UNIVERSITAT ROVIRA i VIRGILI

Complex Networks
Final project

Cañas Tarrasón, Eric
García García, Alba María
Kharitonova, Ksenia

Tarragona, Master in Artificial Intelligence

7th June 2020

Contents

1 Introduction **2**

2 Source data: *FreeAssoc* **2**

3 Network analysis proposal **3**

3.1 The *Random Inheritance Model* (RIM) . . . 3

3.1.1 The *Relevance Index* . . . 3

3.1.1.1 *Relevance Index* vs *Betweenness Centrality* . . . 4

3.1.1.2 Are there other candidates for *Relevance Index*? . . . 5

3.1.1.3 Other proposals for exploiting RIM's output space . . . 6

3.1.2 Using RIM for clustering . . . 6

3.2 Defining semantic relations through the WordNet model . . . 7

3.3 From relations to information . . . 7

4 Results **8**

4.1 General network information . . . 8

4.2 RIM results: analysis of *Relevance Index* . . . 9

4.3 *WordNet* results: analysis of lexical associations . . . 11

4.3.1 Synonym vs Antonym associations . . . 11

4.3.1.1 Is it possible to find communities of words responding to these associations? . . . 12

4.3.2 Similarity associations . . . 13

4.3.2.1 Analyzing the semantic similarity . . . 13

4.3.2.2 Analyzing the string similarity . . . 14

5 Conclusions **15**

1 Introduction

The **purpose** of this final project is to put in practice some of the knowledge acquired in the *Complex Networks* course on a topic and methodology up to the students involved amongst three broad proposals: *analysis*, *models* and *algorithms*. Our proposal was to **study the *Random Inheritance Model* in order to develop some new tool for network analysis** (which ended up being the *Relevance Index*) **and test it in the *FreeAsoc* dataset** –a weighted and directed graph that depicts human associations between English words– after enriching it using *WordNet* and other *Natural Language Processing* (NLP) techniques. The development of this project has been carried out in *Python 3* by means of the *PyCharm IDE* and making use of `NetworkX` and `nltk` libraries for graph and NLP functionalities, respectively.

The **structure of this document** is formed by four main sections: "Source data: *FreeAssoc*", "Network analysis proposal", "Results" and "Conclusions". The first details any relevant information about the source data, the second explains the complete approach proposed to analyze the target network focusing on our *Relevance Index*, the third provides the results obtained in the experimentation and our comments about them and the last one draws some conclusions about what we could infer about our research motivation in this project.

2 Source data: *FreeAssoc*

FreeAsoc [1] is a dataset from the *University of South Florida* that contains the **largest free association database in the United States up to date**. *Free association* is a common technique in psychology (which comes from Freud's psychoanalysis) that looks for the first concept that comes to the mind of the participant when asked about another concept. Thus, its goal is to know what are some of the participant's concept associations with the less possible impact of any kind of *rational* thinking; but when applied to a *massive* amount of participants –as the case at hand–, it can provide information about association patterns in a community.

In order to **collect the data**, the authors asked more than 6,000 participants to write the first word that came to their mind after reading each one of the English words that contained the booklet they were given. Each booklet had around 100-120 randomized words from a total of 5,019. Researchers ended up with three-quarters of a million responses, which were edited later on so that there were not spelling errors nor several words with the same *stem* –responses sharing their *stem* were grouped under the most predominant of them–.

The **results** of this research study were adapted to `Pajek` format in 2007 by V. Batagelj in the **following files**:

- `PairsP.net` and `PairsFSG.net`. Graphs containing 10,617 vertices and 72,168 weighted and directed edges so that each vertex is a word and each edge is a word association. The first graph sets the weights as the number of times some word A was associated to word B by the participants while the second provides this metric *normalized* (also known as *Forward or Cue-to-Target Strength*).
- `clue.clu` and `pofs.clu`. Partition of the previous graphs according to two different criteria: *cue word*[1] and *not cue word* and *part of speech* (using the *undefined*, *noun*, *verb*, *adjective*, *adverb*, *pronoun*, *preposition*, *interjection* and *conjunction* categories).

This project has used only `PairsP.net` and `pofs.clu` files.

[1] Connective expressions such as *now*, *meanwhile*, *anyway*, etc.

3 Network analysis proposal

We propose **two main techniques for extracting the latent information about vertex associations which could be present in a network**. The first approach can be applied to any network and consists on exploiting the properties of the output space obtained by the *Random Inheritance Model* (RIM) in order to extract indexes that state the relevance of each network's node or to apply any unsupervised learning algorithm. The second approach is more specific and aims to extract high level semantic information of human free associations relating them through *WordNet*, a language ontology widely used in the *Natural Language Processing* field.

3.1 The *Random Inheritance Model* (RIM)

Random Inheritance Model (RIM) is a model proposed by Javier Borge-Holthoefer and Alex Arenas [2]. It consists on the simulation of a **naïve navigation** in the network through **uncorrelated random walks** from node to node. This way, the algorithm is able to transform the network information to an equivalent **feature-based space** which has some interesting properties; such as the implicit representation of each node relevance, the possibility of applying any *unsupervised learning* algorithm over it or its equivalence with the *Word2Vec* [3] output space, which is widely studied in the *Natural Language Processing* field for extracting word embeddings from text.

The *Random Inheritance Model* has the following steps:

1. **Initialization of the output space as a diagonal matrix of size N.** First, as there is no information about the word associations at the beginning, each word is assigned to an orthogonal vector in the canonical space.

2. **Generation of each word vector navigating through the graph.** We start a random walk for each word w_i in the graph (represented by the vector v_i) . Then, being R the $S \times N$ matrix which contains the vector v_j of each node that has been visited during the S steps random walk, we apply:
$$v_i = v_i + \sum_{r \in R} r \tag{1}$$
All random walks are performed before modifying any vector. This way, the final vector v_i will count the number of times that this node has been visited during the random walk for any dimension, giving to the output interesting properties that will be commented later. Note that if we divide the output matrix by $S+1$, each vector v_i in it will represent the probability of reaching the word w_j from w_i in a random walk of S steps.

