

## RESEARCH ARTICLE

# Leveraging word embeddings to enhance co-occurrence networks: A statistical analysis

Diego R. Amancio<sup>1\*</sup>, Jeaneth Machicao<sup>2</sup>, Laura V. C. Quispe<sup>1</sup>

**1** Institute of Mathematics and Computer Science – USP, Avenida Trabalhador São-carlense, nº 400, CEP 13566-590, São Carlos, SP, Brazil, **2** Escola Politécnica da Universidade de São Paulo (EPUSP), São Paulo, Brazil

\* [diego@icmc.usp.br](mailto:diego@icmc.usp.br)

## Abstract

Recent studies have explored the addition of virtual edges to word co-occurrence networks using word embeddings to enhance graph representations, particularly for short texts. While these enriched networks have demonstrated some success, the impact of incorporating semantic edges into traditional co-occurrence networks remains uncertain. In this study, we investigate two key statistical properties of text-based network models. First, we assess whether network metrics can effectively distinguish between meaningless and meaningful texts. Second, we analyze whether these metrics are more sensitive to syntactic or semantic aspects of the text. Our results show that incorporating virtual edges can have both positive and negative effects, depending on the specific network metric. For instance, the informativeness of the average shortest path and closeness centrality improves in short texts, while the clustering coefficient's informativeness decreases as more virtual edges are added. Additionally, we found that including stop-words affects the statistical properties of enriched networks. Our results, derived from enriching networks with FastText embeddings, offer a guideline for identifying the most appropriate network metrics for specific applications, based on typical text length and the nature of the task.

## OPEN ACCESS

**Citation:** Amancio DR, Machicao J, Quispe LVC (2025) Leveraging word embeddings to enhance co-occurrence networks: A statistical analysis. *PLoS One* 20(7): e0327421. <https://doi.org/10.1371/journal.pone.0327421>

**Editor:** Manuel Jesús Cobo-Martín, Universidad de Granada, SPAIN

**Received:** October 28, 2024

**Accepted:** June 13, 2025

**Published:** July 11, 2025

**Copyright:** © 2025 Amancio et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

**Data availability statement:** The data underlying the results presented in the study are available from Project Gutenberg (<https://www.gutenberg.org/>).

**Funding:** We acknowledge financial support from the São Paulo Research Foundation (FAPESP), grant no. 2020/06271-0. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

**Competing interests:** The authors have declared that no competing interests exist.

## Introduction

Word co-occurrence networks have been widely used in various text analysis studies, including authorship attribution [1], distinguishing real text from generated text [2], summarization [3], and language clustering [4,5]. In these networks, words from a text are represented as nodes, with edges connecting adjacent words [6]. However, this approach requires large textual corpora to provide sufficient graph-based data. Without enough data, the network topology becomes too simple, reducing its effectiveness in classification tasks.

The recent introduction of semantic edges (also known as virtual edges) enhances the discriminability of networks derived from shorter texts by avoiding linear structures [7]. Virtual edges are hypothetical connections between non-adjacent nodes, based on the notion that semantically similar words can be linked. This enrichment of word co-occurrence networks,

achieved by incorporating virtual edges, can be implemented using word embeddings, where similar words have similar embeddings. This approach has shown significant advances in text classification [7,8] and keyword extraction [9,10]. However, the implications of integrating virtual edges into word co-occurrence networks remain still unexplored.

We propose to analyze the effects of integrating virtual edges into word co-occurrence networks by examining two important properties of the metrics derived from these networks. The first property, referred to as informativeness, is the ability of the metrics to distinguish between real and meaningless texts. The second property aims to determine whether the metrics are more sensitive to syntactical/stylistic or semantic textual factors [11]. Although it is well-established that most metrics derived from traditional co-occurrence networks are informative and more dependent on syntax and style than on semantics, no study has investigated whether this property holds or varies with the number of edges added to the model. Motivated by this gap, this work is driven by the following research questions:

1. How do virtual edges impact the informativeness of the metrics and their ability to differentiate between meaningful and nonsensical text?
2. How do virtual edges affect network features in capturing syntactic or semantic aspects of texts?
3. Do the answers to the above questions depend on the pre-processing steps in network construction, such as stopword filtering?

Our analysis revealed that including virtual edges can indeed affect the statistical properties of networks, particularly in shorter texts. In these cases, the informativeness of metrics such as average shortest path length and closeness centrality is enhanced. Conversely, some metrics, like the clustering coefficient, experience a decrease in informativeness. Regarding the ability to capture syntactic features, our analysis showed that virtual edges in short texts increase the sensitivity of the average shortest path to semantics, while metrics like eigenvector centrality are largely unaffected and show no clear preference for syntactic or semantic features. In some cases, the sensitivity of metrics can shift from syntax to semantics with the addition of virtual edges, as observed with betweenness. We also found that filtering stopwords can affect the informativeness of metrics.

## Related works

Complex networks have been used to analyze texts in various contexts, including uncovering language patterns and performing text classification tasks [6,12–17,17–22]. The most commonly used network for text analysis is the word co-occurrence model. The work conducted in [19] applied co-occurrence networks to analyze six languages, revealing language-specific patterns. The study compared the co-occurrence networks of Chinese and English texts, including essays, novels, articles, and reports. They found that all the networks exhibited scale-free and small-world properties. In particular, in networks derived from Chinese texts, the average shortest path length could distinguish between different content types, making it a style-dependent measure. Additionally, Chinese texts exhibited higher clustering coefficients, while English texts generally had shorter average path lengths. The authors also found that Chinese networks are assortative, while English networks are disassortative.

In [17], the authors analyzed patterns in syntactic dependency networks, where connections are based on syntactic dependencies. They demonstrated that networks derived from different languages—Czech, German, and Romanian—share complex statistical properties, such as the small-world phenomenon, degree distribution scaling, and disassortative mixing.

Additionally, the authors noted that co-occurrence networks provide a simplified approach to constructing syntactic networks, as most syntactic links occur between adjacent words.

Co-occurrence words have also been used to study the statistical properties of unknown manuscripts. Using a co-occurrence strategy without any pre-processing step, the authors in [11] aimed to determine whether texts, such as the Voynich Manuscript, exhibit characteristics of natural language or are merely random character sequences. By analyzing word frequency, intermittency, and network properties, the authors provided valuable insights into the nature of the text, contributing to the understanding of unknown or undeciphered writings. Additionally, this research proposed a framework for evaluating the statistical properties of network models, including the ability of network metrics to distinguish between real and nonsensical texts. Furthermore, the study examined which linguistic properties – syntactic or semantic – the metrics are likely to capture. This framework forms the foundation for the analysis conducted in this study.

More recently, enriched word co-occurrence networks have been proposed to capture information that traditional word co-occurrence networks may miss [7,8,10,23]. [7] employed an enriched network to identify cognitive disorders from short text analysis by incorporating virtual edges. Including virtual edges was crucial not only for capturing additional linguistic information from the texts but also for creating a more complex topology in networks derived from short transcripts. Using a similar strategy, word embeddings and community detection were used for the problem of word sense induction, outperforming competing algorithms and baselines [23].

Enriched networks have also been applied to stylistic tasks, such as authorship recognition [8]. Upon employing different strategies to threshold the networks, the study evaluated various word embedding techniques, including Word2Vec, GloVe, and FastText [24–26]. The authors found that combining FastText with a global strategy yielded the best performance for short texts in the context of authorship recognition. In a similar approach, the study conducted in [10] analyzed the BERT model [27] for adding virtual edges in networks for keyword extraction, achieving better results compared to other word embedding models.

