Graduate Programs in Computing at the University of Brasília: Comparison of Academic Papers and Collaborations by Gender

Mariana Alencar do Vale
Depart. of Computer Science
Universidade de Brasília
Brasília, Brasil
0009-0004-8313-190X

Maristela Holanda
Depart. of Computer Science
Universidade de Brasília
Brasília, DF, Brazil
0000-0002-0883-2579

Celia G. Ralha
Depart. of Computer Science
Universidade de Brasília
Brasília, DF, Brazil
0000-0002-2983-2180

Aleteia Araujo
Depart. of Computer Science
University of Brasília
Brasília, Brazil
0000-0003-4645-6700

Dilma Da Silva
Depart. of Computer Science and Engineering
Texas A&M University
College Station, TX, U.S.
0000-0001-6538-2888

***Abstract*—The computing field has a low level of gender diversity, being predominantly male. This diversity gap is reflected at different academic levels, ranging from undergraduate degrees to master's and doctoral qualifications. As in other parts of the world, the University of Brasília, one of the top 10 universities in Brazil, has a low rate of women in Computing in its graduate programs. According to data from CAPES (Coordination for the Improvement of Higher Education Personnel in Brazil), the computing area is one of the areas of Exact Sciences with the lowest number of women proportionally in Brazil. In this context, this paper aims to present an analysis of the number of publications and scientific collaborations related to the gender of researchers in the Graduate Program in Informatics (PPGI) at the University of Brasília. To develop this research, technologies such as scraping were used to collect data from the program's professors, articles published (only full papers in conferences and academic journals) by them and names of the people who collaborated in the preparation of the articles, a database graph-based noSQL was used to generate the relationship networks. A relationship network was created, in which it was possible to analyze the relationship level of each professor in the program. As initial results, the average number of articles published is the same by gender, we did not find significant differences in the number of publications between men and women. In the study we only counted the number of publications, we did not analyze the impact of the publications. Regarding collaborations, initial results indicate that women are more collaborative than men in the program, with a higher degree of collaboration than the average for male researchers. In general, in the partial results, collaboration networks for journals and conferences present the same result, that female researchers proportionally collaborate more than men, both internally (publications co-authored with members of the University of Brasília) and externally (publication with co-authorship outside the University of Brasília), in academic publications.**

***Index Terms* —NoSQL, Graphs, DB, Women in Computing, Graduate, Master, PhD**

I. Introduction

The computing field is still largely dominated by men [1], [2]. In 2019, women represented only 13.6% of graduates in courses focused on computing and information technology, according to INEP (National Institute of Educational Studies and Research) in Brazil [3]. To try to combat this lack of diversity, there are several initiatives to encourage women to enter the field, such as female groups that come together to teach computing in high schools, an example of this is the Meninas.comp (in English, Girls.comp) at the University of Brasília [4].

Another initiative is research that aims to disseminate the actions of women in the area, such as the research carried out by the authors in [5]. Even so, the number of women enrolled in graduate programs in Computing continues to be one of the lowest in Exact Sciences programs [6]. This lack of gender diversity is reflected in the academic environment in the number of publications by researchers in computing.

This paper aims to characterize the work of advisor professors in the Graduate Program in Informatics (PPGI) at the University of Brasília, at master's and doctoral levels, in order to understand whether the gender of professors is related to the production of papers and the creation of collaboration networks. For this, data collected from the University of Brasília PPGI website [7] and the professors' Lattes Curriculum are used for analysis through graphs and collaboration networks. Lattes Curriculum is a platform with curricula vitae of people associated to Brazilian scientific communities. The Brazilian National Council for Scientific and Technological Development (CNPq) manages the Lattes platform [8].

The remainder of this paper is divided into the following sections: Section II, where theoretical references are presented;

Section III, where related works are presented; Section IV, where the applied methodology is explained; Section V, where results are presented; and, finally, Section VI, which contains the conclusions.

II. Theoretical Reference

This section presents some of the methods and technologies used in the research, which involve collaboration networks and graph databases.