3. **Monte Carlo averaging.** As the process implies a highly stochastic method (random walk), the output obtained is only a high-variance estimation. In order to reduce the absolute error of the estimation, we apply a Monte Carlo simulation so that this error is reduced to $\frac{1}{\sqrt{N}}$, where N is the amount of simulations performed.

Finally, it is possible to reduce the dimensionality of the output by transforming the matrix from the canonical basis to a N-dimensional space. However, in order to maintain some interesting properties of the output space we will avoid this final step.

3.1.1 The *Relevance Index*

If we do not reduce the dimensionality of the output matrix, each value $v_{i,j}$ of the RIM output represents the amount of times a S-step random walker which started on w_i ended in the word w_j. Taking advantage of this property, we can calculate the amount of times each node was stepped in the whole RIM process as:

$$Stepped(i) = \sum_{j=1}^{N} w_{j,i} \tag{2}$$

And normalize this measure in order to obtain an index representing the relevance of each node in the network, the *Relevance Index*:

$$Relevance(i) = \frac{Stepped(i)}{\max_{j=1}^{N} |Stepped(j)|} \tag{3}$$

3.1.1.1 *Relevance Index* vs *Betweenness Centrality*

The *Relevance Index* is a measure that can have some similarities with *Betweenness Centrality*. However, they are totally different and it is important to denote these main differences:

- ***Relevance Index* is a probabilistic-based measure of paths.** However, the *Betweenness Centrality* of some node v computes the amount of shortest paths from s to t which include v:

  $$g(v) = \frac{\sum\limits_{s \neq v \neq t} \sigma_{st(v)}}{\sigma_{st}} \tag{4}$$

  where σ_{st} is the number of shortest paths from s to t and $\sigma_{st}(v)$ the amount of those paths that include v.

  This definition implies that any path which is not minimal will be not taken into account. On the contrary, *Relevance Index* includes all the possible paths, since it allows the possibility that the random walker arrives to a node coming from non-minimal paths. Figure 1 shows a simplified example of this issue.

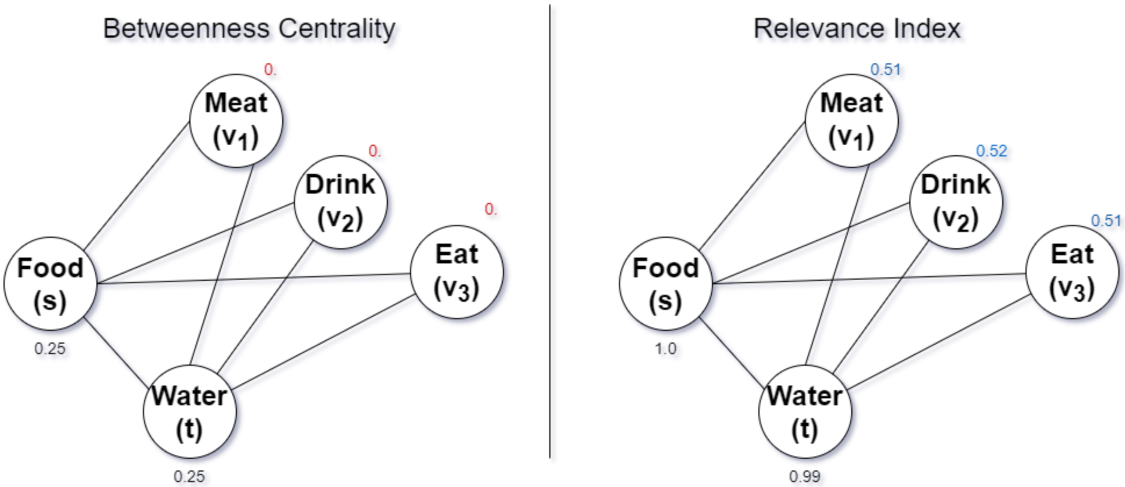

Figure 1: Comparison between *Betweenness Centrality* and our *Relevance Index*. The latter has been computed for ten steps ($S = 10$) and 100 repetitions ($N = 100$). V_i nodes show those nodes *Betweenness Centrality* does not provide any information about their relevance.

  As shown in Figure 1, *Relevance Index* finds that *Food* and *Water* are the most relevant nodes of the network; but also that *Meat*, *Eat* and *Drink* are important, since in half of the cases the transition will not be direct from s to t and will be passing through any v_i edge.

- ***Relevance Index* has the *number of steps* parameter.** While *Betweennness Centrality* does not implement a maximum number of steps in its definition, *Relevance Index* does and takes advantage of it. This way, we can model better the influence between communities by setting a low number of steps S, as Figure 2 exemplifies.

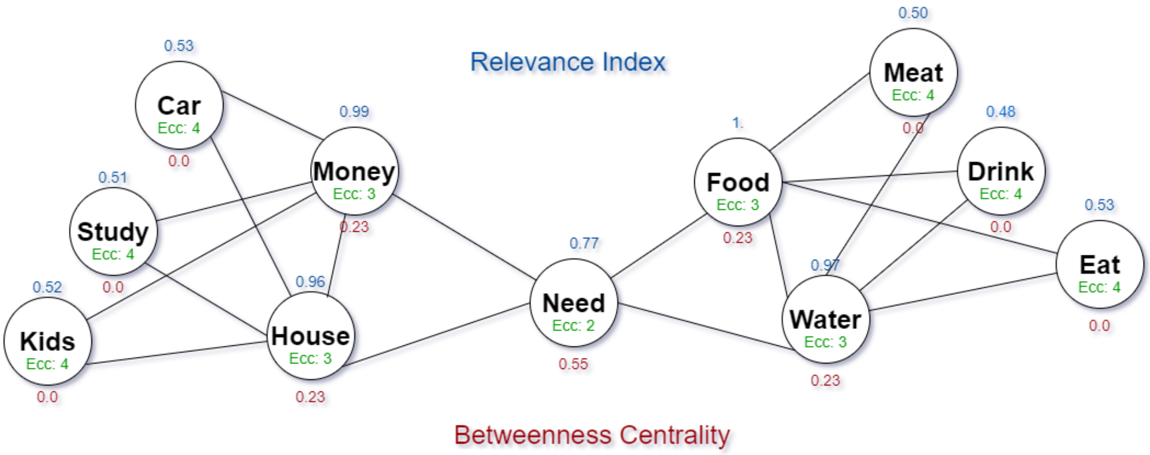

Figure 2: Comparison between *Betweenness Centrality* and our *Relevance Index* over a graph with two clear communities. The latter has been computed for two steps ($S = 2$) to avoid community contamination and 100 repetitions ($N = 100$). On blue, the results of Relevance Index, on red, the results of Betweenness Centrality. Additionally, each node includes in green its *Eccentricity*.

