Visual Social Network Analysis Based on Deep-Walk Graph-Embeddings and Self-Organizing Maps

Thiago Ciodaro[1], Vitor do Carmo[1], Fernando Ferreira[1], Felipe Grael[1], Debora Salles[1], Marie Santini[1]

[1] Internet and Social Network Laboratory – NetLab,
Federal University of Rio de Janeiro – RJ – Brazil

thiago.ciodaro@netlab.eco.ufrj.br, vitor.carmo@netlab.eco.ufrj.br

fernando.ferreira@netlab.eco.ufrj.br, felipe.grael@netlab.eco.ufrj.br

debora.salles@netlab.eco.ufrj.br, marie.santini@eco.ufrj.br

Abstract. *The task of identifying communication patterns on social networks poses a significantly complex challenge. These networks are inherently complex and are characterized by sparsely connected graphs. This study introduces an analytical model that combines the topological representation capabilities of graph-embedding techniques, such as DeepWalk, with the structure identification hability of neural networks based on self-organizing maps. The paper outlines the outcomes of testing the proposed analytical model with data from retweets on Twitter/X concerning topics related to vaccination in Brazil.*

Resumo. *A identificação de padrões de comunicação em redes sociais representa um desafio de grande complexidade. Essas redes, de natureza intrinsecamente complexa, são caracterizadas por grafos esparsamente conectados. Esse estudo propõe um modelo de análise que combina a capacidade de representação topológica de técnicas de graph-embeddings, como o DeepWalk, com o poder de identificação de estruturas dados de redes neurais baseadas em mapas auto-organizáveis. O artigo descreve os resultados da experimentação do modelo de análise proposto com dados de retweets do Twitter/X em temas relacionados à vacinação no Brasil.*

1. Introduction

The size and complexity of social networks, such as Twitter/X, have increased in recent years, along with their impact on public discourse and general elections in democracies [Salles et al. 2023]. Understanding how information propagates across these networks is crucial in any strategy aimed at mitigating the spread of misinformation. Various techniques are employed in the analysis of social networks using graphs, which include specific tools for visualization that emphasize the graph's key features.

To facilitate the inspection of complex networks, numerous graph visualization tools have been developed, including Gephi and Cytoscape [Faysal and Arifuzzaman 2018]. These tools are capable of processing relatively large networks, featuring hundreds of thousands of nodes and edges. However, the memory requirements for calculating and rendering the layouts of such extensive networks can significantly increase the computation time.

DeepWalk is a graph-embedding model that captures the social representations of a graph's vertices by simulating a series of short random walks [Perozzi et al. 2014]. This method uncovers latent features of the vertices that reflect neighborhood similarities and community affiliations, effectively encoding social relationships within a low dimensional continuous vector space. As a result, DeepWalk enhances interpretability, enabling the analysis and comprehension of complex social network structures in light of their intrinsic social dynamics. This improvement paves the way for simpler statistical modeling and examination of the depicted social connections.

Self-Organizing Maps (SOM) based on neural networks provide a method to both grasp the data structure of graph embeddings and enhance visualization via codebook vectors. Introduced by [Kohonen 1990], SOM is an unsupervised neural network utilizing competitive learning. It finds use in a variety of fields, including visualization, clustering, and classification, establishing itself as a multifaceted instrument for data analysis. Applying SOM to these graph embeddings does not just reduce the dimensionality of the data but also augments the analysis and interpretation of intricate graph data through straightforward visualization.

This study presents the results of the analysis of social networks from retweets in X (former Twitter) combining graph-embeddings and self-organizing maps. This article is described as follows: Section 2 reviews similar works in the analysis oif social network graphs, how the DeepWalk embeddings can be used to map complex graphs to a densed dimensional space and, finally, how SOM models are used to identify structural patterns in data. Section 3 describes the experimental setup, from the data collection to the application of the proposed analysis model. The experiment results are discussed in Section 4, while conclusions and future works are derived in Section 5.

