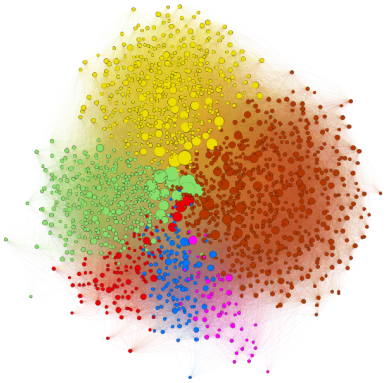
TÉCNICO LISBOA



The Shape of Collaboration Networks in Citizen Science Projects

Guilherme Crespo Rodrigues Antunes Correia

Thesis to obtain the Master of Science Degree in

Information Systems and Computer Engineering

Supervisor(s): Prof. Francisco João Duarte Cordeiro Correia dos Santos
Prof. Ana Gualdina Almeida Matos

Examination Committee

Chairperson: Prof. António Manuel Ferreira Rito da Silva
Supervisor: Prof. Ana Gualdina Almeida Matos
Member of the Committee: Dr. Patrícia Maria Nunes Tiago

September 2021

Dedicated to my family for always supporting me.

Acknowledgments

Throughout the writing of this dissertation I have received a great deal of support and assistance.

I would first like to thank my supervisors for their encouraging guidance: Prof. Francisco Santos, for his ingenious suggestions and invaluable insight into network science; Prof. Ana Matos, for her guidance and useful contributions; Prof. M. Rosário Oliveira and Prof. Ana Subtil Garcia for their constructive criticism and advice. I gratefully acknowledge the assistance of Patrícia Tiago from BioDiversity4All for her availability and helpful contribution.

I would also like to extend my deepest gratitude to my family for their love, patience and profound belief in my abilities. Special thanks to my girlfriend for her patience, unwavering support and encouragement. Thanks also to my friends, for the quality moments of comfort and joy.

Resumo

Tem existido um aumento substancial no interesse pela ciência cidadã, abrangindo várias áreas, desde o estudo da biodiversidade à monitorização da qualidade do ar e da água. Estas plataformas são particularmente eficientes em tarefas de monitorização impossiveis de serem tratadas por pequenas equipas de especialistas. A aquisição e validação da informação resulta de um esforco cooperativo baseado em redes de participantes organizadas expontaneamente. Apesar disso, pouco é conhecido acerca dos padrões estruturais destas redes de colaboração. Aqui, analizamos uma rede de colaboração de uma conhecida plataforma de ciência cidadã, centrada em mapear e partilhar observações de biodiversidade. Nós mostramos que a rede de colaboracao temporal obtida mostra uma dependencia em lei de potência ao nível da connectividade, que persiste o tempo inteiro analisado, apesar das significativas diferenças em numero de participantes ao longo dos anos. Este resultado sugere a existencia de propriedades topologicas independentes de tempo ou escala. Além disto, ainda mostramos que estas redes de colaboração retratam comunidades bem definidas associadas às preferencias de *taxon* dos utilizadores. Finalmente, mostramos que o cargo ou tipo de participaçã de cada utilizador tende a evoluir com o tempo - quanto mais tempo na rede, mais esperado é a adopção do papel de validador das observações de outros utilizadores, e mais provável a ocupação de posições centrais. A metodologia aqui desenvolvida demonstra a possibilidade de analisar, comparar e potencialmente modelar a evoluçao das redes sociais associadas a plataformas de ciência cidadã.

Palavras-chave: Ciência Cidadã, Ciência de Redes, Cooperação, Biodiversidade, Redes Sociais, Redes de Cooperação

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 analyze 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 in the network. The methodology developed here demonstrates the potential of network since in analyzing, comparing and potentially shaping the time-evolution of social networks associated with collaborative science platforms.

Keywords: Citizen Science, Network Science, Cooperation, Biodiversity, Social Networks

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Citizen science is not recent. Before science was a paid profession, dating from the later part of the nineteenth century, there were already many citizen scientists — individuals who were passionate about science and dedicated their time on scientific research but made their living in some other profession [1]. Today, citizen science has become more formalized and gained popularity and acceptance as a mainstream approach to collect information and data [2, 3] in a wide range of research topics - from invasive species monitoring (e.g. [4, 5]), to water quality monitoring, as well as projects on climate change [1] and weather logs transcription [6]. This is mainly due to its capability to address large monitoring/cataloging tasks, impossible to be handled by small teams of experts in a cost effective way [2, 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]. Each user in these platforms is connected to every other he cooperated with in a representative self-organized network. However, while citizen science projects (CSPs) are evolving into massive networks of users, little is known about the structural patterns of these collaborative networks. In this thesis, we perform a *network* analysis on a online CSP - BioDiversity4All [15], one of the citizen science biodiversity databases composing the iNaturalist network [16]

In the BioDiversity4All platform, users report observations on organisms at a particular time and location [17]. These observations can be identified and commented by other users, creating a network between the users and the observations they have participated in (BipartiteCoop). From this network, it's possible to extract the intrinsic collaboration network of users that have cooperated in the identification of the same observation (CoopNet). By analysing these networks, its possible to evaluate their properties and unveil information on the users behaviour, interaction and organization into such structures.

Network science studies complex networks created from empirical data. It is a multidisciplinary area borrowing knowledge from subjects such as graph theory, statistical physics, control and information theory, statistics, but also computer science, including algorithms, database management, data mining and data science [18]. It should not be mistaken for graph theory, however. Graph theory is more abstract, its purpose is to develop mathematical tools to describe a graph's property. Network science is

more data-driven, as Barabási stated [18]: "*Each tool [...] is tested on real data and its value is judged by the insights it offers about a system's properties and behavior*".

1.1 Objectives and Deliverables

In this thesis, we aim to understand interaction, behaviour and evolution of users in the BioDiversity4All platform. To this end, we propose to create and analyse the respective network of users and observations (BipartiteCoop), the social network composed of cooperating users (CoopNet), and the weighted directed versions of these networks. Concerning analysis, we wish to study:

1. both networks' basic properties to understand their size and node's distance;
2. the BioDiversity4All 's observations regarding the number of cooperating users to understand how observations differ;
3. users in terms of number of connections to identify the most social users as well as understand how users differ;
4. how sensible the network is on the withdrawal of the most connected users;
5. the evolution of the network throughout the years.
6. spontaneously occurring communities obtainable from user interaction;
7. users regarding their behaviour and evolution;

We believe that by evaluating these characteristics it is possible to retrieve valuable information regarding: the users, their interactions and interests as well as their evolution; the BioDiversity4All 's underlying communities; the platforms evolution; as well as potentially model the future of the platform. Furthermore, we will compare the obtained values with other known networks, allowing for a deeper insight on the networks nature and resemblances.

1.2 Thesis Outline

This thesis is organized as follows: Chapter 2 gives insight into Graph Theory and Network Science and presents some related work on Citizen Science and Collaboration Networks. Chapter 3 reveals the data and its intricacies, as well as the networks' creation methods. Next, in Chapter 4, the networks' analysis is performed, evaluating their properties, evolution and resilience. Chapter 5 shows an analysis of the networks community structure, and a possible dividing factor between communities is studied. Chapter 6 pictures an analysis of the users' form of interaction, namely, the users most common behaviour and evolution. Finally, Chapter 7 summarizes the achievements of this thesis and illustrates future work.

Chapter 2

Background

In this chapter, we present some theoretical background from Graph Theory and Network Science necessary to understand the methods used and the results obtained (Sections 2.1 and 2.2), as well as some of the previous work on Citizen Science and Collaboration Networks (Section 2.3).

2.1 Graph Theory

In order to better understand how networks are represented, it is necessary to acknowledge some basic notions of graph theory such as the structure and characterization of graphs.

A graph is constituted by nodes connected between each other by links. Nodes are objects with multiple characteristics, like their name (e.g. if the node represents a user, its name is the user's name), degree and local clustering coefficient, explained below. Links connect nodes and can have weight(importance or value), direction or other attributes.

Graphs can be directed, weighted, both or neither depending on whether their links have direction, weight, both or not:

- Directed graphs (or Digraph) have links with direction. For example, in a graph representing the roads in a city, some roads are two-way (link directed both ways) but some roads are one-way (link directed only one way).
- Weighted graphs have links with weight. For example, in a graph representing the roads in a city, the weight could be the length of the road in meters.
- Weighted directed graphs have links with both weight and direction. For example, in a graph representing the roads in a city, the link's weight could be the length of a road, and the direction the way of the street.

2.1.1 Degree

A node's degree is the number of links connected to that node. For example, in the research collaborations network, the degree of an individual is the number of scientists he/she co-authored a paper

with. The **Weighted Degree** is used in weighted networks and is calculated by summing all the weights of each link connected to a specific node.

In-Degree and Out-Degree

For directed networks/graphs only, a node's in-degree is the number of links a node has pointing inwards. Contrarily, a node's out-degree is the number of links pointing outwards. Similarly, the **weighted** in-degree is the sum of all weights of links pointing inwards. The out-degree is the sum of all weights of links pointing outwards.

2.1.2 Average Shortest Path

Average Shortest Path(ASP) refers to the average of all shortest paths between any two nodes in the networks. The shortest path between two nodes i and j is defined as the least number of links between i and j.

2.1.3 Clustering Coefficient

A network's Clustering Coefficient represents the fraction of vertices for which the following is true: if vertex α is connected to vertex β and vertex β to vertex δ, then α is also connected to vertex δ [19]. For example, in a social network this value represents the mean probability that your friend's friend is also your friend. This value has been defined in more than one way [19]:

1. As the mean probability that two vertices that are neighbors between each other are neighbors with a third vertex themselves:

$$C = \frac{3 * \text{number of triangles in the network}}{\text{number of connected triples of vertices}} \tag{2.1}$$

2. As the average of the local clustering coefficients. The local clustering coefficient is defined for each node as [19]:

$$C_i = \frac{\text{number of triangles connected to vertex i}}{\text{number of triples centered on vertex i}} \tag{2.2}$$

   The same value can be calculated in a different manner, as defined in the NetworkX python package [20], and first introduced in [21]:

$$C_i = \frac{2 * \text{number of triangles connected to vertex}i}{\text{degree(i)} * \text{(degree(i)-1)}} \tag{2.3}$$

The values obtained depend on the method used. For this project, since we are using NetworkX, we will calculate the clustering coefficient of the network as the average of all values of C_i as defined in equation 2.3.

2.2 Characterization of Complex Networks

Networks science allows for the extraction of intrinsic information from real networks. In order to model, characterize, analyze, classify and validate our network it is essential to acknowledge the existent connectivity and topology measurements [22].

2.2.1 Average Degree $< k >$

The Average Degree, $< k >$, represents the average number of links that each node has and it can be calculated by averaging the degrees of all nodes. For undirected networks it can also be calculated by the following equation:

$$\langle k \rangle = \frac{1}{N} \sum_{i=1}^{N} k_i = \frac{2L}{N} \tag{2.4}$$

2.2.2 Degree Distribution $P(k)$

A network's Degree Distribution, $P(k)$, is the fraction of users $P(k)$ with k links in the network. In other terms, it estimates the probability that a randomly selected node has degree k. For real networks, the degree distribution is heavy-tailed due to finite-size effects [19, 23]. One possible way to have a cleaner plot is to instead calculate the cumulative version of $P(k)$, $D(k)$ with:

$$D(k) = P(k' \geq k) = \sum_{k'=k}^{\infty} (P(k')) \tag{2.5}$$

For real networks the tail of $P(k)$ often follows a power-law with exponent γ, called degree exponent [19, 23, 24]. The cumulative distribution also follows a power-law but defined by $\gamma - 1$ [19]:

$$D(k) \sim k^{-(\gamma-1)} \tag{2.6}$$

However, the degree distributions of networks may follow an exponential distribution instead [19]. In order to verify if a degree distribution follows a power-law or exponential distributions one should plot the cumulative degree distribution on a log-log scale (power-law) or a log-linear scale (exponential), and sequentially perform a fit to the pretended distribution.

2.2.3 Scale-free Networks

Scale-free Networks are networks with a power-law degree distribution [25]. This is due to the fact that the spread around the average degree can be arbitrarily large(free of scale). For many scale-free networks the exponent of the power law γ) is between 2 and 3 [19]. Taking in mind that the second moment of the degree distribution $< k^2 >$ is used to calculate the variance, these networks have a degree distribution with a variance that tends to +infinity when there is not an upper limit to the number of nodes (i.e if $N \mapsto +\infty$ then $< k^2 > \mapsto +\infty$).

However there are networks for which γ is not between 2 and 3. Generalizing, for a scale-free network the n^{th} moment of the degree distribution is (as shown in [18])

$$\langle k^n \rangle = \int_{k_{min}}^{k_{max}} k^n p(k) dk = C \frac{k_{max}^{n-\gamma+1} - k_{min}^{n-\gamma+1}}{n-\gamma+1} \tag{2.7}$$

where $p(k)$ is the degree distribution, and $p(k) = Ck^{-\gamma}$.

From equation 2.7 it can be deduced that all moments that satisfy $n \leq \gamma - 1$ are finite and, likewise, all moments larger than $\gamma - 1$ diverge [18]. The value $\gamma = 3$ is called the *critical point* due to its theoretical interest: it represents the threshold where the second moment (variation) no longer diverges.

2.2.4 Small-World Property

Small-World Property refers to the fact that any two nodes are very close in a network. This property has been shown by Milgram through the use of letters [26]. Later, Watts and Strogatz [27] have shown that small world networks (networks that present the small world property) have a relatively higher clustering coefficient and a smaller average shortest path (ASP) when compared to random networks with the same number of nodes and edges. Recently this notion has become more formalised: networks are called small world if the value of the ASP increases slower or equal than $log\ n$, where n is the number of nodes and the network's size increases with fixed average degree [19].