Figure 2 shows that we cannot adapt the influence range of the metric in *Betweenness Centrality* as we do in *Relevance Index* even after combining it with other measures like *Eccentricity* (distance to the farthest node).

- **Efficiency.** The complexity of *Betweenness Centrality* using the Floyd-Warshall algorithm [4] is $O(V^3)$. *Relevance Index*, calculated through RIM –a stochastic method implying only random walks– can be executed in an order of $O(VSN)$, where S is the amount of steps and N the amount of Monte Carlo simulations. Increasing this N minimizes the absolute error of our relevance estimation as much as desired ($\epsilon = \frac{1}{sqrt(N)}$).

As shown, both measures are totally different and provide different information. It can be tempting to infer the relevance of a node by its centrality or by searching nodes with low eccentricities. However, in many cases, it could not match the reality that is modelled by the network.

3.1.1.2 Are there other candidates for *Relevance Index*?

As shown in Figures 1 and 2, *Relevance Index* is extremely useful to estimate the relevance of a concrete node in the network. If we look to Figure 2, we can see that it is obvious that people thinks more usually in the words *Food*, *Water*, *Money* or *House* than in the *Need*, since they could be considered as the hubs of their communities. But we could think of other measures which could give us an idea of the relevance of a node:

- **The strength of a node or the degree of its neighbors.** This can be sometimes a good approach for searching hubs in networks. However, both measures are extremely local and are not able to model the complexity of large complex networks with high radius.

- **Other centrality measures like *Eigenvector Centrality*.** This measure scores the influence of a node in the network taking into account the local neighborhood of the node, but not characterizing as well as the *Relevance Index* an stochastic world where non-explicit relations can emerge easily through the properties of the neighbors.
- **Iterative algorithms like *PageRank*[5]** This algorithm introduces a stochastic and iterative procedure taking into account the non-direct neighborhood, and thus it is probably the closer method to the *Relevance Index*. However, it does not implement the flexibility that the parameter S gives to our proposal of managing the limits of the influence and the idea of random jumps can have sense on some domains like the Internet, but not in others.

We can summarize the advantages of our *Relevance Index* in contrast to the aforementioned alternatives as follows:

- **It allows non-shortest paths to influence the relevance proportionally to their probability,** which is a better way of modelling a stochastic reality (telecommunications, the Internet, human relations, word associations, etc).
- **It allows to parameterize its range of influence.** By setting S, it is possible to model scenarios where a node can be influenced by several communities and not only the one which is in (i.e. Does the person only speak to their relatives and close friends? Is it possible to discover new communities that also influence them?).
- **A feature-based space is obtained at the same cost.** The *Relevance Index* is computed from the output of the RIM algorithm and we can even take more advantage of this output as we will see in the following sections.
- **It has a linear complexity order and allows parallelism.** Its complexity is of O(VSN), being N a parameter which can manage the trade-off between the absolute error of the estimation and the computational cost.

3.1.1.3 Other proposals for exploiting RIM's output space

We have proposed how to generate a *Relevance Index* from the RIM output space. However, as it includes the explicit information about how frequent is to visit a node B starting a random walk from a node A (when we do not apply the final step that reduces the dimensionality), it is possible to extract more indexes from it.

For example, we could calculate an **uncertainty index**, computing the standard deviation of each column of the matrix from the RIM output space. This way, we would obtain a measure of the variability of the probability of arriving to each node from a random position. As higher this index is, higher is the uncertainty about the probability of stepping by a concrete node so that it is a node responding to strange patterns or only stepped by its community members. The lower it is, the more probable is of being a node which is almost always or never stepped by.

3.1.2 Using RIM for clustering

In addition to the *Relevance Index*, one of the most useful RIM properties is that, as in word embedding techniques, it transforms the network to a feature-based space where the vectors have an implicit relation with the similarity between nodes. This is extremely useful because over these feature-based continuous spaces we can to apply most of the unsupervised learning techniques for clustering (like *K-means*) or for visualizing the data (*PCA* [6], *TSNE*, *SOM*, etc). For instance, we could calculate with these clusters the behaviour of the defined indexes inside an outside the cluster it belongs.

3.2 Defining semantic relations through the WordNet model

If we move from the general domain techniques to the concrete domain of the target network, we can take advantage of all the knowledge from the *Natural Language* field. This way, it will be possible to research if human free association rules have any special relation with the semantic and structural relations defined by the language. For this reason, we propose to enrich the human associations of the network with the lexical language associations defined by *WordNet* [7]. *WordNet* is a lexical ontology of the language which has been widely used in *Natural Language Processing*. This ontology allows to find relations between words according to their similarity, how specific they are (for example, *dog* is less specific than *Siberian Husky*), if are they synonyms, antonyms, etc. In this case, we have enriched all the edges in the network with the following information:

- ***WordNet* information:**
  - **Path similarity.** There are several ways to calculate similarities between words in the *WordNet* ontology [8]. Path similarity is the most general one and provides a normalization of the minimal path which connects the nodes of the two words within the ontology.
  - **Are they synonyms?** 1 if the words are synonyms, 0 if they are not.
  - **Are they antonyms?** 1 if the words are antonyms, 0 if they are not.
- **Other *Natural Language* features:**
  - **Edit distance.** Edit distance [9] is an algorithm used to measure the similarity between two strings. In *Natural Language Processing* it is used for checking spelling mistakes, but in our case, it will be used as an heuristic for searching if there are free associations responding to words which are similar in form.

