frequently used for determining the importance or influence of actors in networked communities based on centrality measures or to determine subnetworks of particularly high connectivity to identify specific subcommunities. The most basic centrality measure is the number of links (and thus neighbours) associated to a given node (degree) whereas closeness or betweenness centralities require the analysis of paths in the overall network. Eigenvector centrality [3] is a recursive version of degree that does not only consider the number of neighbours but again their "weight", which corresponds to the Page rank measure for web search [6]. According to Hollenbeck and Jamieson, such operationalisations using centrality measures can support recruiting and teamwork for the purpose of developing of human capital [12].

Computer-mediated communication activities can be interpreted by SNA techniques in many different ways, always starting from taking certain communication actions as indicators (such as sending a message from A to B) as basic links in a network structure. In discussion forums, "replying" and "mentioning" can be conceived as two basic mechanisms that establish relations between actors. The example to be elaborated on originates from a discussion forum in the context of crowd-sourced citizen science activities.

Networks with directed edges or links allow for considering the directivity of interactions, which is often important in discourse analyses. Here, we would distinguish indegree and outdegree are measures to represent the quantitative involvement of an actor in the discourse. While the outdegree states how many times an actor actively contributed to a discussion in relation to a receiving actor (reply or mention), the in-degree quantifies the receptive characteristics of an actor, namely how many times the actor has been addressed or referred to.

Data In this study, we extracted data from a discussion forum which is known as "Talk" page of the Chimp&See project hosted on the Zooniverse platform of CS projects. The talk pages are categorised into 3 main categories namely: 1) Help Boards; 2) Chat Boards; and 3) Science Boards. Each board constitutes a number of sub boards that serve as a space for specific discussions between scientists and non-scientists. For instance, there are 4 sub boards under the Help Board namely: 1) Announcements; 2) Frequently Asked Questions (FAQs); 3)

Technical Support; and 4) The Objects. Likewise, the Science Board consists of 22 sub boards and the Chat Board consists of 6 sub boards [3]. Each sub board consisted of a number of conversations (e.g., around 490 conversations in Help Board, around 500 conversations in Chat Board, more than 2000 conversations in Science Board) and each conversation consisted of several posts.

To evaluate and analyse the discourse, all the conversations and posts have been retrieved using a crawling approach. This section describes the data format and metadata that has been stored by the crawler. In addition, the communication in the whole Chimp&See talk pages is modelled as a social network. Following section describes the rules of the network extraction from the forum data, particularly regarding the construction of edges and the assignment of edge weights. The whole process of the data collection consists of the (a) crawling of the Zooniverse talk pages, (b) the extraction of the discourse dataset, (c) and the construction of the social networks. The whole dataset consists of 3218 forum conversations with 24531 individual posts. The forum involved a total of 575 unique user accounts, which represents 10.1% of all the active volunteers of the Chimp& See project. The number of accounts splits up in the following (system) roles: 8 moderators, 25 scientists, 542 volunteers. The time span of the data is from 2015/04/07 to 2019/05/12. Table 1 describes the collected discourse dataset.

Table 1. Description of the dataset

Board Category	The specific forum and sub-forum (i.e., the board), e.g. “Chat Board/The Objects/”, or “Help Board/Announcements/”
Post Number	Sequential post id
Title	Title of the conversation
User Type	Moderator — Scientist — User (‘Volunteer’) — Team (‘Zooniverse admin’) [0..n]
User	Name of the user who contributed, mostly pseudonyms
Response To	(OPT) Name of the user replying to in the post. A reply is determined by the use of the “reply” button on a specific post.
References to Users	(OPT) Mentioning a user within a post text using the ‘@’ sign. [0..n]
References to Objects	(OPT) Mentioning an object within a post text using the ‘@’ sign. [0..n]
Timestamp	Time of the post in the format “dd.mm.yyyy hh:mm:ss”, e.g. “31.10.2018 20:35:00”
Post Content	Textual representation of the post content. Special markup has been removed.
URL	URL of the conversation.
Full HTML	Raw HTML of the post.

[3] Chimp& See Talk Pages: https://talk.chimpandsee.org. Retrieved: 2021-04-10.

