Link Prediction on Paraguay's Public Tender Participants

Matias Romero Moriya
Master in Data Science, Rice University
Houston, TX 77005
msr10@rice.edu

Abstract

In many countries, the government stands as one of the largest entities engaging in the procurement of services and the purchase of products. The initiation of this procurement phase typically involves a tendering process. Unfortunately, there are instances where the number of companies participating in these processes is limited. This challenge could be addressed by implementing a strategy wherein companies already providing services to other government institutions are made aware of additional tender opportunities. This approach can be conceptualized as a link prediction problem, where the goal is to forecast future or plausible links between a supplier and an institution. To operationalize this concept, we extract features from the public tender information in Paraguay and represent the supplier-institution relationship using a bipartite graph structure. We harness the capabilities of two contemporary models, GraphSAGE and GAT, to enhance the classification task. The experimental results demonstrate a high level of precision following the application of these methods. The project's code can be found in this github repo.

1 Introduction

Paraguay, situated in the heart of South America, boasts a population of approximately 6.1 million people [1]. As is often the case, the state serves as one of the largest employers within the country [2], consequently becoming a major player in the procurement of goods and services.

Whenever there is a need for a service or product from any government institution, a public tender is initiated. All companies that meet the specified requirements are eligible to participate in the selection process [3].

Some products or services may be specifically tailored for particular uses, making them exclusive to certain companies. On the other hand, there are products with general applications across various institutions, enabling a provider to cater to different state entities.

However, it is not guaranteed that a supplier offering products and services to one institution will be aware about tenders from another institution, even if the product or service is very similar.

The ability to anticipate connections between institutions and suppliers could be highly advantageous, particularly in scenarios where the participation of companies in a specific tender is limited.

By forecasting these connections, the government could formulate strategies to increase the number of companies engaging in a tender, thereby fostering competition. Competition plays a crucial role in driving down prices and enhancing the quality of goods and services [4].

Procurement is defined as *the process by which an organization buys the products or services it needs from other organizations* [5], that encompasses the tendering process. The procurement environment can be conceptualized as a bipartite graph or bigraph [6], wherein suppliers comprise one part, while institutions receiving the services or products constitute the other. The challenge of identifying potential connections between each part can be structured as a link prediction task.

In this project, we utilized public data provided by the Dirección Nacional de Contrataciones Públicas (DNCP), the agency overseeing the procurement process for all government institutions in Paraguay. We constructed a bipartite graph from this data and extracted relevant features. These features were then transformed into an embedding vector using a Graph Neural Network (GNN). Subsequently, these embedding vectors were employed for link prediction.

This approach allows us to predict the presence of links that may be physically existent but are missing in the dataset. The primary contributions of this project include:

1. Proposing meaningful features derived from the available tender data in Paraguay, with the potential for reproducibility in other countries.
2. Developing a model for link prediction on bipartite graphs within the supply chain context, leveraging the capabilities of Graph Neural Networks (GNNs).

2 Literature Review

2.1 Procurement data as a graph structure

Supply chains, along with their principal component, the procurement process, can be perceived as complex networks [6]. From a mathematical perspective, we could represent this network as a graph G(V, E), where V is a set of companies (nodes), and E denotes the procurement relations between them (edges) [7].

Building upon this concept, Kosaih et al. [7] utilized a real automotive dataset, encompassing over 18,750 firms and 89,175 procurement relationships, to create a graph structure for link prediction. It's worth noting that this graph was homogeneous, meaning that every node could have an edge to any other node in the graph.

In reality, numerous networks deviate from a homogeneous structure, featuring diverse types of nodes and edges. An example of such non-homogeneous networks is the bipartite graph [8].

A bipartite graph is a graph where the nodes are divided into two separate sets. All edges connect a node in one set to a node in the other set, and there are no edges between nodes within the same set [9]. This kind of network can be found in a variety of settings like, scientific collaboration network [10], actors and films [11], and artistic collaboration network [12].

