Text and topology in human interaction networks

Renato Fabbri[1, a] and Osvaldo Novais de Oliveira Junior[b]
Instituto de Física de São Carlos, Universidade de São Paulo (IFSC/USP)

(Dated: 22 February 2017)

This paper explores textual production in interaction networks and its relation to topological metrics. Texts from the email messages were grouped by source: peripheral, intermediary and hub sectors. Correlation of textual and topological measures were observed for the entire network and for each os such sectors. The formation of principal components is used for further insights of how measures are related. Network sectors presented discrepant linguistic features and each principal component exhibit predominance of textual or topological measures. Textual differences, correlation and principal components corroborate the stability of such interaction networks reported in previous works.

PACS numbers: 89.75.Fb,05.65.+b,89.65.-s
Keywords: statistical physics, complex networks, text mining, pattern recognition, social network analysis

I. INTRODUCTION

Sentiment analysis and vocabulary compilation are among a number of examples of the considerable attention that textual production has received from the social network analysis research community[?] . The relation of topological and textual metrics is the subject of this article for the following reasons:

- This relation has been set aside in literature, with scattered and vague suggestions of mutual implications of the text produced and topological characteristics of the agents in the network[?] .
- The results ease understanding of human interaction, which is useful for the observation of personality and cultural types[?] .
- There are hypothesis about verbal differentiation of network sections, derived from a previous article by the same authors[?] , some of which are herein confirmed.

Next section exposes materials used for this research, its textual and network aspects. Section III explains the analysis framework including. Section IV is dedicated to detailing results and discussion. Section V holds concluding remarks and further work envisioned. The Supporting Information presents further tables and figures.

II. MATERIALS

Email list messages were obtained from the public Gmane email archive[?] , which consists of more than 20,000 email lists and more than 130,000,000 messages[?] . These lists cover a variety of topics, mostly technology-related. The archive can be described as a corpus with metadata of its messages, including sent time, sender name and email address. The Gmane usage in scientific research is reported in studies of isolated lists and of lexical innovations[? ?] .

Typos, *leetspeak*, slang and invented words pose some challenges to current analysis which influenced the methodology to employ numerous metrics for the texts. Future work might bring these entries to forefront as neologisms and other linguistic innovations.

A. Availability

Data and scripts needed to derive results, figures, tables and this article itself are publicly available. Email messages are downloadable from the Gmane public database[?] . Computer scripts are delivered through a public domain Python PyPI package and an open Git repository[?] . This open approach to both data and scripts reinforces the scientific aspect of the contribution[?] and mitigates ethical and moral issues of researching systems constituted of human individuals[? ?] .

III. METHODOLOGY

A. Network formation, topological measures and Erdös sectioning

Figure 2 is illustrative of the formation of interaction networks. Avoiding identical repetition of content, Please refer to[?] for:

- further details on network formation.
- A concise consideration of the basic topological measures of vertex i: degree k_i, in-degree k_i^{in}, out-degree k_i^{out} strength s_i, in-strength s_i^{in}, out-strength s_i^{out}, betweenness centrality bt_i, clustering coefficient cc_i.

[a] Electronic mail: fabbri@usp.br
[b] www.polimeros.ifsc.usp.br/professors/professor.php?id=4; Electronic mail: chu@ifsc.usp.br; Also at IFSC-USP

- A specification of the symmetry measures for a vertex i: asymmetry asy_i, mean of asymmetry of edges μ_i^{asy}, standard deviation of asymmetry of edges σ_I^{asy}, disequilibrium dis_i, mean of disequilibrium of edges μ_i^{dis}, standard deviation of disequilibrium of edges σ_i^{dis}.

- The sectorialization of the real network in the periphery, intermediary and hub sectors through a comparison of the real network against an Erdös-Rényi model with the same number of vertices and edges.

Such sectorialization of the network is herein called "Erdös sectioning" and performed with degree k_i unless stated otherwise.

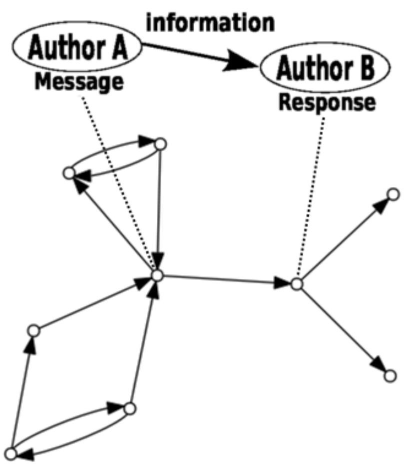

FIG. 1. The formation of interaction networks from email messages. Each vertex represents a participant. A reply message from B to a message from A is regarded as evidence that B has received information from A and yields a directed edge. Multiple messages add "weight" to a directed edge. Further details are given in[?] .