Network Extraction The network for the discourse in the Chimp& See talk pages is modelled as an weighted, directed graph. The set of nodes contains a unique node for each user who contributed to the discussion by creating a post in a conversation. The prescribed role of moderators, scientists and volunteers is modeled as a node attribute. An edge (u, v) is created if one of two conditions is satisfied:

- Reply: if u responds to v explicitly by using the reply functionality of the forum, an edge (u, v) is created
- Reference: if u references v by using the @ notation ('@v') within the message body of the post, (u, v) is created.

Each edge (u, v) is assigned an edge weight w_{uv} which represents the number of references and replies from u to v. Each node a is assigned a node weight v_a, which quantifies the number of contributions (i.e. posts/replies) from user a. As a side note, a post does not necessarily lead to establishing an edge. This is specifically the case if there is no interaction with a user (no replies), thus the degree centrality of such a node can still be zero. For the purpose of visualizing the data with Gephi [2] the particular node weight is multiplied with 100.

The dataset has been partitioned by board category (help, science and chat) to better distinguish the different kinds of communication. In addition, the communication has been sliced into four time slices, whereas each time slice has the span of one year. Due to this partitioning of the data into time slices, a "drop out" occurs in this context when a user does not post anything in subsequent time windows.

2.2 Epistemic Network Analysis

ENA is a novel co-temporal technique that takes into account the temporality in discourse data which avoids limitations of classical coding-and-counting approaches in modeling social interactions over time [20, 8]. Different elements present in discourse, e.g., knowledge, skills, communication, that can be labelled following a pre-specified coding scheme is used in ENA to generate weighted dynamic network models that visualise the structure of connections among codes in discourse [20]. ENA has been used to model discourse in different domains such as education and health care [21, 8]. However, we are unaware of studies that use ENA to analyse discussions in CS project forums although it has a great potential to model how different user roles, i.e., volunteers, moderators, and scientists interact within such discussion spaces. Hence, in this study, we applied ENA to model, visualise, and quantitatively compare the potential differences in the discussion participation across user roles to broaden our knowledge regarding their collective knowledge building processes.

In ENA network models, the nodes represent the codes and the edges reflect the relative frequency of co-occurrences between codes. In these weighted network models, thicker edges indicate that connections occur more often and thinner edges represent less frequent connections [20]. As ENA positions nodes

in a fixed location in the projection space, it enables a visual comparison of different networks. Further, this technique enables to generate difference networks that could highlight salient differences between two networks [20].

Data As ENA requires the coding of the datasets we had to limit the number of posts considered in this study to an amount that is manageable for manual coding. At the same time, we wanted to select a subset that represents all three types of discussion boards. To this end, first, we randomly selected three sub boards for a given discussion board. Then for a given sub board we again randomly selected three conversations. Then an analysis of the distribution of the posts under each conversation was performed to identify which conversations will be selected for ENA. It was seen that the number of posts under conversations ranged from 2 to 40. To acquire a sufficient number of posts to investigate co-occurrences and to remove any bias resulting from the unequal distribution of the posts for a given conversation we selected conversations that included posts within the range of 10 - 20 resulting in a manageable, relevant dataset of 130 posts. Table 2 provides a summary of the conversations selected for ENA and the total number of posts included under a given conversation.

Codebook We followed a bottom-up approach for data coding [9]. After several iterations we came up with a coding scheme that consisted of nine different codes (see Table 3) that are also in alignment with the activities proposed in CS literature [4]. Two authors of this study coded the dataset, any disagreements were resolved by discussion.

Table 2: Conversations selected for manual coding

Board	**Sub Board**	**Conversation**	**Number of Posts**
Help	Announcements	Chimp&See's 3rd anniversary: Meet the neighbours of the "Ngogo chimps"	13
	Frequently Asked Questions (FAQs)	How to start a picture collection (not whole films) and tag them individually? How to build a "database"	15
	Technical Support	Has classification format changed?	17
Chat	Chat	Links for Identification and Further Reading	13
	Elephant Discussion Board	Elephants and insects	14
	Ask Us Anything-2nd Chimp&See Anniversary Board-April 25th, 2017	Thank You!!!	15

Science	Questions for the Science Team	A civet is not a cat??	12
	Chimp&See General Discussion	New hashtagging guidelines for number of chimps -please read carefully :-)	14
	The Objects	Hunting discussion - Man with gun carries dead Diana Monkey	17