While most studies involving enriched networks apply the model to various tasks, this study investigates how including virtual edges affects the statistical properties of the resulting networks. In addition to determining whether these networks can detect gibberish text, we evaluate, metric by metric, whether the model is more sensitive to stylistic/syntactic features or semantic aspects of the texts.

## Materials and methods

The methodology adopted in this work to analyze the statistical properties of enriched co-occurrence networks consists of two main phases: (i) network creation and (ii) topology characterization and analysis. In the network creation phase, we first pre-process the text, followed by constructing the network and adding virtual edges, effectively enriching the network with virtual links. During the topology characterization and analysis phase, metrics are extracted and classified based on two criteria: (i) *informativeness*, referring to the metric's ability to distinguish between meaningful and meaningless texts, and (ii) *variability ratio*, which evaluates whether the metric is more sensitive to syntactic or semantic features of the text.

### Network construction

To construct networks from texts, we employed several pre-processing steps, including tokenization (i.e., identifying individual words), removing punctuation, and filtering out irrelevant information such as spaces and non-alphabetic characters. The removal of stop-words is

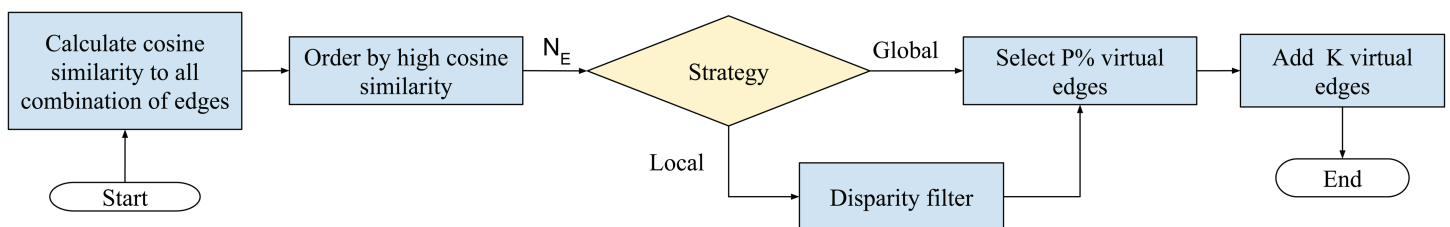
considered an option; therefore, we have two types of text: one with the complete text including stop-words and the other with stopwords removed. After pre-processing, the texts are transformed into co-occurrence networks. In these networks, each unique word in the text becomes a node, and edges are created between nodes representing words that appear adjacent to each other in the text. The total number of edges in the network at the end of this phase is denoted by  $N_E$ .

Originally, word co-occurrence networks were designed to connect only adjacent words. However, this approach proved inefficient for modeling shorter texts. Recent research has shown that incorporating virtual edges – connections between similar but non-adjacent words – provides valuable information, enabling more effective analysis of shorter texts [23]. In a co-occurrence network extracted from very short text, the topology is almost linear. The inclusion of virtual edges increases the complexity of the network, enhancing its usefulness for classification tasks [23].

Fig 1 illustrates the process of adding virtual edges to enrich the network’s structure with semantic information. First, we obtain the word embeddings for each word in the network and calculate the cosine similarity between words that are unconnected in the original co-occurrence model (i.e., the co-occurrence network without virtual edges). This cosine similarity then becomes the weight of the potential virtual edge. While all these pairs of nodes represent possible virtual edges that could be added to the network, including all of them would result in a network (clique) that lacks meaningful topological information, especially when analyzing unweighted metrics. To address this, we applied different criteria to retain only the most significant links: the global and local strategies.

- *Global Strategy (G)*: potential virtual edges are ranked by their weights, and the top  $K$  edges are selected for inclusion. According to [8],  $K$  is selected as a percentage  $P$  of the total edges in the original co-occurrence network ( $N_E$ ), i.e.  $K = PN_E$ . In this case, we selected the following values:  $P = \{25\%, 50\%, 75\%, 100\%\}$ .
- *Local Strategy (L)*: Unlike the global strategy, this method considers the local topology of the final network after adding virtual edges. It retains only the most significant virtual edges, removing non-relevant ones while preserving the key features of the network. This approach, known as the disparity filter [28], establishes a null model to quantify the probability of a node being connected to an edge with a given weight, considering its other connections. This probability is then used to evaluate the significance of the edge. Specifically, the significance of an edge  $e_{ij}$  is measured by  $\alpha_{ij}$ :

$$\alpha_{ij} = 1 - (k_i - 1) \int_0^{\pi_{ij}} (1 - x)^{k_i - 2} dx \tag{1}$$



**Fig 1. Methodology used in this study analyzes enriched networks.** All pairs of similarities are calculated, and the similarity weights are sorted in decreasing order. To filter the edges, the global strategy selects those with the highest weights across the entire network, while the local strategy evaluates the importance of an edge based on the local structure of each node. The total number of included edges is a parameter that varies throughout the analysis.

<https://doi.org/10.1371/journal.pone.0327421.g001>

$$\pi_{ij} = w_{ij} \left( \sum_{ik \in E} w_{ik} \right)^{-1}, \quad (2)$$

where  $w_{ij}$  is the weight of  $e_{ij}$ ,  $k_i$  is the degree of the  $i$ -th node and  $E$  is the set of all edges connected to node  $i$ . The high significance of an edge is represented by low values of  $\alpha_{ij}$ . In this context, the weights of the edges are computed as the embedding similarity using cosine similarity (for virtual edges). To preserve the edges obtained from co-occurrence, we assigned them the maximum similarity value (i.e., 1). The disparity filter removes the edges with the lowest significance, leaving  $K$  edges. For comparison,  $K$  is set to match the number of edges included in the global strategy (i.e.,  $K$  is a fraction of the total edges in the co-occurrence network, excluding virtual edges).

We used the pre-trained FastText model [29] to map words to embeddings. These embeddings were trained on large-scale corpora, including Common Crawl and Wikipedia, providing comprehensive coverage of general language usage. FastText operates at the character level, eliminating the need for lemmatization. It is available for multiple languages and captures semantic information by projecting words into high-dimensional vector spaces, using 300-dimensional vectors. We opted not to evaluate other embedding models, as FastText offers strong cross-linguistic versatility and delivers performance comparable to models such as Word2Vec in practical applications involving enriched networks [8].

## Network analysis

Once the networks with virtual edges are constructed and the top  $K$  virtual edges are identified, the network can be analyzed using network metrics. The network analysis consists of the following steps:

1. *Network metrics extraction*: network metrics are extracted from the enriched networks.
2. *Metrics normalization*: to avoid bias in the metrics caused by network size, the extracted metrics are normalized.
3. *Statistical properties analysis*: this is the most important part of the study, as it assesses the informativeness of the metrics and determines whether they are more effective in capturing syntactic or semantic features of texts.

**Network metrics.** In our analysis, we focused on the most commonly used network metrics for analyzing networks derived from text. These metrics include average shortest path length ( $L$ ), clustering coefficient ( $C$ ), closeness centrality ( $CC$ ), betweenness centrality ( $B$ ), PageRank ( $PR$ ), and eigenvector centrality ( $EV$ ) [11,30]. For metrics calculated at the node level, summarization is necessary. In other words, for each local metric computed for individual nodes in the network, we aim to condense the information into a single value that represents the entire network. Following [11], two types of summarization can be considered:

1. The average measure across all nodes, denoted as  $\bar{X}$ .
2. The average measure for the most important words, denoted as  $\bar{X}^*$ . The most important words are defined as the top 10 most frequent words in each text.

