Graph Analysis Applied to Word Networks

Fabiana Alejandra Rossi Magali Rodrigues Pires Eliana Inés Padula Álvaro López Malizia

ABSTRACT

The following practical work results from the comparison between graphs constructed from interactions between words. On one hand, there are the relationships between words that arise from an experiment of free associations, and on the other hand, the semantic distances calculated between the same words (the embedding used is word2vec).

For both sets of interactions, the corresponding graphs were plotted, and their characterization was performed. Comparisons were made based on the number of edges, centrality measures, network diameter, connections between elements, and community detection.

Finally, the construction of prototypical networks was carried out for the set of words derived from the experimentation.

The data processing and figure creation were done using the following code:colab.

I. Introduction

Networks are characterized as complex systems where diverse elements are connected or interact with each other. For each system, there is a network representation that defines the relationships between its components.

There are multiple examples of intricate systems spanning biology, linguistics, sociology, academia, and even contemporary computer and communication systems (Igual, L. Segui, S. (2017)). For instance, biological systems are inherently complex, involving interactions with the external environment, metabolic pathways, and processing of chemical and electrical signals. Additionally, emergent properties manifest in these processes as a result of these interactions.

Currently, human society functionally relies on increasingly complex networks that coherently cooperate with each other. From electrical systems, city signaling, global positioning to waste management, each of these networks is susceptible to failure, resulting in cascading failures where one or more seemingly independent systems are catastrophically affected (Albert-Laszlo Barabasí 2006).

The objective of studying networks can be as diverse as their composition. They can be studied to find relationships between their elements, determine efficient ways to traverse them, or identify their weak points to evaluate their robustness against attacks.

With the purpose of addressing the analysis of networks, their processing, and information generation, the following practical work was conducted, focusing on linguistics as an application field for these mathematical and computational tools.

Through association rules, connections between different words that we know and use routinely can be experimentally inferred (Elias Costa, M.). A simple model assumes that words can be represented as nodes in a network. The vertices and the distance between each of the nodes in this mental network of words are calculated through the process of evoking a ”response word” or ”R1” from a ”given word” or ”cue.” Elements in the network that are better interconnected through their relationship will naturally emerge more frequently and represent a shorter distance measure between them (Jones, M. N).

This work uses one of the datasets collected by the Small World of Words project 1. In this experiment, each participant is presented with a word and must provide up to three related words. This process constructs a network of free associations, where each word is a node, and the edges arise from connections made by the participants. Different ways of constructing and analyzing the network will be explored throughout the study.

Furthermore, it is possible to define a semantic distance between words based on what is known as embeddings. Embeddings allow representing each word in a vector space, within which a distance between words can be defined. Typically, cosine distance is used, and the embedding used in this work is word2vec (Mikolov, T. 2013). These distances can be utilized both to interpret the data directly (e.g., examining the cohesion of communities) and to construct a second network for comparison.

II. Data

The data belonging to the Small World of Words project was obtained. The ”SWOW-EN2018: Preprocessed” dataset, containing English words, was used. This dataset was previously preprocessed, where special characters and nonsensical words were removed.

On the other hand, Google’s word2vec was used. It consists of a vector representation of a word model, obtained from Google News dataset (with a total of 100 billion words). The resulting model consists of a vector with

300 dimensions and includes 3 million words and phrases (https://code.google.com/archive/p/word2vec/).

III. Methodology

A. Preprocessing

The data underwent preprocessing. Initially, the decision was made to work only with R1, i.e., the first word that a person responded to the stimulus (cue) during the experiment. Subsequently, words with a length less than two letters, ”stopwords,” and terms that did not appear simultaneously in cue and R1 were eliminated. In order to retain only the most frequent responses, all cue words whose frequency fell below the 98.5th percentile were first removed. This operation was repeated for cue-R1 pairs normalized by the cue frequency (cue-R1/cue) that fell below the same percentile. Finally, it was ensured that those words were within the body of word2vec.

For the construction of the semantic distance graph (word2vec), only the words resulting from the aforementioned preprocessing were retained.

B. Connected elements

Additionally, a filtering process was carried out in which a subgraph was generated, considering only the connected elements.

C. Visualization

The graphs were represented using the ”Spring” layout configuration.