Table 3: Codes used to label discussion posts

Code	**Description**
Data Collection	Activities considered within this code include: Volunteering to perform CS tasks, sharing availability, discussion of data collection methodologies, sharing data sources and personal/best practices to follow, sending reminders, and making announcements.
Data Analysis	Activities considered within this code include: expressing doubts and concerns related to data analysis and requesting others' opinions about the doubts expressed.
Giving Help	Activities considered within this code include: offering help in the form of additional resources, e.g., web pages, recommendations, tutorials, or as written instructions or as comments, asking further details about technical problems reported by volunteers, and proposing workarounds to solve technical problem
Request Help	Activities considered within this code include: Asking questions in the form of what?, what do you think?, do you know?, how?, why? etc., describing problems, requesting clarifications or more information, and making technical requests.
Discussion	Activities considered within this code include: encouraging discussion, providing arguments and opinions, agree/disagree on new ideas/suggestions.
Updates	Activities considered within this code include: updates related to volunteer's tasks and introduction to new members who are joining the project, e.g., moderators.
Initiating	Activities considered within this code include: Proposals of new ideas/ suggestions and requesting others opinions regarding those new ideas.
Organising Activities	Activities considered within this code include: details about specific events that are being organized within the context of the project e.g., year-end activities, gift distributions etc. or events planned for the future, organising how to distribute tasks, and invitations for collaboration.

Sharing Knowledge	Activities considered within this code include: sharing knowledge from a previous or from a current CS project, lessons learned, sharing experiences, and providing reasoning to support an argument or an action taken.

Modelling discourse using ENA ENA webtool[4] was used to model the discourse. In ENA, a network model is generated for a given unit of analysis considering the co-occurrences of the codes within a defined conversation. Conversations include lines of data from which we identify or "dig" connections for a given unit of analysis. Links can only be established within a defined conversation.

The three user roles, i.e., volunteers, moderators, and scientists were set as our units of analysis. The conversations were defined by the types of discussion boards. We chose the moving stanza window method to model discussions. Using this method, the discourse elements (codes) are segmented according to their temporal proximity. The size of the window determines the co-occurrences that are considered and thus limits the possible connections [22]. We selected a moving stanza window size of 4. A window of size 2 would lead to a Markov type model with only one-step dependencies. The size of the moving stanza window was chosen after a qualitative assessment of the posts to capture meaningful connections in discourse [22].

3 Analysis Results

3.1 Social Networks and Role Dynamics

The sociogram of the extracted network is shown in Fig. 1. To highlight a relevant portion of the network, 10-core filtering has been applied. Except a single volunteer, the densely connected component shows mainly moderators and scientists that have a high outdegree.

Following the conceptualisation of social capital mentioned before, this might indicate the feeling of importance to communicate with people of higher reputation. Therefore, we investigate the direction of communication, in particularly, who references whom in terms of affiliated forum role. The following analysis considers the whole communication structure and is not restricted to the superusers.

Fig. 2 shows the relative amount of references, normalized by the total number of references over all user roles in the specific forum. A reference is either a direct mentioning of a user (with the '@' symbol) or a post reply in the conversation structure of the discussion forum. In the help forum, most references are

[4] Epistemic Network Analysis, Wisconsin Center for Education Research: https://www.epistemicnetwork.org. Retrieved: 2021-04-26.

