Language differences in human interaction networks

Renato Fabbri[a,*], Osvaldo N. Oliveira Jr.[b]

[a] *Institute of Mathematics and Computer Sciences (ICMC/USP) - Avenida Trabalhador So-carlense, 400 - Centro, So Carlos, SP, Brazil*
[b] *So Carlos Institute of Physics (IFSC/USP) - Avenida Trabalhador So-carlense, 400 - Centro, So Carlos, 13566-590, SP, Brazil*

Abstract

Social networks has been widely considered in recent scientific literature. A central question remains, though: how are the linguistic and topological features of social actors related? In this work, we describe a fundamental association between connectivity-defined sectors of networks and the language used therein. The hubs, intermediary and peripheral sectors of a network, defined though the connectivity distribution, present remarkable statistical distances between the texts they produce. Such distances are greater then e.g. distances between segments of a literary novel, and are often greater than between texts of different networks. This result holds for networks related to very distinct subjects, such as programming libraries and national elections, and are valid for raw textual features (such as token size), grammatical features (such as fraction of adjectives), and semantic features (such as obtained through the canonical Wornet). Furthermore, the measurements obtained allowed for a first tentative characterization of the language used in the networks and among the sectors. The findings herein reported should assist further studies relating language and topology in social networks, e.g. by relating communities or special groups of agents, such as of brokers and those with a high clustering coefficient, by performing analyses using other measures or comparison techniques, and by the consideration of other social networking systems.

*Corresponding author
Email address: fabbri@usp.br (Renato Fabbri)

Keywords: social networks, complex networks, language, text mining, Erds sectors

2019 MSC: 00-01, 99-00

1. Introduction

The first studies dealing explicitly with human interaction networks date from the nineteenth century while the foundation of social network analysis is generally attributed to the psychiatrist Jacob Moreno in mid twentieth century [1, 2]. With the increasing availability of data related to human interactions, research about these networks has grown continuously. Contributions can now be found in a variety of fields, from social sciences and humanities [3] to computer science [4] and physics [5, 6], given the multidisciplinary nature of the topic. One of the approaches from an exact science perspective is to represent interaction networks as complex networks, with which several features of human interaction have been revealed [5, 6]. For example, the connectivity distribution of human interaction networks tends to obey a power-law, which points to the existence of a small number of highly connected hubs and a large number of poorly connected nodes. Text mining is a multidisciplinary field, it is an extension of data mining to (often unstructured) textual data with the goal of discovering structure and meaning [7]. A general outline of a text mining endeavor involves structuring input text, deriving patterns and the evaluation of the output. There are actually numerous models of such outline, as e.g. considering document collection and obtaining a final report in the start and end respectively [8]. Text mining tasks include document summarization, sentiment analysis and natural language processing techniques such as part-of-speech tagging [9]. Among the applications one may include social media monitoring, automated ad placement, and development of tools for semantics [8]. It is believed that applying text mining to social media can yield interesting findings in human behavior [7]. Although there is no clear cut, text mining is sometimes divided into linguistic and non-linguistic [7], and in this study we use

both perspectives. In the first case, techniques borrowed from linguistics are present, such as the analysis of discourse and part-of-speech tagging, and it is often mingled with natural language processing or computational linguistics. In the non-linguistic text mining, text is analyzed by means of statistical features derived from e.g. the size of tokens and sentences, and might be more easily related to the intuitive concept of data mining of text. In this paper, we report on striking differences among the language used by the hub, intermediary and peripheral sectors of human social interaction networks. Such contrasts were found in networks in diverse scales and using a number of text-related statistics, from the usage of individual characters to token sizes, part-of-speech tags and Wordnet synsets. These results potentially encompass the first direct report on the association of topological features of individuals in human interaction networks to the language they employ.

This document is organized as follows. Next subsections briefly discuss related work and terminology remarks. Section 2 describes the data analyzed and the methods employed to derive the networks and sectors, and to obtain the measurements. Section 3 holds the results achieved. Section 4 is dedicated to conclusions and further work. Moreover, the Supplementary document provides all measurements achieved, within hundreds of tables, in order to support current findings and furnish the reader with the means to draw additional hypotheses.