Expanding this idea, Herrera et al. [13] use bipartite graphs to illustrate the tender process in Chile. In their representation, one set of nodes signifies the tenders, while the second set represents the various suppliers applying for any of these tenders. They later project the original graph into a monopartite graph representation to identify the main supplier's communities and compare it over time.

2.2 Graph Link Prediction

Link Prediction is a typically framed as a classification task in graph and network analysis designed to predict missing or potential connections between nodes in a network. When dealing with a partially observed network, the primary objective of link prediction is to deduce which links are most likely to be added or missing [14].

Kosasih et al. [7] applied link prediction to a homogeneous graph within a supply chain context. In this study, their goal was to uncover connections among different companies that might not be publicly disclosed due to confidentiality policies and a lack of willingness on the part of companies to share such information.

They utilized the topological role of nodes in the subgraph as features. Additionally, a GCN was employed to transform the node features matrix into an embedding vector, which was subsequently utilized in the classification task.

In bipartite graph, link predictions is best known in recommender systems [15]. Recommender systems suggest items to users, with these items encompassing products and services like restaurants, movies, music, books, etc. [16]

Zhang et al. [17] present a music recommendation model. In their approach, the recommendation data is represented as a bipartite graph, incorporating user and item nodes. The link weight is expressed as a complex number, portraying users' dual preferences in terms of both liking and disliking, thereby enhancing the similarity between users.

3 Tender and award data

3.1 Data Extraction and Wrangling

In recent years, Paraguay has enhanced access to public information from various sources. One prominent source is information on public tenders, which is readily available on the DNCP website [18].

This information is structured in compressed folders, each containing eight to ten CSV files, and is organized by the awards granted in specific years. The available data spans from 2013 to the present, and our study focuses on the period from January 2013 to September 2023.

Despite the absence of a glossary or instructions on file usage provided on the public website, we meticulously inspected each file to examine the column names and content. This examination enabled us to determine how to merge the tables in the CSV files and identify the relevant columns for our study. In Figure 1, we depict the tables used and illustrate the relationships between each table.

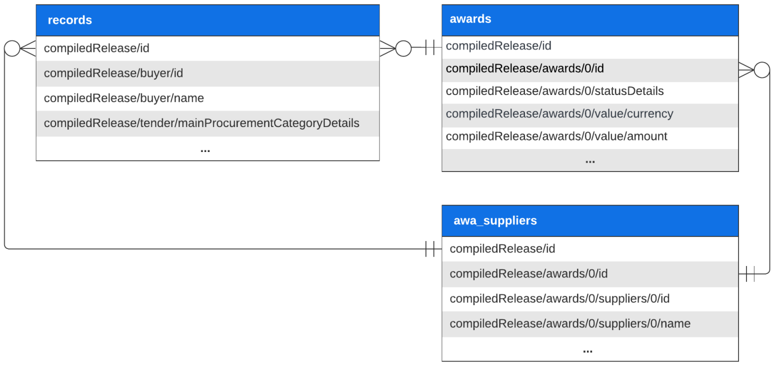
records
compiledRelease/id
compiledRelease/buyer/id
compiledRelease/buyer/name
compiledRelease/tender/mainProcurementCategoryDetails
...
awards
compiledRelease/id
compiledRelease/awards/0/id
compiledRelease/awards/0/statusDetails
compiledRelease/awards/0/value/currency
compiledRelease/awards/0/value/amount
...
awa_suppliers
compiledRelease/id
compiledRelease/awards/0/id
compiledRelease/awards/0/suppliers/0/id
compiledRelease/awards/0/suppliers/0/name
...

Figure 1: ER diagram of relevant DNCP files.

We excluded all columns not displayed in the diagram and combined the three tables. This consolidation resulted in the following information: the government institution initiating the tender, the category of the requested product/service, the amount and currency of the award, the status of the award, and the awarded supplier.