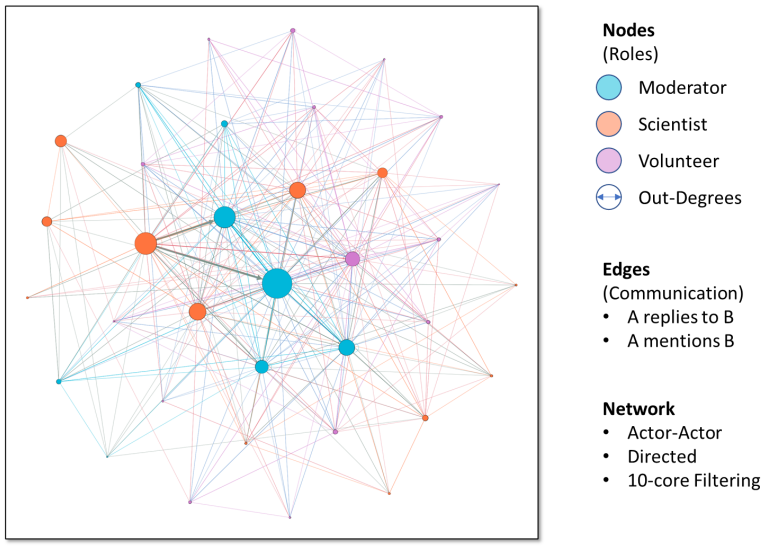

Fig. 1. The extracted sociogram of all boards over all time slices shows the relative importance of moderators in contrast to volunteers. Node labels have been hidden due to privacy reasons.

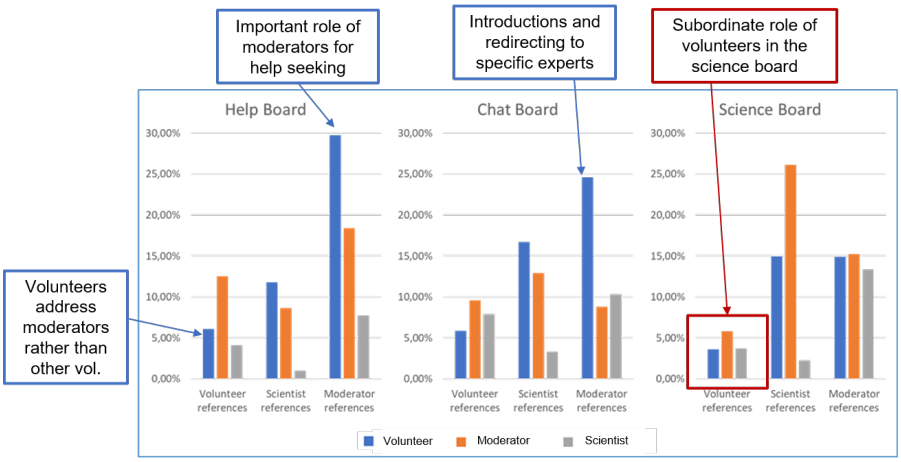

Fig. 2. Who talks to whom? Distribution of messages grouped by source role and board.

made by moderators. When volunteers seek for help, they typically do not know whom to address, whereas moderators might point to scientists and / or mention the user who asked a question. The investigation of epistemic aspects in section 2.2 is dedicated to this. Volunteers typically reference moderators to say "thanks" regarding the prior reply to the help seeking. In the chat board, the references are similar, except that there is less need for moderators to direct to scientists, which explains the lower bar for this reference. Volunteers are mentioned quite often in this forum, usually because the moderators and scientists welcome them. The chat forum is sometimes used by users to introduce themselves. This can serve for further analyses to deepen the understanding of the incentives and backgrounds of volunteers, and particularly their motivation to participate in CS activities. Interestingly, the figures indicate that scientists mostly communicate with persons from other roles, not with other scientists. Such patterns are reasonable under the premise that the communication is well-coordinated, for example, when help seeking is directed towards scientists by moderators, who know their levels of expertise. Particularly in professional contexts, such a knowledge awareness is quite important for successful teamwork. Overall, these distributions of forum communication reveal that particularly moderators play an important role in the mediation of citizen science activities. On the opposite, volunteers play a subordinate role particularly in the content-related forum, the science board.

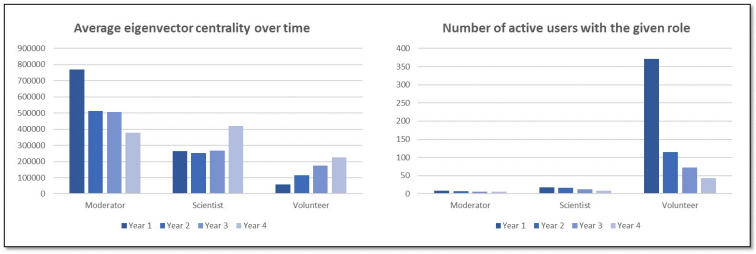

Fig. 3. Eigenvector centrality over time compared to the number of active users.