1.1. Related work

In [10], the authors described the remarkable stability of the overall topological structure of social networks, a stability that holds across network temporal evolution and across different networks. In the same work, a sound method for deriving the hubs, intermediary and peripheral sectors of a network is thoroughly described. As it is fundamental to this present paper, such method is summarized in Section **??**. The studies performed considering the interaction networks of participants and their language focus on specific phenomena or features, such as rumor propagation [11, 12, 13, 14] and language dynamics (e.g.

language shift, emergence or learning) [15, 16, 17, 18, 19, 20, 21, 22, 23]. Linguistic networks have been addressed, such as in [24] where authors reported on the practicability of using word-adjacency networks for (sentiment) polarity detection, but no interaction network of human agents are considered in them. The studies most closely related to this present article, found by the authors, are: 1) [25], where both linguistic and topological features are used to predict socioeconomic attributes, but no relation among topological and linguistic features is reported. 2) [26], where linguistic and topological features are used to inspect the role spammers hold in Twitter and email networks, but, also, no relation among linguistic and topological features is reported. In summary, the work here presented is potentially unique in observing the relation, in real systems, of basic topological features of the participants in their interaction networks to the basic characteristics of the language they use therein. A thorough presentation of the methods and resulting measurements is lengthy and tiresome, and does not add to the core insights provided. Thus, we here summarize the findings and, if the reader finds necessary, she will find a more comprehensive consideration of the procedures and outcomes in the doctoral thesis which motivated this paper [27]. Furthermore, extensive listings of the measurements, encompassing hundreds of tables, are found in the Supplementary document of this article.

1.2. Terminology remarks

A major issue in current vocabulary arises when considered the hubs, intermediary and peripheral sectors. In most cases, the sectors are also partitions (i.e. they are non-overlapping sets whose union is the complete superset), but it depends on the criteria for obtaining the sectors [10], thus the use of the more general term *sector*. Language shift is the process by which speakers start to use different linguistic features or a different language altogether. Despite the fact that a language shift may be regarded as a type of language differentiation, this document is only concerned with the contrasts found in the language used by distinct sets of agents, more specifically, to the differences in the language

found to be employed by the hubs, intermediary and peripheral participants. Finally, complex and social networks comprehend a multidisciplinary field, and e.g. nodes, participants, actors, and agents are used to refer to analogous concepts. We strived to keep the vocabulary consistent.

2. Materials and methods

2.1. The data

Email list messages were obtained from the Gmane email archive, which consists of more than $20,000$ email lists (discussion groups) and more than 130×10^6 messages [28]. These lists cover a variety of topics, mostly technology-related. The archive can be described as a corpus along with message metadata, including sent time, place, sender name, and sender email address. The usage of the Gmane archive in scientific research is reported e.g. in studies of isolated lists and of lexical innovations [29, 4]. After analyzing dozens of the networks (derived from Gmane and data from Twitter, Facebook and Participabr), we randomly selected 18 email lists for the thorough measurements that are available in the Supplementary document. We also selected five of the email lists to illustrate the results in this paper. These lists are as follows:

- Linux Audio Users list[1], with participants from different countries with artistic and technological interests. English is the prevailing language. Abbreviated as LAU from now on.
- Linux Audio Developers list[2], with participants from different countries; a more technical and less active version of LAU. English is the prevailing language. Abbreviated as LAD from now on.
- Developer's list for the standard C++ library[3], with computer programmers from different countries. English is the prevailing language. Abbre-

[1] gmane.linux.audio.users is list ID in Gmane.
[2] gmane.linux.audio.devel is list ID in Gmane.
[3] gmane.comp.gcc.libstdc++.devel is list ID in Gmane.

