Mesoscopic representation of texts as complex networks

Henrique Ferraz de Arruda, Vanessa Queiroz Marinho, and Diego Raphael Amancio
Institute of Mathematics and Computer Science,
University of São Paulo, São Carlos, SP, Brazil.

Filipi Nascimento Silva and Luciano da Fontoura Costa
São Carlos Institute of Physics, University of São Paulo, São Carlos, SP, Brazil

I. INTRODUCTION

The availability of an ever growing amount of complex data brought up by the age of information has strongly impacted science, giving rise to a novel perspective on data analysis. The use and development of systematic approaches to analyze complex data has already become mandatory for many problems in a wide range of knowledge areas, such as physics [1], biology [2, 3], medicine [4] and even humanities [5, 6]. This also includes techniques devoted to the systematic analysis of texts, known as text mining [7]. Traditionally, processes involving text analytics were solely based on simple statistics considering mostly the frequency of words occurrence [8, 9], which are, in general, suitable for the task of text classification [10]. However, more sophisticated methods have been devised for more complex tasks, such as to quantify the words relevance [11, 12] in a document. Techniques can be employed to detect, for instance, important topics in a given text [13, 14]. Even more complex are the methods used to study the relationships among words or topics in a document or a set of documents. Such a kind of analysis can be undertaken by considering semantic similarities [15] or linguistic characteristics [16]. By using these new techniques, many other applications could be performed, e.g., automatic summarization [17], sentiment analysis [18] or authorship detection [19]. Apart from the mentioned approaches, texts datasets can be analyzed in terms of the relationships among their elements, such as words and paragraphs. In this context, texts can be regarded as a complex structure, rendering complex networks a suitable representation for such structures.

Complex networks are a discrete representation of interactions among elements in a complex system. In such a representation, elements are referred to as nodes and pairwise interactions are represented via edges. A well-known approach to construct complex networks from a text dataset is the word-adjacency (or co-occurrence) technique [20, 21], which is based on connecting pairs of words that are immediately adjacent. Such a strategy of mapping texts into networks is a simplification of networks formed by syntactical links [22]. Despite this seeming limitation, word adjacency networks have been employed successfully to address a great variety of natural language processing problems. This includes sentiment analysis [23], authorship detection [24, 26], stylometry [27], text classification [28], word sense disambiguation [29, 30], text summarization [32, 33], machine translation [34, 35] and others.

Previous works employing the co-occurrence network representation of texts have pointed out some of its limitations. Perhaps the most critical disadvantage associated with such a representation is its inability to portray the topical structure present in many kinds of texts. At a glance, the topical structure of a text is expected to naturally emerge from its network representation through a pronounced community structure, however, this hardly ever happens on typical co-occurrence networks, which...
present no community structure at all [36]. This suggests that this representation does not effectively capture the information lying in the mesoscopic structure of the text, such as topics and subtopics. Besides, in co-occurrence networks, the information regarding the temporal evolution along a text is also overlooked, as the emphasis of such models is to analyze language/stylistic dependent features.

The rest of the paper is organized as follows: Section II describes in details our approach to create the mesoscopic network from a given document. Section III provides a case study of our approach with the analysis of a book. Finally, Section IV shows a synthesis of our paper and the perspectives for further studies.

II. METHODOLOGY

This section describes, in details, the procedure to obtain such mesoscopic complex networks from sets of organized text, which includes books and other classes of text with paragraph structure. Here, we also briefly present the technique employed to construct visualizations for such networks.

In recent years, a new set of techniques has been introduced to form networks from unstructured documents, which takes into consideration its mesoscopic structure [36]. Such networks are generated by connecting words existing in the same context, which is defined in terms of fixed window length. This approach was found to be able to produce networks with strong community structures, with each community corresponding to the main contextual topics or subtopics of the text [36]. Even though the semantical organization of texts is captured in such models, they are unable to represent the temporal evolution of the story being told.

Here, we extend the concepts introduced in [36] to derive a new technique to construct networks from organized texts. Our methodology accounts for the two main characteristics overlooked by the traditional approaches: (a) the mesoscopic structure and (b) its evolution along a text. To consider (a), unlike traditional models linking adjacent words, we use larger pieces of texts as the simplest text unit, which are connected according to the similarity among themselves. The temporal evolution of ideas and concepts in (b) is recovered from our model because, by construction, nodes (i.e. adjacent pieces of texts) are always connected.

