DOI: 10.5585/GEAS.V8I3.15777

METHODOLOGICAL PROPOSAL FOR THE USE OF PATENTS IN RESEARCH OF INTERORGANIZATIONAL COLLABORATION NETWORKS FOR INNOVATION

Ricardo Cruz Gomes[1]; Geciane Silveira Porto[2]; Priscila Rezende da Costa[3]

ABSTRACT: The application of Social Network Analysis (SNA) techniques to investigate interorganizational networks for the development of innovations has attracted growing scientific and empirical interest. In this sense, patents are recognized as an important measure of innovation with valuable information, which are publicly available worldwide, allowing them to be applied to investigate these collaborative innovation networks. However, the information on patents made available by patent authorities in different countries varies in format, which could compromise the application of SNA and the interpretation of analyses. To solve this problem, this study aims to select, develop and present techniques for organizing and preparing a large amount of patent data, allowing data enrichment and enabling the construction and analysis of these networks. This study contributes to the dissemination of studies and applications of SNA in interorganizational innovation networks, supporting researchers and professionals.

KEYWORDS: social network analysis; networks; interorganizational; patent; data preparation.

RESUMO: A aplicação de ARS para investigar redes de parcerias para o desenvolvimento de inovações apresenta grande interesse teórico e empírico. Nesse sentido as patentes são reconhecidas como importante indicador de inovação dispondo de informações valiosas, que estão publicamente disponíveis no mundo todo permitindo sua aplicação para investigar essas redes de colaboração para inovação. Entretanto, as informações sobre patentes disponibilizadas pelas autoridades de patentes dos diversos países, geralmente apresentam diferentes tipos de formatos, o que pode comprometer a aplicação de ARS e a interpretação da análise. Para resolver esse problema, esse estudo se propõe a selecionar, desenvolver e apresentar técnicas para organizar e preparar um grande volume de dados sobre patentes, viabilizando a construção e análise dessas redes. Essa pesquisa contribui para difusão de estudos e aplicações de ARS em redes de inovação interorganizacionais, apoiando pesquisadores e profissionais.

PALAVRAS-CHAVE: análise de redes sociais; redes; interorganizacional; patentes; preparação de dados.

RESUMEN: La aplicación de ARS para investigar redes de asociaciones para el desarrollo de innovaciones es de gran interés teórico y empírico. En este sentido, las patentes son reconocidas como un indicador importante de innovación con información valiosa, que está disponible públicamente en todo el mundo, lo que permite su aplicación para investigar estas redes de colaboración para la innovación. Sin embargo, la información sobre patentes proporcionada por las autoridades de patentes en diferentes países generalmente tiene diferentes tipos de formatos, lo que puede comprometer la aplicación de ARS y la interpretación del análisis. Para resolver este problema, este estudio tiene como objetivo seleccionar, desarrollar y presentar técnicas para organizar y preparar un gran volumen de datos de patentes, lo que permite la construcción y el análisis de estas redes. Esta investigación contribuye a la difusión de estudios y aplicaciones de ARS en redes de innovación interorganizacionales, apoyando a investigadores y profesionales.

PALABRAS-CLAVE: análisis de redes sociales; redes; interorganizacionales; patentes; preparación de datos.

* Agradecimento à FAPESP pelo financiamento da pesquisa (processo 2017/25364-6).

1 Ricardo Cruz Gomes - Doutorando do Programa de Pós-Graduação em Administração de Organizações da Faculdade de Economia, Administração e Contabilidade da USP Ribeirão Preto , E-mail: ricruzgomes@usp.br

2 Geciane Silveira Porto - Docente da Faculdade de Economia, Administração e Contabilidade da USP Ribeirão Preto e Pesquisadora do Instituto de Estudos Avançados da USP, E-mail: geciane@usp.br

3 Priscila Rezende da Costa - Docente do Programa de Pós-graduação em Administração da Universidade Nove de Julho (Uninove), E-mail: priscilarc@uni9.pro.br

1. INTRODUCTION

As organizations at diverse levels are articulated in partnership networks for the development of innovations, they have been the focus of numerous studies (Ahuja, 2000; Borgati & Halgin, 2011), including the evaluation and proposition of public policies (Bender et al., 2015; João, Porto & Galina, 2012; Ruby, 2013; Vanderelst, 2015). In this sense, Social Network Analysis (SNA) techniques are useful because they enable the interpretation and visualization of these interorganizational innovation networks (Van Der Valk & Gijsbers, 2010). Furthermore, patents are an important source of data on technological innovations, providing valuable public information that is available from patent authorities around the world, allowing them to be applied to investigate these collaborative innovation networks (Zuniga et al., 2009).

Patent documents have a unique structure and contain information such as the title, abstract, state of the art and the proposed solution, which is described as a novelty, the technological classification and data on the owner(s) and inventor(s) (CNIPA & WIPO, 2019). However, this information that is made available by patent offices in different countries usually has different types of format in their respective databases. The lack of standardization regarding how data on owners and inventors are displayed in patents, according to the patent authority, can lead to heavy restrictions in terms of comparative analyses, compromising the application of SNA and the interpretation of these networks (Kumar, 2015; Wong, Ho, Saini, Hibbs, & Fois, 2015). To solve this problem, this work aims to select, develop and present techniques for organizing and preparing a large volume of data on patents, enabling the enrichment of content and the construction and analysis of these networks.

