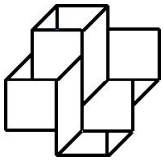

National Laboratory for Scientific Computing

Computational Modeling Graduate Program

New perspectives on analyzing data from biological collections based on social network analytics

Pedro Correia de Siracusa

Petrópolis, RJ - Brazil

June 14, 2018

Pedro Correia de Siracusa

New perspectives on analyzing data from biological collections based on social network analytics

Dissertation submitted to the examining committee in partial fulfillment of the requirements for the degree of Master of Sciences in Computational Modeling.

National Laboratory for Scientific Computing

Computational Modeling Graduate Program

Supervisor: Artur Ziviani

Co-supervisor: Luiz Manoel Rocha Gadelha Júnior

Petrópolis, RJ - Brazil

June 14, 2018

XXXX

Siracusa, Pedro Correia de

New perspectives on analyzing data from biological collections based on social network analytics / Pedro Correia de Siracusa. – Petrópolis, RJ - Brazil, June 14, 2018-

110 p. : il. ; 30 cm.

Orientador(es): Artur Ziviani e Luiz Manoel Rocha Gadelha Júnior

Dissertation (M.Sc.) – National Laboratory for Scientific Computing
Computational Modeling Graduate Program, June 14, 2018.

1. Biodiversity Informatics. 2. Networks. 3. Data Science. I. Ziviani, Artur. II. LNCC/MCTIC. III. Title

CDD: XXX.XXX

Pedro Correia de Siracusa

New perspectives on analyzing data from biological collections based on social network analytics

Dissertation submitted to the examining committee in partial fulfillment of the requirements for the degree of Master of Sciences in Computational Modeling.

Approved by:

Prof. Artur Ziviani, Dr.
(Presidente)

Prof. Fábio André Machado Porto, D.Sc.

Prof. Antonio Mauro Saraiva, D.Sc.

Prof. Marinez Ferreira de Siqueira, D.Sc.

Prof. Eduardo Couto Dalcin, Ph.D.

Petrópolis, RJ - Brazil
June 14, 2018

Dedication

To my wife Adriana and my child Nicole
To my parents Ana and Arquimedes, and sis Clara
and to all my family and friends.
All beloved.

Dedicatória

À minha esposa Adriana e minha bebê Nicole
Aos meus pais Ana e Arquimedes, e irmã Clara
e a toda minha família e amigos.
Todos muito amados.

Acknowledgements

I would like to profoundly thank all who have, somehow, contributed to the execution of this work. In first place, I thank my family for their love, encouragement, emotional and financial support during all stages of my work. Also, I thank all my friends at the Laboratory for the intense knowledge-sharing during these two years. I especially thank my flatmate Fábio Fernandes for the nice company during my stay in Petrópolis and Yasmmin Côrtes for so many coffee-intensive good moments.

I would also like to express my most sincere gratitude to my research supervisors, Dr. Artur Ziviani and Dr. Luiz Gadelha, for their commitment, assistance and patience; all my professors at LNCC, in special to Dr. Fábio Porto, for the research supervision during my first year in the program; and all staff at LNCC, in special to Ana Néri and Roberta Machado for being always so nice and willing to help.

Finally, I thank all hardworking Brazilian taxpayers for supporting science in our country and for my scholarship, granted to me through the National Council for the Scientific and Technological Development (CNPq); Dr. Eduardo Dalcin and Dr. Marinez Siqueira for providing insightful comments along the development of this work; Dr. Carolyn Proença and Dr. Cássia Munhoz for some thoughts in the case study; and the researchers who accepted our invitation for composing the dissertation defense committee.

Gostaria de agradecer imensamente a todos que, de alguma forma, contribuíram para a execução deste trabalho. Primeiramente, agradeço minha família pelo amor, incentivo e apoios emocional e financeiro durante todas as etapas de meu trabalho. Também agradeço a todos meus amigos no Laboratório pela intensa troca de conhecimento durante estes dois anos. Em especial, agradeço a meu colega de apartamento Fábio Fernandes pela excelente companhia durante minha estadia em Petrópolis e Yasmmin Côrtes por todos os bons momentos regados a café.

Também expresso minha mais sincera gratidão aos meus orientadores, Dr. Artur Ziviani e Dr. Luiz Gadelha, pelo comprometimento, assistência e paciência; a todos meus professores no LNCC, em especial ao Dr. Fábio Porto, pela orientação durante meu primeiro ano no programa; e a toda a equipe de apoio no LNCC, em especial à Ana Néri e Roberta Machado, por serem pessoas tão agradáveis e sempre dispostas a ajudar.

Finalmente, agradeço a todos os Brasileiros trabalhadores e pagadores de impostos, pelo apoio à ciência nacional e pela minha bolsa, concedida a mim através do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq); Dr. Eduardo Dalcin e Dra. Marinez Siqueira por terem fornecido comentários valiosos ao longo do desenvolvimento deste trabalho; Dra. Carolyn Proença e Dra. Cássia Munhoz por algumas ideias no estudo de caso; e aos pesquisadores que aceitaram o convite a compor a banca de avaliação desta dissertação.

“Ahhrrrr urrrrghh uhrrrr aaaarrrghh”
(Chewbacca)

Abstract

Biological collections have been historically regarded as fundamental sources of scientific information on biodiversity, supporting a wide range of scientific and management initiatives in the scope of natural resources conservation. As they are typically composed of punctual records of specimens (most of which derived from non-random and opportunistic sampling), biological collection datasets are commonly associated with a variety of biases, which must be characterized and mitigated before data can be consumed.

In this dissertation, we are particularly motivated by taxonomic and collector biases, which can be understood as the effect of particular recording preferences of key collectors on shaping the overall taxonomic composition of biological collections they contribute to. In this context, we propose two network models as the first steps towards a network-based conceptual framework for understanding the formation of biological collections as a result of the composition of collectors' interests and activities. Both models extend the well-established framework of social network analytics, benefiting from a whole set of metrics and algorithms for characterizing network topological features.

Species-Collector Networks (SCNs) model the interests of collectors towards particular species, and are structured by linking collectors to each species they have recorded in biological collection datasets. From complementary perspectives, SCNs allow one to investigate which collectors share common interest for sets of species; and conversely, which species are usually recorded by similar sets of collectors.

Collector CoWorking Networks (CWNs) are a special type of collaboration networks, structured from collaboration ties that are formed between collectors who record specimens together in field. Such collaborative ties are created between pairs of collectors whenever they are both included as collectors in the same record.

Building upon the defined network models, we also present a case study in which we use our models to explore the community of collectors and the taxonomic composition of the University of Brasília herbarium. We describe general topological features of the networks and point out some of the most relevant collectors in the biological collection as well as their taxonomic groups of interest. We also investigate the collaborative behavior of collectors while recording specimens. Finally, we discuss future perspectives for incorporating temporal and geographical dimensions to the models. Moreover, we indicate some possible investigation directions that could possibly benefit from our approach based on social network analytics to model and analyze biological collections.

Keywords: Biodiversity Informatics; Biological Collections; Social Networks; Complex Networks.

Resumo

Coleções biológicas são consideradas fundamentais fontes de informação científica sobre biodiversidade, tendo historicamente suportado uma ampla gama de iniciativas para conservação de recursos naturais. Por serem tipicamente compostas de registros pontuais de espécies (muitos dos quais derivam de amostragem não-aleatória e oportunística), dados de coleções biológicas são comumente associados a uma variedade de vieses, que precisam ser caracterizados e mitigados antes que dados possam ser consumidos.

Nesta dissertação temos como principal motivação os vieses taxonômico e de coletor, que podem ser compreendidos como o efeito de preferências pessoais de coletores-chave na composição taxonômica das coleções com as quais eles contribuem. Neste contexto, propomos dois modelos de redes como um primeiro passo para um modelo conceitual, com o objetivo de compreender a formação de coleções biológicas como resultado da composição dos interesses e atividades de seus coletores. Os modelos extendem o campo bem estabelecido da análise de redes sociais, beneficiando de uma variedade de métricas e algoritmos para a caracterização de aspectos topológicos.

Redes Espécie-Coletor (SCNs) modelam os interesses dos coletores em espécies, e se estruturam por meio de enlaces entre coletores e espécies que eles registram. De forma complementar, SCNs permitem tanto a investigação de coletores compartilhando interesses comuns em conjuntos de espécies; quanto de espécies normalmente coletadas por conjuntos similares de coletores.

Redes Colaborativas de Coletores (CWNs) são um tipo especial de redes de colaboração, estruturadas a partir de enlaces colaborativos que se formam entre coletores que registram espécies em conjunto em campo. Tais relações de colaboração são criadas entre pares de coletores caso ambos tenham sido incluídos como coletores responsáveis pelo mesmo registro.

Com base nos modelos definidos, nós também apresentamos um estudo de caso em que exploramos a comunidae de coletores e a composição taxonômica dos herbário da Universidade de Brasília. Descrevemos aspectos toplógicos gerais das redes e indicamos alguns dos coletores mais relevantes na coleção, bem como grupos taxonômicos de seus respectivos interesses. Nós também investigamos o comportamento colaborativo de coletores durante a coleta de espécimes. Ao final, discutimos perspectivas futuras para a incorporação das dimensões temporal e geográfica nos modelos. Também indicamos algumas possíveis direções de investigação que poderiam se beneficiar de nossa abordagem para a modelagem e análise de coleções biológicas.

Palavras-chave: Informática na Biodiversidade; Coleções Biológicas; Redes Sociais; Redes Complexas.

List of Figures

List of Tables

List of abbreviations and acronyms

SCN	Species-Collector Network
CWN	Collector CoWorking Network
NHC	Natural History Collection
UB	University of Brasília Herbarium
ICN	International Code of Nomenclature for algae, fungi, and plants
BI	Biodiversity Informatics
GBIF	Global Biodiversity Information Facility
TDWG	Taxonomic Databases Working Group
DwC	Darwin Core
DQ	Data Quality

Contents

1 Introduction

As a side effect of the rapid human population growth over the past century, we currently face an alarming scenario of biodiversity crisis, with species being extinct at rates that by far exceed natural background rates (CEBALLOS et al., 2015). Several human activities—most remarkably habitat modification and destruction, the indiscriminate use of fertilizers and pesticides, and the introduction of exotic organism—have been identified as important causes of massive biodiversity loss, besides their direct influence on global climate change (WILCOVE et al., 1998). The high number of species being extinct over a relatively short period suggests the imminence of a new event of mass extinction (also referred to as the sixth extinction), in a magnitude comparable to the previous "big five" mass extinction events: The Ordovician-Silurian, Late Devonian, Permian-Triassic, End Triassic and the Cretaceous-Tertiary, all of which strongly related to the effects of global climatic variations (WAKE; VREDENBURG, 2008). In face of this scenario, understanding how environmental changes—and ultimately human activities—affect natural communities has been a central concern in ecology and biodiversity conservation research.

In this context, **biological collections** stand as invaluable sources of primary biodiversity information, physically storing biologic materials that testify to the existence of living organisms over time and geographic space. Regarded as important natural history repositories, biological collections have been increasingly used for a multitude of ecological and conservationist investigations, including the description of patterns of geographical distribution of organisms and their response to climate change, the selection of areas of high priority for conservation, the construction of red lists of threatened species, and the study of routes of biological invasion, just to cite a few (PYKE; EHRLICH, 2010; NUALART et al., 2017; KEMP, 2015; CHAPMAN, 2005b). As many of these initiatives require intensive use of biodiversity data, typically covering wide geographic areas and long periods of time, they would become impracticable without biological collections, due to the high costs associated with collecting new data in field on demand. Besides, more species have been recently discovered by taxonomists by inspecting unidentified materials at biological collections than by exploring and collecting at new locations (KEMP, 2015). One important limitation, however, is that biological collections provide only a sampled partial view of the actual biological diversity within their actuation regions. Furthermore, applications aiming at investigating wider ecological and biogeographic processes should be able to combine data from multiple biological collections.

Recent efforts towards large-scale digitization of biological collections, associated with a gradual shift in the mindset of data curators towards open-science, are leading many institutions to publish and provide open access to their biodiversity datasets. Data

aggregators, such as GBIF[1], iDigBio[2], and SpeciesLink[3] are also playing a key role in this scenario by providing a centralized and transparent access to primary biodiversity data from many collections worldwide, through Web-based graphical and programmatic interfaces. By facilitating biodiversity researchers to consume data from multiple institutions, such initiatives have boosted the scientific investigation of broader and more complex aspects of biodiversity, which would be otherwise infeasible (JAMES et al., 2018; NEWBOLD et al., 2015). However, simply having access to large amounts of data on species occurrences is not necessarily sufficient for carrying out comprehensive biodiversity studies. Understanding biodiversity patterns often requires the integration of many distinct types of environmental and biological data, coming from diverse sources, with varying levels of complexity and associated with many caveats. Biodiversity research is therefore becoming a *data-intensive science* (KELLING et al., 2009), dealing with the main challenge of how to transform massive amounts of heterogeneous raw data, most of which were not collected for any specific purpose, into valuable knowledge.

Tackling this challenge requires an important analytical paradigm shift in the biodiversity research community, with the adoption of *data-driven approaches* for analyzing biological data, in addition to more traditional, *hypothesis-driven* ones (KELLING et al., 2009). Instead of using data for statistically corroborating or refuting an initial set of hypotheses posed by an investigator (and thus hypothesis-driven), a data-driven approach aims at systematic unraveling hidden patterns from data, eventually leading to insights and to the generation of new domain-specific hypotheses. Moreover, the viability of a data-based endeavor depends on properly dealing with data during multiple stages of its *life cycle*, requiring the wide adoption of guidelines, standards, protocols, and documentation routines. The use of *scientific workflows* has been of great value in this regards, as they allow researchers not only to organize and document each step of their own progress, but also make it reproducible and shareable (KELLING et al., 2009; TALBERT et al., 2013; REICHMAN; JONES; SCHILDHAUER, 2011). Failing to observe and meet the requirements in any stage of the data life cycle can lead to a variety of limitations, hampering the use of biodiversity data for many applications.

According to Michener and Jones (2012), the life cycle of ecological data is composed of eight steps: (*i*) *data management planning*, in which the researcher outlines how data should be collected, stored, and shared; (*ii*) *data collection*, during which one should properly use recording devices and follow protocols in order to avoid the introduction of errors and uncertainties in collected data; (*iii*) *data quality assurance and control*, which involves the definition of standards and mechanisms for preventing and monitoring errors and inconsistencies in datasets; (*iv*) *data description*, which consists of documenting

[1] <https://www.gbif.org>
[2] <https://www.idigbio.org>
[3] <http://splink.cria.org.br>

data with metadata; (*v*) *data preservation*, or the storage of data in a properly curated repository; (*vi*) *data discovery*, which is the process of searching for and gathering relevant data for an intended application; (*vii*) *data integration*, in which data from diverse sources domains should be made structurally compatible; and (*viii*) *data analysis*, the process in which information and knowledge on natural phenomena are extracted from data.

The application of information technology for assisting researchers at each stage of the biodiversity data life cycle has been the main concern of the **Biodiversity Informatics** (BI) community, which has undergone a significant expansion over the last two decades (SOBERÓN; PETERSON, 2004). Notable advances have been achieved in many of the stages listed above, although many still pose important unresolved challenges to be addressed within the next decade (PETERSON; SOBERÓN; KRISHTALKA, 2015). Among those challenges, issues regarding the *data quality* (DQ) and *fitness* of primary biodiversity data for their intended use have been thoroughly explored by the BI community (CHAPMAN, 2005a), leading to the development of many methods and tools for assisting the process of data cleaning. In this context, a conceptual framework for assessing and managing data quality has been recently proposed by Veiga et al. (2017), providing a mechanism for improving the collaborative development and sharing of DQ solutions by the BI community. Data *interoperability*, which encompasses the complexities of discovering and integrating data from multiple heterogeneous sources and disciplines (BISBY, 2000), has also received historical attention, with efforts of groups and organizations towards developing taxonomic backbones (*e.g.* ITIS[4], Species2000[5]), data aggregators (*e.g.* GBIF, SpeciesLink), and data standards and vocabularies (TDWG[6] and Darwin Core[7]).

In this dissertation, we are particularly motivated by the challenge of characterizing *sampling biases* in data, defined as systematic errors that are introduced in data as an effect of not using random sampling designs (DARU et al., 2017; CHRISMAN, 1991). Sampling biases are typically introduced in biodiversity datasets when collectors record specimens in the field in an opportunistic fashion, deploying uneven sampling efforts throughout the studied area and recording preferentially organisms with particular characteristics over others. As observed by Nelson et al. (1990), most collecting activity in the herbarium of National Institute of Amazonian Research (INPA) were, at that time, clustered around previously postulated endemism centers. In addition, collectors consider the accessibility of potential sampling sites while selecting them, and thus locations such as roadsides and the proximities of urban centers are often oversampled (DARU et al., 2017), while others that are more remote remain poorly represented. Sampling biases are in fact one of the main limitations of biological collections, and have been observed to strongly impact the overall

[4] <https://www.itis.gov/>
[5] <http://www.sp2000.org/>
[6] <http://www.tdwg.org/>
[7] <http://rs.tdwg.org/dwc/>

quality of models in case they fail to account for them (NEWBOLD, 2010; ARAÚJO; GUISAN, 2006; KRAMER-SCHADT et al., 2013).

As biological collections are typically composed of a variety of specimen records, which are collected opportunistically by multiple collectors at distinct locations and in distinct contexts (DARU et al., 2017), they provide no accurate representation of the biological diversity within their actuation areas (FUNK et al., 1999). For instance, common species are usually underrepresented in biological collections (NELSON et al., 1990), eventually with fewer representatives than rare species, which are more thoroughly searched by experienced collectors (STEEGE et al., 2011). Also, collectors tend to preferentially sample organisms of their direct interests, especially those that are more conspicuous or charismatic, such as large vertebrates or flowering plants (NEWBOLD, 2010; GRAHAM et al., 2004). As a result, the taxonomic composition and the temporal and geographic coverage of records in biological collections are strongly biased towards the interests, behavior and activity periods of the main collectors who contribute to them. Characterizing bias in such datasets would therefore require a systematic analysis of how the complex arrangements of the perceptions, interests and interactions of collectors shape the overall composition of the collections.

Within this scope, we propose the first step towards a novel modeling approach, based on **social network analysis**, for investigating the assemblage of biological collections as a *social process*, resulting from the collecting activities of collectors and their collaborative interactions. Networks have been used in a wide range of domains for the investigation of complex systems of interacting entities, from studies of the World Wide Web (ALBERT; JEONG; BARABÁSI, 1999) to ecological interaction webs (BASCOMPTE; JORDANO, 2007). However, to the best of our knowledge, network analysis has not yet been applied in BI for investigating the assemblage of biological collections. The most similar study we could find investigates the formation of botanical exchange clubs from the $19th$ and early $20th$ century in Britain and Ireland, in which botanists corresponded with each other by exchanging plant specimens (GROOM; REILLY; HUMPHREY, 2014). Another recent study uses network analytics to investigate the connectivity and roles of many organizations in the BI landscape, in terms of how they exchange information (BINGHAM et al., 2017). Grounded on recent advances in network science theory (BARABÁSI, 2016; NEWMAN, 2010) and social network analytics (BARBIER, 2011; STORK, 2015), in this dissertation, we introduce two classes of *network models* for structuring collaborative relations involving pairs of collectors; and interest relations involving collectors and species.

Species-Collector Networks (SCNs) are a particular type of interest networks, representing the interests of *collectors* towards the *species* they have recorded in field. Interest relationships are directly derived from a species occurrence dataset, and necessarily involve a collector and a species. The strengthness of the ties are given by the number of

times the corresponding collector-species associations are observed in the dataset. Interest relationships are represented in the network model as *edges*, while collectors and species are modeled as *nodes* belonging to distinct sets. A *bipartite constraint* in this model ensures that all edges necessarily connect nodes from distinct sets, avoiding the introduction of semantic inconsistencies in the model (for instance, a collector cannot collect another collector). From the topology of SCNs, collectors can be characterized in terms of their preferred taxonomic groups and, conversely, species can be systematically characterized in terms of which types of collectors are typically interested on recording them. Moreover, a multitude of metrics and algorithms from the *network science* domain can be readily applied for extracting insights from the network structure, such as identifying the most relevant specialist and generalist collectors; species that are widely collected and those which are exclusive of particular groups of collectors; groups of collectors who have similar taxonomic interests (*i.e.*, communities of common interests); and groups of species that best distinguish the interests of collectors.

Collector CoWorking Networks (CWNs) are a particular type of collaboration networks, structured from *collaboration* (or *coworking*) ties between *collectors* who have collected specimens together in field. Collaboration relationships are represented as edges in the network model, each of them involving a pair of collectors (represented as nodes of a single type). As opposed to SCNs, species are not represented in this model. Ties are extracted from a species occurrence dataset by linking, in a pairwise fashion, all collectors who were included as authors for each record. The strength of collaboration ties between a pair of collectors is proportional to the number of times they are observed co-authoring records in the dataset. Our justification for CWN models is that as it happens in many social systems, the behavior and interests of collectors may influence and be influenced by those of colleagues with whom they interact. We consider coworking ties to be good indicatives of the extent to which collectors interact, thus providing the structure for the spread of behaviors and ideas. The relative influence and roles played by collectors can therefore be assessed from their position in the network, and the formation of *coworking groups* from the topology of CWNs.

We demonstrate the practical use of our network models by carrying out a case study, using the species occurrence dataset from the University of Brasília Herbarium (UB), downloaded through the GBIF platform. Before building the network models, we first briefly explore the taxonomic, geographic, and temporal coverages of the records in the dataset; and then perform a cleaning routine, in order to improve the quality of the resulting networks. Once the network models are built, we explore their basic topological features and investigate the formation of communities (interest communities in SCNs and coworking communities in CWNs). We also investigate the relative relevance of collectors in the herbarium, regarding both their taxonomic contributions and their social positions.

Finally, we believe our network models open new perspectives for research in BI, specifically for applications that rely on data from biological collections. With further developments from our work, we expect to provide a mechanism for systematically classifying collectors according to their expertises, their behaviors and their social roles in the collections they contribute to. This could be achieved by using network-based routines for assigning discrete profiles to collectors (*e.g.* experienced *vs.* novice, specialist *vs.* generalist). Another perspective is to enrich species occurrence datasets with contextual information, inferred by observing the composition of collectors associated with each record (and their respective profiles). Moreover, although we have not yet incorporated the temporal and geographical dimensions to the structure of our networks in this work, we believe this would be a fundamental advance, allowing to investigate how collectors interact and which species they record through time and geographic space.

In order to encourage and facilitate others to analyze SCNs and CWNs from other biological collections, we make publicly available a *Python* package, developed during this study. Our package *Caryocar*[8] is built on top of the *NetworkX*[9] package, and provides classes and methods for building SCNs and CWNs from species occurrence datasets.

The remainder of this dissertation is organized as follows. Chapter 2 is an overview of the structure of species occurrence data (which is used for building our network models), with a brief discussion about aspects of data quality that are most relevant for this work. In Chapter 3, we start by reviewing general concepts from network science, as well as some of the most relevant metrics that have been used for characterizing the topology of our resulting networks. Next, we briefly describe the social network analytics framework and exemplify applications of network analysis on the field of biodiversity research. We conclude the chapter by formally describing both SCN and CWN models. Chapter 4 is the case study with the UB herbarium, as mentioned above. We conclude our work in Chapter 5 by pointing out directions for further development and new potential perspectives of applications for our network models.

[8] <https://github.com/pedrosiracusa/caryocar>
[9] <https://networkx.github.io/>

2 Background on Biological Collections

Biological collections are scientific repositories where biological materials, in the form of physical specimens, are systematically deposited and preserved to be used for scientific purposes. Throughout this text we use the term "biological collections"[1] as a synonym for *Natural History Collections (NHC)*, the later being more widely adopted in biodiversity informatics literature. Such biological collections are typically hosted and curated by institutions like herbaria and natural history museums, which provide appropriate physical infrastructure and human resources for ensuring both the long-term preservation of the collections and their accessibility to the scientific community.

In this chapter, we provide an overview of how data in biological collections is structured. Before delving into the characterization of biodiversity data, we first review in Section 2.1 the definitions of some domain-specific terms that will be used throughout this text. We then describe the semantics of species occurrence records in Section 2.2. In Section 2.3, we discuss some aspects regarding the quality and limitations of such data. Finally, in Section 2.4, we present the main data quality requirements for our modeling approach.

2.1 Biodiversity-related terms and definitions

Throughout this work we use definitions from the *International Code of Nomenclature for algae, fungi and plants* (ICN) (MCNEILL, 2012). This document outlines a set of rules and guidelines for scientifically naming and grouping plants, fungi, and algae, consisting of a universally adopted reference by the botanical scientific community. Nomenclature best-practices for other groups of organisms are governed by other (though similar) documents.

Taxonomy. Within the domain of biology, taxonomy is, in a general sense, the science of classification of organisms. Organisms are classified according to their shared characteristics and grouped at distinct levels of specificity (or *taxonomic ranks*) using a hierarchical system, in which groups that are more specific are nested within broader ones. For an analogy with set theory, a taxonomic classification system can be thought as being similar to a hereditary (or pure) set, in that all members in a set are, recursively, also required to be sets.

[1] We occasionally also use the term "museums", for short.

Taxonomic Rank. A taxonomic rank refers to the level of the taxonomic hierarchy at which a a group of organisms is defined. For instance, *Fabaceae* is the scientific name assigned to a large group of economically relevant flowering plants, which is defined at the taxonomic rank of family. We also refer to it as family *Fabaceae*. The most relevant ranks adopted in botany (in descending hierarchical order) are *Kingdom*, *Phylum* (or *Division*), *Class*, *Order*, *Family*, *Genus*, *Species*, as stated in *Art.* 3.1 of ICN.

Taxonomic Resolution. The taxonomic resolution of a biological sample is the rank of the most specific taxonomic determination that has been assigned to it. For instance, if a sample has been determined up to the level of *species*, this rank is also its taxonomic resolution. As taxa relate to each other in a tree-like hierarchical structure (with each child taxon having exactly one parent, while a parent taxon can have one or more children), taxonomic identities of a specimen at ranks higher than its resolution can be directly determined. Although this term is not included in the ICN document, we use this definition throughout this text.

Taxon. A taxon is a taxonomic group of organisms at the level of any rank, which are considered by professional taxonomists to form a *taxonomic unit*. Plural is **taxa**.

Species. Species is one of the taxonomic ranks in which organisms can be classified. It is regarded to be a basic unit of taxonomic classification, although organisms can be further classified in lower-hierarchy taxonomic ranks (*i.e.*, infraspecific ranks). Differently from other ranks, the name of a species is composed using a binomial nomenclature system, composed of the name of the genus followed by a *specific epithet*, *e.g. Caryocar brasiliense*, *Myrcia guianensis*, or *Solanum lycocarpum*. The formal definition is given in *Art. 21* of ICN.

Specimen. When botanists sample organisms in the field, they either collect part of the organism (*e.g.* a branch of a tree), the entire organism (*e.g.* the entire body of a weed), or multiple individuals of the same type (*e.g.* a bunch of identical, very small-sized mosses). Any of these collected biological materials is an evidence of the existence of a particular organism at some place and time, and should be properly deposited in a biological collection for being preserved as a reference. A specimen is defined as one of such evidences, and refers to a punctual observation of a single kind of organism. The formal definition is given in *Art.* 8.2 of ICN. Although a specimen could be classified by a taxonomist as being a representative of a given species, this is not a requirement for it to be included in scientific collections. Although taxonomists classify specimens in a best effort manner (the most taxonomically precise as possible), sometimes only higher ranks can be determined. The highest taxonomic rank at which the specimen could be identified

is known as its **taxonomic resolution**. After properly deposited in a biological collection, each record receives a taxonomic identification that assigns the individual to a **taxon**.

2.2 Species occurrence data

Physical specimens stored in biological collections (also referred to as *vouchers*) are often associated with complementary information, either annotated by the responsible collectors during the collecting act; or annotated at later stages, after the specimen is deposited in the collection (CHAPMAN; SPEERS, 2005). Information from the collection event include the *date, time*, and the *geographic location* where the specimen was collected; the names of the *collectors* who were involved in the collection event; and eventual *field notes* describing contextual remarks, such as weather conditions, habitat features, or the sampling method used. Other crucial piece of information is the *taxonomic identity* of the specimen, which can be determined by the collectors themselves or by professional taxonomists once the biological material is incorporated to the collection (although some materials eventually remain unidentified). The taxonomic identity of a specimen includes not only the taxon name assigned to the sample, but also its nomenclatural status and authorship, the name of the person who has provided the identification, and in some cases, information regarding the certainty of identification. As the taxonomic identity of a specimen can be re-evaluated by specialists several times after the first determination (though it requires that the investigator has access to the physical specimen), a history of determinations for specimens is usually stored in a collection. Vouchered specimens, together with their associated data, is what scientifically testifies a punctual observation of a species by a collector, at some location and at some point in time, and is thus referred to as a *species occurrence* record (Figure 1).

Over the last 40 years, many institutions have adopted digital database management systems—or in the simplest cases, electronic spreadsheets—for improving the management and accessibility of their specimens-related data (SUNDERLAND, 2013). Occurrence-related information is digitally stored in relational databases, while keeping references to the physical vouchers. Each row corresponds to an occurrence, and includes contextual information such as the collectors' names, the geographic coordinates, date and time of the occurrence and the taxonomic identity of the recorded specimen. Some initiatives (*e.g.* Reflora[2]) have also devoted significant effort towards the large-scale digitalization of physical specimens from biological collections, making high-resolution photos of the specimens digitally available for researchers.

One important upside of storing species occurrence information digitally is that institutions can more easily replicate and share their data with others, thus supporting

[2] <http://reflora.jbrj.gov.br>

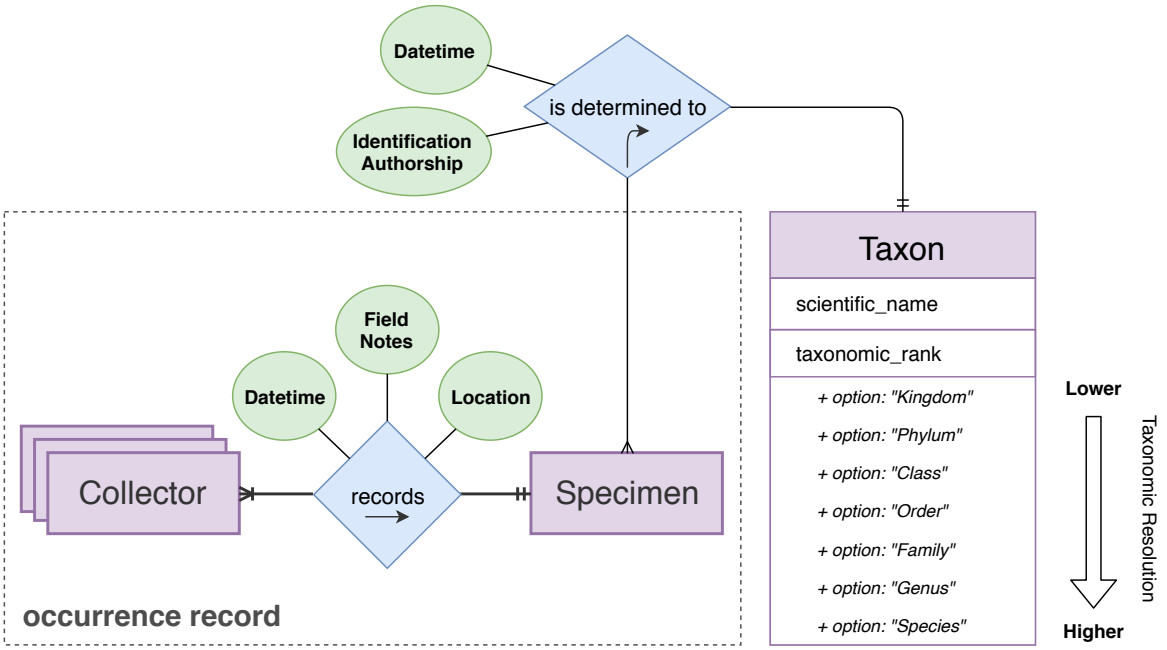

Figure 1 – Entity-relationship diagram illustrating the main features of species occurrence records. The cardinality of relationships is represented using the Crow's foot notation.

wide ranges of applications using biodiversity information (NEWBOLD, 2010). Some data aggregation initiatives, such as the Global Biodiversity Information Facility (GBIF), have provided technological solutions for serving data from multiple collections to researchers through programmatic and visual interfaces (GBIF, 2018). In this context, it becomes relevant to represent data in an standardized and interchangeable format, such that records from distinct collections can be more seamlessly integrated.