A. Collaboration networks

Collaboration networks are defined as an intra- or inter-organizational set with a common objective, obtaining collective solutions [9]. One of the main objectives of collaborative networks is the sharing of knowledge and learning by individuals in the network. An application of this concept serves to study interactions between individuals.

Collaboration networks are interesting subjects for research as they allow us to understand the behavior of the study sample. It is important to generate the collaboration network of the desired context, which in this case would be scientific production in the aspect of authorship and co-authorship, which involves external and internal collaboration networks.

Internal collaboration networks mean, in the case of research, collaborations between professors advising the PPGI at the University of Brasília, while external collaboration networks would mean collaborations between professors from the PPGI with people who are not professors advising it.

B. NoSQL Database

For a long time, relational databases were the main option for data storage. However, the popularity of non-relational databases, called NoSQL (not only SQL) databases is increasing as they present an alternative with great scalability and flexibility [10].

There are several non-relational storage alternatives on the market today, which use data model approaches, such as document (such as MongoDB), graph (such as Neo4j), key-value (such as Redis), among others.

One of the advantages of a graph-oriented database is, as concluded in the article [11], a shorter query time compared to relational databases when it is necessary to perform deep queries that have data crossing from the same entity, since a relational database requires several joins between entities, while the graph-oriented database uses the relationships between graphs to perform such queries. Furthermore, graph-oriented databases are useful for data that is interconnected and also for visualizing the connection between each item of data, that is, it greatly values the relationship, which is essential for creating a collaboration network.

Among these approaches, the graph approach was chosen for this study, specifically, the Neo4J DBMS (database management system). Neo4J was chosen for the following reasons: Neo4J is one of the largest and most active graph communities [12], it has good performance and is easy to use with the Cypher language. Cypher is an SQL-based graph query language created for use with the Neo4J non-relational database. In addition to being a very intuitive language, providing a visual way to create patterns and relationships, it also has extensive online documentation, starting with the guide provided by Neo4J itself [20], as well as guides from the community.

In Neo4J, nodes represent entities. A node would be similar to a row in a relational database. Two nodes can be connected through a relationship, where the relationship can be of the type *incoming* or *outgoing*, i.e. arriving from the node and leaving the node. Although there are no clear rules regarding the use of these types of relationships [21], nodes with relationship *outgoing* are the ones that execute the action, or the node with the greatest importance for the scenario.

For example, Figure 1 has the Tom Hanks node and the Cloud Atlas node. These nodes have the relationship *Acted_In*, meaning "acted in", which is *incoming* to the Cloud Atlas node and *outgoing* for the Tom Hanks node. In this study we used the nodes 'authors' and 'articles' that have the relationship 'wrote'. This relationship would be *outgoing* for the node 'author' and *incoming* for the 'article' node. This approach benefits the study because its focus is on the authors themselves, and not on other nodes such as research and institution.

Fig. 1. Example of incoming and outgoing relationship.

III. Related Work

In the literature it is possible to find several papers that present the disparity of the male gender in different sectors of computing. These topics are necessary for creating the research collaboration networks that are the focus of this article.

Paper [14] presents a digital platform with data from all over Brazil to disseminate the panorama of female performance in computing, such as data from women enrolled in computing courses, women employed in the area, women who work in the area and are satisfied, among others. It is shown that there are still differences between female and male performance in the area, such as admissions, graduates and average salary.

Paper [6] aims to discover the state of female representation in graduate programs in the areas of Exact Sciences at the University of Brasília. For this, open data from CAPES was used, choosing data from 2007 to 2017 referring to the University of Brasília. There was a selection of interesting fields to be analyzed, and, finally, the average number of students studying master's degrees in exact sciences areas divided by gender and year was obtained. With this, the proportion of postgraduate women in each course was obtained, and showed that computing is the area in the sample with the lowest proportion of women.

Similar to [6], [17] used graduate program data from the University of Brasília, but focused on the area of computing. They obtained the proportion of women in each graduate program and the areas that seem to attract more women in professional and academic master's degrees. This study also observed a gender discrepancy in the master's and doctorate degrees in computing at the University of Brasília, with less than 15% of women included.