The feeling to communicate with people of higher reputation is backed by the change of eigenvector centrality over time. For all roles, we could observe a drop in active users over time, the most drastic on the part of the volunteers (cf. Fig. 3). Although, this implies that the change in eigenvector centrality is caused by the drop of users and just an anomaly of the data, the tracing of the change over time on the individual level is twofold: While the average for some volunteers is decreasing over time, on the individual level for the 10 volunteers with the highest eigenvector centrality, it needs to be differentiated. For some volunteers, the centrality is increasing, for others it is decreasing and even leading to a drop out (cf. Fig. 4). This seems to be in line with the previous assumption that volunteers are aspiring to enter a network of higher reputation and thus boost

their own influence, or when they can not enter such a network decrease and / or drop out. On the part of the moderators, a decrease in centrality could be observed on the individual level for most of the 8 moderators. In the two cases without a decrease, the moderators stopped posting and thus dropped out. For the scientists, we could not observe any comparable patterns.

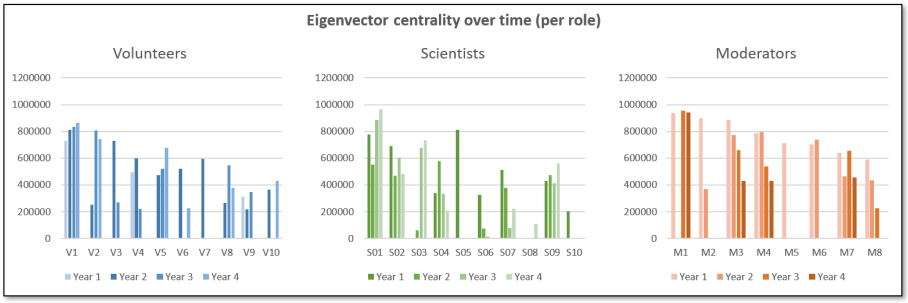

Fig. 4. Eigenvector centrality over time for the highest ranking actors.

3.2 Epistemic Networks and Discourse Structures

Epistemic Networks (EN) generated for different user roles considering different discussion boards are shown in Fig. 5. As shown in Fig. 5, in the Help Board volunteers and moderators exhibit similar behaviours as they often made connections among help-seeking and help-giving (strong connections between *Request Help* and *Giving Help* codes are visible). Notably, the EN of scientists in the Help Board does not exhibit connections to *Request Help* code rather strong connections between *Giving Help*, *Discussion*, *Sharing Knowledge*, and *Updates* are visible. This indicates that the scientists act as knowledge providers who bring knowledge from previous CS projects, sharing relevant experiences and lessons learned (see Table 3 description for *Sharing Knowledge* code). Moreover, connections to the code *Updates* indicate that in these discussion spaces scientists contribute with important updates that are necessary to carry out the intended CS task.

When considering the Chat Board (cf. Fig. 5) the strong connections observed for volunteers (e.g., between *Discussion*, *Giving Help* and *Initiating* codes) can be interpreted as discussions in terms of providing opinions and (dis)agreeing with others opinions lead to proposal of new ideas (code: *Initiating*). Similarly, giving help (e.g., sharing links to related web pages, tutorials, etc.) also leads to discussion and initiation of new ideas in the chat board. As it can be seen in Fig. 5 the networks generated for moderators and scientists for Chat and Science Boards show a similar network structure. Difference networks were generated to disentangle the differences in discourse between moderators and the scientists in Chat and Science Boards (cf. Fig. 6).