IV. Task 1: Construction of the graphs

For the selected set of words from the Small World of Words experiment, a weighted and undirected graph (Gsww) was constructed. In this graph, nodes represent cue and R1 words, edges represent the presence of cue-R1 pairs, and weights represent the frequency of cue-R1/cue. The resulting graph (Gsww) has 219 nodes and 221 edges (Figure 1).

For the second set of words, the similarity matrix was calculated, and then the adjacency matrix was derived to build a weighted and undirected graph (Gw2v). The initial version of Gw2v had 219 nodes and 23,871 edges. A filter was applied to retain only links referring to distances between nodes greater than 0.92, reducing the number of edges to 1918 (Figure 2).

Brief descriptions of both graphs are provided below (Table 1).

Graphs	Gsww	Gw2v
Number of nodes	219	129
Number of edges	221	1918
Directed	No	No
Weighted	Yes	Yes
Loops	No	No

Table 1: *Characteristics of the graphs*

A. Gsww: Small World of Words

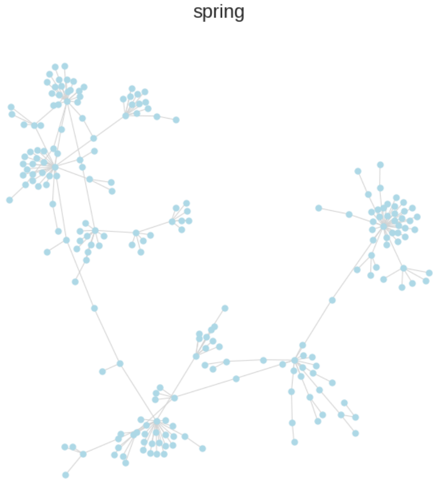

Figure 1: *Spring type graph representation for the dataset obtained from Small World of Words*

Se observa de manera sencilla la presencia de subcomunidades dentro de la red.

B. Gw2v: word2vec

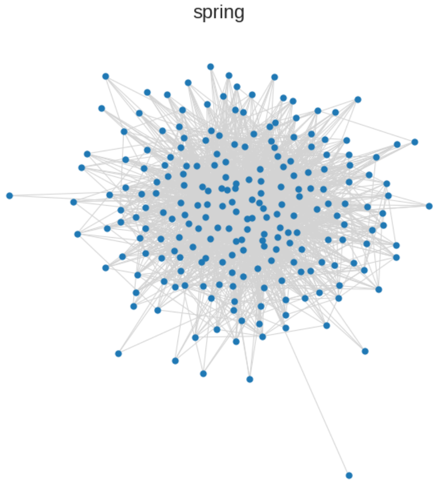

Figure 2: *Spring type graph representation for the dataset obtained from word2vec.*

The network does not exhibit a clear structure, and communities are not readily discernible from its visualization.

Task 2: Characterization of the graphs

In this section, a characterization of the constructed graphs was carried out based on their degree distribution, diameter, density, average shortest path, clustering measures, assortativity, and centrality measures.

Firstly, an analysis of the degree distribution of each of the studied graphs was conducted (Figure 3). Although both graphs are weighted and undirected, the result in Figure 3 shows a clear difference in their structures.

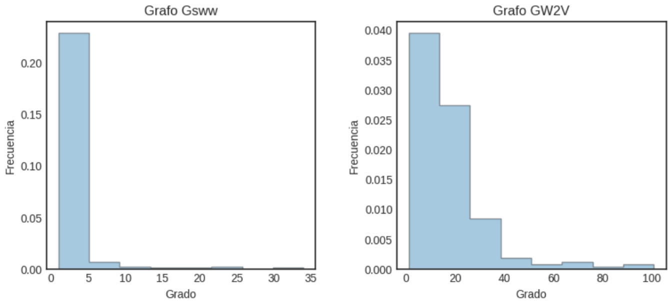

Figure 3: *Distribution of the degrees belonging to the nodes of the graphs Gsww (left) and GW2S (right).*

In the case of Gsww, most nodes have fewer than 10 links each, and a few hubs are observed with between 20 and 25 or between 30 and 35 links. This type of network, characterized by having a few highly connected nodes and the rest of the nodes with low or medium degrees, is called ”scale-free.” In comparison, the Gw2v graph is much more connected: most nodes have up to 40 links each, and a greater number of hubs are observed than in the previous graph. In this case, there are strongly connected nodes with up to 100 links each.