In the following, we consider an organized text as a sequence of words segmented by paragraphs. In our analysis, paragraphs can be innate from the source material, such as the paragraphs in a book, or can be inferred from the text own structure, for instance, by considering each paragraph as a sequence containing a fixed number of words.

Our approach starts with the typical pre-processing step employed for semantical-based text analysis. First, punctuation marks and numbers are removed. We also remove words conveying low contextual meaning, i.e. the stopwords. Examples of stopwords are articles and prepositions. If a lemmatization technique [7] is available for the language being considered, it is used to normalize concepts. In this step, words are reduced to their canonical forms, so that inflections in verbal tense, number, case or gender are disregarded. For example, the sentence “‘Oh, I’ve had such a curious dream!’ said Alice” becomes “curious dream say alice”, after being preprocessed. Next, we employ the tf-idf (term frequency-inverse document frequency) technique [2], which defines a map tf-idf(w, d, D) quantifying the importance of each word w in a given document d from a set of documents D. The tf-idf(w, d, D) map is computed as

$$\text{tf-idf}(w, d, D) = \text{tf}(w, d) \times \text{idf}(w, D),$$

(1)

where tf(w, d), the term-frequency component, accounts for the relevance of w ∈ d and idf(w, D), the inverse document frequency, quantifies the frequency of w in all d ∈ D. Many variations of both tf and idf terms have been proposed [7]. In this paper, we consider tf(w, d) as the raw frequency of a given word w in a document d. The idf(w, D) was calculated as

$$\text{idf}(w, D) = 1 + \log \left( \frac{|D|}{f_w} \right),$$

(2)

where |D| is the total number of documents in D and fw is the frequency of occurrence for a given word w in D, i.e. fw is the number of documents in which w occurs at least once. Note that, when fw = |D|, log(|D|/fw) = 0. If the additional term were not used in equation (2) the value of tf-idf would always be zero, despite of any raw-frequency observed in the current document (tf term).

The mesoscopic network is generated from the preprocessed text, hereafter referred to as organized text O. The organized text O consists of a sequence of paragraphs \(O = (p_0, p_1, p_2, \ldots)\) with each paragraph \(p_i\) comprising a sequence of words \(p_i = (w_{i0}, w_{i1}, w_{i2}, \ldots)\). Differently from the co-occurrence model, where nodes represent words, here, we model entire paragraphs or sequences of consecutive paragraphs by nodes. In particular, for a choice of window size \(\Delta\), each possible subsequence comprising \(\Delta\) paragraphs in O, \(P_k^\Delta = (p_k, p_{k+1}, \ldots, p_{k+\Delta-1})\), is given a node representation in the devised mesoscopic network. Fig. 1(a) illustrates the process to obtain the nodes of the mesoscopic network.

The edges of the mesoscopic network are devised considering a contextual similarity measurement obtained among all pairs of the sequences of paragraphs \(P_k^\Delta\) in the investigated document. Here, we employed the traditional bag of words combined with the cosine similarity measurement [7]. Bearing in mind that the number of words in paragraphs can vary significantly along the investigated document, the choice of cosine similarity was due it not depending on the length of the chunks of text being compared [27]. First, for each considered sequence of paragraphs \(P\), a vector \(W_P\), spanning the
FIG. 1: Illustration of the presented methodology. Initially, the text is organized in sets of subsequent and overlapping windows $P_k^3$, each containing 3 structural paragraphs, as shown in (a). Next, the cosine similarity is calculated among all pairs of text windows (illustrated by the width of the lines in b). The mesoscopic network is obtained by maintaining only connections among pairs of windows with similarity higher than a threshold value $T$. This is illustrated by the network visualization shown in (c).

The same number of words present in $O$, is obtained from the tf-idf map $\text{tf-idf}(w, P, O)$ applied to each word $w$ in $O$. Note that, when a certain word $w$ is not present in $P$, $\text{tf-idf}(w, P, O) = 0$. The content similarity measurement $S(P_A, P_B)$ between two paragraph windows $P_A$ and $P_B$ is obtained using

$$S(P_A, P_B) = \frac{\sum_{w \in O} \text{tf-idf}(w, P_A, O) \times \text{tf-idf}(w, P_B, O)}{\sqrt{\sum_{w \in O} \text{tf-idf}(w, P_A, O)^2} \sqrt{\sum_{w \in O} \text{tf-idf}(w, P_B, O)^2}}.$$  

(3)

As a result, a fully connected network is created (see Fig. 1(b)), where edge weights correspond to the similarity $S(P_A, P_B)$ among each pair of nodes. The final mesoscopic network is obtained by pruning the weakest connections, i.e. the links whose weight takes a value below a given threshold $T$. After such a procedure, edge weights are ignored, resulting in a unweighted network (see Fig. 1(c)).