**Metrics normalization.** Network metrics derived from texts can have their metrics dependent on text size (such as the number of tokens or vocabulary size). For this reason, we

adopted the following procedure to normalize the metrics. For each original text, we generated 10 shuffled versions, where the shuffling process is performed at the word level. Let  $\tilde{X}^R$  represent the value of a specific measure computed for each shuffled text. If  $\mu(\tilde{X}^R)$  denotes the average value computed across the shuffled versions, we define the normalized value of the metric,  $X$ , as

$$X = \frac{\tilde{X}}{\mu(\tilde{X}^R)}. \quad (3)$$

Considering the inherent uncertainty in the random shuffling process, the error associated with  $X$  is quantified as

$$\epsilon(X) = \frac{\sigma(\tilde{X}^R)}{\mu(\tilde{X}^R)} X, \quad (4)$$

where  $\sigma(\tilde{X}^R)$  is the standard deviation computed over the shuffled versions.

**Statistical properties analysis.** The first key property addressed in this paper is the *informativeness* of the metric. A metric is considered informative if it can effectively differentiate between real and shuffled texts. This property is crucial as it indicates the metric's ability to identify texts where semantic words are combined in a nonsensical manner. An informative metric is valuable not only for detecting such anomalies but also for capturing subtle nuances in text styles. If a metric is not informative, its normalized value is expected to be close to  $X = 1$ . To quantify informativeness, we measure the distance ( $D$ ) of  $X$  from 1 for each text (network) and normalize this distance by the error  $\epsilon(X)$ , i.e.:

$$D = \frac{X - 1}{\epsilon(X)}. \quad (5)$$

If the distance  $D > 1$ , it indicates that the measure is informative for the text being analyzed. To assess the informativeness of the dataset, we used a criterion that measures the proportion of texts where  $D > 1$ . This is defined as:

$$I = \frac{|D > 1|}{N_T}, \quad (6)$$

where  $|D > 1|$  denotes the number of texts for which the condition  $D > 1$  is satisfied, and  $N_T$  represents the total number of texts analyzed in the dataset.

When virtual edges are introduced into the network, the structure becomes richer because semantically similar words that were not previously connected are now linked. This can enhance the network's ability to capture semantic relationships; however, it may also introduce drawbacks in terms of informativeness. If informativeness remains high, it often indicates that word order continues to play an important role and that adding virtual edges does not significantly impair the model's ability to recognize structural patterns in the text. Conversely, if informativeness decreases, it suggests that the metric's ability to distinguish meaningful from shuffled texts is weakened, meaning that changes in word order are less perceptible to the system.

The property being captured by the metrics can shift when virtual edges are added. Most metrics in traditional co-occurrence networks tend to reflect syntactic and stylistic features, such as local word order and grammatical patterns [11,31]. However, with the addition of virtual edges, the network increasingly encodes semantic similarity, which may cause the metrics to capture semantic information rather than syntactic/stylistic structure. In other words, the addition of virtual edges may make the model more sensitive to semantics than to syntax.

Determining whether a model captures syntactic or semantic information helps define the types of tasks to which the model/metric can be applied. For instance, in tasks where style or language plays a crucial role, metrics that depend on syntax are more suitable. This is particularly relevant when identifying the nature of an unknown sequence of symbols or when determining the authorship of a text. Conversely, metrics that are more dependent on semantics could be used to identify, for example, shifts in semantic flow within texts, as in tasks like topic segmentation [23].

To determine whether a measure  $X$  is more dependent on linguistic structure (syntax) than on content (semantics), we used two datasets. The first dataset consists of the same text translated into different languages. For this, we used the New Testament, translated into different languages (NLANG dataset, as presented in Sect). This dataset allows us to assess the variability of a metric when the semantics remain constant, but the syntax changes with each language. The second dataset was used to measure the variability of semantics across different texts in a single language. For this purpose, we used a collection of texts (novels) in English (NEN dataset, as presented in Sect).

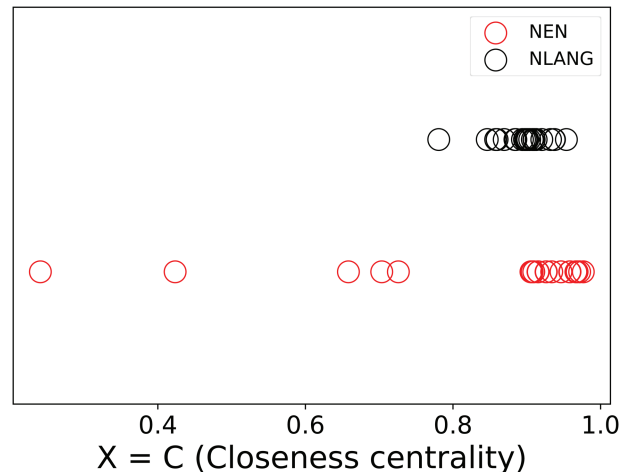
By analyzing both datasets, we were able to measure, for each metric  $X$ , the variability across syntax and semantics. The variability across syntax is computed using the coefficient of variation  $v_{(t=nt,l)}$ , where  $t = nt$  indicates that the variation is computed for a specific text (the New Testament), and  $l$  indicates that the dataset comprises texts in different languages (NLANG dataset). The variability across semantics is computed using the coefficient of variation  $v_{(t,l=en)}$ . This means that we measured the variability within a dataset where the language is constant (English) and the textual content (i.e. the semantics) differ. This corresponds to the NEN dataset. To compute the nature of a metric  $X$  we compute the following variability ratio between  $v_{(t=\tau,l)}$  and  $v_{(t,l=\lambda)}$ :

$$V_R = \frac{v_{(t=nt,l)}}{v_{(t,l=en)}}. \quad (7)$$

$V_R$  can be used to determine whether a metric is more dependent on syntax or semantics. If  $V_R > 1$ , it indicates that the variability across syntax is greater than the variability across semantics, meaning the metric is more dependent on syntax [11]. To illustrate this concept, Fig 2 shows the behavior of the metric  $X = C$  using hypothetical values from both the NEN and NLANG datasets. The figure clearly shows that the variability across texts in the same language is greater than the variability of the same text across different languages. This suggests that the metric is more influenced by semantics than by syntax.

## Dataset

Two different datasets were used in this study, selected to enable direct comparison with previous research evaluating the statistical properties of texts modeled as traditional co-occurrence networks [11]. The first dataset, referred to as the NEN dataset, was extracted from Project Gutenberg (<http://www.gutenberg.org>) and consists of English subtexts extracted from various novels, with varying text lengths. The books included are David Copperfield (Charles Dickens); Dracula (Bram Stoker); Evelina, Or, the History of a Young Lady's Entrance into the World (Fanny Burney); Great Expectations (Charles Dickens); History of Tom Jones, a Foundling (Henry Fielding); Moby Dick; Or, The Whale (Herman Melville); Persuasion (Jane Austen); Pride and Prejudice (Jane Austen); The Life and Adventures of Robinson Crusoe (Daniel Defoe); and Ulysses (James Joyce). The text sizes used for each book are 200, 400, 800, and 1,000 words, corresponding to the first tokens of the texts. This dataset is used to measure the variability of metrics extracted from different texts in the same language.