Next, various parameters were calculated to characterize the found networks. The results are shown in Table 2.

Measures	Gsww	Gw2v
Diameter	19	4
Density	0.0093	0.0803
Average Clustering	0.0051	0.0564
Asortativity	-0.4769	-0.3679
Average Minimum Path	8.26	2.093

Table 2: *Main measurements of the Gsww and Gw2v networks*

The results in Table 2 indicate that the Gw2v network has a smaller diameter, higher density, and average clustering than the Gsww network, suggesting that it has greater potential for connection between nodes. It is a more compact and interconnected graph, indicating that two words are more likely to be connected in the Gw2v semantic network than in the analyzed association network (Gsww).

Assortativity is a measure that allows analyzing whether nodes of the same class cluster only among themselves or with nodes of a different class. Both graphs have negative coefficients (Gsww: -0.47 / Gw2v: -0.36), indicating that they are disassortative networks, where high-degree vertices preferentially connect with low-degree nodes.

Next, various node centrality measures were studied to compare the Gsww and Gw2v graphs. These measures indicate the relevance of a node in a network, usually based on how it transmits information, i.e., how much it collaborates in the network’s flow and its proximity to the rest of the nodes (which refers to the graph’s cohesiveness). The centrality measures calculated for both graphs include degree centrality, where the most important node has the highest degree (the hubs); closeness centrality, a distance-based measure where nodes with higher values transmit information more quickly; betweenness centrality, where the more important the node, the more short paths pass through it; and eigenvector centrality, where the most relevant node has the highest eigenvalue, i.e., the node with the highest degree whose neighbors also have higher degrees. These results can be found in the provided URL, which contains the link to the Google Colab used to generate the networks and their characterizations. Figure 4 shows the result of the betweenness clustering calculation.

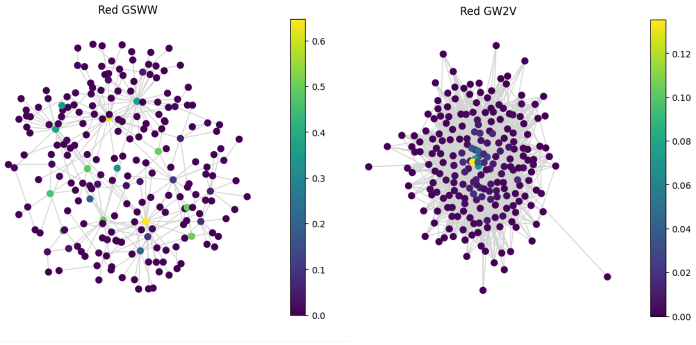

Figure 4: *Gsww (left) and Gw2v (right) network graph, where the nodes are colored according to their betweenness clustering coefficient value.*

It can be observed that the nodes most relevant according to degree centrality are generally the same nodes that gain greater importance according to betweenness centrality, indicating that the nodes through which more short paths pass are also the most connected. These nodes are likely to be polysemous words that facilitate connections between dissimilar concepts.

Regarding the closeness centrality measure (see Colab URL), it is observed that generally, most nodes in both networks have nonzero values. The opposite occurred in the case of eigenvector centrality: for the Gsww graph, only one node stood out for its relevance, while in the case of the Gw2v graph, nodes that obtained higher values when calculating degree and betweenness centrality measures were highlighted.

Task 3: Communities

In a subsequent step, we aimed to assess the presence of communities in the generated networks. For this purpose, the Louvain algorithm was employed. This algorithm uses modularity as a strategy to detect connected and dense communities.

The result of the application of this algorithm can be observed in Figures 5 and 6.

The communities found in the associative network (Gsww) were ten and are well-defined (Figure 5).

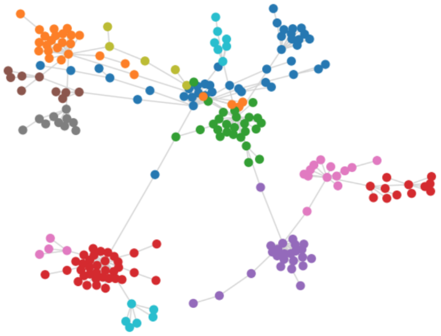

Figure 5: *Gsww network graph. The nodes are colored according to the identity of the different communities found by the Louvain method.*

On the other hand, the semantic network (Gw2v) was represented and colored according to the communities obtained with the Louvain method. Likewise, the color distribution in the same network was compared, but considering the communities found for the Gsww graph (Figure 6). It can be observed that the distribution of communities in the Gw2v network does not coincide with the communities found in the Gsww network. Additionally, the Louvain method finds a lower number of communities in the Gw2v network compared to the Gsww network.

