Does network complexity help organize Babel’s library?

Juan Pablo Cárdenas,1 Iván González,2 Rosa María Benito,2 and Juan Carlos Losada2

1Instituto de Sistemas Complejos de Valparaíso, Subida Artillería 470, Valparaíso, Chile.
2Grupo de Sistemas Complejos, Dept. Física y Mecánica Fundamentales aplicadas a la I. Agroforestal, E.T.S.I. Agrónomos. Universidad Politécnica de Madrid (UPM). Ciudad Universitaria, s/n, 28040, Madrid, Spain

In this work, we study properties of texts from the perspective of complex network theory. Words in given texts are linked by co-occurrence and transformed into networks, and we observe that these display topological properties common to other complex systems. However, there are some properties that seem to be exclusive to texts; many of these properties depend on the frequency of words in the text, while others seem to be strictly determined by the grammar. Precisely, these properties allow for a categorization of texts as either with a sense and others encoded or senseless.

**keywords:** complex networks, word networks, Voynich manuscript, quantitative linguistics

## I. INTRODUCTION

In 1941, Jorge Luis Borges wrote a short story about a custodian of a peculiar repository that contains all the possible arrangements of letters that can be written. In this story, the eponymous Library of Babel[5] contains books of a certain size and specific characteristics. The books contain no pictures, only text: in addition, each book has 410 pages, each page has 40 rows, and each row, 80 characters. The alphabet mentioned in the story consists of only 25 orthographical symbols, including the space, the comma, and the point. Considering only these initial conditions, the library houses $25^{131,2000}$ different books.

The library thus stores all the “sensical” texts that could ever be written, in any language, and any of their possible variations. However, the number of sensical books is minimal in comparison to the huge number of possible, “nonsensical” combinations of words that lack meaning in any language. The librarians of this enormous collection are both charged with its custody and obsessed with finding those books that “say something”.

Although Borge wrote this narrative with different intentions, let us continue the Library of Babel’s fictional game. In order to find those texts with meaning, some reported statistical properties of sensical texts might be used as a first filter[17]. One of the most well-known and basic, but no less important, properties comes from Zipf’s law[31]. Roughly speaking, this law says that the number times a word appears in a text is a function of its ranked frequency of occurrence. The apparent ubiquitous applicability of the law to natural languages would reveal languages functioning[12]. The problem is that a text that follows Zipf’s law does not necessarily have to make sense. For illustration, in the library there is a book called the “Voynich manuscript” that fits the law[15] and other properties of “real” books[13], but has been unreadable and illegible for centuries. The manuscript can not be classified either as a hoax[26, 27] or a text with some sort of sophisticated encoding. The reason is that a seemingly nonsensical or encrypted text such as this manuscript could actually be a readable, but scrambled, text (keeping the frequency of words, but changing their order). Thus, although the text makes no sense for a reader who reads it word for word, the co-occurrence of the words would still comply with Zipf’s law.

Another problem in the Library would be the existence of texts that only make sense to certain machines. This texts were written for machines using formal programming languages. Formal languages correspond to the set of strings of symbols and may be constrained by grammatical rules. Their alphabets, frequently required to be finite[25], are the set of symbols, letters, or tokens from which the strings of the language may be formed. Strings formed from this alphabet can be found as words in texts present in the Library. Although formal languages have been developed intentionally and their words (called well-formed formulas) are linked by grammar specific to each (in contrast to natural ones, which evolved as biological systems), they appear very similar to those linked in a text written in a natural language, making identification as strictly “machine codes” very difficult.

Problems such as those previously mentioned was precisely what motivated this work: searching for statistical properties that are unique to sensical texts. Using techniques borrowed from the field of complex networks, this work looks for the topological properties of co-occurrence word networks that depend exclusively on the sense of a text. These must capture the use of a certain grammar that would be absent in senseless or encoded texts. Thus, by using corporuses written in different natural and formal languages, we studied the networks that represent those texts in order to obtain their common properties and to detect which of these properties might allow the librarians of Babel to organize their books.

The work is structured as follows. In the next section, we present the method for constructing word networks. In Section 3, we present the analysis and results of the networks. In the last section, we discuss the major implications of the obtained results and present the conclusions of our work.
II. NETWORK CONSTRUCTION

A text can be represented as a graph $G(W,E)$, where $W$ is the set of different words contained in the text, and $E$, the set of undirected edges between them. In this work, we define an edge as the link that joins two co-occurrent (adjacent) words in the text.