**Fig 2. Example of hypothetical closeness centrality values obtained from the NEN and NLANG datasets.** Because the variability in the NEN dataset is greater than that in the NLANG dataset, we have  $V_R < 1$ . This indicates that the metric is more dependent on semantics than on syntax.

<https://doi.org/10.1371/journal.pone.0327421.g002>

The second dataset, referred to as the NLANG dataset, consists of subtexts of varying sizes extracted from the New Testament of the Bible. The texts are translated into Arabic, English, Esperanto, German, Hebrew, Hungarian, Korean, Latin, Portuguese, Russian, and Vietnamese. This dataset is used to measure the variability of metrics extracted from the same content across distinct languages.

## Results and discussion

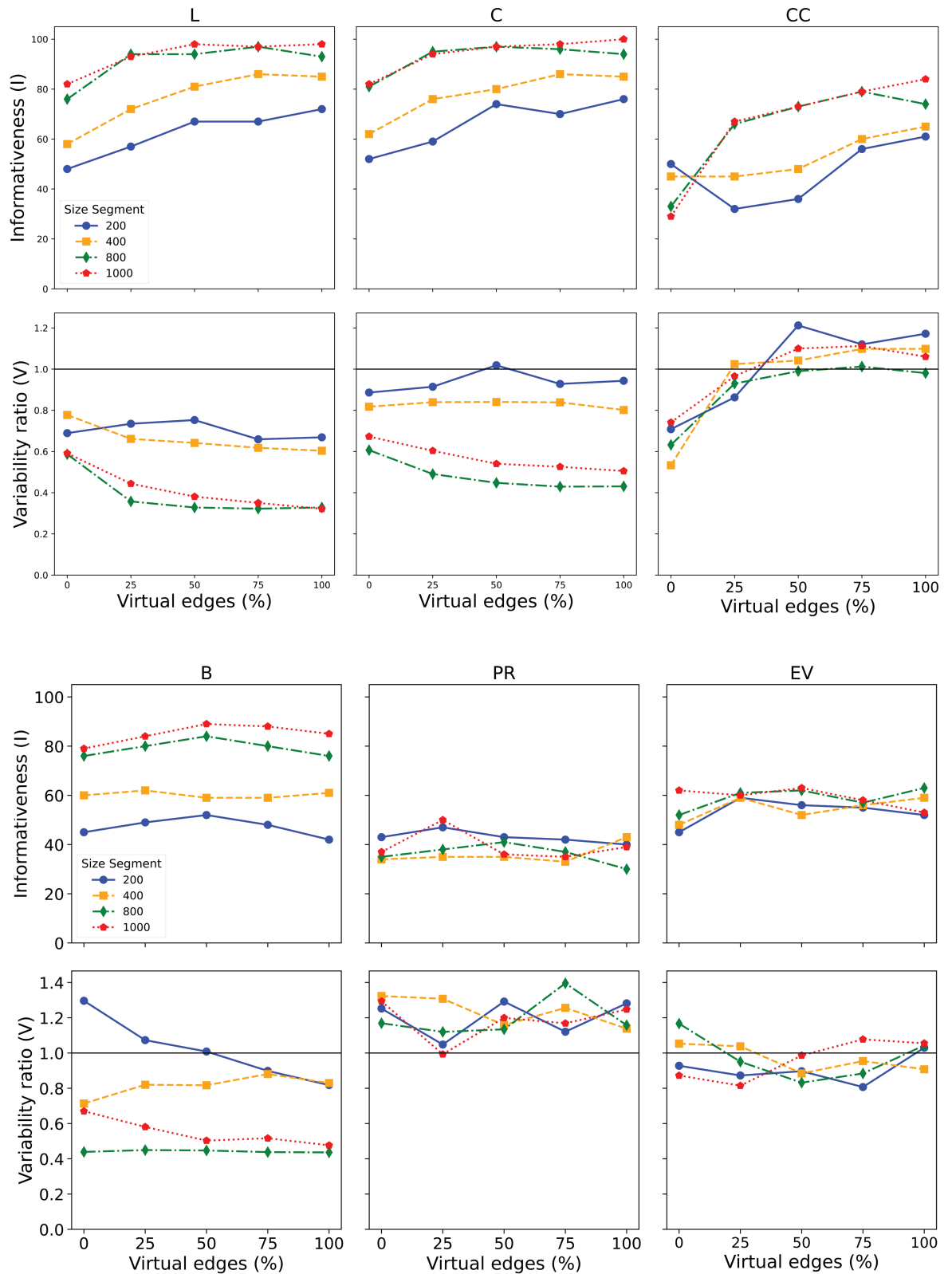
### Informativeness and variability ratio analysis

In this section, we analyze the behavior of informativeness and the variability ratio for the selected complex network metrics. Our analysis focuses on networks derived from texts *without stopwords*. Additionally, we focus our discussion on the global thresholding strategy, as the local approach yielded similar results.

Fig 3 illustrates the informativeness and variability ratio of network metrics, considering different text sizes and the network weight thresholding based on the global strategy. We begin by analyzing the network metrics computed across all the nodes in the network (i.e.,  $X$  as opposed to  $X^*$ ). The results obtained for the local strategy are shown in the Supporting information.

The main findings in Fig 3 are summarized below:

- *Average shortest path (L)*: The informativeness of  $L$  clearly depends on text size. Without virtual edges,  $L$  tends to be more informative for longer text segments. Adding virtual edges enhances the informativeness of the metric, especially for shorter texts. For texts longer than 800 tokens, the inclusion of virtual edges ensures that informativeness is maintained across all texts in the dataset. The variability ratio suggests that the normalized metric, as defined in our methodology, is more influenced by semantics than syntax in shorter texts. This effect becomes even more pronounced when virtual edges are included in shorter texts.
- *Closeness centrality (C)*: The informativeness behavior is similar to the average shortest path length, as the metrics are correlated. However, regarding the variability ratio, in very short



**Fig 3. Global Strategy Analysis.** Distribution of Informativeness and Variability measures for the Average Shortest Path (L), Closeness Centrality (C), Clustering Coefficient (CC), Betweenness Centrality (B), PageRank (PR), and Eigenvector Centrality (EV), with the addition of virtual edges in networks generated with variate text sizes and with filtering stop-word.