And all the nodes with the following parameters:

- ***WordNet* information:**
  - ***WordNet* Depth.** *WordNet* has a graph structure similar to a tree (but allowing multiple parents). The depth can be defined in this model as the lower amount of levels that must be descended until arriving to a concrete word (*Living Being* → *Animal* → *Dog* → *Siberian Husky* would have depth 4).
- **Other *Natural Language* Features:**
  - ***Part of Speech.*** We can extract each word's *Part of Speech* tag (verb, adjective, noun, etc), which could give us more information about how words are usually related.

3.3 From relations to information

There are several ways to transform the proposed analysis approaches into useful information:

- **Histograms of distribution.** In the same way that degree histograms are extracted for weighted graphs, it is possible to use any node measure to generate the same histograms (Histograms of average path similarity of responses, histograms of average antonyms on the responses, etc). These graphs can give information about how is the distribution of any of these lexical associations in the network.
- **Measuring relations inside and outside communities.** We can compose communities by usual community detection algorithms or by clustering the results of RIM. Once these communities (in case of existing) are defined, we can compare the average of any measure for the edges within the community and for the edges which connect the community to the outside.

- **Subgraphs maximizing some measures.** Finally, another possibility is to generate subgraphs with the nodes which maximize some measures. For example, we can build subgraphs containing only the most relevant nodes (to know which words the people more think about), or with the words that incite the most to say synonyms or antonyms.

4 Results

This section presents the **most relevant results obtained from the experimentation** performed over the *FreeAsoc* network according to the analysis proposed in Section 3 and collected as defined in Section 3.3. We can find this section structured in three main topics: "General network information", "RIM results: analysis of the *Relevance Index*" and "*WordNet* results: analysis of lexical associations"; which focus on the two main approaches developed in our analysis plus a general overview of the properties of the graph.

4.1 General network information

The ***FreeAssoc* graph** contains 10,617 nodes and 72,168 weighted and directed edges. The degree distribution of the graph approximately follows a power law with estimated $\gamma = 1.432$, therefore, it can be considered a **scale-free network**. This characteristic is consistent with the origin of the network due to the fact that many socially and anthropologically originated networks exhibit a similar behaviour, especially the semantic networks that represent relationships between concepts. This network can also be considered as such since the semantic relationships between concepts are based on the frequency of associations between them in the human mind.

The degree distribution has only into account the amount of connections and not the weights, as the strength distribution does. As we can see in Figure 3, there is a high amount of nodes with very low degree while a few of them have extremely high degrees. This fact reflects that while the majority of words provoke general associations in the popular culture, there are some words which are not managed by any general free association rule.

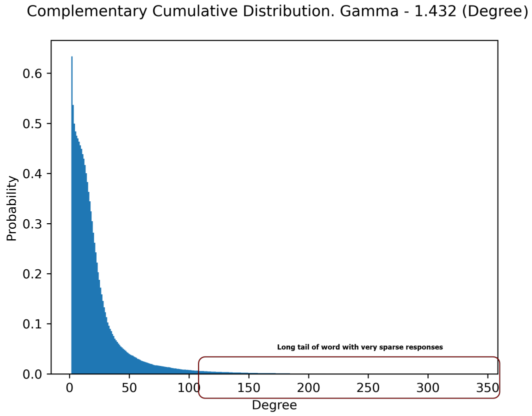

Figure 3: Degree complementary cumulative probability distribution for the *FreeAssoc* graph.

The **strength distribution** of this graph also follows closely a power law with an almost identical value of $\gamma = 1.351$. The long tail of this power law alerts us of the presence of hubs –hyper-connected vertices with extremely high degrees–, which would represent the common associations that are foremost on the people's mind. Further analysis allows us to discover these hubs with the RIM model's *Relevance Index*.

Indeed, the diameter of the network is 7 and the radius is 4, which is consistent with the scale-free behaviour of the graph. In such networks, the distance between the different nodes is usually very small due to the the presence of hyper-connected hubs that serve as shortcuts.

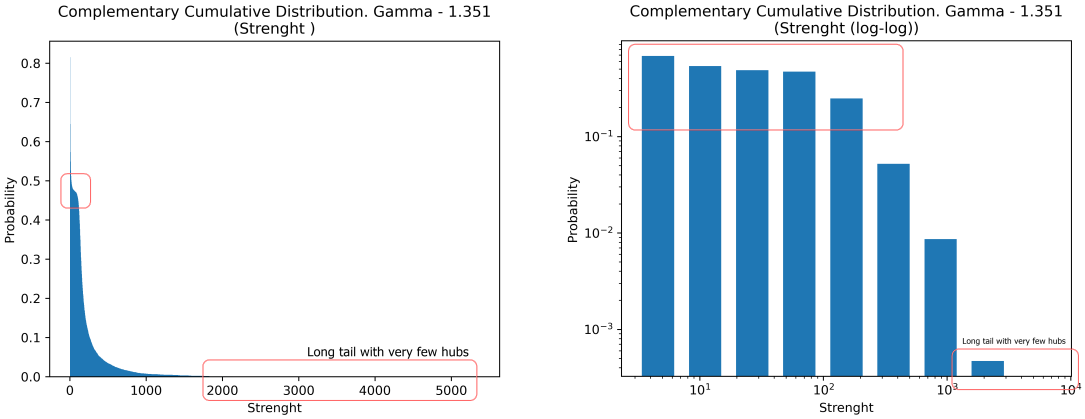

Figure 4: Strength complementary cumulative probability distribution for the *FreeAssoc* graph. It is important to denote the area between nodes with strength 10 and 100 with a lack of responses, indicating that people do not use to have the same rare associations (They could be considered as noisy associations) and also the large tail of the distribution, indicating that some associations are quite installed in the popular culture or have a latent strong rule behind them.