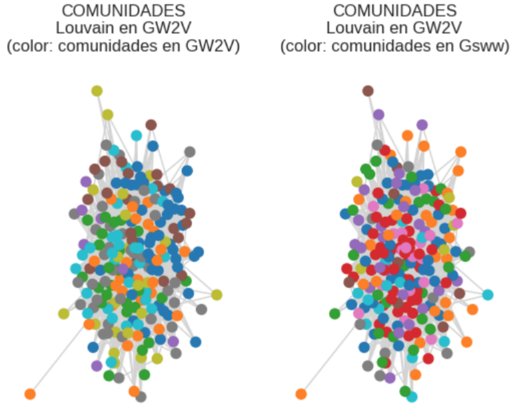

Figure 6: *Gw2v network graph. The colors represent the communities found using the Louvain method in the Gw2v (left) or Gsww (right) network.*

C. Optional 4: Grouping by MDS TSNE

Next we generate a visualization of the words of the Gw2v network using the MSD (Multidimensional scaling) and TSNE (t-Distributed Stochastic Neighbor Embedding) techniques (Figure 7).

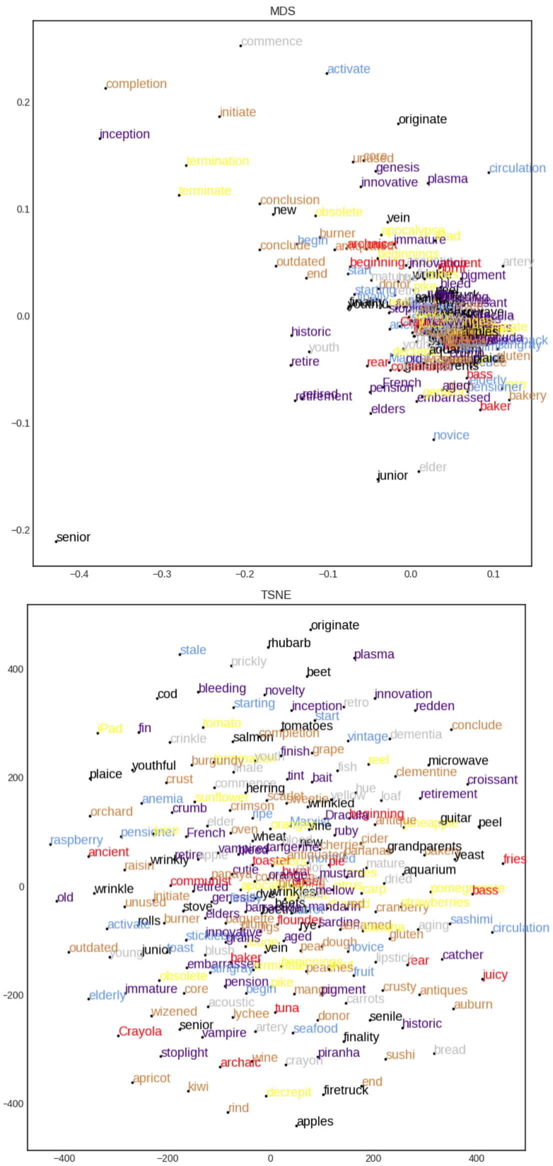

Figure 7: *Graph of words from the Gw2v network, grouped according to distances via MDS (top) or TSNE (bottom). Each color represents a community.*

Due to the large amount of data, there is an overlap of terms, making interpretation difficult. However, it can be observed that the communities (represented with different colors) show a logical structure within their identity. In the MDS representation, there is a community with elements such as ”retired,” ”retirement,” ”elders,” ”pension,” ”retire.” This grouping is coherent as it represents elements related to retirement or leaving the workforce. They are close in the representation of the first two components of MDS. The same occurs for the terms ”conclude,” ”begin,” which form a single community and are semantically related to the duration of activities.