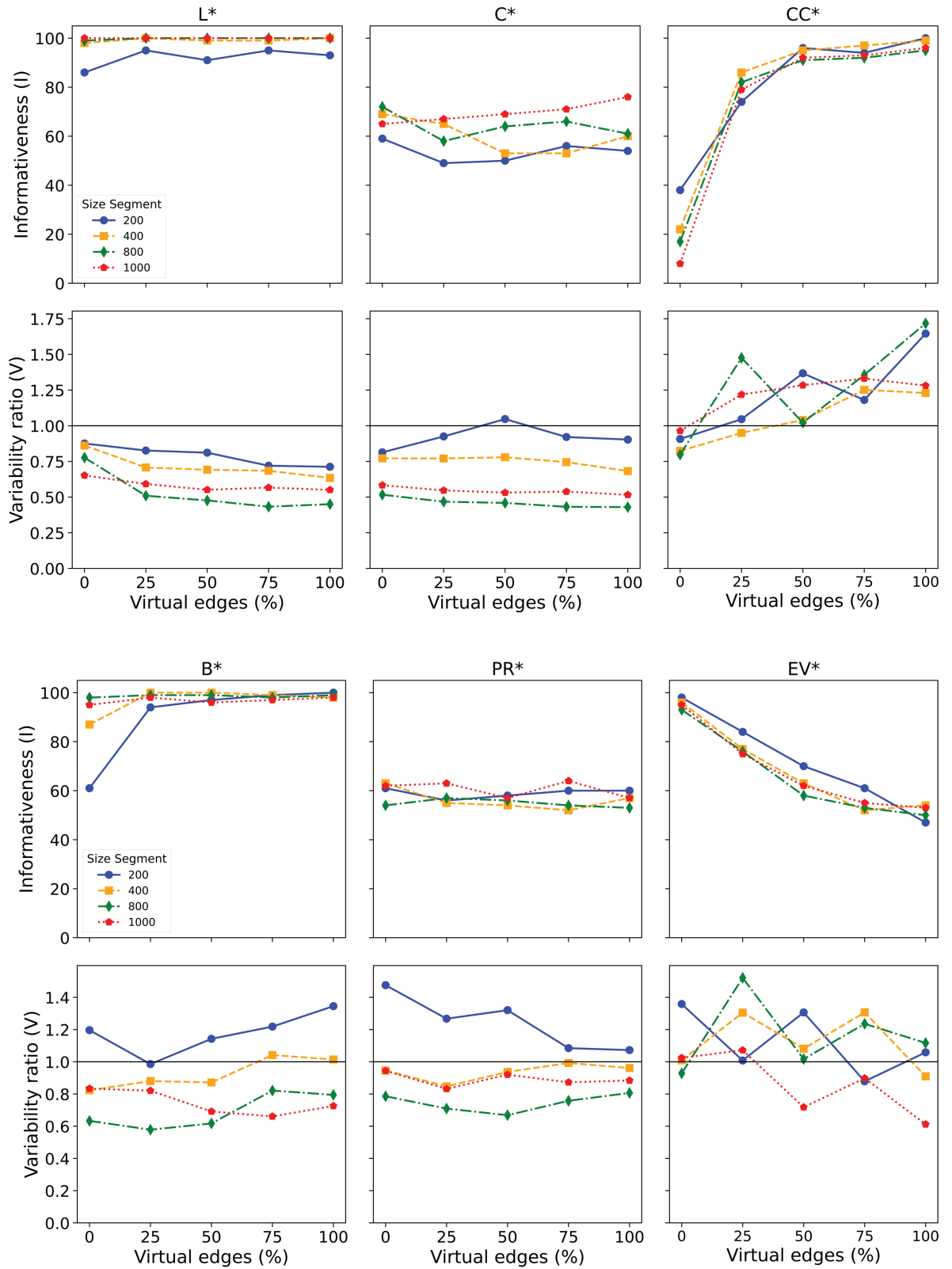
<https://doi.org/10.1371/journal.pone.0327421.g003>

segments (200 tokens), there is no clear dominance in the ability to capture either syntax or semantics, even with the inclusion of virtual edges.

- *Clustering Coefficient (CC)*: the clustering coefficient behaves differently from  $L$  and  $C$ . For short texts, an increase in the number of edges can lead to a decrease in informativeness. However, for longer segments (over 400 tokens), the inclusion of edges enhances the informativeness metric. In these cases,  $I$  can rise from approximately 30% to 80% when virtual edges are included. The variability ratio also shows interesting behavior. In very short texts,  $CC$  seems more influenced by semantic features in traditional co-occurrence documents. However, with the addition of virtual edges, this behavior reverses, and  $CC$  becomes more dependent on syntax.
- *Betweenness Centrality (B)*: the inclusion of virtual edges has a minimal effect on informativeness. However, the results indicate that the betweenness informativeness for longer texts tends to be significantly higher than for shorter documents. Conversely, the behavior of the variability ratio depends on the size of the text. Larger texts remain unaffected, while for shorter texts, the variability ratio shifts from over 1.30 to 0.80 when 100% of virtual edges are included, suggesting that the addition of edges increases the metric's sensitivity to semantics.
- *PageRank (PR)*: the informativeness of PageRank consistently remains low across all text sizes, highlighting its limited sensitivity to incorporating virtual edges. This suggests that this metric may not be effective in distinguishing between gibberish and more detailed stylistic information in unweighted co-occurrence networks. The variability ratio exhibits oscillatory changes with the addition of virtual edges showing a consistent dependence on syntax across all text sizes.
- *Eigenvector Centrality (EV)*: the informativeness of this metric is similar to that of PageRank, though the values for  $EV$  are slightly higher. Once again, the informativeness does not change significantly with the inclusion of virtual edges. Similarly, the variability ratio is also minimally affected by the addition of virtual edges. The variability ratio values suggest that there is no dominant dependence on either syntactic or semantic features.

We now analyze the network metrics that are computed exclusively based on the most frequent words in the text (i.e.,  $X^*$  as opposed to  $X$ ). The results for the global thresholding strategy are presented in Fig 4. The main results are summarized below:

1. *Average shortest path ( $L^*$ )*: The informativeness of  $L^*$  is less sensitive to the inclusion of virtual edges compared to  $L$ , resulting in a relatively flat behavior across most text sizes. Very short text segments, however, show a greater effect from the addition of virtual edges. Despite this lower sensitivity,  $L^*$  still maintains high informativeness even without enrichment. Similarly, the variability ratio indicates that  $L^*$  is only weakly influenced by the inclusion of virtual edges.
2. *Closeness centrality ( $C^*$ )*: the closeness shows little dependency with the inclusion of virtual edges, even for short texts. The informativeness for larger texts increases by a small margin. The variability ratio seems to be more dependent on text size than on the inclusion of virtual edges. Larger pieces of texts tend to capture more semantical features, similarly to the behavior of  $C$ .
3. *Clustering Coefficient ( $CC^*$ )*: The inclusion of virtual edges has a significant impact on informativeness. Even a minimal inclusion can lead to a substantial increase in informativeness across all text sizes. Additionally, the variability ratio is influenced by the presence of virtual edges, with values exceeding the threshold of 1 for all text sizes. When compared to  $CC$ , the gain in informativeness appears to be stronger in this case.



**Fig 4. Global Strategy Analysis.** Distribution of Informativeness and Variability measures for the Average Shortest Path (L\*), Closeness Centrality (C\*), Clustering Coefficient (CC\*), Betweenness Centrality (B\*), PageRank (PR\*), and Eigenvector Centrality (EV\*), with the addition of virtual edges in networks generated with variate text sizes and with filtering stop-word.

<https://doi.org/10.1371/journal.pone.0327421.g004>

4. *Betweenness* ( $B^*$ ): The informativeness of betweenness primarily affects shorter texts. After adding 50% virtual edges, all text sizes become 100% informative. This behavior differs from what was observed with  $B$ , where the informativeness curve appeared to be strongly dependent on text size. Regarding the variability ratio, there is a clear dependence on text size. Differently from  $B$ , adding 50% or more virtual edges to very short texts seems to increase the variability ratio.
5. *PageRank* ( $PR^*$ ): The informativeness of PageRank appears to remain unaffected by the inclusion of virtual edges, regardless of text size. This behavior is consistent with what has been observed for  $PR$ . Conversely, the variability ratio is more impacted in very short texts, leading to decreased variability.
6. *Eigenvector Centrality* ( $EV^*$ ): Unlike  $EV$ , informativeness is significantly affected. We observe that the informativeness values decrease sharply with the inclusion of virtual edges, and this trend is consistent across all text sizes. This contrasts with  $EV$ , where informativeness is only mildly affected. However, the variability ratio exhibits an oscillatory pattern, indicating this metric may be very sensitive to including virtual edges.

