The Shape of Collaboration Networks in Citizen Science Projects

Guilherme Correia

Instituto Superior Técnico, Universidade de Lisboa, 1049-001 Lisboa, Portugal

Abstract. There has been a substantial increase in interest in citizen science, spanning a wide range of areas from biodiversity to water and air quality monitoring. These platforms are particularly efficient in monitoring tasks impossible to be handled by small teams of experts. The acquisition and validation of information emerge as a cooperative effort grounded on large self-organized networks of participants. Despite this, little is known about the structural patterns of these networks of collaboration. Here, we provide a preliminary analysis of a representative collaborative network of a major citizen science platform aiming at mapping and sharing observations of biodiversity. We show that the resulting temporal collaborative network exhibits a power-law dependence on the connectivity that outlasts the entire period investigated, despite significant differences in the number of participants throughout the years. This result suggests the existence of time and scale-invariant topological properties in citizen science platforms. We further show that these collaboration networks portray a well-defined community structure associated with users' taxon preferences. Finally, we show that each participant's role or type of participation tend to evolve in time — the longer at the network, more likely the adoption of the role of Validator of others' observations, and the higher the chances of occupying central positions. The methodology developed here demonstrates the possibility of analyzing, comparing and potentially shaping the time-evolution of social networks associated with collaborative science platforms.

Keywords: Citizen Science · Network Science · Cooperation

1 Introduction

Citizen science can be defined as the participation of citizens in scientific projects. Recently, citizen science has gained popularity and acceptance as a mainstream approach to collect and analyze information [1, 2] in a wide range of research topics [3–6]. This is mainly due to its capability to address large monitoring tasks, impossible to be handled by small teams of experts in a cost effective way [1, 7, 8].

In particular, online citizen science projects, such as iNaturalist [9], eBird [10] or Zooniverse [11] have hundreds of thousands and even millions of users cooperating to acquire and validate large amounts of information [12–14]. In

these platforms, each user is connected to others she/he cooperated with in a representative self-organized network.

However, while citizen science projects (CSPs) are evolving into massive *networks* of users, it remains illusive how volunteers interact to organize into such structures. Here, we perform *network* analysis on a CSP - BioDiversity4All [15] - to describe the structural patterns of these collaborative networks.

BioDiversity4All is a Portuguese citizen science project started in 2010. It is one of the citizen science biodiversity databases that compose the iNaturalist network [9]. In this platform, users report observations on organisms at a particular time and location [16]. These observations can be identified and commented by other users, creating a network between the users and the observations they have participated in - we named this network BipartiteCoop. From this network, it is possible to extract the collaboration network of users that have cooperated in the identification of the same observation - we named this network CoopNet.

Studies on citizen science's origins and future (e.g. [5]) as well as users motivation and social nature (e.g. [17]) is vast, however little work has been done on the analysis of the nature of interactions of citizen science projects. Here, we analyse these networks regarding users' interaction, behaviour and evolution as well as organization into communities inside the platform.

2 Results and Discussion

2.1 Network Analysis

Since we are mainly focusing on user interaction, we started by analyzing the network of user cooperation, CoopNet. First, we analyzed the users regarding the number of connections. Figure 1 (left pane), displays the cumulative degree distribution of the CoopNet network in a double logarithmic plot and a power-law fit[1], following the procedure proposed in [19], with degree exponent $\gamma = 2.7$. This plot shows that most users present a small number of connections, but a small fraction of users have an extreme amount of connections. The power-law fit obtained is particularly interesting, not only emphasizes the heterogeneity of users regarding their number of connections, but also suggests that the CoopNet is scale-free.

We also calculated the average clustering coefficient (CC) for the whole network and the average shortest path (ASP) for the giant component and obtained 0.59 and 2.70, respectively. These values infer that the users in the BioDiversity4All are highly connected and very close to each other.

