VOL.39 Nº2 2022
https://doi.org/10.22491/IMECTA.10853.en

Comparative analysis of topological metrics in co-authorship networks for the link prediction problem

***Mariana Magalhães de Mattos Coelho, Claudia Marcela Justel**
Engenharia de Computação, Instituto Militar de Engenharia
Praça General Tibúrcio 80, 2229-270, Praia Vermelha, Rio de Janeiro, RJ, Brasil.
***mariana@ime.eb.br**

ABSTRACT: The link prediction problem consists in estimating the appearance of edges between nodes of graphs representing a network of interconnected elements (e.g. a co-authorship network whose vertices and edges represent, respectively, the authors and the publications between them). In the last years, several approaches to solve this problem were proposed. Among them, one line of work is considering the topology of the network. This paper is about different metrics used to solve the topological approach of the link prediction problem. Our goal is to compare 4 different metrics by conducting experiments in a collaboration network. We present the results and conclusions obtained with the experiments for a real network developed by students, working in the research project Graph Algorithms.

KEYWORDS: social network analysis, link prediction, graph applications.

RESUMO: O problema denominado predição de links *consiste em estimar o surgimento de arestas entre nós de um grafo que representa uma rede de elementos interligados (e.g., uma rede de coautoria cujos vértices e arestas representam, respectivamente, os autores e as publicações entre eles). Diversas abordagens para resolver esse problema foram propostas nos últimos anos. Dentre as diferentes abordagens existentes, neste trabalho consideramos a abordagem topológica, para a qual foram definidas diferentes métricas. O objetivo deste trabalho é comparar quatro métricas topológicas por meio de experimentos em uma rede de coautoria. Apresentamos os resultados e as conclusões obtidas a partir dos experimentos executados em uma rede real desenvolvida por alunos participantes do projeto de pesquisa Algoritmos em Grafos.*

PALAVRAS-CHAVE: análise de redes sociais, predição de links, aplicações de grafos.

1. Introduction

This study addresses the link prediction problem, which is a fundamental problem in the area of social network analysis, with applications in different domains, such as predicting the evolution in dynamic networks, indicating new friendships in social networks, recommending products and services [1], among others. A co-authorship network is represented by vertices and edges, where nodes are the authors and edges are the publications between them.

The link prediction problem aims to identify the link between pairs of nodes for which this connection does not exist. Different solutions can solve this problem. Some of them use the characteristics or attributes of the nodes while others use only the structural graph information. The former are known as characteristic-based approaches and the latter topological approaches. Other approaches use both characteristics of the nodes and structural information—they are called hybrid approaches [2].

In their study, Liben-Nowell and Kleinberg [3] introduces different topological metrics to solve the problem of predicting links that associate a coefficient, which is called score (x,y), with a pair of unconnected nodes x, y of a G graph at a time t. After that, a list ordered by the score (x,y) values is produced to create a predictor of new connections.

Nassar et al. [4] propose a new topological approach to predict links called “pairwise prediction,” which, instead of considering a pair of nodes, determines which node is most likely to form a triangle with an existing edge. Thus, they introduce a new version of the metrics, as proposed by [3].

Both types of metrics, which we refer as “traditional” and “pairwise” versions, are topological metrics to predict links for homogeneous networks[1][5]. However, to our knowledge, the traditional and pairwise versions of metrics proposed by [3] and [4], respectively, were not compared between them; no study compares the performance of these two versions of the metrics in the same real network. Thus, we develop experiments that allow identifying advantages and disadvantages of the two versions (traditional and pairwise) is important.

This study aims to conduct a comparative analysis of the performance of different topological metrics in traditional and pairwise versions in co-authorship networks. We used the dataset of a co-authorship network, which was generated from information of the Lattes platform of the National Council for Scientific and Technological Development (CNPq), to perform experiments that allowed us to compare the two versions and, later, we analyzed the results obtained. These experiments showed that the pairwise version has a small advantage over the traditional version.