Table 1: Columns $date_1$ and $date_M$ have dates of first and last messages from the 20,000 messages in each email list. N is the number of participants (number of different email addresses), Γ is the number of discussion threads (count of messages without antecedent), $\overline{M}$ is the number of messages missing in the 20,000 collection ($100\frac{23}{20000} = 0.115$ percent in the worst case).

list	$date_1$	$date_M$	N	Γ	$\overline{M}$
LAU	2003-06-29	2005-07-23	1147	3374	5
LAD	2003-07-03	2009-10-07	1232	3114	4
MET	2005-08-01	2008-03-07	477	4607	23
CPP	2002-03-12	2009-08-25	1036	4506	7
ELE	2002-03-18	2011-08-31	302	6070	54

viated as CPP from now on.

- List of the MetaReciclagem project[4], a Brazilian email list for digital culture. Portuguese is the prevailing language, although some messages are written in Spanish and English. Abbreviated as MET from now on.

- List for discussion of the election reform[5]. English is the prevailing language. Abbreviated ELE from now on.

Table 1 holds an overview of the messages in each of these lists. MET was not used for the textual measurements because most messages are not written in English. The selection of these lists was performed in preliminary stages of this work, in compliance with [10, 27], as a convenient, small and diverse set of lists to be considered in depth, and has no impact on the results as they are used only for illustration in Table 1 and Section 3.2.

[4] gmane.politics.organizations.metareciclagem is list ID in Gmane.
[5] gmane.politics.election-methods is list ID in Gmane.

2.2. Derivation of interaction networks and of hubs, intermediary and peripheral sectors

Networks in this paper are directed and weighted, the most informative of the usual models used for interactive networks [4, 30, 31]. Moreover, all results hold for directed unweighted, undirected weighted, and undirected unweighted representations of the interaction networks, with possible exceptions for Sections 3.10 and 3.11. The derivation of such networks are carefully described in [10], and in essence is yield by the representation of authors and replies as nodes and links, respectively. For the results here reported, it is crucial to discern the nodes which are hubs, intermediary and peripheral. Although it is qualitatively described in the basic literature, a quantitative method for such classification is described in [10], and yields the Erdös sectors by comparing real networks against the Erdös-Rényi random network model.

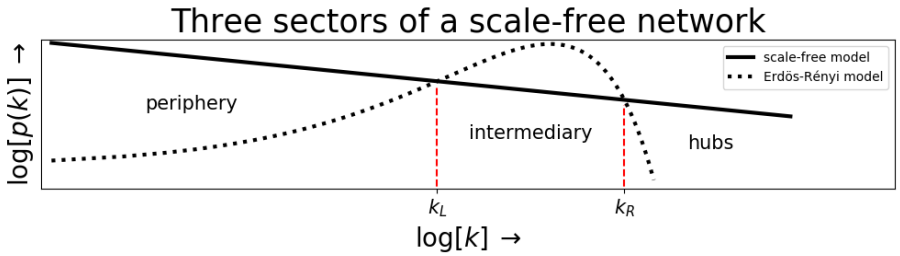

Figure 1: Classification of nodes by comparing degree distributions. The binomial distribution of the Erdös-Rényi network model has more intermediary nodes, while a scale-free network, associated with a power-law distribution of connectivity, has more peripheral and hub nodes. The sector borders are defined with respect to the intersections of the distributions. Characteristic degrees are in the compact intervals: $[0, k_L]$, $(k_L, k_R]$, $(k_R, k_{max}]$ for the peripheral, intermediary and hubs sectors: the "Erdös sectors". Accordingly, the connectivity distribution of empirical interaction networks, e.g. derived from email lists, can be sectorialized by comparison against the associated binomial distribution with the same number of nodes and links. Further information is provided in Section 2.2.

2.3. Topological measures

The derivation of the sectors, described in the previous subsection, relies only on the very basic measures of degree and strength. In order to avoid unnecessary intricacy, the correlation and principal component analysis were performed using olny degree, strength, and clustering coefficient. These measures are defined as follows:

- Degree k_i: number of links incident to node i.
- Strength s_i: sum of weights of all links incident to node i.
- Clustering coefficient cc_i: fraction of pairs of neighbors of i that are linked, i.e. the standard clustering coefficient for undirected graphs.