Figure 1: Network extracted from the phrase: “En un lugar de La Mancha.”

Figure 1 shows a simple word network constructed from a single sentence, extracted from the masterpiece of Miguel de Cervantes y Saavedra, “El ingenioso hidalgo don Quijote de la Mancha”. The sentence, “En un lugar de La Mancha,” contains six different words. Word $w_3$ = “lugar” has as input word $w_2$ = “un”, and as output word $w_4$ = “de”. Now, in the created graph $G$, the connectivity of $w_3$ is $k_3 = 2$. If the text were composed only of different words, as in this example, the network would be a simple lineal chain of $W$ different words and $W - 1$ links. In this type of network, except for the words located at the beginning and at the end of the text, all the words have connectivity $k_i = 2$. Nevertheless, this scenario is quite improbable in long texts because in natural languages, as mentioned above, words are used with different frequency. Thus, if we introduce a new sentence to the previous text, for example “de cuyo nombre no quiero acordarme, no a mucho tiempo que vivía un hidalgo de los de lanza en astillero”, there are both new and repeated words added. In our process of network generation, repeated words maintain the same number $w_i$ that corresponds to its first appearance in the text; however, that word can be subsequently connected to different words or connected many times to the same word (imagine a character name, composed of two words, that appears many times throughout the text, e.g., Don Quijote). The network generated after the incorporation of the new sentence is shown in Figure 2. Notice that the words “de”, “no”, “en” and “un” have a higher connectivity than the rest of the words.

It must be pointed out that the method proposed in this work is case-insensitive, eliminating capitalization effects. Furthermore, due to the fact that this work is concerned with analyzing grammar in language, it is assumed that (any) punctuation cuts any grammatical relationship between two words. For this reason, in the example above, the words “mancha” and “de”, and “acordarme” and “no”, are not connected.

III. RESULTS

In order to analyze the word networks generated, we used a set of metrics to characterize the whole system as well as local relationships between words. One such basic metric used in network topology characterization is degree distribution $P(k)$. This probability distribution represents, in this case, the probability of finding a word with $k$ edges in the network. Degree distribution is one of the most important characteristics of networks, especially due to the fact that the distribution of node connections is indicative of the underlying network formation mechanisms. In fact, random networks, whose graphs show randomly-chosen relationships between nodes, show homogeneous distributions of connectivity, whereas (so called) complex topologies display inhomogeneous distributions. This nonuniformity denotes the presence of rules (e.g., grammar rules) or mechanisms that distribute the links unequally, and where some few nodes may concentrate the bulk of connections.

Figure 3 shows this behavior consistent with complex topologies, as the distribution of the number of words with $k$ connections, $N \cdot P(k)$, for three classic books: “La
that the distributions of the three books follow a power law form, \( N \cdot P(k) \sim k^{-\gamma} \), where \( \gamma \) is the scaling exponent. This scaling, typically observed in complex systems \[24\], reveals a high level of inhomogeneity in the number of connections among words in the network. Thus, in networks constructed from natural languages, like the Spanish presented in these examples, most of the words have few connections, while there are a handful of words responsible for a large majority of connections (hubs). These hubs play an important role \[20\]. All the same, the most interesting property of these word networks, where the probability distribution of degree \( k \) asymptotically decays to the power of a constant \( \gamma \), is that they are scale-free.

For contrast, we randomized (i.e., located randomly \[32\]) the same words and frequencies from the same books and compared their random distributions to those obtained from the previously constructed networks. It can be observed that the distributions of these “randomized texts” are similar to those of the originals, which implies that inhomogeneous distribution of connectivity depends not on the way words are linked, but rather on the frequency of words in the text. The message is clear: degree distribution in word networks gives no information on the meaning of a text, it is but a projection of Zipf’s law.

For example, “La Ruta”, “El Ingenioso Hidalgo Don Quijote de la Mancha” \( W = 3895 \) by Miguel de Cervantes y Saavedra and “La Metafomosis” by Franz Kafka \( W = 3895 \). As can be seen, the distributions show word networks with a wide range of connectivities: many words are poorly connected, while a few are densely connected. Moreover, the log-log plots of the distributions tend towards a straight line, which indicates the asymptotic power law behavior.