In this study, Section 2 presents studies related to our research, Section 3 addresses the link prediction problem, as well as the metrics introduced by [3] and [4] and the definitions of the concepts used in this study, Section 4 describes the methodology used to compare the metrics in the two versions aforementioned, Section 5 presents the experiments with their respective results, and, finally, Section 6 presents the conclusion of this study.

2. Related studies

Liben-Nowell and Kleinberg [3] published an important study on link prediction, in which they analyze a co-authorship network using topological characteristics of the network to predict the edge formation between two unconnected nodes. In this case, the co-authorship network is homogeneous.

The authors of [6] proposed an adaptation of the traditional metrics in the homogeneous networks addressed by [3] in order to use them in heterogeneous bipartite networks, that is, networks in which nodes are of two different types (a bipartition of the set of nodes) and all edges have ends in different sets of the bipartition. The authors proposed to transform the set $\Gamma\ (u)$ of neighbors of u into $\Gamma\ '(u) = \cap_{\nu\ \in\ \Gamma(\upsilon)} \Gamma(\nu)$ (neighbors of the neighbors of node u), that is, replace $\Gamma\ (u)$ by $\Gamma\ '(u)$ when estimating the traditional topological metrics.

In 2010, [7] presented a solution different from [6] to deal with heterogeneous bipartite networks. In this case, the authors used a projection of the graph representing the network over one of the two sets of the bipartition and defined the metrics according to this projection. Figure 1 shows two projections as an example of a bipartite graph, according to [7].

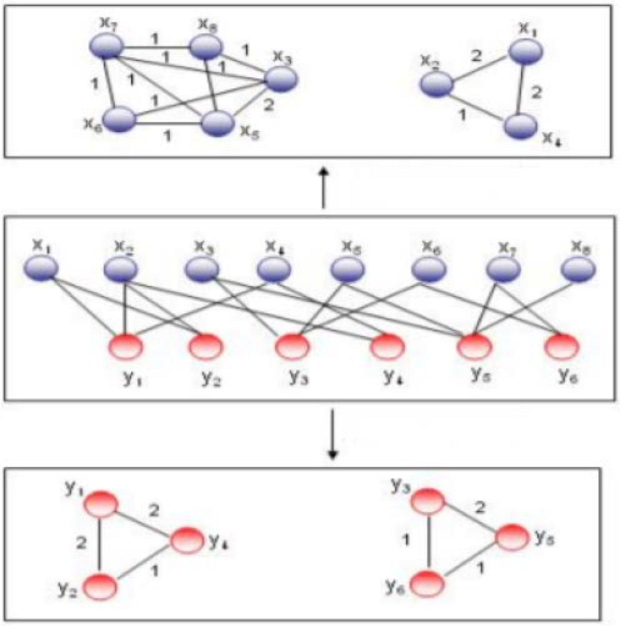

Figure 1. Two projections of the bipartite graph $G = (X \cup Y, E)$.
Source: [7]

The authors of [8] developed a solution for a heterogeneous multipartite network. They proposed the MRLP (multi-relational link prediction) method, in which the main component uses a weight scheme for different types of combinations of edges, from the number of subgraphs formed by three nodes existing in the network.

In 2019, Nassar *et al.* [4] proposed to predict the edge formation considering one node and one edge

[1] Homogeneous networks have a single type of nodes and links.

existing in the network. Studies [3] and [4] address triadic closure in different ways when analyzing subgraphs formed by three nodes in homogeneous networks (Figures 2 and 3).

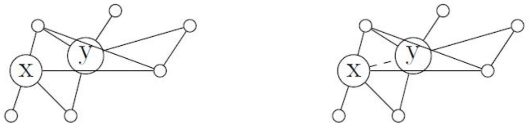

Figure 2. The drawing on the left represents *G* at a time *t* and the existing links until that instant (*x* and *y* are non-adjacent vertices); the drawing on the right represents *G* at a time *t'* and the existing links until that instant, for $t < t'$. The dashed edge would be a possible link between *x* and *y* at the instant *t'*, using metrics for the traditional link prediction. **Source: [4]**