Therefore, the main purpose of this study is to demonstrate how patent data can be structured and standardized to allow the application of SNA to investigate interorganizational collaboration networks. To achieve the central goal, this work begins with the literature on SNA and its application in studies of interorganizational innovation networks. The result of a survey is then presented on the means of collecting, organizing, preparing, standardizing and enriching data, as well as the construction and analysis of these networks.

Therefore, this study contributes to the diffusion of studies and applications of SNA to investigate interorganizational collaborative innovation networks, providing an aid to researchers and professionals. Moreover, it presents new techniques that enable the enrichment of information to analyze a large volume of data by classifying patent owners and extracting of their attributes, such as their nationality.

2. APPLICATION OF SNA IN STUDIES OF INNOVATION NETWORKS

The application of the SNA method to investigate cooperative networks, both scientific and technological, is not only an innovative proposal but a way of observing this phenomenon more fully, since comparing different networks and their structural characteristics using different metrics enriches these analyses from diverse perspectives. Variants of these applications have been used for wide-ranging purposes, such as describing, comparing and explaining the evolution of networks and the profile of actors (Cantner & Graf, 2006; Bazzo & Porto, 2010; Gomes, Galina, Vicentin, & Porto, 2017; João et al., 2012; Ruby, 2013; Souza, Moraes, Dal Poz, & Silveira, 2015), identifying and suggesting opportunities for collaboration (Bender et al., 2015), demonstrating the complementarity between technological and scientific activities (Wang & Guan, 2011), and analyzing the individual effects of collaboration networks on organizations (Kim, 2019), as summarized in Table 1.

3. FUNDAMENTALS OF SNA

SNA addresses relationships between actors and enables the identification of characteristics and structures in a set of complex relationships through statistical calculations and graphic systems. Its origins are in mathematics, specifically graph theory. It is founded on the relationship between sets of actors and allows a better interpretation of networks (Newman, 2010). The actors in a network, also known as nodes or vertices, may be people, organizations, institutions, objects, articles, patents or others, depending on the objective of the analysis. It is possible to characterize each node according to its attributes and characteristics, such as universities, companies, governments

Table 1 – Studies that use SNA to investigate collaborative innovation networks

AUTHORS	OBJECTIVE	RESULTS
Cantner & Graf (2006)	To describe the evolution of the innovative network in the region of Jena in Germany	The revolution of the network of innovators can be explained by the mobility of the scientists and the technological overlap (patents classified in the same field), rather than cooperation that occurred in the past.
Wang & Guan (2011)	To measure the relationship between scientific and technological productivity in the field of nanotechnology in China	The most active inventors and the most frequently cited authors mostly belonged to the same group, suggesting complementary effects between technological and scientific activities.
João et al. (2012)	To compare the innovation networks of the Bioen program of FAPESP in Brazil and the Biomass Program in the EUA	The study identified different stages of maturity in these programs, mapping the main actors and pointing out potential weaknesses in the Bioen network due to its high level of fragmentation and low density, whereas the Biomass Program proved to be more robust.
Ruby (2013)	To compare the cooperation networks for R&D in seven technological fields in terms of energy efficiency in Denmark	The networks showed differences in their structural features, with some areas being denser and with shorter paths, while others were less interconnected, with a greater diversity of actors and relationships. The research centers were highlighted as directors of the network in each technological field.
Bazzo & Porto (2013)	To analyze the evolution of the cooperation network for technological development of Petrobras with universities, research institutes and other companies	The network evolved in size and density throughout the period under study. Universities and research centers played an important role in strengthening the collaboration network, helping to plug structural holes.
Souza et al. (2014)	To describe the global scientific collaboration network in the field of cellulosic ethanol	The actors of American origin showed higher rates of centrality and performance levels in publications, especially the universities.
Bender et al. (2015)	To analyze and map the global scientific collaboration network with institutions from Germany in the field of Neglected Tropical Diseases (NTD)	The study identified opportunities for collaboration through knowledge hubs and suggested strengthening the research capacity in low and medium income countries, which have little involvement in partnerships with rich countries.
Gomes et al. (2017)	To compare the evolution of social networks of biotechnology companies in Brazil and Spain	The network of biotechnology companies in Spain was denser and enjoyed greater diversity of partners, both local and foreign, and the R&D centers in Spain played a central role in attracting new entrants during the evolution of the network. In Brazil, universities were the main partner of the biotechnology companies.
Kim (2019)	To investigate the effects of the structure on the innovative performance of an organization.	The position of centrality in the structure of the interorganizational collaboration network positively affected the individual innovative performance of an organization.

Source: Prepared by the authors

and individual people, either locals or foreigners. Actors and their actions are viewed as interdependent rather than independent (Wasserman & Katherine, 1994).

The existing relationships between actors in the network are considered the channels for the transfer or flow of tangible and intangible resources. They are the so-called ties or links, and can represent a wide variety of relationships, including friendships, family, partnerships, alliances, cooperation, citations, exchanges, monetary flow, properties and affiliations (Jackson, 2010).

Links have different characteristics according to the types of relationships between the actors. In SNA, they can be characterized according to the intensity or direction of these relationships. A directed tie represents a flow in a single direction between actors, as in the case of a citation between patents or articles, or a monetary flow or flow of assets between countries. Directed ties are represented on a graph by an arrow indicating the direction of the flow from its origin to its destination. A set of directed ties can also determine the paths to reaching a certain actor in the network. On the other hand, undirected ties represent a two-way flow, as in the case of a friendship, family relationship, partnerships, collaborations or alliances. Meanwhile, indirect ties establish relationships that are non-redundant, in which certain actors are not directly interconnected, but a relationship between them could be established through one or more intermediary nodes (Wasserman & Katherine, 1994).