2.4. Text-related measures

This work focuses on simple measurements derived from texts. The measures are:

- Frequency of characters: letters, vowels, punctuations and uppercase letters.
- Number of tokens, frequency of punctuations among tokens, frequency of known words, frequency of words that have 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, nouns and other POS (Part-Of-Speech) tags. We implemented a Brill POS tagger because of the massive amount of textual data that we had to analyze (the Brill tagger is very fast). We used the "universal" tagset described in [32] (developed to account for many languages) and trained the tagger with

both Brown and Treebank corpus[6] divided into 80/20% of sentences for training and evaluation. The tagger achieved 94.95% of accuracy.

- Fraction of words in each Wordnet [34] top-most hypernyms, such as abstraction and physical entities for nouns or act for verbs.

- Mean and standard deviation of the number of Wordnet synset relations, such as holonyms and meronyms, domains, lemmas and verb groups.

This selection of measures is based on: 1) the lack of such information in the literature, to the best of our knowledge; 2) potential relations of these incidences with topological aspects, such as connectivity; 3) the interdependence of textual artifacts suggests that simple metrics should mirror complex and more subtle aspects. A preliminary study, with the complete works from the Brazilian writer Machado de Assis [35], made clear that the measurements vary with respect to style.

2.4.1. About Wordnet

We made use of the statistics derived from the incidence of synsets from the caninical Wordnet [34] and their characteristics. This calls for a brief description of what Wordnet is, what is a synset and what are such characteristics:

- Wordnet groups synonyms into synsets, provides a number of relations among the synsets, definitions and use examples.

- A "synset" is defined by Wordnet documentation as a set of one or more synonyms that are (somewhat) interchangeable without changing the truth value of the proposition in which they occur.

- Synset characteristics: a synset is associated to other synsets by a number of semantic relations which differ with the POS tag attributed to them. Examples of such relations include hyponyms and hypernyms (respectively

[6] We used the full Brown corpus while only 5% of the Penn Treebank (as included in NLTK [33]).

less and more general terms), meronyms and holonyms (respectively "is part of" and "has part"), lemmas (canonical forms of a word), similar words by semantic criteria, entailments. Some characteristics are derived from the synset in relation to the whole Wordnet network. In this respect, we used only the maximum and minimum depth of the synset, which are respectively the maximum and minimum number of hypernyms from the synset to the root synset (e.g. 'thing' for nouns).

2.4.2. Relating text and topology

The topological and textual measures were related by:

1. textual measures in each of the Erds sectors, which are delimited by topological criteria as described in Section 2.2.
2. Correlation of metrics of each vertex, facilitating pattern detection involving topology of interaction and language.
3. Principal components formation derived from the usual Principal Component Analysis.

2.5. Statistical distances

An adaptation of the Kolmogorov-Smirnov test [36] was used to observe differences in textual content, as follows. Let F and F' be 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, that the samples are drawn from the same underlying distribution, if:

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

where $D_{F,F'} = sup_x[F - F']$ is the Kolmogorov-Smirnov statistic (illustrated in Figure 2) 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

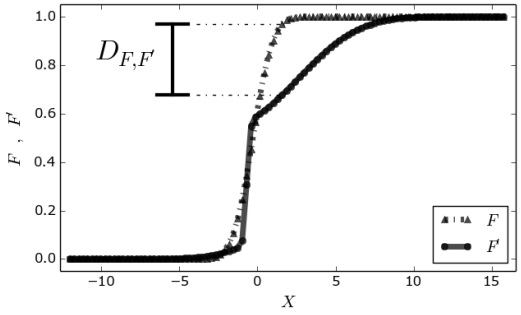

Figure 2: The Kolmogorov-Smirnov statistic $D_{F,F'}$: the maximum difference between two cumulative distribution functions.