The strength states the total frequency of responses for the nodes, and it can be seen clearly that the network consists of nodes with either very rare associations (no more than 10) or with very frequent (more than 100) indicating that there is no middle ground between these thresholds. **Concepts can be either rare**, where few respondents use them, **or extremely common**; which is consistent with the behaviour of the natural language, and it suggests the parallel with the Zipf's law where the frequency of the words in the language is inversely related with their rank [12]. The same holds approximately true for the associations.

If we compare the degree distribution and the strength distribution of the network, despite the fact that they have different scales, the form of distributions is very similar, which is to be expected.

4.2 RIM results: analysis of *Relevance Index*

The concept of relevance based on the RIM model introduced in the Section 3.1 also allows us to make some interesting conclusions about the network.

First of all, the *Relevance Index* allows us to extract the **nodes that have the utmost importance for the network** following the Figure 5. They have among all of them 76 neighbors, which means that they are associated with 76 concepts more. The weights of the connections with their neighbors can exceed 120, a very high frequency of responses for the network. The high indirect connectivity of these nodes indicates that they serve as the key concepts on the respondents' mind even though they do not have the direct connections between themselves. For many concepts, whatever the people say gets them back to the same basic things: Food, Money, Car, Water.

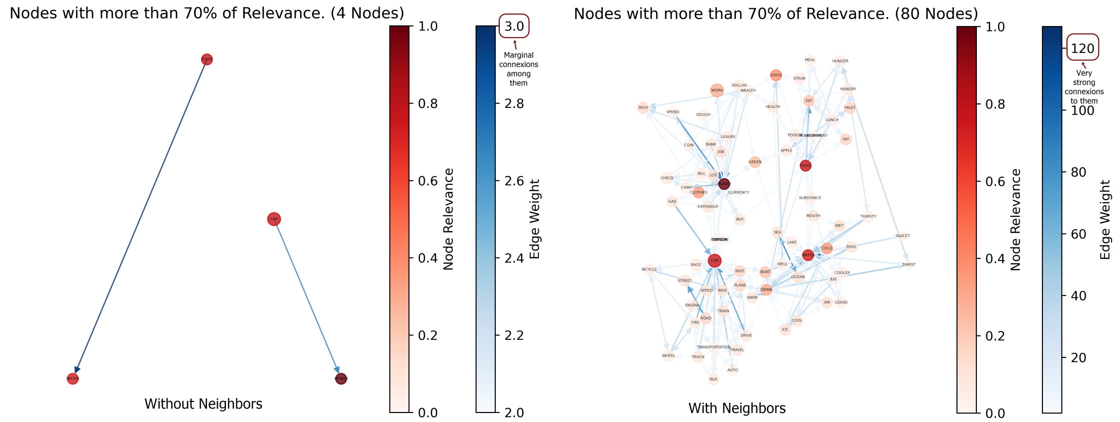
Nodes with more than 70% of Relevance. (4 Nodes)
Without Neighbors
Node Relevance
0.0
0.2
0.4
0.6
0.8
1.0
Edge Weight
2.0
2.2
2.4
2.6
2.8
3.0
Marginal connexions among them
Nodes with more than 70% of Relevance. (80 Nodes)
With Neighbors
Node Relevance
0.0
0.2
0.4
0.6
0.8
1.0
Edge Weight
20
40
60
80
100
120
Very strong connexions to them

Figure 5: Nodes (words) with a relevance greater than 70% ($S = 10$) in the network. The diameter of the node codifies its eccentricity (as higher the diameter, lower the eccentricity). Using $S > NetworkDiameter$ allows the algorithm to extract a general *Relevance Index* giving to all nodes the opportunity to be visited and to influence the rest. The differences between left and right subgraphs show how both pairs of nodes, in spite of having a poor direct connectivity (max $|EdgeWeight| = 3$), have strong connectivity among them through its indirect connections. Each one is the hub of its own community, and there are associations which build cycles communicating these communities. For example, $Food \rightarrow Health \rightarrow Wealth \rightarrow Money \rightarrow Buy \rightarrow Car \rightarrow Boat \rightarrow Water \rightarrow Food$.

Comparing the hubs obtained with the ***Relevance Index*** to the key concepts extracted by other measures of centrality and importance such as Betweenness Centrality, Eigenvector Centrality and even PageRank [5], we can observe how our method is more centered in detecting the general recurrence of the associations, since it is based in the Monte Carlo simulation of the random walks in the network. This property allows the method to **discover better those latent concepts which are hidden below large sequences of associations**. That is, neither the most central nor the most pointed, but the most recurrent.

Ordered Relevance				
Criteria/Order	**1st**	**2nd**	**3rd**	**4th**
Betweenness Cent.	Money	Work	Bad	Animal
Eigenvector Cent.	Push	Pull	Shove	Out
PageRank	Cold	Money	Water	Good
Relevance Index (Ours)	Money	Car	Food	Water

Table 1: Most relevant nodes found by each algorithm. In blue, those words which are related to human basic needs. In red, those words which are related to the means that can satisfy those basic needs. In the *Relevance Index*, nodes from 5th onwards have an important descent of relevance from 74% (4th) to 42% (5th). In the rest of algorithms, the differences are lower and it is less clear where to put a threshold.

For this reason, the concepts that the *Relevance Index* extracts have a high relation both with the human basic needs and the means of satisfying these basic needs (even, the concept of the Car can be considered as such if we take into account that the majority of respondents are Americans), whereas the concepts extracted by other measures are more abstract. This confirms the **robustness of our method**, that can be related to the idea of the free association itself. For example: imagine a person who starts from a certain word and makes free associations with other words until they stop. If many people do it a sufficient amount of times (like in our Monte Carlo simulation), **the**

most visited words will be the ones that are most important for the respondents in general. Therefore, our method can be considered the most suitable for those kind of a network which respond to this nature.

4.3 *WordNet* results: analysis of lexical associations

The analysis of the network enriched with the *WordNet* information has certain limitations: **53.39% of the nodes**, that is slightly more than the half, **do not exist in the *WordNet* ontology** and therefore it does not have information about them (non-computable nodes). Nevertheless, the half of the network that does can provide some important insights about the psycho-linguistic characteristics of the underlying information.