2.2.5 Node Centrality

Node centrality is a measure on how significant a node is within the network. This concept of significance depends on the network, originating a number of measures possible to evaluate centrality. In the context of community structures we are going to focus on degree, eigenvector and betweenness centrality.

Degree centrality

A graceful way to determine centrality is using the node's degree. The higher the nodes degree the more significant is the node i.e the nodes with the most connections are considered the most significant. In social networks the most popular people (who have the highest number of friends), with the highest degree centrality, are usually considered the most important.

Eigenvector centrality

One may argue that a node's significance not only depends on the number of connections but also on the value of those connections. In other words eigenvector centrality focuses on the node's neighbors' centrality. For node i the eigenvector centrality is the i^{th} element of the vector x defined by the equation $Ax = \gamma x$, where A is the adjacency matrix, and γ the eigenvalue. A variant of this method is used by Google to rank Web pages [28].

Betweenness centrality

The structure of the network depends fundamentally on the nodes that have the most control over the flow of information. Betweenness centrality is a simple way to find these nodes - it is calculated by counting the number of shortest paths that go through a node. However, this measure does not account for the fact that information may not flow through the shortest path, giving rise to some variations such as “flow betweenness” and “random walk betweenness” [28].

2.2.6 Community partitioning

Community partitioning of a network consists in the segregation of nodes according to their connections, forming sub-units of highly interconnected nodes [29].

Two major types of community partitioning exist, distinguishable by the type of communities each generates. The first allows to obtain crisp communities, where a node can only belong to a single community. The second yields overlapping communities, where a node can belong to more than one community.

In this thesis, we will only use crisp community finding algorithms. For that reason this section will focus on this kind of methods only.

Multiple algorithms to obtain the before-mentioned crisp communities exist. The Louvain, the Girvan-Newman and the Leiden methods [30–32] are examples of such algorithms.

Concisely, the Louvain consists of a heuristic algorithm for efficient modularity optimization to extract the network’s community structure [30]. The Girvan-Newman method removes edges having the highest edge-betweenness value (a generalization of the betweenness centrality for edges) since these are more likely to be between communities [31]. The Leiden method is an improvement on Louvain, introducing a refinement stage in the algorithm, where previously established communities may be dismantled into sub-communities [32].

2.2.7 Network Robustness

To understand the meaning of network robustness to random failures, we must understand some basic concepts on percolation theory [18]. The main question we need answered is: ”what is the disruption threshold of a scale-free network?”. Or, more specifically, ”What is the minimum amount of nodes that one needs to remove to disrupt a network whose degree distribution follows a power-law?”

First, we need to define what it means for a network to be disrupted - a network is called disrupted if it doesn’t have a noticeable giant component. A network has a giant component if it meets the Molloy-Reed Criterion [33], defined as:

$$\kappa = \frac{\langle k^2 \rangle}{\langle k \rangle} > 2 \tag{2.8}$$

Where $\langle k^2 \rangle$ is the second moment of the degree distribution and $\langle k \rangle$ the average degree.

Random Removal

The threshold of randomly removed nodes beyond which the giant component disappears can be calculated by the following formula, as introduced by Cohen et al. [34]:

$$f_c = 1 - \frac{1}{\frac{\langle k^2 \rangle}{\langle k \rangle} - 1} \tag{2.9}$$

For a scale-free network generated by the configuration model defined by the degree exponent, γ and the minimum and maximum degrees, k_{min} and k_{max} [18]:

$$f_c = \begin{cases} 1 - \frac{1}{\frac{\gamma-2}{3-\gamma} k_{\min}^{\gamma-2} k_{\max}^{3-\gamma} - 1} & 2 < \gamma < 3 \\ 1 - \frac{1}{\frac{\gamma-2}{\gamma-3} k_{\min} - 1} & \gamma > 3 \end{cases} \tag{2.10}$$

Attack Resistance

The percentage of removed (targeted) nodes of a scale-free network (generated by the configuration model with parameters $k_{\min}$, $k_{\max}$ and γ) beyond which the giant component disappears can be calculated by the following equation (introduced in [35, 36] according to [18]):

$$f_c^{\frac{2-\gamma}{1-\gamma}} = 2 + \frac{2-\gamma}{3-\gamma} k_{\min} \left(f_c^{\frac{3-\gamma}{1-\gamma}} - 1 \right) \tag{2.11}$$

Experimental Analysis

In most networks, the procedure to estimate the critical threshold involves removing nodes in a specific order (randomly or by degree, for example) and registering the fraction and number of nodes belonging to the giant component. Usually, a good approximation of the critical threshold can be determined by the fraction of nodes needed to reduce the size of the giant component to 1% of it's original size [18].

2.3 Related Work

Studies on citizen science's origins and future (e.g. [1]) as well as users motivation and social nature (e.g. [37]) is vast, however little work has been done on the analysis of the nature of interactions of citizen science projects. As far as we could find, only two studies have been published applying networks to the study of citizen science.

In 2015, Aristeidou et al. [38] performed a social network analysis on the participation and collaboration in a citizen inquiry community, Weather-it. First, the authors determined the most cooperative and outgoing members by analyzing the network of cooperation between users based on the nodes': 1) degree and betweenness centrality and 2) directed weighted degree centralities. Next, they studied the networks created from Data Creation, Comments and Forum Posts in order to deduce the preferred

ways for users to contribute. Finally, the authors examined the network created from the co-membership of missions (projects), determining two types of users: those interested in a single project and those into investigation in general. Furthermore, they separated the network into different mission types and applied the Louvain method to determine the strongness of ties within types of mission.

Later that year, the same authors used social network analysis to interpret the evolution of the members and their interactions in the same citizen inquiry network (Weather-it) over a period of 14 weeks [39]. From their results the authors suggested that the sustainability of the community requires the engagement of an administrator. This person should be active on the moderation and discussion on the website as well as promoting activities through social media (e.g. Facebook).

More recently, Herodotou et al. [40] performed a social network analysis on the participation by young people in the Zooniverse platform [11]. In a network representing projects as nodes, and co-chosen projects as ties, the authors applied both betweenness centrality and degree centrality to identify the most chosen projects among young people. Next, they divided the network into sub-communities, using the Louvain method, to identify which Zooniverse projects tend to be co-chosen by young users. From their results, they suggested that young users tend to choose projects based on their interests or due to targeted publicity.

However, neither of these projects uses the network's characteristics (e.g Degree distribution, Average Shortest Path) to retrieve information regarding the community. In this thesis we will show how evaluating these characteristics may suggest valuable information regarding the users, their interactions and interests in citizen science projects.

2.3.1 Citizen Science and Collaboration Networks

Network analysis can be used in a variety of fields, from the study of drugs (e.g. [41, 42]), investigation on protein interactions (e.g. [43]) all the way to social relations (e.g. [44]) and disease spreading (e.g. [45, 46]). Within the range of possible networks that have been analysed, the network extracted from the BioDiversity4All database appears to be a **collaboration network** since in these networks collaborators cooperate between each other to achieve a certain objective. In our case, different users are cooperating to identify an organism.

Collaboration networks have been analysed as social networks, i.e collaborators are considered connected if they have cooperated together, thus, disregarding the object of collaboration (e.g research papers in scientific collaboration, observations in BioDiversity4All) and highlighting the social nature of the network. Newman alerted other researchers to the presence of a valuable source of collaboration data in bibliographic databases. Furthermore, he unveiled that the collaboration networks he studied are small world, with an average shortest path scaling logarithmically with the number of authors, and scale-free, with a degree distribution that follows a power-law [47, 48]. Another more recent example is a study on the biomedical research collaboration network constructed from the research grants collected at University of Arkansas for Medical Sciences, which was conducted using social network analysis. In this work a weighted network model was created to represent collaboration strength, allowing the

measurement of its characteristics, recognizing key authors and suggesting potential collaborations. The authors concluded that collaboration networks at UAMS are small-world but not scale-free [49].

Barabâsi et al. [50] also analysed a scientific co-authorship network, but focusing on its dynamic structural properties by evaluating the network's topology and characteristics over time. Concluding that the network's key characteristics (e.g. the diameter, the clustering coefficient, as well as the average degree) variate over time, but can show at any given time its Small World characteristics. Additionally, they have shown that the degree exponent does not variate over time.

Network science can also be used to expose the reliability of the network, as well as its resistance against attacks. It has been shown that scale-free networks are error tolerant, but not attack resistant [51]. According to Albert et al. [51] the weakness of these networks is that they depend on a few nodes to maintain connectivity. These nodes can be identified using the centrality measures seen in section 2.2. More recently, an article on the optimal influence problem [52] has shown the deep correlation between the concept of influence and the network's reliability, and introduced the concept of collective influence (CI) of a node. This measure can be calculated by multiplying the reduced degree of a node i ($k_i - 1$) and the sum of the reduced degrees of nodes within distance l from node i ($\sum_{j \in \partial B(i,l)} k_j - 1$) within a time order of $O(NlogN)$ [52] by using a max-heap data structure [53]. The authors have also compared this solution to other known solutions using algorithms such as high-degree and centrality measures, concluding that the removal of nodes with highest CI value is more effective at disrupting the network than any other known solution. As reported by [54], CI's effectiveness on real networks requires further testing and an effective way to determine the value of diameter l must be developed.

Network analysis has been described as an organizational X-ray, mapping relationships that are not readily evident, allowing for a deepper insight on the qualities, weaknesses and potential of a network [55]. In this work we aim at finding the most important nodes for the networks connectivity using multiple known solutions. Also, we aim at testing the networks dependence on the hubs. Then we will proceed with an analysis on the evolution of the network over time. Furthermore we aim at identifying and explaining the network's community structure. Finally we will perform an analysis on the behaviour of users regarding interaction as well as its evolution.

Chapter 3

Materials and Methods

In this chapter, we describe the raw data. We also show its intricacies and problems as well as the adaptations we made to overcome them. Furthermore, we show how we created the main networks used throughout the project as well as how we adapted them to create their weighted and directed versions.

3.1 Data

The BioDiversity4All team provided us with the information on 568403 different observations produced by 16795 users from the BioDiversity4All database. This includes all data from the BioDiversity4All platform, excluding observations on rare species to protect endangered organisms. Also, to avoid biases, we excluded any observations outside Portugal or without a taxon attributed.

The retrieved data consisted of two CSV data files. The first has every observation (identified by *observation_id*), its creator (*user_id*) and date of creation (*user_id*) as well as other information. The second CSV file consists of the participations of users and is organized in the following manner: for each observation, there are as many lines as the number of identifications that were made; each line has multiple values referring to each participation, such as the observation identifier (*observation_id*), the user who participated in the observation (*user_id*), the taxon of the species that was identified (*taxon_id*), the date of creation (*created_at*), the date of observation (*observed_at*) and others.

In our analysis we used the *observation_id*, *user_id* columns for the creation of the networks, as well as the *created_at* and *observed_at* to obtain time information for the time dependant networks. For the community analysis we used the *taxon_id* column to correlate the communities obtained with the users' taxon participation.

3.1.1 Observations Date of Creation

Before 2018, the BioDiversity4All was not a part of the iNaturalist network. When these two databases merged, the users that were in the BioDiversity4All, had to confirm the import of all their observations to the new iNaturalist platform. However, every observation a user had in the BioDiversity4All was added

to the new platform with the date of import, instead of the original creation date in the BioDiversity4All. Consequently, there are a number of observations after 2018 that were actually created in the years ranging from 2010 to 2018. This has a great effect on analysis involving the use of the creation date.

To get a more approximate date of creation, we used the date of observation, since this date is usually close to the original date of creation. This adaptation was applied to observations whose date of creation was after 01/01/2018 and whose date of observation was after 01/01/2010. This little adjustment has some rather large implications. In table 3.1 we show the number of users before and after the aforementioned adaptations were performed.

Table 3.1: The number of users before and after the alterations on the date of creation to circumvent the "date of creation = date of import" data intricacy.

	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020*
Users Bef.	2	1	1	13	26	65	106	197	381	760	2880	4992	1708
Users Aft.	2	1	593	750	641	562	719	924	1322	1727	3311	4586	1150

3.2 Networks

In this section we show the meaning of each of the networks created. Concretely, we present the fundamentals behind each network - why they were built and how.

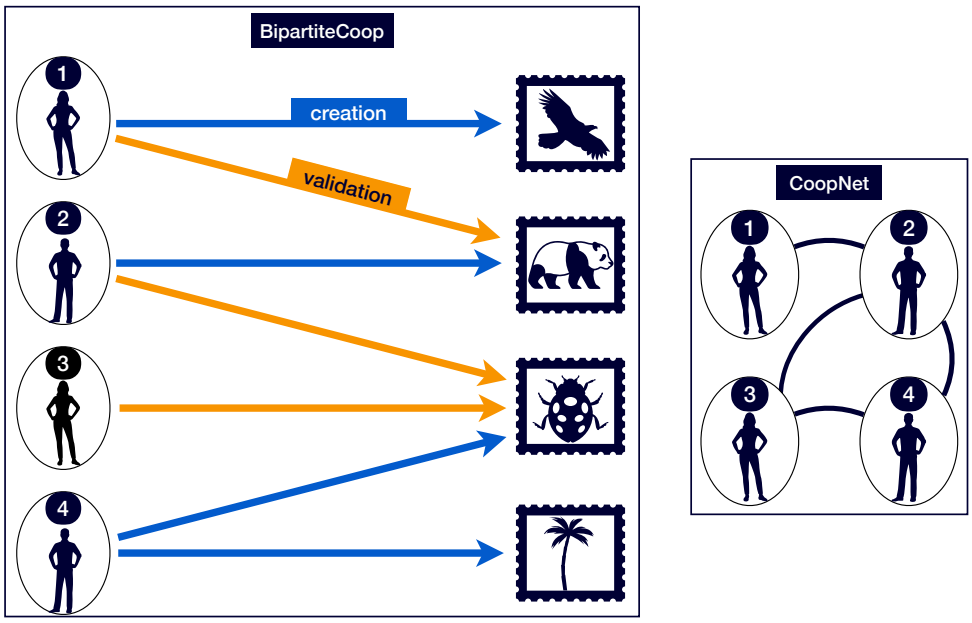


Figure 3.1: The creation of the BipartiteCoop and CoopNet Networks. The users in the platform can create or validate observations, represented in blue and orange respectively. Thus, creating a network of users and the observations they created/validated (BipartiteCoop). Each user can then be connected to the users he cooperated with in a cooperation network (CoopNet), in black, on the right.

