Analyzing the Collaboration Network of Real Options Authors

Hernandes Coutinho Fagundes

Petróleo Brasileiro S.A.(Petrobras) / Universidade Estadual do Norte Fluminense Darcy Ribeiro (UENF), Brazil

hernandesfagundes@gmail.com

Rodrigo Tavares Nogueira

Universidade Estadual do Norte Fluminense Darcy Ribeiro (UENF), Brazil

nogueirart@gmail.com

Abstract

Bibliometrics is a powerful tool to obtain scientific production data, with applications in a wide range of knowledge fields. When the data visualization and analysis, due complexity, requires techniques that are more advanced, the Graph Theory application provides a great benefit. This study develops bibliometric methods in association with the Graph Theory to construct and analyze the collaboration network among Real Options Articles Authors in a world scale and pointing out Brazilians particulars. The paper concludes that the existent network is complex and notices the occurrence of various isolated communities, which are inwardly well connected. Few professionals compose most of these groups, but it was possible to identify a giant component joining 11% of the world researchers.

Keywords: Bibliometrics, Graph Theory, Real Options Theory, Collaboration Networks.

1. Introduction

In literature, the definition of bibliometrics is consensual as a set of laws and principles applied on knowledge production mapping and based on mathematical and statistical methods. (Café; Bräscher, 2008 and Guedes; Borschiver, 2005). Bibliometrics was known in the first instance as Statistical Bibliography and its application can be tracked until 1922, in the work of E. Wyndham Hulme. After that, it stood in a latency period, being applied in 1944, twenty-two years later, by Gosnell and in 1962 by L. M. Raisig (Guedes; Borschiver, 2005).

In 1969, Pritchard discussed in his article the term "bibliography", commonly employed, arguing that it was not suitable and suggesting the adoption of "bibliometrics", which became the official designation (Groos; Pritchard, 1969).

The distribution of the written knowledge, focus of the bibliometrics, shows that most prominent members of the community have increased chances to achieve further development when compared to those less connected. This finding was named "Matthew Effect" as an allusion to a biblical story and was applied on Science by Merton (1968).

The bibliometrics ability to evaluate knowledge generation, in an objective manner, has inspired many researchers to consider it a valuable tool. It help to achieve a reliable bibliography, determine the "state of the art" of a particular field and study the relationship network related to scientific communities.

The Graph Theory belongs to Discrete Mathematics field and makes use of "dots and lines" representation to identify relations and make problem solving easier to understand (Ostroski; Menoncini, 2009).

Collaboration Networks represent the visual approach and metrics developed in Graph Theory together with oriented data survey professed by bibliometric fundamental. This combination has the potential to improve significantly the researcher understanding.

This paper addresses the study of articles generation concerning Real Options Theory, which influences greatly Economics, Engineering and Management. This subject arose from the Financial Options valuation problem to permit the scientific analysis of flexibility and managing action on investment projects. Real Options Theory has expanded, fast-paced over the 90's and can be comprehended as an optimization under uncertainty problem (Dias, 2014). The growing importance and wide reach justifies the subject choice for the present research.

2. Bibliometrics

The main bibliometric laws are (Café; Bräscher, 2008):

(i) The Bradford's Law, for journals' influence study about a specific theme;

(ii) The Lotka's Law, for researchers' contribution about a Field of Knowledge;

(iii) The Zipf's Law, for words frequency inside scientific publications as its representativeness.

Beyond these three distinguished laws, other bibliometric principles have been used in specific applications to evaluate mathematical relations with the available data (Guedes; Borschiver, 2005). These principles are summarized as shown in Table 1.

Table 1 – Bibliometric laws and principles.

Information Science - Bibliometrics	
Laws and Principles	**Study Focus**
Bradford's Law	Jounals
Lotka's Law	Authors
Zipf's Law	Words
Goffman's Transition Point	Words
Invisible Colleges	Citations
Impact Factor	Citations
Bibliographic Coupling	Citations
Co-citation	Citations
Literature Obsolescence	Citations
Mid-Life	Citations
Goffman's Epidemic Theory	Citations
Elitism Law	Citations
Research Front	Citations
80/20's Law	Information Demand

Source: Guedes; Borschiver, 2005

For a Brazilian perspective of Real Option publications, it is possible to consider as reference a 2014 study by Costa (2014), which lists thesis and dissertations published by the country's universities. It has shown a clear concentration in the Southeast region, which has the highest Gross Domestic Product (GDP) per capita among all Brazilian regions. The study revealed as well that four Universities are responsible for 74% of the thesis and dissertations published in Brazil: PUC-RIO, FGV São Paulo, USP and IBMEC. It is also possible to verify the wide range of interest in Real Options Theory, as mentioned before, since there were publications identified in 16 different graduation courses.