For the scope of this study, we focused on awards granted in the local currency and restricted the analysis to national companies. Additionally, we filtered out awards that were in progress or cancelled.

Furthermore, we refined the dataset by excluding suppliers that participated in only one tender. Ultimately, the dataset comprised 5,190 suppliers that had been involved in supplying to at least one of the 374 government institutions.

3.2 EDA and feature engineering

We explored deeper into the analysis of the data obtained from the processed dataset. Figure 2a illustrates that the majority of companies in the dataset received awards to provide products/services to 2-5 government institutions. During this analysis, we identified certain cases that raised suspicions and merit further investigation. Notably, there were instances where some companies obtained awards to supply to over a hundred different government institutions.

On the other hand, as shown in Figure 2b, it is noteworthy that the majority of institutions received products/services from 10-50 different suppliers within the studied timeframe.

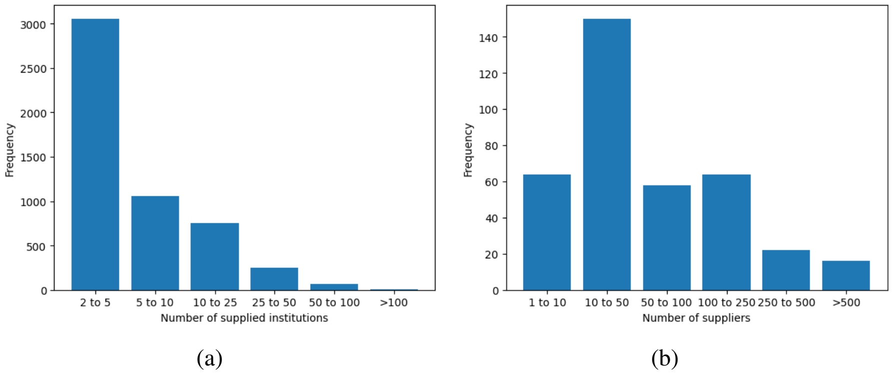

Figure 2: 2a) Distribution of the number of institutions supplied by a single company. 2b) Distribution of the number of suppliers each institution collaborates with.

We created a bipartite graph representation from this information, as shown in Fig. 3. The graph comprises two types of nodes: suppliers and government institutions. We establish a connection between a supplier node and an institution node if the supplier has won awards enabling them to supply products/services to the institution.

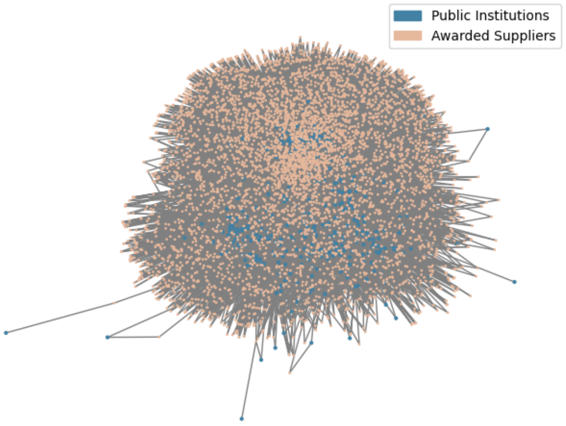

Figure 3: Bipartite graph representing the DNCP awards.

In Fig. 4, we illustrated the adjacency matrix of the graph. Each value on the y-axis corresponds to a single institution, while the suppliers are represented on the x-axis. As observed from the image, the matrix is highly sparse.

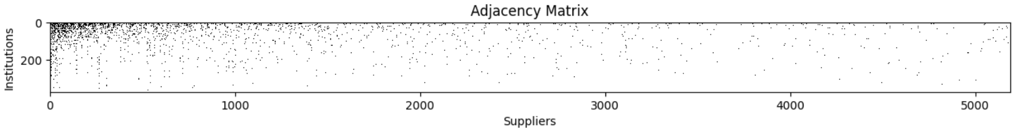

Figure 4: Adjacency matrix representation of the bipartite graph.