Darwin Core (DwC)[3] is a body of standards designed to provide a consistent vocabulary for describing biodiversity-related information and making it exchangeable, accessible, reusable and interoperable (WIECZOREK et al., 2012). An extension of the Dublin Core Metadata Initiative (DCMI)[4], DwC has been specifically created and adopted as a standard for publishing biodiversity data, composed of many terms to designate *classes* of entities and their *properties*. The *Occurrence* class includes terms for representing many aspects of the gathering act *per se*, documenting the existence of organisms at particular places and times. Taxonomic information of organisms is represented under the *Taxon* class. Throughout this chapter, we refer to darwin core terms by prepending '*dwc:*' to names (*e.g.* the *Occurrence* class becomes *dwc:Occurrence* and the *Taxon* class, *dwc:Taxon*). However, the set of terms and definitions provided by DwC does not allow for the representation of semantic structures involving multiple classes of concepts. The development of biodiversity *ontologies*, which consist of domain-specific concepts, data

[3] <http://rs.tdwg.org/dwc/>
[4] <http://dublincore.org>

and entities linked in an hierarchical structure, is a promising step towards overcoming such limitation (WALLS et al., 2014).

Last, it is worth noting that, although biological collections have been traditionally regarded as the main sources of species occurrence data, recent advances in mobile computing technology, associated with the increasing connectivity of electronic devices to the World Wide Web, have leveraged the participation of informal groups of nature observers towards recording and sharing biodiversity data in online platforms (SILVERTOWN, 2009). Many of these communities also collaborate with researchers by sharing large amounts of records of their target organisms, in extents that would be otherwise impracticable to obtain, without enormous financial support. The nature of such records is similar to those from biological collections in that they are punctual observations of specimens in nature, but also have some important distinctions. First, nature observers usually materialize their observations through digital recordings (audio, photo or video) instead of collecting physical biological samples. The absence of physical specimens in many cases makes the identification of the observed organisms inaccurate, thus limiting the use of such data for scientific purposes. In addition, identifications are often provided by a collaborative community of individuals with practical experience, not necessarily professional taxonomists and specialists.

2.3 Limitations of data

Besides the remarkable relevance of biological collections as sources of biodiversity information, they are far from being adequate for investigating every aspect of natural systems. This is partially a consequence of the fact that detailed information is unavailable or very scarce for most known organisms. This scenario, referred to as the *Wallacean Shortfall* (LOMOLINO, 2004), is even more critical for megadiverse countries, such as Brazil, which still remain largely unexplored for many regions and taxonomic groups (SOBERÓN; PETERSON, 2004). The lack of sufficient data for threatened species is even more concerning, as designing efficient programs for their conservation require knowledge on their geographic distribution and ecological requirements. This shortage of data, combined with the non-systematic sampling and insufficient quality, limits the use of data from biological collections for many intended applications, many of which require an intensive amount of data to be available (GUISAN et al., 2007). Failing to account for the inherent limitations of such data while posing and investigating their hypotheses, researchers may obtain erroneous or misleading results, eventually impacting the success of management policies that rely on such information (CHAPMAN; SPEERS, 2005).

2.3.1 Sampling biases

One important aspect that often limits the usability of primary data from biological collections concerns the way in which it is gathered in the field. In general, most species occurrence data composing biological collections derive from exploratory field expeditions, in which organisms are recorded in a non-systematic *observational* fashion by different collectors, using different methods and at distinct circumstances (though records resulting from experimental studies are eventually incorporated in museums as well). As a result, the distribution of the sampling effort in such datasets is uneven and rarely quantified, leading to *sampling biases*.

Building models without accounting for biases in data has been observed to strongly impact their performance, leading to spurious results which can be misinterpreted and, ultimately, lead to wrong decisions. For instance, assessing patterns of species richness from species occurrence datasets has been shown to be particularly challenging due to geographical bias in data (HORTAL; LOBO; JIMENEZ-VALVERDE, 2007; REDDY; DÁVALOS, 2003), as higher diversities tend to be observed at more accessible sites due to higher sampling effort.

As defined by Chrisman (1991), biases are uniform shifts in measured values, resulting from systematic errors that are introduced by some measurement system. They are expressed as unrealistic tendencies in data, and can usually be mitigated with the adoption of random sampling designs. Sampling bias in biodiversity data can be classified into several distinct categories, depending on the aspect of data under investigation (DARU et al., 2017). Here we briefly present four of them, the first two (collector and taxonomic biases) being the most relevant in the scope of our work.

Collector bias. Not all collectors contribute to the same extent to biological collections. In fact, it has been observed that a considerable percentage of records in biological collections are gathered by only a small subset of very productive collectors, while the vast majority of collectors contribute with just a few records each (DARU et al., 2017; BEBBER et al., 2012). This imbalance in the representativity of collectors is what defines the *collector bias* in biological collections. As the overall taxonomic composition of collections—as well as the geographical and temporal distribution of their records—tend to reflect the particular interests and collecting behavior of their most representative collectors, collector bias propagates to other categories of biases in the collection. Therefore, we consider that characterizing collector bias is a fundamental step towards understanding other types of biases.

Taxonomic bias. Not all taxa are quantitatively represented in biological collections in the same proportions as they occur in natural systems. Instead, the taxonomic composition

of collections reflect the interests and collecting behavior of the communities of collectors contributing to them. Moreover, rare species tend to be overrepresented in biological collections, as experienced collectors tend to prioritize collecting them over those which are more common (NELSON et al., 1990). In addition, as collectors usually avoid collecting more than one exemplary of each species at the same place in a given expedition (STEEGE et al., 2011), common species tend to be underrepresented. As the overall taxonomic composition of biological collections tend to reflect the interests and collecting behavior of their most productive collectors, taxonomic bias is intrinsically related to collector bias.

Geographic bias. Collection sites are not randomly selected in geographic space, nor they are all sampled to the same extent. As features of the landscape make some areas more accessible for collection activities than others (HIJMANS et al., 2000), collectors tend to prioritize those to maximize their productivity while minimizing costs. Geographic bias thus arises as a consequence of non-uniform collecting effort in geographic space, and tends to reflect the preferred locations of the most productive collectors. Some regions that are more accessible being thoroughly sampled (such as areas near urban centers, roadsides and margins of rivers); while others that are more inaccessible, such as rainforests, being only poorly or not sampled at all. Geographic bias is also observed at broader scales. A compilation of the representativity of plants in GBIF by Meyer, Weigelt and Kreft (2016) has shown that among the most representative countries and regions are the United States (mainly the west coast), Central America, countries in Europe (including the nordic countries), Australia, Japan, and New Zealand.

Temporal bias. The patterns of the recording activities of collectors are not uniform over time. Instead, collectors often show preferences towards performing field work in periods when they can get more productive, have more financial resources, or can find more organisms of their interests. For instance, wet seasons possibly impact the performance of collectors in the field, and thus it would be natural to observe a relative drop on collecting activity during these periods. Nevertheless, some organisms are better collected during wet periods (such as *Rivulidae* temporal fish), pushing collectors towards performing fieldwork despite unfavorable conditions. Further, the availability of a naturalist for spending time in the field as a collector varies with their age, position, and stage in career. While professional collectors would be available for collecting during weekdays, amateurs or students would be more available in weekends or vacation periods (DARU et al., 2017). The historical of records in biological collections therefore reflects the periods of activities of the most productive collectors.

2.3.2 Data quality

Besides biases, another important limitation in biological collection datasets concerns the *quality* of data. A definition for data quality (DQ) based on its *fitness for the intended use* was first proposed in the context of geographical information systems (CHRISMAN, 1984), and became widely adopted by the BI community. According to this definition, quality is not an absolute attribute of a dataset, but is rather given by its potential to provide users with valuable information, in specific contexts. Assessing quality attributes of data is a fundamental step for any applications that might use it, and requires that users previously delimit the purpose, scope, and requirements of their investigation. Data is considered to be of high quality if it is suitable for supporting a given investigation. Depending on the application, users might need to improve the fitness of the data they have in hand, which is part of the data quality management process.

Loss of quality in biodiversity data can occur during multiple stages of its life cycle (CHAPMAN; SPEERS, 2005), including the moment of the recording event, its preparation before it is incorporated in the collection, its documentation, digitalization, and storage. Some of the most common problems are observed in the taxonomic and geospatial data domains, including the wrong identification of specimens, in part due to using outdated taxonomy; and bad georreferencing of records (SOBERÓN; PETERSON, 2004). In some cases, errors can be manually corrected by referring to supplementary information, such as the field notes of collectors (GRAHAM et al., 2004). However, this approach turns out impractical for large datasets, and methods for systematically assessing quality issues are required.

Dalcin (2005), Veiga, Jr. and Saraiva (2014) have assessed the most common data quality issues in species occurrences datasets, and identified eight recurrent patterns of problems. During our study, we have identified five of them, which we considered to be most relevant: (*i*) **domain value redundancy**, when multiple distinct values in the dataset redundantly represent the same real-world entity; (*ii*) **non-atomic data values**, in case a value semantically contains multiple instances of the fundamental piece of information it should represent (an atomic value is regarded as being indivisible); (*iii*) **inconsistent data values**, in case they do not follow a strict standard, eventually leading to contradictory information; (*iv*) **incorrect data values**, when erroneous information are inserted in the dataset; and (*v*) **missing data value**, in case values at some field are absent (*i.e.* null values).

A framework for systematically assessing and managing the "fitness for use" of biodiversity data from a user-centered perspective was recently proposed by Veiga et al. (2017), and is based on three main components: **DQ Needs**, **DQ Solutions**, and **DQ Report**. While defining DQ needs, users specify their *use case*, comprising the main goals and scope of their investigation; the relevant *information elements* that should be explicitly

represented; and the *dimensions* of data that should be assessed while measuring its quality. Users also specify *criteria* for defining acceptable measurements in each dimension; and activities for improving the suitability of data for the use case (*enhancement*). Within the scope of DQ solutions, users specify methods and implements tools for measuring, validating and enhancing data quality. Finally, DQ reports are produced as documentations of the data quality assessment and management process in different contexts. Users can therefore refer to such reports in order to evaluate whether their data is fit for their intended use, providing a basis for the collaborative development of data quality solutions (which authors refer to as a "Fitness for Use Backbone" (FFUB)).

2.4 Data requirements for network modeling

In this section, we briefly characterize aspects of data quality that are relevant for the network models presented in this dissertation. Our models are constructed based on two **information elements**: the names of collectors and the taxonomic identity of the specimen at each occurrence record. In a dataset following the DwC standards body, collector names and taxonomic identities should be found under the terms *dwc:recordedBy* (class *dwc:Occurrence*) and *dwc:scientificName* (class *dwc:Taxon*), respectively. In SCNs, each record containing a set of collectors and a taxon. Interest ties are extracted from each record, linking each collector to the taxon. More details on the construction process are provided in Section 3.2.2. In CWNs, only the collector field is required. For each record, collaboration ties are created (or reinforced) by combining all collectors in a pairwise fashion. More details of the construction process are provided in Section 3.3.2. Below we describe both information elements and highlight some of the main issues and requirements associated with them.

2.4.1 Collector names

This information element contains the names of all collectors who were responsible for a species occurrence record. In a DwC-compliant dataset, collectors names should be found under the term (or field) *dwc:recordedBy* (within class *dwc:Occurrence*), formatted as a list of collectors names concatenated into a string, using the vertical bar (' | ') as the delimiter. For instance, a record authored by 'M.A. Silva' and 'N.T. Souza' would contain a string with value "M.A. Silva | N.T. Souza". However, storing multiple names into a single string makes values in this field *non-atomic*, thus requiring additional processing for the retrieval of individual names.

We refer to the process of extracting multiple names by splitting the string on a delimiter character as the **name atomization** routine, which is mandatory for our use case. Applying the name atomization process to an entire dataset should be a trivial task

if a consistent and non-conflicting delimiter (though not necessarily the vertical bar) is adopted throughout the entire dataset. However, we suspect this is rarely the case for most species occurrence datasets. Atomization issues arise when the atomization routine is unable to systematically distinguish names in a string, leading to the representation of unrealistic entities in the model.

Using inconsistent (or ignoring) **naming conventions** while registering collectors names in a dataset is also potentially problematic, as it makes it more difficult to systematically interpret names and extract their component parts. Naming conventions are rules used for shortening (or simplifying) names before they are inserted in the database. One common practice is to abbreviate the first and middle names of collectors, while keeping the last name unabbreviated. Under this convention, for instance, the name 'João Souza Silva' would be mapped to 'J.S. Silva'. As collectors are not always aware of the naming convention adopted at the collection, they often include collectors names in their field notes using their own convention or, in the worse case, not using conventions at all. The task to assure that all records are properly formated is therefore assigned to system managers, who must inspect and fix each entry manually before including them in the dataset. As a result, names are eventually registered with typographical errors or being inconsistent with the adopted convention, although some database management systems are able to assist users in this aspect by parsing input names or by suggesting similar names that have already been registered.

Another negative effect of inconsistently formating collector names is the insertion of multiple **name variants** referring to the same real-world entity (domain value redundance). This can also be the result of typos (e.g. souza becoming sousa), or simply omitting parts of the name (e.g. if there are two collectors, A. M. Souza and A. P. Souza, omitting the middle initial makes their names indistinguishable). The **entity-resolution** problem concerns the mapping of name variants that refer to the same real-world entity. Groom, Reilly and Humphrey (2014) came across the same problematic, considerable variations in the formats of names. They used a name cleaning routine in which they merged all variants of names that could be unambiguously mapped to a single person, while excluding those which could not. This problem was also tackled by Silva (2016) by using a data mining methodology, based on association rules analysis, for identifying possible name variants.

Assuming that all records in a dataset are consistent with some naming convention and that they use the same character for delimiting names (*i.e.*, names can be properly atomized), distinct real-world entities may still end up being registered under the same name. These entities, which we refer to as being **homonymous**, are also problematic for our use case, as they are incorrectly represented in the network models as a single entity. In case the naming convention requires the abbreviation of parts of the names, homonymous can be created in the database from two names that are originally distinct. For instance,

'João Souza Silva' and 'Jorge Soares Silva' would be mapped to a homonymous 'J.S. Silva', if both the first and middle names were abbreviated. Some collections include mechanisms in their naming conventions for disambiguating homonymous. For instance, a complementary field containing unique identifiers for collectors can be appended to each record in the dataset, allowing the identification and resolution of homonymous entities. Also, suspect homonymous entities that are already introduced in the dataset can be screened by applying anomaly detection techniques to other collector-related fields. For instance, an entity in the model displaying very improbable patterns of collecting activity (*e.g.* activity peaks temporally spaced by 70 years) is likely to refer to multiple real-world entities.

Finally, some collectors only include their own names in records in which they are first collectors, omitting the names of secondary collectors or, eventually, aggregating them under the expression 'et al.'. We refer to this issue as a **name omission** (within the *completeness* DQ dimension), which strongly impacts as it fails to represent collaborative ties and collecting activities of the collectors who are not listed.

2.4.2 Taxon name

This information element contains the taxonomic identity assigned to the specimen from an occurrence record, being included into the taxonomic data domain. In a DwC-compliant dataset, the taxonomic identities (or taxon) assigned to specimens at the highest resolution as possible are stored under the term *dwc:scientificName*, which is part of the *dwc:Taxon* class. Other relevant elements within this class concern the common name of the taxon (*dwc:vernacularName*); the rank of the taxon (*dwc:taxonRank*); the names of higher-rank taxa within which the the taxon is classified (*dwc:kingdom*, *dwc:phylum*, *dwc:order*, *dwc:family*, *dwc:genus*, *dwc:specificEpithet*); the name of the author of the scientific name of the taxon[5] (*dwc:scientificNameAuthorship*). Two important dimensions that possibly impact the fitness of taxonomic information for our use case concern the *accuracy* and *precision* of identifications; as well as *misspelling errors*.

As exposed by Chapman (2005a), the **accuracy of taxonomic identifications** depends on the level of expertise of the professionals providing them. For instance, a determination of taxonomic identity can be provided by the collector; by a professional taxonomist who is not an expert in the taxonomic group; or by a professional taxonomist who is either a regional or world specialist in that group. Determinations provided by experts are, in general, more credible than those by non-experts, though even experts cannot always guarantee the highest degree of certainty on every identification. Although information about the level of expertise of the determiners is typically not included in datasets from biological collections, many institutions adopt nomenclatural terms to convey

[5] not to be confused with the author of the identification, in *dwc:identifiedBy* (*dwc:Identification* class).

the level of certainty in identifications, for instance 'aff.' (for *'affinis'*, suggesting affinity but not necessarily identity to other taxon) or 'cf.' (for *'confer'*, which is usually placed between the genus and epithet in the name of a species, indicating an uncertainty in the identification due to technical difficulties).

Misspelling errors are also included in the accuracy dimension of taxonomic data quality (VEIGA; JR.; SARAIVA, 2014), and are introduced in the datasets in a variety of ways, including typographic errors (typos), encoding issues, and from the incorrect transcription of names from how they sound (DALCIN, 2005). Taxonomic authority files (files containing accepted taxonomic names) have been widely used for preventing the insertion of new invalid names during data entry; or for checking the validity of names already included in databases (CHAPMAN, 2005a). In addition, Dalcin (2005) has proposed the use of both phonetic and string similarity algorithms for automatically screening potential spelling errors, by matching pairs of slightly different names with high degrees of similarity. Further development towards this direction has led to the development of *Taxamatch* algorithm, having achieved remarkable results (REES, 2014).

The **precision dimension of taxonomic data** is related to the most specific rank at which a specimen can be identified, *i.e.*, its taxonomic resolution (VEIGA; JR.; SARAIVA, 2014). Depending on the modeling requirements, a minimal taxonomic resolution may be necessary for the construction of the networks, making datasets more or less adequate for use. Some applications using our models might require the representation of taxa at higher taxonomic resolutions (*e.g.* at the level of species), whereas for others it could suffice to use lower resolutions, such as the family level. The precision of identifications is also related to the level of expertise of the determiners, although some groups of organisms are naturally more difficult to identify than others. For instance, identifying mammals to the rank of species is far easier than insects. Furthermore, higher levels of expertise are often required for identifying organisms within larger families, which comprise a higher number of species. Therefore, assessing the proportion of records in a dataset that qualify for our network-based approach requires first defining the taxonomic resolution to be adopted during modeling. In that sense, the **completeness of taxonomic information** (the proportion of records that are usable) is intrinsically associated with data precision.

Last, there are cases in which a taxon is referred to by to distinct names, although only one of them is accepted at any given time, according to the *code of nomenclature* adopted. Such names are said to be **synonyms**, and are an example of the domain value redundancy problem in the taxonomic data domain. Similarly to redundancies in collector names, this issue leads to semantic inconsistencies in the network models (more especifically in SCNs), with entities represented more than once. The insertion of synonyms in datasets can also be avoided during the data entry process, by using authority files based on the current code of nomenclature (VEIGA; JR.; SARAIVA, 2014).

3 Network Models

In this chapter, we present and formally describe two classes of network models that were developed during this study. **Species-collector networks** (SCNs) are built based on associations between collectors and the species they have recorded based on their field activities, whereas **collector coworking networks** (CWNs) describe direct collaborative associations between collectors when recording specimens in the field.

Although structurally distinct, SCNs and CWNs provide complementary perspectives on the recording behavior of collectors from a given species occurrence dataset. From SCNs, we retrieve information on which collectors have recorded which species and, conversely, which species were recorded by which collectors. On the other hand, CWNs allow us to investigate which collectors team up with whom during fieldwork, although here species are not represented as entities. As both network models were elaborated based on a framework for social network analytics, we first review some general concepts from the network science and graph theory domains that are used in this study.

3.1 Network science: a theoretical background

Network science refers to a relatively new domain of scientific investigation, which aims at describing emergent properties and patterns from complex systems of interacting entities. Such relational systems are naturally represented as networks, in which interactions are represented as pairwise connections (*links*) between entities (*nodes*) and assume particular semantics depending on the nature of the modeled phenomenon. The rise of this field is strongly associated with recent advances of information technology, which provided scientists with novel tools for collecting, storing, and processing data from many knowledge domains in a more efficient way and at larger scales. Although a variety of networked systems in many disciplines have been studied long before that, technological advances allowed the modeling of real-world systems in much more details, from large volumes data that are often public or easily accessible for the investigators.

In a seminal paper that has inspired many researchers to engage into the field of network science, Albert, Jeong and Barabási (1999) built a model of the World-Wide Web from data collected by a web crawler. Their model represented the web as a set of interconnected documents, where a connection between document a and document b existed if either a contained hyperlinks pointing to b (outgoing links of a) or if a was referred by b through hyperlinks (incoming links of a). By exploring some of the model topological features, they estimated that although the web is composed of a huge number of documents (by that time there were around 8×10^8 documents online), two randomly

chosen documents are, in average, separated by a relatively small number of links (19 in their study).

Although surprising, this finding was in fact consistent with a generative model proposed by Watts and Strogatz (1998). In that paper, authors argued that real-world networked systems are neither completely randomly nor completely regularly connected, and demonstrated that more realistic topologies could be derived by combining features from both extremes. Following this approach, they were able to generate network models which became known as the **small-world networks**, in which nodes are separated from others by very few connections, even for very large systems, while still keeping a relatively high clustering structure. This kind of networks was named after a phenomenon known as *small-world*, which became popular after the experimental work of Milgram (1969) that also led to the expression "six degrees of separation". The concept referred to the experimental finding that two arbitrary individuals in a large population are separated by at most 6 connections, following an acquaintances chain (i.e. a relatively low number of connections compared to the population size).

Besides the small-world network model, other models have been proposed in literature to explain mechanisms by which real world networks grow and acquire topologies with particular properties. A *preferential attachment mechanism* was proposed by Barabási and Albert (1999) as ruling network growth by preferentially connecting new nodes to those in the network that already have many connections. This phenomenon is also referred to as "the rich gets richer" effect and allows the appearance of few heavily connected nodes in the network, called *hubs*. The majority of nodes in the network are, in contrast, very poorly connected, and thus unlikely to be linked with the new edges. Large networks with such characteristics are also known as **scale-free networks**.

Those recently proposed models represent advances if compared with a **random network model** as the $G(N, p)$ model by Erdos and Rényi (1959). In the $G(N, p)$ model, random networks are generated by creating N distinct nodes and each pair of nodes is connected with a uniform probability p, independently of the number of connections a node already has. In random networks, topologies in which few hubs coexist with a massive amount of very poorly connected nodes are then extremely unlikely to be observed.

Empirical results thus indicate that links are not randomly established in many real-world networks. Instead, plenty of system-specific mechanisms occurring in smaller scales are thought to be key in the process of link creation. Network growth is thus not governed by a global mechanisms, but results from the orchestration of many local processes that occur between entities within the system. We refer to such networks, in which non-trivial topological features emerge from the dynamics of subcomponents, as **complex networks**. Uncovering some of such mechanisms and characterizing their effects in network topology compose the core of network science.

Network science has been applied to model networked systems in a variety of knowledge domains, including the Internet, movie actors costarring networks, scientific collaboration networks, networks of human sexual contacts, cellular networks, and ecological networks just to cite a few (ALBERT; BARABÁSI, 2002). Given their relational structure, network models are formally represented as **graphs**.

Next we briefly review some fundamental definitions from *graph theory*. For the purpose of this work, this is not intended to be a thorough review on the subject, but rather a quick overview for not familiarized readers. For a slightly more comprehensive introduction we recommend the reader to refer to *Barabási*'s and *Newman*'s books on network science (BARABÁSI, 2016; NEWMAN, 2010). In the sequence, we briefly describe the social network analytics framework and, finally, we give some examples of applications of network thinking in the biodiversity literature.

3.1.1 Some concepts from graph theory

Intuitively, **graphs** are mathematical structures composed of discrete objects holding pairwise connections to each other. In graph theory terminology, objects are referred to as *vertices* and the connections between them as *edges*. Graph structure provides a natural representation for networked systems for a couple of reasons. First, they allow representing entities and their interactions in a structured way, as a "map of connections" through which information flows. Second, graph theory provides a bunch of well established data structures, metrics, and algorithms for systematically representing, characterizing, and exploring the topologies of complex networks. Finally, the availability of a considerable amount of graph visualization tools and techniques helps analysts to obtain insights from network structures by allowing them to focus on different aspects of the networked complex system depending on their own interests.

A graph is formally defined as a pair $G = (V, E)$, where V is the set of vertices and E is the set of edges that compose the graph (Figure 2). Edges typically involve pairs of vertices and are thus represented as (u, v), for $u, v \in V$. Pairs of vertices that are directly connected by an edge are said to be *adjacent*, and the set of all vertices that are adjacent to a given vertex composes its *neighborhood*. Depending on the nature of the modeled connections, edges can be either classified as *undirected* or *directed*. Undirected edges model symmetric connections, in which case information is allowed to flow equally in both directions. There are situations, however, in which information necessarily flows from a source to a target vertex, thus making connections asymmetric. Those are represented in the graph as directed edges. Graphs strictly composed of directed edges are known as *digraphs* (or *directed graphs*), whilst *undirected graphs* only have undirected edges. In undirected graphs, we should note that edges (u, v) and (v, u) are equivalent. Additionally, connections in a networked system can be modeled as either being all identical or distinct

in terms of their strength or relevance, which is incorporated in the graph as *weight* in the edges. Graphs are also classified regarding the nature of their edges. A graph in which all edges are weighted is known as a *weighted graph*, whilst a *unweighted graph* is exclusively composed of unweighted edges. Given the nature of the networks modeled in the context of this project, we specifically focus on properties of **undirected weighted graphs** for the rest of this section.

Attributes allow representing non-topological features that are relevant in the context of the modeled networked system, and can be optionally assigned to both edges and vertices in a graph. For example, we might want to model a social network in which age and gender information about individuals are relevant for the intended analyses. Such features, which assume distinct values for each individual in the network, are typically stored as attributes of the vertices. Similarly, edge attributes store particularities regarding each individual connection between a pair of vertices. In principle, any numerical edge attribute could be used for weighting edges.

Graphs can be computationally represented using a variety of data structures, the most appropriate one heavily depending on the set of algorithms one expects to run using the graph. Figure 2 shows an illustrative undirected weighted graph and three of its possible representations: an edge list, an adjacency matrix, and an adjacency list, which are explained next.

Edge lists are possibly the simplest and most intuitive graph representation, consisting of a list storing all edges as ordered pairs (u, v), for $u, v \in V$. Given its simplicity, this representation is more appropriate for situations in which the user directly interacts with the data, such as when manually creating or editing graphs without the aid of specialized software. This representation is also compact enough to be adopted for graph storage and data exchange in some situations. Most algorithms, however, perform better by using alternative representations, such as *adjacency lists* or *adjacency matrices*.

Adjacency matrices are a way of representing graphs in matrix form, in which elements store adjacencies between pairs of vertices. An adjacency matrix A is defined as having non-zero entries a_{ij} *if, and only if* $(i, j) \in E$ which, in weighted graphs, correspond to the weight assigned to edge (i, j). Given the equivalence between edges (i, j) and (j, i) in undirected graphs, adjacency matrices in this case are always symmetric, meaning information is stored redundantly in its structure. Besides, the size of an adjacency matrix grows quadratically with the number of vertices in the graph, which would make loading and processing large graphs in computer memory problematic. However, real-world networks have been observed to be naturally sparse, meaning that only a very small percentage of all possible edges do in fact exist. Adjacency matrices can thus be optimized for storage and performance by adopting standard sparse matrix representations, such as compressed row storage (CRS) (SAAD, 2003). An alternative way to represent sparse

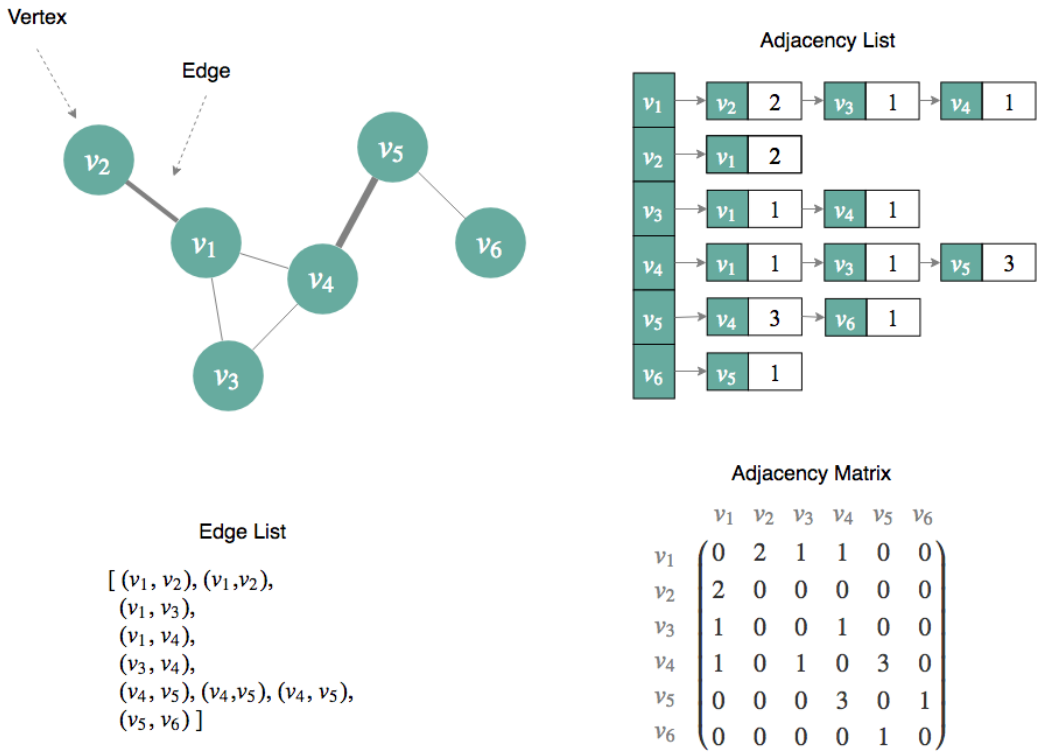

Figure 2 – Undirected weighted graph and three of its possible representations. The graph is composed of six vertices and six edges, with edge thickness representing its weight.

networks is to use the *adjacency list* representation, which in short consists of storing a list of neighbors for every single vertex in the network. The adjacency list illustrated in Figure 2 is structured as an array for which each entry represents a vertex. Moreover, each entry holds a reference for a linked list, containing all neighbors of the vertex. Each element in the list keeps a reference to the next element and stores the weight of the edge connecting it to the vertex in the array.

As previously mentioned, one of the main benefits of adopting graphs for representing networks is the availability of a whole set of well-established metrics and definitions that allow analysts to explore and characterize the topologies of their networked systems. We next review some of the main concepts for network analysis that are useful in this dissertation.

Path. A path can be thought of as a chain of adjacent vertices composing a route through which information may flow in a graph. For undirected graphs, there exists a *path* of length n between a pair of vertices if they are mutually reachable by following a sequence of n edges. A pair of vertices is said to be *connected* if at least one path exists between these vertices. The shortest path between a pair of vertices is defined as the *distance* between them. Finally, the *diameter* of a graph is given by the largest distance

between any pair of vertices, although some authors alternatively define *diameter* of a graph as the average length of all the shortest paths between the pairs of vertices.

Connected component. A connected component is a subset of vertices in the graph in which all included vertices are *connected* to every other one by at least one path. Moreover, this subset is required to be maximal, meaning that all vertices in the network for which the inclusion property holds should also be included in the component. A graph can be composed of many distinct connected components completely disconnected from each other. In the case of a graph with more than one distinct connected component, if one of these components comprises the majority of the nodes in the graph, then it is called the **giant component**.

Clique. A clique is a structure composed of a subset of vertices in which all included vertices are *adjacent* to each other. The set of edges that composes a clique is obtained by computing the pairwise combination of all its n vertices, being the total number of edges in the clique given by the binomial coefficient $\binom{n}{2}$.