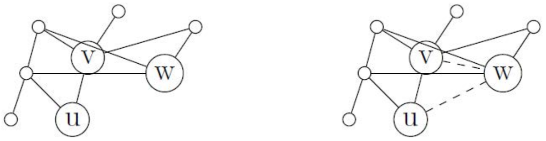

Figure 3. The drawing on the left represents *G* at a time *t* and the existing links until that instant (*u* and *v* are adjacent); the drawing on the right represents *G* at a time *t'* and the existing links until that instant, for $t < t'$. Dashed edges would be possible links between *u* and *w* and between *v* and *w*, using metrics for the pairwise link prediction. **Source: [4]**

3. Two topological approaches

Social networks are highly dynamic objects, which grow and change rapidly over time by the addition of new edges, according to the emergence of new interactions in the original network. The link prediction problem is related to the evolution of a social network over time. Given an snapshot of a social network at a time *t*, the link prediction problem seeks to predict with some precision edges that will be added to that network during the time interval from *t* to a future time *t'* [3].

All metrics considered by [3] for link prediction associate a coefficient with pairs of non-adjacent nodes *x, y*, which is called score *(x, y)*, from an input graph and produce an ordered list in non-ascending order of these coefficients. Coefficients can be considered a measure of proximity or similarity between a pair of nodes, being called **traditional metrics**. These metrics were adapted from some techniques used in graph theory and social network analysis. They were generally not designed to estimate the similarity between nodes in a graph, thus, modifying them for their new purpose was necessary.

The notation used was: *G=(V,E)* an undirected graph; $x \in V$ a node; Γ *(x)*= $\{y \in V : (x,y) \in E\}$ the set of neighbors of the node $x \in V$; and $|\Gamma(x)|$ the cardinality of the set Γ *(x)*. Table 1 presents four of the traditional topological metrics used by [3].

Table 1 – Metrics in the traditional version.

Traditional metrics	
CN(x,y) =	$\lvert\Gamma(x) \cap \Gamma(y)\rvert$
JS(x,y) =	$\frac{\lvert\Gamma(x) \cap \Gamma(y)\rvert}{\lvert\Gamma(x) \cup \Gamma(y)\rvert}$
AA(x,y) =	$\sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{log\lvert\Gamma(z)\rvert}$
PA(x,y) =	$\lvert\Gamma(x)\rvert \, \lvert\Gamma(y)\rvert$

Source: [3]

Nassar et al. [4] states that the traditional link prediction, as presented by [3], can be described by the following question: "given a node *x* in the network, which nodes are most likely to be linked to it?" They consider a new version of the metrics from the following question: "given an edge *(u, v)* in the network, which nodes are most likely to connect to the ends of the edge (the vertices *u* and *v*)?" Aiming to define a new version of the proximity or similarity between each vertex from the ends of the edge and the node—which the authors call "pairwise" and we call **pairwise metrics**—the following notation was used:

$\Gamma*((u,v)) = \{z \in V : \Gamma(u) U \Gamma(v) - \{u,v\}\}$ the set of neighbors in the edge $(u,v) \in E; |\Gamma*((u,v))|$ the cardinality of the set $\Gamma*((u,v))$. The metrics adapted by the pairwise version were common neighbors (*CN**), Jaccard similarity (*JS**), Adamic-Adar (*AA**), and Preferential Attachment (*PA**). Table 2 presents the four pairwise topological metrics.

Table 2 – Metrics in the pairwise version.

Pairwise metrics	
$CN^*\,(w,(u,v)) =$	$\lvert \Gamma(w) \cap \Gamma^*((u,v) \rvert$
$JS^*\,(w,(u,v)) =$	$\dfrac{\lvert \Gamma(w) \cap \Gamma^*((u,v)) \rvert}{\lvert \Gamma(w) \cup \Gamma^*((u,v)) \rvert}$
$AA^*\,(w,(u,v)) =$	$\sum_{z \in \Gamma(w) \cap \Gamma^*((u,v))} \dfrac{1}{\log \lvert \Gamma(z) \rvert}$
$PA^*\,(w,(u,v)) =$	$\lvert \Gamma(w) \rvert\lvert \cap \Gamma^*((u,v)) \rvert$

Source: [4]