In comparing two empirical distribution functions, $D_{F,F'}$ is given, as are n and n'. All terms in equation 1 are positive and $c(\alpha)$ can be isolated:

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

When c' is high, corresponding low values of α favor rejecting the null hypothesis. For example, when c' is greater than ≈ 1.7, one might assume that F and F' differ. We used c' as a measure of how much the distributions differ and for deriving hypotheses about how different are the underlying mechanisms of generation of texts. In [36], systematic measurements of the c' statistic illustrate that c' and $D_{F,F'}$ are consistent and useful in considering the similarity or difference in the distributions underlying sets of observations. A note of caution should be given here: what is a difference in distributions might vary with context. A $D_{F,F'}$ of only 0.0001 will yield a very large c' if n and n' are large enough. On the other hand, a large $D_{F,F'}$ with a small c' is not a strong evidence that the distributions differ. Therefore, we consider both $D_{F,F'}$ and c' simultaneously in our analysis. Each text-related measure yields a histogram for the whole network, and one for each of the Erds sectors. To obtain the histograms, one value was obtained through each message.

3. Results and discussion

The consideration of all measures and all distances results in a vast number of measurements, comparisons and tables. Also, a thorough discussion of these outcomes becomes lengthy and tiresome. We here present a summary of the findings and a more comprehensive presentation is available at [27] if the reader finds necessary. Furthermore, the Supplementary document of this paper provides all measurements in all 18 randomly selected email lists, together with analytical considerations. Each email lists is related to a numerical TAG, as given in the Supplementary document. Such numerical tags are employed in the tables of the following sections.

The most important result is the strong statistical evidence of differentiation of each Erdös sector with respect to the language used. This conclusion can be reached by observing the differences in the measurements of textual features in each sector, and is supported on firm theoretical grounds by using the adapted Kolmogorov-Smirnov test presented in Section 2.4.2.

Other relevant results are:

- the identification of patterns in the frequency of use of nouns, adverbs, sizes of words, depth of Wordnet synsets and other linguistic features, in the network as a whole. We did not find in the literature any indication for such values, thus we understand useful to report e.g. that about 15% of the characters are spaces and that nouns often account for more than 25% of the words. These values are available in the Supplementary document and not in the body of this paper, since the focus here is the evidence that texts from distinct sectors differ.

- Evidence of what is different in the language used by each Erdös sector. For example: hubs were found to use more contractions, more common words, and less punctuation if compared to the rest of the network, especially the peripheral sector. In general, the rise or fall of a text-related metric is not relevant or is monotonic along connectivity, but some of them reach extreme values in the intermediary sector.

The next sections summarize results of immediate interest and further insights can be obtained by skimming through the tables of the Supplementary document of this paper. We illustrate with just one table of each kind, and from networks obtained with 2000 messages. After analysing the networks in many scales, we considered the networks obtained by constructed using 1000 and 2000 messages as representative for the phenomena being characterized. The findings with 1000 messages are the same as with 2000 messages. This motivated the exclusion of the tables with measurements in networks of 1000 messages.

3.1. General characteristics of activity distribution among sectors

In order to support a relevant consideration of findings derived from text-related measurements, this section provides a glance at the general structure of these networks, with emphasis on the activity of the Erds sectors. In almost all our observations, the peripheral sector was responsible for starting most of the discussion threads, i.e. for sending messages to the list which are not replies. This is surprising since the peripheral sector is responsible for fewer messages. It suggests a complementarity between peripheral diversity and hub specialization, which, on its turn, emphasizes the understanding of the interaction network as a meaningful system. These assertions are condensed in Table 2. Less often, the intermediary sector is responsible for the largest number of messages and of threads. Also meaningful is that the hubs sector is responsible for most of the messages, which is not obvious: hub participants are far more active but way less numerous. Interestingly, in such a setting where every characteristic differs with respect to distinct sectors, there was no evidence of difference on the size of the threads started by each sector.