4.3.1 Synonym vs Antonym associations

In order to get insights about how human free association rules are managed, it is important to relate them with the existent lexical information. Maybe the **strongest way to relate two words in lexical terms** is to categorize their relation by **equality** or **opposition**.

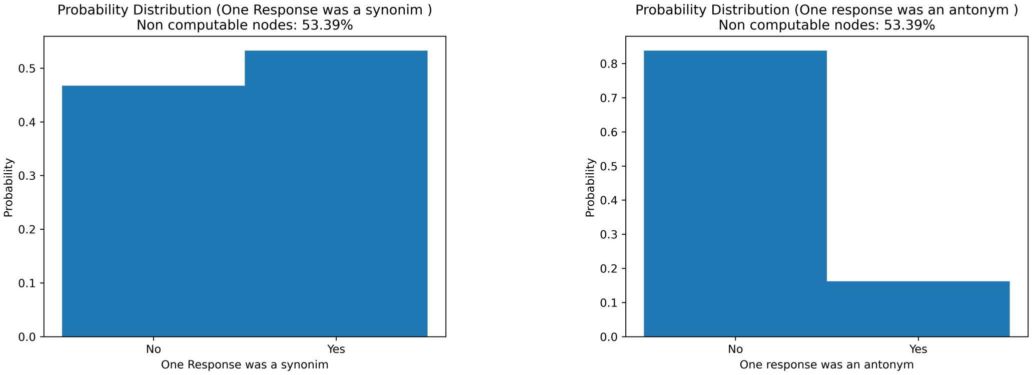

Figure 6: Amount of (computable) nodes which produced that one or more subjects answered with a synonym (left) vs nodes which produced one or more antonym responses (right). This comparison show how the human *Free Association* rules are managed more by similarities than oppositions.

Figure 6 shows how more than the 45% of computable nodes in the network are connected to at least one synonym, while only the 15% of them are connected to an antonym. This fact leads us to suggest that the **similarity and familiarity is more important for the human mind than the contrariness**. Moreover, if we analyze the complete distribution of synonyms and antonyms (taking into account the weight of the connections) these results are clearly confirmed.

If we look at Figure 7, this hypothesis is clearly confirmed: much more nodes produced the synonym responses than the antonym responses. However, it is important to denote that, while in Figure 6 we can see that there was 30% more nodes with a synonym response than an antonym one, in Figure 7 the difference between full synonym responses and full antonym responses decreases a 21%. This effect, that is exemplified in Figure 8, lead us to suggest that **for those concepts for which exist a clear antonym association** (*East* → *West*, *Right* → *Left*), **this association is clearly more rooted in our mind than synonym associations**, which tends to be more flexible and sparse.

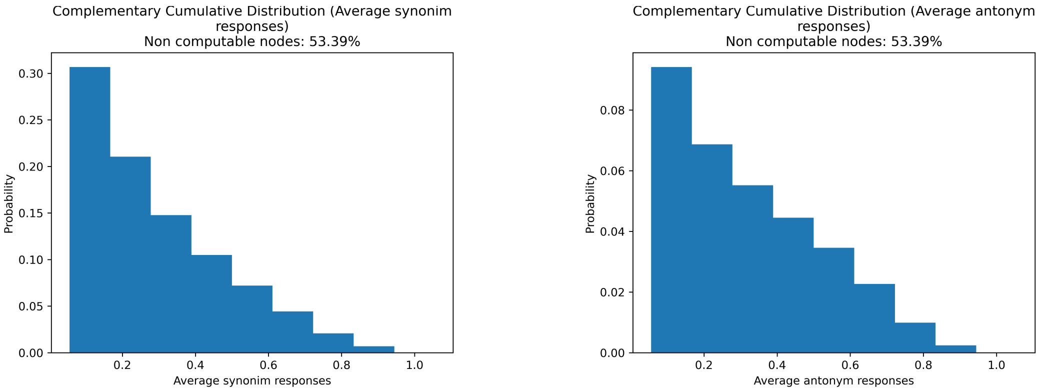

Figure 7: Complementary cumulative distribution explaining the average synonym / antonym responses of the subjects. In spite of the similarity of both plots in shape, it is important to take attention on the Y axis range. The 30% of nodes within the network produced more than a 90% of synonym responses, while less than the 10% produced the same amount of antonym responses. Figure 8 exemplifies the concrete cases which maximize these kind of connections.

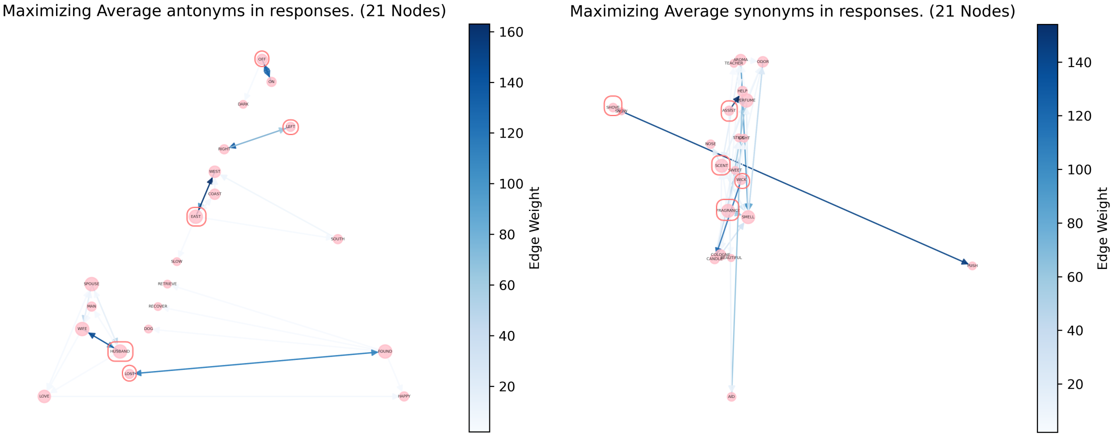

Figure 8: Subgraphs showing the neighbors of those five words which maximized synonyms and antonyms responses. The words which maximize that response have been circled and their diameter represents its eccentricity (as higher the diameter, lower the eccentricity). It is important to take attention to how sparse the response is. Antonyms responses were quite precise (*On* → *Off*, *East* → *West*), synonym responses were more fuzzy (*Scent* → {*Odor*, *Perfume*, *Fragance*...}).