Table I: Topological properties of The Universal Declaration of Human Rights (HR) written in different languages (ID=[1,17]), classic books (ID=[18,26]) and computer codes (ID=[27,36]). Number of different words \( W \), number of Edges \( E \), mean degree \( k \), mean clustering coefficient \( C \) and average path length \( l \) for original and randomized texts \((C_r)\), \((l_r)\).

<table>
<thead>
<tr>
<th>#</th>
<th>Description</th>
<th>Network</th>
<th>( W )</th>
<th>( E )</th>
<th>( k )</th>
<th>( C )</th>
<th>( l )</th>
<th>( C_r )</th>
<th>( l_r )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>HH</td>
<td>Dutch</td>
<td>494</td>
<td>852</td>
<td>1.99</td>
<td>0.154</td>
<td>0.055</td>
<td>2.90</td>
<td>3.76</td>
</tr>
<tr>
<td>2</td>
<td>HH</td>
<td>English</td>
<td>430</td>
<td>835</td>
<td>1.95</td>
<td>0.177</td>
<td>0.060</td>
<td>3.16</td>
<td>3.42</td>
</tr>
<tr>
<td>3</td>
<td>HH</td>
<td>Spanish</td>
<td>563</td>
<td>772</td>
<td>1.37</td>
<td>0.040</td>
<td>0.031</td>
<td>4.41</td>
<td>4.92</td>
</tr>
<tr>
<td>4</td>
<td>HH</td>
<td>German</td>
<td>517</td>
<td>892</td>
<td>1.73</td>
<td>0.135</td>
<td>0.060</td>
<td>3.66</td>
<td>4.05</td>
</tr>
<tr>
<td>5</td>
<td>HH</td>
<td>Greek</td>
<td>576</td>
<td>1035</td>
<td>1.80</td>
<td>0.109</td>
<td>0.076</td>
<td>3.04</td>
<td>3.95</td>
</tr>
<tr>
<td>6</td>
<td>HH</td>
<td>Italian</td>
<td>512</td>
<td>579</td>
<td>0.57</td>
<td>0.110</td>
<td>0.093</td>
<td>5.60</td>
<td>3.89</td>
</tr>
<tr>
<td>7</td>
<td>HH</td>
<td>Arabic</td>
<td>554</td>
<td>778</td>
<td>1.40</td>
<td>0.036</td>
<td>0.022</td>
<td>4.99</td>
<td>5.12</td>
</tr>
<tr>
<td>8</td>
<td>HH</td>
<td>Maori</td>
<td>342</td>
<td>1197</td>
<td>3.50</td>
<td>0.325</td>
<td>0.180</td>
<td>2.60</td>
<td>2.99</td>
</tr>
<tr>
<td>9</td>
<td>HH</td>
<td>Mapudungun</td>
<td>321</td>
<td>736</td>
<td>2.27</td>
<td>0.230</td>
<td>0.127</td>
<td>2.95</td>
<td>3.75</td>
</tr>
<tr>
<td>10</td>
<td>HH</td>
<td>Nahuatl</td>
<td>523</td>
<td>1181</td>
<td>2.15</td>
<td>0.190</td>
<td>0.096</td>
<td>5.25</td>
<td>3.61</td>
</tr>
<tr>
<td>11</td>
<td>HH</td>
<td>Portuguese</td>
<td>464</td>
<td>827</td>
<td>1.78</td>
<td>0.140</td>
<td>0.071</td>
<td>3.10</td>
<td>3.92</td>
</tr>
<tr>
<td>12</td>
<td>HH</td>
<td>Quechua</td>
<td>407</td>
<td>921</td>
<td>1.39</td>
<td>0.080</td>
<td>0.026</td>
<td>4.99</td>
<td>5.12</td>
</tr>
<tr>
<td>13</td>
<td>HH</td>
<td>Russian</td>
<td>586</td>
<td>971</td>
<td>1.66</td>
<td>0.091</td>
<td>0.037</td>
<td>3.73</td>
<td>4.18</td>
</tr>
<tr>
<td>14</td>
<td>HH</td>
<td>Romanian</td>
<td>401</td>
<td>657</td>
<td>1.43</td>
<td>0.070</td>
<td>0.043</td>
<td>4.02</td>
<td>4.63</td>
</tr>
<tr>
<td>15</td>
<td>HH</td>
<td>Spanish</td>
<td>486</td>
<td>915</td>
<td>1.88</td>
<td>0.183</td>
<td>0.090</td>
<td>3.29</td>
<td>3.76</td>
</tr>
<tr>
<td>16</td>
<td>HH</td>
<td>Tamil</td>
<td>432</td>
<td>1015</td>
<td>2.35</td>
<td>0.437</td>
<td>0.188</td>
<td>2.16</td>
<td>2.26</td>
</tr>
<tr>
<td>17</td>
<td>HH</td>
<td>Zeqit</td>
<td>562</td>
<td>670</td>
<td>1.08</td>
<td>0.015</td>
<td>0.009</td>
<td>7.33</td>
<td>7.67</td>
</tr>
</tbody>
</table>