3.2.1 BipartiteCoop

BioDiversity4All is an online citizen science project, where users can broadcast observations on organisms. These observations usually have a picture, a time and a location. Each observation in the dataset can have multiple users cooperating, providing suggestions and comments, validating the identification of an organism's taxon. Ultimately, this will result in multiple users cooperating on to identify multiple observations, creating a collaboration network. This collaboration network is a bipartite network where users are connected to observations they participated in, but never to other users. We called this network the **BipartiteCoop**[*]. An illustration of this network can be seen in Figure 3.1, on the left.

3.2.2 CoopNet

From the BipartiteCoop network it is possible to perform a deeper analysis on the users by extracting and analyzing the underlying social network present. This social network is formed from the collaborations of users, i.e each two users that participate in an observation must be connected in the new social network. At a micro level, this can be done by iterating every observation node, verify who are its neighbours (users that participated in that observation) and then create a link for every possible combination of any two neighbours (users). We called this network the **CoopNet**[*]. An illustration of this network can be seen in Figure 3.1, on the right.

3.2.3 Weighted/Directed Versions

With the BipartiteCoop and CoopNet networks, we are considering that users who have co-participated in an Observation multiple times have the same relationship than those who have only done it once. Another issue is how to distinguish between users that identify observations and users who create them. To differentiate such cases we created a weighted directed version of these networks.

In the weighted directed BipartiteCoop[*] a user's links have the same weight as the number of interactions she/he had with a given observation. Concerning direction, links pointing towards an observation mean the user participated in it, a link pointing from the observation to the user means she/he created it.

In the weighted directed CoopNet[*], cooperating users have links with weight equal to the number of times they have cooperated in the same observation (users who have shared the most observations have the strongest bond). Regarding direction, a link pointing outwards from an user (A) means A has cooperated in another user's (B) observation, links inward are exactly the opposite - B has cooperated in A's observation.

[*]The networks were created using the NetworkX python package [20].

Chapter 4

Network Analysis

In the BioDiversity4All platform, users interact by creating observations and participating in other's. This interaction forms a bipartite network - with users connected to observations - which can then be used to extract a social network - users are connected to users interacting in the same observations.

In this Chapterwe aim at characterizing these networks using some of the measures seen in Sections 2.1 and 2.2 We show that such analysis allows for a more profound understanding of the network by uncovering the nature of user's interactions. We then compare the obtained characteristics of our network to other previously studied networks.

4.1 BipartiteCoop

The first network created consists on a bipartite graph with only users and observations (BipartiteCoop). This network has 585198 nodes (16795 users and 568403 observations) and 1294774 edges connecting users to observations they created or validated.

To understand how much participation differs from user-to-user and observation-to-observation we plotted the fraction of users connected to a given number of observations as well as the number of users per observation (see Figures 4.1 and 4.2).

Figure 4.1-a shows that most users have participated on a small number of observations, while one user is connected to a rather high number of observations (over 75000 connections). More concretely, approximately 72% of users have participated in less than 10 observations, only 7% have over 100. Moreover, it seems that this distribution follows a power-law distribution. A possible fit is shown in 4.1-b, with a power exponent $\gamma = 2.02$ and an optimal minimum x value of $x_{min} = 466.0$ (method of Clauset et al. [57]). These results show that the network is highly heterogeneous concerning the users' number of observations.

Next, we analysed the graph's homogeneity concerning the number of users per observation (see Figure 4.2). We used a log-linear plot in order to better analyse the obtained distribution (Figure 4.2-b).

This figure shows that the number of users per observation is not homogeneous, presenting approximately 66% of observations with one or zero participations, but, also, a single observation with

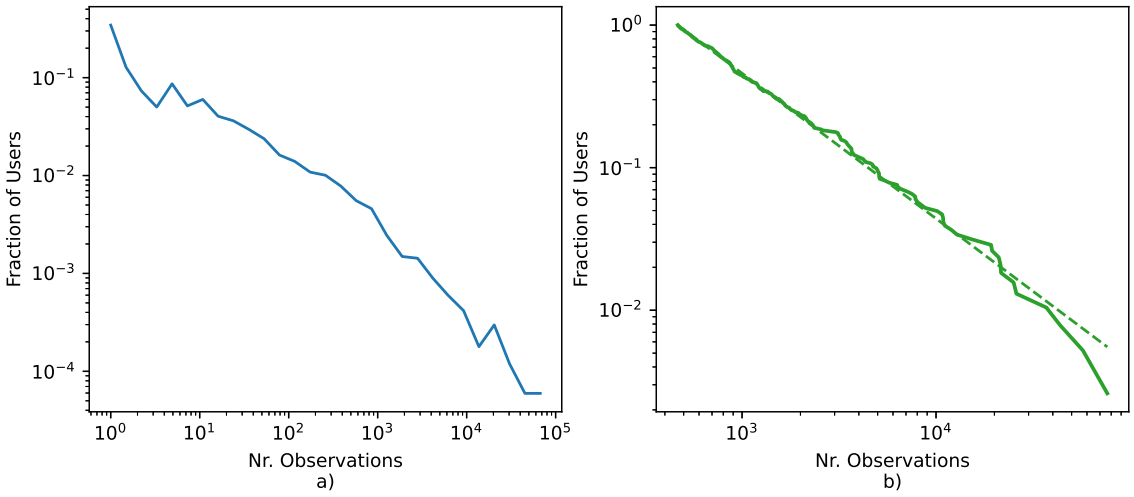


Figure 4.1: a) The fraction of users connected to a given number of observations in a log-log plot using logarithmic bins. b) A power-law fit to the same plot, regarding users with over 466 observations. Plot b) was obtained using the *powerlaw* python package [56]. The values demonstrate a high heterogeneity concerning the users' number of observations, with most users connected to only a few observations and a rather small number of users with thousands of observations.

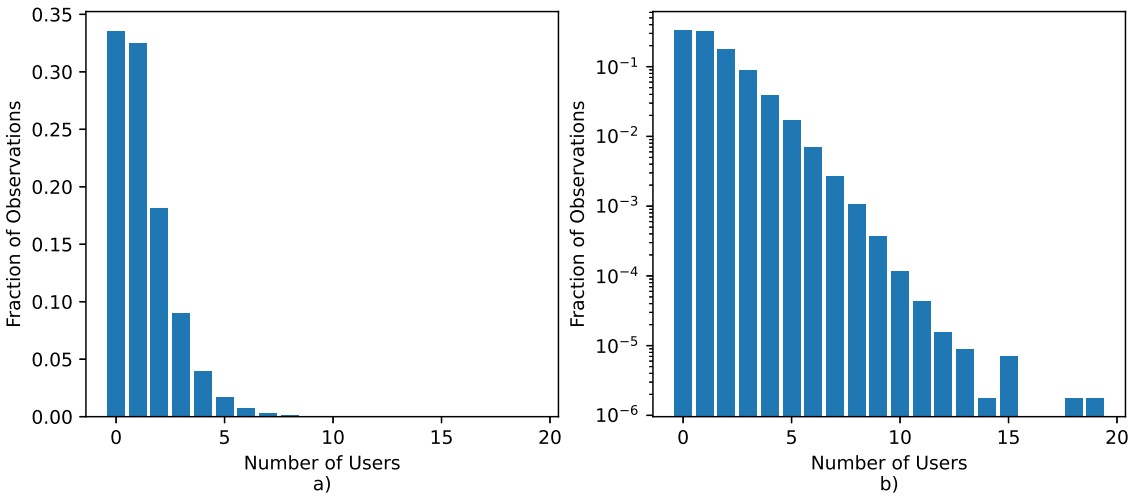


Figure 4.2: The number of users per observation on the BipartiteCoop network: a) distribution of users by observations on a linear plot and b) on a log-linear plot.

19 participations. This suggests that some few observations have higher interest on participating than others. Some explanations for this fact might be the popularity of a species or the discordance on the organism's taxon. For example, the reason for the high popularity of some observations seems to be the inability of users to agree on a taxon, resulting in multiple opinions and high participation.

4.2 CoopNet

Based on the cooperation between users, the CoopNet, on the other hand, focuses on the social part of the network. All the nodes in this network are users and a user has a link to another if they have cooperated in the same observation. This network has a total of 16795 nodes connected by 236153 edges.

In terms of sub-components, this network presents 1 giant component with 16005 nodes (representing approximately 95.3% of the network), 17 components with 2 nodes and 756 isolated nodes.

4.2.1 User Connectivity

To understand user interaction, we must understand how the number of connections of users in the network varies. In average, each user is connected to 28 other users, with a standard deviation of 127.8. This rather high value for standard deviation indicates that the number of connections is, at the least, wide-ranging. To further explore this apparent heterogeneity, we then analyzed the network's degree distribution. in a double-logarithmic plot (Figure 4.3-a). In this figure, we can see that there is a very high heterogeneity associated with the number of connections of each user. Most users have few connections but a small number, the hubs, have a high degree. For instance, the top 10 users (degree-wise) are connected to $\sim 66.0\%$ of the network.

By using logarithmic bins as well as the cumulative version of the degree distribution it's possible to eliminate the fat-tail and fit a power law to the curve (see Figure 4.3). The cumulative version of $p(k)$ fits a power law of the form $k^{-(\gamma-1)}$ for values of k greater than k_{min}. For our network $\gamma = 2.70$ and $k_{min} = 491$. These values were obtained using *powerlaw*, a python package which provides fitting and statistical analysis on heavy tailed distributions [56].

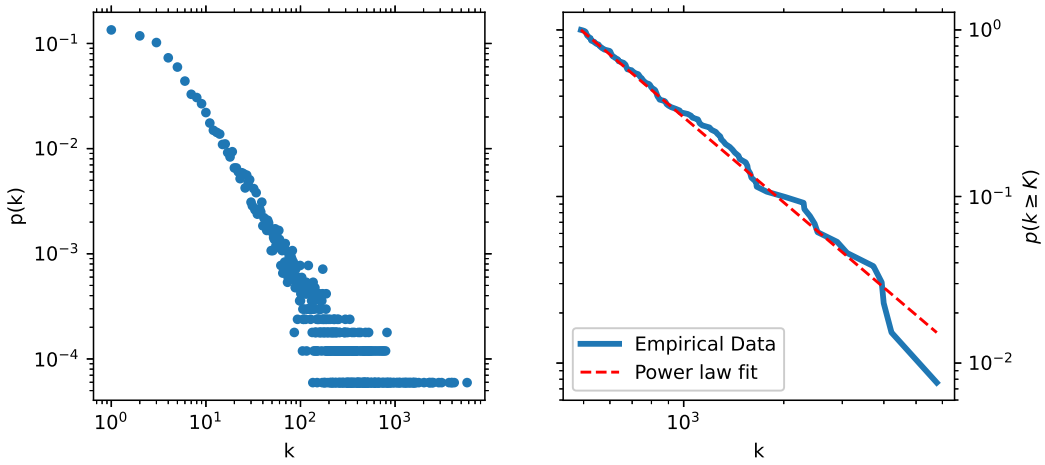


Figure 4.3: On the left: Distribution of node linkages $p(k)$. The fat tail of the distribution is due to finite-size effects [18, 58]. On the right: The cumulative version of $p(k)$, $P(K \geq k)$ and a power law fit. Most nodes have a very low degree but some few nodes (the hubs) have a significant high degree, suggesting a power-law dependence on the degrees.

4.2.2 Average Shortest Path and Clustering Coefficient

The Average Shortest Path (ASP) is the average distance between any two nodes in the network (see Sub-Section 2.1.2). However, this value is impossible to calculate for networks not fully connected, as is the case with ours. Instead, the same calculation is often performed on the giant component. We calculated the ASP for the giant component and the average clustering coefficient for the whole network

and obtained 2.70 and 0.59, respectively.

We tested these values against a random graph with the same number of nodes and edges as our network and found that the BioDiversity4All network has a smaller ASP (2.70 *vs.* 3.22) and a 199% higher clustering coefficient (0.59 *vs.* 0.0015). This comparison provides a point of reference - the values obtained show that the BioDiversity4All community is highly structurally interconnected.

4.2.3 Comparison to other networks

Table 4.1: Basic statistics of some real networks and the CoopNet for comparison. The tabulated properties are: the number of nodes (N); the number of links (L); the average degree ($< k >$); the average shortest path (ASP); the degree exponent (γ); and the average clustering coefficient (C). The "–" means that data was unavailable or not found. All statistics were retrieved from [19].