A bipartite network or affiliation network represents the relationships between actors that are members of the same group or community of any nature. It has at least two types of actors, with one representing the original node and the other the community of which they are part. It could, for instance, be a network of co-authors of the same article, co-owners or co-inventors of a patent, advisors to a company or the purchases of an organization. A bipartite network can be transformed into a one-mode network, establishing the relationships between the actors that are members of a common network, making it possible to understand and analyze the relationships (Newman, 2010).

The concept of component has to do with the set of all the nodes that are directly or indirectly interconnected, allowing a determined actor to reach any other actor in the network via some path of connections through the network. This plays an important role in cases such as diffusion, learning and contagion. A portion of nodes in a network are part of a component if there is a path through the ties, where any node can reach another node. The giant represents the largest cluster in the network, interconnecting the highest possible number of nodes (Jackson, 2010).

4. PATENT DATA

A granted patent represents a legal monopoly for a limited time on a regional basis (every country is sovereign), designed to protect and encourage new inventions that have an industrial application. In return for the sufficient disclosure of the idea claimed in the patent, its owner receives the exclusive right to produce and sell the innovation for twenty years, thus assuring investments in R&D and stimulating inventions by reducing the risk of imitation (Lindberg, 2008).

Patent documents are recognized as indicators of technological innovation, and analyses based on this information have a series of applications that range from measuring technological development to measuring the dynamism of the innovation process through analyses of cooperation or technological routes. Much of the information on technologies is only found in patent documents, to which all of society has access, helping to advance knowledge and the development of new technologies. Diverse information can be found in a patent, such as a description of the innovation in question, state of the art and the claims regarding the novelty of the idea. The name(s) and address(es) of the owner(s) and inventor(s) are also included, along with patent and non-patent citations that indicate the origin of the invention, and the technological classification in keeping with and standardized by the International Patent Classification (IPC), along with other information (Zuniga et al., 2009). The recommendation for the IPC of the patent is made by the applicant and confirmed at the time of analysis by the technician of the patent authority. Most countries use the IPC, which was established in 1971 through the Strasbourg Agreement, which is constantly updated and expanded. The IPC uses a hierarchical classification of sections, classes, subclasses groups and subgroups. It is a powerful tool, which can be used by many patent offices, facilitating searches and identifying patents in certain fields of knowledge (WIPO, 2019).

The inventor of a patent is always an individual person, who directly participated in conceiving the idea and/or developing the technology. The owner of a patent retains the rights to that invention. Owners can be inventors or a public or private organization where the inventors work. (Zuniga et al., 2009). Co-ownership means that a patent has two or more owners, indicating that they effectively cooperated in the development of the invention and intend to have a joint share in its results (Fischer, 2005).

As a patent can be filed in different countries, these documents are then classified according to the priority number and can be unified either through the first application or the latest. This unification is known as a patent family (INPADOC), and it is carried out by the International Patent Documentation Center. In the scope of research on collaboration networks, this classification is useful because it avoids including in the analysis duplicate inventions regarding coverage of protection in several countries simultaneously.

5. METHODOLOGICAL ASPECTS FOR SNA STUDIES WITH PATENTS

In this article, a descriptive approach is used to propose a structured methodology to analyze secondary data, i.e., patents filed in different bases that were previously collected and organized.

The first stage, data collection, includes methods for defining the scope of the analysis (technological fields, organizations, geographical delimitation and others), as well as public and private sources for data collection on patent filings that include a number of offices on a global basis.

The next stage consists of organizing these data using databank construction techniques and selecting which information is useful for analyzing networks. This depends on the aim of the study and the means of identifying technologies that originated through partnerships between two or more organizations.

In the data preparation phase, techniques for cleansing the base are used, which includes removing duplicate information and excluding information that is not pertinent to a specific type of analysis, such as individuals when the focus of the network analysis is only at the interorganizational level. It also goes deeper in data refinement methods by applying specific algorithms in the OpenRefine software, enabling the massive standardization of owners' names. This stage avoids the presentation of duplicate organizations in the analysis, which tends to compromise the application of SNA and the interpretation of the results. This occurs because among the collected data there are many errors and variations in the spelling of the name of the same organization, which ranges from orthographic inconsistencies to the use of abbreviations.

This stage also makes advances in terms of techniques to enrich information for analysis, such as classifying patent owners according to their nature (e.g., universities, R&D centers, organizations and individuals) and extracting the nationality of these owners, adding valuable information to the analysis of a large volume of data. Finally, means are presented of structuring the data and constructing networks at the intraorganizational level, and these data should be fed for analysis in Gephi software.

6. THE PROCESS OF COLLECTING DATA ON PATENTS

Following the conclusion of the construction of the theoretical framework and the proposed methodology, it is time select and collect data, prepare them and afterwards analyze them. Regarding the data selection and collection, this work presents some repositories with a global scope, some free and others paid, as well as strategies for defining the sample.