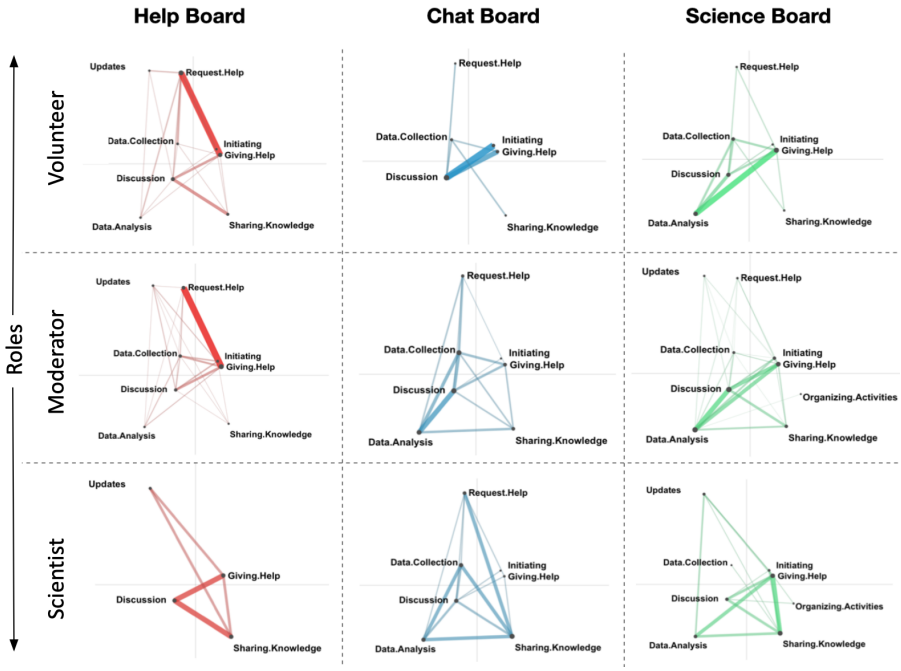

Fig. 5. Discourse patterns by user groups considering three boards.

As shown in Fig. 6(a) the strong connections between *Data Analysis* and *Discussion* for moderators in the Chat Board indicate that as they engage in data analysis they share doubts or concerns and engaged in discussion to solve those. The connection between *Data Collection*, *Giving help* and *Discussion* indicate that as they engage in data collection, e.g., sharing observations, tagging etc. they often request others opinions and sometimes offer help all the while moderating discussion. As shown in the difference network (see Fig. 6(a)) in the Chat Board a number of connections to *Sharing Knowledge* indicate that scientists engage in sharing knowledge regarding various aspects. This behaviour is similar to their behaviour observed in Help Board.

Finally, in the Science Board (cf. Fig. 5), the strong connections between *Data Collection*, *Discussion*, *Data Analysis*, *Giving help* indicate that as volunteers engage in CS activity it leads them to express doubts. Expressing doubts (*Data Analysis*) co-occur with strong help giving (e.g., links to relevant web pages, Wikipedia articles etc.). Moreover, *Data analysis*, *Data Collection*, and *Giving Help* are connected with *Discussion* showing that sharing doubts, sharing observations, and sharing additional resources lead to collective discussion.

Regarding moderators, a strong connection between *Data Analysis* and *Discussion* (cf. Fig. 6(b)) shows that in the Science Board moderators expressed doubts or concerns regarding *Data Collection* and engaged in the *Discussion* to solve those doubts. Similar connections in the EN diagrams are observed for

Fig. 6. Difference network for moderators (in purple) and scientists (in orange) considering discourse in (a) Chat Board and (b) Science Board.

moderators in the chat board as well. The scientists seem to share their own experiences, previous knowledge as well as directions to additional resources (connections between *Sharing Knowledge* and *Giving Help* codes) (cf. Fig. 6(b)). Moreover, connections between *Giving Help*, *Updates*, and *Data Analysis* in Fig. 6(b) indicate that in situations where additional knowledge or resources shared can not solve doubts and concerns related to CS task, scientists discussed those doubts with other scientists and attempted to facilitate the activity. The following excerpt indicate such connections.

Moderator: *I am still a bit unhappy about the vocal tags, because I am still missing the case: there is an animal seen (maybe even a chimp or baboon), but the vocalisation is off camera (by same species). I would like to make that clearer, as both tags - for me - do not transport that message. Especially with the baboons, but sometimes also with chimps, the individuals are reacting to that vocalization not seen and that's interesting.*

Scientist: *Hi @anonymous(name of moderator removed)- as far as I can remember, I've only used #0_chimponce. I discussed the matter with @anonymous(name of scientist removed) afterwards and subsequently removed the tag. You must have come across my post within this short time window. I am sorry for any confusion this may have caused!.*

4 Discussion

The results obtained using SNA and ENA revealed interesting findings about participation and collaboration patterns of different stakeholder groups in a CS online community that could not be obtained using either of the techniques alone. ENA results indicated that in general volunteers and moderators engaged in help-seeking and help-giving in the Help Board. However, SNA further indicated that in many instances volunteers were unaware to whom the help requests need

to be addressed. In certain situations when moderators lacked the knowledge to handle such help requests they were seen to notify and obtain help from the science team. These findings are aligned with findings from previous similar research [24, 19].