4.3.1.1 Is it possible to find communities of words responding to these associations?

One of the properties of the RIM model, which was proposed as interesting for working with, was the potential that its output has for being used as input for any usual clustering method of unsupervised learning. This application could be considered as a method for extracting **communities** in a different way, in which they were more likely to be **separated according to the behaviour detectable by RIM**.

Figure 9 presents a comparison among this clustering method and other usual method used for community detection, as Asynchronous Fluyd Modularity [10].

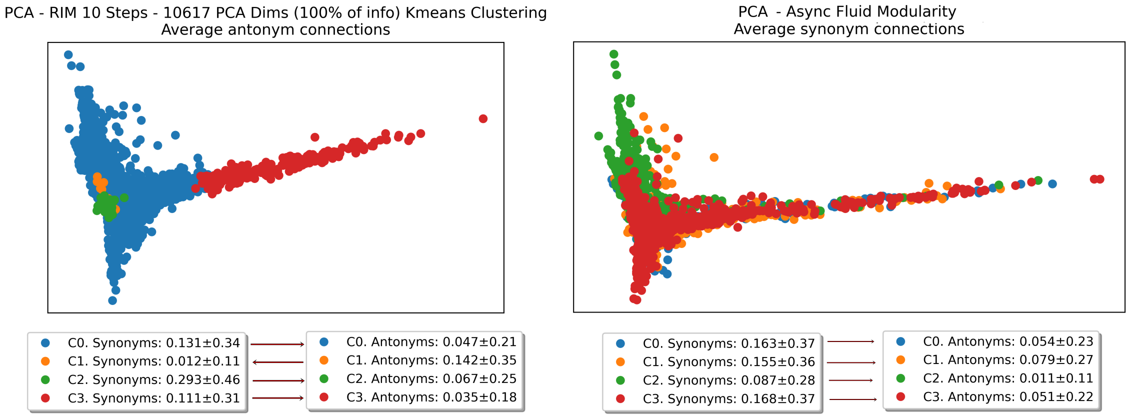

Figure 9: Comparison between frequency of antonyms and synonyms inside each cluster of a K=4 K-means clustering of the RIM results with 10 steps (left) and within each community detected by Asynchronous Fluyd Modularity [10] (right). These plots show specially how it is not possible to find a low number of huge communities in this graph, since Free Association rules are extremely fuzzy and present a lot of sparse connections. For this reason, algorithms like Greedy Modularity [11] are able to find up to 147 communities within it.

In this figure, we can find some differences. First of all, while in Asynchronous Fluyd Modularity all the different communities detected follow (in terms of the mean) the distribution between antonyms and antonyms expected from a random sample; in the case of RIM clustering, cluster *C1* does not follow this expectation. Taking this into account, we could cautiously assume that we are more able to clusterize these differences using the RIM clustering approach. However, the standard deviation of the resulting communities or clusters in both cases is so high that these conclusions cannot be taken as relevant. This fact is due that the used **network is too sparse for being classified using a low number of communities**, this is the reason why some algorithms that are able to determine the number of communities like Greedy Modularity [11] are able to find up to 147 communities within the graph.

4.3.2 Similarity associations

4.3.2.1 Analyzing the semantic similarity

If we relax the binary terms of equality opposition, another important *WordNet* information that is possible to associate with the nodes in the *FreeAssoc* network emerges: the semantic **similarity between the nodes**. Responding to this purpose, the path similarity method returns a score denoting how similar two word senses are, based on the shortest path that connects both senses in the is-a (hypernym/hyponym) *WordNet* taxonomy. This method gives a score is in the range 0 to 1, in which similarities higher than 0.125 define a clear proximity between both concepts, and similarities higher than 0.5 means that one concept is the direct hyponym or hyperonym of the other. If we look at Figure 10, we can observe how the prevailing similarities are usually in the higher range of the scale. This **confirms** the conclusions of the previous section, **that the respondents tend to associate similar things**.

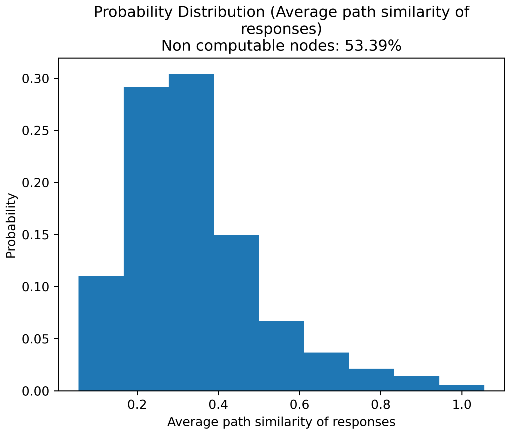

Figure 10: Path similarity probability distribution shows a high level of lexical similarity associations, with a high concentration of (high) path similarities in the range (0.2, 0.4). Putting in context *WordNet* path similarity ranges: $Path_{sim}(Car, Bus) = 0.125$, $Path_{sim}(Car, Motorbike) = 0.25$, $Path_{sim}(Car, MotorVehicle) = 0.5$, $Path_{sim}(Car, Automobile) = 1$.

If we look more in detail the distribution, we can see how practically 90% of the computable responses had a similarity higher than 0.2 (which could be, for example, the similarity between *Car* and *Motorbike*). This is a relevant conclusion which shows that there is a **very strong linkage between the semantic associations defined by a language and the free associations learned by its speakers**. Rarely, the responses given by the surveyed were from a different lexical domain. If we think about it, this rule is expected to be the one which defines the best these *Free Association* rules, since this is the natural way on how we understand concepts: the whole thing by the composition of its parts. For example: If we ask someone to think about a person, what concepts do we expect to appear in their mind? *Head, body, arms, etc* or the concepts of *building, clouds, physics, etc*? It is clear that concepts composing the parts of an element are the first to come to our minds when asking about that element.