Density. Density is one of the simplest metrics for assessing graph connectivity and indicates the proportion of pairs of vertices that are connected by edges. For a graph with n vertices, its density is given by $\frac{2E}{n(n-1)}$, where E is the number of edges in the graph. Possible values for density are thus limited from 0 to 1, where 1 is obtained for a *complete graph* (or a *clique*) and 0 would be obtained for a graph with no edges. This concept is the inverse of graph *sparsity*.

Clustering. Density alone is seldom a sufficient metric for investigating relevant patterns of connectivity in real-world networks, as their topologies usually display *clustering patterns*. *Clusters* are regions in the graph in which vertices are significantly more densely connected than average and more loosely connected with other vertices from the remainder of the graph. Clustering analysis is particularly useful in the context of social network analysis for identifying communities. *Clustering coefficients* are used to characterize how clustered vertices are in a graph, and can be defined as either local or global metrics. The local clustering metric c_i for a given vertex i is calculated by evaluating how close to a clique is a subgraph composed of the vertex i itself and all its k neighbors (which is known as the *ego network* of vertex i). Thus, $c_i = \frac{2L_i}{k_i(k_i-1)}$, where L_i is the number of edges connecting direct neighbors of i and k_i is the degree of i.

To illustrate with an intuitive example within the context of social networks, a local clustering coefficient is an answer to the question "To what extend pairs of my friends are also friends themselves?". The global clustering metric c_Δ, on the other hand, computes the fraction of triplets in the graph that compose cliques of three nodes (*i.e.*, triangles).

A triplet is a set of three nodes connected by either two (open triplet) or three (closed triplet) undirected ties. A triangle therefore includes three closed triplets, one centered on each of the nodes. In short,

$$c_{\Delta} = \frac{3 \times numOfTriangles}{numOfTriples}. \tag{3.1}$$

Transitivity. The transitivity relation for a triple of nodes $\{a, b, c\}$ is observed if the fact that node a is connected to b and b to c implies that a is connected to c. In social networks, transitivity can be interpreted as the phenomenon in which "a friend of my friend is also my friend". Network transitivity is often used as a synonym for *clustering*, and thus the overall transitivity of a network can be quantified through the global clustering coefficient from Equation 3.1.

Degree. The degree (denoted k) of a vertex in a undirected graph is given by the total number of vertices that are adjacent to it or, in other words, the size of its neighborhood. We compute the degree of a particular vertex i from the graph's adjacency matrix A as $k^{(i)} = \sum_j bool(a_{ij})$, where a_{ij} are elements containing the weights of each edge (i, j); and the function $bool(x)$ returns 1 if X is greater than zero; and 0 otherwise. We can similarly compute the *weighted degree* for vertex i as $k_w^{(i)} = \sum_j a_{ij}$. The *average degree* $\langle k \rangle$ is a global property of the network with N vertices, computed by averaging the degree values for each individual vertex i: $\langle k \rangle = \frac{1}{N} \sum_{i=1}^{N} k_i$.

Degree distribution. The probability distribution of the degrees of vertices over a graph is known as its *degree distribution*. From the degree distribution we retrieve the probability p_k that a randomly selected vertex from the graph has degree equal to k. In random graphs, the degree distribution is typically well approximated by a Poisson model, which peaks at $\langle k \rangle$. Thus, vertices with degree equal to $\langle k \rangle$ are most likely to occur, followed by vertices with degree close to the this value, whereas those with very high and very low degrees are very unlikely. As mentioned in the previous section, however, many real-world networks are mainly composed of low-degree nodes, coexisting with few hubs. Such a scenario is better described by a *power law*, such that $p(k) \sim k^{-\alpha}$.

Degree correlation. A positive degree correlation indicates that edges are more likely to connect pairs of either high-degree or low-degree vertices. On the other hand, a negative degree correlation indicates that connections are more likely to exist between high-degree and low-degree vertices. A variety of metrics have been proposed in the literature for assessing degree correlation in a graph (NEWMAN, 2003), being the *Pearson correlation coefficient* one of the most widely used choices. Generally, correlation values range from -1 for maximal negative correlation; to 1 for maximal positive correlation.

Mixing patterns. Mixing patterns describe the influence of node characteristics on the formation of links in networks. In the case associations are more likely to be established between nodes which are similar on some characteristic, the pattern is known as having an *assortative mixing* or, in social sciences, *network homophily*. In contrast, *dissortative mixing* is observed when nodes preferentially connect to other nodes with opposite characteristics. If node characteristics show no significant influence on link formation, we say the network is *neutral*. Degree correlation can be understood as a specific type of mixing pattern, evaluated based on the degree of the vertices.

Centrality. Centrality is a measure of the relative importance of vertices in a network. Different definitions of network centrality typically refer to ranking influential nodes under different perspectives. *Degree centrality* is perhaps the simplest centrality measure, and defines that the relevance of a node depends only on its own degree. Nodes with many connections are thus more influential than those with fewer connections. If we consider, however, that the relevance of a node also depends on the relative importances of their neighbors (and that some neighbors are more influential than others), the *eigenvector centrality* becomes a more realistic metric (BONACICH, 1987). In this case, nodes holding fewer links to more influential nodes might turn out to be more influential than those holding more links to less influential nodes. Other centrality metrics emphasize the importance of other topological aspects instead of node degree. *Closeness centrality* measures the importance of a node by computing its mean distance to all other nodes in the network, and thus nodes are more influential if they can reach their peers more efficiently by taking shorter paths. *Betweeness centrality* identifies nodes which are important strategical intermediators, if they are included in the shortest path between many pairs of nodes in the network.

Bipartite graphs. Bipartite graphs, also known as bigraphs or two-mode graphs, are a special class of graphs composed of two distinct sets of vertices U and V, with the constraint that no vertices within the same set are allowed to be adjacent to each other. Bipartite graphs are defined as triples $B = (U, V, E)$, where E is the set of edges between vertices from U and V. Using set theory terms, U and V are both *disjoint* and *independent* sets. This means vertices must be assigned to exactly one vertex set and, moreover, all edges in E necessarily connect vertices from opposite sets. Such features make bipartite graphs particularly useful for representing interactions that only make sense to exist between entities of different classes. Figure 3(a) presents an example of bipartite graph.

The *biadjacency matrix* (or *incidence matrix*) is the matrix representation of a bipartite graph, being equivalent to adjacency matrices defined for one-mode graphs (those composed of a single vertex set). Differently from one-mode graphs for which the adjacency matrix is necessarily squared, biadjacency matrices are usually rectangular, unless both

vertex sets have the same size.

As most algorithms and metrics in literature are primarily designed for one-mode graphs, it is often convenient to summarize a bipartite graph by directly linking vertices that belong to the same set based on indirect associations they might have. This operation, in which vertices from one set get directly linked if they are intermediated by at least one vertex from the opposite set, is called a *bipartite projection*, as in Figure 3(b). Projections thus compress a bipartite graph into a one-mode graph by only representing vertices from the projected set, while those from the opposite set are omitted. Therefore, bipartite projections into each vertex set provide complementary perspectives on the modeled relationships.

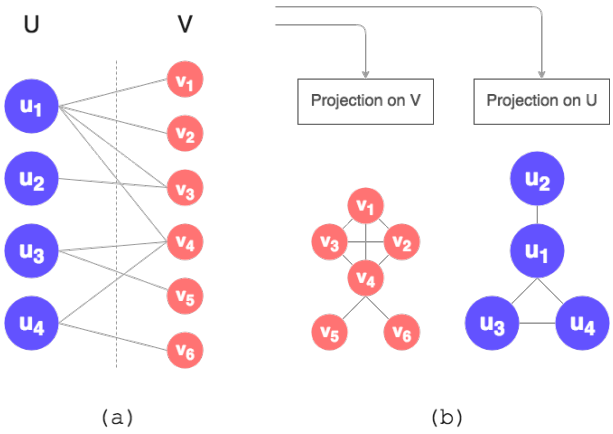

Figure 3 – General aspect of a bipartite graph (a). All vertices in the graph belong to exactly one of U and V vertex sets. In addition, edges are only established between vertices from distinct sets. (b) Bipartite projections. Projections onto each node set are constructed by linking together vertices that are at a length-2 distance in the bipartite graph, while omitting vertices from the opposite set.

3.1.2 Social network analytics

In the context of this work, we refer to the *social network analytics* framework as a set of concepts, methods, and algorithms that can be directly applied, or slightly adapted, for modeling and analyzing social networks in diverse and independent contexts and knowledge domains. In our definition, we refer to *social networks* as systems in which entities (mostly people) interact with each other through some type of *social tie*.

As social networks are themselves a particular class of networks, part of the theoretical foundation used to study them is directly inherited from network theory. Social networks, however, display two important particularities if compared to other types of networked systems (NEWMAN; PARK, 2003). First, entities tend to organize themselves in groups within social systems, such that those who are members of the same group tend

to interact more intensely with themselves than with non-members. This contributes to the formation of community structures, making social networks *highly clustered*. Second, as entities belonging to larger groups tend to interact with many peers—and, conversely, entities belonging to smaller groups tend to interact with fewer peers—, social networks tend to be assortatively mixed.

A variety of social network models within many distinct areas have been proposed in literature. Most authors have primarily focused on obtaining domain-specific insights from statistical and topological properties of network models such as *degree distribution*, *degree correlation*, *mixing patterns*, *node centrality*, and *clustering* (NEWMAN, 2003). Next we describe some of these network models for giving the reader a more solid intuition on how they are structured before we introduce our own network models.

Affiliation networks are a particular type of social network in which individuals are associated with events, groups, or institutions (we will refer to all those as organizations) by being members or participating in them (BORGATTI; HALGIN, 2015). Such networks are considered relevant in the context of social network analytics due to the fact that individuals belonging to the same groups or attending the same events are more likely to become acquainted and develop social ties than those who are not affiliated to common organizations. The most conservative way to structure affiliation networks, so that we keep the complete information on who have affiliated to which organizations, is achieved by representing both individuals and organizations as distinct types of entities in a bipartite graph model.

A classic example of affiliation network is the movie actors network, introduced as an empirical example by Watts and Strogatz (1998) while describing the small-world property in real-world networks. This network was built from the Internet Movie Database (IMDB)[1] by linking actors to movies they have starred in. Actors and movies are thus modeled as distinct entities in a *bipartite* network, such that an actor gets connected to all movies he/she has participated in, whilst a movie gets connected to all actors who have composed its cast. From complementary perspectives, one could distinguish for instance movies that are most similar in terms of their cast composition and, on the other hand, actors that are most related by having starred in the same set of movies. Similarly, Davis, Gardner and Gardner (1941) have studied women social circles from data on their attendance to public social events, and related how their attendance behavior could be influenced by their own social casts.

Another well-known application of affiliation networks are scientific paper authorship networks, in which authors get connected to papers they've authored. In this case, however, the focus has been mostly on deriving co-authorship relationships between paper authors, and in many cases the individual papers which originated the co-authorships

[1] <http://www.imdb.com/>

are not very relevant (NEWMAN, 2004; BORRETT; MOODY; EDELMANN, 2014). A simplified one-mode network, in which only authors are represented as nodes and get directly linked by co-authorship relations, can be derived from the bipartite model by computing its projection onto the author's node set. In the resulting network, authors who have collaborated at least once in paper production get directly connected. Networks where edges assume this semantic are also known as **collaboration networks** (or, alternatively, *coworking networks*) (RAMASCO; DOROGOVTSEV; PASTOR-SATORRAS, 2004). They can either be directly constructed from data or be obtained by projecting bipartite affiliation networks.

Apart from affiliation networks, the bipartite structure is also suitable for modeling other types of social networks. One of those, which we here refer to as **interest networks**, model relations of interest from entities towards objects. Besides their structural similarity to affiliation networks, relationships modeled in interest networks are conceptually distinct from the former, as the fact of two individuals sharing interests does not necessarily provide them differentiated opportunities for developing social ties. Instead, interest networks tie together individuals with objects they are interested in, independently of the social groups or communities they are members of. Thus, instead of social communities, interest networks are more appropriate for revealing *communities of interests*. Interest networks have been used, for instance, to characterize collective music listening habits among users of media streaming platforms (LAMBIOTTE; AUSLOOS, 2005). From a network of listeners interests towards music groups, the authors found evidence that the traditional music genres classification by the music industry is not the best way to categorize listeners in terms of their listening behaviors.

3.1.3 Networks and biodiversity

Network modeling has been widely adopted in the context of biodiversity research, especially for investigating ecological and evolutionary aspects of ecosystems and natural communities. Efforts towards this goal have led to the creation of the field of *network ecology*, which has undergone a noticeable growth over the last few years (BORRETT; MOODY; EDELMANN, 2014). Network ecology has traditionally focused on describing general aspects of the entangled networks by which organisms interact. As ecological interactions are regarded as key processes modeling ecosystems functioning and structure, unraveling their architecture and dynamics is essential for understanding a variety of ecosystem features, such as stability and energy flow. Interaction networks can be broadly classified as *food webs*, *host-parasitoid webs*, or *mutualistic webs* (INGS et al., 2009), being food webs the first ones described in literature, since two classical papers by Lindeman (1942), Odum (1956).

Besides ecological interactions, network thinking has also been applied for modeling

other aspects of natural systems. Patterns of animal movement can be investigated in a structured way, for instance, by means of *movement networks* (JACOBY; FREEMAN, 2016). These networks represent geographical space as a set of discrete and interconnected locations, forming a mesh of possible routes through which animals (or groups of animals) transitate. Links between each pair of locations are weighted according to their geographical connectivity. Animal movement is thus regarded as dynamic processes composed of sequences of discrete movement steps running through the network structure. As the spatial feature is key in this type of network, they are also referred to as *spatial networks* (BASCOMPTE, 2007).

Others have applied network science to investigate biogeographical patterns, such as species co-occurrence. The so called *co-occurrence networks* model species associations in terms of their geographical distributions, such that species which are often observed occurring together in the same set of localities are considered to be strongly connected to each other. Similarly to other networked systems, co-occurrence networks are composed of a majority of the species holding co-occurrence links to very few others, whilst only few species are connected to many others (ARAÚJO et al., 2011). Co-occurrence network analysis has been used for many applications in biodiversity studies, such as for selecting subsets of species to be used as surrogates for the characterization of biological communities (TULLOCH et al., 2016); for assessing the resilience of biotic communities towards climate change (ARAÚJO et al., 2011); and for identifying modularity (clusters of overlapping species ranges) in biological communities from animal-location bipartite networks (THÉBAULT, 2013).

The social network analytics framework has also been applied in some biodiversity studies, though in most cases for modeling animal social behavior (FAUST, 2011). An alternative perspective is to look at communities of biodiversity data producers and consumers, in order to better understand the myriad of contexts in which data is collected, shared and used. Mapping data flow within the community of biodiversity informatics initiatives, for instance, could help prioritizing and improving the coordination of collaborative actions, leading to more effective biodiversity data-based policies (BINGHAM et al., 2017). Also, important scientific communities and gaps could be identified and characterized by exploring collaborative paper authoring networks and scientific topic networks (BORRETT; MOODY; EDELMANN, 2014). Another interesting example of a collaboration network in biodiversity is given in Groom, Reilly and Humphrey (2014), where a correspondence network of $19th$-$20th$ century botanists was structured from digitized data from the British Herbaria. Botanists composing this network corresponded to each other by exchanging specimens, a practice that has led to the formation of exchange clubs. Many aspects regarding the particular ways botanists used to work as well as the roles they assumed could be investigated with the aid of exchange networks.

Finally, a better understanding of the factors and processes influencing the composition of species occurrence datasets would be invaluable for improving data usability, especially for species distribution modeling (DARU et al., 2017). As biological collections are typically composed of an ensemble of opportunistic species occurrence records, each of which having been gathered in a particular context by a different collection team, their datasets do not necessarily reflect the biological diversity from the areas in which the collections are physically located. Rather, they best reflect the interests of their most active and relevant collectors, i.e. those who have contributed to the collection to larger extents.

In our work, we understand the assemblage of a species occurrence dataset as a *social process*, resulting from a multitude of complex interactions between individual collectors, each of them having particular preferences towards recording taxonomic groups and collaborating with other collectors. The network models we propose in the following section help characterizing collectors in terms of their interests and recording behavior, as a proxy for a better understanding on the assemblage of biological collections.

3.2 Species-Collector Networks

In this section, we introduce species-collector networks (SCNs), which provide the structure for modeling associations between collectors and species they record. We first give an overview on the semantics of the modeled relationships and provide a formal definition of SCNs using graph theory. We then describe how such networks can be built from species occurrence datasets. Finally, we define attributes and operations for SCNs that facilitate obtaining domain-specific insights from the network structure.

3.2.1 General description

Species-Collector networks are a particular type of *interest networks* describing relationships of type "**collector** samples **species**" or, conversely, "**species** is sampled by **collector**" (Figure 4). The network is thus composed of collectors holding links to every single species they have ever recorded or alternatively, species holding links to every collector who have ever recorded them. An important semantic aspect of this model worth emphasizing is that here we model collectors recording *species* rather than *specimens*. As exposed elsewhere in this text, the term *species* refers to a grouping of individuals (or *specimens*) which share a set of features and are thus considered to be taxonomically equivalent at that level. Each occurrence record that is used to build the network includes a single specimen, which is a representative of a species. Thus, while collectors are represented in the network at the individual level (each collector is a person), species are instead represented as entities comprising groups of individuals. Nevertheless, each species is

uniquely represented as a node in the network.

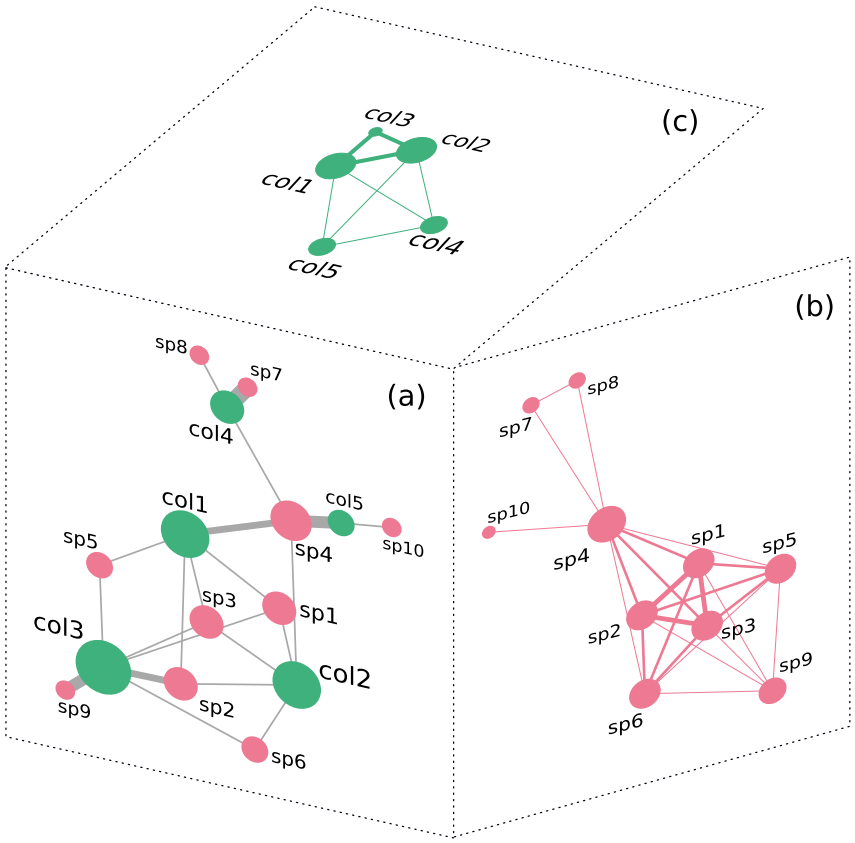

Figure 4 – Multiple perspectives of a Species-Collector Network (SCN). (a) Unprojected bipartite network, where collectors (green nodes) are linked to the species (red nodes) they have recorded. The total number of records of a given species by some collector is reflected in the strength of their link. (b) SCN projection onto the species set. Species are linked together if they have been collected by common collectors. The strength of links between two species is proportional to the number of collectors they share. (c) SCN projection onto the set of collectors. Collectors are linked together if they have recorded species in common. The strength of links between two collectors is proportional to the number of species they share. Link strength for both projections were obtained using the *simple weighting* rule (Eq. 3.2), and are graphically displayed as edges thickness. The sizes of collector and species nodes reflect their degrees, in each perspective.

As collectors and species refer to distinct entities in our system, we represent them in our network as two disjoint node sets. Moreover, as in our case the relationships of interest can only possibly exist between collectors and species, we impose an additional constraint that all edges in the network must necessarily connect nodes from distinct sets. This matches the description of a bipartite network

$$SCN = (S_{col}, S_{sp}, E),$$

where $S_{col} = \{u_1, u_2, ..., u_n\}$ is the node set representing the collectors group; $S_{sp} = \{v_1, v_2, ..., v_m\}$ is the node set representing the species group; and E is the set of undirected

edges between members of S_{col} and S_{sp}. The bipartite graph can also be represented as a rectangular biadjacency matrix $A^{n \times m}$ for which $a_{ij} \neq 0$ *iff* $(u_i, v_j) \in E$. Non-zero a_{ij} elements represent the edges (u_i, v_j) in the network, as described below. For a general overview of bipartite graphs the reader should refer to Section 3.1.1.

3.2.2 SCN model construction from data

A SCN model is built from a species occurrence dataset by using basically two fields. The first is the collectors field, containing the names of all collectors that were responsible for the record; and the second one is the species field, storing the species identity assigned to the specimen in the record. Following terms on the Darwin Core standard,[2] we should expect to find the names of the collectors in the *recordedBy* field; and the species name in a field named *species*. As not every biological collection dataset uses Darwin Core standards though, these fields might be occasionally found under different names.

The network is built up from the dataset in an iterative process, in which weighted edges linking collectors to species are structured from rows in the dataset. For each new record containing n collectors, n links connecting each individual collector to the recorded species are created (or strengthened, in case they already exist). In the end of the process, the strength of each link is equivalent to the number of times each species-collector association appears in the original dataset. The more often a given collector records a particular species, the stronger becomes the link between them. Moreover, as at least one additional link is necessarily either created or reinforced for each new row, the construction process guarantees that no disconnected species or collector nodes can possibly exist in a species-collector network. For a concrete example, the dataset used for creating Figure 4 is given in Table 1.

We keep the record of the number of times each link occurs in the network by assigning a *count* attribute to them, which is initially set to 1 and is increased by one every time a new occurrence of the link is observed. Link strength is proportional to this attribute, and is graphically represented by the edge thickness, as it can be observed in Figure 4. Edges' *count* values are stored in the biadjacency matrix A, and thus the value of element a_{ij} is the number of times the edge (u_i, v_j) occurs. The biadjacency matrix for the example SCN network in Figure 4 is thus

$$A = \begin{array}{c} \\ col1 \\ col2 \\ col3 \\ col4 \\ col5 \end{array} \begin{array}{c} \begin{array}{cccccccccc} sp1 & sp2 & sp3 & sp4 & sp5 & sp6 & sp7 & sp8 & sp9 & sp10 \end{array} \\ \left[\begin{array}{cccccccccc} 1 & 1 & 1 & 2 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 2 & 1 & 0 & 1 & 1 & 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 4 & 1 & 0 & 0 \\ 0 & 0 & 0 & 3 & 0 & 0 & 0 & 0 & 0 & 1 \end{array} \right] \end{array}.$$

[2] <http://rs.tdwg.org/dwc/terms/>

Table 1 – Species occurrence dataset from which the SCN model in Figure 4 was built.

id	recordedBy	species
0	col1; col2; col3	sp1
1	col1; col2; col3	sp2
2	col1; col2; col3	sp3
3	col1; col2	sp4
4	col1	sp4
5	col1; col3	sp5
6	col3; col2	sp6
7	col3	sp2
8	col4	sp4
9	col4	sp7
10	col4	sp7
11	col4	sp7
12	col4	sp7
13	col4	sp8
14	col3	sp9
15	col3	sp9
16	col3	sp9
17	col5	sp4
18	col5	sp4
19	col5	sp4
20	col5	sp10

An homonymous attribute is also set to graph nodes, which is increased whenever a new link involving the node is either added or strengthened. As a result, the node's *count* attribute keeps a record of how many times a given species or collector occurs in the dataset. The reader should note, however, that *count* attributes for nodes and edges are conceptually distinct and are not to be confused.

3.2.3 SCN definitions

Given an overall description on the structure and semantics of the SCN model, we now define a set of attributes and operations that provide higher-level abstractions for dealing with the system here modeled. By using such field-domain abstractions we improve the data exploration process, eventually making it more insightful for the analyst. We introduce both the *species bag* and the *quorum vector* as specific attributes of collectors and species, respectively. Moreover, we define the process of taxonomy aggregation as a model summarization routine for grouping together species nodes into higher-rank taxa.

Species bag. The entire set and counts of species a collector has recorded in a dataset, which can be thought as a collector's species signature, composes his/her *species bag*. This

attribute is therefore exclusively derivable for collector nodes. As species bags are directly obtained as row-vectors of the graph's biadjacency matrix, they are a convenient structure for comparing collectors in terms of the composition of their records, i.e. their respective species bags. For that task, a high variety of well-known distance algorithms for vectors in literature can be readily applied. The species bag σ_{u_i} for collector u_i is thus defined as

$$\sigma_{u_i} = \left[a_{i1}, a_{i2}, ..., a_{im}\right],$$

where m is the length of the species set and each a_{ij} is the total number of records of species v_j by collector u_i. The sum of all elements in a collector's species bag, which is equivalent to the vector's *L1 norm* $||\sigma_{u_i}||_1$, corresponds to the total number of records for that collector u_i.

Quorum. The entire set and counts of collectors who have recorded a particular species in a dataset comprise its *quorum*, an exclusive attribute of species nodes. This concept can be thought as the inverse of a species bag, being the collector-based signature of a species. The quorum vector ι_{v_j} of a species v_j is directly obtained from the graph's biadjacency matrix as the j^{th} column-vector, i.e.,

$$\iota_{v_j} = \left[a_{1j}, a_{2j}, ..., a_{nj}\right],$$

where n is the length of the set of collectors and each element a_{ij} is the total number of times collector u_i has recorded species v_j. The total number of occurrences of species v_j in the entire dataset can be obtained as the sum of all elements in its quorum vector $||\iota_{v_j}||_1$.

Taxonomic aggregation and resolution. In some contexts, it might be desired to simplify SCNs by grouping species nodes into higher taxonomic ranks (or levels), such as a *genus* or a *family*. This process is defined as a *taxonomic aggregation*, and is performed by (i) obtaining a grouping of species using some taxonomic rank; (ii) obtaining quorum vectors for each species; (iii) summing up quorum vectors for all species in each group; and (iv) building a new SCN model, aggregated on taxonomic rank T. The *taxonomic resolution* of a SCN is thus the taxonomic rank at which species are aggregated in the model. For the sake of model interpretability, all nodes in S_{sp} must necessarily be taxa belonging to the same taxonomic rank as the taxonomic resolution adopted for the model.

For a more formal description, let $G_T = \{g_1, g_2, ..., g_n\}$ denote a taxonomic grouping at rank T, containing a set of n rank-T taxa. In addition, let each taxon $g_i \in G_T$ itself be a set of nodes $S_{sp}^{(i)} \subseteq S_{sp}$, with the conditions that there are no empty $S_{sp}^{(i)}$ and that every node $v \in S_{sp}$ is a member of exactly one set from $\{S_{sp}^{(1)}, S_{sp}^{(2)}, ..., S_{sp}^{(n)}\}$. Such a grouping rule makes G_T a *set partition* of S_{sp}, and thus the entire set S_{sp} can be recreated by simply

computing the union of elements in G_T. This guarantees that no entities are duplicated or eliminated on aggregations using it.

We then use grouping G_T for obtaining quorum vectors for each of its taxa $g_i \in G_T$, which will be represented as nodes in the new aggregated graph. Quorum vectors are computed as $\iota_{g_i} := \sum_j \iota_{v_j}$ for $v_j \in S_{sp}^{(i)}$. Finally, the rank-T aggregated graph $SCN_T = (G_T, S_{col}, E)$ is created from a biadjacency matrix, which is constructed by stacking quorum vectors for each taxon g_i as row-vectors. The set of collector nodes remain the same in the aggregated graph.

3.2.4 SCN projections

Bipartite projections on each one of the SCN's node sets allows one to investigate indirect associations two entities from the same class might have with each another, as intermediated by a third entity from the opposite class. Figure 4 illustrates projections of a SCN onto the species set in Figure 4(b) and onto the collector set in Figure 4(c). Overall, each projection gives us complementary perspectives of transitive relationships in the SCN, either from the point of view of the collectors or of the species.

From a *species-centric perspective* (Figure 4(b)), connections are formed between species that have been recorded by at least one collector in common, with link strength being proportional to the number of different collectors they share. Although collectors are used during projection for determining the existence of links between species they are not represented as nodes in this projection. In general, strongly connected species can be interpreted as being both included in the species bags of many collectors, whereas weakly connected or isolated species are seldom or never recorded by the same collectors. The second perspective (Figure 4(c)) is *collector-centric*, i.e., only collectors are represented as nodes whilst species are omitted. Analogously to the species-centric perspective, here collectors are linked together if they have recorded at least one species in common, with link strength depending on the number of commonly recorded species between them. From this perspective, we could identify collectors having similar recording profiles.

As previously discussed, projections are a mechanism for summarizing bipartite into more convenient one-mode graphs, where only one class of entity is represented. Projections, however, come with the cost of information loss, as any relationships or attributes of nodes from the omitted set are not represented in the projection (BORGATTI; EVERETT, 1997). Moreover, relevant associations between entities eventually become obfuscated by others of lower relevance, since projections tend to generate graphs that are much denser than the original bipartite model (LAMBIOTTE; AUSLOOS, 2005). Choosing an appropriate weighting rule for the aspects one wants to investigate is thus crucial for separating relevant from less-relevant associations, so that the latter ones can be subsequently removed by applying weighting filters. In the following, we first describe the simplest weighting rule

with its limitations and, in the sequence, we discuss some alternative rules for overcoming such limitations.

Simple weighting. This rule assigns weights to links between pairs of collectors (u_s and u_t) or species (v_s and v_t) by simply counting the total number of species the pair of collectors share on their species bags or the total number of collectors the pair of species share in their quorum vectors. The simple weighting rule is thus mathematically defined as:

$$\begin{aligned} w_{(u_s,u_t)} &= \sum_{j=1}^{m} \delta(\sigma_{u_s}^{(j)}, \sigma_{u_t}^{(j)}), \text{ for the projection onto } S_{col}; \\ w_{(v_s,v_t)} &= \sum_{i=1}^{n} \delta(\iota_{v_s}^{(i)}, \iota_{v_t}^{(i)}), \text{ for the projection onto } S_{sp}, \end{aligned} \tag{3.2}$$

where $n = |S_{col}|$, $m = |S_{sp}|$, $\sigma^{(i)}$ and $\iota^{(i)}$ are the i^{th} element of a species bag and a quorum vector, respectively; and $\delta(u, v) = 1$ if both u and v are non-zero and 0 otherwise.

In order to obtain the weights for every pair of projected nodes more efficiently we can use a vectorized implementation of this rule. First we derive a $n \times m$ logic matrix A_{bool} from the SCN biadjacency matrix A by simply replacing its non-zero elements by ones. Then a $n \times n$ adjacency matrix with edges weights for the S_{col} projection is obtained by calculating the dot product $A_{bool}A_{bool}^T$. Conversely, for the S_{sp} projection, the $m \times m$ weight matrix is obtained by calculating $A_{bool}^T A_{bool}$.

The simple weighting rule has, however, an important limitation when applied to SCNs. This limitation arises from the fact that the weight assigned to edges linking pairs of nodes in the projection only reflects the number of distinct intermediate neighbors from the complementary set they shared in the non-projected graph. The number of times each species is recorded by each collector is therefore ignored while computing the strength of links in the projections, underestimating the importance of recurrent relationships. Consequently this weighting rule tends to make very prolific and generalist collectors or very attractive species strongly connected to many others in a disproportional way, as an effect of their high degrees in the non-projected model. The opposite happens in the case of specialized nodes, which typically hold fewer—although recurrent—distinct links to their neighbors. In order to mitigate these effects two alternative weighting rules are proposed next.