3. Graphs and Collaboration Networks

Graphs are defined as mathematical representations of a communication network. Formally, they can be described as a collection of vertices (V) and a collection of edges (E). The graph G is written as: $G = (V, E)$ (Van Steen, 2010).

The number of edges linked to a vertex defines the vertex degree. To sort the graph's vertices degrees in decreasing order, in the shape of a plot or list, represents the Graph's Degree Distribution, as exemplified in Figure 1.

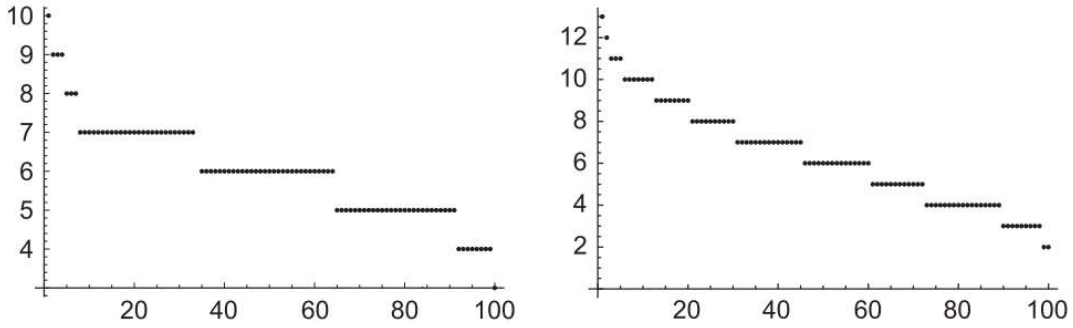

Figure 1-Two different Graph's Degree Distributions. In the ordinates axis is represented the number of vertices with the degree shown in the x-axis. Source: Van Steen, 2010.

The metrics developed for graph analysis, such as degree distributions, reveal important information about the network it represents. For the purpose of Collaboration Networks Analysis it is possible to highlight among these valuable properties the Clustering Coefficient and the Size Distribution. They will be introduced with other relevant concepts.

A vertex's Clustering Coefficient is defined as the odds that two different vertex's neighbors are also neighbors between each other (Latapy, 2008). For the whole Graph (Global View), the clustering coefficient is the arithmetic mean of the vertices' clustering coefficients. It is important to notice that the clustering coefficient exists only for vertices that have degree two or greater.

Two different vertices of graph G are considered connected if there is a path between them. The Graph G is considered connected if all possible pairs of vertices are connected.

Another, important definition is about components. A subgraph of G is considered a component of G if it is connected and is not restrained inside another subgraph of G with more vertices or edges (Van Steen, 2010). In Collaboration Networks, it is particularly common that the representative Graph has more than one component, hence being disconnected. The component's size distribution shows, graphically, how large these components are, considering the amount of vertices in each one.

"Non-social" network's vertex degrees exhibit Gaussian or Poisson distribution. However, Social Networks regularly have other distribution like the Power-Law. According this distribution, there are few individuals (represented as vertices) with high degree and a very large number of individuals with low degree (Cervantes, 2015), as shown in Figure 2.