Network	N	L	$< k >$	ASP	γ	C
CoopNet	**16795**	**236153**	**28.12**	**2.70***	**2.7**	**0.59**
Film actors	449913	25516482	113.43	3.48	2.3	0.78
Email Messages	59912	86300	1.44	4.95	$1.5/2.0$	0.16
Citation Network	783339	6716198	8.57	–	3.0	–
WWW nd.edu	269504	1497135	5.55	11.27	$2.1/2.4$	0.29
Internet	10697	31992	5.98	3.31	2.5	0.39
Protein Interactions	2115	2240	2.12	6.80	2.4	0.071

* Value associated with the giant component of the network.

In Table 4.1 we show the obtained values for our network and other known networks, according to values obtained from [19]. It is remarkable that such different networks in size and nature, with nodes and links representing such distinct entities (people, documents, routers and proteins), present such similar values for the degree exponent. Every network in the table presents a degree distribution following a power law with γ between 2 and 3, indicating that these networks are highly heterogeneous degree-wise.

4.3 Hub Dependence

As we have seen before, our network's inter-connectivity is highly uneven - while most users have only one connection, the top 10 most connected users are connected to 62% of the network. To understand how this disparity influences the BioDiversity4All community, we must evaluate how important the most connected users (the hubs) are. A way to do this is to focus on the reliability of the network. In other words, we evaluate the difference between removing a 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 in Sub-Section 4.2.1 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.

Focusing on the network's reliability, we may calculate the theoretical and practical values for the percentage of nodes that we need to remove to disrupt the network.

The network's theoretical critical thresholds of disruption were calculated using the equations in Sub-section 2.2.7. These have shown that 99,8% of random nodes need to be removed to disrupt our network. However, we find that only 4,4% of the hubs need to be removed for the same effect, according to equation 2.11*.

The theoretical values obtained show that our network is very susceptible to the hubs leaving the platform, but not to random dropouts. Next, we tested these conclusions by analysing how the network responds to the removal of nodes. First, we analysed how the random removal of nodes (simulating accidents) influences the size of the network's giant component. Secondly, we performed the same analysis but removing the nodes with the highest degree (simulating an attack).

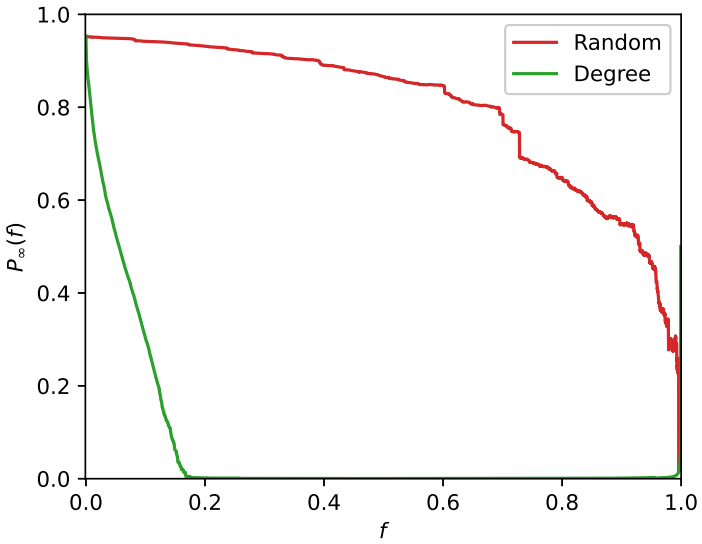


Figure 4.4: Estimated probability that a given node belongs to the giant component after a f fraction of nodes has been removed: in random order (in red) and by decreasing degree (in green).

Figure B.1 shows the 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 network shows to be able to withstand most users leaving the platform, since we need to remove almost 100% of the randomly selected nodes for the giant component to effectively disappear. However it cannot endure the disengagement of the most connected nodes - removing the top 16,7% nodes with highest degree reduces the percentage of nodes belonging to the giant component to below 1%. In other words, the BioDiversity4All's connectivity is highly dependant on the most connected users - and

These results are somehow expected. Has shown before our network presents a great heterogeneity degree-wise, with most nodes presenting small degrees. Consequently, random removal of nodes will most likely remove a node with very few linkages, doing little damage to the network's connectivity. On

*The results may not be as accurate since equation 2.11 is meant to be applied to the theoretical model of a scale-free network, defined by the parameters k_{min} and γ.

the other hand, by having only a few nodes (with high degree) maintaining its connectivity, the targeted removal of these hubs is very effective at disrupting the network.

4.4 Evolutionary Analysis

To better understand the evolution of the network, we divided the original data year by year, from 2008 to 2021*. From each dataset, we then extracted the CoopNet graphs corresponding to each time interval.

To evaluate the evolution of the CoopNet's topology, we analyzed its growth - in the number of users (its nodes) - and the evolution of its properties. As the network grows we expect the number of users and the average degree to vary but the degree exponent not to variate as much due to the network's scale-free properties.

In Figure 4.5 we show the evolution of the number of users in the BioDiversity4All platform throughout the years. Figure 4.6 shows the variation of the average degree and degree exponent. It is clear that since the year 2015 the network has grown in the number of users, and the average degree has increased but the degree exponent barely varies (Figure 4.6-b), confirming our expectations. Particularly interesting is the variation of the degree exponent - since 2010 the value barely changed - showing that the network's connectivity does not variate with the number of participants. Also, the degree exponents remained between two and three for the entirety of the time analysed, suggesting that this network preserves its scale-free property.

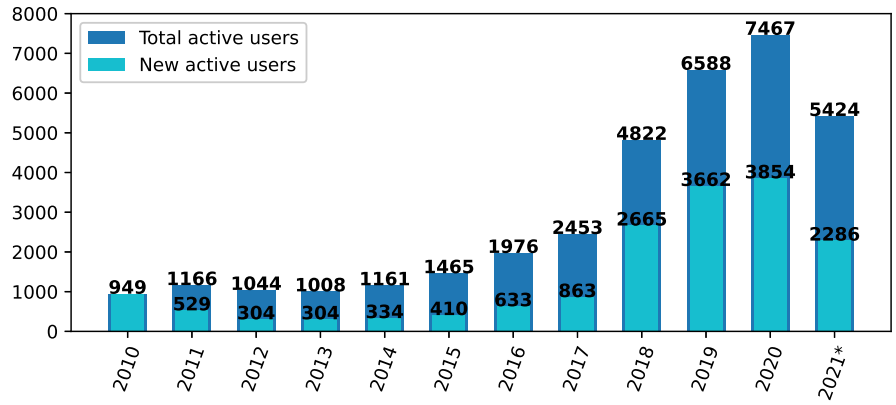


Figure 4.5: The number of users in the network by year. There is a significant increase in the number of users in the network.

To test the evolution of the "distance" between nodes, we analysed the average shortest paths's (ASP) evolution throughout the years. We also compared the values obtained with $ASP_{max} = \frac{log(N)}{log(<k>)}$ - representing the maximum value of the ASP for the network to be considered small-world - for each year (Figure 4.7). For each year the value barely changes despite the significant increase in the number of users. Furthermore, for each year, $ASP \leq ASP_{max}$, meaning that the network is small-world as defined by Newman [19].

*The values for the year 2021 are incomplete: last entry in 02-06-2021

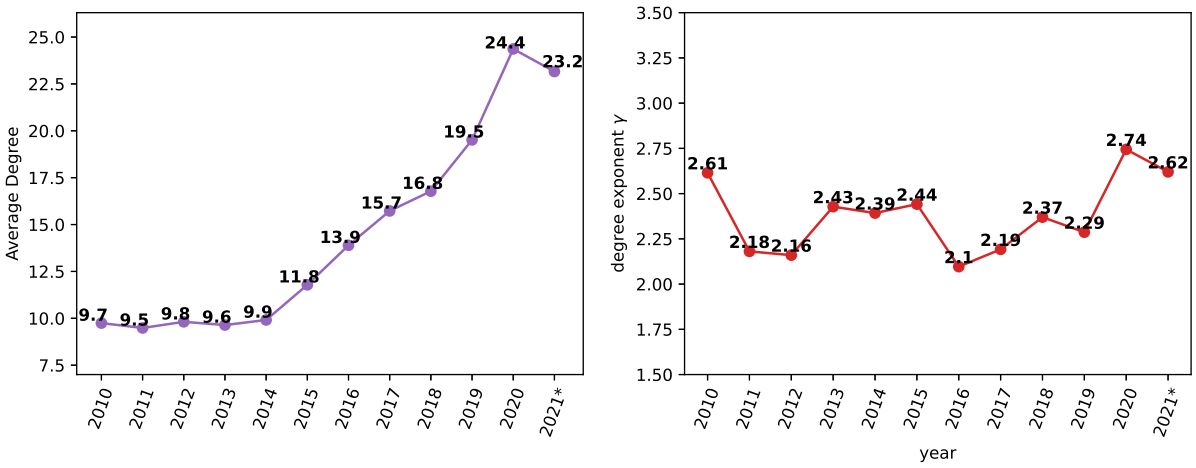


Figure 4.6: Values characterizing the nodes' connectivity by year: the average degree (left pane); b) exponents of the powerlaw distributions best fit to the degree distributions of the networks by year (right pane). Despite the variation in the average degree of users, the degree exponents barely change throughout the years. In other words, while the network grows, its nodes connectivity distribution is invariable.

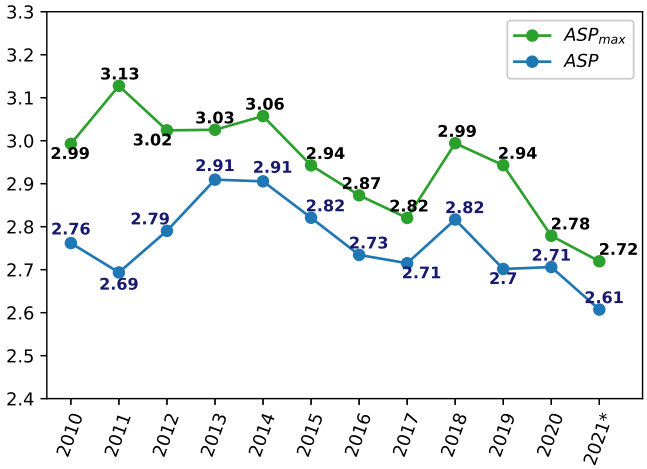


Figure 4.7: In blue, the values for the average shortest path (ASP) by year. In green the values of $ASP_{max} = \frac{log(N)}{log(<k>)}$, representing the maximum value of the ASP for the network to be considered small-world, by year. For each year $ASP \leq ASP_{max}$, meaning that the network is small world as defined by Newman [19].

Table 4.2: The represented properties are specific for each year since each is represented by a different network. The tabulated values are: the number of nodes (N); the number of links (L); the average degree ($< k >$); the average shortest path(ASP); the average clustering coefficient (CC); the values of the exponents for the degree distribution (γ);

Year	N	L	$\langle k \rangle$	γ	ASP	CC
2010	949	4624	9.74	2.61	2.76	0.58
2011	1166	5530	9.49	2.18	2.69	0.58
2012	1044	5122	9.81	2.16	2.79	0.56
2013	1008	4859	9.64	2.43	2.91	0.58
2014	1161	5753	9.91	2.39	2.91	0.59
2015	1465	8633	11.79	2.44	2.82	0.59
2016	1976	13726	13.89	2.1	2.73	0.61
2017	2453	19288	15.73	2.19	2.71	0.6
2018	4822	40451	16.78	2.37	2.82	0.57
2019	6588	64285	19.52	2.29	2.7	0.58
2020	7467	90968	24.37	2.74	2.71	0.55
2021[§]	5424	62809	23.16	2.62	2.61	0.54

[§] The values for the year 2021 are incomplete: last entry in 02-06-2021

Table 4.2 summarizes all the values obtained for each year. It should be noted that, as with the ASP and degree exponent, the clustering coefficient barely changes. This means that the users in the BioDiversity4All platform keep the same degree of clustering together throughout the years

Overall, the platform has grown in users throughout the years. Accordingly, the number of connections between users has also increased. However, the BioDiversity4All maintains its degree exponent, the ASP as well as the clustering coefficient. These values are coherent with [18], and show that the platform has inherent characteristics that do not variate through time, allowing it to maintain its small-world scale-free properties throughout the years.

4.5 Citizen science during COVID-19 pandemic

During the time this thesis was written, the COVID-19 pandemic was happening. Countries around the world applied confinement measures, to guarantee people would not leave their homes.

Supposedly, this period would be of great disturbance to the BioDiversity4All , since people would be unable to create observations. In this way, we expected that the COVID-19 confinement measures would have a negative effect on the evolution of the network. To test this, we evaluated the network's properties from observations one year before and after 22-03-2020 - the start of confinement in Portugal.

To get an overall picture of the variation of participation throughout the years, we plotted the weekly number of observations observed between 01/01/2018 and 02/06/2021 (see Figure 4.8). This figure shows an increase in the number of observations in 2020. Contrarily from what we expected, it would seem that the confinement - represented by the red line - has increased volunteer participation.