Figure 3: Distribution of the number of words with \( k \) connections, \( N \cdot P(k) \), for the texts. Top: “La Ruta” \( W = 3895 \) by Homer. Mid: “El Ingenioso Hidalgo don Quijote de la Mancha” \( W = 3895 \) by Miguel de Cervantes y Saavedra. Bottom: “La Metamorfosis” by Franz Kafka \( W = 3895 \). Original text in black squares and randomized version in red squares.

To further characterize the word networks, we compared...
puted other classic topological measures of networks: the average path length \( \langle l \rangle \) between pairs of words; and the clustering coefficient of the network \( \langle C \rangle \), as the average coefficient of all the words in a network. Table I shows these metrics for different words networks constructed from The Universal Declaration of Human Rights written in different languages (ID=[1,17]), classic books (ID=[18,25]), the text of the “Voynich manuscript” (ID=26), and a set of computer codes written with formal grammatical rules in two different programming languages [33] (ID=[27,36]).

The top plot of Figure 4 shows that for most of the studied networks, except for the “Voynich manuscript” and programming codes (red and yellow squares, respectively), there is a much higher mean clustering coefficient than would be expected given only their randomized versions. However, their average path length values are practically the same as the ones observed in the randomized texts (Fig. 4, bottom plot). This behavior, also observed in other words networks [11, 16], is a property of small-world networks [30] and denotes high transitivity of word connections in the original word networks. It also shows that the well known small-world effect, associated with a short average path length, is not only a property of networks with complex topology, but also networks with random connections [10].

The clustering coefficient, then, would seem to be a possible first step in solving the sensical text classification problem. At the very least, focusing on the clustering coefficient might allow one to find differences between an original text and its randomized version; however, this is both insufficient and misleading. As Table I shows, the clustering coefficient varies widely among the same text written in different languages. For example, the mean clustering of The Universal Declaration of Human Rights, written in German, Spanish, or English, is on the order of the randomized text written in Mahori (0.18); or the Basque text to the randomized Russian Declaration (0.04).

That said, an interesting property was found in this study: a correlation between the mean clustering coefficient and the average path length that follows a potential function, \( \langle C \rangle \sim \langle l \rangle^{-3.56} \). This property, far from trivial, seems to be exclusively a property of texts, differing from correlations found in other complex networks.

In Figure 5 we can see that all the word networks follow the potential function, and their position on this fit depends on their mean degree \( \langle k \rangle \) (see Table I). Thus, texts with high \( \langle k \rangle \), are those with high clustering and low average path length. The text with the lowest mean degree, the Zulu Declaration, is the one furthest to the right of the curve (\( \langle C \rangle = 0.015 \) and \( \langle l \rangle = 7.33 \)). It is interesting to note that when these texts are randomized, they follow the same potential fit (see inset of the figure), but shift their positions to the right of the curve. This may be due to the way in which we made the randomization, but it does suggest that nonsensical texts can also satisfy this relation. It is also interesting to highlight that both formal programming languages (yellow squares) and the “Voynich manuscript” (red square) also fit this function. As to the exclusivity of this function to texts, other complex networks of a different nature in the figure (social networks: a, c and d [19,21]; e [30]; f and g [3], language networks: b [1], co-occurrence network: green square [1], and technological networks: blue stars [7,8], i [30]) do not display this behavior, save network b [19,21]. Moreover, it is noteworthy that the Barabási network [2] correlation...
Figure 5: Mean clustering coefficient \( \langle C \rangle \) and average path length \( \langle l \rangle \) correlation. Color code as in Fig. 4. Other networks: social networks (a, b, c, d \[19–21\]; e \[30\], f, g \[3\]), language network (h \[1\]), technological networks (blue stars \[7, 8\] and i \[30\]), co-occurrence network (green square \[1\]) and different networks generated by Barabási model (pink crosses).