In [15] an assessment of the female presence in computing areas is presented based on participation in computing event program committees. To do this, they selected events held by the Brazilian Computing Society (SBC), choosing the study period between 2011 and 2019. The research concludes that the percentage of women in national and international events is similar, which confirms that there is low female representation in computing worldwide.

For data storage, paper [11] analyzes database approaches, comparing a relational database and a graph-oriented database. It concluded that for queries of great depth and that require the crossing of data from the same entity or node, a graph database would be a better alternative compared to the relational method, in terms of execution time requirement.

Finally, paper [16] provides ways to use graph databases to obtain interesting information from a relationship. The study focuses on investigating collaboration networks between professors from graduate computing programs at five Brazilian universities, aiming to generate recommendations for partnerships based on work already published by researchers.

Unlike the other papers presented, this paper aims to create a gender-based collaboration network to analyze the participation of PPGI faculty advisors at the University of Brasília in the publication of articles and in internal and external collaborations.

IV. Methodology

The development of this research consisted of the following steps: definition of the data model, data collection, identification of professors' genders, ETL process (*Extract, Transform, Load*) for the graph database (Neo4j) and data analysis, which was divided into journal data, conference data, and combined journal and conference data.

The database model is based on the one proposed in "Scientific Collaboration Network Analysis as a Tool in the Management of Postgraduate Programs" [16], however in our paper the gender field was added to the authors, represented by Figure 2. The model presents three entities: the Author, the published Articles (Paper) and the Institution. The Author has the attributes name, lattes (which would be the link to the author's Lattes resume) and gender. The Institution has its name, and the Article has its title.

After designing the database model, data collection begins. The names of the PPGI professors at the University of Brasília were obtained from *program website* [7], using *webscraping*, with the Python BeautifulSoup library[1], to automate the process.

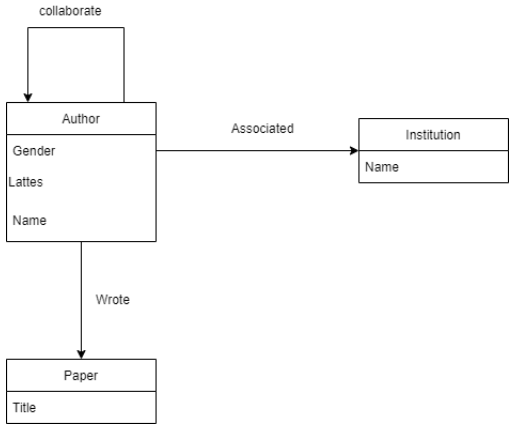

Fig. 2. Database Model.

With the list of professors, the next step is to create the gender role. Initially, the gender-guesser[2] library was used, as in the [15] search, but it was noted that this library was not very effective for Brazilian names. To remedy this, Brazilians from the 2010 Census and classified by gender, were added to the dataset of names, as guided by [18]. The function to infer gender based on name helps automate the task of defining the gender of authors of academic articles, a fundamental point for the objective of this research.

The next step is to obtain the published articles of each professor advising the program. The *webscraping* process was not carried out on the Lattes web page, since the page's *firewall* blocks such operations. The solution found was to manually download the professors' Lattes curriculum from the Lattes platform in the *.xml* extension, which allows the data *scrapping* process with the same libraries used to extract the names of the professors from the program. In this way, the articles published by the program's professors and the people they collaborated with were obtained, as well as the relationship between the authors and the articles.

Now that we have data on the program's professors, articles published in journals and full papers published in conference annals by these professors and the names of the co-authors of these articles, the accuracy of the gender function for the sample can be verified, obtaining the results present in Table I. This table is made up of four columns, namely hits, errors, unidentified and total; and by 4 lines of authors by affiliation, these being the University of Brasília, other authors who appear in journals, other authors who appear in conference data and Total. From the University of Brasília, 29 professors were correctly classified and 2 were not identified. In other affiliations, 3,009 authors were classified correctly, 1 was misclassified, and 237 were not identified. This totals 3,038 hits, 1 error and 239 unidentified authors from the 3,269 authors, which made the function accurate to around 93%. Names not identified, or incorrect, by the function were inserted manually, after confirming the author's gender on

[1]https://pypi.org/project/beautifulsoup4/

[2]https://pypi.org/project/gender-guesser/

Google.