### Impact of stop-words

In Table 1, we present the impact of including or filtering stopwords in the network analysis. Specifically, we focus on how the highest informativeness values—across varying numbers of virtual edges—are affected by the inclusion or exclusion of stopwords. Similarly, when analyzing the variability ratio, we analyze how the metric changes in the scenario where the network shows the greatest dependence on either syntax or semantics, corresponding to the minimum or maximum variability ratio values. The analysis covers both global ( $G$ ) and local ( $L$ ) thresholding strategies. For comparison, we also provide the results for the original ( $O$ ) network (i.e., the network without virtual links). Additionally, the results are segmented by text size. We focus our discussion, however, on the shortest and longest text sizes.

For texts of size 200 tokens, we observe that in the global strategy, filtering stopwords significantly enhances the informativeness of the  $L$  and  $C$  metrics, increasing their values by over 20%. However, a decrease in informativeness is observed for both  $PR$  and  $EV$ . The only metric that remains practically unaffected is betweenness, indicating that it is relatively stable regardless of the presence of stopwords. Using the local strategy does not seem to strongly affect the margin of gain (or loss) in terms of informativeness. In the global thresholding analysis, the largest changes in the variability ratio are observed for  $L$ ,  $CC$ , and  $B$ . These metrics decrease in value and eventually shift their dependency from syntax to semantics.

For texts comprising 1000 tokens and using the global thresholding strategy,  $L$ ,  $C$ , and  $B$  show a significant increase in informativeness, while  $PR$  experiences only a slight improvement. Conversely,  $EV$  displays a large decrease. This is an intriguing finding, as  $PR$  and  $EV$  are typically correlated, yet they exhibit distinct behavior when stopwords are included in the analysis. Once again, the local strategy does not strongly affect this analysis. Regarding the variability ratio, apart from  $PR$  and  $EV$ , the values decrease significantly, indicating that the metrics become less dependent on syntax when stopwords are filtered.

We now shift our analysis to metrics based solely on the most frequent words in the text. The results are presented in Table 2. For shorter text lengths (200 tokens), the informativeness improves significantly when stopwords are filtered for  $L^*$ ,  $CC^*$ ,  $B^*$ , and  $EV^*$ . However, informativeness slightly decreases for  $C^*$  and drops considerably for  $PR^*$ . The variability ratio decreases in most cases when applying the global strategy, except for  $PR^*$  and  $EV^*$ .

When analyzing larger text segments (1,000 tokens), excluding stopwords improves informativeness only for  $L^*$  and  $B^*$ . Significantly, the informativeness of  $PR^*$  decreases. Regarding

**Table 1. Comparison of informativeness and variability metrics. We compared the Original network (O), Global strategy (G), and Local strategy (L), with and without stop-words filtering (SW). High values of informativeness measures in networks incorporating virtual edges are highlighted, along with their respective variability.**

	Informativeness						Variability					
	$\mathcal{O}$	$\mathcal{O}_F$	$\mathcal{G}$	$\mathcal{G}_F$	$\mathcal{L}$	$\mathcal{L}_F$	$\mathcal{O}$	$\mathcal{O}_F$	$\mathcal{G}$	$\mathcal{G}_F$	$\mathcal{L}$	$\mathcal{L}_F$
Text size 200												
L	42.0	48.0	52.0	72.0	52.0	72.0	2.3	0.7	1.4	0.7	1.4	0.7
C	48.0	52.0	55.0	76.0	53.0	76.0	2.3	0.9	1.5	0.9	1.5	0.9
CC	47.0	50.0	76.0	61.0	76.0	61.0	1.7	0.7	1.9	1.2	1.9	1.2
B	40.0	45.0	51.0	52.0	51.0	54.0	2.2	1.3	1.7	1.0	1.7	1.1
PR	58.0	43.0	60.0	47.0	47.0	51.0	1.1	1.3	1.1	1.0	1.0	1.2
EV	46.0	45.0	66.0	59.0	67.0	52.0	1.1	0.9	0.8	0.9	1.3	1.0
Text size 400												
L	46.0	58.0	59.0	86.0	59.0	85.0	2.1	0.8	1.6	0.6	1.6	0.6
C	45.0	62.0	54.0	86.0	58.0	85.0	2.1	0.8	1.6	0.8	1.7	0.8
CC	46.0	45.0	83.0	65.0	83.0	65.0	2.0	0.5	1.9	1.1	1.9	1.1
B	43.0	60.0	53.0	62.0	53.0	61.0	1.9	0.7	1.7	0.8	1.7	0.8
PR	41.0	34.0	43.0	43.0	45.0	46.0	0.9	1.3	1.1	1.1	0.9	1.0
EV	54.0	48.0	74.0	59.0	76.0	59.0	1.2	1.1	0.7	1.0	0.7	0.9
Text size 800												
L	52.0	76.0	59.0	97.0	59.0	93.0	1.9	0.6	1.2	0.3	1.2	0.3
C	50.0	81.0	60.0	97.0	60.0	94.0	2.0	0.6	1.3	0.4	1.3	0.4
CC	38.0	33.0	96.0	79.0	96.0	74.0	2.2	0.6	2.0	1.0	2.0	1.0
B	45.0	76.0	60.0	84.0	58.0	77.0	1.4	0.4	1.3	0.4	1.2	0.4
PR	55.0	35.0	55.0	41.0	37.0	45.0	1.2	1.2	1.2	1.1	1.3	1.1
EV	56.0	52.0	78.0	63.0	83.0	63.0	1.2	1.2	0.7	1.0	0.7	1.0
Text size 1000												
L	68.0	82.0	74.0	98.0	74.0	99.0	2.3	0.6	1.6	0.4	1.7	0.4
C	62.0	82.0	73.0	100.0	72.0	100.0	2.4	0.7	1.7	0.5	1.6	0.5
CC	54.0	29.0	91.0	84.0	90.0	84.0	2.0	0.7	1.7	1.1	1.5	1.1
B	59.0	79.0	70.0	89.0	66.0	88.0	1.8	0.7	1.6	0.5	1.6	0.5
PR	41.0	37.0	41.0	50.0	41.0	46.0	1.0	1.3	1.0	1.0	1.0	1.0
EV	67.0	62.0	79.0	63.0	85.0	62.0	1.5	0.9	0.8	1.0	0.8	0.9

<https://doi.org/10.1371/journal.pone.0327421.t001>

the variability ratio, excluding stopwords generally leads to decreased values. The ability to capture linguistic features clearly shifts from syntax to semantics for  $L^*$ ,  $C^*$ , and  $B^*$ . Interestingly, despite a significant decrease in the variability ratio for  $CC^*$ , this metric continues to capture syntactical features.

All in all, the results show that filtering stopwords may have a strong affect in both informativeness and variability ratio. This may happen independently of the chosen thresholding strategy.

### Conclusion

In this paper, we analyze the statistical properties of enriched complex networks for text analysis. Although enriched co-occurrence networks have been applied in various contexts, the impact of including virtual edges has not yet been studied. Our focus was on two main properties: (i) *informativeness*, which refers to a metric’s ability to distinguish between meaningful and nonsensical texts; and (ii) *variability ratio*, which reflects a metric’s ability to capture syntactic or semantic variations. We also evaluated two strategies for filtering virtual edges in enriched networks. While both the global thresholding and local (backbone-based) strategies were considered, our analysis focused on the global approach, as the local strategy yielded similar results in terms of informativeness and sensitivity to syntactic and semantic features.