3.2.1 Institution node's features

With the goal of obtaining features to enhance the predictive capabilities of the model, certain characteristics were extracted from the available data regarding institutions.

It's reasonable to assume that specific institutions will require products or services that fall within certain categories. For instance, the Ministry of Health might need to procure particular oncology medications for the treatment of patients. In contrast, it wouldn't make much sense for the Senate Chamber to acquire products under the same category.

The DNCP database categorizes awards into 63 different categories based on the specific product or service supplied to the institution. To reduce the number of categories, some related categories were merged, resulting in a total of 25 categories.

Subsequently, all the categories are listed: 1) Acquisition and Lease of Real Estate; rental of furniture 2) Agricultural and Forestry Goods and Supplies 3) Training and Workshops 4) Fuels and Lubricants 5) Construction, Restoration, Reconstruction, Remodeling, and Repair of Real Estate 6) Consultancies, Advisory, and Research. Investment Studies and Projects 7) Cleaning Elements and Supplies 8) Military and Security Equipment. Security and Surveillance Services 9) Computing, Office, Educational, Printing, Communication, and Signaling Equipment, Accessories, and Software 10) Medical and Laboratory Equipment, Products, and Instruments. Health Care Services 11) Machinery, Equipment, and Major Tools - Transportation Equipment 12) Electrical, Metallic and Non-Metallic Materials and Supplies, Plastics, Rubbers. Spare Parts, Tools, Cameras, and Tires 13) Minerals 14) Furniture and Furnishings 15) Tickets and Transportation 16) Food Products 17) Chemical Products 18) Advertising 19) Insurance 20) Ceremonial, Gastronomic, and Funeral Services 21) Cleaning, Maintenance, and Minor and Major Repairs of Facilities, Machinery, and Vehicles 22) Technical Services 23) Textiles, Clothing, and Footwear 24) Kitchen and Dining Utensils, Porcelain, Glass, and Earthenware Products 25) Office Supplies, Paper and Cardboard Products, and Printed Materials

3.2.2 Supplier node's features

Similar thought can be applied to the suppliers, as it is also reasonable to think that a supplier is specialized on producing or offering certain kind of products/services. Therefore, the categories of their product/service were utilized as features, employing the same categories as for the institution node's features.

During the categorization process, additional suspicious cases were identified, which could be further analyzed in future investigations. Notably, there are suppliers that provide services and products in 18 different categories, ranging from military and security equipment to kitchen and dining utensils, porcelain and glasses.

To capture other features that might be relevant in the prediction process, we also extracted a feature indicating the type of institutions a supplier typically provides to. Expanding on this idea, we could hypothesize that certain companies primarily seek tenders related to institutions with specific characteristics, such as institutions of a certain size, location, or relationship to the company. We assumed that this implicit categorization or preference was immersed in the historical data.

In order to keep this institution's classification process as objective as possible, the categorization method of Contraloría Nacional de la República, Paraguay's National Audit Office, was employed [19]. After classifying all the institutions in the dataset, 9 categories were left; 1) University 2) Municipality 3) Ministry 4) Decentralized Administration 5) Retirement Fund 6) Independent Institution 7) Judicial Branch 8) Legislative Branch 9) Department Government.

Both the institution node's features and the supplier node's features were one-hot encoded to generate the feature matrix. In the end, we were left with 25 feature columns for the institution nodes and 34 feature columns for the supplier nodes.

4 Modeling

4.1 Graph Neural Network

Graph Neural Networks (GNNs) are a class of neural network architectures. Unlike traditional neural networks that process grid-like data (such as images or sequences), GNNs are tailored to handle data organized in graphs, which consist of nodes connected by edges. The key idea behind GNNs is to learn representations of nodes in a graph that capture both their local and global structural information [20].

In a GNN, the network typically consists of multiple layers, and each layer is responsible for updating the representation of each node based on information from its local neighborhood. The process of updating node representations involves aggregating information from neighboring nodes and incorporating it into the current node's representation [21].