2. Related Work

Graph embeddings offer an innovative means of representing complex graph structures as dense vector data and have been tailored for a broad range of applications [Chen et al. 2020]. Graph-based neural networks, a category of graph representation models, have shown promising outcomes in identifying patterns within large-scale datasets. However, their utility in fraud detection within data networks is still being explored and refined [Pereira and Murai 2021]. Additionally, graph embeddings have been employed to process Twitter data within the context of British politics, transforming the graph into a dense space, which is then utilized as input for an SVM classifier [Won and Fernandes 2022].

SOM models have been leveraged to create graphs for modeling complex networks in social network analysis [Rolemberg and Silva 2021], where each graph node is represented by a vector from the SOM's codebook. The findings indicated that the codebook was capable of mapping network characteristics, such as centralities. Nonetheless, representing larger networks would necessitate a more extensive codebook. A similar methodology for integrating SOMs with graph analysis was introduced in [Bonabeau and Hénaux 1998]. In both instances, a graph embedding was not utilized as an intermediary step between the graph network and the SOM model.

Our proposed model contributes to the discussion by experimenting topological mapping over graph-embeddings of large social networks. While the DeepWalk model fo-

cus on representing the social network onto a dense vector space, the SOM model encodes this dense space onto a 2D codebook. The codebook is a map of vectors fixed in the grid-space, but whose values in the embeddings space were adjusted according to its density. Each codebook vector, then, is responsible for mapping a set of graph nodes that are close in the embeddings space. Neighbouring vectors in the grid-space distant from each other represent regions of low density in the data. Additionally, the codebook vectors themselves can be used in clustering applications effectively [Vesanto and Alhoniemi 2000], reducing the input space to the number of vector in the codebook.

3. Experimental Setup

The analysis model was experimented using real, large and complex networks from retweets in X and its implementation is described in the following sections.

3.1. Data Collection

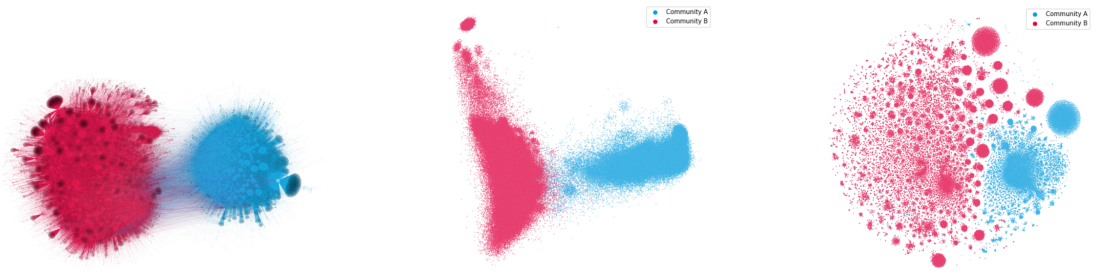
The Twitter/X dataset was gathered within a specific timeframe, spanning from January 01, 2022, to June 30, 2022. The process of data collection was executed through focused-queries and hashtags pertinent to the ”vaccine” topic in Brazil. Following the preprocessing phase to enhance data quality, a network was constructed based on retweets. In this setup, the original tweeter was designated as the source node, while the individual retweeting assumed the role of the target node. Each retweet instantiated a directed edge from the source to the target nodes. To quantify the strength of the connections, edges were assigned weights reflecting the frequency of retweets between users, with higher weights denoting stronger user connections.

3.2. Model Implementation

The experiment was implemented in Python, using open-source libraries such as sklearn [Pedregosa et al. 2011], igraph [Csardi and Nepusz 2006] and SOM [Vettigli 2018]. A C implementation of DeepWalk algorithm was used to generate the graph-embeddings. The mode training was developed and executed over an i5-intel CPU and 16 GB of RAM.