To better understand the overall structure of mesoscopic networks, we visualized the network structure.
using a technique based on force-directed nodes placement. In particular, we use a technique based on the Fruchterman-Reingold (FR) [37] algorithm, in which the network is regarded as a system of nodes behaving like particles that interact among themselves by the action of two types of forces: attractive forces, existing only between connected nodes, and repulsive forces, that exist between all pairs of nodes. By minimizing the energy of such a system, the network organizes itself in a graphically appealing layout. Such a visualization technique naturally highlights many aspects of the topological structure of networks [37].

To illustrate the proposed methodology, we investigated the obtained properties from the mesoscopic network modelling a well-known book: “Alice’s Adventures in Wonderland”. This story revolves around the adventures of a little girl, called Alice, after she falls in a hole and arrives in a unknown and fantasy world. The book was written in 1865 by Charles Lutwidge Dodgson under the pseudonym Lewis Carroll. It is divided into the following twelve chapters:

1. Down the Rabbit-Hole
2. The Pool of Tears
3. A Caucus-Race and a Long Tale
4. The Rabbit Sends in a Little Bill
5. Advice from a Caterpillar
6. Pig and Pepper
7. A Mad Tea-Party
8. The Queen’s Croquet-Ground
9. The Mock Turtle’s Story
10. The Lobster Quadrille
11. Who Stole the Tarts?
12. Alice’s Evidence

III. RESULTS AND DISCUSSION

Two mesoscopic network, \( G_1 \) and \( G_2 \), were constructed from the book under investigation by applying the proposed methodology, for which two distinct thresholds, respectively, \( T_1 = 0.46 \) and \( T_2 = 0.24 \) were considered. For such a process, we have chosen a fixed window size \( \Delta = 20 \).

We start the analysis of the obtained mesoscopic structures by initially investigating the properties of the \( G_1 \) network, which is simpler. In the following, the analysis is undertaken considering a 2D visualization of the \( G_1 \) network, which is shown in Figure 2 and was obtained by employing the FR algorithm. Since nodes sharing the same paragraphs become strongly connected among themselves, a pronounced chain-like structure naturally emerges on the devised mesoscopic network. In addition, such a structure is aligned to the ordering of the nodes along the book. This property is better observed in Figure 2(a), where the color of each node indicates its position along the text. Nevertheless, the existence of connections among distant nodes along the text indicates regions of high contextual similarity that are not a result of overlapping sequences of paragraphs. In the obtained visualization, such a structure creates bridges of connections among contextually similar regions of nodes, which in turn brings them more closely together along the chain-like structure of the network.

To better understand the relationship between the contextual information of the book and its mesoscopic structure, we used the chapter organization of the book to segment the obtained network. Such aspect is visualized in Figure 2(b), where the chapter of each node is indicated by a color according to the legend. By taking into consideration the connectivity among the chapters of the book, derived the following observations:

- In chapter 1, we note that there is no strong connection among its paragraphs and those from other chapters, except for chapter 2, which is explained by the aforementioned overlap between subsequent paragraphs. The lack of long range connections among the nodes of chapter 1 may happen because the main subject of this chapter is substantially different from every other in the book. In this chapter, the story unfolds in a realistic scenario and has no descriptions of the fantasy locations and creatures found in the rest of the book, except for the Rabbit;

- Chapters 2, 3, and the beginning of chapter 4 are connected among themselves. This may be a consequence of the fact that all these chapters describe the period of the story when Alice, the protagonist, was very frightened about the world she just jumped in. In addition, all these chapters mentioned when she cried and formed a pool of tears;

- In chapter 5, there are strong connections between regions lying in the same chapter. This probably happens because there is an long conversation between Alice and the Caterpillar, in which they discuss the many sizes she had during the previous chapters;

- The connection between chapter 7 and chapter 11 can be related to the character The Hatter, who is drinking tea in both chapters. Furthermore, in both situations, he talked about some specific kinds of food, which are related to the tea party, eg. bread and butter;