However, this increase in participation was not reflected in the network's degree exponent, as we

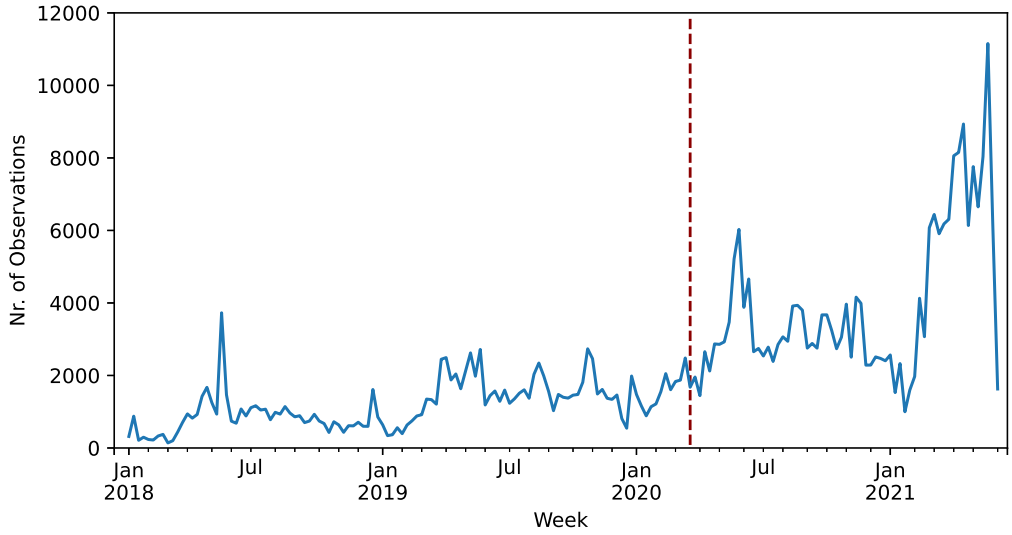


Figure 4.8: The weekly number of observations observed between 01/01/2018 and 02/06/2021. The vertical red line marks the first day of confinement measures, against COVID-19, in Portugal - 22/03/2020. The pandemic has driven a significant increase in participation.

have seen in the previous Section (Section 4.4). In fact, this result further demonstrates the conclusion obtained — the network's connectivity is time and scale invariant — with the network's degree exponent barely changing during the pandemic (see Table 4.2).

4.6 Summary

In this Chapter, we started by analyzing the BipartiteCoop network. The values obtained show that the popularity of observations is not homogeneous. In fact, some observations have more interest or are more polemic in the identification of the organism. Causing these observations to have a higher participation rate.

Next, we focused our analysis on the CoopNet. The results show that the users cooperating in the BioDiversity4All platform interact to form a scale-free network. This translates into a small fraction of users presenting a very high participation rate - either by creating observations or participating on others' - but most users simply performing one or two interactions in the platform. This kind of participation pattern has been found in other citizen science projects [6, 59].

Next, we discussed the importance of the highly connected users in the BioDiversity4All platform. There is a clear contrast in the users' influence on the network's connectivity. We have demonstrated how the absence of the most connected users (the hubs) would have an extreme impact on the platform. However, if a random node were to be removed, little impact would be noticed. These results show that this community greatly depends on the hubs in order to remain connected.

Finally, we focused our analysis on the evolution of the CoopNet's topology. We concluded that our network's connectivity *is independent* of time - in spite of significant differences in the number of users throughout the years, the network exhibits the same approximate degree exponent. Furthermore, the

cluster coefficient, the average shortest path as well as the degree exponent values obtained show that the CoopNet keeps its scale-free and small-world properties throughout the years.

Overall, the results presented in this chapter indicate that the users in the platform are highly interconnected and very ”close”. The next chapter focuses on the identification of existent communities in the network, in order to identify possible dividing factors between users, in the BioDiversity4All platform.

Chapter 5

Community Analysis

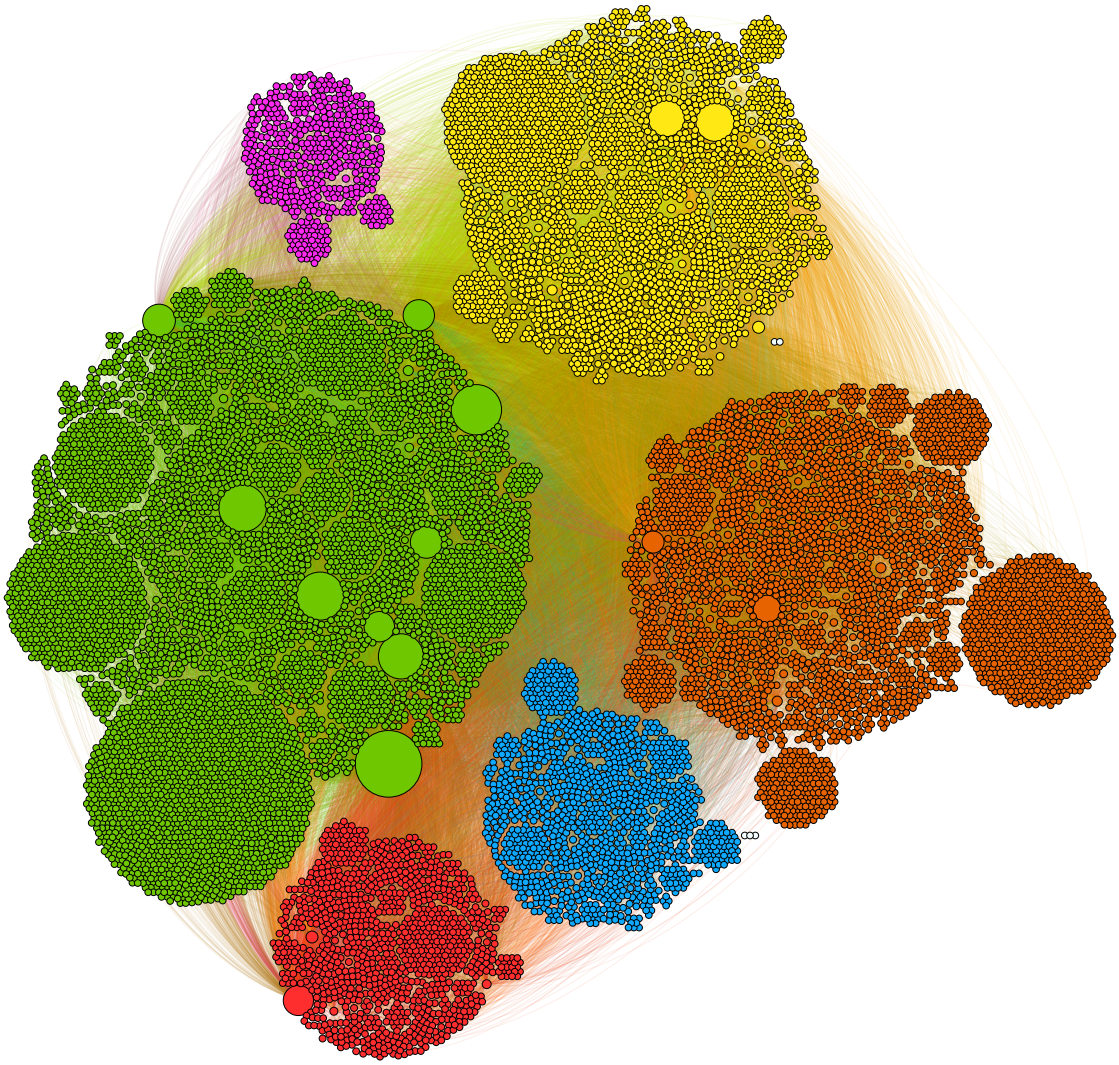


Figure 5.1: Visual representation of the BioDiversity4All network subdivided into communities obtained using the Louvain method. For readability purposes, only 20% of the links are represented.

The BioDiversity4All platform allows for the publication and identification of multiple kinds of organisms, including birds, plants, insects, reptiles, fishes and others. This feature defines the BioDiversity4All as fundamentally multifaceted, attracting users with various knowledge and interests.

In this chapter, we will investigate how the diversity in organisms and users influences the community structure of the platform. To achieve this, we will study the network's topological partitioning obtained from the connections between users.

5.1 Community Partitioning

There are several methods that can be used to identify the communities of a network (see Section 2.2.6). We tried both the Louvain and Leiden methods [30, 32] to obtain a node partitioning without community overlap (crisp communities). These yielded similar results in modularity (see Table A.1) as well as in the accuracy of partitioning by taxon.

For this paper, we used the Louvain method, since it is the most used in the literature [38, 40]. However, we will use the results obtained from the Leiden method to further show the dependability of the networks community structure in the users' taxon.

Both the Louvain and Leiden methods are heuristic, becoming impossible to obtain a precise value for the number of communities. To circumvent this, we applied the algorithm 100 times, registering the number of communities obtain in each run. Then, we averaged all values to reach a final approximate value.

The average number of communities detected by the Louvain method from the cooperation network of users (CoopNet) is 8.9 with a standard deviation of 2.3. These values are deceptive, nonetheless, considering the reasonably small size of some of the communities found. If we only consider partitions with more than ten nodes, this average becomes reduced to 6.0 communities with a standard deviation of 0.9.

In short, the most frequent result presents six communities with over 500 users and an additional two to four communities with under 10.

5.2 Partitioning by Taxon

Understanding the community structure of the network is one of the main goals of this thesis. A possible hypothesis is that users group according to their interests (and knowledge).

One hypothesis for the obtained partitioning is that users group according to their interest and knowledge on specific organisms. For example, a plant enthusiast would have more interest in observations of plants, thus interacting more with other users with the same interest — creating a community of plant enthusiasts. To evaluate this, we characterized the frequency of users' *taxon* interests by community (see Table 5.1; Figure 5.2 is a visual representation of the same values): Then, for each community, we averaged these values for each *taxon*, obtaining an approximation of the level of participation of the community's users around each *taxon*.

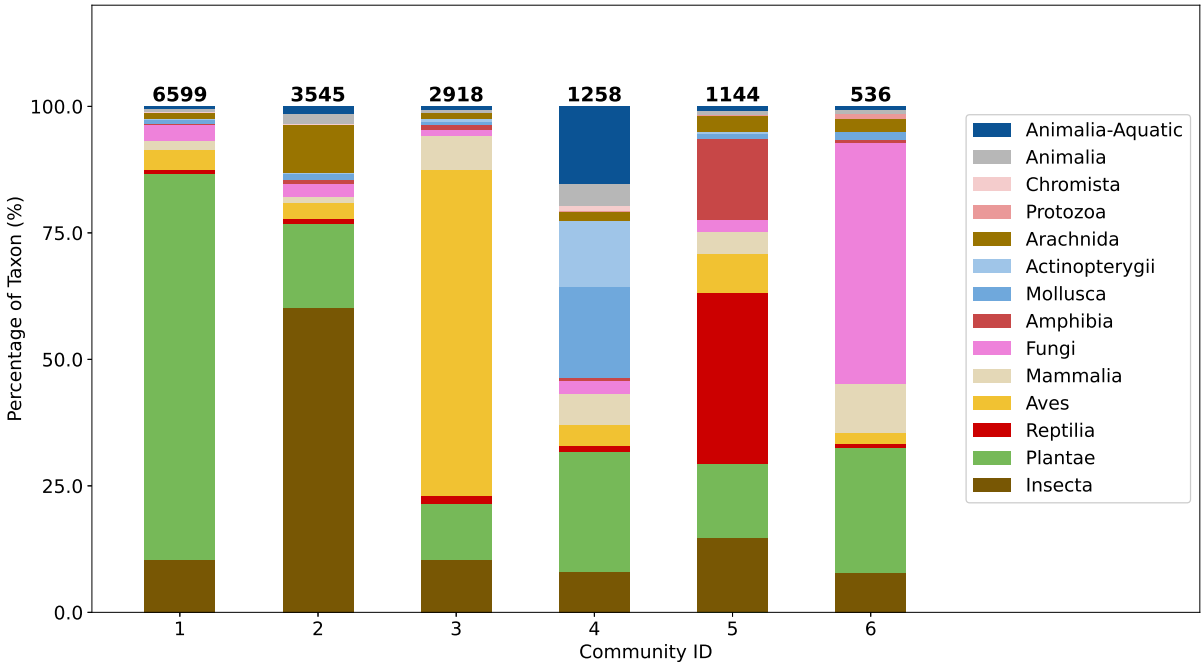


Figure 5.2: Cumulative bar plot representing the percentage of each taxon in each community. The respective values, are displayed in Table 5.1. Communities with less than ten users are not represented. Each community presents a very distinct pattern concerning its users' taxon preferences. Also, it shows that communities tend to have a predominant taxon or multiple prevailing taxa.

Table 5.1: Average percentage of the level of participation of users around each *taxon*, for each community. Each line in this table represents the average of the *taxon* interests of a community's users. The values obtained show that each community presents a very distinct pattern concerning its users' taxon preference.

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

Figure 5.2 is quite revealing - it shows that communities tend to have a predominant *taxon* or multiple prevailing *taxa* (values were obtained from Table 5.1):. Overall, every community presents a very distinct pattern concerning its users' taxon preferences. The three largest communities —— communities 1, 2 and 3 in Table 5.1 —— 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:

Table 5.2: Average percentage of the weight of each community in the participations around each *taxon*. Each column represents the percentage of users with a given taxon interest in each community.

Comm.	Size	Insect	Plant	Rept	Aves	Mam	Fungi	Amphi	Mollu	Actin	Arach	Proto	Chrom	Other	Other-Aqua
1	6599	20.0%	76.9%	9.3%	11.1%	20.8%	32.0%	8.2%	13.1%	8.7%	14.2%	42.7%	13.7%	19.4%	13.0%
2	3545	62.2%	9.0%	6.3%	4.7%	8.9%	13.6%	9.4%	12.1%	2.9%	65.0%	14.4%	12.9%	35.3%	17.3%
3	2918	8.9%	4.9%	8.9%	77.9%	36.6%	6.2%	9.8%	4.5%	8.0%	6.4%	3.0%	3.9%	10.0%	6.4%
4	1258	2.9%	4.6%	2.6%	2.2%	14.4%	5.0%	3.3%	64.9%	77.4%	4.7%	3.8%	66.9%	27.9%	59.2%
5	1144	4.9%	2.6%	72.0%	3.7%	9.4%	4.2%	67.9%	3.1%	2.8%	7.0%	6.4%	1.9%	5.1%	3.0%
6	536	1.2%	2.0%	0.9%	0.5%	9.9%	39.0%	1.3%	2.4%	0.3%	2.7%	29.7%	0.8%	2.4%	1.1%
7	3	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
8	2	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Total	16005	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%

- 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*.