4.3.2.2 Analyzing the string similarity

In computational linguistics and computer science, **Edit Distance** is a way of quantifying how dissimilar two strings (e.g., words) are to one another by counting the minimum number of operations required to transform one string into the other. If we analyze how this property is distributed among the edges of the network, we can find a normal bell-shaped distribution indicating a **random character of the relationship between how phonetically similar the associations are** (the preference is in the range of not too similar and not to dissimilar).

This distribution makes clear that human ***Free Association* rules work in a high level rather than a lower**. Making the obvious evident, that human mind works not in the world of words but in the word of the concepts which represent these words.

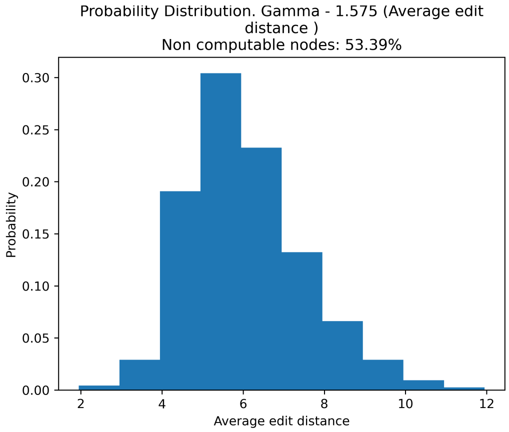

Figure 11: Edit Distance probability distribution shows a normal distribution with the mean in $\sigma = 5$. To put in context, an Edit Distance of 5 could be, for example, the distance between *Mobile* and *Phone*. This distribution shows that people tends to prefer, generally, lexical high level association than low level phonetic associations.

5 Conclusions

In this final project, we have proposed **two main techniques for network analysis**. The most specific has been to perform a lexical analysis of the network, trying to relate the samples of *Free Associaton* rules in the network with a semantic ontology of the language like the one proposed by *WordNet*. The other technique, which is maybe the most interesting since it is applicable to any network independently of its scope, has been to exploit the properties of the RIM model output space.

Taking advantage of these properties, we have proposed a way for extracting the most relevant latent information of a network provided that this information can be represented as the **probability of a node to be stepped by in a random walk across the network**. We have named this measure as ***Relevance Index***. To evince how powerful this proposal is in a network like the analyzed here, we have compared it against other algorithms, like Betweenness Centrality, Eigenvector Centrality and PageRank (Table 1); proving that our method is able to extract not only those nodes which matter the most to human beings but also to establish a well-defined threshold about which ones had capital relevance and which ones were only important. However, it is important to remark that our method will only obtain its best results when the concept of random free walks which sustains it has sense in the network, penalizing its performance in networks which represent deterministic or static systems.

We have tested both techniques in the ***FreeAsoc* graph**, which contains weighted relations between words, in order to analyze how we humans relate concepts. Such **analysis** led to the following conclusions:

- *Free Association* rules work preferably in higher levels of the language (*lexical level*) rather than lower levels like phonetics, since we prefer to associate words by their meaning and not by their sound. We do not think this hypothesis could be influenced by the language.
- Synonyms are more common in *free association* rules that antonyms; but when the opposition is the relation that prevails, it tends to depict *complementary* terms (i.e., *mortal-immortal*, *black-white*) thus reducing in comparison the range of answers we observe in synonyms (i.e., *scent* ⟶ *odor*, *fragrance*, *aroma*).

- *Free association* rules have a strong connection to semantic rules. Despite that in some cases we observe a consequent association (*nose* → *scent*, *scent* → *flower*), most of the times this association responds to a close semantic relation (*car* → *motorbike*, *play* → *game*).
- We have evidenced that the most relevant concepts for the people are those that denote basic needs for survival (such as *food* and *water*) as well as the tools that can be used to satisfy them. Among the latter, the most relevant without any doubt is *money* (the second more relevant falls to an index of 76%), as it is the only concept able to satisfy any kind of need we could have in the society we live nowadays.

References

[1] Nelson, D.L., McEvoy, C.L., Schreiber, T.A. “The University of South Florida word association, rhyme, and word fragment norms” *Behavior Research Methods, Instruments, Computers* 36: 402–407 (2004)

[2] Borge-Holthoefer, J., Arenas, A. “Navigating word association norms to extract semantic information” COGSCI 2009.

[3] Mikolov, T., Chen, K., Corrado, G., Dean, J. “Efficient Estimation of Word Representations in Vector Space” (2013)

[4] Floyd, R W. “Algorithm 97: Shortest Path” *Communications of the ACM* 5 (6): 345 (1962)

[5] Page, L. Brin, S. “The PageRank Citation Ranking: Bringing Order to the Web” (1998)

[6] Pearson, K. “On lines and planes of closest fit to systems of points in space” *Philosophical Magazine* 2: 559-572 (1901)

[7] Miller, G A. Beckwith, R. Fellbaum, C,. Gross, D., Miller, K. “Introduction to WordNet: An On-line Lexical Database” (1993)

[8] Pedersen, T., Patwardhan, S., Michelizzi, J. “WordNet: Similarity - Measuring the Relatedness of Concepts” (2004)

[9] Navarro, G. “A Guided Tour to Approximate String Matching” *ACM Computing Surveys* 33 (1): 31–88 (2001)

[10] Parés F., García-Gasulla D. et al. “Fluid Communities: A Competitive and Highly Scalable Community Detection Algorithm”. *Complex Networks & Their Applications* VI: 229-240. Springer, Cham (2017)

[11] Clauset A., Newman M. E., Moore C. “Finding community structure in very large networks.” *Physical Review E* 70, Issue 6. American Physical Society (2004)

[12] Zipf, G. K. “The Psychobiology of Language” New York, NY, USA: Houghton-Mifflin (1935)