- There is a group of highly connected nodes in the end of chapter 9 and in the beginning of chapter 10. This probably happens because in the final paragraphs of chapter 9, Alice met The Mock Turtle and their talk ended only in chapter 10.
Similarly to Figure 2, Figure 3 displays a visualization of the $G_2$ network, which was constructed using a lower threshold value $T_2 = 0.24$. By varying this threshold, we illustrate the potential of our method in describing the characteristics of the network in a multi-scale fashion. From Figure 3(a), we can observe that the network maintains a chain-like structure similar to that found in the $G_1$ network. However, such a network presents much more connections among nodes representing paragraphs belonging to different parts of the book. This is a consequence of the $G_2$ network capturing more fine-grained details about the relationships among the paragraphs. By a closer look at the differences between $G_1$ and $G_2$, we note, for instance, that, while chapter 1 is connected only to chapter 2 in $G_1$, in $G_2$ it becomes connected with many other parts of the book, in particular, with chapters 2, 3, 4, 6, 7, and 8. The analysis of more fine-grained networks, however, may present some disadvantages. Since it tends to incorporate much more local characteristics, it may include noise and relationships not driven by a strong contextual content.

IV. CONCLUSION

In order to investigate the mesoscopic properties of texts modelled as networks, we proposed a novel approach that considers the semantical similarity between textual segments. Distinctly from previous studies, we modeled sequences of adjacent paragraphs as nodes in a network, whose links are established by content similarity. By doing so, we could capture two important features present in written texts: long-range correlations and the temporal evolution of the story. The main feature of our approach concerns its ability to analyze documents in a multi-scale fashion. Specifically, two parameters control the scale: (i) $\Delta$: the number of paragraphs in each window, and (ii) $T$: the threshold used to prune connections among nodes with low contextual similarity.

We tested our approach in a well-known book, “Alice’s Adventures in Wonderland”, by employing network visualization techniques on the generated mesoscopic networks. Many insights can be drawn from the visualizations by tracing a parallel between its structure and the story presented in the book. In particular, we investigated the correspondence between the content of each chapter and underlying network structure arising from the proposed model.

Our model uncovered many relationships among different contexts sharing the same topics, such as similar characters or places appearing throughout the story. For example, the high contextual similarity found between chapters 7 and 11 can be explained by the fact that both chapters share a recurrent subject revolving around the character “The Hatter” approaching the tea party thematic. Note that similar textual inferences could not be drawn from models solely based on local features, as it is the case of traditional word-adjacency or syntactical networks, as they emphasize only stylistic textual subtleties.

The proposed network representation paves the way for developing new techniques that could be applied to automatically analyze the mesoscopic structure of documents. These techniques could be used to improve traditional approaches used to tackle typical text mining problems under a new perspective. This capability should be further explored in future works, for instance, by measuring the efficiency of our model in text classification, summarization and similar applications where an accurate semantic analysis plays a prominent role in the characterization of written texts.

Acknowledgements

The authors acknowledge financial support from Capes-Brazil, São Paulo Research Foundation (FAPESP) (grant no. 2015/08003-4, 2015/05676-8, 2014/20830-0 and 2011/50761-2), CNPq-Brazil (grant no. 307333/2013-2) and NAP-PRP-USP.
FIG. 2: Visualization of the network $G_1$, representing the book Alice’s Adventures in Wonderland with a threshold $T_1 = 0.41$. Each node indicates a sequence of paragraphs. The order of the network nodes according to the story are shown in (a). Note that the first nodes of the story appears in blue, while the last nodes are represented in a reddish color. In (b), the chapters to which paragraphs belong are represented with distinct colors.

FIG. 3: Visualization of the network $G_1$, representing the book Alice’s Adventures in Wonderland with a threshold $T_1 = 0.24$. Each node indicates a sequence of paragraphs. The order of the network nodes according to the story are shown in (a). Note that the first nodes of the story appears in blue, while the last nodes are represented in a reddish color. In (b), the chapters to which paragraphs belong are represented with distinct colors.
[1] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D.-U. Hwang, Physics reports **424**, 175 (2006).

[2] A.-L. Barabasi and Z. N. Oltvai, Nature reviews genetics **5**, 101 (2004).

[3] H. F. de Arruda, C. H. Comin, M. Miazaki, M. P. Viana, and L. da Fontoura Costa, Journal of neuroscience methods **245**, 1 (2015).

[4] A.-L. Barabási, N. Gulbahce, and J. Loscalzo, Nature Reviews Genetics **12**, 56 (2011).

[5] Y. Moreno, M. Nekovee, and A. F. Pacheco, Physical Review E **69**, 066130 (2004).