In Table 5.1, the big size of the communities obfuscates the taxa with the least participation. For example, even though *Arachnida* is only 9,6% of community 2, this represents 65% of all users with interest in *Arachnida*. To identify this incoherence, we multiplied each value by the community's size and divided it by the total of each taxon, yielding Table 5.2. This table represents the percentage of the taxon present in the community, as opposed to the percentage of the community of a given taxon.

Overall, Table 5.2 shows that users with the same interests are often in the same community. Community 3 is an example of this - most of the *Arachnida* aficionados are put together with most Insect. Likewise, every community (apart from 7 and 8) shows a relationship between users and their interests:

- there is a community for plant enthusiasts (community 1 with 76.9% of all *Plantae*), but also 32.0% of all *Fungi*;
- another for Insects (community 2 with 62.2% of all *Insecta* and 65.0% of all *Arachnida*);
- and another for Birds (community 3 with 77.9% of all *Aves*);
- there is also a community that has the most aquatic animals enthusiasts (community 4, 64.9% of *Mollusca*, 77.4% of *Actinopterygii* and 59.2% of other aquatic animals).
- another with most Reptiles and Amphibians (community 5, with 72.0% and 67.9% respectively);
- community 6 presents 39% of all *Fungi*. These users appear to be split in two communities - community 1 alongside *Plantae* and in community 6 representing nearly half of the community.

These values show that most users are grouped in communities based on their taxon interests.

It should be noted that it is possible to obtain similar results using the Leiden method [32] (see Tables A.2 and A.3). This method did not create a separate community for *Fungi*. Instead it allocated these users to community 1, alongside *Plantae*.

Both methods partition the network based on the number of connections of nodes. This means that users allocated to the same community are more connected. In this manner, both methods have shown that, overall, users are more likely to connect to those presenting the same *taxon interests* and *knowledge*.

COVID-19 The community structure obtained since the beginning of confinement due to COVID-19 in Portugal (22/03/2020 $\sim$ 02/06/2021), is shown in the appendix — Table A.4 shows the percentage of the community of a specific *taxon*, and Table A.5 shows the percentage of a taxon present in each community. Overall, during the pandemic, the network presents a similar community structure organization by taxon.

5.3 Summary

In this chapter, we studied the network's topological community structure and it's correlation with the users' taxon preference. Overall, every community presented a very distinct pattern concerning its users taxon. Concretely, we found six distinct communities: a **Bird**, **Plant**, **Insect** communities with most *Aves*, *Plantae* and *Insecta*/*Arachnida* enthusiasts respectively, in the same manner, we also found a **Reptiles**/**Amphibians**, **Aquatic** and ***Fungi*** communities. These results suggest that the users with similar interests tend to interconnect the most.

In conclusion, although the BioDiversity4All integrates all sorts of organism enthusiasts, the users are highly connected and very close to each other (as concluded in the previous chapter). Nonetheless, the results in this chapter have shown that there are divisions between users - their interests and knowledge.

In the next chapter we aim to evaluate the behaviour and evolution of users users in the BioDiversity4All .

Chapter 6

Triggers And Validators: Behaviour Analysis

Observations are composed of different users, each adopting a particular role: those that create the observations, the Triggers, and those that participate in other users' observations, the Validators. In this chapter, we analyze how these preferences are distributed through the population of users and how they may evolve in time.

We believe that categorizing users allows for an easier understanding of their behaviour in the network. In this manner, 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

$$B(i) = \frac{v_i - t_i}{\max(t_i, v_i)} \tag{6.1}$$

Where the values of t_i and v_i are obtained from the user's $InDegree$ and $OutDegree$ of the Bipartite-Coop's weighted directed version described in Sub-Section 3.2.3.

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. Finally, *Hybrids* are equally likely to propose or validate observations —– for simplicity, we adopt $-0.1 < B < 0.1$, in this case.

The mean value for the Behaviour Value of all users in our network is -0.21, with a standard deviation of 0.96. The very high value for standard deviation shows that users' Behaviour Value does not tend to be in the midsection of the scale, as the mean value infers. To better study these values, we plotted the Behaviour Value's relative frequency (Figure 6.1).

Figure 6.1 shows two peaks at -1 and 1 behaviour values. Concretely, approximately 54% of users only create observations and nearly 37% only participate in others' observations. In total, around 91% of all users are in the extremities.

This *polarization* seem like a consequence of bias caused by the fact that most users have a small

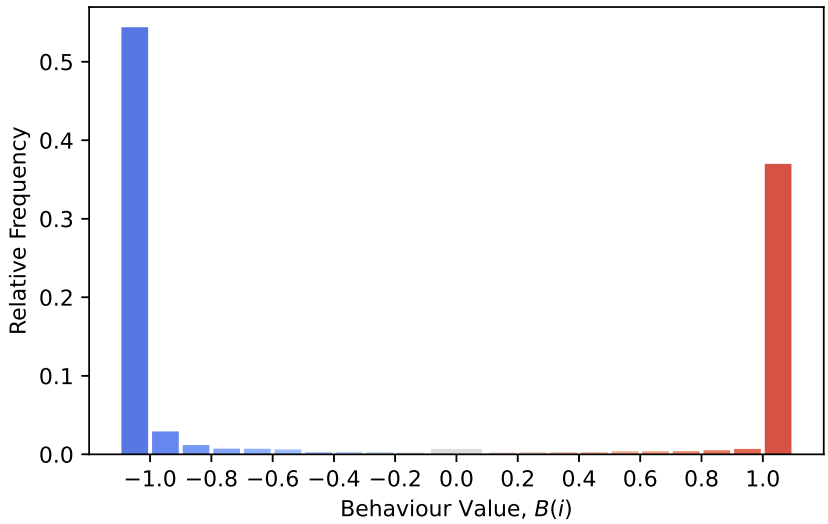


Figure 6.1: Relative frequency of Behaviour Values - the fraction of users with a given behaviour value. The closer a user is to the extremities, the more specialized she/he is in a type of activity. Utmost, a user can be a Pure Trigger or Pure Validator (the darkest red/blue colours) if $B(i) = -1$ or $B(i) = 1$, respectively.

number of participations. More concretely, as seen in section 4.1, roughly 72% of users have less than ten participations. For example, users with a single participation (34% of users) affect the probability distribution in fig. 6.1 drastically towards the extremities, since they could only have performed a trigger or a validation but never both.

To test this theory, we created a heatmap of behaviour values by degree (Figure 6.2), allowing us to analyze how user behaviour differs according to it's number of interactions with the platform.

Figure 6.2 shows the same two peaks in the -1 and 1 behaviour values as Figure 6.1. As expected,

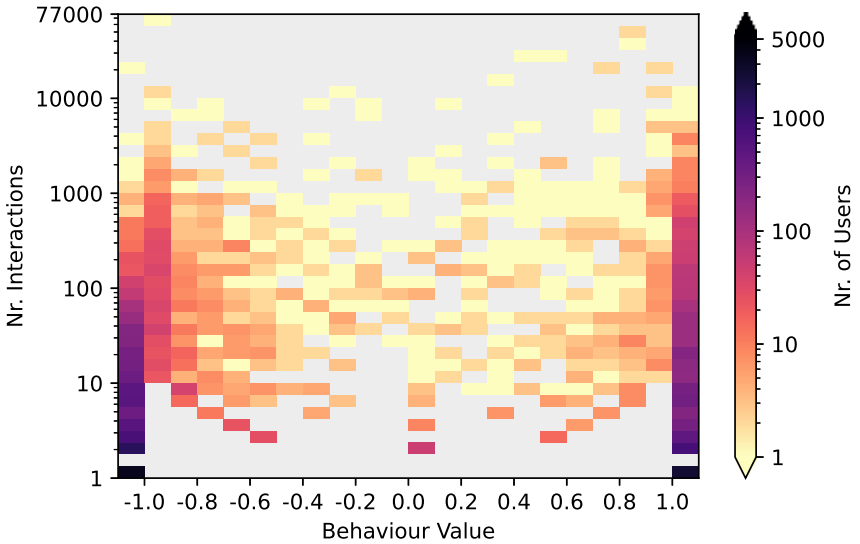


Figure 6.2: Linear-log heatmap representing the number of users (in a color log-scale) with a given cooperation and degree. Most Pure Triggers/Validators are users with a rather small number of connections.

most Pure Triggers/Validators are users with a rather small number of connections. However, it seems that users with more observations also have this tendency towards the extremities. Overall, this proves that the polarization of user behaviour is not biased from the high number of users with small degree.

However, by analyzing Triggers and Validators separately we notice a difference. At approximately 1000 interactions, there seems to be a sudden stop on the leftmost bar. In fact, less than 4% of users with 1000 or more interactions are Pure Triggers. This infers that users with a high number of triggers also explore the other side of the spectrum by validating other users' observations.

On the other hand, users with a high number of validations do not seem as wilful to experiment triggering. Above 1000 interactions, 27% of users are Pure Validators, never creating an observation - this is rather suspicious. After analysis on some of the more engaging Pure Validators in the platform's website, we concluded that a number of users have, contrarily to what is shown in the figure, created observations. However, every observation these users created was outside of Portugal. Since we are only analysing observations in this country, these users translate into Pure Validators (BehaviourValue $=$ 1) - even though they are, in the perspective of the whole platform, only Validators (0.1 $<$ BehaviourValue $<$ 1). This fact contributes even more to user behaviour polarization.

Overall, it seems that most users are either Pure Triggers or Pure Validators. The number of interactions seems to be have a higher effect on the probility of a user being a Pure Trigger than of being a pure Validator. As we have seen, Triggers with a high number of interactions, have tried validating at least once. However, the same does not happen for high degree validators. A possible bias contributing to this, is that users are able to validate observations in Portugal from anywhere in the world. Consequently, many Pure Validators are foreigner users that validated observations in Portugal, but have never created observations there.

6.1 Evolution of Users' Behaviour

Since the tendency to validate or trigger new observations may depend on the expertise of a 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 6.3. 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 the longest tend to specialize in the validation of other people's observations. But, they also alternate 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 are likely to be mostly Triggers, explaining the rather high Trigger percentage in the first year, as well as the sudden rise in the percentage of Validators in the second year.

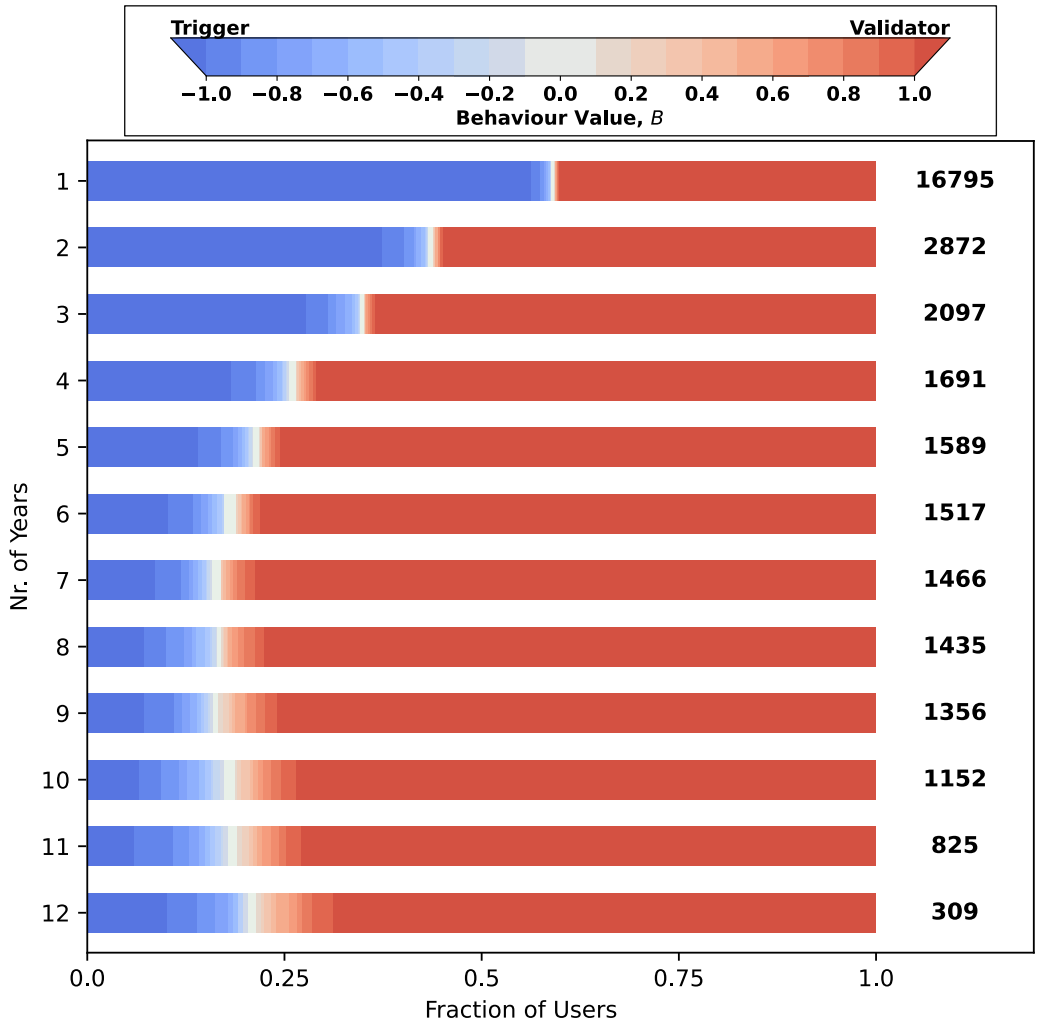


Figure 6.3: Fraction of Triggers, Validators and Hybrids, by the number of years since their first participation. The users are represented in a color spectrum that lies within the -1 to 1 range (top pane). Utmost, a user can be a Pure Trigger or Pure Validator (the darkest red/blue colours) if hers/his Behaviour Value is exactly -1 or 1, respectively. The increasing percentage of validators by the number of years suggests that users tend to specialize more in validation the more time they spend on the platform.