Average weighting. This rule is a slight modification of the simple weighting rule since it also considers the total number of times entities interact through each neighbor-intermediated path in the non-projected network. The rule is expressed using the same equations from (3.2), but changing the δ function to

$$\delta(u,v) = \begin{cases} \frac{u+v}{2} & \text{if both } u \text{ and } v \text{ are non-zero,} \\ 0 & \text{otherwise.} \end{cases}$$

In case every distinct path in the non-projected SCN only occurs once, then both simple and average weighting rules lead to the same result. This modified rule enhances the effect of recurring edges from the non-projected graph on computing edge weights in the projection, thus reducing weighting asymmetries from generalist and specialized nodes.

Nevertheless, the average weighting still has the drawback that nodes with high degrees in the non-projected graph tend to become much more strongly connected with themselves in the projection than average-degree ones, simply because they have many more connections than average. Additionally, without a superior limit for the δ function it turns out to be hard to determine a proper threshold when filtering relevant from non-relevant edges. The next weighting rule is designed to reduce the effects of node degrees on their edges' weights, outputting values which are bounded to the $[0,1]$ interval.

Similarity-based weighting. This weighting rule uses a similarity (or correlation) matrix that is computed for each projection of the SCN. Edges' weights are given by the similarity between their nodes. The similarity matrices for the projections on collectors and on species are constructed by computing the *cosine similarity* of species bags and quorum vectors for each pair of nodes in the respective projection. The *species bag similarity* for collectors u_s and u_t; and the *quorum vector similarity* for species v_s and v_t are then defined as

$$\begin{aligned} sim(\sigma_{u_s}, \sigma_{u_t}) \equiv \cos\theta_{u_s,u_t} &= \frac{\sigma_{u_s} \cdot \sigma_{u_t}}{||\sigma_{u_s}||_2||\sigma_{u_t}||_2}, \\ sim(\iota_{v_s}, \iota_{v_t}) \equiv \cos\theta_{v_s,v_t} &= \frac{\iota_{v_s} \cdot \iota_{v_t}}{||\iota_{v_s}||_2||\iota_{v_t}||_2}. \end{aligned} \tag{3.3}$$

Therefore, each element in the similarity matrix holds the edge weight for a pair of nodes, with a value ranging within the interval $[0,1]$. Edge weights are zero-valued if no direct link exists between two nodes, whereas nodes linked by edges with a weight of 1 have identical species bags or quorum vectors. Intermediate values reflect the cosine similarity measure obtained for each node pair. As this rule outputs weight values that are ranged in a known interval, filtering less relevant links becomes much more straightforward. Depending on the aspects regarding the species-collector system an investigator might be interested in, a filtering threshold ϕ can be set based on the minimum similarity value he/she considers acceptable, such that only the most relevant relationships for that particular analysis are kept.

3.3 Collector Coworking Networks

In this section, we describe Collector Coworking Networks (CWNs) that model collaborative associations between collectors. We first define CWNs using graph theory and then we describe how such models can be structured from species occurrence datasets.

3.3.1 General description

Collector Coworking networks are a particular instance of *collaboration networks* describing coauthoring relationships between collectors from species occurrence records (see Figure 5). We consider two collectors to be coauthors in a given record if they are both included in the collector field for that record. The collector field holds a list of the names of the collectors who have authored each record in the dataset. This field is equivalent to the *recordedBy* field in a dataset following the terms of *Darwin Core* standards (check an example in Table 2). We refer to each distinct list of collectors in this context as a **team**.

As opposed to SCNs, which basically describe the interests of collectors towards species, relationships in CWNs are directly formed between collectors who have effectively worked collaboratively in the field. These collaborative relationships are semantically described as "**collector** records specimen with **collector**". Each individual species occurrence record with at least two collectors (i.e., team size greater than 1) is thus considered a distinct collaboration act, originating new pairwise connections between all the involved collectors. For records with team size equal to 1, which we refer to as non-collaborative records, no connections are created.

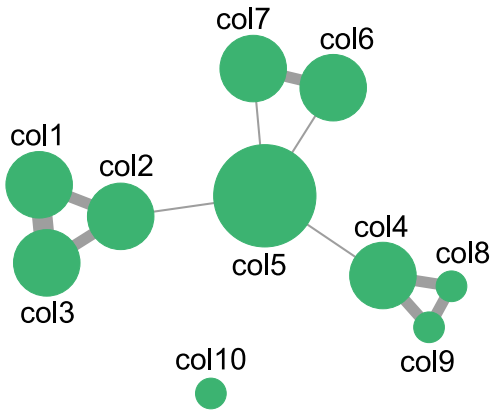

Figure 5 – General aspect of a Collector CoWorking Network (CWN). Coauthoring relationships between collectors (nodes) are structured as edges in the graph, graphically represented as gray lines. Nodes are sized according to the absolute number of records collectors have authored; and edge weight represents the strengtheness of collaborative ties between two collectors.

Differently from SCNs, where two classes of entities are represented in the graph as disjoint node sets with the bipartite constraint, CWNs exclusively model direct relationships

between entities from a single class (collectors), with the only connectivity restriction that a collector should not hold collaborative ties to itself. The model is thus formally described as an unipartite (or one-mode) undirected graph

$$CWN = (S, E),$$

where $S = \{u_1, u_2, ..., u_n\}$ is the graph's node set of collectors and E is the set of undirected edges linking members of S.

Analogously to SCNs, both nodes and edges hold homonymous, but conceptually distinct, *count* attributes. The node's count attribute stores the total number of records—including non-collaborative ones—a collector has authored, whereas the edge's count attribute stores the total number of times an association between two collectors was observed in the dataset. Thus, although each node and edge respectively are uniquely included in S and E, their recurrence patterns are registered in the model as attributes. Another important edge attribute is the *species list*. Although species associated with each occurrence record are not represented as entities in this model, edges can optionally keep a list of species that are shared by two collectors through that link, which is stored in this attribute.

Weights are assigned to edges in the CWN as a measure of their overall relevance in the network structure. Edges with higher weight values represent stronger collaborative ties between collectors, pointing out the main groups of collectors who are most willing to collaborate. The simplest rule is to set the edge weight as the total number of occurrences of the tie it represents in the dataset. However, as pointed out by other authors studying social networks (NEWMAN, 2001a), in reality not all collaboration acts should contribute the same way for a collector's network. Collectors tend to hold weaker collaborative ties with each other when they collaborate in larger teams than when they collaborate in smaller teams.

The **hyperbolic weighting** rule accounts for this fact, while also considering the total number of collaborations between two collectors as a factor contributing to the strength of their common link. According to this rule, not every new occurrence of the link increases the edge weight equally. The contribution of each new link depends on the number of the collectors $n^{(k)}$ included in record k or, in other words, the team size. This rule follows a hyperbolic growth function

$$w_{(i,j)} = \sum_k \frac{\delta_i^{(k)} \delta_j^{(k)}}{(n^{(k)} - 1)} \ , \tag{3.4}$$

where $\delta_u^{(k)} = 1$ if collector u is in record k and 0 otherwise. As the hyperbolic function above has singularity at 1, it gets ill-defined for records with only one collector. Therefore, in this case, only records with two or more collectors are used to compute edge weights.

The maximum weight contribution of 1 is assigned to records with two collectors, whereas records with larger cliques yield smaller contributions.

Relationships in the CWN graph can be represented in a symmetric adjacency matrix $A^{n\times n}$ for which $a_{ij} \neq 0$ *iff* $(u_i, u_j) \in E$. Values of non-zero elements depend on the weighting method adopted for representing link strength, being the absolute counts of edge recurrence (i.e., the edge *count* attribute) the simplest one. Additionally, the model's connectivity constraint states that all diagonal elements in A are necessarily equal to 0, thus ensuring that no self-loops are formed. To give a concrete example, the adjacency matrix for the graph in Figure 5 is

$$A = \begin{array}{c} \\ col1 \\ col2 \\ col3 \\ col4 \\ col5 \\ col6 \\ col7 \\ col8 \\ col9 \\ col10 \end{array} \begin{array}{c} \begin{array}{cccccccccc} col1 & col2 & col3 & col4 & col5 & col6 & col7 & col8 & col9 & col10 \end{array} \\ \left[\begin{array}{cccccccccc} 0 & 2 & 2 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 2 & 0 & 3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 2 & 3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 2 & 2 & 0 \\ 1 & 0 & 0 & 1 & 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 2 & 0 & 0 & 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 2 & 0 & 0 & 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right] \end{array},$$

where each element is the count of the total number of recurrences of collector associations, which are represented in the graph as edge weights.

3.3.2 CWN model construction from data

We build CWNs from species occurrence data in an iterative process that is similar to the one described for SCNs (see Section 3.2.2). In this case, however, the only field that is strictly required for structuring relationships is the one containing collector names, which in a database following terms on Darwin core standards should be named "*recordedBy*". The field containing species identities, although not required, can be optionally used during model construction in case the user decides to set the edges' *species list* attribute. Table 2 shows the species occurrence dataset that was used to build the graph in Figure 5.

For each row in the dataset, a *clique* structure is formed by creating edges between all collectors included in the record's team, in a pairwise fashion. Each clique thus represents one collaborative act, where every collector gets an additional collaborative tie with every other collector included in that collaborative record. The clique size, which is the number of nodes included in the clique structure, is equivalent to the team size. For non-collaborative records the clique is composed of the collector node itself, and therefore no edges are

Table 2 – Species occurrence dataset from which the CWN model in Figure 5 was built. The *species* field is not strictly required for building CWN models.

id	recordedBy	species
0	col1; col2; col3	sp1
1	col3; col1; col2	sp2
2	col1; col3	sp3
3	col5; col4	sp3
4	col5; col2	sp3
5	col5; col6	sp5
6	col5; col7	sp4
7	col6; col7	sp6
8	col6; col7	sp7
9	col4; col8; col9	sp4
10	col4; col9; col8	sp5
11	col10	sp6
12	col10	sp6

created. As a user might want to distinguish the relevance of links originated from teams with distinct team sizes, the hyperbolic weighting rule described in the previous section can be used for weighting links in each clique. In case the species field is included in the building routine, each clique also gets associated with the name of the species to which the recorded specimen belongs to.

The CWN model is finally composed of the combination of all cliques together into a single undirected graph. In this process, edges that occur in multiple cliques have their weights summed up. In case the species list attribute is set, combining edges also merges their respective species lists.

4 Case Study: The University of Brasília Herbarium (UB)

In this chapter, we use the network models proposed in Chapter 3 to understand aspects regarding the taxonomic preferences and the collecting behavior of collectors who have contributed to the University of Brasília Herbarium with specimens records. By exploring basic topological features of the networks, we investigate the formation of (*i*) groups of collectors with similar taxonomic interests; (*ii*) groups of taxa which are often recorded by similar sets of collectors; and (*iii*) groups of collectors who collaborate by recording specimens together in collecting expeditions. From the resulting network structure, we also identify collectors who are the most relevant for the herbarium, and how they contribute for the representativeness of each taxonomic group in the collection. Taxa that are either widely collected or collected by very specific groups of collectors are also easily identified from the network structure.

The University of Brasília Herbarium (UB) is a reference collection for the flora of the Cerrado biome, being noticeably representative for families *Cyperaceae*, *Myrtaceae*, and *Fabaceae*; as well as for cryptogams (algae, mosses, and lichens). The herbarium is physically located at the Biology Department of the University of Brasília (UnB-IB) since its foundation in 1963. Table 3 lists some historically relevant collectors who have intensively contributed to the herbarium, according to the *Florescer* project (PEDROSA; GALLANT, 2018).

Table 3 – Historically important collectors for the University of Brasília Herbarium (UB).

	Activity years	Contribution
William R. Anderson	1962-1976	Central Brazil Expedition (NYBG)
Howard S. Irwin	1962-1976	Central Brazil Expedition (NYBG)
George Eiten	1955-1975[1]	Cerrado biome, mainly in MA state
James Alexander Ratter (1st period)	1968-1976	Xavantina-Cachimbo expedition
James Alexander Ratter (2nd period)	1996-2006	Cerrado biome
Joseph Harold Kirkbride Junior	1976-1983	Flora of DF state
Ana Lúcia Tostes Leite	1982-1984	Continental algae of DF state
Carolyn Elinores Barnes Proença	1981-current	Cerrado biome
Maria das Graças Machado de Souza	1982-current	Continental algae of DF and GO states

Source: Florescer Project (<http://www.florescer.unb.br>)

In this case study, we have used the entire digitized collection of records from

[1] Collecting period inferred from Figure 19.

the UB herbarium (MUNHOZ et al., 2018), which is publicly available for download through the Global Biodiversity Information Facility (GBIF) data portal (GBIF, 2018). After downloading the entire dataset through the portal, we first performed a quick exploratory analysis, for a general overview on its taxonomic composition, the temporal and geographical distribution of the records, as well as the main issues associated with the dataset (Section 4.1). We have also performed data cleaning and transformation routines for atomizing and mapping variants of collectors' names, improving the quality of data from which the network models are constructed (Section 4.2.1). Finally, in Sections 4.2.2 and 4.2.3, we present the SCN and CWN models built from the UB dataset and explore some of their topological features. We have used the *Python v.3.6* language loaded with packages *Pandas*, *Numpy*, and *Matplotlib* for exploring the occurrence dataset; and the *Caryocar* package (designed and implemented by us, in the context of this dissertation) for programatically constructing the SCN and CWN models from occurrence data.

4.1 Dataset exploration

At the time of this study, the entire occurrences dataset from the UB herbarium had a total of 185311 records and 235 fields. For our application, however, only a small subset of those fields were considered to be relevant and were thus included in our exploratory analysis. Most of these fields (except for *issue*), which we briefly describe below, follow *Darwin Core* terms standards.[2] The relevant fields we take into account in this work are:

recordedBy. A string containing names of people or groups who have authored an occurrence record, separated by some delimiter character. Although the vertical bar (' | ') bar is officially recommended by *TDWG*, names are separated by a semicolon character (' ; ') in the UB dataset. The UB convention for collector names makes each of them composed of two parts, separated by a comma (' , '). The first part is the collector's last name, with the first character capitalized; and the second part corresponds to the first initials of the names, all capitalized and appended with a period. A collector named '*João da Silva Simão*', for instance, would be included as '*Simão, J. S.*'.

eventDate. A date-time string representing the moment when the recording act happened.

stateProvince. The name of the state where the occurrence was recorded.

countryCode. The code of the country where the occurrence was recorded.

[2] <http://rs.tdwg.org/dwc/terms>

decimalLatitude. The geographic latitude coordinate of the place where the occurrence was recorded. Datum is assumed to be *WGS84* for all records in the UB herbarium.

decimalLongitude. The geographic longitude coordinate of the place where the occurrence was recorded. Datum is assumed to be *WGS84* for all records in the UB herbarium.

issue. A sequence of data issues that have been identified in the record. This field is included by GBIF during data preprocessing, and is not part of the *Darwin Core* standard.

scientificName. The scientific name assigned to the specimen, at the lowest taxonomic resolution as it can be determined. The authorship of the name is also included for many records, although not relevant for this study.

taxonRank. The taxonomic rank of the name in the *scientificName* field.

Taxonomic composition. For exploring the taxonomic composition of the dataset we first removed 10 records with missing values for the *scientificName* field. From the remaining 185301 records, most had a taxonomic resolution at the *species* level (75.96%), followed by *genus* (13.17%), and *variety*(4.82%), as shown by absolute metrics in Table 4. However, the percentage of records which can be determined at the rank of *species* is slightly higher (82.23%), and is given by the cumulative metrics in the table. This happens because the *species* identity of records at higher taxonomic resolutions (in this case *form*, *variety* and *subspecies*) are directly determined, as explained in section 2.1. Similarly, although only 1.08% of the records have the taxonomic resolution at the *kingdom* level, all records are determined at that rank.

Table 5 shows the number of distinct taxa at each rank and lists those with the highest amounts of records. The herbarium is mostly composed of plants (kingdom *Plantae*, representing 96.72% of the records), followed by *Chromista* (including some phyla of algae), *Funghi*, and *Bacteria*. Vascular plants (phylum *Tracheophyta*) compose 82.89% of the herbarium, although *Bryohpyta* (mosses), *Ochrophyta* (algae from kingdom *Chromista*) and *Charophtya* (algae from kingdom *Plantae*) are also representative. At the family rank, *Fabaceae*, is the most representative one (10.39%), followed by *Myrtaceae* (6.93%), *Asteraceae* (6.08%) and *Rubiaceae* (5.10%). The most collected species are *Myrcia splendens*, *Myrcia guianensis*, *Eugenia punicifolia*, and *Sematophyllum subpinnatum*, the first three flowering plants and the latter a species of moss.

Table 4 – Number of records with taxonomic resolution at each rank. Ranks are ordered hierarchically, being *FORM* the most restrictive (higher resolution) and *KINGDOM* the broader (lower resolution) one. Absolute metrics show the number and percentage of records at each taxonomic resolution, while cumulative metrics show the number and percentage of records that are taxonomically determined at each rank.

	count	%	cumulative count	cumulative %
FORM	1000	0.5397	1000	0.5397
VARIETY	8935	4.8219	9935	5.3615
SUBSPECIES	1681	0.9072	11616	6.2687
SPECIES	140763	75.9645	152379	82.2332
GENUS	24397	13.1661	176776	95.3994
FAMILY	6223	3.3583	182999	98.7577
PHYLUM	294	0.1587	183293	98.9164
KINGDOM	2008	1.0836	185301	100.0000
Total	185301	100.0000		

Table 5 – Number of distinct taxa at each taxonomic rank in the dataset. For each rank a list with its top-4 most recorded taxa is included with their respective counts.

	num of taxa	top-4 taxa	num of records	% of records
kingdom	5	Plantae	179218	96.72
		Chromista	4204	2.27
		Fungi	1391	0.75
		Bacteria	342	0.18
phylum	12	Tracheophyta	153589	82.89
		Bryophyta	16485	8.90
		Ochrophyta	4133	2.23
		Charophyta	3500	1.89
class	29	Magnoliopsida	126288	68.15
		Liliopsida	23004	12.41
		Bryopsida	15899	8.58
		Bacillariophyceae	4133	2.23
order	136	Myrtales	24312	13.12
		Fabales	20846	11.25
		Poales	17159	9.26
		Malpighiales	16188	8.74
family	507	Fabaceae	19254	10.39
		Myrtaceae	12833	6.93
		Asteraceae	11271	6.08
		Rubiaceae	9447	5.10
genus	3374	Myrcia	4654	2.51
		Eugenia	3750	2.02
		Mimosa	2992	1.61
		Miconia	2402	1.30
species	15379	Myrcia splendens	696	0.38
		Myrcia guianensis	560	0.30
		Eugenia punicifolia	462	0.25
		Sematophyllum subpinnatum	438	0.24

Geographic distribution. We then explored the geographic distribution of the herbarium records. From the set of 185,301 records, approximately 48% (a total of 89,216) were interpreted and tagged as having geospatial issues during internal preprocessing routines in the *GBIF* platform (GBIF.org, 2018). Routines perform geospatial validation by checking, for each record: (*i*) if geographical coordinates are erroneously assigned to zero latitude and longitude; (*ii*) if geographic coordinates are in fact placed within the boundaries of the country, if the country is indicated; and (*iii*) if the lat/long values are likely to having been accidentally swapped or negated during recording. If inconsistencies are detected during the execution of such routines, issues are registered for each record, prior to making the dataset available for download or query via an API.

As shown in Figure 6, most geospatial issues are due to records being assigned to zero coordinates (zero latitude and longitude), which also makes them laying outside the country boundaries. As a consequence, they are tagged as having both the "Zero Coordinates" (ZC) and "Country Coordinates Mismatch" (CCM) issues, being represented in the Figure 6 as the intersection between CCM and ZC. Some records with only the CCM issue were not misplaced in coordinates $(0, 0)$, but were still improperly placed outside country frontiers. On the other hand, records with only the ZC issue are those which were misplaced in coordinates $(0, 0)$, but were lacking information about the country, and therefore couldn't be classified as being outside the country frontiers. Records that were not interpreted as having geospatial issues comprise approximately 52% of the dataset, and the coordinates of those records that are located around the Brazilian territory are shown in Figure 7.

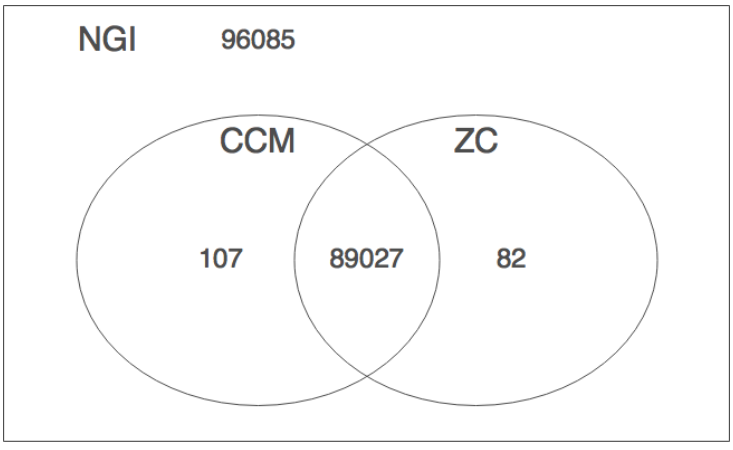

Figure 6 – Number of occurrences from the UB herbarium dataset classified within each geospatial issue class. *NGI*: Records with no geospatial issues; *CCM*: Country Coordinates Mismatch issue; *ZC*: Zero Coordinates issue.

Most records deposited in the UB herbarium were collected in Brazil (94.47%), from which the Federal District (DF) is the most sampled federative unit despite its relatively small area (Figure 8(b)). Records in the Federal District comprise 30.56% of those falling within the Brazilian territory. Other states with relatively high collecting effort are Goiás

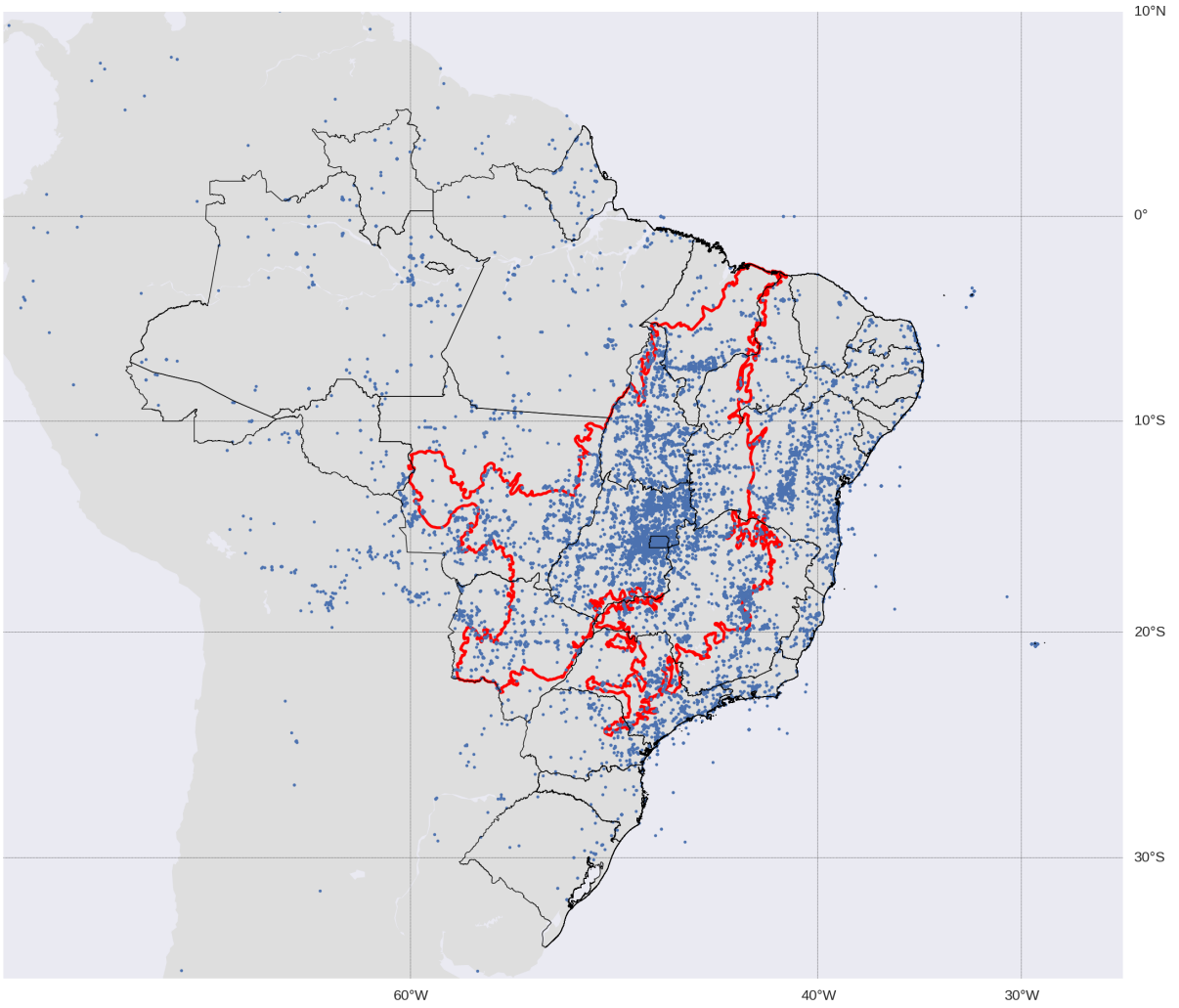

Figure 7 – Geographic distribution of the occurrences from the UB Herbarium dataset near the Brazilian territory. Records without geospatial issues are placed in the map as blue dots. The area outlined in red represents the boundaries of the Cerrado biome.

(22.10%), Minas Gerais (14.51%), and Mato Grosso (7.86%), which, together with the Federal District, correspond to 75% of all records from Brazil in the UB herbarium. Such a geographical bias in the occurrence distribution can be explained by the fact that UB is physically located at the University of Brasilia, making it more viable for associated collectors to perform collecting expeditions in nearby locations. Moreover, as UB is a national reference herbarium for species occurring the Cerrado biome, botanists working in Cerrado areas nearby might become more inclined towards depositing their recordings in that institution. Figure 7 shows that UB records are more densely concentrated within the Cerrado biome, mostly in the Central Brazil region. The UB herbarium also includes a total of $10,252$ records from other countries (Figure 8(a)), which are derived from exchanges with foreign herbaria (under the curatorship of *Dr. George Eiten* (PEDROSA; GALLANT, 2018)) and from international recording expeditions performed by collectors associated with the UB herbarium.

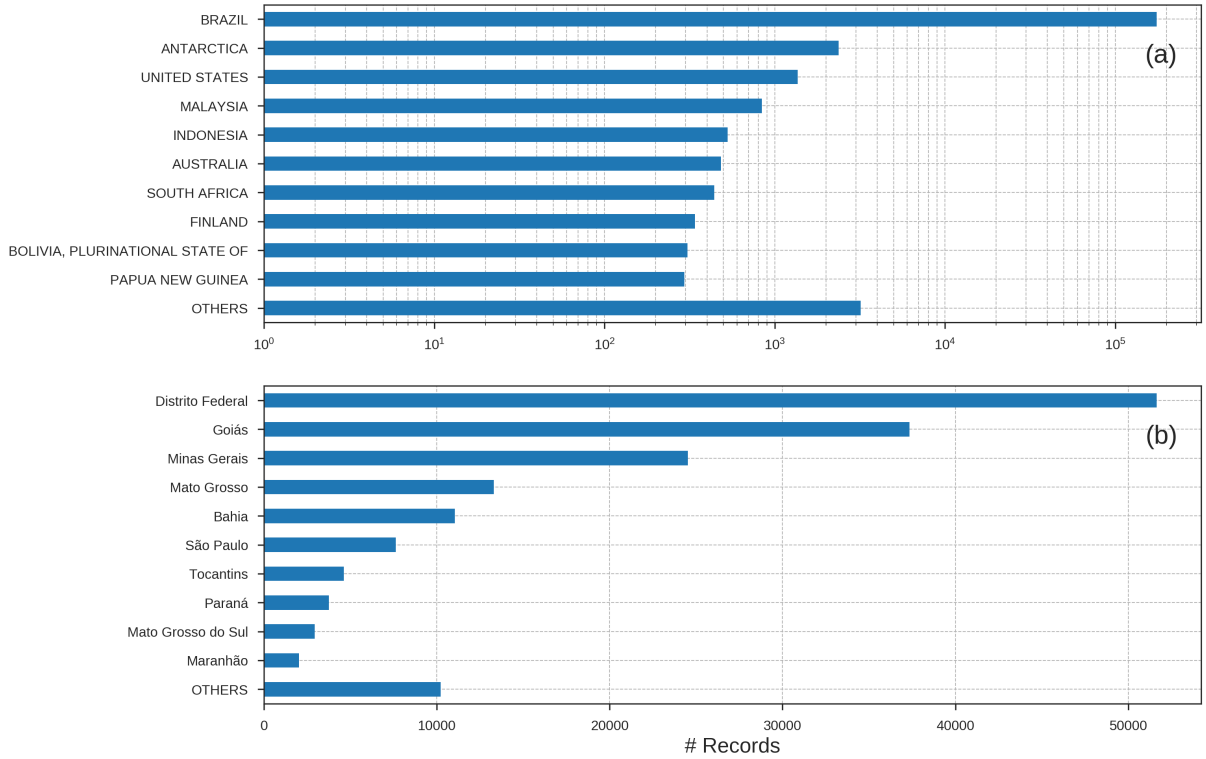

Figure 8 – Top-10 countries (a) and top-10 Brazilian states (b) with most occurrence records deposited in the UB herbarium. Records from countries and states beyond the $10th$ position in the respective ranks are summed and assigned to *OTHERS*.

Temporal distribution. Although preprocessing routines in GBIF also include the interpretation of date-time values in the *eventDate* field, no flags regarding date-time inconsistencies have been assigned to any of the records from the UB dataset. A total of 181, 254 records, comprising 97.86% of the 185, 301 records we explored, contain information about their collection dates. The temporal distribution of the records spans the period from the year 1800 up to 2017, although a more intensive collection activity starts around 1960 (Figure 9). In fact, the vast majority of records (96.11%) are concentrated within the period from 1960 to 2017 (an average of 3, 000 records per year), whereas the period from 1800 to 1959 includes only 3.9% of the records, with an average of 45 records per year. Moreover, the number of records accumulated during the last 30 years (from 1988 to 2017) is approximately equal to the number of records from 1800 to 1987 (188 years). Three peaks of activity are most pronounced in Figure 9. The first is around years 1966 and 1968, the second around 1988, and the third around 2012.

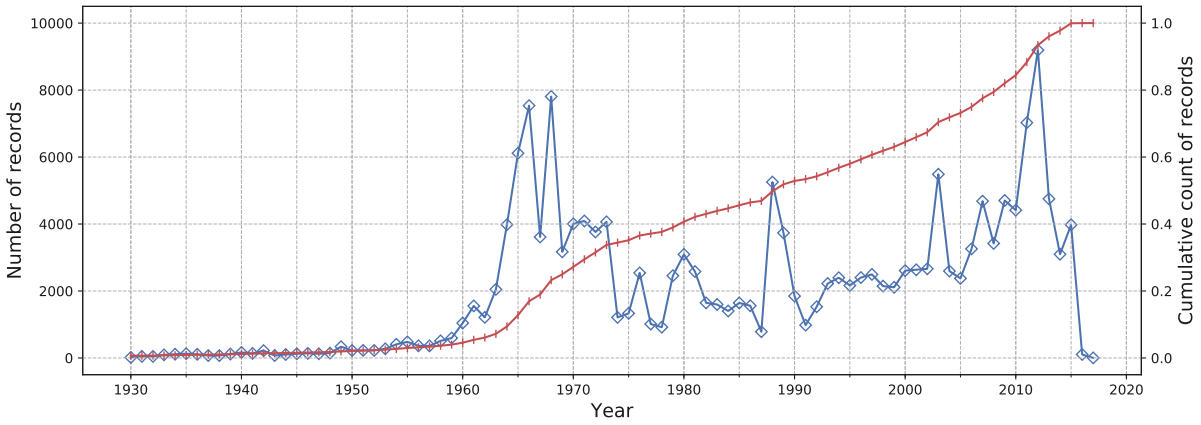

Figure 9 – Recording activities registered in the UB herbarium aggregated by year, since 1930. Both the absolute (blue line) and the cumulative (red line) record counts for each year are shown.