4. Methodology

In this study, a graph $G=(V,E)$, in which each edge $e \in E$ represent an interaction between two nodes u and v at an instant of time $t(e)$ in a social network, was not considered. Multiple interactions between u and v were considered. For a given instant of time t, G_t showed a subgraph of G with all edges, for $t(e) \leq t$. The mathematical formulation of the problem is given below. Two instants of time, $t < t'$, were chosen and an algorithm that accesses the graph that represented the network until instant t, G_t, and returns a list of pairs of elements (two non-adjacent nodes or a node and an edge in G_t), which are predictions of edges to $G_{t'}$, was considered. The intervals $(0,t]$ and $(t,t']$ were considered training and test intervals, respectively.

Each predictor p considered returns an L_p ordered list of pairs in $V \times V$, which is formed by predictions of new interactions in $G_{t'}$ in non-increasing trust order.

The performance measure for predictor p is assessed by choosing the first k pairs of predictions of new interactions from the L_p ordered list (top-k). Therefore, to compare each metric M_i in the traditional ($i = 1$) and pairwise ($i = 2$) versions, the top-k, that is, the first k elements of the $L_p(M_i)$ list, shall be built from the predictor $p(M_i)$, for $i = 1, 2$. Finally, for each $L_p(M_i)$ list, the measures of the quality of classification of the predictor $p(M_i)$, $i = 1, 2$, are determined using Equations 1 to 4:

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

$$F1 = \frac{2\ Precision.Recall}{Precision + Recall} \tag{4}$$

where

- TP = true positive (a link is classified, or predicted, as positive and exists in graph $G_{t'}$);
- FP = false positive (a link is classified, or predicted, as positive and does not exist in graph $G_{t'}$);
- FN = false negative (a link is classified, or predicted, as negative and exists in graph $G_{t'}$);
- TN = true negative (a link is classified, or predicted, as negative and does not exist in graph $G_{t'}$);

According to the type of network in which using the traditional and pairwise metrics performed in Section 3 is possible, a dataset was chosen, corresponding to a co-authorship network, to perform the comparative experiments. The dataset description is presented below.

When determining which dataset to use in the experiments, both traditional and pairwise metrics showed not to produce good results in heterogeneous bipartite networks. Some preliminary experiments were performed, showing that the adaptation proposed by [6] does not work with pairwise metrics. Therefore, a co-authorship network, which is a homogeneous network, was chosen as dataset. The co-authorship network can be understood as the projection presented by [7], in the set of authors of heterogeneous bipartite networks that relates authors and studies.

In [9], a co-authorship network was created from data collected on October 11, 2010, in the CNPq Lattes platform. The nodes of the graph representing the network correspond to the authors, the edges correspond to at least one joint publication between two authors (thus, the graph is undirected, without multiple edges), and *t(e)* corresponds to the year of the first joint publication between vertices, which are the ends of edge *e*.

Later, the network, created in [9], was updated with data until October 10, 2014 [10]. Both the creation of the network and its first update were developed by undergraduate students in the Graph Algorithms project as part of their final course work and scientific initiation project. Figures 4 and 5 show graphs G_{2011} and G_{2014}, which represent the two co-authorship networks aforementioned.

In the context of this study, a new data update was performed on January 5, 2021. Table 3 shows the sizes of the sets of nodes and edges of graphs G_{2011}, G_{2014}, and G_{2020}, with information corresponding to the periods (2010, 2014] and (2014, 2020], respectively. Figure 6 presents graph G_{2020}.

Table 3 – Information of co-authorship networks obtained from the CNPq Lattes platform.

$G = (V, E)$	$\lvert V \rvert$	$\lvert E \rvert$
G_{2011}	207	520
G_{2014}	280	756
G_{2020}	425	1,350

All experiments were performed on an Intel Core i7 computer, CPU 1.80GHz, 8GB RAM. The Python programming language (version 3.9.4 [11]) was used to implement metrics, determine predictors, and obtain the top-*k* and the NetworkX 2.5.1 library was used to manipulate graphs [12].