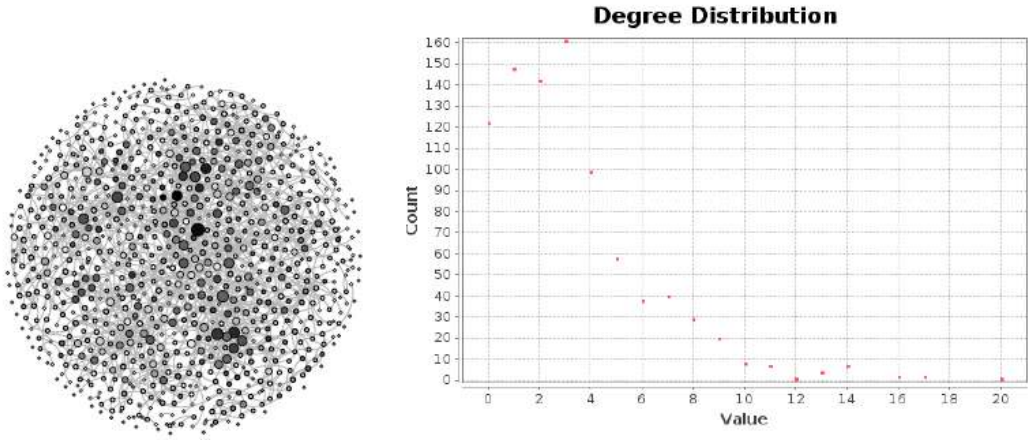

Figure 2 – An Example of a real scientific collaboration network showing the existence of few vertices with high degree and many vertices with low degree. Source: Cervantes, 2015.

The roots of Collaboration Graphs are established inside the mathematical community. In fact, one of the first parameters, the author's Erdös Number, is defined as the distance, in a collaboration Graph, between the author and the remarkable mathematician Paul Erdös. A study by Grossman and Ion (1995), underscore that the maximum Erdös Number of Field and Nevanlinna Awards winners, between 1986 and 1994, was nine, discussing the probability and implications.

4. Data Gathering Methodology

For data mining, the Elsevier's SciVerse SCOPUS database was used, restraining the search for the period between January 2011 and November 2016. As search criteria it was considered articles with "Real Options" included among the keywords. The result was a sample of 1034 articles as shown in Figure 3.

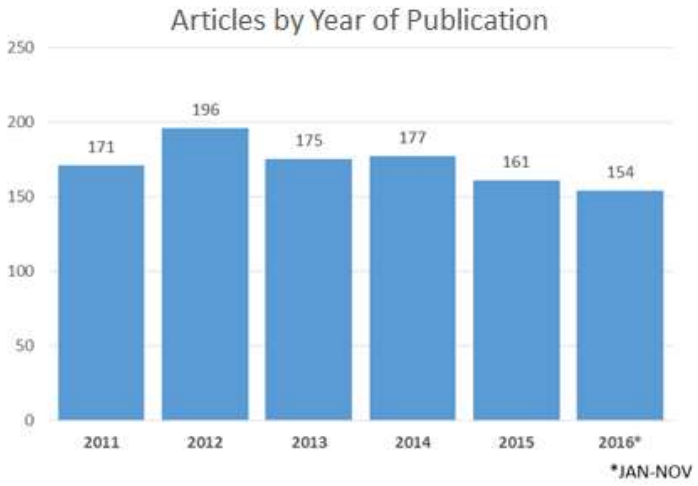

Figure 3- Articles by year of publication. Source: Author.

For data manipulation, the result were directly exported to the bibliographies management software Thomson Reuter's EndNote in order to easily separate the information in fields like Title and Author. After that, all data was moved to Microsoft Excel in a Comma-separated Values (.CSV) file.

The first step was to distinguish the authors responsible for the 1034 articles. In a first approach 2560 authorships were found, notice that one single author may be responsible for several authorships. With more treatment, it was possible to distinguish that 1829 authors wrote the 1034 articles.

To recognize that different publications belong to same author is not a simple task in some cases. Four authors presented subtle grammatical discrepancy. In those occurrences, the articles were considered as belonging to one single author. The mentioned cases were:

- ShahNazari, M. and Shahnazari, M.
- Al Sharif, A. A. A. and Al sharif, A. A. A.
- Musshoff, O. and Mußhoff, O.
- Van Reedt Dortland, M. and van Reedt Dortland, M.

Other entries of notorious similarity were also matched. They were merged when considered that the subjectivity of judgement represented a smaller possibility of mistake when compared to disregard it. As an example the following four versions: (i) Brandao, L. E.; (ii) Brandão, L. E.; (iii) Brandaõ, L. E. T. and (iv) Brandão, L. E. T. were all considered as a single author.

Figure 4 presents the authors with the larger number of publications. All scientists with seven or more works fitting the search criteria are listed.

Figure 4- Articles by authors, with seven or more published works. Source: Author.

The data analysis reveals a significant contribution from the University PUC-RIO, in Brazil, with nineteen articles published in less than six years. Considering the sample in study, it is the institution with most published works followed by the University of Manchester, which had eighteen published articles. Figure 5 shows the top ten Universities according to the amount of published works.