As we have seen before, our network's inter-connectivity is highly uneven - most users have few connections but a small number, the hubs, have a high degree. To understand how this disparity influences the BioDiversity4All community, we evaluated the importance of the the most connected users. A way to do this is to compare the impact on the network's connectivity of removing a

[1] Fit obtained using the *powerlaw* python package [18].

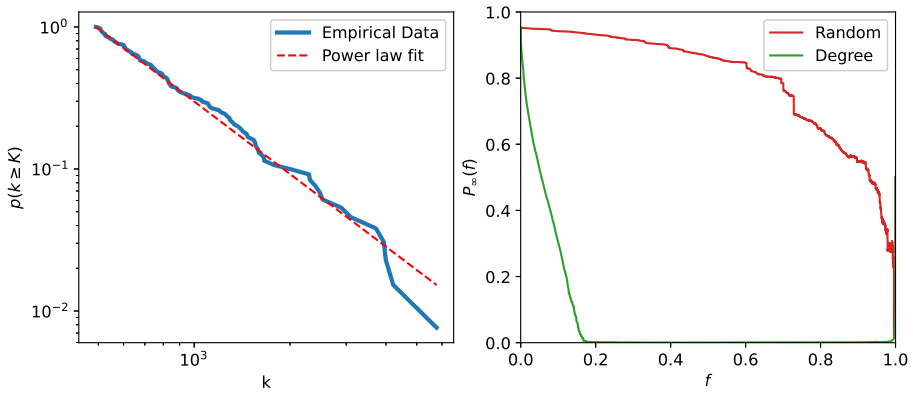

Fig. 1. Left pane: Cumulative degree distribution of the cooperation network. The straight red line provides a power-law fit following the procedure proposed in [19]. Right pane: Estimated probability that a given node belongs to the giant component after a f fraction of nodes has been removed: in random order (in red) and by degree (in green).

fraction of the hubs and a random fraction of users. From there, we may conclude how much our platform depends on the most connected users.

The scale-free property indicated in the degree distribution above represents good news in terms of resilience against the removal of users. Indeed, these platforms should be robust to users that decide (or are obliged) to cease to contribute. Figure 1 (right pane) illustrates this effect. It shows the estimated probability that a node belongs to the giant component as we remove nodes in random order (simulating random dropouts) and in order of their Degree Centrality (simulating the most influential users leaving the platform). The results show that the network is able to remain connected when most users leave the platform - we need to remove nearly all of the randomly selected nodes for the giant component's effective disappearance. However, it cannot endure the disengagement of the top 16,7% nodes with highest degree. From these point onward the network is fully disconnected. These results illustrate a classic property of scale-free networks: robustness against random removal of users come with a price. In this case, it shows that the BioDiversity4All platform greatly depends on the hubs in order to remain connected.

To study the evolution of the CoopNet's topology, we analyzed its growth, in the number of users (its nodes), and the evolution of its degree exponent (see Figure 2). What stands out in Figure 2 is that despite significant differences in the number of users throughout the years, the network exhibits the same approximate degree exponent, suggesting that our network's connectivity *is independent* of scale. Regarding other properties of the network, the average degree increases throughout the years but the cluster coefficient and the average shortest path remain invariant. These values are coherent with [19] and show

that the CoopNet keeps its scale-free and small-world properties throughout the years.