4. Results

The collected data comprised nearly 2.45 million tweets from approximately 416,000 accounts, resulting in about 1.57 million connections. Figure 1a presents a Gephi visualization of the vaccine-related retweet network, where colors represent the two main communities identified by the Leiden algorithm [Traag et al. 2019]. This visualization reveals the presence of two principal communities within the retweet graph, alongside some localized neighborhoods in both. Community A, depicted in blue on the right side of the graph, encompasses 105k Twitter accounts and 760k connections, whereas Community B, shown in red on the left side, includes 311k Twitter accounts and 803k connections (these numbers are approximations). Despite its larger size, Community B exhibits a lower average number of connections per account. Conversely, the smaller Community A demonstrates greater connectivity, averaging about 7.2 connections per account, in contrast to 2.6 for Community B.



(a) Gephi (b) PCA (c) t-SNE

Figure 1. Gephi (a), PCA (b) and t-SNE (c) visualizations of the retweet graph.

4.1. DeepWalking Embeddings

After constructing the network, we proceeded to generate node embeddings using the DeepWalk algorithm. This involved setting a walk length of 100 to define the steps taken during a random walk originating from each node in the network. To ensure a broad capture of neighborhood information, we conducted 250 random walks for each node (with a window size of 10). The decision to perform a higher number of random walks from each node, combined with the specified walk length, was aimed at capturing both the local and global structural features of the network. This approach was intended to provide a thorough understanding of the network's connectivity. Moreover, to represent each node within a low-dimensional space effectively, we set the dimensionality of the node embeddings to 128.

Figure 1 also illustrates the projection of the retweet-graph with popular visualization tools and methods [Anowar et al. 2021], such as PCA (Figure 1b) and t-SNE (Figure 1c). While PCA appears capable of representing the high-level characteristics of the two communities, the t-SNE projection uncovers more detailed structures within these communities. Although t-SNE provides interpretability, it is impossible to represent new data without retraining the entire database. SOM, on the other hand, can be applied to data unseen on training.

4.2. SOM Model

The subsequent step in our analysis involved utilizing the generated embeddings as input to train a SOM model. Considering the number of nodes in our network, we determined the dimensions of the SOM map to be 140x69. The SOM grid weights were initialized using PCA. To guide the training process, we specified a sigma value of 7.5 and a learning rate of 0.1. The neighborhood function was set to Gaussian, which allowed the smooth adaptation of neighboring nodes during training. For calculating the activation distance, we utilized the cosine distance metric and the topology was configured as hexagonal. Finally, the SOM was trained for a total of 200,000 iterations, ensuring convergence and stability in the learning process.

Figure 2 illustrates the distance matrix [Kohonen 1990] (or d-matrix) determined using the Euclidean distances between the codebook vectors. It represents the mean distance (calculated in the embeddings space) between the vector and its neighbors in the grid. Areas in the codebook grid that are distant from each other in the graph-embeddings space indicate stretched regions on the map. Conversely, densely populated areas in the

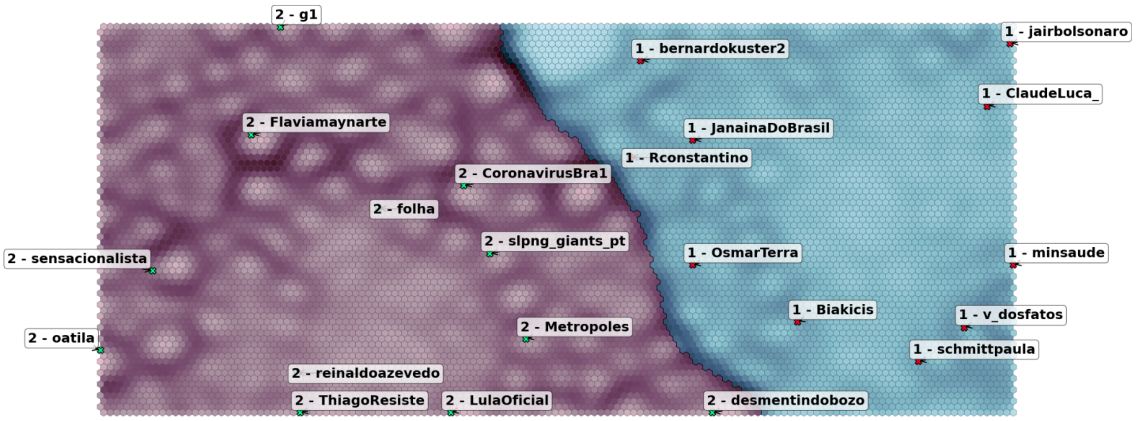
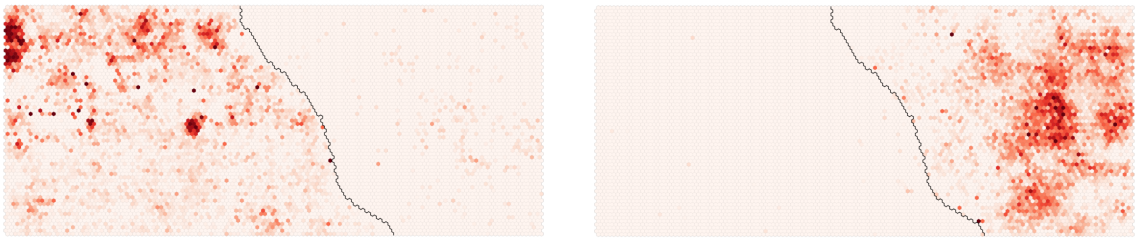