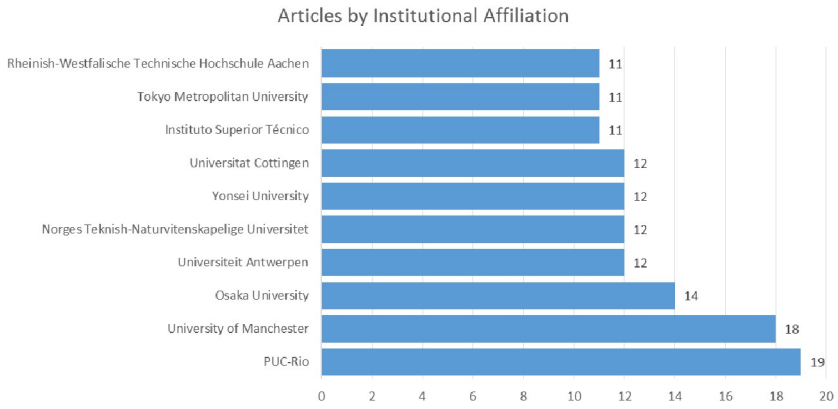

Figure 5- Articles by institutional affiliation. Source: Author.

Considering the publications country of origin, the United States has a special place as the nation with most published works and a wide lead, 237 articles in total. Brazil occupies the 12th position with 34 articles.

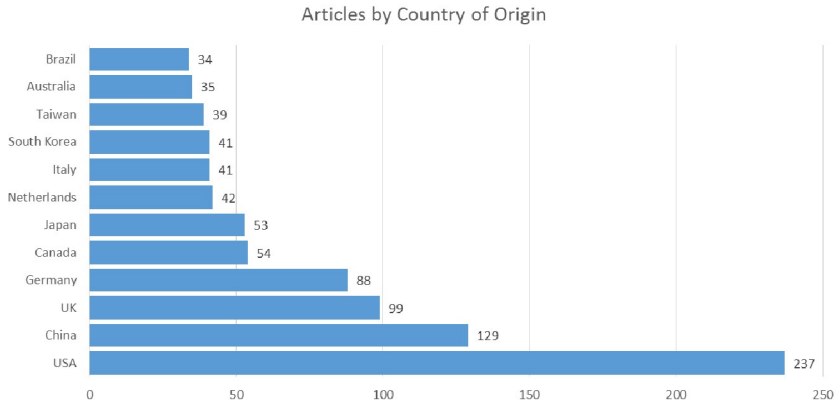

Figure 6- Articles by country of origin, 1st to 12th position. Source: Author.

5. Collaboration Graph Construction

A vertex number was assigned for each one of the 1829 authors identified in the data sample, following alphabetical order. As such, the researcher Aabo, T. was assigned as vertex 1 and Zwart, G. as vertex 1829.

In the vertices list, a Visual Basic for Applications (VBA) routine was implemented in order to link the vertex number with the article identification number and, after that, list the articles assigning the vertex number to each work. As a result, it was produced a list of articles with the authors data processed and coded, which could be used to evaluate relationships, like shown in Figure 7.

Figure 7- Articles by number of authors. Source: Author.

The amount of vertices was previously known as the quantity of authors, which is 1829. Nevertheless, the 2-combination of all authors assigned for the same article compose the list of edges, this way, it is possible to find out that the overall number of edges seems to be 2347.

It is important to realize that this first calculation can consider more than one edge between the same pair of authors, when both researchers published more than one article together. However, a collaboration network is represented as a Simple Graph, which admits no multiple edges. Eliminating these redundancies, a final number of 2103 edges composes the Collaboration Network.

The 244 eliminated edges provide information about the behavior of articles publication as the recurrence of works with same co-authorship shows. It is possible to notice that 134 pairs of authors published two articles together in the period of study and 33 pairs of authors published three works. The maximum co-authorship recurrence found counted five works as represented in Figure 8.