In this project, two different architectures of GNNs are utilized to try the link prediction task in the generated dataset; GraphSAGE [22] and GAT [23].

In both cases, the output embeddings from the models are employed in the link prediction task. This involves calculating the dot product of the embeddings of the two nodes, from which we intend to predict the link. The resulting predictions then undergo a sigmoid activation, mapping them to the [0, 1] range and making them be interpretable as probabilities.

4.1.1 GraphSAGE

This method is designed to capture both local and global information within a graph. In the context of GraphSAGE, each node's representation is updated by sampling a fixed-size neighborhood and aggregating information from these sampled neighbors. This process enables the model to generate embeddings that capture the structural characteristics of the graph efficiently [22]. The utilized model employs two layers of GraphSAGE convolutional operations with ReLU activation.

The model is optimized using the Adam optimizer with a learning rate of 0.001. Moreover, as the task is a binary classification task, the loss function employed for training is the binary cross-entropy with logits loss.

4.1.2 GAT

GAT introduces the concept of attention mechanisms to the realm of graph-based learning. Unlike traditional graph convolutional layers that aggregate information uniformly from neighboring nodes, GAT allows nodes to assign different levels of importance (attention scores) to their neighbors during aggregation. This attention mechanism enables the model to focus on relevant nodes while aggregating information, allowing for more expressive and adaptive node representations [23].

The utilized model uses two GATConv layers, with 4 attention heads. The model is trained using the Adam optimizer with a learning rate set to 0.001. Moreover, the training procedure employs the binary cross-entropy with logits loss function.

5 Results

The training and testing split process involved allocating 20% of the data as validation, 20% as test data, and the remaining for training. Both models were trained for 4 epochs.

For result evaluation, link prediction was performed on the test data multiple times after training the model 20 different times with different seeds. In each training iteration, the following metrics were calculated: ROC-AUC, accuracy, precision, and recall. The mean of these values is presented in Table 1. We can observe that GraphSAGE outperforms GAT in most of the metrics.

Table 1: Performance of GraphSAGE and GAT on dataset.

	GraphSAGE	GAT
ROC-AUC	**0.882**	0.798
Accuracy	**0.815**	0.761
Precision	**0.786**	0.693
Recall	**0.741**	**0.741**

To conduct a more in-depth comparison of precisions, the precision@K plot was employed using the model with the best performance among all the iterations for both types of models. This plot illustrates precision utilizing the k percent true linked pairs. To streamline the representation, the x-axis in the plots was restricted to the percentage of positive predicted links, as precision remains constant beyond that point.

In Fig. 5, it is observable that GAT maintains a precision of nearly 100% until k reaches 5%. However, as the percentage of k continues to increase, GraphSAGE exhibits more stable performance, ultimately surpassing GAT in precision.

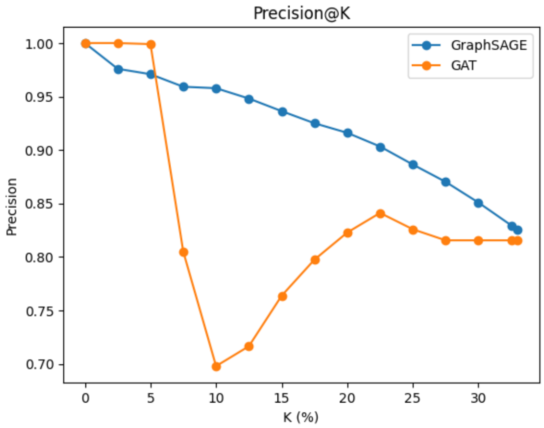

Figure 5: Adjacency matrix representation of the bipartite graph.