Figure 2. SOM codebook's distance matrix from the retweet graph-embeddings.

graph-embeddings correspond to neighboring vectors on the map with minimal distances between them. The intensity of the color in the grayscale reference on the map visualization increases with the distance, with darker colors denoting greater distances. A pronounced vertical stretch from the top to the bottom of the map reveals the presence of two primary retweet patterns in the context of the Brazilian vaccine discussion. These findings are in alignment with the community analysis conducted on the retweet graph. The map's color coding, red and blue, represents the codebook vectors associated with each of these two communities, determined by the users linked to each vector.

Furthermore, the map's left side is characterized by more granularity, with distinct clusters, indicating users who interact more frequently within local neighborhoods. This pattern is not as pronounced on the right side of the map, where the clusters are fewer and lack well-defined boundaries. This discrepancy can be attributed to the higher average number of connections per node in Community A, despite Community B having a larger total number of edges and nodes. The map additionally highlights some of the most active accounts in terms of the number of retweets. Notably, Community A is associated with vaccine discourses related to conservative and right-wing politicians, such as Federal Deputies Bia Kicis and Osmar Terra, as well as the Brazilian President in 2022, Jair Bolsonaro. On the other hand, an examination of user accounts in Community B reveals that this community aligns with accounts identified with progressive discourse and left-wing politicians, including the ex-president Lula (in 2022). Interestingly, the Brazilian Health Ministry's account is situatedwithin Community A, suggesting that the government institution's discourse was in sync with the information propagated within this community.

The SOM codebook also facilitates the visualization of which communities and local neighborhoods retweeted specific topics of interest in the social analysis. The tweets pertaining to the vaccine context underwent topic analysis [Grootendorst 2022] to extract keywords from specific subjects. Two prevalent topics were selected for emphasis on the map: 1) Child Vaccination, containing keywords from tweets discussing the vaccination of children against COVID-19; and 2) Sanitary Passport, containing keywords from tweets contrary to the sanitary passport, which would be demanded from people to use public services and closed spaces (a governamental measure to restrict the virus dissemination). The regex terms were, respectively, ['criancinhas; vao; vacinar thread; vacinacao; infantil — boomerang; crianca; sendo; vacinada'], which identified a total



(a) (a) (b) (b)

Figure 3. Codebook for topics (a) Child Vaccination and (b) No Sanitary Passport.

of 40,250 tweets and 36,478 accounts, and [# passaportesanitarionao], which identified 15,498 tweets and 8,960 accounts.

Figure 3a displays a heatmap showcasing the number of user accounts that retweeted at least one tweet pertaining to the Child Vaccination topic, corresponding to each codebook vector on the map. This heatmap reveals that tweets about this topic are predominantly found in localized clusters within Community B. The detailed pattern aligns with the blobs observed in the SOM's distance matrix, indicating that certain discourses are concentrated within specific clusters. Meanwhile, Community A exhibits some engagement with this topic, though it is confined to certain vectors and lacks a discernible pattern.