Moreover, ENA indicated that in the Chat and Science Boards volunteers and moderators not only engaged in assigned CS task, e.g., tagging, classification (coded under *Data Collection*) within the duration of the project, rather they also engage in sharing opinions, doubts and concerns which seems to lead the initiation of new ideas that often matches with their own interests. These types of initiations are referred to as the *citizen led inquiries* [24] that opens new pathways for knowledge production and eventually leads to citizen initiated discoveries [19]. Initiation was not seen to co-occur frequently in the case of scientists indicating that the knowledge they shared is in a different format as also shown in previous research [19].

The findings of the SNA also confirmed the important role moderators play in the Chat Board mediating CS activities in these discussion spaces. Findings of SNA also indicated that in Chat Board volunteers were referred more often by scientists and moderators to welcome and introduce them in the discussion forum. The analysis of the centralities of the different actors has shown that volunteers seem to aspire establishing a network of higher reputation. Being connected to prestigious people, either scientists or moderators, might be an important key to understand motivational factors for volunteers participating in citizen science activities. The nature of Zooniverse projects foresees different models to recruit moderators, particularly from the volunteers. Unfortunately, the data has been extracted ex-post and thus we could not answer the question based on evidence, whether some volunteers have been casted to moderators. Future qualitative studies might investigate those aspects further.

When considering the role of scientists ENA indicated that it is common for the scientists to intervene in talk pages in all three boards to share knowledge.

The findings of the study need to be interpreted with caution given the limited number of talk pages selected to generate EN. It is known that CS projects are unique not only based on its participants, methodology, goals, design, etc. but also in terms of generated data and knowledge production [19, 17]. Hence in the future it is important to automate the manual coding process and to extend our analysis considering more than one CS project to produce generalizable study findings.

5 Conclusions and Future Work

Understanding the dynamics in discussion spaces of CS projects is important not only to characterise and distinguish different forms of participation and expertise of different user groups, but it also opens the prospect of improving the practice of CS in a wide variety of application fields. These include formal educational contexts as well as strategic support for policy makers [19, 11].

As a caveat, we have to consider that the analysed data and obtained results are based on forum discussions. However, such crowd-sourced citizen science activities have their focus in the collection of data and classification of objects by volunteers. The active participation in forum discussions only captures a small portion of the users activities, not directly including the collection and classification of source data. Due to limited access to the Zooniverse platform, we do not have any insights in the active participation beyond the forum activity. Future studies might take this into account, particularly to assess to what extent communication and knowledge building are interdependent with the active participation in the core of the citizen science activities.

Acknowledgements This work was partially funded by the European Union in the context of the CS Track (Grant Agreement no. 872522) under the Horizon 2020 program. This document does not represent the opinion of the European Union, and the European Union is not responsible for any use that might be made of its content. We thank all CS Track team members for the fruitful interactions that facilitated this work.

References

1. Aristeidou, M., Scanlon, E., & Sharples, M.: Profiles of engagement in online communities of citizen science participation. Computers in Human Behavior, **74**, 246–256 (2017).
2. Bastian, M., Heymann, S., & Jacomy, M.: Gephi: an open source software for exploring and manipulating networks. In: Proceedings of the 3rd International AAAI Conference on Web and Social Media **3**(1) (2009).
3. Bonacich, P.: Power and centrality: A family of measures. American journal of sociology, **92**(5), 1170–1182 (1987).
4. Bonney, R., Ballard, H., Jordan, R., McCallie, E., Phillips, T., Shirk, J., & Wilderman, C. C.: Public Participation in Scientific Research: Defining the Field and Assessing Its Potential for Informal Science Education. A CAISE Inquiry Group Report. Online Submission. ERIC (2009).
5. Borgatti, S. P., Mehra, A., Brass, D. J., & Labianca, G.: Network Analysis in the Social Sciences. Science **323**(5916), 892–895 (2009).
6. Brin, S., & Page, L.: The anatomy of a large-scale hypertextual web search engine. Computer Networks and ISDN Systems **30**(1–7), 107–117 (1998).
7. Chang, H. H., & Chuang, S. S.: Social capital and individual motivations on knowledge sharing: Participant involvement as a moderator. Information & management **48**(1), 9–18 (2011).
8. Csanadi, A., Eagan, B., Kollar, I., Shaffer, D., W., & Fischer, F.: When coding-and-counting is not enough: using epistemic network analysis (ENA) to analyze verbal data in CSCL research. International Journal of Computer-Supported Collaborative Learning **13**(4), 419–438 (2018).
9. Glaser, B.G., & Strauss, A.L.: The Discovery of Grounded Theory: Strategies for Qualitative Research. Aldine Transaction, New Brunswick (1967).
10. Haklay, M. M., Dörler, D., Heigl, F., Manzoni, M., Hecker, S., & Vohland, K.: What Is Citizen Science? The Challenges of Definition. K. Vohland et al. (eds.), The Science of Citizen Science, pp. 13–33. Springer (2021).