4.2 Construction of the network models

In this section, we describe the main features of the Species-Collector Network (SCN) and the Collector CoWorking Network (CWN) constructed from the UB herbarium occurrences dataset. Using these models we explore the diversity of recording behaviors of the collectors who have contributed to UB, in terms of their taxonomic preferences and collaborativeness during collecting expeditions. We start by describing our data preparation routine, necessary for improving the quality of the network models.

4.2.1 Data preparation

Before we could use the UB occurrences dataset for actually building the network models, we submitted the tabular dataset to some data filtering and transformation routines. The data preparation process consisted of (*i*) selecting occurrence records from which relevant social ties could be derived for both network models; (*ii*) extracting atomized collector names from the *recordedBy* field, which originally contains a string of names; (*iii*) normalizing the extracted collector names to obtain their id's; (*iv*) resolving inconsistencies on collector names and mapping name variants to entities; and (*v*) filtering out inadequate collector names.

In the first step, we initially removed all records with missing values for either fields *recordedBy* or *scientificName*, as these are both critical for the construction of the network models. In order to build the SCN at the taxonomic resolution of *species* we selected records for which the taxonomic identity could be resolved at that level, *i.e.*, at the resolutions of *form*, *variety*, *subspecies*, and *species*. This subset of records corresponds to approximately 82% of the original dataset (a total of 152, 379 records, as shown in Table 4). All records with taxonomic resolution higher than *species* (from *genus* to *kingdom*) were

not used for the construction of the networks. Although the taxonomic resolution of records is irrelevant at all for the construction of the CWN model (edges are built exclusively using cliques of collectors, a process described in Section 3.3.2), we chose to use the same set of records we used for building both models.

The name atomization step consists of iterating each record from the UB dataset and splitting the corresponding string containing collector names, originally stored in the *recordedBy* field. The process originates a list of strings for each record, containing atomic names for each collector associated with it. A new field named *recordedBy_atomized* is created for holding the lists of atomic names, and is later used to construct the network models. We split the strings on the semicolon character (‘ ; ’), as it is used as the delimiter for the majority of entries. In some few occurrences, however, names are delimited by other characters, for instance ‘ , ’, ‘ / ’ and ‘ & ’. We dealt with these cases by replacing inconsistent delimiters manually and storing the changes in a separate file, without modifying the original dataset.

In order to obtain unicode identities for collectors, we defined a unicode normalization function, which is executed on the strings forming the names of collectors upon network construction (step *iii*). Each name is initially split in its two component parts (see Section 4.1), using the comma character as the delimiter. Next, each part of the name is passed into the normalization function, which forces all characters to lowercase and removes any accents, periods, and spaces. Finally the name is recomposed by joining the first initials to the last name with a comma. For instance, the name ‘*Simão, J. S.*’ would be normalized to ‘*simao,js*’. Names were mapped to their respective normalized forms and were stored in a *names map* file, which was later passed in to the constructor methods that built the network models. A table mapping the identities of some of the main collectors from UB to their respective names is presented in Appendix A. However, we still faced the problem of resolving names inconsistencies (step *iv*). If a collector is associated with multiple name variants, he/she gets consequently represented as multiple entities in the network, which should be avoided as to ensure the semantic correctness of the models. For dealing with these issues, we included new entries in the same *names map*, keying variants of each name to their “correct forms”. Finally, after the networks were built, the last step was to remove entities which we considered to be “noise” in the network, such as ‘*etal*’ (from *et. al*), ‘*incognito*’, ‘*ignorado*’, ‘*ilegivel*’ or ‘*?*’.

4.2.2 The UB Species-Collector Network

The SCN model of the UB herbarium has a total of 6,768 collectors and 15,344 species nodes, with a total of 142,647 undirected edges connecting nodes from opposite sets. The average degree for the collector and species set is 21.08 and 9.30, respectively.

Connected components. The network is composed of a total 351 connected components, the largest of which (the giant component, or c_1) contains the majority of nodes in the network (93.6% of the collectors and 95.0% of the species). From a collector's perspective, the requirement for it to belong to the giant component is that it must have collected at least one species in common with another collector who is already included in the giant component. The same reasoning applies to species nodes, by observing the inverse relationship. Apart from the giant component, most other connected components contain as few as two or three nodes, representing collectors who have never recorded a species in common with any collector from c_1; and conversely, species that have never been collected by any of those collectors that belong to c_1. One of those 350 remaining connected components, however, is considerably larger than the others, with a total of 3 collectors and 141 species. We refer to it as the second largest component (c_2) throughout this section.

Among all the records that form the giant component, 95.2% are from Brazil, out of which 53% were recorded either in the Federal District or in the state of Goiás. Further, c_1 is mostly composed of species from phylum *Tracheophyta* (88% of all records), followed by phylum *Bryophyta* (mosses), comprising 8% of the records. Component c_2, on the other hand, is represented by algae (phyla *Charophyta*, *Chlorophyta*, comprising 91.6% of all records), bacteria (phylum *Cyanobacteria* (4.3%)), and other microscopic eukariotic organisms (phyla *Euglenozoa*, *Myzozoa* (4.1%)), which are taxonomically distinct from the vast majority of species in the herbarium. In addition, all records composing c_2 were collected in the Federal District. The remaining components ($c_3, c_4, ..., c_{351}$) include a total of 431 distinct collectors and 446 distinct species, resulting from records from many other countries than Brazil (79% of the records). The most representative country is the United States, with 23.9% of the records, followed by Brazil (with 21%). Considering the records from within Brazil, only 7.8% of them were collected in the Federal District. Most of them (59.2%) are from the southeast region, among which the state of São Paulo is the most representative (52.5%).

We have hypothesized some possible explanations for the existence of so many small connected components in the UB herbarium SCN with such characteristics. First, they could be a consequence of specimen exchange, a practice that is widely adopted among herbaria for collaboratively diversifying and distributing their collections (GROOM; REILLY; HUMPHREY, 2014). Exchange materials are typically duplicated exsiccates collected in field that a sender institution dispatches to a receiver institution. The receiver institution then incorporates those materials to its scientific collection, and ocasionally sends back to the other institution some duplicate exsiccates of its own. From the receiver herbarium viewpoint the inclusion of records from exchanges usually adds new species and collectors, for they are a sample reflecting the sender institution's purpose and the interests of people associated with it. If both the collectors and species from such

records were previously inexistent in the receiver collection, no links to c_1 or to any other connected component that already exists are formed. These records thus get included into its SCN network as nodes and edges composing new small connected components. We should overstate, however, that including exchange records does not necessarily create new connected components. New species or collectors could be included in c_1 in case either the species or at least one of the collectors associated with the exchange record are already linked to it.

A second situation that could potentially lead to isolated components in the network is when there are groups of very specialized collectors sampling very specific and distinct groups of organisms. Cryptic organisms such as algae, fungi, and mosses are examples of groups that tend to be overlooked by botanists who are not directly interested on them. If collectors who record such groups also show a very high specificity towards them, it gets more likely that they lack links to other species that are more commonly recorded by the rest of the collectors, thus forming structures that are weakly linked (if not detached) to the giant component. In the case these collectors regularly deposit their materials in the herbarium, it would be natural to observe them composing connected components with a relatively large number of species. The connected component c_2, for instance, includes *Ana Lúcia Tostes Leite* (*leite,alta*), one of the herbarium's most relevant continental algae collector as shown in Table 3. She holds a total of $2,757$ records, from 87 distinct species, all of them being charophytes, a division of green algae (Figure 12, within shape *ii*).

Number of species per collector. Collectors contributing to the UB herbarium have, on average, recorded and successfully deposited approximately 21.1 distinct species in that collection during their careers (Table 6). This value is obtained by computing the average degree $\langle k_{col} \rangle$ from all nodes in S_{col} set, and would be equivalent to the expected number of species for a randomly selected collector, if the SCN were a random network (ALBERT; BARABÁSI, 2002). However, by visually inspecting the degree distribution of collectors in Figure 10(b) and (d), we realize that the vast majority of collectors have recorded very few species. In fact, around 86.7% of the collectors have degrees below or equal $floor(\langle k_{col} \rangle)$, and around 79% have recorded 10 or fewer species. On the other hand, there are also some few collectors who have recorded much more species than the average, as it is the case of *Howard S. Irwin* (*irwin,hs*), with a total of $4,535$ records of distinct species; and *Carolyn E. B. Proença* (*proenca,ceb*), with $1,888$ distinct species. Such nodes holding a very high number of connections in the network are called hubs.

As opposed to random networks, in which extremely well connected nodes are unlikely to coexist with a large number of extremely poorly connected ones, the degree distribution observed for collectors in the UB SCN is better approximated by a power law $p(k) \sim k^{-\alpha}$, typically observed in many large real-world networks with the scale-free property (BARABÁSI; ALBERT, 1999). The dashed line in Figure 10(b) is a power law

Table 6 – Degree centrality metrics for the UB SCN model. For each nodes set the total number of nodes, average degree $\langle k \rangle$, top-10 highest-degree nodes, and their respective degree k, weighted degree k_w, and normalized degree k^* are listed.

	num of nodes	$\langle k \rangle$	top-10	k	k_w	k^*
collectors	6768	21.08	irwin,hs	4535	18065	0.30
			heringer,ep	2586	6495	0.17
			anderson,wr	2156	4710	0.14
			proenca,ceb	1888	4803	0.12
			ratter,ja	1803	4728	0.12
			faria,jeq	1681	4693	0.11
			eiten,g	1586	3048	0.10
			souza,rr	1549	3887	0.10
			harley,rm	1514	2564	0.10
			santos,rrb	1502	3587	0.10
species	15344	9.30	*Myrcia splendens*	388	1284	0.06
			Myrcia guianensis	335	1061	0.05
			Eugenia punicifolia	266	796	0.04
			Casearia sylvestris	258	662	0.04
			Palicourea rigida	241	456	0.04
			Myrcia tomentosa	239	634	0.04
			Qualea parviflora	232	567	0.03
			Solanum lycocarpum	228	340	0.03
			Piper aduncum	209	371	0.03
			Miconia albicans	201	425	0.03

fit for the degree distribution of nodes from S_{col}, with $\alpha = 1.52$. An apparently similar heavy-tailed distribution was recently reported by Daru et al. (2017), while investigating sampling bias in the digitized datasets of three distinct herbaria from Australia, South Africa, and the United States.

Other useful degree centrality metrics shown in Table 6 are the weighted degree k_w and the normalized degree k^*. The weighted degree for each collector node is computed by summing up weights assigned to each of the node's edges. Edges can be weighted according to any chosen attribute of its own, depending on the aspect of the connection the user wants to emphasize. Here we use the *count* attribute, which holds the number of times the association between a collector and a species occurs. Formally, $k_w^{(u)} = \sum_v count(e_{u,v})$ for $v \in neighbors(u)$. Differently from the degree, which gives the number of distinct species that have been recorded by a collector, the weighted degree of a collector can be interpreted as the total number of specimens (or occurrence records) collected by him/her. It is also worth noting that the value of k_w of a node is equivalent to its *count* attribute (not to be confused to the homonymous attribute of edges).

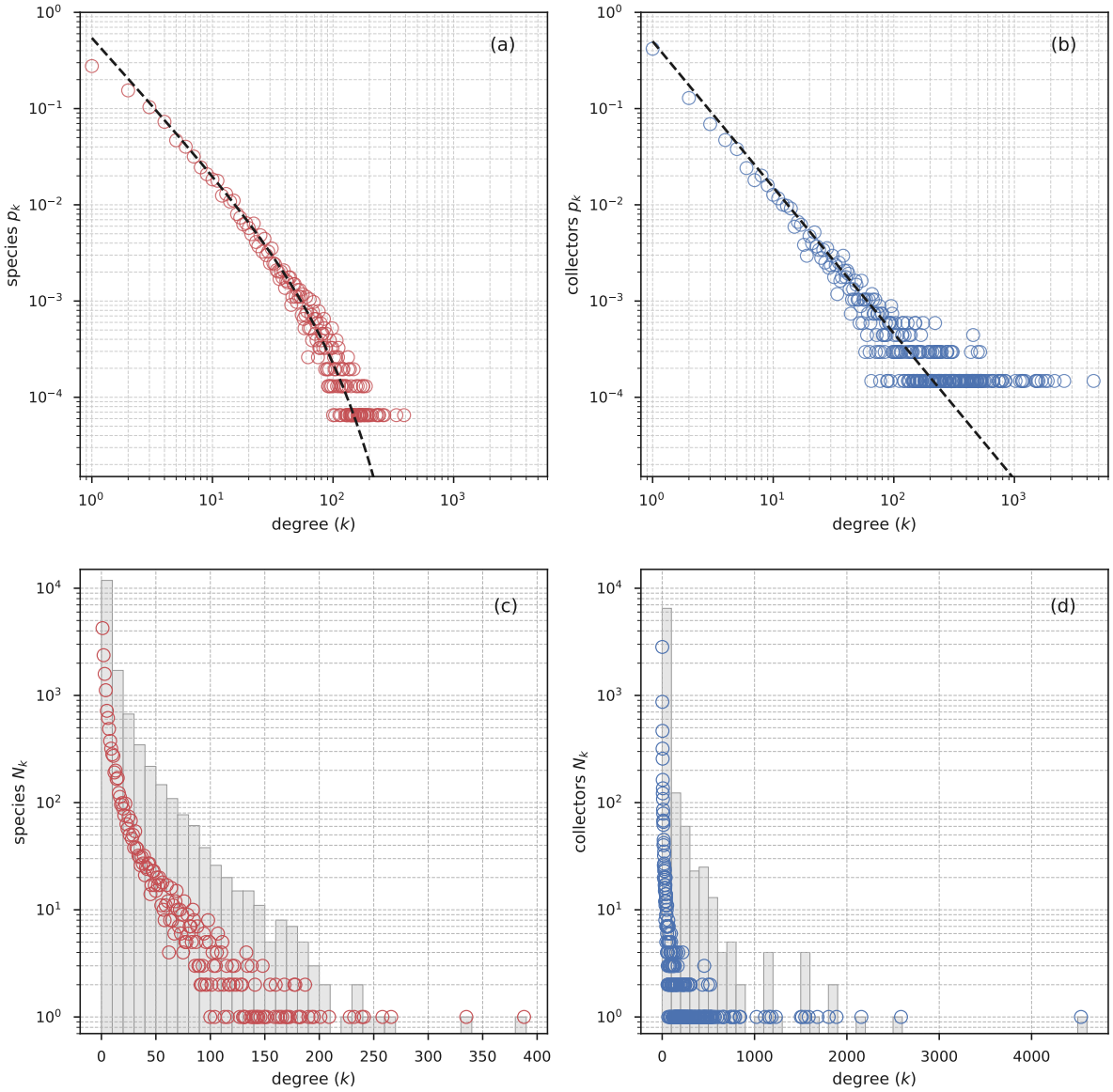

Figure 10 – Degree distribution for both species nodes, in red in (a) and (c), and collector nodes, in blue in (b) and (d) in the UB SCN. The upper row plots show the probability p_k of finding nodes with each degree value, using a log-log scale. The dashed curves represent power laws (note that (a) has a tail cutoff) that best fit the SCN data, with $\alpha = 1.38$ and $\alpha = 1.52$ for plots (a) and (b), respectively. The exponential cutoff in (a) is obtained by using $\lambda = 0.014$. Plots in the lower row show the total number N_k of nodes with each degree value, using a lin-log scale for enhancing interpretability. Histograms in the background group collectors using bin sizes $s = 10$ and $s = 100$ for plots (c) and (d), respectively.

The intuition behind the normalized degree k^* of a collector is the fraction of species from the entire S_{sp} set it is connected to. We compare the degree of the node to the maximum degree value that would be possible for nodes in the same set. In a bipartite graph model, it turns out to be the node's degree divided by the size of the opposite set. Therefore, $k^{*(i)} = \frac{k^{(i)}}{|S_{sp}|}$, for a node $i \in S_{col}$. As stated by Borgatti and Halgin (2015), the

advantage of using this normalized metric over the non-normalized degree k is that it allows us to numerically compare centrality scores across nodes sets independently of their respective sizes. In our case, hubs in the set of collectors then tend to be more central than hubs of species. The top collector *irwin,hs* is linked to 30% of the total diversity of the herbarium species, whilst the top species *Myrcia splendens* has been recorded by 6% of the herbarium collectors (Table 6).

Number of collectors per species. For the species node set (S_{sp}), the average degree $\langle k_{sp} \rangle \approx 9.5$, which can be interpreted as that species in the UB herbarium have been collected by approximately 9.5 distinct collectors, on average. The total number of times each species has been recorded and the percentage of collectors that have recorded them at least once are given by the k_w and k^* metrics, respectively (see Table 6).

The degree distribution for species nodes (Figure 10(a) and (c)) is similar to the distribution for collector nodes, with a majority of very low degree nodes coexisting with a few hubs. Most species (around 76.7%) hold degree values below or equal $floor(\langle k_{sp} \rangle)$, meaning they have been recorded by 9 distinct collectors or less. Yet, there are also some species that have been recorded by many more collectors than the average, as it is the case of *Myrcia splendens* and *Eugenia punicifolia*. The top-10 species (Table 6) are in fact reasonably well-known, common, and easily detectable. Typical from cerrado physiognomies, *Solanum lycocarpum* (*lobeira*), *Palicourea rigida* (*chapéu-de-couro*), *Qualea parviflora* (*pau-terra*) are examples of species that are both very conspicuous and easy to identify.

A particularity that can be observed in the species degree distribution is the existence of a sharp tail decrease in the probability curve towards the highest degree nodes (Figure 10(a)), known as a tail cutoff. This behavior can be better adjusted by using a slight variant of the power law function $p(k) \sim e^{-\lambda k} k^{-\alpha}$, which is obtained by simply including an exponential term in the function. We call this variant a truncated power law function, as the degree distribution of nodes is not scale-independent. Truncated power-laws have been also reported by Newman (2004) for scientific collaborations networks. As there is an evident limitation on the number of papers scientists are able to publish during an interval, Newman has attributed this behavior to the limited time window used in his studies. However, the same behavior has also been observed in other real world networks, in which the formation of a very high number of links on any node is constrained by some physical factor, specific to each system (ALBERT; BARABÁSI, 2002). In the case of SCNs, the exponential cutoff might be due to the fact that as the representativity of species in the herbarium increases, collectors become less inclined towards collecting those species in the future. The dashed curve in Figure 10(a) represents a truncated power with parameters $\alpha = 1.38$ and $\lambda = 0.014$, which was fit to the species degree distribution data.

How densely connected are species and collectors? The density of the network is the ratio between the actual number of edges in the network and the maximum possible number of edges, in case every collector were linked to every species. This metric informs us how far the network is of being complete, in which case its density becomes $d = 1$. For a bipartite network, density can be computed as $d = \frac{|E|}{|S_{col}||S_{sp}|}$, where $|E|$ is the number of edges in the network and $|S_{col}|$ and $|S_{sp}|$ are the number of collectors and species, respectively. The observed density for the overall UB species-collector network is approximately 1.37×10^{-3}, meaning that the probability that we find an edge linking two arbitrary nodes from opposite sets is about 0.14%. Additionally, as shown in Figure 11, network density increases as we perform taxonomic aggregations onto successively higher-hierarchy ranks. In most social networks, however, not all regions are equally dense, as entities usually do not connect to others randomly. Instead, entities with similar attributes or interests are more likely to establish new ties with each other than with dissimilar ones, leading to the formation of communities in an assortative network.

Communities of common interests. Communities in SCNs are formed by groups of collectors who are more interested towards particular subsets of species than are other collectors, external to the group. As the number of edges linking members of a community with other members tends to be larger than those connecting members to non-members, communities can be visually detected as distinguished clusters of nodes in the network when using force-directed algorithms (JACOMY et al., 2014) for graph layout.

As the size of the UB SCN is relatively large for it to be informative in a static figure, we summarized the network in two steps. First we aggregated the SCN onto the *family* rank following the routine described in Section 3.2.3, as it would be impractical to draw relevant conclusions from the network, if every single species were plotted in the figure. By performing the taxonomic aggregation, we reduced the number of S_{sp} nodes (taxa) from $15,344$ to 474, although the number of edges (from $142,647$ to $43,803$) did not decrease in the same proportion (Figure 11). This incurred in a 10× increase in network density, from 1.37×10^{-3} to 1.36×10^{-2}. In the second step of the summarization routine we removed collectors-families ties which occurred less than 20 times throughout the entire dataset by filtering edges based on their *count* attribute. As this edge filtering routine resulted in many isolated collectors, most of which novice collectors with low absolute recording counts, we also omitted them as to improve the figure readability.

Three communities are visually distinguishable from the central region of the network, which we refer to as the network core (Figure 12). Although the network core could be considered a community *per se*, we prefer to think of it as a region that best reflects the overall interests of the majority of the collectors contributing to the herbarium. Nevertheless, collectors from the network core still vary considerably regarding their recording interests, as it can be verified by inspecting the sets of taxa they're linked to

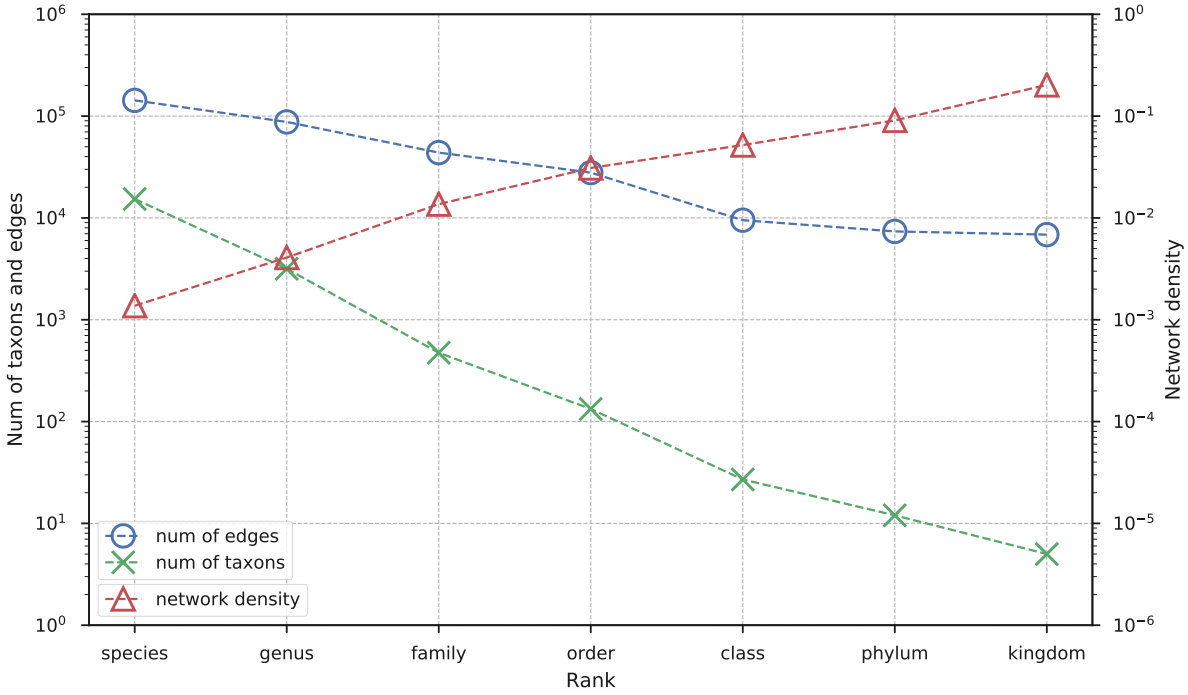

Figure 11 – Number of taxa (S_{sp} nodes), edges and network density for the UB SCN aggregations onto successive taxonomic ranks.

and the strength of their connections. Those who have sampled organisms from many distinct families (and are thus considered to display a more generalist collecting behavior) are placed more centrally in the network core by the layout algorithm, whereas those who are more specialists are consequently pushed towards the borders of the network core, as near as possible to their most recorded taxa.

Howard S. Irwin (*irwin,hs*) is the collector with most records in the network, having intensively collected organisms from many distinct families, especially from the most central ones (illustrated in Figure 12 as the largest pink nodes in the network core). The majority of his records are, in descending order, from families *Fabaceae*, *Rubiaceae*, *Asteraceae*, *Poaceae*, and *Cyperaceae*. He is also the collector holding the highest number of records for those families in the herbarium. An interesting fact is that although *Myrtaceae* is the second most recorded family in the UB dataset (with a total of 10,951 records), it was relatively overlooked by '*irwin,hs*', having himself contributed with only 399 *Myrtaceae* records. The main *Myrtaceae* collector in the herbarium is *Jair E. Q. Faria* (*faria,jeq*), who apparently has a preference towards this family (it comprises 31.0% of his entire set of records). It could be insightful to investigate why a generalist collector such as *Irwin*, who mostly contributed to the herbarium in the context of the Central Brazil expedition (see Table 3), was not very interested in such a predominant family from the Cerrado biome as *Myrtaceae*. *Carolyn E. B. Proença* (*proenca,ceb*) is another key *Myrtaceae* collector, although she also seems interested, to the same extent, in families *Fabaceae* and *Asteraceae*. Moreover, Figure 12 also makes it easy to detect collectors who exclusively (or almost

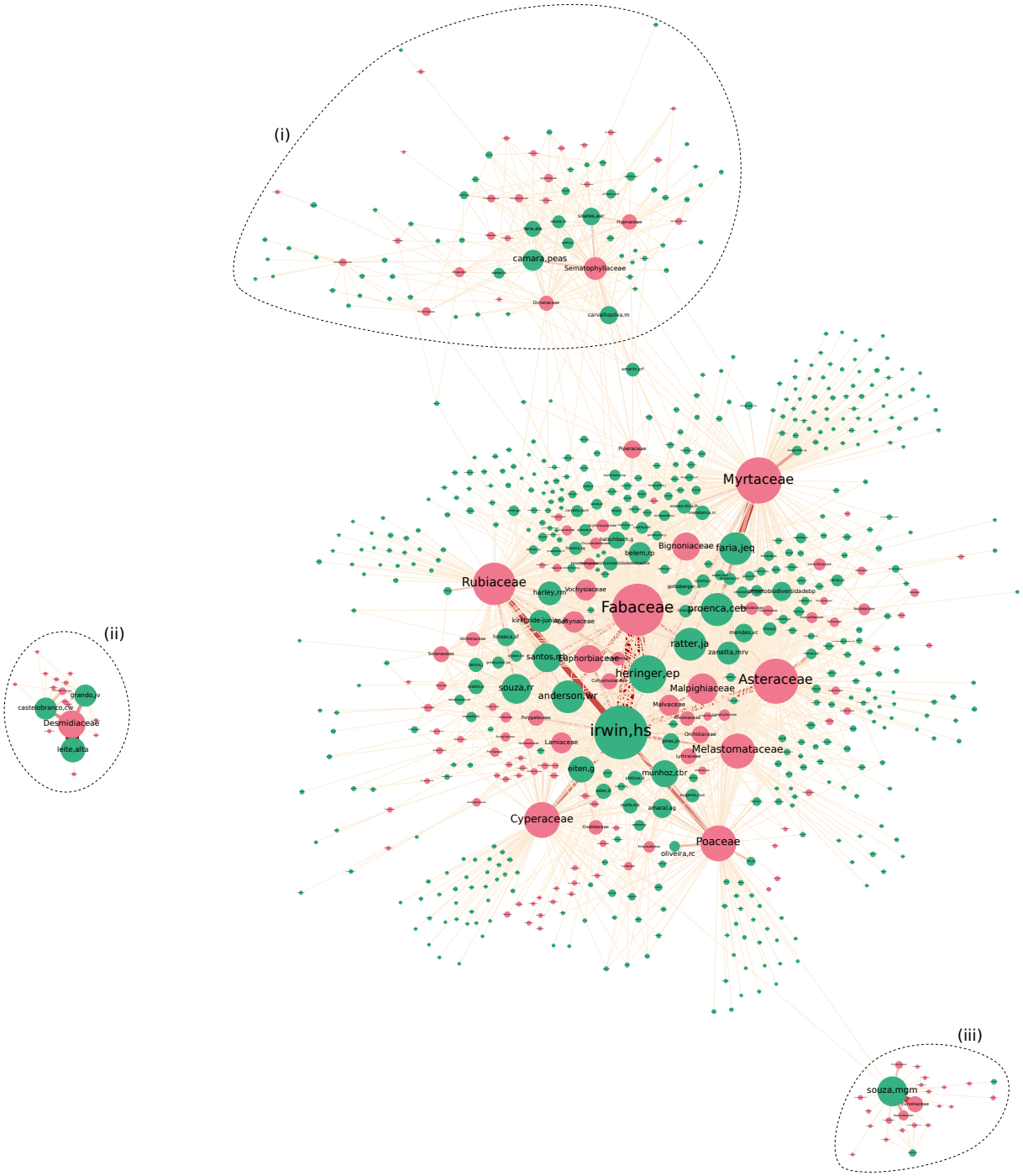

Figure 12 – General aspect of the UB SCN taxonomically aggregated at the family rank. Species and collectors are respectively represented as pink and green nodes, and edges are weighted according to their *count* attribute. Nodes sizing is based on their *count* attribute, whilst edges color and width reflect their weight. For improving visualization we first filtered out edges with weight lower than 20 and then omitted isolated nodes from the resulting graph. Nodes within dashed shapes (*i*), (*ii*), and (*iii*) compose visually distinguishable communities. Graph layout was computed using the *ForceAtlas2* algorithm (JACOMY et al., 2014). For a better visualization experience refer to the interactive version of the graph (<https://lncc-netsci.github.io/pedrocs/networks/ub_scn>).

exclusively) collect each family, as it is the case of *Vanessa G. Staggmeier* (*staggmeier,vg*) for *Myrtaceae* and *Regina C. Oliveira* (*oliveira,rc*) for *Poaceae*, for instance.

Community (*i*) in Figure 12 represents a large part of the collectors from *Cryptogams Lab*[3], together with the taxa they are typically most interested in. The lab is part of the University of Brasília Department of Botany, having *Paulo Eduardo A. S. Câmara* (*camara,peas*), *Micheline C. Silva* (*carvalhosilva,m*), and *Maria das Graças M. de Souza* (*souza,mgm*) as the principal investigators. The first two researchers, included in community (*i*), are mostly interested in bryophytes (mosses and liverworts), mainly those from families *Sematophyllaceae*, *Hypnaceae*, and *Dicranaceae*. *Micheline C. Silva* also shows interest towards *Piperaceae*, a family of flowering plants that is also fairly recorded by some collectors from the network core. Therefore, *Piperaceae* is an important node connecting community (*i*) to the network core, as it intermediates many paths between collectors from both network regions. Some PhD and MSc students from the *Cryptogams Lab* including '*soares,aer*', '*faria,ala*', '*souza,rv*', '*dantas,ts*', '*gama,r*' and '*pinheiro,eml*' have also been placed in community (*i*), and their closeness to their academic supervisor '*camara,peas*' reflects their common taxonomic interests in bryophytes.

Although she is one of the principal investigators of the *Cryptogams Lab*, *'souza,mgm'* was placed in community (*iii*), instead of (*i*), due to her taxonomic interest towards algae, a taxonomic group which is overlooked by the vast majority of collectors in the herbarium, including bryophytes collectors. She is mostly interested in families *Eunotiaceae*, *Naviculaceae*, and *Pinnulariaceae*, which compose a group of algae known as diatoms. There are two more collectors in community (*iii*). Together with *'souza,mgm'*, *Roni Ivan Rocha de Oliveira* (*oliveira,rir*) was a member of a survey on the diatoms acquatic biota of the Paranã River, during years 2002 and 2003. *Drielle dos Santos Martins* (*martins,ds*) was an undergraduate student and a lichen collector, having been supervised by '*souza,mgm*'.

Another community of algae collectors is the one formed by *Ana Lúcia T. A. Leite* (*leite,alta*), *João V. Grando* (*grando,jv*), and *Christina W. Castelo Branco* (*castelobranco,cw*) (community (*ii*)), which also happens to be the connected component c_2 itself. *Ana Lúcia Leite* has intensively collected continental green algae from family *Desmidiaceae* in the Federal District while pursuing her MSc degree at the University of Brasília. During that period, she has also collected some specimens from another green algae family, *Closteriaceae*, which has not been recorded by anyone else from the UB herbarium. She is regarded as one of the UB historically most relevant collectors (Table 3). *João Grando* (*grando,jv*) and '*castelobranco,cw*' were also MSc students at the University of Brasília, both working under the academic supervision of Dr. *Antônio José de Andrade Rocha* (not included as a collector in the UB dataset). Besides green algae (families *Desmidiaceae*, *Scenedesmaceae*, *Chlorellaceae Selenastraceae Hydrodictyaceae*), they have also collected other types of

[3] <http://labcriptounb.blogspot.com.br/>

microscopic organisms, such as euglenophytes (family *Euglenaceae*) and cyanobacteria (families *Nostocaceae*, *Oscillatoriaceae*, *Chroococcaceae*, and *Pseudanabaenaceae*). All those organisms are typical of aquatic ecosystems.