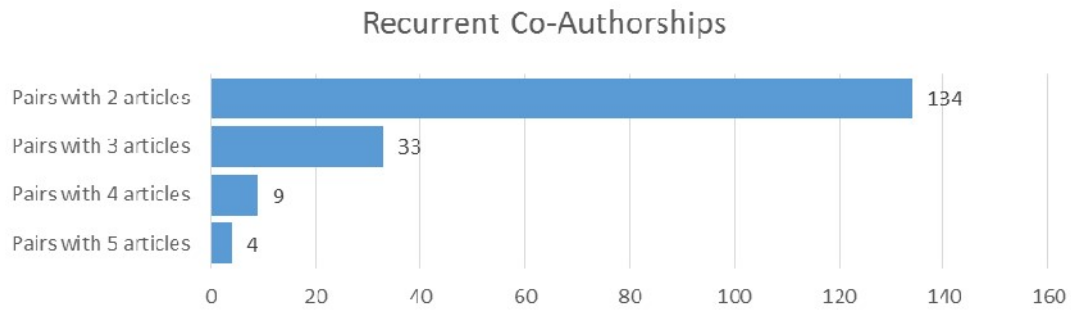

Figure 8- Recurrent co-authorships considering pairs of authors. Source: Author.

With the assistance of a manipulation and visualization software, GNU Gephi, it is feasible to construct a depiction of the Collaboration Network, which can be used for relationships identification and parameters measurement. The software output can be viewed in Figures 9 and 10, with different magnifications.

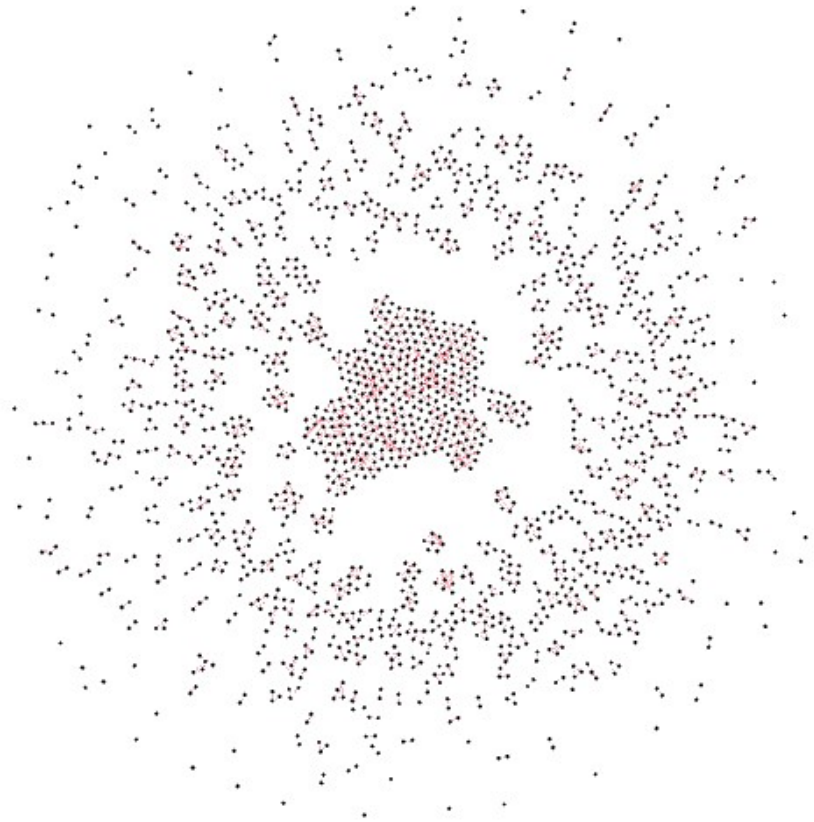

Figure 9 - Collaboration Network as constructed by Gephi software – Overall Network. Source: Author.