TABLE I
TABLE WITH THE RESULTS OF THE FINAL GENDER FUNCTION.

Author by Affiliation	Hits	Errors	Unidentified	Total
University of Brasília	29	0	2	31
Others (Journals)	1105	1	126	1217
Others (Conferences)	1904	0	111	2015
Total	3038	1	239	3269

To insert the data obtained into Neo4J, files were created to store the Cypher commands generated by a *script* in *Python*. These Cypher commands were inserted through the Neo4j Desktop interface, obtaining the graphs represented in Figures 3 and 4, which represent the journal and conference, respectively.

The journal graph is composed of 2004 nodes and 3,092 relationships, represented in Tables II and III. The conference graph is composed of 4,859 nodes and 7,682 relationships, represented in Tables IV and V.

Nodes classified as *'Author'* represent both the professors advising the program and the authors who collaborated with the professors in the program. The *'Institution'* node represents the University of Brasília and, finally, the *'Paper'* node represents all articles published by professors from the program. All these nodes can be seen in Tables II, and IV, which represent the journals and conferences respectively.

The *'Authoring'* relationship represents the relationship of the *'Author'* node with the *'Paper'* node, representing the authors who were involved in creating the article. The relationship *'Associated_to'* represents the relationship between the node *'Author'* and the node *'Institution'*, representing the relationship between the advising professors and the PPGI at the University of Brasília. As there are 31 professors in the program, there are only 31 *'Associated_to'* relationships. All these relationships can be seen in Tables III and V. This node and relationship information is illustrated in Figures 3 and 4.

TABLE II
NODES IN THE GRAPH GENERATED FROM ARTICLES PUBLISHED IN JOURNALS.

Label	Quantities
Author	1282
Institution	1
Paper	729
Total	2004

TABLE III
RELATIONSHIP TABLE IN THE GRAPH GENERATED FROM ARTICLES PUBLISHED IN JOURNALS.

Label	Quantities
Authoring	3061
Associated_to	31
Total	3092

With the generated graphs, it is possible to create Cypher queries to check the research carried out by the University of Brasília researchers, filtering by gender. However, to obtain

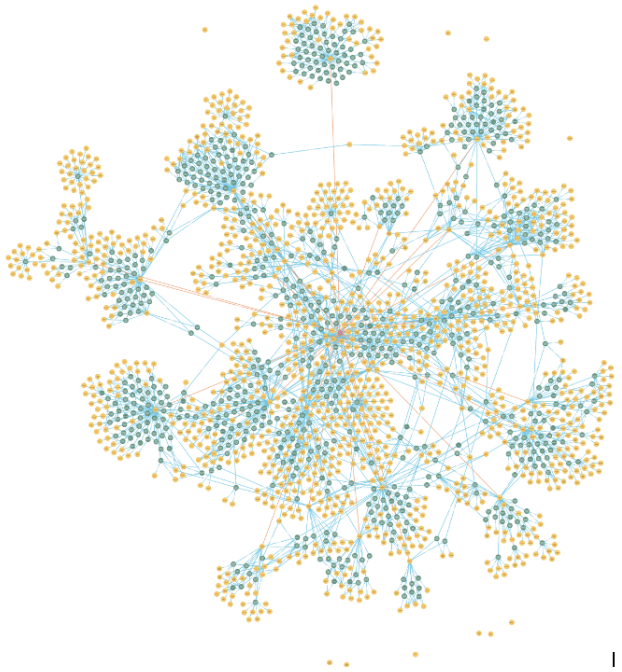

Fig. 3. Graph generated with sample data from articles published in journals.