B. Textual measures

This work focuses on very simple metrics derived from texts as they were sufficient for current step:

- Frequency of characters: letters, vowels, punctuations and uppercase letters.

- Number of tokens, frequency of punctuations among tokens, frequecy of known words, frequency of words that has wordnet synsets, frequency of tokens that are stopwords.

- Mean and standard deviation for word, token sentence and message sizes.

- Fraction of morphosyntactic classes, such as adverbs, adjectives and nouns, represented by POS (Part-Of-Speech) tags.

- Fraction of words in each wordnet[?] top-most hypernyms, such as abstraction and physical entities for nouns or act for verbs.

To such measures are dedicated Tables **??**, **??**, **??**, **??**, **??**, **??**.

This choice is based on: 1) the lack of such information in literature, as far as authors know; 2) potential relations of these incidences with topological aspects, such as connectivity; 3) the interdependence of textual artifacts suggests that simple measures should reflect complex behaviors and more subtle aspects. A preliminary study, with the complete works from the Brazilian writer Machado de Assis[?] , made clear that these metrics vary with respect to style.

C. Relating text and topology

The topological and textual measures were related by:

1. incidences of linguistic traces in hub, intermediary and peripheral network sectors, which are delimited by topological criteria.

2. Correlation of measures of each vertex, easing pattern detection involving topology of interaction and language.

3. Principal components formation derived from usual Principal Components Analysis.

An adaptation of the Kolmogorov-Smirnov test was used to observe differences in textual content, as follows. Be $F_{1,n}$ and $F_{2,n'}$ two empirical distribution functions, where n and n' are the number of observations on each sample. The two-sample Kolmogorov-Smirnov test rejects the null hypothesis if:

$$D_{n,n'} > c(\alpha)\sqrt{\frac{n+n'}{nn'}} \tag{1}$$

where $D_{n,n'} = sup_x[F_{1,n} - F_{2,n'}]$ and $c(\alpha)$ is related to the significance level α by:

α	0.1	0.05	0.025	0.01	0.005	0.001
$c(\alpha)$	1.22	1.36	1.48	1.63	1.73	1.95

We need to compare empirical distribution functions, therefore $D_{n,n'}$ is given, as are n and n'. All terms in the equation 1 are positive and $c(\alpha)$ can be isolated:

$$c(\alpha) < \frac{D_{n,n'}}{\sqrt{\frac{n+n'}{nn'}}} = c' \tag{2}$$

When c' is high, low values of α favor rejecting the null hypothesis. For example, when c' is greater than ≈ 1.7, one might assume that $F_{1,n}$ and $F_{2,n'}$ differ. More importantly for us is that c' can be taken a measure of how much the distributions differ[?] . We use collections of such values for deriving hypotheses about how different are the underlying mechanisms of generation of texts.

IV. RESULTS AND DISCUSSION

The most important result in this article is the extreme differentiation of each Erdös sector with respect to the texts produced. For example: hubs use more contractions, more adjectives, more common words, and less punctuation if compared to the rest of the network, specially the peripheral sector. In general, the rise or fall of a metric is monotonic along connectivity, but some of them reached extreme values in the intermediary sector.

Next sections summarize results of immediate interest and further insights can be obtained by skipping through the tables and figures in the Supporting Information document.

A. General characteristics of activity distribution among participants

Hubs and periphery swap fractions of participants and activity: while peripheral sector has $\approx 75\%$ of participants, it produces $\approx 10\%$ of all messages. Conversely, hubs sector present $\approx 10\%$ of participants and produces $\approx 75\%$ of all messages. Fewer threads are created by the hubs in proportion to total messages sent, while threads created by the periphery are twice as frequent as general messages. This suggests a complementarity between peripheral diversity and hub specialization which, on its turn, deepens the understanding of the interaction network as a meaningful system, notably if yield by online activity. These assertions are condensed in Table **??**.