We could use a whole set of modularity detection algorithms in order to further split the network core and identify communities which are visually less conspicuous. However, if we used general modularity algorithms (appropriate for one-mode graphs) for analyzing non-projected bipartite networks, a set of bipartite constraints and assumptions would be systematically ignored (BORGATTI; HALGIN, 2015). A common practice among the statistical physics community for detecting communities in bipartite networks has been to first obtain bipartite projections and then running unipartite community detection algorithms for each of them, separately. In order to minimize information loss, obtaining weighted bipartite projections is preferred over unweighted, as the latter has been shown to produce poor results (GUIMERÀ; SALES-PARDO; AMARAL, 2007).

In order to investigate more thoroughly communities of common interests under complementary perspectives, we projected the family-aggregated SCN network shown in Figure 12 into both the collector (Figure 13) and the species sets (Figure 15), using the approach described in Section 3.2.4. We used the species bag / quorum similarity rule (Expression 3.3) for weighting the edges in the resulting unipartite networks.

Communities in the collector projection. The collector projection of the family-aggregated UB SCN network has a total of $6,768$ nodes and $5,431,799$ edges, with average node degree $\langle k \rangle = 1,605$. Although the taxonomic aggregation process does not change the number of nodes at all in a collector projection of a SCN, the number of edges connecting collectors together increases significantly from lower-rank towards higher-rank aggregations, which means resulting networks become denser. In fact, the density of our family-aggregated SCN collector projection is 2.37×10^{-1}, while the same projection at the species resolution is 6.5 times sparser, with a density of 3.64×10^{-2}. This happens because higher-level taxa represent more inclusive groupings of organisms, and the aggregation process causes them to incorporate ties from all lower-rank taxa they are composed of. In other words, the lower the taxonomic resolution of the network model is, the more inclusive are taxa represented as nodes, and the more likely it is that collectors record at least one specimen belonging to each taxon.

Another issue is that collectors with very few records in the dataset are more likely to have recorded identical sets of taxa by chance, leading to identical species bags. As a consequence, they are tied by edges with similarity very close to 1. This explains the formation of many fully connected regions containing low-degree nodes, which contribute to increasing the network density. More experienced collectors can be observed to have high similarity with others, but it is unlikely to be very close to 1.

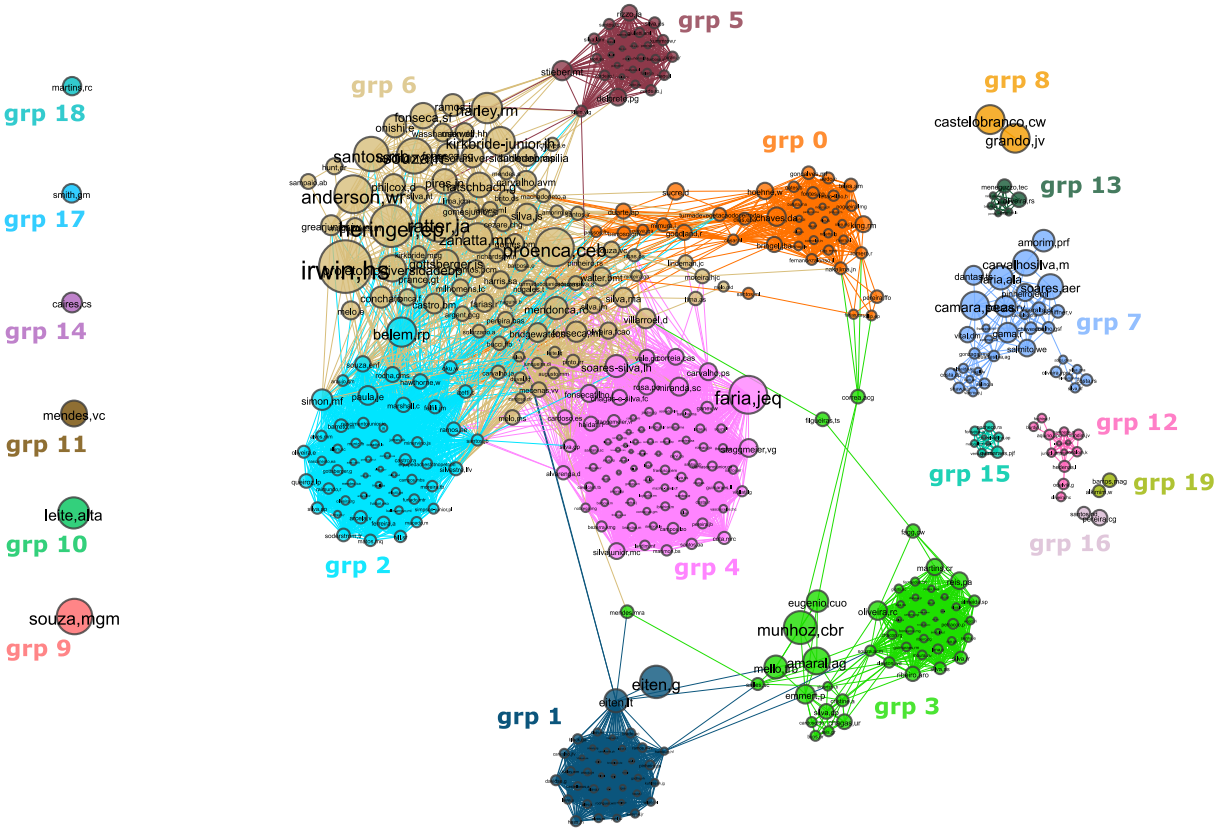

Figure 13 – Communities of common interests in the collectors projection of the family-aggregated UB SCN (from Figure 12). In order to improve visualization, communities with scores lower than 500 were omitted from the figure. Colors are used to distinguish communities, and nodes are sized according to their total number of records in the dataset. Graph layout was computed using the *ForceAtlas2* algorithm (JACOMY et al., 2014). For a better visualization experience refer to the interactive version of the graph (<https://lncc-netsci.github.io/pedrocs/networks/ub_scn_projCol>).

We could reduce the density of the projected network by both ignoring nodes with very low degrees and weaker ties, which arise between collectors who are less similar in their species bags. For that, we define a filtering routine, which consists of iterating through all entries $A_{i,j}$ of the network's biadjacency matrix and setting them to 0 in case their respective values fall below a user-defined filter threshold ϕ. Acceptable values for the filtering coefficient ϕ range within the interval $[0, 1[$, and represent the minimum similarity score for their species bags that should be observed for a pair of collectors for their tie to be considered relevant. Figure 14(a) illustrates the effect of increasing ϕ on the density of the UB SCN collector projection, under three distinct taxonomic resolutions. Although the three curves show a similar behavior, aggregations at lower taxonomic ranks show more pronounced drops, especially in the region where ϕ ranges from 0 to 0.6. In addition, we verify that higher-rank aggregations lead to denser networks for all possible values of ϕ.

In order to investigate the formation of communities in the collector projection we first filtered collectors ties by setting $\phi = 0.8$. This means only edges holding similarity

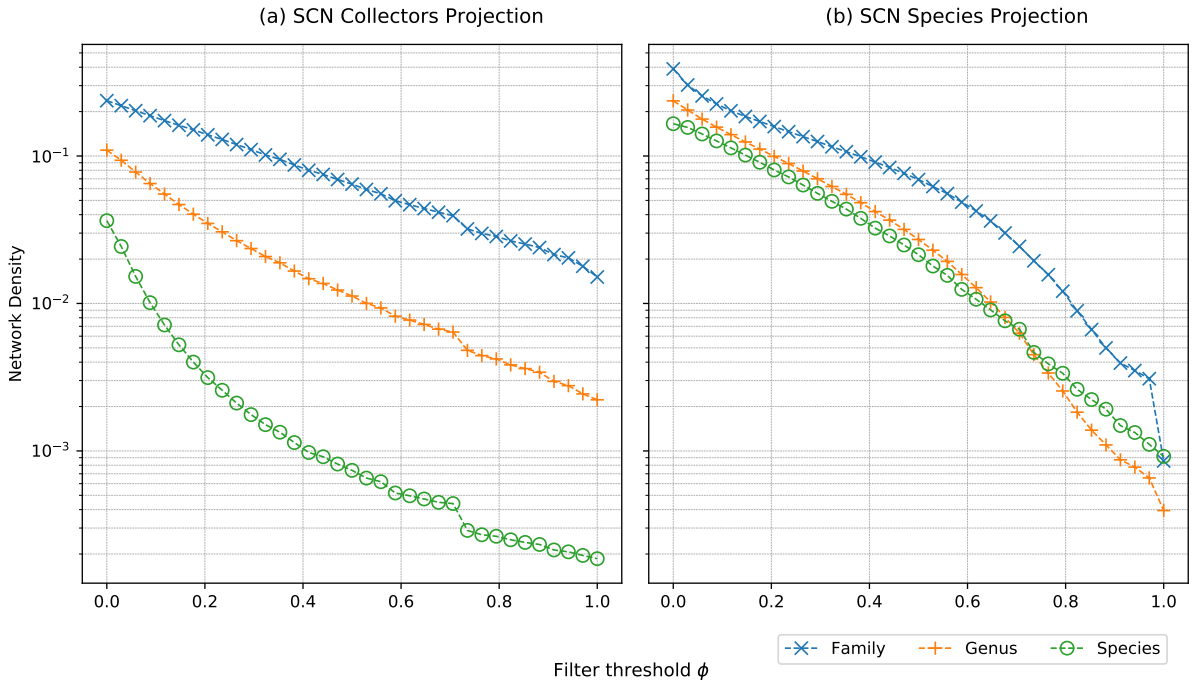

Figure 14 – Reduction in the density of the collectors (a) and species (b) projections of the UB SCN, as a consequence of increasing filtering threshold ϕ. Curves represent the SCN networks aggregated by family (blue), genus (orange), and at the original taxonomic resolution of species (green).

scores equal or higher than 0.8 were kept, while weaker ones were discarded. This process led to the formation of many small connected components (or *islands*), many of which composed of a single collector. An additional step in the filtering process was thus required for removing less relevant connected components from the network. In order to calculate the relevance of a component we defined a simple relevance score which adds up the the values associated to the *count* attribute of each node composing it. We then removed all islands with score lower than 500.

We applied the *Louvain* heuristic method for community detection (BLONDEL et al., 2008), which maximizes modularity scores in the network in successive steps. The result is a partition of the network into modules (communities) within which nodes are more densely connected with each other than with external ones. We also computed the overall taxa composition for each group by summing up the species bags of each of their collectors, which we show as percentages in Table 7.

The collector projection has a total of 20 communities, 7 of which are part of the giant component (Figure 13). Not all collectors from the giant component in Figure 12 are, however, also included in the giant component of the collectors-projected SCN. The largest community in the giant component is *grp* 6, including some of the most important UB collectors, among which *'irwin,hs'*, *'heringer,ep'*, *'anderson,wr'*, *'ratter,ja'*, and *'proenca,ceb'*. These are generalist collectors, having recorded 104 from a total of

Table 7 – Taxa composition for each community as illustrated in Figure 13. The total number of distinct taxa, together with the percentages of the top-3 most recorded ones in each community are given.

community	num of taxa	top taxon	2nd taxon	3rd taxon
0	15	Asteraceae (72.7%)	Fabaceae (9.1%)	Melastomataceae (5.2%)
1	24	Cyperaceae (50.4%)	Fabaceae (9.9%)	Oxalidaceae (7.2%)
2	28	Fabaceae (58.8%)	Myrtaceae (8.6%)	Rubiaceae (6.0%)
3	27	Poaceae (40.4%)	Cyperaceae (12.5%)	Asteraceae (7.5%)
4	37	Myrtaceae (62.0%)	Fabaceae (7.7%)	Asteraceae (4.3%)
5	11	Rubiaceae (82.7%)	Myrtaceae (4.6%)	Fabaceae (3.0%)
6	104	Fabaceae (18.8%)	Rubiaceae (10.1%)	Myrtaceae (8.4%)
7	27	Sematophyllaceae (26.8%)	Hypnaceae (16.0%)	Dicranaceae (13.8%)
8	14	Desmidiaceae (31.4%)	Scenedesmaceae (18.9%)	Chlorellaceae (9.7%)
9	23	Eunotiaceae (35.4%)	Naviculaceae (15.2%)	Pinnulariaceae (11.2%)
10	2	Desmidiaceae (97.5%)	Closteriaceae (2.5%)	
11	18	Poaceae (14.3%)	Fabaceae (13.2%)	Santalaceae (11.1%)
12	4	Amblystegiaceae (63.6%)	Polytrichaceae (22.6%)	Dicranaceae (10.4%)
13	5	Orchidaceae (73.6%)	Myrtaceae (9.8%)	Araceae (7.2%)
14	9	Santalaceae (34.5%)	Loranthaceae (17.4%)	Myrtaceae (13.4%)
15	1	Melastomataceae (100.0%)		
16	6	Fissidentaceae (53.2%)	Calymperaceae (17.3%)	Pottiaceae (16.8%)
17	7	Rubiaceae (20.5%)	Fabaceae (18.8%)	Dicranaceae (16.2%)
18	6	Arecaceae (36.1%)	Myrtaceae (25.4%)	Fabaceae (15.4%)
19	4	Calophyllaceae (38.1%)	Fabaceae (28.1%)	Asteraceae (23.8%)

161 distinct families, being families *Fabaceae* (18.8% of the group's interest), *Rubiaceae* (10.1%) and *Myrtaceae* (8.4%) the most recorded ones (Table 7). The giant component also includes other smaller communities showing predominant interests towards families *Myrtaceae* (*grp* 4), *Fabaceae* (*grp* 2), *Poaceae* (*grp* 3), *Asteraceae* (*grp* 0), *Cyperaceae* (*grp* 1), and *Rubiaceae* (*grp* 5).

Apart from the giant component, there are also 7 other connected components and 6 isolated nodes, each of them making a distinct community. The second largest connected component (*grp*7) includes part of the bryophytes collectors (dashed shape *i* in Figure 12) who are more generalists. The most recorded family is *Sematophyllaceae* (26.8%), followed by *Hypnaceae* (16.0%) and *Dicranaceae* (13.8%). Another part of the bryophytes collectors, who are more specialized towards family *Amblystegiaceae* (64.6%), composes a distinct community (*grp* 12). Algae collectors from shape (*ii*) (Figure 12) are also split into two distinct communities. *João Grando* (*grando,jv*) and '*castelobranco,cw*' are placed in *grp 8*, while *leite,alta*' is an isolated node, forming a community of a single collector (*grp* 10). *Maria das Graças de Souza* (*souza,mgm*), which is included in dashed shape (*iii*) (Figure 12)), is also isolated in the collectors-projected SCN. The last two collectors mentioned above have very particular taxonomic interests, and are represented as isolated nodes in the projection as they show no strong enough species bags similarities with any other collector in the herbarium.

Although Table 7 gives some intuition on the main taxonomic groups that compose the interests of each community, it is not very informative on how community-specific each of those families are. Family *Fabaceae*, for instance, has been relatively more recorded by collectors from community 2 (with 58.8% of the records in that community), although it is also included as a top-3 taxon in communities 0, 1, 4, 5, 6, 11, 17, 18, and 19. This makes it non-specific, as collectors from many communities have demonstrated interest towards recording it, although to distinct extents. In contrast, family *Desmidiaceae* is more community-specific, as it is only recorded by algae collectors composing communities 8 and 10 (both comprising together only 3 collectors). Moreover, it is also relevant to identify groups of families that best split collectors interests in the network by observing which taxa are frequently recorded by common sets of collectors. A better approach for identifying such groupings of taxa is described next, and consists of changing the projection perspective, focusing on the species set instead.

Communities in the species projection. The species projection of the family-aggregated UB SCN network is a one-mode network exclusively representing taxa at the family resolution, with a total of 474 nodes, 43,756 edges, and average node degree of 184. Edges between pairs of taxa are weighted according to their quorum vector similarities, such that strongly connected taxa are those which have been recorded by similar sets of collectors, in similar proportions. The relatively low number of nodes observed in the

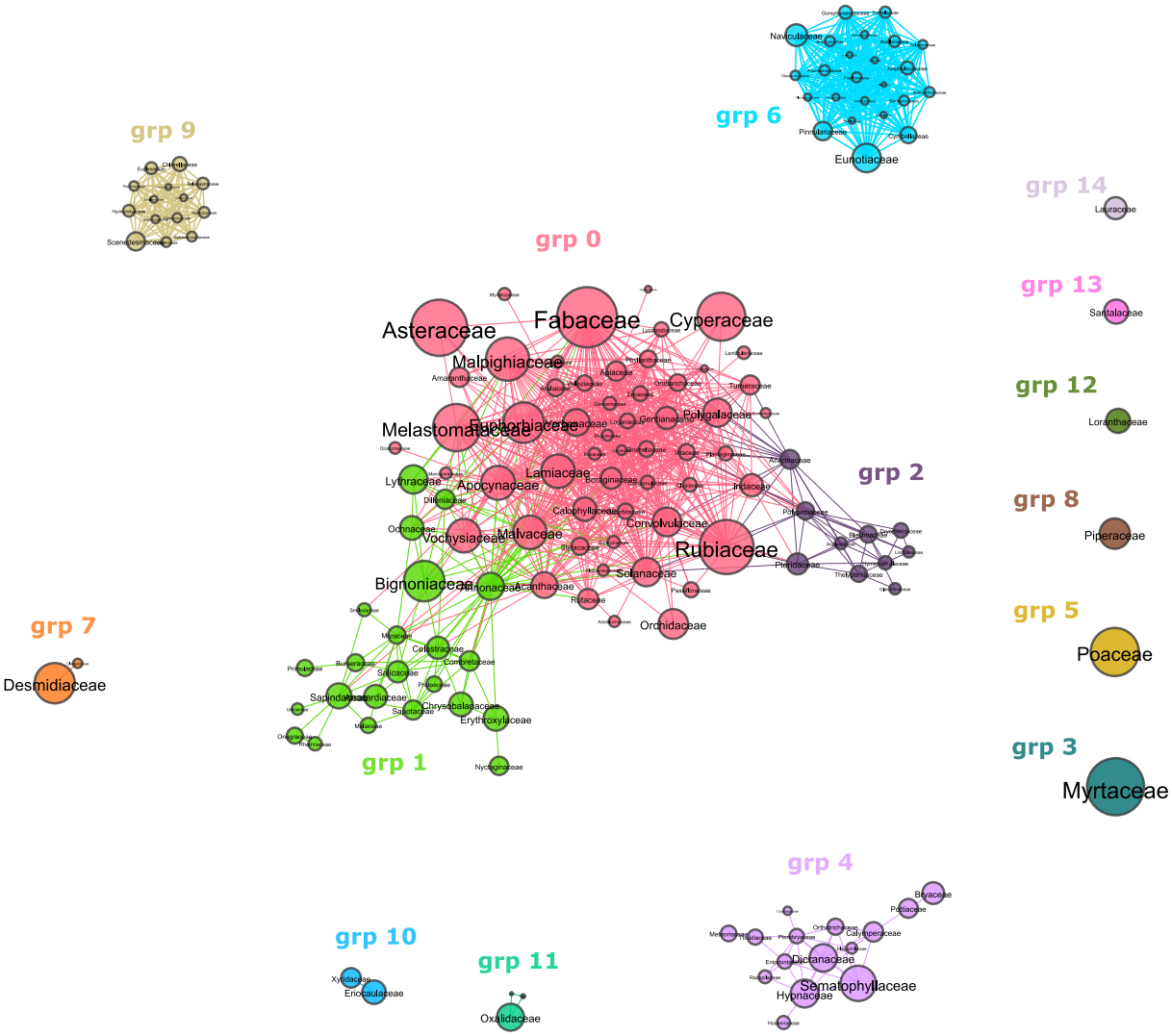

Figure 15 – Communities in the species projection of the family-aggregated UB SCN (from Figure 12). In order to improve visualization, communities with scores lower than 600 were omitted from the figure. Colors are used to distinguish communities and nodes sizes reflect families prevalence in the dataset. Graph layout was computed using the *ForceAtlas2* algorithm (JACOMY et al., 2014). For a better visualization experience refer to the interactive version of the graph (<https://lncc-netsci.github.io/pedrocs/networks/ub_scn_projSp>).

species projection if compared to the collector projection stands from the fact that in this case the taxonomic aggregation process does reduce the total number of nodes in the network. While aggregations at higher taxonomic ranks would result in networks with fewer nodes, network density tends to increase as an effect of the summarization of entities and their ties. It is also important to observe that although the absolute number of nodes and edges in the species-projected network are lower than those from the collector projection, it is still denser than the latter, with a density score of 3.9×10^{-1}.

In order to reduce network density and better visualize the formation of communities we performed a filtering routine to remove less relevant associations of families, analogously

to what we have done for the collector projection. As network density decreases with increasing values for ϕ (Figure 14(b)), we again set the threshold value to 0.8, omitting ties that are weaker than the threshold value. As the process produces many islands, we use the same relevance score from the collector projection for filtering out non-relevant components, though now using a threshold score of 600. This means that islands composed of families whose summed up counts are below 600 are omitted from the figure. The resulting network is shown in Figure 15. Nodes were sized based on the *count* attribute, which indicates how common each family is in the herbarium dataset or, in other words, their *prevalence*. Nodes colors reflect the communities they have been assigned to, using the same approach as we did for the collector projection.

The species projection has a total of 15 communities, three of which belong to the giant component. The largest community is *grp* 0, containing families that are, in general, more widely recorded by herbarium collectors, as it is the case of *Fabaceae*, *Euphorbiaceae*, and *Rubiaceae*. Most of them form a densely-connected structure, containing central families that are all recorded by a very similar set of collectors. However, there are some families included in the same community that are more loosely connected, and would quickly become isolated nodes, if we continued to increase ϕ. This is the case for families *Asteraceae*, *Cyperaceae*, and *Orchidaceae*, which indicates that they are recorded by more specialized collectors. Community 2, which is also part of the giant component, is exclusively composed of fern families (gymnosperms). Some of the included families are more recorded by fern specialists, as it is the case of *Gleicheniaceae*, *Dryopteridaceae*, and *Hymenophyllaceae*. Others as *Pterydaceae*, *Anemiaceae*, and *Polypodiacea* are recorded by collectors who also record flowering plants, such as families *Rubiaceae* and *Iridaceae*.

The remaining 12 communities are all islands, 6 of which formed by isolated nodes. Community 4 contains bryophyte families, mostly recorded by collectors within shape (*i*) in Figure 12. By observing the low clustering and high diameter of this island, we conclude that bryophyte collectors vary significantly regarding their taxonomic interests, as opposed to what we observe for communities 6 and 9, which are very close to being fully connected (cliques). Community 6 contains algae families, mostly diatoms (shape (*iii*)); while communities 7 and 9 are composed by green algae families recorded by collectors within shape (*ii*) in Figure 12. The reason that nodes from shape (*ii*) give origin to two distinct islands in the SCN species projection is that the taxonomic interests of '*leite,alta*' are substantially distinct from those of '*grando,cw*' and '*castelo-branco,*', and thus families *Desmidiaceae* and *Closteriaceae* are separated from the rest of green algae families. Finally, families *Myrtaceae* (*grp* 3), *Poaceae* (*grp*5), and *Piperaceae* (*grp*8) are typically recorded by collectors who are also more specialized in each of those families, and are thus represented as isolated nodes in the network.

Communities discussed above are composed of collectors who share interests in

common taxa and, conversely, taxa that have been sampled by common sets of collectors. Collectors with similar interests, however, do not necessarily work together in field, and thus communities of common interests do not necessarily correspond to coworking teams. The latter can be investigated from CWN models instead, which are structured from collaborative ties between collectors.

4.2.3 The UB Collector CoWorking Network

As the UB CWN was built based on the same set of records as the SCN model explored in the previous section, the number of nodes is $6,768$, equivalent to the number of collectors nodes in the SCN model. A total of $10,391$ edges represent collaborative ties between collectors. The average degree and density for the overall network are, respectively, 3.07 and 4.5×10^{-4}.

Connected components. The UB CWN is composed of a a total $2,991$ connected components, the largest of which (*i.e.*, the giant component c_1) contains 46% of all nodes in the network. Such a relatively low percentage of nodes in the giant component contrasts to most empirical scientific paper-publishing collaboration networks studied by Newman (2001b), with giant components containing as much as 80% to 90% of all nodes. Moreover, only 318 of the connected components in the UB CWN ($c_1, c_2, ..., c_{318}$) are composed of collectors with at least one collaborative tie. The remaining $2,673$ components ($c_{319}, c_{320}, ..., c_{2991}$) are all disconnected nodes (*i.e.* nodes with $k = 0$), which we refer to as *individualist* collectors.

Individualist collectors are those who have never recorded specimens collaboratively—or at least they have not included the names of their collaborators in the records as authors—, and thus are considered to have no structural role in the collaboration network of collectors. They comprise 39.5% of all nodes in the network, and the fact that they lack connections impacts on the overall network density, making it relatively low. In fact, if we instead compute network density by only considering nodes from the giant component c_1, we observe an increase in density from 4.5×10^{-4} to 1.95×10^{-3}. One important example of an individualist collector in the herbarium is '*leite,alta*', with a total of $2,757$ records, none of which recorded as made collaboratively (inset in Figure 16). This comprises 18.14% of all records by individualist collectors. All other $2,672$ individualist collectors have fewer than 400 records each. Moreover, 81.1% of occurrences recorded by individualist collectors have been obtained in Brazil.

How collaborative are collectors? By inspecting the team sizes of all records in the dataset, we find that the average team size for records in the UB dataset is 1.73, as a consequence of the fact that 63% of all records are non-collaborative, *i.e.* recorded by a single collector (Figure 16). The number of records as a function of team size seems to decay

logarithmically. For some reason, which remains unclear to us, team size is constrained

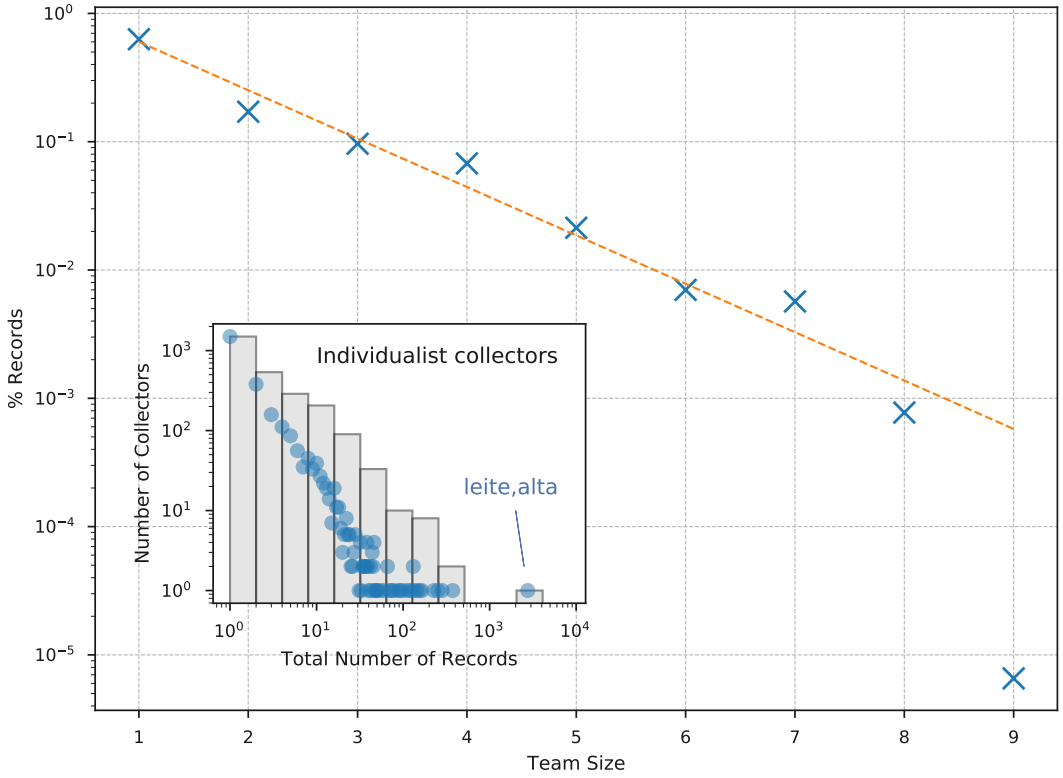

Figure 16 – Percentage of occurrence records in the UB dataset for each team size. The dashed line represents an exponential fit to the data (using team sizes from 1 to 8) using function $f(x) = 10^{\beta_0+\beta_1 x}$ and parameters $\beta_0 = 0.16$ and $\beta_1 = -0.38$. The inset figure shows the number of records individualist collectors typically hold, with records numbers logarithmically binned using base 2.

from 1 to 8 collectors in this dataset (we consider the single record with team size 9 to be an artifact). One hypothesis to be verified is that the data management system adopted by UB might impose a limitation on the maximum number of collectors allowed to be registered for each record or, similarly, on the maximum length of the string containing names of collectors. In fact, we observed that strings containing collectors names are truncated in some records.

Collectors from UB vary substantially regarding their collaborativeness on fieldwork. Whereas few ones have collaborated with a large number of collectors throughout their careers (much more than the average 3.07), many of them hold very few collaborative ties. In fact, almost 40% of them are individualist collectors, having never co-authored a single record. Similarly to the SCN network, the degree distribution of the UB CWN is heavy tailed (Figure 17), and thus is not well fit by a poisson distribution. Such a topology suggests the existence of non-random processes ruling the probability that two collectors get connected, leading to the co-existence of few hubs and many low-degree collectors.