(purple x) clearly shows a different behavior when compared to word networks. This result suggests that the preferential attachment mechanism is not valid for text construction, as otherwise suggests. In that model, increasing the number of nodes (and therefore the links between “words”) shifts the position in the graph to the right (higher values of \( \langle l \rangle \), and lower of \( \langle C \rangle \)). This is the opposite of what happens in the case of word networks. The networks generally move to the right in the graph when the number of links decreases (actually when \( \langle k \rangle \) decreases, but those texts with smaller \( E \) are the same as those with lowest mean degree \( \langle k \rangle \), as shown in the Table \[1\]).

We have therefore found a property that can help discern whether a complex network comes from a text or not. In other words, in the presence of a network generated by linking words from a text based on co-occurrence, that network should necessarily be positioned on the fit. However, this method still does not solve the deeper problem that motivates this work: the search for properties of networks of sensical texts. Randomized texts also adjust to the function, not to mention texts written for machines and the “Voynich manuscript”. Correlation between clustering and average path length is a necessary condition, but still insufficient.

Notwithstanding the inability of that metric alone in describing sensical texts, we found another property that can help us to solve the puzzle: the network assortativity \[22\]. This metric corresponds to the correlation between node properties, like their degree, for example, and seems to be closely related with determining if a text is sensical (or nonsensical). For network assortativity, in the scenario of symmetric connections (undirected network) like the ones studied in this work, if densely connected nodes are connected to other nodes with many connections, then the network is considered assortative, \( r > 0 \). On the other hand, if densely connected nodes are connected with their poorly connected counterparts (or vice versa), then the network is disassortative, \( r < 0 \). If no degree correlation is observed, \( r \sim 0 \), then there is no link preference between nodes, as is the case in random networks or networks generated by the Barabási model \[22\].

Word networks are expected to be disassortative \[13\], since highly connected words, such as articles, are linked with others that appear far less in the text, like nouns. It is necessary to emphasize that networks of the same type as the network b in Figure 5 (the only non-text network that fit the correlation between \( \langle C \rangle \) and \( \langle l \rangle \)) are assortative \[22\]. Figure 6 shows the disassortative character of words networks.

Figure 6: Text (ID) vs Assortativity \( r \). Black crosses represent original texts, red crosses randomized texts and green bars the difference. Color regions: The Universal Declaration of Human Rights (grey), classic books (blue), “Voynich manuscript” (orange), computer codes (yellow).

In the figure, we can appreciate the assortativity values for word networks constructed from texts of Table \[1\] (black crosses), and the mean assortativity of networks constructed from the same texts, randomized 100 times (red crosses). As can be observed, original text network assortativities are negative (i.e., words networks are disassortative). However, the value for this correlation is much higher (and tends to 0) when the texts are random-
ized. An interesting result appears, again, in the analysis of the “Voynich manuscript” (ID=26). It is the only analyzed text that does not present a significant difference between assortativities before and after randomization. One book, “Harry Potter and the Philosopher’s Stone” (ID=23), also displays a smaller change. However, this text also shows other particular properties, possibly related to the target audience to which it was directed; in fact, this text had the highest value of mean degree (see table [1]) and contains approximately 2000 words less than its version in Spanish (ID=22). To check the statistical validity of the change in the assortativity, we calculated the standard score (Z-score) \[1\] according to,

\[
Z_{score} = \frac{r(ID) - \langle r_R(ID) \rangle}{\sigma_R}
\]

where \( r(ID) \) is the assortativity of the text ID (see table [1]), \( \langle r_R(ID) \rangle \) and \( \sigma_R \) the mean and standard deviation of its randomized versions, respectively.

Figure 7: Text (ID) negative Z-score Assortativity. Color regions as in Fig. 6

Significantly, and in all cases, the assortativity values of original text networks are below that of their randomized versions, and particularly so for books.

Although the word networks assortativity now allows us to discriminate between sensical texts (albeit for human or machine) and nonsensical (or randomized) texts, it does not say anything by itself due to the wide range of values observed in all the sensical texts evaluated.