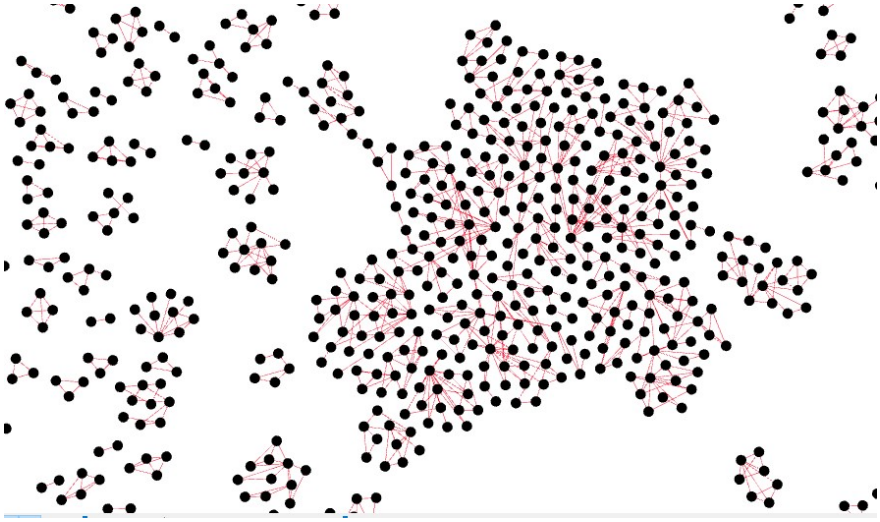

Figure 10 - Collaboration Network as constructed by Gephi software – Core Magnification. Source: Author.

Interesting conclusions can be expressed through Graph metrics analysis. First observation shows that degree distributions is in accordance with forecast expectation, with a high number of low degree vertices and a low number of high degree vertices. The distribution, as shown in Figure 11, is very similar to the one presented by Cervantes (2015).

Results:

Average Degree: 2,300

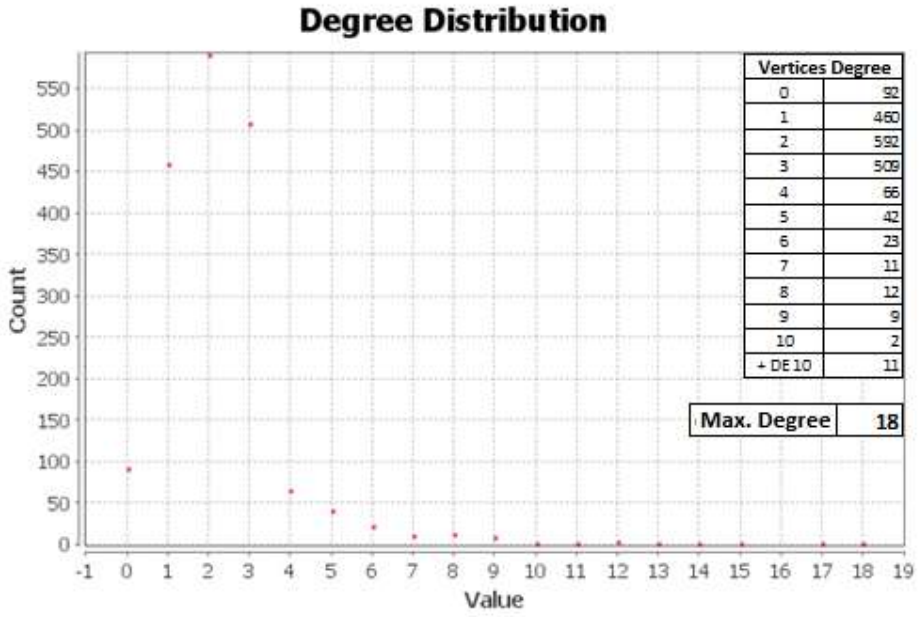

Figure 11 – Degree distribution of the Real Options Authors Collaboration Network. Source: Author.

The components size calculation shows a highly disconnected Graph, there are 548 components in total. Regarding the number of vertices constituting components, the median is 2.5, proving that major of components are quite small. Nevertheless, it is possible to identify a giant component with 197 vertices, as shown in Figure 12, while the second component in size has only 37 vertices. The overall components size distribution is shown in Figure 13.

Figure 12- Giant component with 197 Real Options Scientists. Source: Author.