**Table 2. Comparison of informativeness and variability metrics. We compared the Original network (O), Global strategy (G), and Local strategy (L), with and without stop-words filtering (SW). High values of informativeness measures in networks incorporating virtual edges are highlighted, along with their respective variability.**

	Informativeness						Variability					
	$\mathcal{O}$	$\mathcal{O}_F$	$\mathcal{G}$	$\mathcal{G}_F$	$\mathcal{L}$	$\mathcal{L}_F$	$\mathcal{O}$	$\mathcal{O}_F$	$\mathcal{G}$	$\mathcal{G}_F$	$\mathcal{L}$	$\mathcal{L}_F$
Text size 200												
L*	61.0	86.0	61.0	95.0	65.0	97.0	1.8	0.9	1.8	0.8	1.3	0.8
C*	64.0	59.0	64.0	59.0	69.0	64.0	1.7	0.8	1.7	0.8	1.4	0.8
CC*	79.0	38.0	79.0	100.0	79.0	99.0	1.7	0.9	1.7	1.6	1.7	1.8
B*	39.0	61.0	59.0	100.0	59.0	100.0	1.8	1.2	1.9	1.3	1.9	1.3
PR*	73.0	61.0	88.0	61.0	88.0	78.0	2.1	1.5	1.5	1.5	1.5	1.1
EV*	64.0	98.0	80.0	98.0	80.0	97.0	2.1	1.4	1.0	1.4	1.2	1.3
Text size 400												
L*	68.0	98.0	68.0	100.0	70.0	100.0	1.7	0.9	1.7	0.7	1.2	0.6
C*	74.0	69.0	78.0	69.0	79.0	69.0	1.7	0.8	1.3	0.8	1.2	0.8
CC*	91.0	22.0	91.0	99.0	91.0	99.0	2.1	0.8	2.1	1.2	2.1	1.5
B*	45.0	87.0	64.0	100.0	64.0	100.0	1.4	0.8	1.3	0.9	1.3	0.9
PR*	91.0	63.0	99.0	63.0	99.0	83.0	1.1	0.9	1.0	0.9	1.1	0.9
EV*	77.0	96.0	93.0	96.0	93.0	95.0	1.6	1.0	1.0	1.0	1.1	1.0
Text size 800												
L*	75.0	99.0	81.0	100.0	85.0	100.0	1.6	0.8	0.9	0.5	1.2	0.6
C*	81.0	72.0	85.0	72.0	89.0	72.0	1.7	0.5	1.4	0.5	1.3	0.5
CC*	98.0	17.0	98.0	95.0	98.0	96.0	2.0	0.8	2.0	1.7	2.0	1.2
B*	49.0	98.0	86.0	99.0	86.0	99.0	1.6	0.6	1.5	0.6	1.5	0.6
PR*	100.0	54.0	100.0	57.0	100.0	75.0	1.7	0.8	1.7	0.7	1.7	0.6
EV*	80.0	93.0	95.0	93.0	97.0	94.0	2.2	0.9	0.6	0.9	0.6	0.9
Text size 1000												
L*	78.0	100.0	78.0	100.0	86.0	100.0	1.7	0.7	1.7	0.7	1.2	0.7
C*	80.0	65.0	83.0	76.0	88.0	76.0	1.7	0.6	1.5	0.5	1.3	0.5
CC*	97.0	8.0	97.0	96.0	97.0	95.0	2.5	1.0	2.5	1.3	2.5	1.2
B*	60.0	95.0	85.0	98.0	85.0	98.0	1.4	0.8	1.3	0.8	1.3	0.7
PR*	98.0	62.0	100.0	64.0	100.0	73.0	2.0	0.9	1.6	0.9	1.9	0.8
EV*	78.0	95.0	95.0	95.0	97.0	95.0	2.3	1.0	0.8	1.0	0.9	1.0

<https://doi.org/10.1371/journal.pone.0327421.t002>

Our analysis revealed several interesting results. For instance, the addition of virtual edges can enhance the informativeness of certain metrics, such as the average shortest path—particularly in shorter texts—and closeness centrality. However, caution is warranted, as some metrics, like the clustering coefficient in short texts, may experience a decline in informativeness. Interestingly, other metrics, such as betweenness, appear to be unaffected by the inclusion of virtual edges in terms of informativeness.

Regarding the variability ratio analysis, we found that the inclusion of virtual edges in short texts increases the sensitivity of the average shortest path to semantics. Other metrics, such as eigenvector centrality, showed little effect from virtual edges and did not exhibit a clear dominance between semantic or syntactic features. Interestingly, the nature of certain metrics can change depending on the number of edges added. For instance, in the case of clustering and betweenness in short texts, the variability ratio decreases with the inclusion of virtual edges, shifting sensitivity from syntax to semantics. All the results confirm that the inclusion of virtual edges can play a significant role in determining which metric is best suited for use in specific NLP applications.

Overall, our results provide valuable insights for NLP research that relies on topological and semantic analysis, whether applied to short or longer text segments. While previous studies have shown that enrichment can enhance classification performance [8], our findings emphasize that network metrics must be carefully selected based on the degree of enrichment and the specific nature of the classification task at hand.

The main limitation of the proposed strategy is its application to extremely short texts, such as those commonly found on social networks. In such cases, even with enrichment, the resulting network may be too small to exhibit meaningful topological patterns. Another potential limitation of our method lies in its reliance on word embeddings, which inherently depend on the corpus used for their training. Although we adopted FastText embeddings in this study—due to their strong multilingual coverage, enabling the analysis of texts across diverse languages—the choice of embedding model could influence how virtual edges impact the network structure and, consequently, the behavior of network metrics. Different embeddings might shift the balance between syntactic and semantic sensitivity, potentially affecting the outcomes of enriched network analyses. Therefore, future work should extend the present framework by evaluating the effects of alternative embedding models, such as those trained on domain-specific corpora, which could further enhance the effectiveness of enriched networks, particularly in contexts requiring precise terminology or nuanced language, such as medical literature or historical texts.

We considered semantic edges occurring within the same network as the co-occurrence edges. In future work, distinct layers could be used to better leverage the network representation, with each layer capturing either co-occurrence or semantic similarity. For instance, in tasks where semantics are more relevant — such as topic segmentation — one could design metrics that prioritize the semantic layer. Conversely, tasks focused on style analysis could benefit more from features derived from the co-occurrence layer. Another type of analysis could involve examining the centrality of nodes across different layers and investigating how these variations influence the performance on the target task. Another possibility to improve the model is to integrate word vectors into the nodes using additional information, such as domain-specific data, to further enhance the representational capacity of the networks.

While our focus was on informativeness, this work could be extended to consider other types of noise in texts beyond shuffled words. For example, we could analyze the impact on metrics when authors attempt to conceal their identity in texts, such as during anonymization. This could involve changes in writing style or the introduction of errors, including synonym or antonym replacement, as well as other forms of word substitution. Additional possibilities include simulating typos, sentence fragmentation, or introducing ambiguity. All of these strategies could prove useful in the task of author masking identification.

## Supporting information

[S1 Fig](#) and [S2 Fig](#) illustrate how the informativeness and variability ratios of the metrics behave when applying the local strategy for pruning semantic edges.