The different Z-score ranges shown in Figure 7 are explained by the fact that the texts have different lengths: the statistical models begin to fail when the text has fewer edges, since there are not as many possible variations. However, when a text has many edges, each successive realization in the process described herein gives a very different network, and eventually allows for significant observations of how far a corpus is from a sensical text value. Confirming this, there is a positive correlation between Z-score and number of edges, as shown in Figure 8.

In order to solve this problem with classifying texts as sensical using only assortativity values, we found another correlation that neither an encoded text nor texts written in formal languages seem to follow. The correlation between assortativity and mean degree of sensical texts
written in natural languages follows a log-normal function, while any randomized or encoded version moves away from this function (Fig. 9). This function permits us, then, to find a specific assortativity value for a certain mean degree for which a text is more likely to make sense. Of note, text ID=23, which has the highest mean degree, also fits the function, while the “Voynich manuscript” does not. The anomalous position of the manuscript in Figures 8 and 9, taken with its anomalous behaviors with respect to clustering and assortativity between its original and randomized versions, suggest that the manuscript is an encoded or ciphered text, even with a lower disassortativity. We use the term encoded because the manuscript has a word frequency distribution that follows Zipf’s law, and because the correlation between $\langle C \rangle$ and $\langle l \rangle$ also fits the power function of other texts (see Fig. 5).

Finally, this log-normal function that the correlation between assortativity $r$ and mean degree $\langle k \rangle$ fits, as well as Zipf’s law, seems to be exclusive to texts written in natural languages. In fact, all the computer codes included in this study are outside this regularity, and even more so if they were randomized (Fig. 10).

Figure 10: Degree assortativity $r$ and mean degree $\langle k \rangle$ correlation for texts ID=[27,36]. Color code as in Fig. 9.

IV. CONCLUSIONS

In this work we searched for a method to discern between sensical texts and nonsensical/encoded texts or texts written using formal grammar. Our approach was to analyze the topological properties of co-occurrence word networks.

Our results suggest that a set of metrics related to network assortativity are able to solve this classification, as long as the word network has passed a series of pre-filters. This means that if a network has been constructed from a sensical text, irrespective of its origin, we should expect that the network displays a “proper” position in the space $\langle C \rangle$ vs $\langle l \rangle$, is disassortative, and that its correlation between assortativity and mean degree fits a log-normal function. However, if a word network has been constructed from a sensical text written in a formal programming language, then the last condition will most likely not be fulfilled.

This allows us to speculate on a desirable sorting method for the books in Library of Babel. It might be possible to separate sensical texts from those that no one could understand using the following method: (i) calculate four macroscopic statistical properties of word networks: mean degree, mean clustering coefficient, average path length, and network assortativity; (ii) check for the position of the word networks in plane $\langle C \rangle$ vs $\langle l \rangle$. If the network does not fit the power function of Figure 5 it is a network that is not a text, and therefore should be discarded. If the text fits the function, the network has a text structure, but is not necessarily sensical, then; (iii) two alternatives are possible: (a) order these word networks by mean degree, and then from lowest to highest values of disassortativity. Those with higher values of disassortativity have a higher probability of being sensical texts. (b) Check for the position of these networks in plane $r$ vs $\langle k \rangle$ of Figure 9. Those networks that are closer to the function, are more readable (i.e., make more sense) than versions located away from the fit line. Networks located far away from the fit are texts that might correspond to a ciphered text or texts written for a machine.

The result of our study leads us to propose the following hypothesis: for all variations of a text (with the same number and frequency of words), the most disassortative version will have the highest probability of making sense to a reader. Hence, this method could help in deciding when a decoded version of a text is correct or not, without the necessity of “translating” the encoded text. Simply put, the results from a decoding process will be more effective than another if the resulting word network has a higher disassortativity.

Finally, our study has also allowed us to find evidence supporting the thesis that the “Voynich manuscript” is a written text which has been ciphered.

This work opens up interesting questions that will be addressed in future works.
19. M. E. J. Newman. 2001c. Random text is composed by new edges ei between co-occurrent words, ei = (wi, w′ i), where i = [1, ..., E] and wi and w′ i are the words that appear adjacent in the original text. To randomize the text, we maintain fixed column wj while the column w′ j is moved R places, where R is a random number. Thus, the randomized text is composed by new edges eij = (wij, w′ j+i), where wE+i = w′ i. These random texts are different to the ones described in [12].