3.2. Evidence that the texts from Erdös sectors differ

This is the most important result reported in this paper: there is strong enough evidence to support the assertion that the language used by distinct Erdös sectors are different. Figure 3 illustrates, and Tables 3-**??** exemplify three results:

Table 2: Participants, messages and threads among each Erdös sector: (**p.** for periphery, **i.** for intermediary, **h.** for hubs) in a total timespan of 0.72 years (from 2003-11-30T20:21:32 to 2004-08-19T18:11:24). N is the number of participants, M is the number of messages, Γ is the number of threads, and γ is the number of messages in a thread. The % denotes the usual 'per cent' with respect to the total quantity (100% for **g.**) while μ and σ denote mean and standard deviation. TAG of list: 10

	g.	**p.**	**i.**	**h.**
N	131	80	46	5
$N_{\%}$	100.00	61.07	35.11	3.82
M	1000.00	136.00	361.00	503.00
$M_{\%}$	100.00	13.60	36.10	50.30
Γ	292.00	76.00	147.00	69.00
$\Gamma_{\%}$	100.00	26.03	50.34	23.63
$\frac{\Gamma}{M}\%$	29.20	55.88	40.72	13.72
$\mu(\gamma)$	2.74	2.76	2.81	2.58
$\sigma(\gamma)$	0.44	0.43	0.39	0.49

- The evidence that the textual production of the Erdös sectors is different. This can be noticed from the high values of c' on these tables, clearly above reference values used for the acceptance of the null hypothesis (the null hypothesis being that the probability distributions generating the samples are the same). Notice that the texts are distinct and the authors are different. In this context, it is suitable to rely in reference values. As provided by [36], in well known documents from the English literature, $c' \geq 1$ is rare for the same author when using the size of tokens to build the histograms. In contrast, notice the high values in table 3.

- The distance (as measured by c') between sectors on the same network is often greater than between the same sector from distinct networks. This is exemplified by Tables 4 and 5.

- Intermediary sectors sometimes exhibit greater differences against the pe-

riphery or hubs than these extreme sectors between themselves, as exemplified in Tables 3 and 4. This differentiation of the three sectors is a further indicative that the Erdös Sectioning described in Section 2.2 reveals meaningful sectors of the networks.

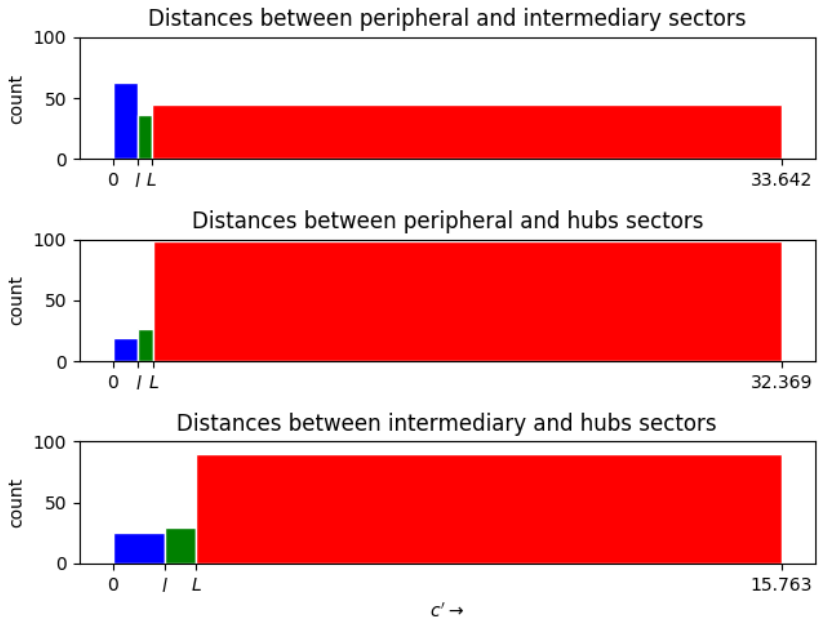

Figure 3: Histograms of the c' statistic between the Erdös sectos for all the language-related measurements performed. The histogram bins are limited by the values $l = 1.22$, and $L = 1.95$, which are related to the significance levels of 0.1 and 0.001, respectively, as described in Section 2.5. In qualitative terms, one may associate the bars to: measurements that are not evidence of difference (blue), measurements that are evidence of difference (green), and measurements that are strong evidence of difference (red). Notice that, although the c' values obtained by comparing the peripheral and intermediary sectors present less values in the red bar, the values reach the magnitudes comparable to the highest values obtained by comparing the peripheral and hubs sectors. Further information is provided in Section 3.2.