The selection and extraction of data on patents should consider the objective of the analysis. Thus, the first step is to define the search strategy, which may be structured using specific or combined criteria, such as:

a) Technological classification codes (IPC), which allow studies of wide technological fields, or selection of specific technological niches;
b) Names of owners or inventors;
c) Priority country, which enables the verification of the correlation between the technology and its geographical origin or destination, i.e., the target markets where the owners wish to have exclusive rights in the production or commercialization of the technology;
d) Key words in the title, abstract or claims to prove that something unprecedented was created and protected;
e) Date of publication, priority or filing;
f) Combination of above criteria using Boolean codes (AND, OR, NOT).

Table 2: Comparison of the characteristics of the main patent databases:

CHARACTERISTICS	PATENTSCOPE	ESPACENET	DERWENT INDEX – CAPES	DERWENT INNOVATION – CLARIVATE	PATSNAP	CAS
COVERAGE	PCT, EPand another 52 offices	110 patent offices worldwide	40 patent offices worldwide	90 patent offices worldwide	128 patent offices worldwide	63 patent offices worldwide
ACCESS	Free	Free	Free through the Capes portal	Subscription	Subscription	Subscription
SCOPE OF TECHNOLOGICAL FIELDS	All fields of knowledge	All fields of knowledge	Exclusive for chemistry, electric, electronics and mechanical engineering	All fields of knowledge	All fields of knowledge	All fields of knowledge
CLUSTERING BY PATENT FAMILY	No	Yes	Yes	Yes	Yes	N/I
EXPORT CAPACITY	10,000	Only the first 500 (1)	Only the first 500	60,000 patent families	Yes	N/I
GRAPHIC ANALYSIS OF THE RESULTS	Yes - Summarized (2)	No	Yes - Summarized (2)	Yes – with edit option	Yes	Yes
ENGLISH TRANSLATIONS	Machine Translation	Machine translation	Expert's translation	Expert's translation	N/I	N/I

Notes:
1. The research criterion should be limited each time, such as reducing the period to reach up to 500 patents.
2. Analysis with information only on the main applicants, technological field, inventors and geographical coverage
N/I: Information is not publicly available.

The temporal range of the analysis should also be defined, as the term of a patent is 20 years, after which it enters the public domain. It should be highlighted that this period must be considered in accordance with the research goal, since there are themes in which it is not a restrictive factor, such as in the study of collaboration networks. Some

of the databases with information on patents with global coverage are shown in Table 2.

7. PROCESS OF PREPARING DATA ON PATENTS

The data preparation stage presents the result of the search, development and application of a set of techniques and tools that enable an analysis of networks through an extensive volume of data. This stage can generally be divided into four steps: organization of data, data cleansing, standardization of owners' names and construction of networks.

7.1 ORGANIZATION OF THE DATA

To analyze each field of technology, it is necessary to handle hundreds of tables. Thus, the broader the overview of the analysis, the greater the demands in terms of the "usability" of the base to be consulted.

After the data are extracted, they still need to be correctly prepared to enable the application of SNA techniques. In addition to requiring considerable effort to manipulate the data, a limiting factor of this stage preceding the analysis is the operational capacity of some software. To overcome such problems, it is necessary to create a database to store, organize and enable the standardization of a large quantity of data on patents.

Therefore, once the data have been extracted from the platform, the first step in the data organization is the consolidation of the tables in a single database, containing only the information necessary to conduct the analysis. This task can be achieved, for example, through the use of SQL alongside Microsoft Excel.

The second step is the creation of tables that allow the selection and treatment of data using data modeling techniques. As a model, the standard data extracted from the Derwent Innovation platform (Clarivate) was used (Table 3). Several items of information are found in the same cell using some kind of separator (e.g., "|"), which does not allow the establishment of relationships or the effectuation of standardization and cleansing. Therefore, for each item of information that is useful for the analysis, it is necessary to construct a separate table: IPC, inventor, owner and others. In these tables, the data should be placed in lines to expand the data consolidated in a single cell (Table 4). For this purpose, OpenRefine tools can be used: command "Split Multi-valued cells" in the columns with consolidated data and the separator can be used as a criterion, followed by "Fill down" in the other columns to fill in the new lines created with the above values.

7.2 DATA CLEANSING

The data cleansing involves the exclusion of owners who are individuals (people) when the analysis level of the study is interorganizational, as patents can be found that are exclusively owned by one individual or a group of individuals. Therefore, this information could affect the interpretation of the innovation networks at the interorganizational level. Due to the large volume of data, it is not feasible to perform this task manually. To address this limitation, it is assumed that the owners, when they are individual people, are listed as the inventors because they participated in the concept and invention. Therefore, when inventors are identified who are also listed as owners, these can be removed from the database so that only organizations remain in the field of "owners". A representative sample of the data should also be checked manually to identify false inventors[1].

7.3 STANDARDIZATION OF OWNERS' NAMES

The lack of data standardization, specifically the names of the owners of the patents, is a serious problem when it comes to applying SNA, as it hinders an accurate analysis of the network (Wong et al., 2015; Kumar, 2015). In the collected data, there are many ambiguities due to the variations in the spelling of a name, which range from minor orthographic inconsistencies to the use of different variations, resulting in duplicities in the same "node" in the network. This tends to skew the application of SNA, compromising the interpretation of the results.

Since a wide variety of studies use SNA for large volumes of data, the standardization of the names of the owners of patents, if done manually, would require a great deal of time and effort and could also lead to some inconsistencies being maintained. In this context, algorithms were sought that could overcome this restriction. OpenRefine software was selected, a free software

1 This kind of check is important because some inconsistencies have already been identified in the data, such as names of organizations that are listed as inventors.