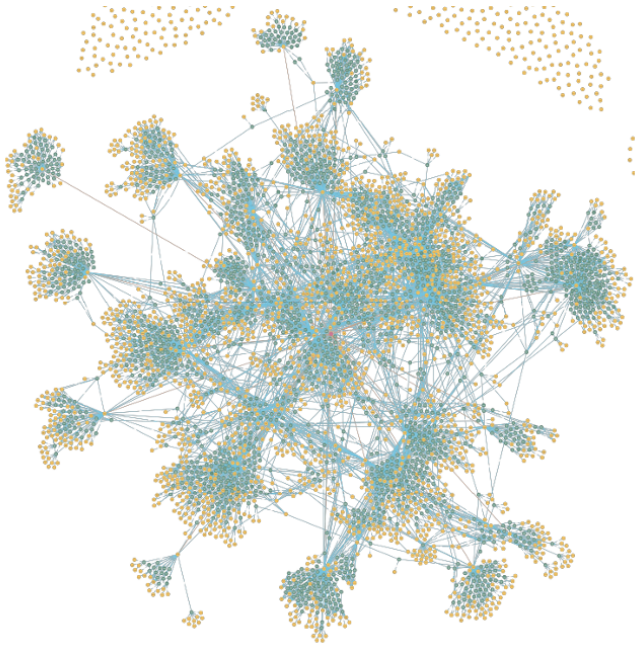

Fig. 4. Graph generated with sample data from papers published in conferences.

TABLE IV
NODES IN THE GRAPH GENERATED FROM PAPERS PUBLISHED IN CONFERENCES.

Label	Quantities
Author	2783
Institution	1
Paper	2075
Total	4859

TABLE V
RELATIONSHIP TABLE IN THE GRAPH GENERATED FROM PAPERS PUBLISHED IN CONFERENCES.

Label	Quantities
Authoring	7651
Associated_to	31
Total	7682

the collaboration network, it was necessary to create a new relationship, the co-authorship relationship, which links an Author node regardless of its affiliation with an Author node affiliated with the University of Brasília, as long as these two have collaborated somehow. In the case of two University of Brasília authors who have collaborated together, the relationship is created twice, so that each author has an *outgoing* relationship.

The relationship generated 2,891 connections for journals, 7,090 connections for conferences. This makes it possible to analyze author collaborations.

V. RESULTS

As the objective of the research is to verify whether, regardless of the participant's gender, the number of articles produced is similar, the average search metric by gender is important. The average of a graph can be calculated by its average degree, represented by Equation 1. The equation presented calculates the average degree of the nodes, represented by d(G), dividing the graph nodes, represented by the symbol n and 'nodes' in the equation, by the number of relationships that these nodes have, represented by the symbol 'rn' and by 'Relationships(nodes)' in Equation 1.

$$d(G) = \frac{relationships(nodes)}{nodes} = \frac{rn}{n} \quad (1)$$

It starts by looking for the average by gender. The first query created retrieved the list of Participants (PPGI professors), filtering by gender, and their research.

For journals, the data present in Table VI. For this type of paper, the average graph degree of women is slightly higher than that of men, but this difference is not very significant. For conferences, the data presented in Table VII. For conferences, the average number of searches per woman is 107 while that of men is 67.4.

To understand the Participants' co-authorship pattern, the external collaboration network and the internal collaboration network are obtained. For the external network, all authors related to the Participant must not be associated with the University of Brasília. The internal network is only the authors

TABLE VI
LIST OF PARTICIPATING AUTHORS AND THEIR RESEARCH PUBLISHED IN JOURNALS.

	Female	Male
University of Brasília	10	21
Researches	256	523
Authorship Relationships	294	548
d(G)	29,4	26,1

TABLE VII
LIST OF PARTICIPATING AUTHORS AND THEIR RESEARCH PUBLISHED AT CONFERENCES.

	Female	Male
University of Brasília	10	21
Researches	908	1300
Autorship Relationships	1070	1416
d(G)	107	67,4

associated with the University of Brasília and related to the Participant. Based on the results of the journals' internal network, only 17 of the 21 male Participants collaborate internally to the program, while 9 out of 10 female Participants collaborate internally. In other words, with reference to journals, of the program participants, 5 do not contribute internally to the program, and only one of these five is female.