[6] M. Kalimeri, V. Constantoudis, C. Papadimitriou, K. Karamanos, F. K. Diakonos, and H. Papageorgiou, Journal of Quantitative Linguistics **22**, 101 (2015).

[7] C. D. Manning and H. Schütze, *Foundations of Statistical Natural Language Processing* (MIT Press, Cambridge, MA, USA, 1999), ISBN 0-262-13360-1.

[8] U. Y. Nahm and R. J. Mooney, in *AAAI 2002 Spring Symposium on Mining Answers from Texts and Knowledge Bases* (2002), vol. 1.

[9] E. G. Altmann, J. B. Pierrehumbert, and A. E. Motter, PLoS One **4**, e7678 (2009).

[10] T. Joachims, in *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval* (ACM, 2001), pp. 128–136.

[11] J. Ramos, in *Proceedings of the first instructional conference on machine learning* (2003).

[12] A. Hotho, A. Nürnberger, and G. Paas, in *Ldv Forum* (2005), vol. 20, pp. 19–62.

[13] D. M. Blei, A. Y. Ng, and M. I. Jordan, the Journal of machine Learning research **3**, 993 (2003).

[14] L. AlSumait, D. Barbará, and C. Domeniconi, in *Data Mining, 2008. ICDM’08. Eighth IEEE International Conference on* (IEEE, 2008), pp. 3–12.

[15] T. K. Landauer, P. W. Foltz, and D. Laham, Discourse processes **25**, 259 (1998).

[16] V. Hatzivassiloglou, L. Gravano, and A. Maganti, in *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval* (ACM, 2000), pp. 224–231.

[17] Y.-L. Chang and J.-T. Chien, in *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on* (IEEE, 2009), pp. 1689–1692.

[18] A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts, in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1* (Association for Computational Linguistics, 2011), pp. 142–150.

[19] X. Chen, P. Hao, R. Chandramouli, and K. Subbalakshmi, in *Machine Learning and Data Mining in Pattern Recognition* (Springer, 2011), pp. 375–386.

[20] D. R. Amancio, O. N. Oliveira Jr., and L. F. Costa, Physica A **391**, 4406 (2012).

[21] A. Kulig, S. Drożdż, J. Kwapień, and P. Oświęcimka, Phys. Rev. E **91**, 032810 (2015).

[22] R. Ferrer i Cancho, R. V. Solé, and R. Köhler, Phys. Rev. E **69**, 051915 (2004).

[23] R. Feldman, Commun. ACM **56**, 82 (2013).

[24] A. Mehr, A. H. Darooneh, and A. Shariati, Physica A **391**, 2429 (2012).

[25] S. Segarra, M. Eisen, and A. Ribeiro, IEEE Transactions on Signal Processing **63**, 5464 (2015).

[26] D. R. Amancio, Journal of Statistical Mechanics: Theory and Experiment **2015**, P03005 (2015).

[27] D. R. Amancio, PLoS ONE **10**, e0136076 (2015).

[28] H. F. de Arruda, L. d. F. Costa, and D. R. Amancio, EPL (Europhysics Letters) **113**, 28007 (2016).

[29] R. Mihalcea, P. Tarau, and E. Figa, in *Proceedings of the 20th International Conference on Computational Linguistics* (Association for Computational Linguistics, Stroudsburg, PA, USA, 2004), COLING ’04.

[30] T. C. Silva and D. R. Amancio, EPL (Europhysics Letters) **98**, 58001 (2012).

[31] D. R. Amancio, O. N. Oliveira Jr., and L. da F. Costa, EPL (Europhysics Letters) **98**, 18002 (2012).

[32] L. Antiqueira, O. N. Oliveira Jr., L. F. Costa, and M. G. V. Nunes, Information Sciences **179**, 584 (2009).

[33] D. R. Amancio, M. G. V. Nunes, O. N. Oliveira Jr., and L. F. Costa, Physica A **391**, 1855 (2012).

[34] Q. Xuan and T.-J. Wu, Phys. Rev. E **80**, 026103 (2009).

[35] D. R. Amancio, L. Antiqueira, T. A. S. Pardo, L. F. Costa, O. N. Oliveira Jr., and M. G. V. Nunes, International Journal of Modern Physics C **19**, 583 (2008).

[36] H. F. de Arruda, L. d. F. Costa, and D. R. Amancio, Chaos **26**, 063120 (2016).

[37] T. Fruchterman and E. Reingold, Software-Practice & Experience **21**, 1129 (1991).