It would also be interesting to evaluate if users that are in the network the longest are also those that participate the most. Furthermore, how do users evolve regarding the number of participations? To answer these questions we plotted a histogram of the users' number of observations in order of the years they have been in the network (see Figure 6.4). It shows that, in their first years users do not tend to participate as much, and that users that last in the network tend to engage in more observations.

Overall, it seems users tend to participate more the longer they are in the network. This result suggests that users' engagement in the BioDiversity4All platform manifests both in the number of observations as well as in the continuity of participation over time. Thus, creating/validating more observations, the longer they engage with the platform.

Following this train of thought, it would be expected that the hubs — individuals with the most connections and observations — would be in the network for a long time. Figure 6.5 confirms this idea, showing that most hubs — including the highest number of connected user — have been in the network for several years.

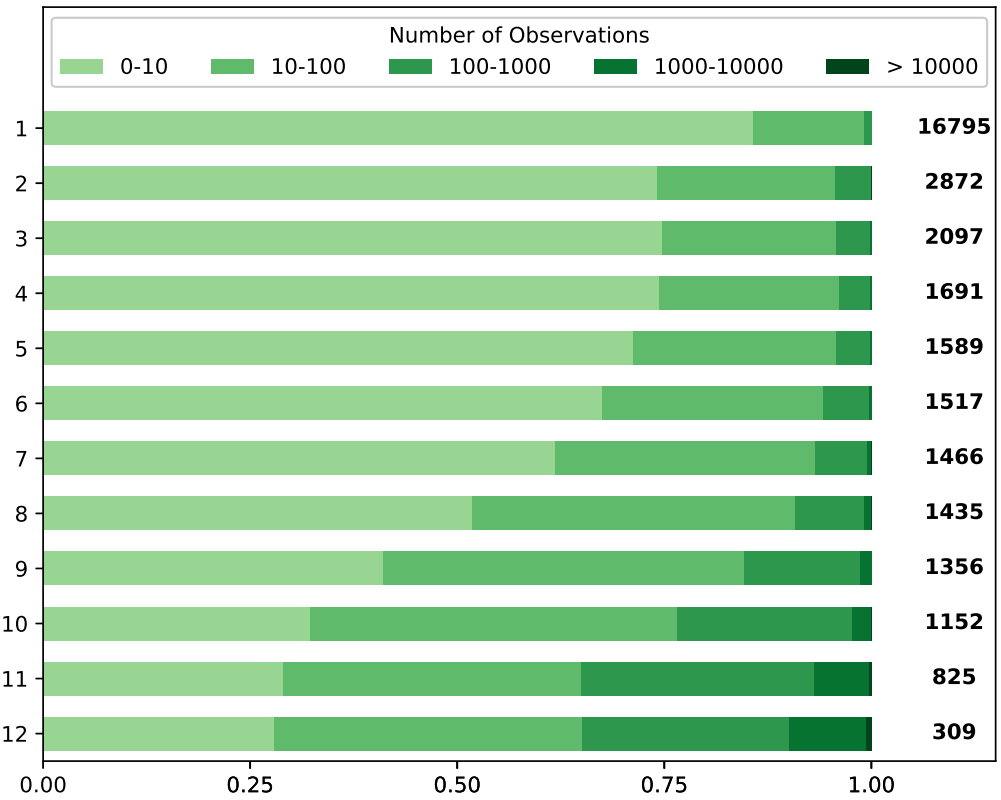


Figure 6.4: Fraction of users with a number of observations — represented with colors as described in the legend — by the number of years since their first participation. It seems users tend to participate more, the longer they are in the network.

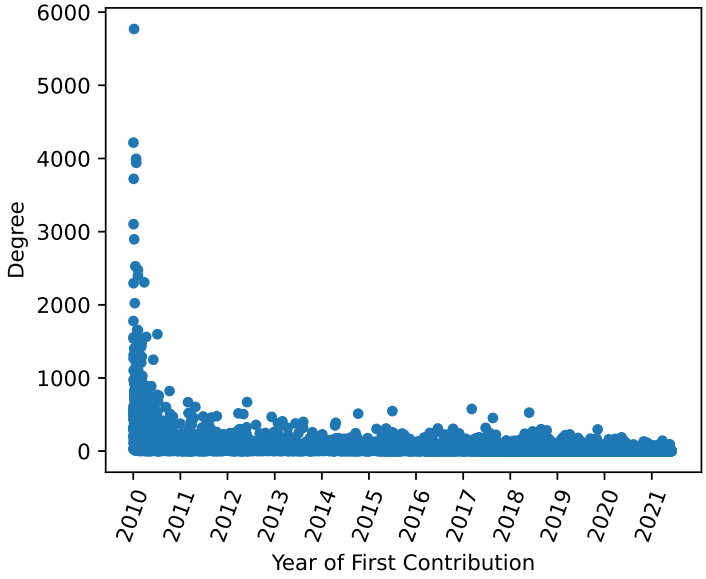


Figure 6.5: Users' degree by the date of their first contribution. Users with the most connections - the hubs - have been in the network for several years.

6.2 Summary

In this chapter we studied the behaviour of users regarding their interaction with the network. In this manner, we divided users into those that create observations, the Triggers, and those that participate in other users' observations, the Validators.

We concluded that, in general, people tend to be attracted to one of the extremities, with a very high number of users choosing to only trigger or validate. However, this tendency seems to be related to the users number of interactions - the higher, the more probable it is for a user to have tried validation. In other words, it is highly unlikely that a high degree user is a Pure Trigger.

Concerning the evolution of users' behaviour, the fraction of users adopting each role highly varies with the number of years they have been in the network. Considering users that have been in the network for one year, there is a high fraction of users that are Triggers. Taking into consideration users that have been in the network more years, the fraction of Triggers decreases, but the percentage of Validators and Hybrids increases.

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 with creating observations.

Chapter 7

Conclusions

This thesis set out to evaluate users' interaction, behaviour, and evolution in the BioDiversity4All platform by studying its participants' underlying collaborative networks.

We started by analyzing the network of users and observations they created or validated. We studied the number and types of users' contributions, creating a typical profile for each participant. We showed that users have a strongly heterogeneous contribution rate: While most users developed or validated few observations, a tiny fraction of users has contributed with over 100 observations (7% of users). Concerning the popularity of observations, the obtained values showed that it is also not homogeneous: while most observations have none or a single identification, others may reach 19 identifications.

Next, we studied the actual cooperation network among users, which we called CoopNet. Our results show that users in the BioDiversity4All form a scale-free network, portraying a well-defined power-law degree distribution, similar to most online social networks, such as Facebook or Twitter, or other large-scale human endeavours such as the WWW the Internet. These results translate into most users barely cooperating with others while a small fraction of users presents a very high participation rate, interacting with thousands of others. This small fraction of users plays a crucial role in the network by keeping the network connected (see Section 4.3).

One of the distinctive features of this network is that it evolves in time, following users preferences and observations. Thus, we also evaluated the time-evolution of this network, including its network properties. We measured its average degree, average shortest path, clustering coefficient and degree exponent values by year. The results obtained showed that the BioDiversity4All 's user collaboration network - CoopNet - keeps its small world and scale-free properties throughout the years, despite significant differences in the number of users, technical transformations occurred in the meantime, or even pandemic outbreaks. This apparent structural invariance in time was also observed in all network measures. This result suggests the existence of time and scale-invariant topological properties in citizen science platforms, and in the BioDiversity4All platform, in particular.

We then focused on the identification of sub-communities and possible dividing factors between users in the BioDiversity4All . First, we identified the existent communities using two different network community partitioning methods - the Louvain and Leiden methods. Next, we determined the taxon

preferences of users in each community. The values obtained showed that each community has a predominant taxon (or taxa) and a distinct pattern concerning its users' taxon preferences. Specifically, we found six communities: Bird; Plant; Insect; Reptiles/Amphibians; Aquatic; and *Fungi*. These results suggest that users connect the most with those that have the same interests and knowledge.

Lastly, we studied the behaviour of users concerning their preferred from of interaction. We classified users into three categories - those that create observations, the Triggers, those that validate observations, the Validators, and those with a mixed profile, the Hybrids. Users that only create or validate observations, we called Pure Triggers and Pure Validators, respectively. We conclude that most users either only create observations (Pure Triggers) or only validate observations (Pure Validators). We also concluded that Pure Triggers are typically associated with low degree nodes, and that high degree nodes portray a mixed profile or are pure validators. Finally, we evaluated the time evolution of users' preferences, showing that most users adopt the role of Trigger in the first year, but, as time passes, they tend to increase their role as Validators. In other words, new users tend to create observations rather than validate; in contrast, users that have been in the network longer tend to validate more while still making novel observations. Furthermore, we analysed user engagement evolution and concluded that users tend to have a more central role in the network the longer they are in the platform, with most hubs being in the platform since 2010.

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 and knowledge. And, (most) users evolve, starting by creating observations but eventually assuming the role of validators. Through extrapolation, these results may prefigure future activity in the BioDiversity4All platform.

We hope to have illustrated how a network science perspective on citizen science may be of use in designing and understanding these platforms. As such, it would be interesting to automatize this type of analysis by creating a visualization platform that would automatically obtain the different plots shown in this thesis from a given database. Namely, the degree distribution and power-law fit; a table showing the various properties of the network by a given frequency of time (e.g. by year as in table 4.2); a graphic showing the existing communities in order of the users' taxon interests; a graphic layout showing the behaviour of users and its evolution.

Bibliography

[1] J. Silvertown. A new dawn for citizen science. *Trends in ecology & evolution*, 24(9):467–471, 2009.

[2] E. Aceves-Bueno, A. S. Adeleye, M. Feraud, Y. Huang, M. Tao, Y. Yang, and S. E. Anderson. The accuracy of citizen science data: a quantitative review. *Bulletin of the Ecological Society of America*, 98(4):278–290, 2017.

[3] G. Newman, A. Wiggins, A. Crall, E. Graham, S. Newman, and K. Crowston. The future of citizen science: emerging technologies and shifting paradigms. *Frontiers in Ecology and the Environment*, 10(6):298–304, 2012.

[4] J.-C. Streito, M. Chartois, É. Pierre, and J.-P. Rossi. Beware the brown marmorated stink bug! *IVES Technical Reviews, vine and wine*, 2020.

[5] J.-L. Justine and L. Winsor. First record of presence of the invasive land flatworm platydemus manokwari (platyhelminthes, geoplanidae) in guadeloupe. 2020.

[6] A. Eveleigh, C. Jennett, S. Lynn, and A. L. Cox. "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*, pages 79–82, 2013.

[7] E. J. Theobald, A. K. Ettinger, H. K. Burgess, L. B. DeBey, N. R. Schmidt, H. E. Froehlich, C. Wagner, J. HilleRisLambers, J. Tewksbury, M. A. Harsch, et al. Global change and local solutions: Tapping the unrealized potential of citizen science for biodiversity research. *Biological Conservation*, 181:236–244, 2015.

[8] V. Devictor, R. J. Whittaker, and C. Beltrame. Beyond scarcity: citizen science programmes as useful tools for conservation biogeography. *Diversity and distributions*, 16(3):354–362, 2010.

[9] *iNaturalist. Accessed 12 May 2020*, . URL `https://www.inaturalist.org`.

[10] B. L. Sullivan, C. L. Wood, M. J. Iliff, R. E. Bonney, D. Fink, and S. Kelling. ebird: A citizen-based bird observation network in the biological sciences. *Biological conservation*, 142(10):2282–2292, 2009.

[11] *The Zooniverse Home page. Accessed 25 January 2021*, . URL `https://www.zooniverse.org`.

[12] *The iNaturalist Stats page., Accessed 25 January 2021"*, . URL https://www.inaturalist.org/stats.

[13] *The eBird About page. Accessed 25 January 2021*, . URL https://ebird.org/about.

[14] C. J. Torney, D. J. Lloyd-Jones, M. Chevallier, D. C. Moyer, H. T. Maliti, M. Mwita, E. M. Kohi, and G. C. Hopcraft. 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. doi: https://doi.org/10.1111/2041-210X.13165. URL https://besjournals.onlinelibrary.wiley.com/doi/abs/10.1111/2041-210X.13165.

[15] *BioDiversity4All Home Page. Accessed 12 May 2020*, . URL https://www.biodiversity4all.org.

[16] *The iNaturalist Network page. Accessed 12 May 2020*, . URL https://www.inaturalist.org/pages/network.

[17] *BioDiversity4All Help page. Accessed 12 May 2020*, . URL https://www.biodiversity4all.org/pages/help.

[18] A.-L. Barabási et al. *Network science*. Cambridge university press, 2016.

[19] M. E. Newman. The structure and function of complex networks. *SIAM review*, 45(2):167–256, 2003.

[20] A. Hagberg, P. Swart, and D. S Chult. Exploring network structure, dynamics, and function using networkx. Technical report, Los Alamos National Lab.(LANL), Los Alamos, NM (United States), 2008.

[21] D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world'networks. *nature*, 393(6684): 440–442, 1998.