5.1 Ablation study

To assess the impact of the supplier feature indicating the type of institution each supplier provides to, we conducted training and testing of the GraphSAGE model on a graph that did not include those nine one-hot-encoded categories as node features for the suppliers. In Table 2, we observe that predicting with the model that included these features performs better in almost all metrics, especially in precision, where there is a gap of 0.036 between the two models. In the recall metric, the model without these features performs slightly better.

Table 2: Performance of GraphSAGE and GAT on dataset.

	GraphSAGE	
	w/ supplier feature	w/o supplier feature
ROC-AUC	**0.882**	0.866
Accuracy	**0.815**	0.8
Precision	**0.786**	0.75
Recall	0.741	**0.748**

While looking at the precision@k (Fig. 6) for the best model in all the iterations for each type, we observe that the model that includes all the supplier features demonstrates higher precision across all K values.

6 Conclusion and main challenges

In this project, we presented a method for representing suppliers and institutions involved in the tender process using bipartite graph representations. While Paraguay has made strides in improving access to public data, there are still significant steps that could be taken to facilitate understanding and utilization of the available information. Beyond the modeling aspect in this project, the understanding and wrangling process posed substantial challenges, requiring meticulous attention and cross-referencing data from multiple sources to validate the correct interpretation of tables.

Furthermore, we successfully extracted meaningful features from Paraguay's available public tender data. We showcased the capabilities of Graph Neural Network (GNN) architectures in generating embeddings that proved useful for predicting links in the graph.

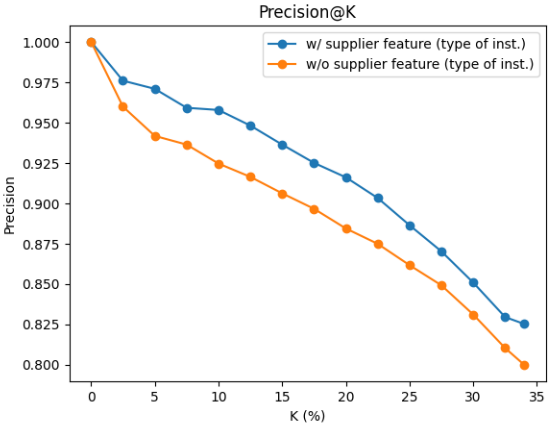

Figure 6: Adjacency matrix representation of the bipartite graph.

7 Future Work

For future work in this domain, exploration of additional node features sourced from other open platforms, such as news sources, could provide valuable insights. Additionally, a broader comparison among various models could be undertaken to assess the performance of different architectures.

Furthermore, a promising direction for future research involves exploring deeper into the suspicious cases highlighted throughout this report. Investigating ways to detect abnormalities in the current public tender process could have significant social impact, fostering increased transparency in the public sector's procedures. This direction of study holds the potential to contribute to a more robust and accountable public procurement system.

References

[1] Instituto Nacional de Estadística. (2023). Resultados Preliminares del Censo Nacional de Población y Viviendas 2022. Instituto Nacional de Estadística del Paraguay. Retrieved November 22, 2023, from https://www.ine.gov.py/censo2022/

[2] López, M. (2023). Gasto salarial en la era Mario Abdo sumó G. 86 billones y aumentó casi 30%. Market Data. Retrieved November 22, 2023, from https://marketdata.com.py/noticias/nacionales/gasto-salarial-en-la-era-mario-abdo-sumo-g-86-billones-y-aumento-casi-30-103681/

[3] Dirección Nacional de Contrataciones Públicas. (2023). Preguntas Frecuentes. Dirección Nacional de Contrataciones Públicas. Retrieved November 22, 2023, from https://www.contrataciones.gov.py/dncp/preguntas-frecuentes/

[4] Boushey, H., & Knudsen, H. (2021). The Importance of Competition for the American Economy. The White House. Retrieved November 22, 2023, from https://www.whitehouse.gov/cea/written-materials/2021/07/09/the-importance-of-competition-for-the-american-economy/