Number of Weakly Connected Components: 548

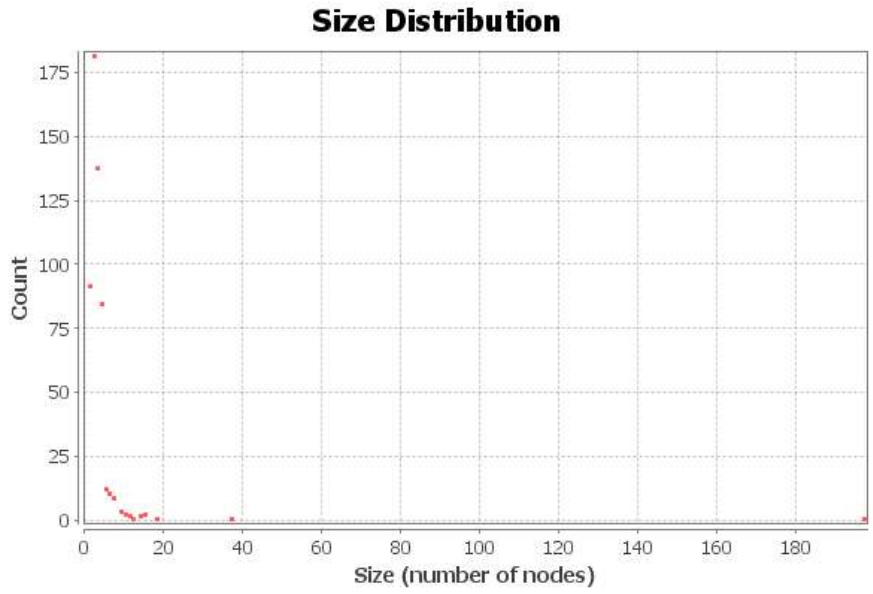

Figure 13 – Components Size distribution in Real Options Authors Collaboration Network. Source: Author.

Other interesting measure that arises from Graph metrics is the clustering coefficient. In this case, the measured clustering coefficient is 0.893, suggesting a strong connection between neighbors or, in other words, that researchers that had a co-authorship relation with a particular scientist have a high chance to have this relationship among themselves as well. This outcome is biased by the existence of a high number of low degree components, authors who have published with just several researchers, since it is easier for small groups to achieve a high clustering coefficient.

To exemplify this statement, from the 548 components, 274 are constituted by less than three vertices and don't contribute for clustering coefficient calculation. From the 274 remaining components, 138 have exactly three vertices and 94% of the 3-vertices components have a clustering coefficient equal to one.

Less predictable is the observation that larger components have also a high clustering coefficient. Even with 197 vertices, the giant component has a 0.712 clustering coefficient and the second largest component with 37 vertices has a 0.813 clustering coefficient. The clustering coefficient distribution is shown in Figure 14.

Results:

Average Clustering Coefficient: 0,893
Total triangles: 930
The Average Clustering Coefficient is the mean value of individual coefficients.

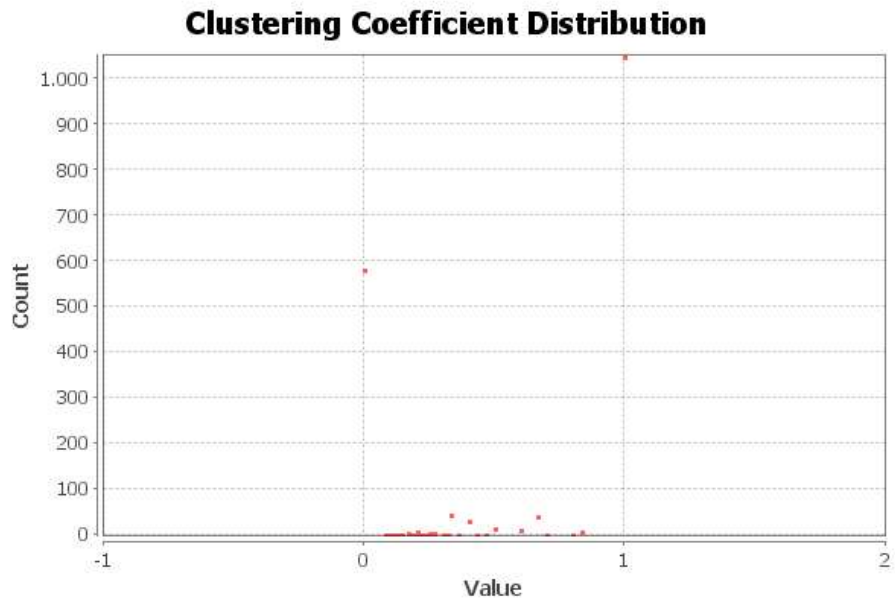

Figure 14- Clustering Coefficient distribution in Real Options Authors Collaboration Network. Source: Author.

6. Conclusion

This paper exhibits the applicability of bibliometric methods to understand Real Options knowledge production. The statistical approach provides reliable data gathering and objective mathematical evaluation. The sample contained a lot of information with complex associations, usefully addressed by Graph Theory methodology.