[22] L. d. F. Costa, F. A. Rodrigues, G. Travieso, and P. R. Villas Boas. Characterization of complex networks: A survey of measurements. *Advances in physics*, 56(1):167–242, 2007.

[23] S. N. Dorogovtsev and J. F. Mendes. *Evolution of networks: From biological nets to the Internet and WWW*. OUP Oxford, 2013.

[24] R. Albert and A.-L. Barabási. Statistical mechanics of complex networks. *Reviews of modern physics*, 74(1):47, 2002.

[25] A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *science*, 286(5439): 509–512, 1999.

[26] S. Milgram. The small world problem. *Psychology today*, 2(1):60–67, 1967.

[27] D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world'networks. *nature*, 393(6684): 440–442, 1998.

[28] M. E. Newman. The mathematics of networks. *The new palgrave encyclopedia of economics*, 2 (2008):1–12, 2008.

[29] S. Fortunato and C. Castellano. Community structure in graphs. *arXiv preprint arXiv:0712.2716*, 2007.

[30] 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.

[31] M. Girvan and M. E. Newman. Community structure in social and biological networks. *Proceedings of the national academy of sciences*, 99(12):7821–7826, 2002.

[32] V. A. Traag, L. Waltman, and N. J. Van Eck. From louvain to leiden: guaranteeing well-connected communities. *Scientific reports*, 9(1):1–12, 2019.

[33] M. Molloy and B. Reed. A critical point for random graphs with a given degree sequence. *Random structures & algorithms*, 6(2-3):161–180, 1995.

[34] R. Cohen, K. Erez, D. Ben-Avraham, and S. Havlin. Resilience of the internet to random breakdowns. *Physical review letters*, 85(21):4626, 2000.

[35] R. Cohen, K. Erez, D. Ben-Avraham, and S. Havlin. Breakdown of the internet under intentional attack. *Physical review letters*, 86(16):3682, 2001.

[36] B. Bollobás and O. Riordan. Robustness and vulnerability of scale-free random graphs. *Internet Mathematics*, 1(1):1–35, 2004.

[37] P. Tiago. Social context of citizen science projects. In *Analyzing the Role of Citizen Science in Modern Research*, pages 168–191. IGI Global, 2017.

[38] M. Aristeidou, E. Scanlon, and M. Sharples. Weather-it missions: a social network analysis perspective of an online citizen inquiry community. In *Design for teaching and learning in a networked world*, pages 3–16. Springer, 2015.

[39] M. Aristeidou, E. Scanlon, and M. Sharples. Weather-it: evolution of an online community for citizen inquiry. In *Proceedings of the 15th International Conference on Knowledge Technologies and Data-driven Business*, pages 1–8, 2015.

[40] C. Herodotou, M. Aristeidou, G. Miller, H. Ballard, and L. Robinson. What do we know about young volunteers? an exploratory study of participation in zooniverse. *Citizen Science: Theory and Practice*, 5(1), 2020.

[41] E. Guney, J. Menche, M. Vidal, and A.-L. Barábasi. Network-based in silico drug efficacy screening. *Nature communications*, 7(1):1–13, 2016.

[42] F. Cheng, I. A. Kovács, and A.-L. Barabási. Network-based prediction of drug combinations. *Nature communications*, 10(1):1–11, 2019.

[43] I. A. Kovács, K. Luck, K. Spirohn, Y. Wang, C. Pollis, S. Schlabach, W. Bian, D.-K. Kim, N. Kishore, T. Hao, et al. Network-based prediction of protein interactions. *Nature communications*, 10(1):1–8, 2019.

[44] J. Kim and M. Hastak. Social network analysis: Characteristics of online social networks after a disaster. *International Journal of Information Management*, 38(1):86–96, 2018.

[45] R. Pastor-Satorras and A. Vespignani. Epidemic spreading in scale-free networks. *Physical review letters*, 86(14):3200, 2001.

[46] J. Zhang and D. Centola. Social networks and health: new developments in diffusion, online and offline. *Annual Review of Sociology*, 45:91–109, 2019.

[47] M. E. Newman. Scientific collaboration networks. i. network construction and fundamental results. *Physical review E*, 64(1):016131, 2001.

[48] M. E. Newman. Scientific collaboration networks. ii. shortest paths, weighted networks, and centrality. *Physical review E*, 64(1):016132, 2001.

[49] J. Bian, M. Xie, U. Topaloglu, T. Hudson, H. Eswaran, and W. Hogan. Social network analysis of biomedical research collaboration networks in a ctsa institution. *Journal of biomedical informatics*, 52:130–140, 2014.

[50] A.-L. Barabâsi, H. Jeong, Z. Néda, E. Ravasz, A. Schubert, and T. Vicsek. Evolution of the social network of scientific collaborations. *Physica A: Statistical mechanics and its applications*, 311(3-4): 590–614, 2002.

[51] R. Albert, H. Jeong, and A.-L. Barabási. Error and attack tolerance of complex networks. *nature*, 406(6794):378–382, 2000.

[52] F. Morone and H. A. Makse. Influence maximization in complex networks through optimal percolation. *Nature*, 524(7563):65–68, 2015.

[53] F. Morone, B. Min, L. Bo, R. Mari, and H. A. Makse. Collective influence algorithm to find influencers via optimal percolation in massively large social media. *Scientific reports*, 6:30062, 2016.

[54] I. A. Kovács and A.-L. Barabási. Network science: Destruction perfected. *Nature*, 524(7563):38–39, 2015.

[55] O. Serrat. Social network analysis. In *Knowledge solutions*, pages 39–43. Springer, 2017.

[56] J. Alstott, E. Bullmore, and D. Plenz. powerlaw: a python package for analysis of heavy-tailed distributions. *PloS one*, 9(1):e85777, 2014.

[57] A. Clauset, C. R. Shalizi, and M. E. Newman. Power-law distributions in empirical data. *SIAM review*, 51(4):661–703, 2009.

[58] S. N. Dorogovtsev, A. V. Goltsev, and J. F. F. Mendes. Using model on networks with an arbitrary distribution of connections. *Physical Review E*, 66(1):016104, 2002.

[59] L. Ponciano, F. Brasileiro, R. Simpson, and A. Smith. Volunteers' engagement in human computation for astronomy projects. *Computing in Science & Engineering*, 16(6):52–59, 2014.

Appendix A

Further results

Some results were obtained and considered for further analysis.

Community Analysis

Table A.1: Modularity values obtained using both Louvain and Leiden methods

	Newman Girvan	Erdos Renyi	Z Modularity
Louvain	0.26	0.25	0.61
Leiden	0.26	0.25	0.61

Table A.2: Community by *taxon* percentage

Comm.	Size	Insect	Plant	Rept	Aves	Mam	Fungi	Amphi	Mollu	Actin	Arach	Proto	Chrom	Other	Other-Aqua	Total
1	6600	8.1%	76.7%	0.5%	2.2%	1.9%	6.8%	0.3%	0.7%	0.2%	1.2%	0.2%	0.1%	0.6%	0.6%	100.0%
2	3411	62.8%	14.4%	0.9%	2.9%	1.3%	2.4%	0.7%	1.2%	0.2%	9.5%	0.0%	0.0%	2.0%	1.7%	100.0%
3	3187	9.9%	14.6%	1.6%	61.5%	6.1%	1.9%	0.7%	0.5%	0.6%	1.3%	0.0%	0.0%	0.7%	0.7%	100.0%
4	1612	20.8%	17.6%	25.3%	9.6%	5.6%	2.4%	12.3%	1.1%	0.6%	3.0%	0.1%	0.0%	0.9%	0.8%	100.0%
5	1185	8.7%	19.1%	1.3%	4.7%	5.9%	2.5%	0.6%	19.0%	13.7%	2.4%	0.1%	1.1%	4.4%	16.6%	100.0%
6	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%
7	3	16.7%	83.3%	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%
9	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 A.3: *Taxon* percentage by community

Comm	Size	Insect	Plant	Rept	Aves	Mam	Fungi	Amphi	Mollu	Actin	Arach	Proto	Chrom	Other	Other-Aqua
1	6600	15.5%	77.5%	6.2%	6.1%	24.4%	68.2%	6.5%	12.9%	7.0%	15.2%	71.3%	19.6%	20.5%	11.8%
2	3411	62.5%	7.5%	5.5%	4.1%	8.4%	12.3%	9.2%	11.8%	2.6%	61.8%	8.5%	8.2%	35.5%	17.9%
3	3187	9.2%	7.1%	9.5%	81.0%	36.7%	9.0%	8.4%	4.9%	9.2%	8.2%	2.7%	3.9%	10.5%	6.5%
4	1612	9.8%	4.3%	76.0%	6.4%	17.1%	5.9%	73.4%	5.3%	4.3%	9.3%	13.4%	1.9%	7.4%	3.8%
5	1185	3.0%	3.5%	2.8%	2.3%	13.4%	4.6%	2.4%	65.2%	76.8%	5.4%	4.1%	66.4%	26.2%	60.1%
6	3	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
7	3	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
8	2	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
9	2	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Total	16005	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%

Table A.4: Community Structure by *taxon* percentage since 22/03/2020 - the start of confinement due to COVID-19 in Portugal

Comm.	Size	Insect	Plant	Rept	Aves	Mam	Fungi	Amphi	Mollu	Actin	Arach	Proto	Chrom	Other	Other-Aqua	Total
1	3115	10.3%	78.6%	0.5%	2.6%	1.0%	3.4%	0.3%	0.6%	0.2%	1.3%	0.1%	0.1%	0.6%	0.4%	100.0%
2	1963	63.6%	14.0%	0.9%	3.3%	1.0%	2.3%	1.0%	1.2%	0.1%	8.9%	0.1%	0.0%	1.9%	1.6%	100.0%
3	1574	13.3%	14.8%	2.6%	55.8%	5.1%	1.9%	0.9%	0.9%	1.6%	1.4%	0.0%	0.0%	0.7%	1.0%	100.0%
4	1113	23.6%	22.4%	18.3%	10.8%	5.0%	3.5%	9.7%	1.2%	0.5%	3.3%	0.1%	0.0%	0.9%	0.8%	100.0%
5	744	10.3%	22.9%	1.9%	5.1%	7.1%	3.4%	0.5%	17.3%	8.0%	3.4%	0.1%	0.8%	4.3%	14.9%	100.0%
6	413	13.2%	29.3%	1.1%	3.0%	2.4%	40.8%	0.8%	1.5%	1.9%	2.4%	1.1%	0.1%	1.4%	1.2%	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	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%
9	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%
10	2	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%	0.0%	100.0%
11	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	8934	2174	3507	297	1195	249	414	158	205	106	309	11	9	115	184	8934

Table A.5: *Taxon* percentage by community since 22/03/2020 - the start of confinement due to COVID-19 in Portugal

Comm	Size	Insect	Plant	Rept	Aves	Mam	Fungi	Amphi	Mollu	Actin	Arach	Proto	Chrom	Other	Other-Aqua
1	3115	14.8%	69.8%	5.3%	6.8%	12.5%	25.6%	5.9%	9.1%	5.9%	13.1%	27.1%	32.9%	16.3%	6.8%
2	1963	57.4%	7.8%	6.0%	5.4%	7.9%	10.9%	12.4%	11.5%	1.8%	56.5%	17.1%	0.0%	32.5%	17.0%
3	1574	9.6%	6.6%	13.8%	73.5%	32.2%	7.2%	9.0%	6.9%	23.7%	7.1%	0.0%	0.0%	9.6%	8.5%
4	1113	12.1%	7.1%	68.7%	10.1%	22.3%	9.4%	68.3%	6.5%	5.2%	11.9%	9.7%	0.0%	8.7%	4.8%
5	744	3.5%	4.9%	4.8%	3.2%	21.2%	6.1%	2.4%	62.9%	56.0%	8.2%	6.5%	62.8%	27.9%	60.1%
6	413	2.5%	3.5%	1.5%	1.0%	4.0%	40.7%	2.1%	3.0%	7.4%	3.2%	39.6%	4.4%	5.0%	2.7%
7	3	0.0%	0.1%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
8	3	0.0%	0.1%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
9	2	0.0%	0.1%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
10	2	0.1%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
11	2	0.0%	0.1%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Total	8934	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%

Appendix B

Examples

B.1 User Categorization

A user can be determined to be a Validator, a Trigger or a Hybrid, according to the relative difference formula on a user's number of validations and triggers:

$$RelativeDiff(v,t) = \frac{v-t}{\max(v,t)} \tag{B.1}$$

A user can be determined to be:

- A Trigger if: $RelativeDiff(v,t) < -0,1$
- A Validator if: $RelativeDiff(v,t) > 0,1$
- A Hybrid if: $-0.1 \leq RelativeDiff(v,t) \leq 0.1$

For example: Consider a user, John, that completed **13 triggers** and **10 validations**.
Applying the formula:

$$RelativeDiff_{John}(10,13) = \frac{10-13}{\max(10,13)} = -\frac{3}{13} \approx -0,27 < -0.1 \tag{B.2}$$

Therefore, John is a **Trigger** (figure B.1 shows John's location in the behaviour colour spectrum of Figure).

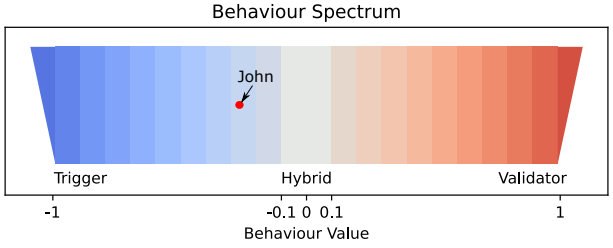


Figure B.1: John in the behaviour colour spectrum of Figure 6.3