Table VIII was generated from all the results obtained throughout this section, both from collaboration networks and research. With these data, it is clear that of the 10 women in the PPGI program, only one does not collaborate internally at all, in either journals or conferences. For men, only 1 does not contribute internally to the program at all, 2 do not contribute internally to conferences and 4 do not contribute internally to journals. Analyzing the collaboration networks, it is clear that, in all cases, female networks have a greater degree of collaboration than male networks, which implies that the average number of female collaborations is proportionally greater than that of men.

TABLE VIII
D(G) RESEARCH AND COLLABORATION NETWORKS

	Female			Male		
	Journal	Conf.	Total	Journal	Conf.	Total
Researches	29,4	107	136,3	26,1	67,4	93,6
Internal	23,8	77,6	101,8	12,5	42	54,5
External	102	310	412	77,1	175,3	242,1
General	141,2	415,3	452,8	101,3	241,9	352

VI. CONCLUSION

This work presented an analysis of data from the Graduate Program in Informatics at the University of Brasília, using a graph-oriented NoSQL database to generate the collaboration network of the PPGI faculty advisors at the University of Brasília.

As shown, the average production of journals for PPGI's female advisors is numerically similar to the production of male advisors, while internally for journals, the proportion of contributions from men is also lower.

For conferences and in the case of all data, the average number of articles produced by women is higher than that of men for this sample. Therefore, for these data, women contribute proportionally more than men.

Collaboration networks for journals, conferences and general data present the same result, which is that women proportionally collaborate more than men, both internally and externally, in academic productions.

Internally, for the types of articles studied, only two participants did not collaborate internally, one of them being a man and the other a woman.

For future work, this study will be expanded to other Brazilian universities to verify whether the average amount of research and collaborations is similar for women and men in the computing area.

REFERENCES

[1] Wu, Jue, and David H. Uttal. "Diversifying computer science: An examination of the potential influences of women-in-computing groups." Science Education 108.3 (2024): 957-980.

[2] Y. Kovaleva, J. Kasurinen, E. Kindsiko and A. Happonen, "State-of-the-Art Review on Current Approaches to Female Inclusiveness in Software Engineering and Computer Science in Higher Education," in IEEE Access, vol. 12, pp. 1360-1373, 2024, doi: 10.1109/ACCESS.2023.3346767

[3] M. Holanda, A. de Araújo, M. Walter, and C. de Oliveira. " Meninas.comp: Um Relato da Experiência de Integração entre o Ensino Médio e a Universidade de Brasília", in Anais do X Women in Information Technology, Porto Alegre, 2016, pp. 78-82, doi: https://doi.org/10.5753/wit.2016.9706.

[4] M. Abrahão Amaral, M. Figueiredo Pereira Emer, S. Amélia Bim, M. Gomes Setti and M. Mikosz Gonçalves, "Investigando questões de gênero em um curso da área de Computação", Revista Estudos Feministas, vol. 25, no. 2, pp. 857-874, 2017. doi: https://doi.org/10.1590/1806-9584.2017v25n2p857.

[5] Gomes, R., 2022. UnB Notícias - Presença de mulheres em profissões 'tipicamente masculinas' é tema de mesa-redonda. [online] Presença de mulheres em profissões 'tipicamente masculinas' é tema de mesa-redonda. Available at: <https://noticias.unb.br/112-extensao-e-comunidade/4837-presenca-de-mulheres-em-profissoes-tipicamente-masculinas-e-tema-de-mesa-redonda> [Accessed 11 April 2024].

[6] T. Nunes, A. Araujo, and M. Holanda. "Mulheres na Pós-graduação nas Áreas de Exatas: Um Estudo de Caso na Universidade de Brasília", in Anais do XIV Women in Information Technology, Cuiabá, 2020, pp. 244-248, doi: https://doi.org/10.5753/wit.2020.11303.

[7] Programa de Pós-Graduação em Informática (PPGI) da University of Brasilia (UnB). 2022. Docentes. [online] Available at: <http://ppgi.unb.br/index.php?option=com_content&view=article&id=78&Itemid=471&lang=pt> [Accessed 8 April 2022].