B. Characters

Texts from peripheral vertices use more punctuation characters, digits and uppercase letters. Hubs use more letters and vowels among letters. The use of white spaces does not seem to have any relation to connectivity, with the exception that the intermediary often presented a slightly higher or lower incidence of spaces than both peripheral and hub sectors. Further information is given in Table **??**.

C. Tokens and words

The longer words used by hubs might be related to the use of a specialized vocabulary. Although the token diversity ($\frac{|tokens\neq|}{|tokens|}$) found in peripheral sector is far greater, this result has the masking artifact that the peripheral sector corpus is smaller, yielding a larger token diversity. This can be noticed by the token diversity of the whole network, which is lower than in any of the sectors. This same results apply to the lexical diversity ($\frac{|kw\neq|}{kw}$).

Punctuations among tokens are less abundant in hubs, and discrepancies here are larger than with character comparisons (subsection IV B). Known words are used more frequently by hubs.

ELE and CPP both exhibit intermediaries with the more frequent production of punctuation, less frequent production of known words, and the highest incidence of words with wordnet synsets among known words. This suggests some peculiarity in network structure, such as authorities in the intermediary sector of such networks, using smaller sentences and a more intensive use of jargons, as made explicit in the following sections.

Words with synsets, among known English words, are less frequent in hubs sector, further evidencing the jargon and specialization hubs develop.

Further information is given in Table **??**.

D. Sizes of tokens and words

Sizes of known words are smaller for hubs, which suggests its use of more common words, although some of the previous results suggests that hubs have a very differentiated and specialized vocabulary. Larger words seems to be related to intermediary sector, which might be related to the use of elaborated vocabulary. Further details are given in Table **??**.

E. Sizes of sentences

Hubs present the lowest average sentence size, both in characters and in tokens. We hypothesize that this smaller sentence use is related to the efficiency of hub specialization. Also, the incidence of usual known words seems to decay with connectivity, as does the number of known words with wordnet synsets. This reflects our view that connectivity is inversely proportional to diversity.

Further information is given in Table **??**.

F. Messages

Connectivity was related to smaller messages in terms of characters and tokens. ELE list displayed an inverse situation: the more connected the sector, the longer the messages are. This was considered a peculiarity of the culture bonded with the political subject of ELE list, to be further verified. Regarding sentences, the size of messages seem to hold steady throughout connectivity. Further information is given in Table **??**.

G. POS tags

Lower connectivity yields more nouns and less adjectives, adverbs and verbs. This suggests that the networks collect issues important to the world by the peripheral sector. These issues are qualified, elaborated about, by the more connected participants. This is a further indicative that peripheral sectors are related to diversity while hubs relate to specialization. Further information is given in Table **??**.

H. Wordnet synsets

I. Differentiation of the texts from Erdös sectors

Results from our adaptation of the Kolmogorov-Smirnov test suggest that the texts produced by each sector are extremely different. Intermediary sectors sometimes exhibit greater differences from periphery and hubs than these extreme sectors themselves (Tables **??** and **??**). This differentiation of the three sectors is a strong indicative that the Erdös Sectioning described in[?] reveals meaningful sectors of the networks.

Tables **??**-**??** illustrate two strong results:

- Differences of textual production of the Erdös sectors are extreme. This can be noticed from the high values on these tables, beyond reference values used for the acceptance of the null hypothesis (see Section III C).
- Differences between sectors on the same network (Tables **??**, **??**, **??** and **??**) are greater than differences between same sector from distinct lists (Tables **??**, **??**, **??** and **??**).

We can summarize these results stating that the extreme difference found between the texts produced the Erdös sectors are greater that that found between that of texts from different networks or from the same sector of different networks.

J. Correlation of topological and textual metrics

Correlation of degree and strength metrics is substantially smaller for intermediary sector. Also noteworthy is the negative correlation of degree and message size (number of characters, tokens or sentences) that intermediaries presented. This and other insights can be drawn from Tables **??**, **??** and **??**. Overall, negligible correlation is found between textual and topological metrics.