The representation of data in the TSNE components (Figure 7) does not show a clear spatial grouping of terms.

Task 4: Small-world and prototypical networks.

Lastly, the clustering and average path length measures were compared against prototypical networks defined according to Barabasi Albert, Erdos Rényi, or Newman, Watts Strogatz. To do this, 1000 prototypical networks were generated with the same number of nodes and edges as the networks in this study. The aim was to assess whether the characteristics measured in Gsww and Gw2v could belong to these distributions. The results are indicated in Figures 8, 9, and 10.

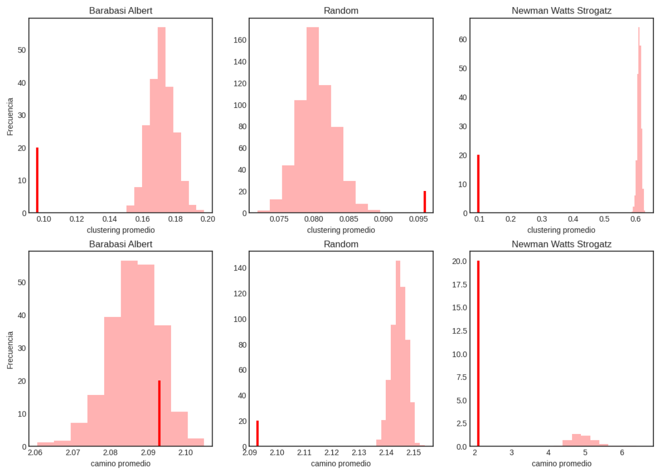

Figure 8: *Clustering distribution and average path of 1000 prototypical Barabasi-Albert, Random (Erdos-Rényi) or Newman-Watts-Strogatz networks. The solid red line represents the clustering and average path of the Gw2v network.*

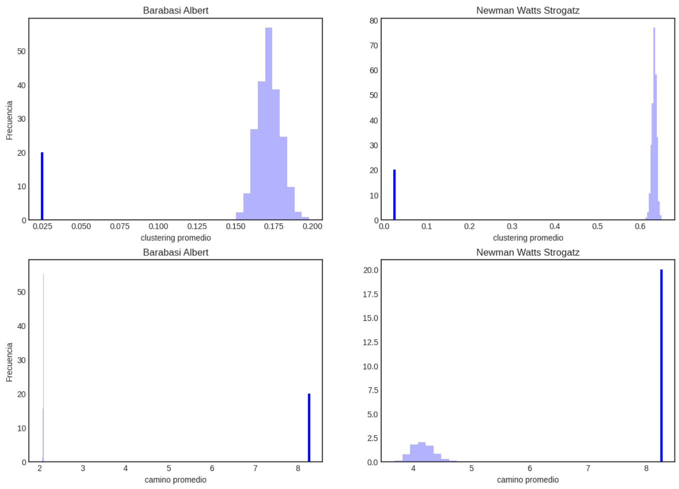

Figure 9: *Clustering distribution and average path of 1000 prototypical Barabasi-Albert, Random (Erdos-Rényi) or Newman-Watts-Strogatz networks. The solid blue line represents the clustering and average path of the Gsww network.*

For the Gw2v and Gsww networks, we can observe that only the average path length measure of the Gw2v network is comparable to one of the prototypical networks (Barabasi-Albert - scale-free network).

On the other hand, given that small-world networks are characterized by low average path lengths and high average clustering, we can conclude that if we compare Gw2v against the Random network (Erdos-Rényi), this concept would be fulfilled, but not when comparing it against the Barabasi-Albert scale-free network and the Newman-Strogatz network. As for the association network Gsww, it does not have a smaller average path length or higher clustering than the Barabasi-Albert and Newman-Strogatz networks.

These results are summarized in Figure 10, where the average path length and average clustering of the 1000 generated networks are represented for each prototypical network.