Table 3 – Model of the format of the raw data

PUBLICATION NUMBER	ASSIGNEE/ APPLICANT	APPLICATION DATE	IPC	INVENTOR
Patent A	Owner Y | Owner X | Owner Z	Date 1	IPC 1 | IPC n	Inventor 1 | Inventor 2 | Inventor n
Patent B	Titular 4 | Titular5	Date 2	IPC 1 | IPC n | IPC n1 | IPC n2	Inventor 4 | Inventor 2

Source: Prepared by the authors

Table 4 – Structure of the table to analyze co-ownership in patent bases

PUBLICATION NUMBER	ASSIGNEE/APPLICANT	APPLICATION DATE
Patent A	Owner Y	Date 1
Patent A	Owner X	Date 1
Patent A	Owner Z	Date 1
Patent B	Owner 4	Date 2
Patent B	Owner 5	Date 2

Source: Prepared by the authors

with an open code, made available to the public by Google in 2010.

OpenRefine has very useful tools for handling lost or duplicate data, making the task of standardization faster and, especially, more reliable. These tools enable the exploration, cleansing and handling of a large volume of data, and allow them to be connected with external data sources through the web. This enables enriched content of databases. To standardize owners' names, clustering tools are used, which are operations that help to locate groups of words with different values that may be alternative representations of the same thing and/or content. For this purpose, the software has introduced a select number of different methods and clustering algorithms that can be effectively and efficiently used with a wide variety of data, dividing them into two groups: key collision methods and nearest neighbor methods.

Key collision methods are based on the idea of creating an alternative representation of a value (key), which contains only the part with the highest value or meaning of a string of characters and comparing it with different strings, based on the fact that the keys are the same (Fingerprint and N-Gram Fingerprint). Two useful methods for addressing minor spelling errors are also included, errors due to misunderstood pronunciation or not knowing how to spell a word. In this case, the idea is that phonetically similar words will be clustered jointly (Metaphone3 and Cologne phonetic).

Fingerprint: this method uses a process to generate a key from a string. It removes leading and trailing whitespace, changes all characters into a lowercase representation and removes all punctuation and control characters. It also splits strings into whitespace-separated tokens, sorts the tokens and removes duplicates, joins the tokens back together, and normalizes characters using a standard western representation (ASCII). This method is considered simple and fast, with a wide range of applications and little probability of generating false positives.

N-Gram Fingerprint: A method similar to the Fingerprint method, the main difference being found when splitting the strings, as it uses as a separator the number of characters (n-grams) instead of whitespace. The process is in steps, as follows. It changes all characters to their lowercase representation, removes all punctuation,

control characters and whitespace and obtains all the string n-grams. It then sorts the n-grams and removes duplicates, joins the sorted n-grams back together and normalizes extended western characters to their ASCII representation.

Metaphone3 *(*Philips, 2009): A phonetic algorithm that generally creates an index of the way words are pronounced, particularly generating, comparing and identifying the intended correspondence, especially for the English language, although it can function with other languages.

Cologne phonetic: An algorithm that attributes phonetic codes to the sounds of words so that words with the same sound have similar codes. It was created especially for German but can also be applied to other languages.

Nearest neighbor methods provide a parameter that represents a distance threshold between words. Thus, any pair of strings that are close to a certain value are clustered. This method requires more in terms of computational processing, as the values of all the strings need to be compared. Of the nearest neighbor methods, the following may be listed:

Levenshtein Distance: This algorithm is the implementation of the method developed by Levenshtein (1966). It measures the minimum number of edit operations (insertions, removals or substitutions of characters) necessary to change one string into another. The edit distances between all the strings are compared in order to relate the nearest ones.

Prediction by Partial Matching (PPM): This algorithm is the implementation of the research conducted by Li, Chen, Li, Ma and Vitányi (2004) on a distance metric suitable for measuring similarities between sequences, based on the notion of Kolmogorov complexity.

The methods and algorithms for clustering words that were presented above are listed in order of complexity and demand for data processing capacity. Thus, to increase the effectiveness of the cleansing, it is advisable to apply all of them, in ascending order of complexity, beginning with the key collision methods that are computationally faster and ending with the nearest neighbor methods.

For each method applied, a list of suggested names is presented that might be duplicates. It then falls to the researcher to conduct a critical analysis of the listed results to gauge and correct false positives. The work of verifying the suggestions of standardization, albeit extensive, is much more efficient than it being done manually, as reported by Wang and Guan (2011). This task is very time consuming and a great deal of effort is required to standardize the names of authors manually followed by an application of SNA.

The tools available in OpenRefine software, despite their high complexity, are easy to apply, and the user does not need advanced knowledge in programming for some of these activities. Therefore, the software has attracted enthusiasts from different fields, including librarians, journalists, analysts and researchers (Stonebraker, 2014). It is also possible to find reports of its application in the standardization of metadata with external databases (Van Hooland et al., 2013), implementation of data standardization protocols for names of drugs using the FAERS database2 (Wong et al., 2015), standardization of the names of researchers and addresses obtained from the SCOPUS base (Bender et al., 2015), and the successful standardization of 25.23% of the names of patent owners in the field of biofuels in a sample of 37,241 participants (Gomes, 2017). The methods used by Bender et al. (2015) and Gomes (2017) are in keeping with the proposal of this study, as they applied the OpenRefine data standardization tools to prepare data for SNA using Gephi software. As shown in the literature, the techniques presented here help to organize and successfully standardize a large volume of data, overcoming one of the main challenges to research of this nature.