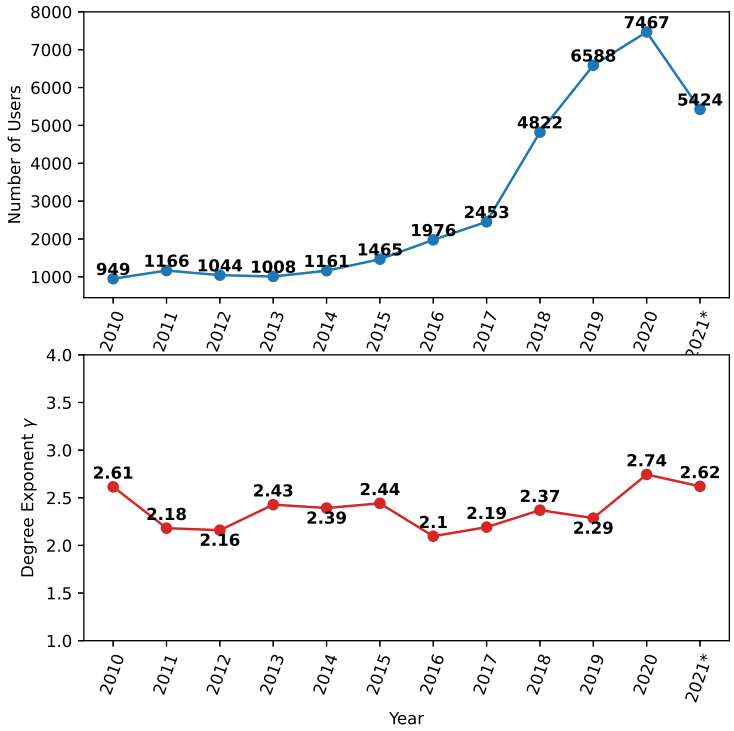

Fig. 2. Values characterizing the network's evolution: the number of nodes per year (top pane); the degree exponents of the network by year (bottom pane)

2.2 Community Analysis

In this subsection, we aim to identify existent communities in the network, to determine possible dividing factors between users in the BioDiversity4All platform. To evaluate the networks community structure, we used the Louvain method. This method consists of a modularity optimization algorithm to segregate nodes according to their connections, thus forming communities of highly interconnected nodes [20].

We obtained eight communities, of varying sizes. One hypothesis for the obtained partitioning is that users group according to their interest or knowledge on specific organisms. To evaluate this, we characterized the frequency of *taxon* interests by community (see Table 1).

Table 1. Communities by *taxon* percentage. Every community presents a distinct pattern concerning its users' *taxon* preferences.

Comm.	Size	Insect	Plant	Rept	Aves	Mam	Fungi	Amphi	Mollu	Actin	Arach	Proto	Chrom	Other	Other-Aqua	Total
1	6599	10.4%	76.2%	0.8%	4.1%	1.7%	3.2%	0.3%	0.7%	0.3%	1.1%	0.1%	0.0%	0.6%	0.6%	100.0%
2	3545	60.1%	16.6%	1.0%	3.2%	1.3%	2.5%	0.7%	1.2%	0.2%	9.6%	0.1%	0.1%	2.0%	1.6%	100.0%
3	2918	10.4%	10.9%	1.6%	64.5%	6.6%	1.4%	0.9%	0.5%	0.6%	1.1%	0.0%	0.0%	0.7%	0.7%	100.0%
4	1258	8.0%	23.7%	1.1%	4.2%	6.0%	2.6%	0.7%	17.9%	13.0%	1.9%	0.0%	1.0%	4.4%	15.4%	100.0%
5	1144	14.6%	14.7%	33.7%	7.8%	4.3%	2.4%	16.0%	0.9%	0.5%	3.2%	0.1%	0.0%	0.9%	0.8%	100.0%
6	536	7.8%	24.5%	0.9%	2.1%	9.8%	47.6%	0.7%	1.5%	0.1%	2.6%	0.8%	0.0%	0.9%	0.7%	100.0%
7	3	0.0%	100.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
8	2	0.0%	100.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
Total	16005	3427	6537	536	2417	527	653	269	346	212	522	15	19	197	327	16005

Table 1 is quite revealing - it shows that communities tend to have a predominant *taxon* or multiple prevailing *taxa*. Overall, every community presented a very distinct pattern concerning its users' taxon preferences. The three largest communities —— communities 1, 2 and 3 in Table 1 as 1, 2 and 3 —— represent about 80% of the entire network: Community 1 is characterized by Plant enthusiasts (76.2% *Plantae*); in Community 2 most users are Insect enthusiasts (60.1% *Insecta*); Community 3 is made mostly by bird watchers, with 64.5% *Aves*. Communities 4, 5 and 6 follow similar patterns, each mostly associated with a particular taxon or taxa. Community 4 is interested in aquatic organisms — 46.3% (17.9% *Mollusca*, 13.0% *Actinopterygii* and 15.4% other aquatic animals), and 23.7% *Plantae*; Community 5 is mainly Reptiles and Amphibians - 49.7% *Reptilia* and *Amphibia*, 33.7% and 16,0% respectively — but also 14.7% *Plantae*. Finally, Community 6 is 47.6% *Fungi* and 24.5% *Plantae*.