[8] Mena-Chalco, J. P., & Junior, R. M. C. (2009). ScriptLattes: an open-source knowledge extraction system from the Lattes platform. Journal of the Brazilian Computer Society, 15, 31-39

[9] N. Platform, "Best practices when choosing relationship direction\ name?", Neo4j Online Community, 2022. [Online]. Available: https://community.neo4j.com/t/best-practiceswhenchoosingrelationship-directionname/7902/2. [Accessed: 25 Apr 2022].

[10] Strauch, C., Sites, U. L. S., & Kriha, W. (2011). NoSQL databases. Lecture Notes, Stuttgart Media University, 20(24), 79.

[11] N. Rodrigues and C. Ralha. "Conhecendo a Comunidade de Sistemas de Informação no Brasil: um Estudo Comparativo Utilizando Diferentes Abordagens de Banco de Dados", in Anais do XI Simpósio Brasileiro de Sistemas de Informação, Goiânia, 2015, pp. 555-562, doi: https://doi.org/10.5753/sbsi.2015.5861.

[12] Neo4j Graph Data Platform. 2022. Why Neo4j? Top Ten Reasons. [online] Available at: <https://neo4j.com/top-ten-reasons/#:~:text=Neo4j%20delivers%20the%20lightning%2Dfast,predictability%20of%20relationship%2Dbased%20queries.> [Accessed 11 April 2022].

[13] L. M. Camarinha-Matos and H. Afsarmanesh, "Collaborative Networks", IFIP International Federation for Information Processing, vol. 207, pp. 26-40, 2006. doi: https://doi.org/10.1007/0-387-34403-9_4

[14] L. Ribeiro, G. Barbosa, I. Silva, F. Coutinho, and N. Santos. "Um Panorama da Atuação da Mulher na Computação", in Anais do XIII Women in Information Technology, Belém, 2019, pp. 1-10, doi: https://doi.org/10.5753/wit.2019.6707.

[15] A. Lorens, J. Botelho, A. Moura, B. Duarte, and M. Moro. "Participação Feminina em Comitês de Programa de Simpósios da Computação", in Anais do XIV Women in Information Technology, Cuiabá, 2020, pp. 90-99, doi: https://doi.org/10.5753/wit.2020.11279.

[16] A. Costa and C. Ralha. "Análise de Rede de Colaboração Científica como Ferramenta na Gestão de Programas de Pós-graduação", in Anais do XI Simpósio Brasileiro de Sistemas de Informação, Goiânia, 2015, pp. 275-282, doi: https://doi.org/10.5753/sbsi.2015.5827.

[17] M. Holanda and A. Araujo. "Pós-graduação em Computação na Universidade de Brasília: Um Grande Desafio na Diversidade de Gênero", in Anais do XIII Women in Information Technology, Belém, 2019, pp. 169-173, doi: https://doi.org/10.5753/wit.2019.6731.

[18] Conselho Nacional de Desenvolvimento Científico e Tecnológico. 2022. Dia Internacional de Mulheres e Meninas na Ciência. [online] Available at: <https://www.gov.br/cnpq/pt-br/assuntos/noticias/destaque-em-cti/dia-internacional-de-mulheres-e-meninas-na-ciencia> [Accessed 2 April 2022].

[19] Blog.brasil.io. 2022. Classificando Nomes por Gênero Usando Dados Públicos — Brasil.IO - Blog. [online] Available at: <https://blog.brasil.io/2019/05/31/classificando-nomes-por-genero-usando-dados-publicos/index.html> [Accessed 6 April 2022].

[20] Neo4j Graph Data Platform. 2022. Cypher Query Language - Developer Guides. [online] Available at: <https://neo4j.com/developer/cypher/> [Accessed 11 April 2022].

[21] A. Corbellini, C. Mateos, A. Zunino, D. Godoy and S. Schiaffino, "Persisting big-data: The NoSQL landscape", Information Systems, vol. 63, pp. 1-23, 2017. doi: https://doi.org/10.1016/j.is.2016.07.009 [Accessed 25 April 2022].