7.4 CLASSIFICATION OF OWNERS

To classify the inventors and/or actors according to their attributes, primarily the nature of the institution to which they are linked (Universities or R&D Institutes or Centers, either public or private, public or private companies, government agencies) or the geographical attributes (nationality according to the mailing address of the owner) techniques were developed to allow automation and application in a large volume of data, maximizing the success rate of the operation.

7.4.1 NATURE OF THE OWNERS

To identify the actors of a scientific nature, such as a university, it is necessary to prepare a list

2 US Food and Drug Administration Adverse Event Reporting System, which is considered by Wong et al. (2015) as one of the largest drug repositories in the world.

of terms in several languages related to learning institutions (academy, school, college, faculty, teaching, university), as well as abbreviations or variations of these terms, identified manually, that may occur due to typing errors or some other related problem. For this purpose, the OpenRefine tool Word-facet was used, which allows the handling and verification of the frequency of single words within names. To identify R&D institutes or centers, either public or private, a criterion was used that encompasses the occurrence of a set of terms (center, institute, laboratory, scientific, technology; research, development, research, investigation, R&D and research institute), also translated into several languages and including abbreviations and variations.

7.4.2 GEOGRAPHICAL ATTRIBUTES

As the information was obtained from different patent authorities, the publication of owners' mailing addresses can be in various formats and inconsistencies, when they are available, which is not always the case. When patents are published, they may include information ranging from the zip code, address, city or town, state or nationality, as the format, availability and order can vary greatly. The USPTO is the patent authority that best displays this information and this explains why studies involving this variable are often restricted to the American database.

Therefore, it was decided that an algorithm would be developed using OpenRefine to identify and extract the sequence of two letters that correspond to the universal standard for nationality acronyms. The technique applied by the algorithm seeks and extracts the first, second and last occurrence of sequences containing the two letters together, which may be located at the initial or final frontier, along with (or not) a symbol or space:

First occurrence:
value.match(/.?[-](\w{2})[-].*?/)[0]*

Second occurrence:
value.match(/.?[-]\w{2}?[-](\w{2})[-].*?/)[0]*

Final occurrence:
value.match(/.[-](\w{2})[-].*?/)[0]*

The next step is to calculate the frequency of the occurrences of nationality for each owner and select only that which represents the highest frequency of occurrences for that owner in all the patents collected for the case of studies in which the intention is to work with the uniqueness of the owner.

After manually verifying a random sample, and among the main actors in the network, this method was found to present the highest success rate for the identification of the nationality of the owner in an automated way, in the format that the data were found and for a large volume of data. In patents filed with the USPTO, the acronyms of states are also published. Therefore, it is necessary to create an index of acronyms of these states and convert them into the acronym that represents their nationality (US). When one of the research goals is to identify the participation of Brazilian actors, a manual verification is recommended of documents filed at the National Intellectual Property Institute (INPI).

7.5 CONSTRUCTING THE NETWORKS

To construct and analyze the collaboration networks, only patents filed in co-ownership are selected. These networks are constituted by patents developed in cooperation, in other words, with two or more owners. This shared filing of a patent shows that the owners cooperated effectively in the development of their innovation and that they intend to share the results. This is usually done through the establishment of formal agreements between the parties (Fischer, 2005).

The data need to be manipulated to allow adequate input for Gephi software, which allows analyses of large networks, designs them and calculates the metrics necessary to interpret them (Bastian, Heymann, & Jacomy, 2009). For this purpose, it is necessary to feed the software with two tables of data, one relational and the other with the attributes of the actors.

To create the relational table of the owners, the data need to be handled and converted as follows: given a patent (A), which has three owners (X, Y, Z), the relationships will then be established among all of them (X-Y, X-Z, Y-Z). Initially, the data are found as shown in Table 2 and are then converted for the model shown in Table 3. In addition to establishing the relationships, it is necessary to calculate the frequency of patents among

Table 5 – Model of a table of relationships between actors for Gephi software.

SOURCE	TARGET	WEIGHT	TIME	TYPE
Owner X	Owner Y	1	Date 1	Undirected
Owner X	Owner Z	2	Date 1	Undirected
Owner Z	Owner Y	10	Date 2	Undirected

Source: Prepared by the authors

all of the actors, which is represented as the weight of this relationship.

To generate the table that establishes the relationship between each owner that shares the same patent, a tool was developed using the VBA in Microsoft Excel, which establishes the relationship between each applicant for the same patent (Gomes & Visnardi, 2019).

With the conclusion of all of these steps, it is possible to begin the data analysis, applying SNA using Gephi, with the respective calculation of the measurements necessary to characterize and describe the structure of the networks and represent the network in graphic form.

8. ANALYSIS OF COLLABORATION NETWORKS

After feeding the SNA software, the network analysis is begun. For this purpose, the nodes of the network represent the patent owners and are classified according to their nature, which can be defined as universities, R&D centers, organizations and individuals, according to the specific objective of each study and the level of classification that is achieved. Furthermore, according to the geographical attributes, the nodes can be classified by the nationality of the mailing address.

The ties in the network represent the sharing of at least one patent. They are undirected because a collaborative relationship represents a two-way flow. Weights can be assigned to the ties according to the number of patents shared by the same owners. This weighting of relationships is in keeping with the concept of Granovetter (1977), in which the strength of a tie represents the frequency of relationships over time. The collaborative interorganizational relationship for innovation in the field of solid biofuels is illustrated in Figure 1.