5. Description of the experiments

5.1 CNPq Lattes Experiment 2011–2014

In this first experiment, we used the Lattes co-authorship network to produce graphs G_t and $G_{t'}$ with $t < t'$, which represent the network, for $t = 2011 < t' = 2014$. G_{2014} presented 32 edges between pairs of non-adjacent nodes in G_{2011}, that is, 32 new edges. The top-*k* values obtained with $k = 3$, 5, 7 showed small differences, thus, we show only the results for $k = 7$. We chose the value $k = 7$ because the traditional Common Neighbors metric presented only nine different score values. Table 4 presents all values of the measures of the quality of classification for the predicted links obtained in this experiment.

Table 4 shows that, for Common Neighbors and Jaccard similarity, the results obtained by the pairwise version were better or equal, except for the recall for Common Neighbors. For Preferential Attachment, the traditional version presented better results in all cases. And for Adamic-Adar, the results were balanced for both versions of the metrics.

Table 4 – Results for G_{2011} and G_{2014}.

	Top-7	
	CN	*CN**
Precision	0.01010101	0.025
Accuracy	0.965818951	0.994732053
F1	0.019310345	0.035087719
Recall	0.21875	0.058823529
	JS	*JS**
Precision	0	0
Accuracy	0.985241094	0.992497372
F1	0	0
Recall	0	0
	PA	*PA**
Precision	0.142857143	0
Accuracy	0.998221239	0.998029793
F1	0.051282051	0
Recall	0.03125	0
	AA	*AA**
Precision	0.142857143	0.058823529
Accuracy	0.998221239	0.996928246
F1	0.051282051	0.058823529
Recall	0.03125	0.058823529

5.2 CNPq Lattes Experiment 2014–2020

In this experiment, we used the Lattes co-authorship graph to produce the information necessary to work with $t < t'$, for $t = 2014$ and $t' = 2020$. Results were evaluated for different top-*k* values ($k = 3, 5, 7$).

G_{2020} presented 105 edges between pairs of non-adjacent nodes in G_{2014}, that is, 105 new edges. The top-*k* values showed small differences between the results obtained, thus, only the results for $k = 7$

were showed. We chose the value $k = 7$ because the traditional Common Neighbors metric presented only eight different score values. Table 5 presents all values of the measures of the quality of classification of the predicted links obtained in this experiment.

Table 5 shows that, for Common Neighbors and Jaccard similarity, the results obtained by the pairwise version were better when compared with those obtained by the traditional version, except for the recall. For Preferential Attachment, the value for the two versions was zero regarding precision, F1, and recall. On the other hand, regarding accuracy, the traditional Preferential Attachment metric was a little better in comparison with the pairwise version. And for Adamic-Adar, the results obtained by the pairwise version showed a small advantage, except for accuracy.

Table 5 – Results for G_{2014} and $G_{2020.}$

	Top-7	
	CN	*CN**
Precision	0.00791526	0.033950617
Accuracy	0.776524645	0.989178834
F1	0.015639374	0.05
Recall	0.647619048	0.094827586
	JS	*JS**
Precision	0.002932551	0.006024096
Accuracy	0.988408521	0.992981544
F1	0.004484305	0.007352941
Recall	0.00952381	0.009433962
	PA	*PA**
Precision	0	0
Accuracy	0.99707602	0.997050532
F1	0	0
Recall	0	0
	AA	*AA**
Precision	0	0.023255814
Accuracy	0.99707602	0.996166584
F1	0	0.013422819
Recall	0	0.009433962

6. Conclusion

In this work, we presented the results of comparative experiments between two versions of the link prediction metrics (traditional and pairwise) in a real co-authorship network.

The results obtained showed that for the four metrics considered (common neighbors, Jaccard similarity, Preferential Attachment, and Adamic-Adar), the two versions presented similar behaviors, with a small advantage to the pairwise version. In all experiments, for common neighbors, the pairwise version showed a slight improvement regarding precision, accuracy, and F1. For Jaccard similarity, the results were equal or better regarding precision, accuracy, and F1 while for Adamic-Adar, F1 and recall improved.