Figure 3b presents the heatmap for user accounts discussing the No Sanitary Passport topic. This topic is notably prevalent within Community A, with minimal representation in Community B. This analysis further reveals that the No Sanitary Passport topic garnered more interactions compared to the Child Vaccination topic, as evidenced by the higher average number of tweets per account associated with the former. This distinction underscores the variance in engagement and interest levels between the two topics across the different communities.

5. Conclusion and Future Work

This preliminary study combined DeepWalk graph embeddings with the SOM model for social network analysis. Tweets about Brazilian vaccination were collected to create a large retweet network. Results demonstrated that both DeepWalk and SOM models effectively identified the two main communities and other local neighborhoods in the retweet network. These patterns were visible in the SOM codebook vectors' d-matrix visualization. Utilizing the SOM codebook to visualize the codebook vectors map and the propagation of vaccine-related topics across the network provided valuable insights. This analysis helped identify user accounts actively engaging in socially relevant topics, showing their distribution across local clusters and their overall network reach.

Representing graph networks as dense vector embeddings allows the use of other machine learning models like clustering. However, updating the analysis for new data is challenging. Adding new nodes or connections alters the possible short-paths in the DeepWalk algorithm, potentially changing the entire graph embeddings. Nevertheless, the SOM codebook can efficiently cluster the graph embeddings, mapping local neighborhoods for deeper analysis.

References

Anowar, F., Sadaoui, S., and Selim, B. (2021). Conceptual and empirical comparison of dimensionality reduction algorithms (pca, kpca, lda, mds, svd, lle, isomap, le, ica, t-sne). *Computer Science Review*, 40:100378.

Bonabeau, E. and Hénaux, F. (1998). Self-organizing maps for drawing large graphs. *Information Processing Letters*, 67(4):177–184.

Chen, F., Wang, Y.-C., Wang, B., and Kuo, C.-C. J. (2020). Graph representation learning: a survey. *APSIPA Transactions on Signal and Information Processing*, 9:e15.

Csardi, G. and Nepusz, T. (2006). The igraph software. *Complex Syst*, 1695:1–9.

Faysal, M. A. M. and Arifuzzaman, S. (2018). A comparative analysis of large-scale network visualization tools. In *2018 IEEE International Conference on Big Data (Big Data)*, pages 4837–4843. IEEE.

Grootendorst, M. (2022). Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Kohonen, T. (1990). The self-organizing map. *Proceedings of the IEEE*, 78(9):1464–1480.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830.

Pereira, R. and Murai, F. (2021). Quão efetivas são redes neurais baseadas em grafos na detecção de fraude para dados em rede? In *Anais do X Brazilian Workshop on Social Network Analysis and Mining*, pages 205–210. SBC.

Perozzi, B., Al-Rfou, R., and Skiena, S. (2014). Deepwalk: online learning of social representations. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '14. ACM.

Rolemberg, T. and Silva, L. (2021). Aplicação de conceitos de redes complexas para a descoberta de formação de grupos em mapas auto-organizáveis. In *Anais do X Brazilian Workshop on Social Network Analysis and Mining*, pages 1–12, Porto Alegre, RS, Brasil. SBC.

Salles, D., de Medeiros, P. M., Santini, R. M., and Barros, C. E. (2023). The far-right smokescreen: Environmental conspiracy and culture wars on brazilian youtube. *Social Media + Society*, 9(3):20563051231196876.

Traag, V. A., Waltman, L., and Van Eck, N. J. (2019). From louvain to leiden: guaranteeing well-connected communities. *Scientific reports*, 9(1):5233.

Vesanto, J. and Alhoniemi, E. (2000). Clustering of the self-organizing map. *IEEE Transactions on Neural Networks*, 11(3):586–600.

Vettigli, G. (2018). Minisom: minimalistic and numpy-based implementation of the self-organizing map.

Won, M. and Fernandes, J. (2022). Analyzing twitter networks using graph embeddings: an application to the british case. *Journal of Computational Social Science*, 5.