[5] Cambridge University Press and Assessment. (2023). Meaning of Procurement in English. Cambridge Dictionary. Retrieved November 22, 2023, from https://dictionary.cambridge.org/us/dictionary/english/procurement

[6] Choi, T. Y., Dooley, K. J., & Rungtusanatham, M. (2001). Supply networks and complex adaptive systems: control versus emergence. Journal of Operations Management, 19(3), 351-366. https://doi.org/10.1016/S0272-6963(00)00068-1

[7] Kosasih, E. E., & Brintrup, A. (2022). A machine learning approach for predicting hidden links in the supply chain with graph neural networks. International Journal of Production Research, 60(17), 5380–5393. https://doi.org/10.1080/00207543.2021.1956697

[8] Fortunato, S. (2010). Community detection in graphs. Physics Reports, 486(3–5), 75–174. https://doi.org/10.1016/j.physrep.2009.11.002

[9] Metcalf, L., & Casey, W. (2016). Graph theory. In L. Metcalf & W. Casey (Eds.), Cybersecurity and Applied Mathematics (pp. 67-94). Syngress. https://doi.org/10.1016/B978-0-12-804452-0.00005-1

[10] Newman, M. E. J. (2001). The structure of scientific collaboration networks. Proceedings of the National Academy of Sciences, 98(2), 404–409.

[11] Moody, J., & White, D. (2003). White, D.: Structural Cohesion and Embeddedness: A Hierarchical Concept of Social Groups. American Sociological Review, 68, 103–127. https://doi.org/10.2307/3088904

[12] Gleiser, P. M., & Danon, L. (2003). Community structure in jazz. Advances in Complex Systems, 6(4), 565–573. https://doi.org/10.1142/S0219525903001067

[13] Herrera, F., Torres, R., Nicolis, O., & Salas, R. (2020). Characterization of the Chilean Public Procurement Ecosystem Using Social Network Analysis. IEEE Access, PP(1), 1–1. https://doi.org/10.1109/ACCESS.2020.3011947.

[14] Papers with Code. (2023). Link Prediction. Papers with Code. Retrieved November 22, 2023, from https://paperswithcode.com/task/link-prediction

[15] Lakshmi, T. J., & Bhavani, S. D. (2021). Link Prediction Approach to Recommender Systems. arXiv preprint arXiv:2102.09185.

[16] Pham, K. (2022). What are Recommendation Systems?. Medium. Retrieved November 22, 2023, from https://medium.com/@khang.pham.exxact/what-are-recommendation-systems-6bb5036042db

[17] Zhang, L., Zhao, M., & Zhao, D. (2020). Bipartite graph link prediction method with homogeneous nodes similarity for music recommendation. Multimedia Tools and Applications, 1–19.

[18] Dirección Nacional de Contrataciones Públicas. (2023). Datos Abiertos. Dirección Nacional de Contrataciones Públicas. Retrieved November 22, 2023, from https://www.contrataciones.gov.py/datos/adjudicaciones

[19] Contraloría General de la República. (2023). Categorías de Archivos. Contraloría General de la República. Retrieved November 22, 2023, from https://www.contraloria.gov.py/index.php/categorias-de-archivos

[20] Zhou, J., Cui, G., Hu, S., Zhang, Z., Yang, C., Liu, Z., ... Sun, M. (2021). Graph Neural Networks: A Review of Methods and Applications. arXiv preprint arXiv:1812.08434.

[21] Sheikh, S. (2023). Exploring SageConv: A Powerful Graph Neural Network Architecture. Medium. Retrieved November 22, 2023, from https://medium.com/@sheikh.sahil12299/exploring-sageconv-a-powerful-graph-neural-network-architecture-44b7974b1fe0

[22] Hamilton, W. L., Ying, R., & Leskovec, J. (2018). Inductive Representation Learning on Large Graphs. arXiv preprint arXiv:1706.02216.

[23] Velickovic, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., & Bengio, Y. (2018). Graph Attention Networks. arXiv preprint arXiv:1710.10903.