K. Formation of principal components

Principal components formation seem to be the less stable of all results reported in this study. First component, with $\approx$ 25% of dispersion, relies heavily on POS tags, and slightly on sizes of tokens, sentences and messages. Second component, with almost 12% of dispersion, blends topology, POS tags and size of textual

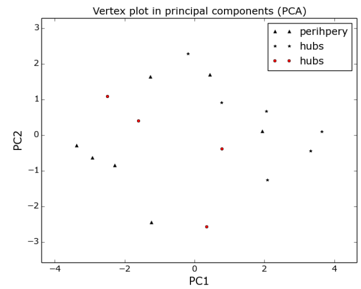

FIG. 2. First two principal components.

units. Third component, with about 8.5% of dispersion is mostly nouns frequency and size of textual units. Fourth and fifth components present less than 5% of total dispersion, but are included in the Supporting Information document for completeness of exposition.

Tables **??**-**??** yield these results and further insights.

L. Results still to be interpreted

Histogram differences of incident word sizes with and without repetition of words are constant. That is, in each email list, when a histogram of word sizes were made with all words written, and another histogram made with sizes of all *different* words, the cumulative absolute difference of the two histograms throughout the bins were found constant for all lists analysed. When all known English words were considered, the difference sums up to ≈ 1.0. When stopwords are discarded, the difference found was different, but still constant, slightly above 0.5. When only stopwords were considered, the difference is ≈ 0.6. When only known English words that does not have wordnet synsets are used, this difference is ≈ 1.2. Appendix SII and Figures S1-S5 are dedicated to this histogram differences.

V. FINAL REMARKS

This is a first systematic exploration of the relation between topological and textual metrics in human interaction networks, as far the author knows. Different textual features were scrutinized and were found to present evident patterns, specially in relation to topological measures and the Erdös sectors. Furthermore, results suggest that less connected participants bring external content and concepts, while hubs qualify the content. For example, periphery sectors present more nouns while hubs use more adjectives and usual words. Such findings have potential applications in the collection and diffusion and information, resources recommendation in linked data contexts, and open processes of document elaboration and refinement[?, ?, ?, ?, ?].

A. Further work

Similarity measures of texts in message-response threads has been thought about by the author, and some results are being organized. These are two hypothesis obtained from recent experiments:

- existence of information “ducts”, observable through similarity measures. These might coincide with asymmetries of edges between vertexes pairs, with homophily or with message-response threads, to point just a few possibilities.

- Valuable insights can be derived from the self-similarity of messages by same author, of messages sent at the same period of the day, etc. This includes incidences of word sizes, incidences of tags and morphosintactic classes, incidences of particular wordnet synset characteristics and wordnet word distances.

Current results suggests that diversity and self-similarity should vary with respect to connectivity. Literature usually assumes that periphery holds greater diversity[?] , which can be further verified, for example through the diversity of entries.

Other potential next steps are:

- The observation of most incident words and word types, such as words related to cursing or to food.

- Interpretation of the results exposed in Section IV L.

- Extend word class observations, e.g. to include plurals, gender, common prefixes and suffixes.

- The observation of date and time in relation to textual production of interaction networks and to activity characteristics (e.g. dispersion of sent time along the day or the week). This was tackled by the author for the topological characterization of interaction networks[?] , but left aside in this article.

- A careful analysis of each textual features distribution which is likely to reveal multimodal outlines and other non-trivial characteristics.

- Extend analysis to the windowed approach along the timelines used in the article where hub, peripheral and intermediary sectors where topologically characterized[?] .

- For ELE list, the more connected the sector, the longer the messages are. This is the inverse of what was found in the other lists, and was considered a peculiarity of the culture bonded with the political subject of ELE list. This hypothesis should be further verified.
- Tackle the same analysis on networks with languages other than English. This is especially important for easing applications[?] and should rely on dedicated implementation of tokenization, lemmatization and attribution of POS tags.

- Observe a broader set of human interaction networks and the resulting types of networks and participants with respect to topological and textual features.

- Analyse interaction networks from other platforms such as from LinkedIn and Facebook, etc.

- Sentiment analysis.

ACKNOWLEDGMENTS

Financial support was obtained from CNPq (140860/2013-4, project 870336/1997-5), United Nations Development Program (contract: 2013/000566; project BRA/12/018) and FAPESP. The authors are grateful to the Gmane creators and maintainers for the public email list data, to the communities of the email lists and other groups used in the analysis, and to the Brazilian Presidency of the Republic for keeping Participabr code and data open. We are also grateful to developers and users of Python scientific tools.

Appendix A: Meaning of acronyms and abbreviations used the tables

Some concepts, such as *contractions*, *token* and *char* are standard in natural language processing, and the reader is invited to visit[?] .