These results suggest that users are more likely to connect to those presenting the same taxon interests and knowledge.

2.3 Distribution and evolution of roles

Observations are composed of different users, each adopting a particular role: those that create observations, the Triggers, and those that participate in other users' observations, the Validators. Here we analyze how these preferences are distributed through the population of users and how they may evolve in time. To classify users we resorted to the normalized relative difference between the number of times a user has participated on others observations (validations, v_i) and created an observation (triggers, t_i), to create a Behavioral Value, $B(i)$ associated to each individual i, given by

$$B(i) = \frac{v_i - t_i}{\max(t_i, v_i)} \tag{1}$$

Intuitively, we may classify *Triggers* as users with a value of $-1.0 \leq B < 0.1$, creating more observations than validations. Similarly, *Validators* have $0.1 < B \leq 1.0$, mostly contributing with expert validation of others' observations.

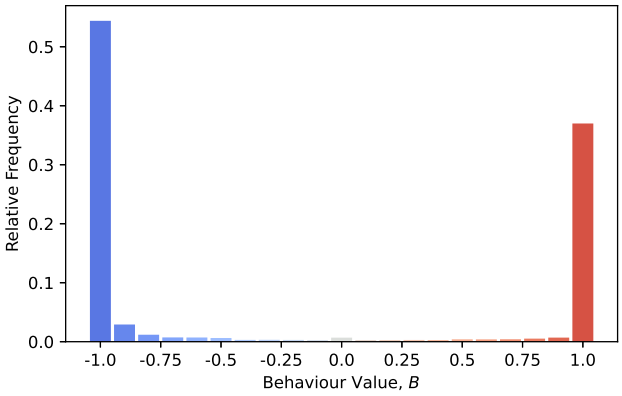

Fig. 3. The fraction of users with a given Behaviour Value.

Finally, *Hybrids* are equally likely propose and validate observations — for simplicity, we adopt $-0.1 < B < 0.1$, in this case. In Figure, 3 we show the Relative Frequency of the Behaviour Value, B, for all years, showing a polarized trend: Users tend to be attracted to one of the extremities, with a significant number of users choosing to only trigger or validate. As shown in Fig. 3 this conclusion is independent of the threshold (in our case 0.1.) we use to classify each typology of user.

Since the tendency to validate or trigger new observations may depend on the expertise of an user, we also analyzed how users behaviour changes throughout the years in the platform. To this end, we plotted the fraction of users of each type by the number of years they have been participating in the network. The results are shown in Figure 4. It shows that the percentage of triggers decreases and stabilizes below 25%, with its highest value being the users' first year (59%). On the other hand, the percentage of validators and hybrids tends to increase.

A possible interpretation for these results is that users that have been on the platform longer tend to specialize in the validation of other people's observations. But, they also contribute by creating their own observations.

The number of users that last also changes. From the first year to the second there is a drastic change - from 16795 to 2872 active users. These users that dropped out seem to be mostly Triggers, explaining the rather high Trigger percentage in the first year, and the sudden rise in Validators in the second year.

Overall, users tend to choose between adopting the role of Trigger and Validator. Although a few users have adopted a hybrid role, in which they create as well as validate observations. In terms of behaviour evolution, it seems people tend to try the network by creating observations (adopting the role of Trigger in their first year). But, as time passes, they become Validators that may alternate by creating observations.