Better understanding of scientific knowledge formation is currently becoming more important as intellectual production rate is increasingly growing and so the quality requirements for thesis, dissertations and articles.

Real Options Collaboration Network has a significant independent portion of independent publications, with a low collaboration propensity. Notice that 92 researchers (5%) published without co-authorship relation during this almost six years period and 460 researchers (25%) published with only one workmate. The increased number of components also demonstrate a decentralized generation of knowledge.

The clustering coefficient analysis proves that the many different groups are inwardly well-connected sharing information and developing co-authorships.

It was particularly interesting to identify a giant component with 197 of global researchers (11%) with a high clustering coefficient. The scientists inside this component have probably a great influence in Real Options Theory global tendencies and evolution pace.

For future evaluations, it is important to monitor both parameters, components size distribution and clustering coefficient, in order to correctly cognize the researchers' involvement.

7. Limitations and Future Work Opportunities

This paper researched the SCOPUS database to find articles with "Real Options" among the keywords, on documents published between January 2011 and November 2016. In this way, there is an opportunity to extend it by changing the search criteria, the period and scientific reference.

Previous bibliometric studies, like Café and Bräscher (2008), indicate that a major error factor is the author's name standardization. In this paper, a lot of caution was taken to avoid this mistake, as mentioned during discussions. Even so, it is possible that a single author was interpreted as different researchers or the opposite. To exemplify, notice the inputs:

- Batkovskiy, A. M. e Batkovskiy, M. A. and
- Noronha, J. C. e Noronha, J. C. C

Both could be easily classified as a single author, but in both cases, they are actually two different researchers in the same article. Their full names are:

- Aleksandr Mikhaylovich Batkovskiy and Mikhail Aleksandrovich Batkovskiy
- Juliana Caminha Noronha and Julia Cristina Caminha Noronha

It is very interesting to investigate the reasons for the decentralization, concerning the knowledge generation. It is possible it arises from the Real Options interdisciplinarity, recent development, geographic factors or other variables.

It is also a great opportunity to identify the expertise of different groups, their influence capacity and their relationship with the giant component.

8. References

CAFÉ, L. M. A.; BRÄSCHER, M. Organização da informação e bibliometria 10.5007/1518-2924.2008 v13nesp1p54. Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação, v. 13, n. 1, p. 54-75, 2008.

COSTA, B. E. D. Estudo bibliométrico sobre opções reais no Brasil. 2014. 254 (Mestrado em Ciências Sociais Aplicadas). Universidade Federal de Uberlândia, Uberlândia.

DIAS, M. A. G. Análise de Investimentos com Opções Reais – Teoria e Prática com Aplicações em Petróleo e Outros Setores – Volume 1: Conceitos Básicos e Opções Reais em Tempo Discreto. 1 ed. Rio de Janeiro: Interciência, 2014.

GROOS, O. V.; PRITCHARD, A. Documentation Notes. Journal of Documentation, v. 25, n. 4, p. 344-349, 1969.

GROSSMAN, J. W.; ION, P. D. On a portion of the well-known collaboration graph. Congressus Numerantium, p. 129-132, 1995.

GUEDES, V. L.; BORSCHIVER, S. Bibliometria: uma ferramenta estatística para a gestão da informação e do conhecimento, em sistemas de informação, de comunicação e de avaliação científica e tecnológica. Encontro Nacional de Ciência da Informação, v. 6, p. 1-18, 2005.

LATAPY, M. Main-memory triangle computations for very large (sparse (power-law)) graphs. Theoretical Computer Science, v. 407, n. 1-3, p. 458-473, 2008.

MERTON, R. K. The Matthew effect in science. Science, v. 159, n. 3810, p. 56-63, 1968.

OSTROSKI, A.; MENONCINI, L. Teoria dos grafos e aplicações. Pato Branco: 2009. P.6

PEREZ CERVANTES, Evelyn. Análise de redes de colaboração científica: uma abordagem baseada em grafos relacionais com atributos. 2015. Dissertação (Mestrado em Ciência da Computação) - Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2015.

ROBERT TARJAN, Depth-First Search and Linear Graph Algorithms, in SIAM Journal on Computing 1 (2): 146–160 (1972)

VAN STEEN, M. Graph theory and complex networks. An introduction, v. 144, 2010.