For Preferential Attachment, in the experiment from 2014 to 2020, both the traditional version and the pairwise version presented values equal to zero regarding precision, F1, and recall. However, in the experiment from 2011 to 2014, the values of these three quality measures were different from zero for the traditional version.

The preliminary results show that the pairwise version, which was recently introduced, can also be used to solve the topological approach of the link prediction problem in the case of three of the metrics used in this article.

Future studies shall continue to analyze the results for the arXiv preprint repository datasets in five astrophysics fields [13], which have already been used in the literature to compare the results of link prediction methods. General Relativity and Quantum Cosmology (gr-qc), Astrophysics (astro-ph), Condensed Matter (cond-mat), High Energy Physics – Phenomenology (hep-ph), and High Energy Physics – Theory (hep-th) are datasets to be considered. In this future analysis, we also intend to include experiments using an altered pairwise version, originally proposed by the same authors of [4] ([14]).

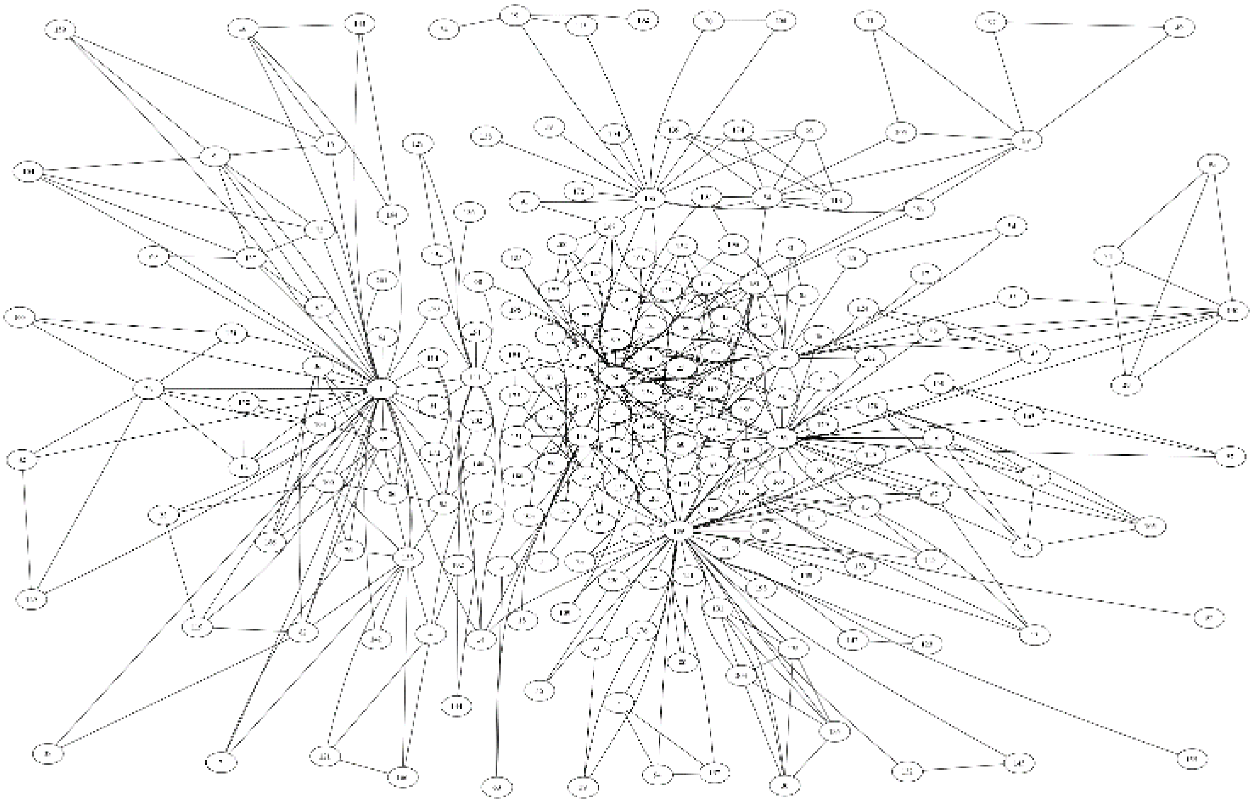