**S1 Fig. Local Strategy: Distribution of Informativeness and Variability measures for the Average Shortest Path (L), Closeness Centrality (C), Clustering Coefficient (CC), Betweenness Centrality (B), PageRank (PR), and Eigenvector Centrality (EV), with the addition of virtual edges in networks generated with variate text sizes and with filtering stop-word.**

(PDF)

**S2 Fig. Local Strategy: Distribution of Informativeness and Variability measures for the Average Shortest Path (L\*), Closeness Centrality (C\*), Clustering Coefficient (CC\*), Betweenness Centrality (B\*), PageRank (PR\*), and Eigenvector Centrality (EV\*), with the addition of virtual edges in networks generated with variate text sizes and with filtering stop-word.**

(PDF)

## Author contributions

**Conceptualization:** Diego R. Amancio.

**Formal analysis:** Diego R. Amancio.

**Investigation:** Diego R. Amancio, Laura V. C. Quispe.

**Methodology:** Diego R. Amancio.

**Project administration:** Diego R. Amancio.

**Software:** Diego R. Amancio, Laura V. C. Quispe.

**Supervision:** Diego R. Amancio.

**Validation:** Laura V. C. Quispe.

**Visualization:** Jeaneth Machicao.

**Writing – original draft:** Diego R. Amancio.

**Writing – review & editing:** Diego R. Amancio, Jeaneth Machicao, Laura V. C. Quispe.

## References

1. Akimushkin C, Amancio DR, Oliveira ON Jr. Text authorship identified using the dynamics of word co-occurrence networks. *PLoS One*. 2017;12(1):e0170527. <https://doi.org/10.1371/journal.pone.0170527> PMID: 28125703
2. Amancio DR. Comparing the topological properties of real and artificially generated scientific manuscripts. *Scientometrics*. 2015;105(3):1763–79.
3. Tixier A, Skianis K, Vazirgiannis M. GoWvis: a web application for graph-of-words-based text visualization and summarization. In: *Proceedings of ACL-2016 System Demonstrations*. 2016. p. 151–6.
4. Liu H, Cong J. Language clustering with word co-occurrence networks based on parallel texts. *Chin Sci Bull*. 2013;58(10):1139–44. <https://doi.org/10.1007/s11434-013-5711-8>
5. Vera J, Palma W. The community structure of word co-occurrence networks: experiments with languages from the Americas. *Europhys Lett*. 2021;134(5):58002.
6. Ferrer I Cancho R, Solé RV. The small world of human language. *Proc Biol Sci*. 2001;268(1482):2261–5. <https://doi.org/10.1098/rspb.2001.1800> PMID: 11674874
7. Borges L, Correa EA, Oliveira ON, Amancio DR, Lessa L, Aluisio SM. Enriching complex networks with word embeddings for detecting mild cognitive impairment from speech transcripts. In: *Proceedings of the 1st Conference on Association for Computational Linguistics*; 2017. p. 1284–96.
8. Quispe LV, Tohalino JA, Amancio DR. Using virtual edges to improve the discriminability of co-occurrence text networks. *Physica A: Statist Mech Appl*. 2021;562:125344.
9. Garg M, Kumar M. The structure of word co-occurrence network for microblogs. *Physica A: Statist Mech Appl*. 2018;512:698–720.
10. Tohalino JA, Silva TC, Amancio DR. Using word embedding to detect keywords in texts modeled as complex networks. *Scientometrics*. 2024.
11. Amancio DR, Altmann EG, Rybski D, Oliveira ON, da F Costa L. Probing the statistical properties of unknown texts: application to the voynich manuscript. *PLOS ONE*. 2013;8(7):1–10.
12. Montemurro MA, Zanette DH. Keywords and co-occurrence patterns in the voynich manuscript: an information-theoretic analysis. *PLoS One*. 2013;8(6):e66344. <https://doi.org/10.1371/journal.pone.0066344> PMID: 23805215
13. Estevez-Rams E, Mesa-Rodriguez A, Estevez-Moya D. Complexity-entropy analysis at different levels of organisation in written language. *PLoS One*. 2019;14(5):e0214863. <https://doi.org/10.1371/journal.pone.0214863> PMID: 31067221
14. Stanisz T, Drożdż S, Kwapien J. Complex systems approach to natural language. *Phys Rep*. 2024;1053:1–84. <https://doi.org/10.1016/j.physrep.2023.12.002>
15. Cong J, Liu H. Approaching human language with complex networks. *Phys Life Rev*. 2014;11(4):598–618. <https://doi.org/10.1016/j.plrev.2014.04.004> PMID: 24794524

16. Wachs-Lopes GA, Rodrigues PS. Analyzing natural human language from the point of view of dynamic of a complex network. *Exp Syst Appl*. 2016;45:8–22. <https://doi.org/10.1016/j.eswa.2015.09.020>
17. Ferrer I Cancho R, Solé RV, Köhler R. Patterns in syntactic dependency networks. *Phys Rev E Stat Nonlin Soft Matter Phys*. 2004;69(5 Pt 1):051915. <https://doi.org/10.1103/PhysRevE.69.051915> PMID: 15244855
18. Liu H, Li W. Language clusters based on linguistic complex networks. *Chinese Sci Bullet*. 2010;55(30):3458–65.
19. Gao Y, Liang W, Shi Y, Huang Q. Comparison of directed and weighted co-occurrence networks of six languages. *Physica A: Statist Mech Appl*. 2014;393:579–89.
20. Liang W, Shi Y, Tse CK, Liu J, Wang Y, Cui X. Comparison of co-occurrence networks of the Chinese and English languages. *Physica A: Statist. Mech Appl*. 2009;388(23):4901–9.
21. Amancio DR, Oliveira Jr ON, Costa L da F. Using complex networks to quantify consistency in the use of words. *J Stat Mech*. 2012;2012(01):P01004. <https://doi.org/10.1088/1742-5468/2012/01/p01004>
22. Amancio DR. Network analysis of named entity co-occurrences in written texts. *EPL*. 2016;114(5):58005. <https://doi.org/10.1209/0295-5075/114/58005>
23. Corrêa EA, Amancio DR. Word sense induction using word embeddings and community detection in complex networks. *Physica A: Statist Mech Appl*. 2019;523:180–90.
24. Mikolov T, Corrado GS, Chen K, Dean J. Efficient estimation of word representations in vector space. *arXiv preprint* 2013. <https://arxiv.org/abs/1301.3781>
25. Pennington J, Socher R, Manning CD. GloVe: global vectors for word representation. In: *Empirical Methods in Natural Language Processing (EMNLP)*, 2014. p. 1532–43.
26. Bojanowski P, Grave E, Joulin A, Mikolov T. Enriching word vectors with subword information. *Trans Assoc Comput Linguist*. 2017;5:135–46.
27. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019. p. 4171–86.
28. Serrano MA, Boguñá M, Vespignani A. Extracting the multiscale backbone of complex weighted networks. *Proc Natl Acad Sci U S A*. 2009;106(16):6483–8. <https://doi.org/10.1073/pnas.0808904106> PMID: 19357301
29. Grave E, Bojanowski P, Gupta P, Joulin A, Mikolov T. Learning word vectors for 157 languages. In: *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*. 2018.
30. Stella M, Beckage NM, Brede M. Multiplex lexical networks reveal patterns in early word acquisition in children. *Sci Rep*. 2017;7:46730. <https://doi.org/10.1038/srep46730> PMID: 28436476
31. Ferrer I Cancho R, Solé RV, Köhler R. Patterns in syntactic dependency networks. *Phys Rev E Stat Nonlin Soft Matter Phys*. 2004;69(5 Pt 1):051915. <https://doi.org/10.1103/PhysRevE.69.051915> PMID: 15244855