Table 8 ranks the top-20 collectors on weighted degree centrality, using the hyper-

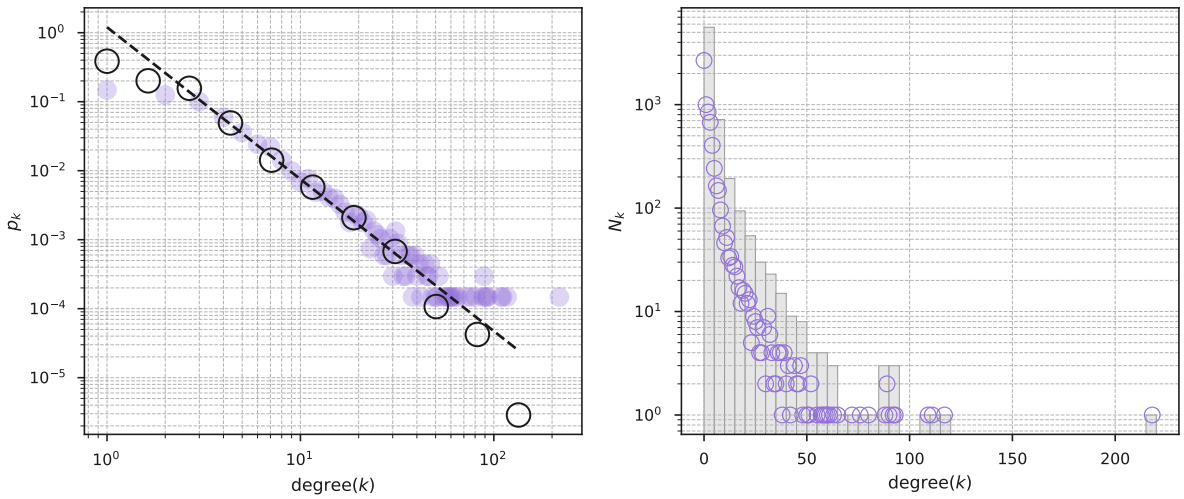

Figure 17 – Degree distribution for the UB herbarium CWN network. Plot (a) shows, using a log-log scale, the probability p_k of finding collectors with each value for k. The black line represents a power law function $p(k) \sim k^{-\alpha}$ fit to the data, with $\alpha = 2.2$. Purple dots are linearly binned, showing the absolute number of nodes with each degree value. Black circles aggregate degree values by using logarithmic binning. Nodes with $k = 0$ were omitted from this plot. Plot (b) shows the absolute number N_k of nodes with each degree (including nodes with $k = 0$) in a lin-log scale. The histogram in the background groups nodes degrees in bins of size 5.

bolic weighting rule. As discussed in Section 3.3, this rule assigns weights to collector ties by considering the sizes of the teams in which collectors collaborated, such that higher weights are assigned to ties which were originated from collaborations involving smaller teams (minimum size of 2). The weighted degree of a node is given by the sum of the weights of all its edges. The top collector is '*irwin,hs*', with a weighted degree score of $5,696$. Although he has not collaborated with many distinct collectors (a total of 39), his position on the rank reflects the absolute number of collaborative records authored by him ($6,169$), even though this only comprises 34.1% of all his records. The second collector in the rank, '*proenca,ceb*', has less than a third of *irwin*'s absolute number of records, although 88.3% of them are collaborative. With a total of 218 distinct collaborators, she would be placed at the first position if the rank were built based on the non-weighted degree score (k) instead, whilst *'irwin,hs'*, in contrast, would be placed at the 43^{rd} position. Another cases worth mentioning are those of '*grando,jv*' and '*castelobranco,cw*', both of which have necessarily and exclusively collaborated to each other in all their records (*i.e.*, team size is 2 for all records). As a consequence their weighted degree scores k_w are equivalent to the total number of records authored by them ($2,256$), whereas their non-weighted degree scores are $k = 1$. Thus, the non-weighted degree score k ranks collectors based on their total number of distinct collaborators, irrespective of the number of times each association has been observed, whilst, on the other hand, the weighted degree score k_w takes into account

Table 8 – Top-20 collectors of the UB CWN, ranked by the weighted degree k_w centrality score. Both the absolute number of records and the percentage of those which have been collected collaboratively (involving at least two collectors) are given for each collector. The degree centrality k represents the number of collaborative ties linking a collector to distinct collaborators, whereas the weighted degree is computed with the hyperbolic weighting function.

collector	num of records	% collaborative	k	k_w
irwin,hs	18065	34.1	39	5696.0
proenca,ceb	4803	88.3	218	4203.7
faria,jeq	4687	82.9	117	3881.0
souza,rr	3885	99.3	37	3835.5
santos,rrb	3587	94.5	41	3382.3
munhoz,cbr	3191	82.8	109	2493.2
zanatta,mrv	2364	95.8	50	2264.0
castelobranco,cw	2256	100.0	1	2256.0
grando,jv	2256	100.0	1	2256.0
eiten,g	3046	61.4	33	1865.5
amaral,ag	1825	99.5	37	1815.0
projetobiodiversidadebp	1780	100.0	19	1780.0
mendes,vc	1696	99.9	89	1695.0
fonseca,sf	1610	99.9	18	1607.0
camara,peas	2076	79.8	47	1486.0
harley,rm	2564	66.0	90	1455.8
carvalhosilva,m	1635	98.5	58	1436.0
eiten,lt	1262	99.6	14	1256.5
mello,trb	1247	100.0	21	1247.0
soares,aer	1557	73.0	29	1135.0

the relevance of each tie, considering that collectors collaborating in smaller teams are relatively more strongly tied.

The *betweenness* centrality metric is also frequently used in social network analytics for ranking collectors who act as "bridges", intermediating a considerable fraction of shortest paths between pairs of nodes in the network. We compute betweenness centrality of a node v by making $c_B(v) = \sum_{s,t} \frac{\sigma(s,t|v)}{\sigma(s,t)}$, where V is the node set; $\sigma(s,t)$ is the number of shortest paths between nodes $s, t \in V$ (both s and v different from v); and $\sigma(s,t|v)$ is the number of shortest paths between s and v that are pass through v (BRANDES, 2008). The 4 collectors with the highest betweenness centrality scores in the UB CWN are '*proenca,ceb*' (0.38), '*faria,jeq*' (0.18), '*mendes,vc*' (0.14), and '*ratter,ja*' (0.13). The first three collectors are also included in the degree centrality rank, in Table 8, although '*ratter,ja*' is not (he occupies the 21^{th} position, with $k = 111$). By inspecting Figure 18, we verify that in fact all those collectors are in strategic positions. *Carolyn E. Proença* (*proenca,ceb*) is located at the center of the network, while '*mendes,vc*' and '*faria,jeq*' also

interconnect nodes from many distinct communities. *James A. Ratter* (*ratter,ja*) is a very important node bridging a very relevant group (the green one, around '*irwin,hs*') to the remainder of the network.

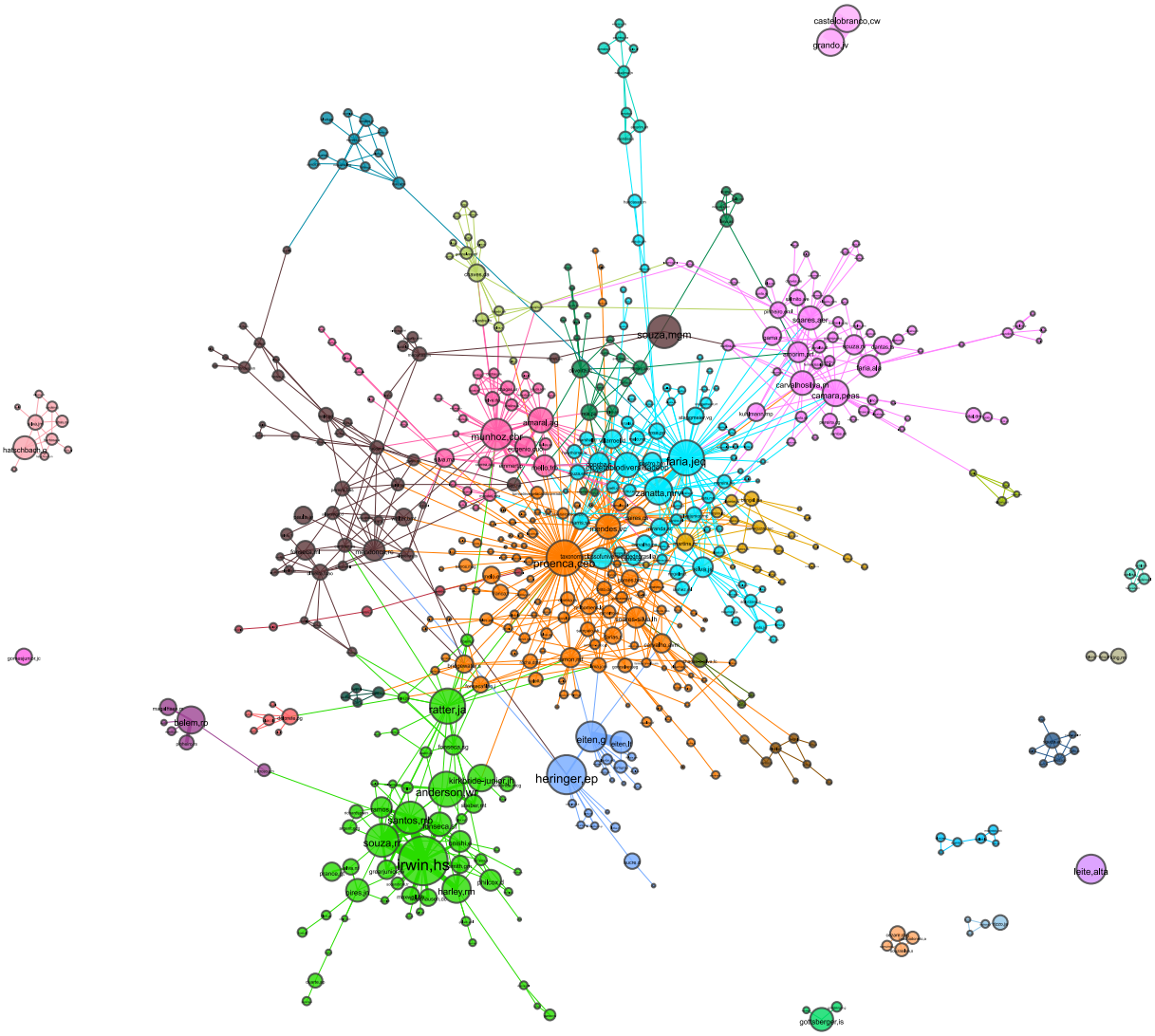

Figure 18 – Coworking groups in a subset of the UB CWN. A total of 30 distinct communities are differentiated by color. Sizes of nodes are proportional to the total number of records of collectors in the dataset. The strength of the ties between each pair of collectors is represented as links thickness. Graph layout was computed using the *ForceAtlas2* algorithm (JACOMY et al., 2014). For a better visualization experience refer to the interactive version of the graph (<https://lncc-netsci.github.io/pedrocs/networks/ub_cwn>).

Coworking groups. One important aspect that can be investigated from the topological structure of the UB CWN is the formation of communities of collectors who co-author specimen records, which we refer to as **coworking groups**. In order to detect such groups we have applied the same algorithm for community detection (BLONDEL et al., 2008) as we did for the SCN projections. We detected a total of 30 distinct communities, 11 of which are detached from the giant component. For graph visualization, we first performed

a filtering routine, similar to the one we applied for plotting the SCN projections. We first filtered out weaker edges (those with hyperbolic weight lower than 10), which resulted in many islands. We assigned scores to each island by summing up the values of the *count* attribute of all nodes composing them. Islands with scores lower than 600 were omitted. The resulting graph has 545 nodes, 1,158 edges, and an average degree of 4.25 (Figure 18).

Some of the communities we found in the CWN are formed by collectors who also compose tight communities of interests in the SCN projection (Figure 13). One such example is the bryophytes research group, which includes '*camara,peas*', '*carvalhosilva,m*', and '*soares,aer*'. Collectors from this coworking group (colored in purple and located in the upper-right region of the giant component in Figure 18) not only mostly collaborate with members from the same group in field, but are also interested on recording the specific taxonomic group of bryophytes. Another similar example is the coworking group of collectors of herbaceous plants (colored in pink), including '*munhoz,cbr*', '*amaral,ag*', '*mello,trb*', and '*eugenio,cuo*', although this group is apparently more open to collaborating with non-members.

Other coworking communities include collectors who happen to be strongly connected with collaborative ties in the CWN despite having distinct taxonomic interests, being weakly connected in the SCN projection. One such case is that of '*faria,jeq*' and '*zanatta,mrv*', both having recorded a total of 1,524 specimens together (comprising 64.5% of all records by *zanatta,mrv* and 32.5% of all records by *faria,jeq*), and thus belonging to the same coworking group (colored in light blue, Figure 18). However, as the overall composition of their species bags are substantially distinct ('*faria,jeq*' is more interested on family *Myrtaceae*, whereas '*zanatta,mrv*' has a more generalist recording profile), they have been included in distinct interest communities (*grp4* and *grp6*, in Figure 13).

There are also cases of collectors who, despite having very similar taxonomic interests (*i.e.*, being strongly tied in the SCN projection), seldom or never collaborate in field, and are therefore not included in the same coworking groups. For instance, collectors '*chaves,da*' and '*bringel,jba*' have the majority of their species bags composed of family *Asteraceae* (54.3% for '*bringel,jba*' and 78.5% for '*chaves,da*'), being both included in the same interest community (*grp0*, Figure 13). However, they have never collaborated on specimens recording, not being adjacent and belonging to distinct communities in the CWN (Figure 18). For another example, the community around '*hatschbach,g*' (at the leftmost region of Figure 18) is an island in the CWN, as it incorporates no relevant collaborations with collectors from the giant component. However, given their taxonomic interests at the *family* rank, members of that community are included in the giant component of the SCN (*grp*6).

We consider some possible explanations for why groups of collectors with very similar taxonomic interests would not also form collecting coworking teams. First, as

coworking relations are based on the physical recording of organisms taking place at some location and date, they can only possibly exist between collectors whose recording activities overlap both in geographical and temporal spaces. As temporal and geographical distances between collectors increase, they consequently become more unlikely of developing collaborations, constrained by the limited lifespan of their own careers.

Second, non-professional relationships could also facilitate or hinder the establishment of new coworking ties between individuals. Positive relationships (*e.g.* friendship) could lead two collectors to collaborate irrespective of their own taxonomic interests, for instance, if they simply mean to be companions of each other and help their friend on field collections. On the other hand, negative relationships, such as rivalry, can prevent potentially fruitful coworking relationships to be established between collectors who record very similar taxonomic groups at the same place and time. Moreover, people have limited information on the activities of other collectors and research groups, and thus might not collaborate simply as a consequence of not being acquainted of each other.

Last, we might fail to grasp a complete picture of how interest groups are truly structured in the SCN depending on the taxonomic resolution we adopt while inspecting it. To exemplify this issue, consider the scenario in which we detect a subset of collectors composing a community in a family-aggregated projected SCN, such as that in Figure 18. At the taxonomic resolution of family, we would naturally assume that all those collectors have very similar taxonomic interests, as the reason they are included in the same community is that they have very similar "family signatures" distinguishing them from the remaining collectors. However, if we increased the taxonomic resolution of the model (*e.g.* to genus), we might observe the community to subdivide, each of them containing collectors who, although interested in the same family, collect distinct sets of organisms belonging to it. Subcommunities that emerge in higher-taxonomic resolution SCNs are thus not detectable in lower-resolution aggregations.

Temporal evolution of the CWN. Finally, we investigated the temporal evolution of communities in the UB CWN by rendering a sequence of frames covering the period from 1955 to 2017, each spanning a period of 7 years (Figure 19). Frames aggregate and display collaborations (edges) occurring within the respective periods, and only collectors showing significant activity within each period are plotted. We consider the activity of a collector to be significant if the number of occurrences recorded within the period comprises at least 5% of his/her absolute number of records. Although our proposed CWN model does not incorporate the temporal dimension in its structure, we were able to build the frames by querying the UB species occurrence dataset from which the network was constructed.

Three pioneer collector groups, which we refer to as the **first generation** of botanists (with most of their recording activities concentrated from 1955 to 1982), were

Figure 19 – Temporal evolution of the UB CWN, shown in Figure 18. Each frame corresponds to a period of 7 years, and only collectors with significant activities in the respective periods are drawn in each frame. Nodes sizes are proportional to the total number of records of each collector up to the end of each period; and the thickness of ties are proportional to how intensively collectors have collaborated during each period (using the hyperbolic weighting rule). The layout of nodes in the figure is the same as from Figure 18. For a better visualization experience refer to the animation in <https://lncc-netsci.github.io/pedrocs/video/ub_cwn_evo.mp4>

historically remarkable for having largely contributed to the early formation of the UB herbarium (frames 1 to 4 in Figure 19). The first group (colored in blue-gray) includes '*heringer,ep*', '*eiten,g*', and '*eiten,lt*'. *Ezechias Paulo Heringer* (*heringer,ep*) is regarded as one of the founders of the UB herbarium, having led the first exploratory expeditions for sampling the flora of the Federal District (WALTER; CAVALCANTI; FILGUEIRAS,

2001). During his earlier activities (around 1955) he has collaborated with '*castellanos,a*', a renowned Argentinian botanist who visited the region during the construction of the city of Brasília. Later in his career (1969 to 1975), '*heringer,ep*' has also teamed up with '*eiten,g*', recording specimens mainly in the states of Minas Gerais and Goiás. *George Eiten* (*eiten,g*) is also considered a historically relevant collector for the herbarium, having contributed with an extensive set of plant records, mainly from the Cerrado biome (GOMES; WALTER; FRANCO, 2012). *Eiten*'s known strong collaboration with his first wife, '*eiten,lt*', is captured by our network model, as shown in the first 3 frames in Figure 19.

A second group which also heavily contributed to the formation of the herbarium is the one composed of '*irwin,hs*', '*anderson,wr*', '*ratter,ja*', '*harley,rm*', '*pires,jn*', among others (colored in green). *Howard Irwin* (*irwin,hs*) was the head of a series of exploratory expeditions to the Central Brazilian Highlands from 1964 to 1971, which were part of a collaborative program between the New York Botanical Garden (NYBG) and the University of Brasília (IRWIN et al., 1996). Since the beginning of the of the expeditions, we observe a strong association between *irwin,hs* and two of his field assistants (*souza,rr* and *santos,rrb*), shown in frames 2 and 3 from Figure 19. Other important collectors, including '*harley,rm*', '*fonseca,sf*', and '*onishi,e*', have later on joined the team around '*irwin,hs*', contributing to the expedition in the period from 1969 to 1975. It is also noticeable the association between '*anderson,wr*', '*kirkbride-junior,jh*' and '*stieber,mt*', in the time frame from 1969 to 1975. *William Anderson* (*anderson,wr*) was a botanist from the NYBG, who continued the expeditions initiated by his colleague '*irwin,hs*' on 1971 (IRWIN et al., 1996), together with '*kirkbride-junior,jh*', a graduate student at the same institution. *James Ratter* (*ratter,ja*) was another important collector of the herbarium, having collaborated with at least two different generations of botanists. We therefore consider him to be a "temporal bridge", connecting two important periods of the time-evolving network that do not overlap temporally. *George Eiten* (*eiten,g*) has also collaborated with second-generation collectors (*simon,mf*, *lima,cjm*, and *carvalho,avm*) within a period from 1997 to 2003, after at least 21 years with no recording activities. The third group (colored in purple) is smaller, composed of 6 collectors including *magalhaes,m* and *belem,rp*, the latter a Brazilian student who has mostly collected at the state of Bahia from 1955 to 1968.

The **second generation** of collectors are those more active within the period from 1983 to 2003 (frames 5 to 7), and are placed more towards the central region of the network. Two relevant collectors from the earlier second generation are '*castelobranco,cw*', '*grando,jv*', and '*leite,alta*', all algae collectors with most of their activity ranging from 1983 to 1989. As we previously pointed out, '*castelobranco,cw*' and '*grando,jv*' have collaborated on all their records, whereas '*leite,alta*' is an individualistic collector, having no collaborative records at all. Other important second-generation collectors are '*proenca,ceb*' and '*souza,mgm*', with most intensive activities concentrated from 1990 to 2003. Whereas '*souza,mgm*' has almost

exclusively collected within the period from 1990 to 1996 and with very few collaborations, '*proenca,ceb*' has extended has extended her collecting activities until current date (2017), having collaborated with a very high number of collectors, including the first-generation collector '*ratter,ja*'. *Cássia Munhoz* (*munhoz,cbr*) starts her activities around 1990, being one of the founders of a coworking community around herbaceous plants.

Finally, we assign collectors whose activities are concentrated from 2004 to 2017 to the **third generation**, the most prominent one being '*faria,jeq*'. As shown in the graph, '*faria,jeq*' has intensively collaborated with important collectors from distinct coworking communities, including '*zanatta,mrv*', '*staggmeier,vg*', '*amorim,prf*', '*proenca,ceb*', '*amaral,ag*', '*bringel,jba*', and '*ribeiro,aro*', mainly in the period from 2011 to 2017. Other important collectors from this generation are '*amaral,ag*', '*eugenio,cuo*', and '*mello,trb*', having collaborated more intensely from 2004 to 2010. They form a community around the second-generation collector '*munhoz,cbr*', who was also their academic advisor during graduate research. An important community which emerged during this period is the bryophytes lab team. *Paulo Eduardo Câmara* (*camara,peas*) initiated his recording activities in $1997 - 2003$, followed by '*carvalhosilva,m*' and '*soares,aer*' in $2004 - 2010$. These three collectors formed the base of the group, which grew significantly in the period from 2011 to 2017, as the group was joined by several students.

5 Conclusion and Perspectives

In this dissertation, we proposed the conceptual basis of a new approach for describing the assemblage of biological collections as a social process, driven by the taxonomic interests of contributor collectors as well as their social interactions. In this context, we provided methods for structuring species occurrence data from biological collections into two main classes of network models, each giving distinct perspectives on the recording behavior of collectors. **Species-Collector Networks** (SCNs) model interest relations between collectors and taxa they record, whilst **Collector CoWorking Networks** (CWNs) represent collaborative ties between collectors co-authoring records of specimens. As a case study, we demonstrated the use of our network models by exploring the species occurrence dataset of the University of Brasília Herbarium (UB). Using the social network analytics framework (BARBIER, 2011; STORK, 2015) as a theoretical foundation, we explored structural properties of the studied networks as well as investigated the formation of communities of collaboration and common interests. We also assessed the distinctiveness of collectors regarding their taxonomic interests, their collaborativity with others, and the temporal evolution of collaborative recording in the herbarium. Although in this study we specifically discuss SCNs and CWNs in the context of scientific biological collections, the same ideas here exposed can be also extended to other communities, such as those of nature observers (*e.g.* wildlife photographers, bird watchers) and citizen scientists.

We believe our network models provide the structural basis for a more realistic understanding on how collector and taxonomic biases arise in biological collections. As stated by Marin and Wellman (2011), network-based approaches allow analysts (*i*) to investigate the effects of interactions between individuals on shaping their own behaviors, rather than simply comparing static attributes of individuals within a population; and (*ii*) to investigate the formation of non-homogeneous communities, composed of individuals interacting with their groups at varying levels of commitment. We consider that these two aspects are particularly relevant in the context of biological collections.

First, collectors often start their careers being supervised by one or more experienced collectors. As it naturally happens in many social systems, the behavior of individuals can be strongly influenced by others at more privileged positions, and this is likely to be the case for collector communities as well. A network-based approach would allow us to investigate, for instance, how the collecting behavior and taxonomic interests of novice collectors are shaped by their association with more experienced ones. Moreover, depending on its position on the network, a collector can interact with multiple groups of collectors at different extents, thus assuming the role of influencer in some cases while being influenced in others. The influential power of a collector depends not only on the absolute

number of connections it holds with others, but also on how strongly it intermediates other connections, how close it is to every other collector in the network, and how influential are its own connections. All these aspects can be assessed using well-known network centrality metrics (including degree, betweenness, closeness, and eigenvector centralities), and could be used for investigating which collectors are the most relevant for shaping the taxonomic composition of a biological collection.

Second, although collectors often define their own taxonomic interests and expertises in terms of natural or functional groups of organisms, those are not necessarily the groups that best split the interests of collectors in the dataset of a given collection. In addition, collectors (even the most specialized ones) are not restricted towards exclusively recording organisms of their expertises, nor they have uniform interest towards all of those organisms. In this context, SCNs provide the structure for discovering groupings of taxa that are better for characterizing and differentiating collectors based on their taxonomic interests (communities in the species projection); and for investigating associations of collectors with groups of taxa in a non-discrete manner, allowing collectors to be linked to taxa at multiple groups and at different intensities. In fact, while some groups of taxa are more specifically recorded by distinctive communities of specialized collectors, others are recorded by a wider range of collectors, with diverse taxonomic interests. For instance, the SCN of the UB herbarium suggests that collectors specialized in families *Myrtaceae*, *Poaceae*, or *Piperaceae* form communities of interests which are in fact more distinctive, as those families are relatively poorly recorded by collectors who are not specialists in each of them (see Figure 15). In contrast, other families, such as *Fabaceae*, *Melastomataceae*, and *Rubiaceae*, are more widely recorded in the herbarium, thereby they do not form a tight community of interest.

The quality of our network models strongly depends on the quality of the species occurrence dataset that is used to build them, more specifically on the fields containing the names of the collectors (*recordedBy*, in TDWG standards) and the taxonomic identity of the specimen (*scientificName*, in TDWG standards) of each record. During this study, we have explored occurrence datasets from other herbaria other than the UB, including the RB (at the Rio de Janeiro Botanical Garden) (FORZZA; DALCIN, 2018), the MBML (Mello Leitão Herbarium) (FERNANDES, 2018), and the Hemilio Goeldi Museum Herbarium (VIANA, 2016). Nevertheless, we decided to only use the dataset from the UB in our case study because of its relatively high quality, specifically for the two aforementioned fields. In all other herbaria, the collectors field (*recordedBy*) was particularly problematic. Our hypothesis is that the low quality of this field is associated with its low value for most uses of species occurrence data, implying that not much effort has been employed by data curators towards improving the data quality of this field. While imprecise taxonomic determinations in the *scientificName* field would also lead to low quality networks, this field is critical for many other applications of occurrence data, and thus improving its

quality has been extensively pursued by the biodiversity informatics community. The most common and impacting issues associated with the collectors field were: (*i*) using inconsistent delimiter characters for separating the names of each collector in a record, leading to many non-atomized names and consequently to the existence of nodes in the network that represent more than one collector; (*ii*) registering collectors names using inconsistent naming conventions, which makes it hard to systematically interpret what are the component parts of a name; (*iii*) using multiple name variations for a collector, leading to collectors being represented by more than one node in the resulting network; and (*iv*) only including the name of the first collector in records (and eventually aggregating all secondary collectors under the expression '*et al.*'), which is interpreted as an absence of collaborative ties and thus does not contribute for the formation of edges in CWNs. Constructing the network models from a low-quality dataset can therefore introduce several semantic imprecisions.

Our network models, as proposed in this dissertation, also have a set of limitations, which should be addressed in the future. First, as the networks are built in a single step using a single dataset (which is usually provided by a single institution), they only represent a partial view of the real interests and collaborations that collectors have accumulated during their careers. For obtaining a more holistic representations, a mechanism for dynamically joining multiple occurrence datasets—and eventually other types of data—should be incorporated to the models. Although one might argue that multiple datasets can be simply merged before they are passed to the constructors of the models, that would still consist of a one-step construction, requiring the availability of all data at the first place. In addition, if any other dataset were to be incorporated in the future, the model would need to be entirely reconstructed, and all necessary preprocessing routines would have to be re-executed in each one of the previous datasets. We believe that the most challenging aspect of joining multiple datasets would be addressing the entity resolution problem across different datasets from multiple sources. A possible solution would be to map entities in each dataset to unique identifiers (such as those from the ORCID[1] initiative or the id in the Lattes platform,[2] the latter widely adopted by the Brazilian scientific community) using a crowdsourcing strategy, described later in this chapter.

Another important limitation of our network models is that they are static and non-spatialized (*i.e.*, they are temporally and geographically invariant), and limited to representing relationships of a single type each. This implies that relationships modeled in both networks are assumed to occur irrespective of temporal and geographic dimensions, which is clearly limiting for the phenomena they model. As the careers of collectors have limited lifespans, they can only possibly collaborate with others if their activity periods overlap in time. In addition, both coworking (between collectors) and interest (involving a

[1] <https://orcid.org/>
[2] <http://lattes.cnpq.br/>

collector and a taxon) relationships derive from collecting events—each happening at a determined geographic location and at some point in time—, being thus temporally and spatially constrained. Incorporating these two dimensions to our models is also central for capturing network evolution in their structure. As in many other social systems, relationships in SCNs and CWNs change in time, as new ties are constantly formed while older ones are broken. It is reasonable to consider that collectors interact with distinct groups of people throughout their careers, assuming distinct roles in each relationship. For instance, we hypothesize that earlier in their careers, collectors are more likely to assume relationships and have their interests influenced by their academic supervisors or other collectors who are more experienced. On the other hand, relationships assumed by collectors later in their careers tend to be the opposite, as they assume the role of the more experienced collector (and thus, the influencer). Further, depending on the stage of a collector's career and the groups of collectors he/she interacts at that moment, we might observe substantial shifts in his/her taxonomic interests (while changes in his/her taxonomic interests can also lead to collaborating with different groups). Other factors can also influence the patterns of recording activity of a collector, including oscillations in the availability of financial resources for field expeditions, and changes in his/her residence location.

As many applications using our network models would need to combine the perspectives of both SCNs and CWNs, we also recognize the importance of adopting an unifying model for seamlessly integrating the two networks into a single structure. The requirement for such a model is that it represents two types of connections (interest and coworking) and two distinct sets of nodes (collectors and species), besides incorporating the temporal and geographical dimensions into its structure. Although the concept of *multilayer networks* provides a solution for modeling dynamic complex systems with many aspects of connectivity, literature around this topic is still incipient, with many proposals though little consensus on the best way to represent them (KIVELÄ et al., 2014). In this context, Wehmuth, Ziviani and Fleury (2016) have introduced the concept of *Multiaspect Graphs* (MAGs) as a graph extended abstraction for high order networks that operates on a structure that is proved to be isomorphic to traditional directed graphs. By using the structure of a MAG, the set of vertices, layers, time instants, and geographic locations can be then represented as 4 distinct *aspects*; and edges as 8-tuples, composed of pairs of elements of each aspect. Moreover, key properties and algorithms that have been widely used for analyzing directed graphs are also extended to MAGs, which makes them a relatively simple representation for higher-order graphs (WEHMUTH; FLEURY; ZIVIANI, 2015).

Finally, for a direction towards further developments of this work, we indicate the incorporation of geographic and temporal dimensions to the networks as top priorities. We also conclude our text by briefly presenting some possible new perspectives of ideas

for applications (some of them still very rough) that could potentially make use of our network models. Many examples below assume that temporal and geographic dimensions are already included.

1. Profiling collectors. An important improvement towards a systematic understanding of the roles, interests, and behaviors of collectors in a biological collection is grouping them into discrete profiles. Profiles aggregate semantic value to the model, as they allow domain analysts to summarize the complex variety of collector features into general classes that are more comprehensible to them. For instance, analysts may be interested on inferring *academic roles* of collectors (*e.g.* professor, student, or field assistant), whether collectors are currently *active* or *retired*, or still whether they are *experienced* or *novice*. Similarly, collectors can be classified as *innovators* if they collect taxonomic groups that have never been recorded by others in the collection, or *followers* if they follow the interests of others; as *visitors* if they contribute to the collection in bursty patterns, or *residents* if they contribute to it a regular basis; *specialists* or *generalists*, regarding how specific their taxonomic interests are; *regionalists* or *travelers*, regarding their interests to collect at many distinct localities. Compositions of each of these aspects, which we refer to as *characteristics* of collectors, are used to define *collector profiles*. Considering that the interests and collecting behavior of collectors change in time as they get more experienced, profiles are naturally dynamic. A simplistic approach for representing such variations would be to associate profile timelines with collectors, composed of multiple discrete events documenting profile changes. Profiling collectors can be generalized to a network problem known as the **node classification problem** (BHAGAT; CORMODE; MUTHUKRISHNAN, 2011). Starting from a subset of nodes that are previously labeled (profiled), the goal is to train a machine learning classifier that learns which compositions of features lead to each profile. Next, in an iterative process, non-profiled nodes have their profiles predicted based on their attributes. Network structure can be useful in this process, as it allows the propagation of labels (profiles) among collectors, using their positional features and patterns of association.

2. Contextual enrichment of occurrences. One of the main complexities of characterizing bias in occurrence records is that they are typically obtained in an opportunistic way, without the adoption of systematic sampling designs. Specimens are also collected in a high variety of contexts, from *botany field classes* mainly composed of naive students to *big survey projects*, involving many teams of expert collectors. Surveys can also be characterized as being *focal*, if individuals from a specific taxonomic group are thoroughly searched; or *generalist*, if the goal is to document the diversity of organisms at a location as comprehensively as possible. Also some records may result from a *herbarium exchange*. Considering that different contexts lead collectors to behave and collaborate differently

during collection activities, enriching occurrence records with contextual information could make them more comparable, potentially helping to characterize biases and sampling efforts. One idea worth investigating is whether the composition of collectors associated with a record convey contextual information about it. For instance, observing groups composed of many novice collectors associated with one or two who are very experienced, recording in areas relatively well explored by others, could indicate that those records have been obtained in the context of a field course. Records in which all collectors are substantially experienced but with distinct interests, on the other hand, could indicate the context of exploratory surveys. Moreover, additional attributes of the recorded species can also be inferred from the composition of collectors. For instance, Steege et al. (2011) observed that experienced collectors tend to explore a higher diversity of vegetation types during expeditions, and thus record more rare species than novice collectors. Thus, the likelihood that a species represented in a record is rare can apparently be correlated with the profiles of collectors associated to it. Assigning contexts to occurrence records could be modeled as a classification problem, analogous to that for assigning profiles to collectors. From a subset of records for which an analyst previously provides some class of context, the algorithm learns patterns of collectors profiles, and predicts the contexts of the remainders.