Figure 4. CNPq Lattes graph G_{2011}. **Source: [9]**

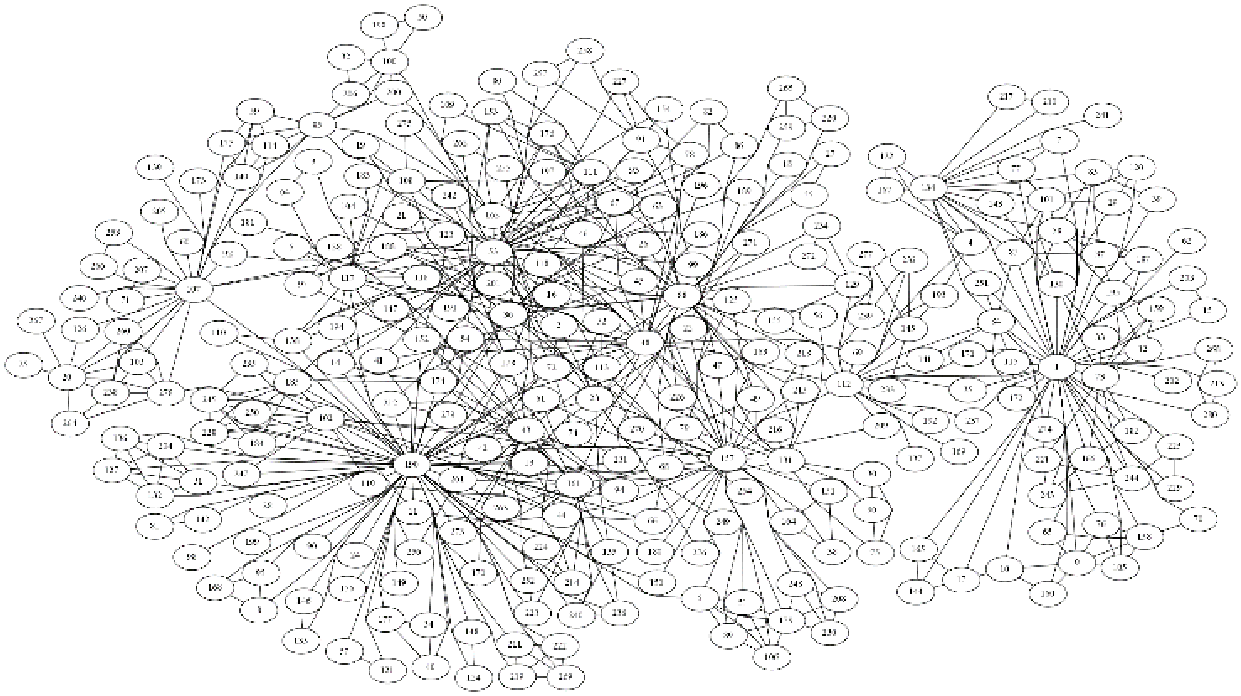

Figure 5. CNPq Lattes graph G_{2014}. **Source: [10]**

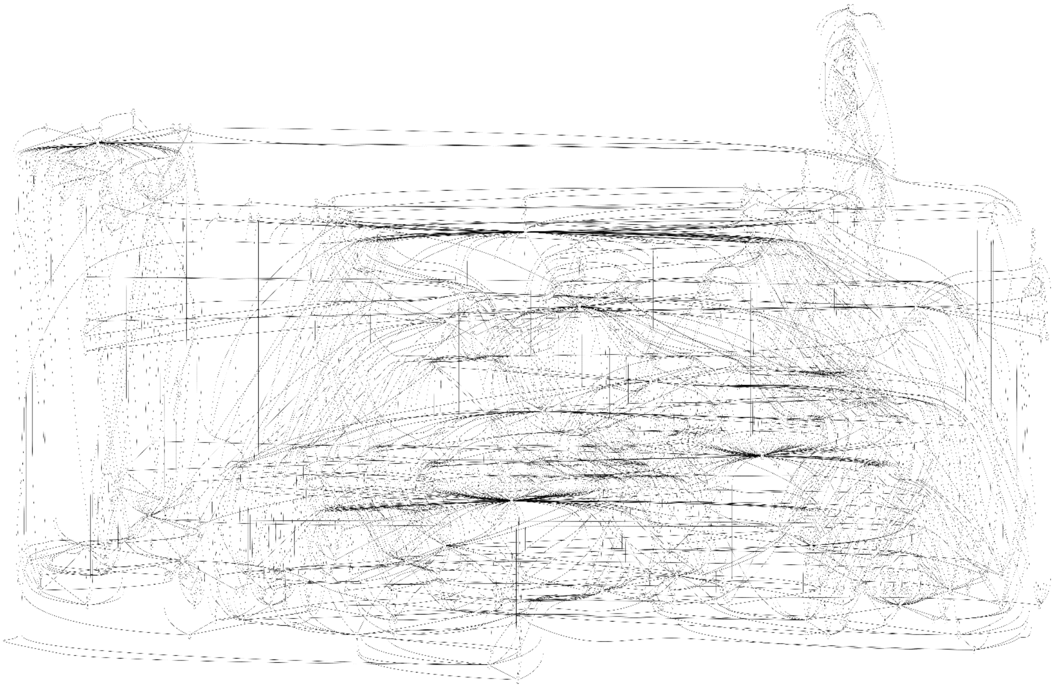

Figure 6. CNPq Lattes graph G_{2020}.

References

[1] ZAREIE, A.; SAKELLARIOU, R. Similarity based link prediction in social networks using latent relationships between the users. Scientific Reports, v. 10, n. 20137, p. 1–11, 2020.

[2] PUJARI, M. Link Prediction in Large-scale Complex Networks (Application to bibliographical Networks). Paris: University Sorbonne Paris Cité, 2015.

[3] LIBEN-NOWELL, D.; KLEINBERG, J. The Link-Prediction Problem for Social Networks. Journal of the American Society for Information Science and Technology, v. 58, n. 7, p. 1019–1031, 2007.

[4] NASSAR, H.; BENSON, A. R.; GLEICH, D. F. Pairwise Link Prediction. arXiv:1907.04503v1 [cs.SI]. 10 Jul 2019.

[5] AL HASAN, M.; ZAKI, M. J. A survey of link prediction in social networks. In AGGARWAL, C. C. Social Network Data Analytics. New York: Springer, 2011. p. 243–275.

[6] HUANG, Z.; LI, X.; CHEN, H. Link prediction approach to collaborative filtering. In ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL). New York: ACM Digital Library, 2005. p. 141–142.