Network analysis can be conducted at two levels, macro and micro. At the macro level, the characteristics of the network as a whole are described, while at the micro level the actors who are best positioned in the relationship structure of the network are identified. To operationalize SNA, Gephi software is recommended, which allows a graphic representation of the network and executes statistical calculations of the network, as it is open to having a discussion community and wide-ranging improvements3. However, other kinds of software can also be used, including Ucinet and the igraph model of R software, each with functionalities that researchers can evaluate to determine the most suitable for their study.

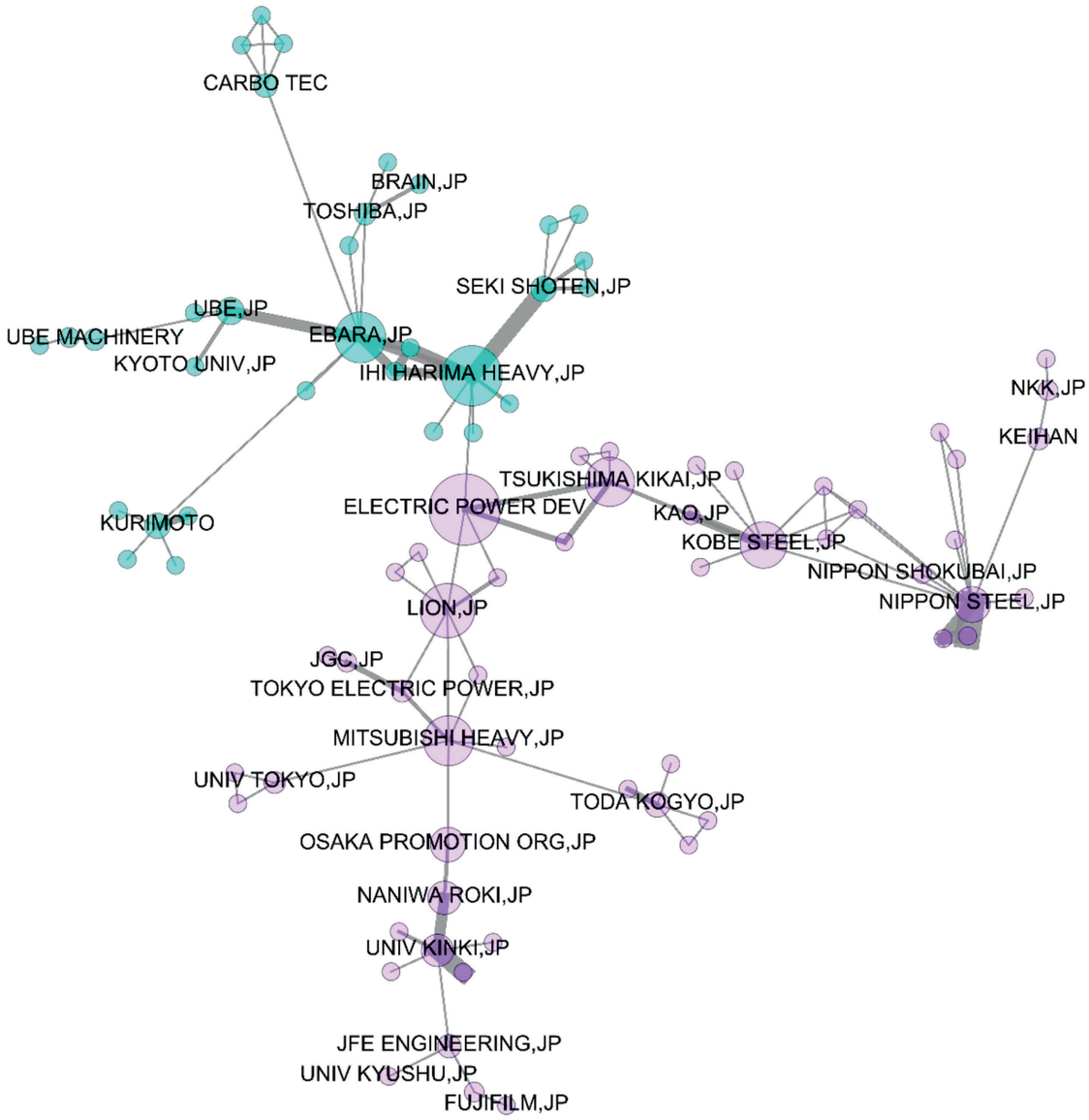
9. FINAL CONSIDERATIONS AND LIMITATIONS OF THE STUDY

This study presents a proposal for organizing the process of searching, selecting, developing and applying tools, algorithms and software to prepare and standardize a large volume of data on patents for the application SNA in interorganizational networks that are formed for the purpose of developing innovative technologies. Thus, this study contributes to the field by presenting means that allow the enrichment of analyses, enabling new research perspectives, such as those regarding the functional and geographical diversity in these networks.

As the tools presented are only intended for application in interorganizational innovation networks, in future studies this proposal should be adapted for applications in different network analysis perspectives, such as technology prospection, inventors, and mapping technological fields. Therefore, depending on the scope of analysis, the tools that have been presented need to be adapted to allow the application of SNA.

3 https://github.com/gephi/gephi

Figure 1 – Interorganizational collaboration network for innovation in solid biofuels



Source: (Gomes, 2017 p. 158)

Note: The colors represent the clustering calculated by the modularity, the size of the node represents the measurement of intermediation and the thickness of the tie represents the volume of patents.