11. Haklay, M.: Citizen science and policy: A European perspective. Washington, DC: Woodrow Wilson International Center for Scholars (2015).
12. Hollenbeck, J. R., & Jamieson, B. B.: Human capital, social capital, and social network analysis: Implications for strategic human resource management. Academy of Management Perspectives **29**(3), 370-385 (2015).
13. Hoppe, H. U., Harrer, A., Göhnert, T., & Hecking, T.: Applying network models and network analysis techniques to the study of online communities. In: Mass Collaboration and Education, pp. 347–366. Springer, Cham (2016).
14. Huang, J., Hmelo-Silver, C. E., Jordan, R., Gray, S., Frensley, T., Newman, G., & Stern, M. J.: Scientific discourse of citizen scientists: Models as a boundary object for collaborative problem solving. Computers in Human Behavior **87**, 480–492 (2018).
15. Kullenberg, C., & Kasperowski, D. (2016). What is citizen science? - A scientometric meta-analysis. PLoS ONE **11**(1) (2016).
16. Lave, J., & Wenger, E.: Legitimate peripheral participation. Learners, learning and assessment. London: The Open University (2016).
17. Lemmens, Rob, et al.: A Conceptual Model for Participants and Activities in Citizen Science Projects. The Science of Citizen Science (2021). https://doi.org/10.1007/978 − 3 − 030 − 58278 − 4_9
18. Newman, G., Wiggins, A., Crall, A., Graham, E., Newman, S., & Crowston, K.: The future of citizen science: emerging technologies and shifting paradigms. Frontiers in Ecology and the Environment **10**(6), 298–304 (2012).
19. Rohden, F., Kullenberg, C., Hagen, N., & Kasperowski, D.: Tagging, pinging and linking–User roles in virtual citizen science forums. Citizen Science: Theory and Practice **4**(1), (2019).
20. Shaffer, D. W., Collier, W., & Ruis, A., R.: A tutorial on epistemic network analysis: Analyzing the structure of connections in cognitive, social, and interaction data. Journal of Learning Analytics **3**(3), 9–45 (2016).
21. Shum, S., B., & Echeverria, V., Martinez-Maldonado, R.: The Multimodal Matrix as a quantitative ethnography methodology. In: Proceedings of the 1st International Conference on Quantitative Ethnography. pp. 26–40. Springer (2019).
22. Siebert-Evenstone, A. L., Irgens, G. A., Collier, W., Swiecki, Z., Ruis, A. R., & Shaffer, D. W.: In Search of Conversational Grain Size: Modelling Semantic Structure Using Moving Stanza Windows. Journal of Learning Analytics **4**(3), 123–139 (2017).
23. Simpson, R., Page, K. R., & De Roure, D.: Zooniverse: observing the world's largest citizen science platform. In: Proceedings of the 23rd International Conference on World Wide Web, pp. 1049–1054. (2014).
24. Tinati, R., Van Kleek, M., Simperl, E., Luczak-Rösch, M., Simpson, R., & Shadbolt, N.: Designing for citizen data analysis: A cross-sectional case study of a multidomain citizen science platform. In: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, pp. 4069–4078. ACM (2015).
25. Wasserman, S., & Faust, K.: Social network analysis: Methods and applications. Cambridge University Press (1999).