3. Crowdsourcing the validation of collectors identities. Given the complexity of resolving the identities of names in the collector field of occurrence datasets, one possible solution would be to use a network-based *collaborative validation*, in which the record validation task is distributed through many collaborators. The general idea of the method consists of using the structure of a CWN, initially built from the non-validated dataset, for propagating a message requesting the collaboration of collectors themselves as information validators. Starting from an initial subset of influential collectors, whose identities are already resolved, collectors recursively may resolve the identities of as few as 5 of their most acquainted colleagues and, in sequence, forward them the collaboration request message. The process of resolving names consists of assigning unique identifiers to them, such as the ORCID or Lattes id. Assuming that the probability that a receiver does not ignore the message depends on its esteem towards the sender, a central requirement for an efficient diffusion of the messages in the network is to start with an initial set of vertices (collectors) that not only occupy central positions in the network, but which are also influential in their respective communities and, moreover, are available and willing to collaborate with the validation process. This is analogous to studying the dynamics of contagion in network systems (GIBSON, 2005) and the network seeding problem, *i.e.* properly choosing a set of nodes to start an efficient diffusion.

4. Formation of teams of specialists. One more possible application of our network models is for recommending teams of specialists for composing biodiversity projects.

In many cases, experts are required not only to have a substantial expertise in their respective areas, but also to be prone to collaborate with others. A generalization of this problem is known as the **expert team formation**, and can be subdivided in two main steps (LAPPAS; LIU; TERZI, 2011). The first one (the *expert location problem*) consists of assessing the level of expertise of individuals in a given set of topics, or for performing a given task. Each individual is assigned a set of skills which characterize their expertises. Once experts have been located, the second step consists of forming teams of experts such that collaborators can effectively communicate and work collectively for achieving a common goal. Locating potentially effective teams of collectors could be useful for better *planning biodiversity surveys*, maximizing the productivity of the expeditions while reducing associated costs. Given the overall goals of a survey, a list of qualified collectors could be obtained from SCNs, while arranging teams with them should considerate how effectively they have collaborated in the past (from CWNs). Another example is the evaluation of conservation status of species for the *elaboration of red lists*. This task requires the collaboration of teams of specialists, more specifically for validating species occurrence data and for evaluating profiles that are assigned to each species (MARTINELLI et al., 2013). Experienced collectors can be important data validators, as they tend to develop good intuition about the biological communities at locations where they collect (NOSS, 1996). In order to better distribute the validation workload over the collaborators, occurrence records should be directed to collectors according to their profiles, considering their experience and taxonomic interests; and according to the regions where they have collected. In this context, potential validators could be indicated by our network models according to their profiles (see *profiling collectors*).

5. Assessing the accuracy in taxonomic determinations. One issue with the *determinavit system* of biological collections (*i.e.* a system where experts review the taxonomic determinations assigned to species) is that the certainty of identifications are not always documented. Depending on the experience of the person who assigns a taxonomic identity to a specimen, it can be more or less reliable. In this context, Chapman and Speers (2005) have proposed that collections should incorporate in their databases an indication of the certainty of identifications, by using a system of ranks of expertise. For instance, a record could be identified by a world expert in that taxa, by a regional expert, a non-expert, or by the person who has collected the specimen himself/herself. Ranking identifiers manually, however, is no trivial task. A network model, analogous to the SCNs we propose in this work, can be defined for modeling the interests of *identifiers* towards *species*. This new class of network (Species-Identifier Network) could be used for helping profiling identifiers, similarly to what we have described for collectors, thus associating a certain identifier reputation to the reliability of the identification.

Bibliography

ALBERT, R.; BARABÁSI, A.-L. Statistical mechanics of complex networks. *Reviews of Modern Physics*, v. 74, n. 1, p. 47–97, 2002. ISSN 0034-6861. Disponível em: <http://dx.doi.org/10.1103/RevModPhys.74.47%5Cnpapers2://publication/doi/10.1103/RevModPhys.74.47%5Cnhttp://stacks.iop.org/1478-3975/1/i=3/a=006?key=crossref.7e041937ef77358d60afe44f40925c5b%5Cnhttp://link.aps.org/doi/10.1103/RevModPhys.74.47>. Cited 3 times in pages 34, 67, and 70.

ALBERT, R.; JEONG, H.; BARABÁSI, A.-L. Diameter of the world-wide web. *Nature*, v. 401, n. 6749, p. 130–131, 1999. ISSN 00280836. Cited 2 times in pages 17 and 32.

ARAÚJO, M. B.; GUISAN, A. Five (or so) challenges for species distribution modelling. *Journal of Biogeography*, v. 33, n. 10, p. 1677–1688, 2006. ISSN 03050270. Cited in page 17.

ARAÚJO, M. B. et al. Using species co-occurrence networks to assess the impacts of climate change. *Ecography*, Blackwell Publishing Ltd, v. 34, n. March, p. 897–908, 12 2011. ISSN 09067590. Disponível em: <http://doi.wiley.com/10.1111/j.1600-0587.2011.06919.x>. Cited in page 43.

BARABÁSI, A. L. *Network Science*. Cambridge University Press, 2016. 475 p. ISBN 1107076269. Disponível em: <http://networksciencebook.com/>. Cited 2 times in pages 17 and 34.

BARABÁSI, A.-L.; ALBERT, R. Emergence of Scaling in Random Networks. *Science*, v. 286, n. 5439, 1999. Disponível em: <http://science.sciencemag.org/content/286/5439/509>. Cited 2 times in pages 33 and 67.

BARBIER, G. *Social Network Data Analytics*. [s.n.], 2011. ISBN 978-1-4419-8461-6. Disponível em: <http://www.springerlink.com/index/10.1007/978-1-4419-8462-3http://link.springer.com/10.1007/978-1-4419-8462-3>. Cited 2 times in pages 17 and 92.

BASCOMPTE, J. Networks in ecology. *Basic and Applied Ecology*, v. 8, n. 6, p. 485–490, 2007. ISSN 14391791. Cited in page 43.

BASCOMPTE, J.; JORDANO, P. Plant-Animal Mutualistic Networks: The Architecture of Biodiversity. *Annual Review of Ecology, Evolution, and Systematics*, Annual Reviews, v. 38, n. 2007, p. 567–593, 12 2007. ISSN 1543-592X. Disponível em: <http://www.annualreviews.org/doi/10.1146/annurev.ecolsys.38.091206.095818http://www.jstor.org/stable/30033872>. Cited in page 17.

BEBBER, D. P. et al. Big hitting collectors make massive and disproportionate contribution to the discovery of plant species. *Proceedings of the Royal Society B: Biological Sciences*, v. 279, n. 1736, p. 2269–2274, 2012. ISSN 0962-8452. Disponível em: <http://rspb.royalsocietypublishing.org/cgi/doi/10.1098/rspb.2011.2439>. Cited in page 25.

BHAGAT, S.; CORMODE, G.; MUTHUKRISHNAN, S. Node classification in social networks. *Social network data analytics*, p. 115–148, 2011. Cited in page 96.

BINGHAM, H. et al. The Biodiversity Informatics Landscape: Elements, Connections and Opportunities. *Research Ideas and Outcomes*, Pensoft Publishers, v. 3, p. e14059, 6 2017. ISSN 2367-7163. Disponível em: <http://riojournal.com/articles.php?id=14059>. Cited 2 times in pages 17 and 43.

BISBY, F. A. The quiet revolution:biodiversity informatics and the internet. *Science*, v. 289, n. September, p. 2309–2311, 2000. Cited in page 16.

BLONDEL, V. D. et al. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, v. 2008, n. 10, 2008. ISSN 17425468. Cited 2 times in pages 77 and 86.

BONACICH, P. Power and Centrality: A Family of Measures. *American Journal of Sociology*, v. 92, n. 5, p. 1170–1182, 1987. ISSN 0002-9602. Disponível em: <http://www.journals.uchicago.edu/doi/10.1086/228631>. Cited in page 39.

BORGATTI, S. P.; EVERETT, M. G. Network analysis of 2-mode data. *Social Networks*, v. 19, n. 3, p. 243–269, 1997. ISSN 03788733. Cited in page 49.

BORGATTI, S. P.; HALGIN, D. S. Analyzing Affiliation Networks. In: *The Sage Handbook of Social Network Analysis*. [S.l.: s.n.], 2015. p. 417–433. ISBN 9781452258225. Cited 3 times in pages 41, 69, and 75.

BORRETT, S. R.; MOODY, J.; EDELMANN, A. The rise of Network Ecology: Maps of the topic diversity and scientific collaboration. *Ecological Modelling*, v. 293, p. 111–127, 2014. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0304380014001136>. Cited 2 times in pages 42 and 43.

BRANDES, U. On variants of shortest-path betweenness centrality and their generic computation. *Social Networks*, v. 30, n. 2, p. 136–145, 2008. ISSN 03788733. Cited in page 85.

CEBALLOS, G. et al. Accelerated modern human – induced species losses: entering the sixth mass extinction. *Science Advances*, v. 1, n. e1400253, p. 1–5, 2015. ISSN 2375-2548. Cited in page 14.

CHAPMAN, A. D. Principles and methods of data cleaning - primary species and species-occurrence data. *Report for the Global Biodiversity Information Facility*, n. version 1.0, p. 72, 2005. Cited 3 times in pages 16, 30, and 31.

CHAPMAN, A. D. Uses of primary species-occurrence data, version 1.0. *Report for the Global Biodiversity Information Facility*, p. 111, 2005. Cited in page 14.

CHAPMAN, A. D.; SPEERS, L. Principles of Data Quality. *Global Biodiversity*, Copenhaen, n. Chrisman, p. 58, 2005. Disponível em: <http://www2.gbif.org/DataQuality.pdf>. Cited 4 times in pages 22, 24, 27, and 98.

CHRISMAN, N. R. The role of quality information in the long-term functioning of a geographic information system. *Cartographica: The International Journal for Geographic Information and Geovisualization*, v. 21, p. 79–88, 1984. Cited in page 27.

CHRISMAN, N. R. The error component in spatial data. *Geographical Information Systems: Principles and Applications1*, p. 165–174, 1991. Cited 2 times in pages 16 and 25.

DALCIN, E. C. Data Quality Concepts and Techniques Applied to Taxonomic Databases. *Life Sciences*, n. February, p. 266, 2005. Disponível em: <http://dalcin.org/eduardo/downloads/edalcin_thesis_submission.pdfhttp://www.dalcin.org/eduardo/downloads/edalcin_thesis_submission.pdf>. Cited 2 times in pages 27 and 31.

DARU, B. H. et al. Widespread sampling biases in herbaria revealed from large-scale digitization: Sampling bias in herbarium specimens. *New Phytologist*, v. 217, n. 2, p. 939–955, 2017. ISSN 14698137. Disponível em: <http://dx.doi.org/10.1101/165480>. Cited 6 times in pages 16, 17, 25, 26, 44, and 68.

DAVIS, A.; GARDNER, B. B.; GARDNER, M. R. *Deep South: A social anthropological study of caste and class*. Univ of South Carolina Press, 1941. 557 p. ISBN 1570038155. Disponível em: <http://books.google.com/books?id=Q3b9QTOgLFcC&pgis=1>. Cited in page 41.

ERDOS, P.; RÉNYI, A. On random graphs, I. *Publicationes Mathematicae (Debrecen)*, v. 6, p. 290–297, 1959. Cited in page 33.

FAUST, K. Animal social networks. In: *The SAGE handbook of social network analysis.* London: SAGE Publications, 2011. cap. 11, p. 148–66. Cited in page 43.

FERNANDES, H. d. Q. B. MBML-Herbario - Herbário Mello Leitão. *Version 1.39. Instituto Nacional da Mata Atlântica. Occurrence Dataset https://doi.org/10.15468/z8djaf accessed via GBIF.org on 2018-05-19*, 2018. Cited in page 93.

FORZZA, R.; DALCIN, E. RB - Rio de Janeiro Botanical Garden Herbarium Collection. *Version 84.146. Instituto de Pesquisas Jardim Botanico do Rio de Janeiro. Occurrence Dataset https://doi.org/10.15468/7ep9i2 accessed via GBIF.org on 2018-05-19*, 2018. Disponível em: <https://doi.org/10.15468/7ep9i2>. Cited in page 93.

FUNK, V. A. et al. Testing the use of specimen collection data and Gis in Biodiversity exploration and consection decision making in Guyana. *Biodivers Conserv*, v. 8, n. 727, p. 751, 1999. ISSN 1572-9710. Cited in page 17.

GBIF. What is GBIF? *The Global Biodiversity Information Facility*, 2018. Disponível em: <https://www.gbif.org/what-is-gbif>. Cited 2 times in pages 23 and 57.

GBIF.org. *Data processing*. 2018. Disponível em: <https://www.gbif.org/data-processing>. Cited in page 61.

GIBSON, D. R. Concurrency and commitment: Network scheduling and its consequences for diffusion. *Journal of Mathematical Sociology*, v. 29, n. 4, p. 295–323, 2005. ISSN 0022250X. Cited in page 97.

GOMES, S. M.; WALTER, B. M. T.; FRANCO, A. C. George Eiten - 1923 - 2012. *Acta Botanica Brasilica*, v. 26, n. 4, p. 725–726, 2012. Cited in page 90.

GRAHAM, C. H. et al. New developments in museum-based informatics and applications in biodiversity analysis. *Trends in Ecology and Evolution*, v. 19, n. 9, p. 497–503, 2004. ISSN 01695347. Cited 2 times in pages 17 and 27.

GROOM, Q. J.; REILLY, C. O.; HUMPHREY, T. Herbarium specimens reveal the exchange network of British and Irish botanists, 1856–1932: New Journal of Botany: Vol 4, No 2. *New Journal of Botany*, v. 4, n. 2, p. 95–103, 2014. ISSN 2042-3489. Disponível em: <http://www.maneyonline.com/doi/pdfplus/10.1179/2042349714Y.0000000041>. Cited 4 times in pages 17, 29, 43, and 66.

GUIMERÀ, R.; SALES-PARDO, M.; AMARAL, L. A. N. Module identification in bipartite and directed networks. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, v. 76, n. 3, p. 1–8, 2007. ISSN 15393755. Cited in page 75.

GUISAN, A. et al. What matters for predicting the occurrences of trees: Techniques, data, or species' characteristics? *Ecological Monographs*, v. 77, n. 4, p. 615–630, 2007. ISSN 00129615. Cited in page 24.

HIJMANS, R. J. et al. Assessing Genebank Potatoes the of Geographic Representativeness the Collections : of Bolivian. *Conservation Biology*, v. 14, n. 6, p. 1755–1765, 2000. ISSN 1523-1739. Cited in page 26.

HORTAL, J.; LOBO, J. M.; JIMENEZ-VALVERDE, A. Limitations of biodiversity databases: Case study on seed-plant diversity in Tenerife, Canary Islands. *Conservation Biology*, v. 21, n. 3, p. 853–863, 2007. ISSN 08888892. Cited in page 25.

INGS, T. C. et al. Ecological networks - Beyond food webs. *Journal of Animal Ecology*, v. 78, n. 1, p. 253–269, 2009. ISSN 00218790. Cited in page 42.

IRWIN, S. et al. Memories of The N e w York Botanical Garden ,. v. 48, n. 3, p. 365–371, 1996. Cited in page 90.

JACOBY, D. M. P.; FREEMAN, R. *Emerging Network-Based Tools in Movement Ecology*. 2016. 301–314 p. Disponível em: <http://linkinghub.elsevier.com/retrieve/pii/S0169534716000264>. Cited in page 43.

JACOMY, M. et al. ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software. *PLoS ONE*, v. 9, n. 6, p. 1–12, 2014. ISSN 19326203. Cited 5 times in pages 71, 73, 76, 80, and 86.

JAMES, S. A. et al. *Herbarium data: Global biodiversity and societal botanical needs for novel research: Global*. 2018. Cited in page 15.

KELLING, S. et al. Data-intensive Science: A New Paradigm for Biodiversity Studies. *BioScience*, University of California Press, v. 59, n. 7, p. 613–620, 7 2009. ISSN 0006-3568. Disponível em: <https://academic.oup.com/bioscience/article-lookup/doi/10.1525/bio.2009.59.7.12>. Cited in page 15.

KEMP, C. Museums: The endangered dead. *Nature*, v. 518, n. 7539, p. 292–294, 2015. ISSN 14764687. Cited in page 14.

KIVELÄ, M. et al. Multilayer networks. *Journal of Complex Networks*, v. 2, n. 3, p. 203–271, 2014. ISSN 20511329. Cited in page 95.

KRAMER-SCHADT, S. et al. The importance of correcting for sampling bias in MaxEnt species distribution models. *Diversity and Distributions*, v. 19, n. 11, p. 1366–1379, 2013. ISSN 13669516. Cited in page 17.

LAMBIOTTE, R.; AUSLOOS, M. Uncovering collective listening habits and music genres in bipartite networks. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, v. 72, n. 6, 2005. ISSN 15393755. Cited 2 times in pages 42 and 49.

LAPPAS, T.; LIU, K.; TERZI, E. A survey of algorithms and systems for expert location in social networks. *Social Network Data Analytics*, p. 215–241, 2011. Cited in page 98.

LINDEMAN, R. L. The trophic-dynamic aspect of ecology. *Ecology*, v. 23, p. 399–417, 1942. Cited in page 42.

LOMOLINO, M. V. Conservation biogeography. *Frontiers of Biogeography: new directions in the geography of nature*, p. 293–296, 2004. Cited in page 24.

MARIN, A.; WELLMAN, B. Social Network Analysis: An introduction. *The SAGE handbook of social network analysis*, v. 11, 2011. Cited in page 92.

MARTINELLI, G. et al. Extinction Risk Assessments of the Brazilian Flora. *Livro Vermelho da Flora do Brasil*, 2013. Cited in page 98.

MCNEILL, J. *International code of nomenclature for algae, fungi and plants (Melbourne code) : adopted by the Eighteenth International Botanical Congress Melbourne, Australia, July 2011*. Koeltz Scientific Books, 2012. 208 p. ISBN 9783874294256. Disponível em: <http://www.iapt-taxon.org/nomen/main.php>. Cited in page 20.

MEYER, C.; WEIGELT, P.; KREFT, H. Multidimensional biases, gaps and uncertainties in global plant occurrence information. *Ecology Letters*, 2016. ISSN 14610248. Cited in page 26.

MICHENER, W. K.; JONES, M. B. *Ecoinformatics: Supporting ecology as a data-intensive science*. Elsevier, 2012. 88–93 p. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/22240191>. Cited in page 15.

MILGRAM, J. T. An Experimental Study of the Small World Problem. *Sociometry*, v. 32, n. 4, p. 425 – 443, 1969. Disponível em: <http://www.jstor.org/stable/2786545>. Cited in page 33.

MUNHOZ, C. B. R. et al. UB - Herbário da Universidade de Brasília. *Version 1.35. Universidade de Brasília. Occurrence Dataset. https://doi.org/10.15468/caq5no accessed via GBIF.org on 2018-03-16*, 2018. Disponível em: <https://doi.org/10.15468/caq5no>. Cited in page 57.

NELSON, B. W. et al. Endemism centres, refugia and botanical collection density in Brazilian Amazonia. *Nature*, v. 345, n. 6277, p. 714–716, 1990. ISSN 0028-0836. Disponível em: <http://www.nature.com/doifinder/10.1038/345714a0>. Cited 3 times in pages 16, 17, and 26.

NEWBOLD, T. Applications and limitations of museum data for conservation and ecology, with particular attention to species distribution models. *Progress in Physical Geography*, v. 34, n. 1, p. 3–22, 2010. ISSN 0309-1333. Disponível em: <http://journals.sagepub.com/doi/10.1177/0309133309355630>. Cited 2 times in pages 17 and 23.

NEWBOLD, T. et al. Global effects of land use on local terrestrial biodiversity. *Nature*, v. 520, n. 7545, p. 45–50, 2015. ISSN 14764687. Cited in page 15.

NEWMAN, M. E. J. Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, v. 64, n. 1, p. 016132, 2001. ISSN 1063-651X. Disponível em: <http://link.aps.org/doi/10.1103/PhysRevE.64.016132>. Cited in page 53.

NEWMAN, M. E. J. The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences*, v. 98, n. 2, p. 404–409, 2001. ISSN 0027-8424. Disponível em: <http://www.pnas.org/cgi/doi/10.1073/pnas.98.2.404http://www.pnas.org/content/98/2/404.abstract>. Cited in page 82.

NEWMAN, M. E. J. The structure and function of complex networks. *SIAM Review*, v. 45, n. 2, p. 167–256, 2003. Cited 2 times in pages 38 and 41.

NEWMAN, M. E. J. Coauthorship networks and patterns of scientific collaboration. *Proceedings of the National Academy of Sciences*, v. 101, n. Supplement 1, p. 5200–5205, 2004. ISSN 0027-8424. Disponível em: <http://www.pnas.org/cgi/doi/10.1073/pnas.0307545100>. Cited 2 times in pages 42 and 70.

NEWMAN, M. E. J. Networks: An Introduction. *Networks: An Introduction*, p. 1–784, 2010. ISSN 1578-1275. Cited 2 times in pages 17 and 34.

NEWMAN, M. E. J.; PARK, J. Why social networks are different from other types of networks. *Physical Review E*, p. 1–8, 2003. ISSN 1063-651X. Disponível em: <http://arxiv.org/abs/cond-mat/0305612%0Ahttp://dx.doi.org/10.1103/PhysRevE.68.036122>. Cited in page 40.

NOSS, R. The naturalists are dying off. *Conservation Biology*, v. 10, n. 1, p. 1–3, 1996. ISSN 08888892. Cited in page 98.

NUALART, N. et al. Assessing the Relevance of Herbarium Collections as Tools for Conservation Biology. *Botanical Review*, v. 83, n. 3, 2017. ISSN 00068101. Cited in page 14.

ODUM, H. T. Primary Production in Flowing Waters. *Limnology and oceanography*, v. 1, n. 1, p. 102–117, 1956. Cited in page 42.

PEDROSA, F.; GALLANT, J. *FLORESCER: Herbário da Universidade de Brasília*. 2018. Disponível em: <http://florescer.unb.br/janela4.html>. Cited 2 times in pages 56 and 62.

PETERSON, A. T.; SOBERÓN, J.; KRISHTALKA, L. A global perspective on decadal challenges and priorities in biodiversity informatics. *BMC Ecology*, BioMed Central, v. 15, n. 1, p. 1–9, 2015. ISSN 14726785. Cited in page 16.

PYKE, G. H.; EHRLICH, P. R. Biological collections and ecological/environmental research: A review, some observations and a look to the future. *Biological Reviews*, v. 85, n. 2, p. 247–266, 2010. ISSN 14647931. Cited in page 14.

RAMASCO, J. J.; DOROGOVTSEV, S. N.; PASTOR-SATORRAS, R. Self-organization of collaboration networks. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, v. 70, n. 3 2, 2004. ISSN 15393755. Cited in page 42.

REDDY, S.; DÁVALOS, L. M. Geographical sampling bias and its implications for conservation priorities in Africa. *Journal of Biogeography*, Blackwell Science Ltd, v. 30, n. 11, p. 1719–1727, 11 2003. ISSN 03050270. Disponível em: <http://doi.wiley.com/10.1046/j.1365-2699.2003.00946.x>. Cited in page 25.

REES, T. Taxamatch, an algorithm for near ('fuzzy') matching of scientific names in taxonomic databases. *PloS one*, Public Library of Science, v. 9, n. 9, p. e107510, 1 2014. ISSN 1932-6203. Disponível em: <http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0107510>. Cited in page 31.

REICHMAN, O. J.; JONES, M. B.; SCHILDHAUER, M. P. Challenges and Opportunities of Open Data in Ecology. *Science*, v. 331, n. 6018, p. 703–705, 2011. ISSN 0036-8075. Disponível em: <http://www.sciencemag.org/cgi/doi/10.1126/science.1197962>. Cited in page 15.

SAAD, Y. *Iterative Methods for Sparse Linear Systems*. Society for Industrial and Applied Mathematics (SIAM, 3600 Market Street, Floor 6, Philadelphia, PA 19104), 2003. xviii+528 p. ISSN 1570579X. ISBN 0898715342. Disponível em: <https://books.google.com.br/books?hl=en&lr=&id=h9nwszYPblEC&oi=fnd&pg=PR2&dq=Y.+Saad.+Iterative+Methods+for+Sparse+Linear+Systems.+SIAM,+Philadelpha,+PA,+second+edition,+2003.&ots=PW-5J5HULQ&sig=OkEGCLNVXQhSqu35RZjYLm2HO_o#v=onepage&q&f=false>. Cited in page 35.

SILVA, L. A Data Mining Approach for Standardization of Collectors Names in Herbarium Database. *IEEE Latin America Transactions*, v. 14, n. 2, p. 805–810, 2016. ISSN 15480992. Cited in page 29.

SILVERTOWN, J. A new dawn for citizen science. *Trends in Ecology and Evolution*, v. 24, n. 9, p. 467–471, 2009. ISSN 01695347. Cited in page 24.

SOBERÓN, J.; PETERSON, A. T. Biodiversity informatics: managing and applying primary biodiversity data. *Philosophical Transactions of the Royal Society B: Biological Sciences*, v. 359, n. 1444, p. 689–698, 2004. ISSN 0962-8436. Disponível em: <http://rstb.royalsocietypublishing.org/cgi/doi/10.1098/rstb.2003.1439>. Cited 3 times in pages 16, 24, and 27.

STEEGE, H. ter et al. A model of botanical collectors' behavior in the field: Never the same species twice. *American Journal of Botany*, v. 98, n. 1, p. 31–37, 2011. ISSN 00029122. Cited 3 times in pages 17, 26, and 97.

STORK, M. G. *The Sage Handbook of Social Network Analysis*. [S.l.: s.n.], 2015. 22–24 p. ISSN 0038-0385. ISBN 9781452258225. Cited 2 times in pages 17 and 92.

SUNDERLAND, M. E. Computerizing natural history collections. *Endeavour*, Elsevier Ltd, v. 37, n. 3, p. 150–161, 2013. ISSN 01609327. Disponível em: <http://dx.doi.org/10.1016/j.endeavour.2013.04.001>. Cited in page 22.

TALBERT, C. et al. *Data management challenges in species distribution modeling*. 2013. 31–40 p. Disponível em: <http://sites.computer.org/debull/A13dec/p31.pdf>. Cited in page 15.

THÉBAULT, E. Identifying compartments in presence-absence matrices and bipartite networks: Insights into modularity measures. *Journal of Biogeography*, v. 40, n. 4, p. 759–768, 4 2013. ISSN 03050270. Disponível em: <http://doi.wiley.com/10.1111/jbi.12015>. Cited in page 43.

TULLOCH, A. I. T. et al. Dynamic species co-occurrence networks require dynamic biodiversity surrogates. *Ecography*, v. 39, n. 12, p. 1185–1196, 2016. ISSN 16000587. Cited in page 43.

VEIGA, A. K.; JR., E. A. C.; SARAIVA, A. M. Data Quality Control in Biodiversity Informatics: The Case of Species Occurrence Data. *IEEE LATIN AMERICA TRANSACTIONS*, v. 12, n. 4, p. 683–693, 6 2014. ISSN 1548-0992. Disponível em: <http://ieeexplore.ieee.org/document/6868870/http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6868870>. Cited 2 times in pages 27 and 31.

VEIGA, A. K. et al. A conceptual framework for quality assessment and management of biodiversity data. *PLoS ONE*, 2017. ISSN 1932-6203. Cited 2 times in pages 16 and 27.

VIANA, P. Museu Paraense Emilio Goeldi Herbarium. *https://doi.org/10.15468/igjr8k accessed via GBIF.org on 2018-05-19*, 2016. Disponível em: <https://doi.org/10.15468/igjr8k>. Cited in page 93.

WAKE, D. B.; VREDENBURG, V. T. Are we in the midst of the sixth mass extinction? A view from the world of amphibians. *Proceedings of the National Academy of Sciences*, v. 105, n. Supplement 1, p. 11466–11473, 2008. ISSN 0027-8424. Disponível em: <http://www.pnas.org/cgi/doi/10.1073/pnas.0801921105>. Cited in page 14.

WALLS, R. L. et al. Semantics in support of biodiversity knowledge discovery: An introduction to the biological collections ontology and related ontologies. *PLoS ONE*, v. 9, n. 3, p. 1–13, 2014. ISSN 19326203. Cited in page 24.

WALTER, B. M. T.; CAVALCANTI, T. B.; FILGUEIRAS, T. S. Coletas botânicas no Distrito Federal. *Flora do Distrito Federal, Brasil*, Embrapa Recursos Genéticos e Biotecnologia, Brasília, p. 43–56, 2001. Cited in page 90.

WATTS, D. J.; STROGATZ, S. H. Collectivedynamics of 'small-world' networks. *Nature*, v. 393, n. 6684, p. 440–442, 1998. ISSN 00280836. Disponível em: <http://www.nature.com/nature/journal/v393/n6684/abs/393440a0.html%5Cnhttp://www.nature.com/nature/journal/v393/n6684/pdf/393440a0.pdf%5Cnhttp://www.nature.com/doifinder/10.1038/30918>. Cited 2 times in pages 33 and 41.

WEHMUTH, K.; FLEURY, E.; ZIVIANI, A. MultiAspect Graphs : Algebraic representation and algorithms. *arXiv*, p. 1–61, 2015. Cited in page 95.

WEHMUTH, K.; ZIVIANI, A.; FLEURY, E. On MultiAspect Graphs. *Theoretical Computer Science*, v. 651, p. 50–61, 2016. Cited in page 95.

WIECZOREK, J. et al. Darwin core: An evolving community-developed biodiversity data standard. *PLoS ONE*, v. 7, n. 1, p. e29715, 1 2012. ISSN 19326203. Disponível em: <http://dx.plos.org/10.1371/journal.pone.0029715>. Cited in page 23.

WILCOVE, D. S. et al. Quantifying Threats to Imperiled Species in the United States. *BioScience*, v. 48, n. 8, p. 607–615, 1998. ISSN 00063568. Disponível em: <https://academic.oup.com/bioscience/article-lookup/doi/10.2307/1313420>. Cited in page 14.

Appendix

APPENDIX A – Collectors IDs

Table 9 – Names and IDs of some of the main collectors from the University of Brasília Herbarium. Names contain hyperlinks in the digital version of this document.

id	name
amaral,ag	Aryanne Gonçalves Amaral
anderson,wr	William Russell Anderson
belem,rp	Romeu P. Belém
bridgewater,s	Sam G. M. Bridgewater
bringel,jba	João Bernardo de Azevedo Bringel Júnior
camara,peas	Paulo Eduardo Aguiar Saraiva Câmara
carvalhosilva,m	Micheline Carvalho Silva
castelobranco,cw	Christina Wyss Castelo Branco
chaves,da	Daniel Augusto Chaves
dantas,ts	Tamara Silva Dantas
eiten,g	George Eiten
eiten,lt	Liene Teixeira Eiten
eugenio,cuo	Chesterton Ulysses Orlando Eugênio
faria,ala	Allan Laid Alkimim Faria
faria,jeq	Jair Eustáquio Quintino de Faria Júnior
fonseca,sf	Sidney F. da Fonsêca
gama,r	Renato Gama Dias Neto
grando,jv	João Vademar Grando
harley,rm	Raymond Mervyn Harley
heringer,ep	Ezechias Paulo Heringer
irwin,hs	Howard Samuel Irwin
kirkbride,mcg	Maria Cristina Garcia Kirkbride
kirkbride-junior,jh	Joseph Harold Kirbride Jr.
leite,alta	Ana Lúcia Tostes de Aquino Leite
magalhaes,m	Geraldo Mendes Magalhães
martins,ds	Drielle dos Santos Martins
mello,trb	Thiago de Roure Bandeira Mello
munhoz,cbr	Cássia Beatriz Rodrigues Munhoz
oliveira,rc	Regina Célia de Oliveira
oliveira,rir	Roni Ivan Rocha de Oliveira
onishi,e	Eunice Onishi
pinheiro,eml	Eliana Marília Lima Pinheiro

pires,jn	João Murça Pires
proenca,ceb	Carolyn Elinores Barnes Proença
ratter,ja	James Alexander Ratter
ribeiro,aro	André Rodolfo de Oliveira Ribeiro
santos,rrb	Raimundo Reis dos Santos
simon,mf	Marcelo Fragomeni Simon
soares,aer	Abel Eustáquio Rocha Soares
souza,mgm	Maria das Graças Machado de Souza
souza,rr	Raimundo Souza
souza,rv	Ronaldo Viveiros de Sousa
staggmeyer,vg	Vanessa Graziele Staggemeier
stieber,mt	Michael Thomas Stieber
zanatta,mrv	Maria Rosa Vargas Zanatta