[7] BENCHETTARA, N.; KANAWATI, R.; ROUVEIROL, C. Supervised machine learning applied to link prediction in bipartite social networks. In International Conference on Advances in Social Networks Analysis and Mining. Piscataway: IEEE, 2010.

[8] DAVIS, D.; LICHTENWALTER, R.; CHAWLA, N. V. Supervised methods for multi-relational link prediction. Social Network Analysis and Mining, v. 3, p. 127–141, 2013.

[9] BARBOSA, D. A. B. L.; AVELINO, L. B.; SOUZA, R. F.; OLIVEIRA, C. C. G. F.; JUSTEL, C. M. Medidas de centralidade e detecção de comunidades em rede de co-autoria. In Anais do XVIII Simpósio Brasileiro de Pesquisa Operacional. Rio de Janeiro: Sobrapo, 2011. p. 2574–2583.

[10] MAGNANI, H. M. Redes Sociais e Comunidades. Relatório Final Projeto Institucional de Iniciação Científica CNPq-IME, 2014.

[11] PYTHON. Disponível em: https://www.python.org/. Acesso em: 6 abr. 2021.

[12] NETWORKX. Disponível em: http://networkx.org. Acesso em: 6 abr. 2021.

[13] ARXIV. Disponível em: https://arxiv.org/. Acesso em: 6 abr. 2021.

[14] NASSAR, H.; BENSON, A. R.; GLEICH, D. F. Pairwise Link Prediction. In IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM). New York: ACM Digital Library, 2019. p. 386–393.