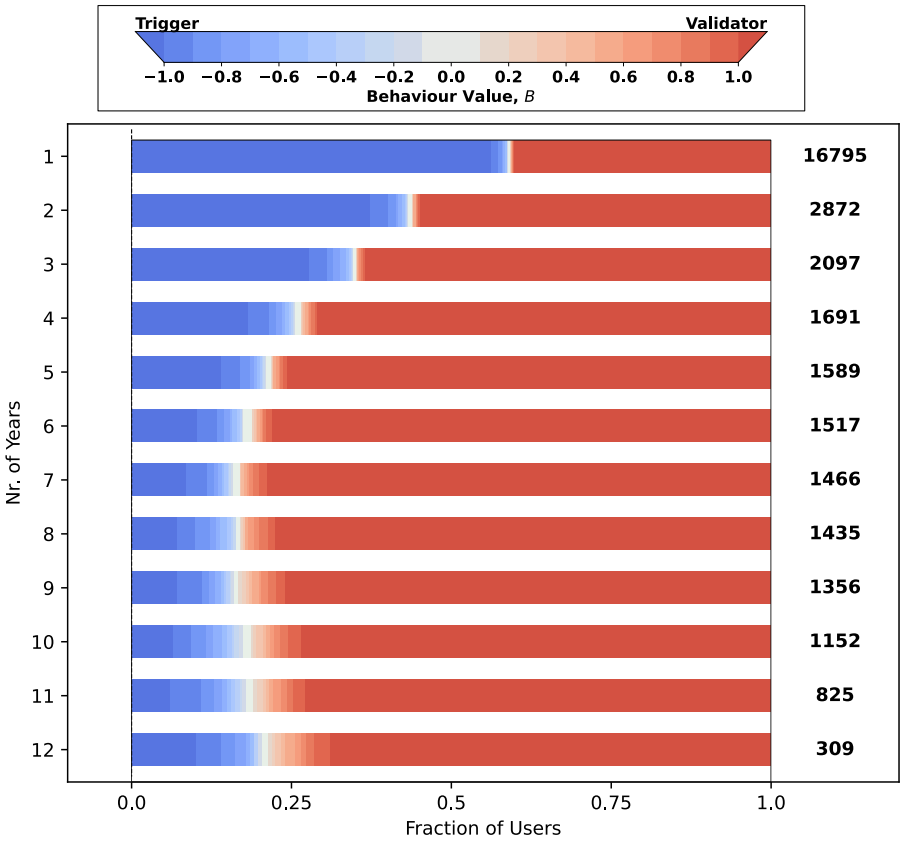

Fig. 4. Fraction of Triggers, Validators and Hybrids, in order of the number of years since their first participation.

Conclusions

The main goal of this study was to describe how citizen science projects' volunteers interact to organize into networks of cooperation. To address this question, we analysed users regarding their connectivity, community organization and interaction with the platform.

First, we found that users organize into a scale and time invariant network, suggesting that despite future variations in the number of users, it will always present a power law dependence on the connectivity. Next, we showed that some users, the hubs, bind the structure together, suggesting that the BioDiversity4All highly depends on these to keep users connected - should the platform take measures to keep the hubs engaged?

Regarding community partitioning, the users tend to group according to taxon interests and knowledge. This result may be explained by the fact that users with the same interests are more connected, since tend to participate in the

same observations - consequently, the partitioning method groups these users in the same community

Another interesting result relates to user behaviour, most users tend to either only create or only validate observations, suggesting that users have a strong preference towards the way that they enjoy the BioDiversity4All platform. A possible interference that cannot be ruled out is the fact that users from anywhere in the world can perform validations in Portugal - if the same users have only created observations outside this country, they are considered to have only validated observations.