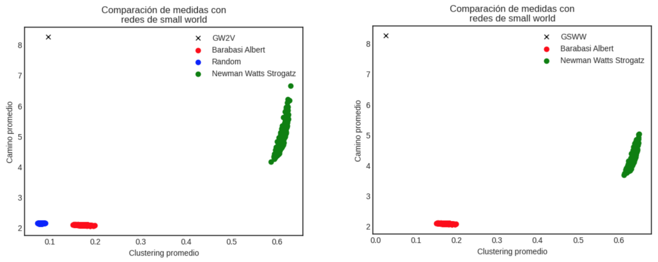

Figure 10: *Average clustering distribution and average path of 1000 prototypical Barabasi-Albert, Random (Erdos-Rényi) or Newman-Watts-Strogatz networks. The cross represents the Gw2v and Gsww networks in the left and right graphs, respectively.*

The comprehensive analysis of Figure 10 indicates that the measures of the Gsww and Gw2v networks are not compatible with the evaluated prototypical networks.

Discussion

During the development of this work, techniques were employed to become familiar with network management. Specifically, two networks of words were explored individually and comparatively. On one hand, a network of connections between terms related by stimulus words and responses belonging to the Small Word of Words project was generated. In contrast, the other network was generated from semantic distances between words provided by Google (using the word2vec embedding).

The visualization of both networks highlighted the presence of nodes whose connections support a large part of the elements that make up the Gsww graph. Differentially, the distribution of elements and connections in the Gw2v graph is more random and does not allow the detection of nodes of greater importance.

This observation could explain the differences between the calculated diameters for the networks. The diametrical length of Gsww is greater than that of Gw2v. In this scenario, it is plausible to think that there are multiple connection paths in the second network that shorten the distance between its most distant nodes. In the case of the Gw2v graph, there would be ”hub” nodes that create shortcuts, allowing for the reduction of distances between the centers. In contrast, for Gsww, as a result of these central nodes, there is a need to traverse these centers to move between the extremes of the entire network.

Similarly, it can be argued that the calculated density is qualitatively inversely proportional to the difference in diameter. While for the Gsww network, the density is higher due to the large number of branches, the Gw2v network is lower due to its more compact structure and greater number of connections.

Additionally, the calculated average shortest path for both networks reinforces the observations made for diameter and density. The minimum path presented by Gsww is higher than that of Gw2v. Again, the lack of connections between the elements that compose it could explain these differences.

Continuing with the characterization of the networks, the tendency to clustering was higher for the Gw2v network, while it was lower for Gsww. For the clustering coefficient, Gw2v has more connected neighbors. In comparison, in Gsww, nearby nodes have fewer connecting edges. The distribution of the latter resembles a hierarchical cluster where there is only a connection with a central node and with the terminal points.

The assortativity index, another network characterization measure, was negative for both graphs. This result indicates that there is no inherent tendency to grouping between nodes of the same class. The obtained measure indicates that even though in the Gsww graph there are nodes of higher degree, the connections between them would not be direct. Consistent with the aforementioned, it is likely that these are polysemous words that allow connecting dissimilar concepts.

Finally, according to the analysis conducted, we can affirm that the Gsww word network presents an intrinsic logical structure and suggests the presence of an associative relationship. On the contrary, the relationships between words obtained through the Google embedding present a more random and less organized structure.

The organization of the word network for Gsww generates a structure more susceptible to failures, where the loss of a connection can isolate large portions of the graph. Without multiple connections facilitating movement between elements, the inability to access a word significantly compromises the integrity of the network.

In a biological sense, the characterization of these cognitive networks could provide information and allow the use of strategies in the face of neurodegenerative diseases where word nodes and memories are progressively lost.

In conclusion, the study and characterization of networks, in general, allows for an understanding of the behavior and relationships of complex systems. In this practical work, based on connections between words, we were able to characterize the grouping structures and speculate on the underlying processes that generate them.

References

Albert-Laszlo Barabasí. Network Science, http://networksciencebook.com/

Igual, L. Segui, S. (2017). Introduction to data science. In Introduction to Data Science

Elias Costa, M., Bonomo, F., and Sigman, M. Scale-invariant transition probabilities in free word association trajectories. Frontiers in integrative neuroscience 3 (2009), 19.

Jones, M. N., Kintsch, W., and Mewhort, D. J. High-dimensional semantic space accounts of priming. Journal of memory and language 55, 4 (2006), 534–552.

Sigman, M., and Cecchi, G. A. Global organization of the wordnet lexicon. Proceedings of the National Academy of Sciences 99, 3 (2002), 1742–1747.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013).