REFERENCES

Ahuja, G. (2000). Collaboration networks, structural holes, and innovation: A longitudinal study. Administrative Science Quarterly, 45(3), 425-455.

Bazzo, K. D. C., & Porto, G. S. (2013). Redes de cooperação da Petrobras: Um mapeamento a partir das patentes (Cap. 6. 163-208). In: Turchi, L. M. O., De Negri, F. O., & De Negri, J. A. O. Impactos tecnológicos das parcerias da Petrobras com universidades, centros de pesquisa e firmas brasileiras. Brasilia

Bastian, M., Heymann, S., & Jacomy, M. (2009, March). Gephi: an open source software for exploring and manipulating networks. In Third international AAAI conference on weblogs and social media.

Bender, M. E., Edwards, S., von Philipsborn, P., Steinbeis, F., Keil, T., & Tinnemann, P. (2015). Using co-authorship networks to map and analyse global neglected tropical disease research with an affiliation to Germany. PLoS neglected tropical diseases, 9(12).

Borgatti, S. P., & Halgin, D. S. (2011). On network theory. Organization science, 22(5), 1168-1181.

Cantner, U., & Graf, H. (2006). The network of innovators in Jena: An application of social network analysis. Research Policy, 35(4), 463-480.

CNIPA - China National Intellectual Property Administration, & WIPO - World Intellectual Property Organization. (2019). Intellectual Property Basics: A Q&A for Students. Switzerland.

Fischer, F. (2005). O Regime de co-propriedade em patentes. Rev. Assoc. Bras. Propr. Intel.

Granovetter, M. S. (1977). The strength of weak ties. In Social networks (pp. 347-367). Academic Press.

Gomes, R. C. (2017). Redes de cooperação para desenvolvimento tecnológico dos biocombustíveis: mapeamento a partir de cotitularidade em patentes (Dissertação de Mestrado, Universidade de São Paulo).

_____, Galina, S. V. R., Vicentin, F. O. D. P., & Porto, G. S. (2017). Interorganizational innovation networks of Brazilian and Spanish biotechnology companies: Dynamic comparative analysis. International Journal of Engineering Business Management.

_____, & Visnardi, F. (2019). Convert two-mode networks to one-mode networks - vba macro (Version V1.0.0) [Software]. Zenodo. http://doi.org/10.5281/zenodo.3475658.

Van Hooland, S., Verborgh, R., De Wilde, M., Hercher, J., Mannens, E., & Van de Walle, R. (2013). Evaluating the success of vocabulary reconciliation for cultural heritage collections. Journal of the American Society for Information Science and Technology, 64(3), 464-479.

Jackson, M. O. (2010). Social and economic networks. Princeton University Press.

João, I. S., Porto, G. S., & Galina, S. V. R. (2012). A posição do Brasil na corrida pelo etanol celulósico: mensuração por indicadores C&T e programas de P&D. Revista Brasileira de Inovação, 11(1), 105-136.

Kim, H. S. (2019). How a firm's position in a whole network affects innovation performance. Technology Analysis & Strategic Management, 31(2), 155-168.

Kumar, S. (2015). Co-authorship networks: a review of the literature. Aslib Journal of Information Management, 67(1), 55-73.

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. In Soviet physics doklady (Vol. 10, No. 8, pp. 707-710).

Lindberg, V. (2008). Intellectual property and open source: A practical guide to protecting code. " O'Reilly Media, Inc.".

Li, M., Chen, X., Li, X., Ma, B., & Vitányi, P. (2004). The similarity metric. IEEE transactions on Information Theory.

Newman, M. E. J. (2010). Networks an Introduction. Nova York: Oxford University Press.

Philips, L. B. F. (2009). U.S. Patent Application No. 11/890,334.

Souza, L. G. A., Moraes, M. A. F. D., Dal Poz, M. E. S., & Silveira, J. M. F. J. (2015). Collaborative Networks as a measure of the Innovation Systems in second-generation ethanol. Scientometrics, 103(2), 355-372.

Stonebraker, I. (2015). Good Library Data Made Better With Technology! Using OpenRefine and Google Fusion Tables in Academic Business Libraries Instruction. Academic BRASS.

Wang, G., & Guan, J. (2011). Measuring science–technology interactions using patent citations and author-inventor links: an exploration analysis from Chinese nanotechnology. Journal of Nanoparticle Research, 13(12), 6245-6262.

WIPO- World Intellectual Property Organization. (2019) Guide to International Patent Classification. Disponível em: <https://www.wipo.int/export/sites/www/classifications/ipc/en/guide/guide_ipc.pdf>

Wong, C. K., Ho, S. S., Saini, B., Hibbs, D. E., & Fois, R. A. (2015). Standardisation of the FAERS database: a systematic approach to manually recoding drug name variants. Pharmacoepidemiology and drug safety, 24(7), 731-737.

Van Der Valk, T., & Gijsbers, G. (2010). The use of social network analysis in innovation studies: Mapping actors and technologies. Innovation, 12(1), 5-17.

Vanderelst, D. (2015). Social Network Analysis as a tool for research policy. PLoS neglected tropical diseases, 9(12).

Zuniga, P., Guellec, D., Dernis, H., Khan, M., Okazaki, T., & Webb, C. (2009). OECD patent statistics manual. França: OECD Publications.