Regarding behaviour evolution in the BioDiversity4All , the results show that new users have a very high tendency to create observations rather than validate. In contrast, users that have been in the network longer tend to validate more, while still creating observations. Two main hypothesis to explain this pattern emerge: users validate more, the longer they spend in the network; users that quit the most are Triggers;

Overall, we hope that this work will have relevant implications for the study of citizen science. While formulating a methodology for studying volunteers' interaction and behaviour, we have shown numerous results using network analysis. Mainly, these results have demonstrated that the BioDiversity4All is a highly connected platform presenting time and scale-invariant topological properties. Also, it exhibits a community structure well defined by its users' interests or knowledge. And, (most) users evolve, starting by creating observations but eventually assuming the role of validators. Furthermore, these results prefigure future activity in the BioDiversity4All platform.

References

1. Aceves-Bueno, E., Adeleye, A.S., Feraud, M., Huang, Y., Tao, M., Yang, Y., Anderson, S.E.: The accuracy of citizen science data: a quantitative review. Bulletin of the Ecological Society of America 98(4), 278–290 (2017)
2. 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)
3. Streito, J.C., Chartois, M., Pierre, É., Rossi, J.P.: Beware the brown marmorated stink bug! IVES Technical Reviews, vine and wine (2020)
4. Justine, J.L., Winsor, L.: First record of presence of the invasive land flatworm platydemus manokwari (platyhelminthes, geoplanidae) in guadeloupe. Preprints (2020)
5. Silvertown, J.: A new dawn for citizen science. Trends in ecology & evolution 24(9), 467–471 (2009)
6. Eveleigh, A., Jennett, C., Lynn, S., Cox, A.L.: "i want to be a captain! i want to be a captain!" gamification in the old weather citizen science project. In: Proceedings of the first international conference on gameful design, research, and applications. pp. 79–82 (2013)
7. Theobald, E.J., Ettinger, A.K., Burgess, H.K., DeBey, L.B., Schmidt, N.R., Froehlich, H.E., Wagner, C., HilleRisLambers, J., Tewksbury, J., Harsch, M.A., et al.: Global change and local solutions: Tapping the unrealized potential of citizen science for biodiversity research. Biological Conservation 181, 236–244 (2015)

8. Devictor, V., Whittaker, R.J., Beltrame, C.: Beyond scarcity: citizen science programmes as useful tools for conservation biogeography. Diversity and distributions 16(3), 354–362 (2010)
9. The iNaturalist Network page (Last accessed 12 May 2020), https://www.inaturalist.org/pages/network
10. Sullivan, B.L., Wood, C.L., Iliff, M.J., Bonney, R.E., Fink, D., Kelling, S.: ebird: A citizen-based bird observation network in the biological sciences. Biological conservation 142(10), 2282–2292 (2009)
11. The Zooniverse Home page (Last accessed 25 January 2021), https://www.zooniverse.org
12. The iNaturalist Stats page (Last accessed 25 January 2021), https://www.inaturalist.org/stats
13. The eBird About page (Last accessed 25 January 2021), https://ebird.org/about
14. Torney, C.J., Lloyd-Jones, D.J., Chevallier, M., Moyer, D.C., Maliti, H.T., Mwita, M., Kohi, E.M., Hopcraft, G.C.: A comparison of deep learning and citizen science techniques for counting wildlife in aerial survey images. Methods in Ecology and Evolution 10(6), 779–787 (2019), https://besjournals.onlinelibrary.wiley.com/doi/abs/10.1111/2041-210X.13165
15. BioDiversity4All (Last accessed 25 January 2021), https://www.biodiversity4all.org
16. BioDiversity4All Help page (Last accessed 12 May 2020), https://www.biodiversity4all.org/pages/help
17. Tiago, P.: Social context of citizen science projects. In: Analyzing the Role of Citizen Science in Modern Research, pp. 168–191. IGI Global (2017)
18. Alstott, J., Bullmore, E., Plenz, D.: powerlaw: a python package for analysis of heavy-tailed distributions. PloS one 9(1), e85777 (2014)
19. Barabási, A.L., et al.: Network science. Cambridge university press (2016)
20. Blondel, V.D., Guillaume, J.L., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. Journal of statistical mechanics: theory and experiment 2008(10), P10008 (2008)