Therefore, reference distance values are derived from literary documents (such as Shakespeare) and from the comparison of different networks. Notice that the highest values in Table 5 are related to the ELE list, which employs very distinct linguistic features, which are noticeable e.g. by the discrepant average

number of sentences per message provided in the Supplementary document of this paper. Moreover, the greatest c' distances are most often found when comparing the hubs to the other sectors, a further confirmation that the language employed by participants in each of the Erdös sectors is somewhat specialized.

Two measurements for each pair of sectors are in Tables 3-**??**. The top value is c' while the bottom value is the Kolmogorov-Smirnov statistic. In Tables5 only c' values are shown to illustrate that the measurements found between networks are comparable to (and often lower than) measurements found between sectors of a same network.

Table 3: Measurements of c' and the KS statistic related to sizes of tokens between the Erdös sectors and the whole network. The abbreviations are: g for global, p for peripheral, i for intermediary, h for hubs. See Section 2.4.2 for and explanation of the measures and Section 3.2 for discussion. TAG of list: 6

	p-g	**i-g**	**h-g**	**p-i**	**p-h**	**i-h**
c'	4.327	17.168	7.851	18.907	7.833	15.540
KS	0.014	0.115	0.044	0.129	0.045	0.129

Table 4: Measurements of c' and the KS statistic related to the frequency of use of nouns between the Erdös sectors and the whole network. The abbreviations are: g for global, p for peripheral, i for intermediary, h for hubs. See Section 2.4.2 for and explanation of the measures and Section 3.2 for discussion. TAG of list: 1

	p-g	**i-g**	**h-g**	**p-i**	**p-h**	**i-h**
c'	0.642	1.791	6.936	1.007	6.970	7.510
KS	0.023	0.067	0.537	0.044	0.560	0.607

3.3. What differs and how the texts from the sectors differ

In the next sections we will look through language-related measurements and summarize the findings about what might be different in the textual features of the sectors. One should keep in mind that our core result is the evidence that the texts from distinct sectors differ. The following discussion of what differs

Table 5: c' values for the frequency of use of nouns. Comparison of the same sector between lists, each author is an observation. See subsection 3.2 for discussion and directions.

	CPP-LAD	CPP-LAU	CPP-ELE	LAD-LAU	LAD-ELE	LAU-ELE
P	1.35	4.05	5.80	3.00	5.41	4.94
I	1.27	0.78	4.01	0.84	3.84	3.94
H	0.98	1.94	3.17	1.32	3.82	4.47

and how it differs is interesting but is both derived from less strong statistical evidence and less crucial for our current stage of researching these structures. Nonetheless, for the sake of clarity, we state the main findings in this respect:

- Peripherals were found to use more nouns while hubs use more verbs and adverbs. The fraction of adjectives did not change as irrefutably with respect to connectivity, but given that nouns are more numerous in the periphery sector, there are more adjectives per noun in the hubs sector texts.

- Sentences and messages were found smaller in the more connected sectors although punctuation was more incident in the less connected sectors.

- The differences found in analyzing Wordnet synset hypernyms were found less well behaved. Often, the sectors exhibit noticeable differences but greater and smaller incidences are found in all sectors (but in different networks). Some of these incidences are more systematic and this analysis assisted by Wordnet is the only semantic analysis we made, which is why we included these results.

Appendix Appendix A presents tables with counts for larger incidence in each sector throughout the networks. In the following discussion we provide example tables for each set of measurements. In the measurements derived from synset hypernyms the example tables were less meaningful because of the greater variability and therefore we present the counts of greater incidence directly. In any case, the measurements for each of the networks are in the Supplementary

document of this paper. The analysis of the measurements is not trivial because of the number of different measures and because differences in measurements are not obviously relevant. Furthermore, there is too much variation among networks which renders worthless the usual global measurements such as mean and variance when obtained from all the systems at once. In order to obtain consistent results, we considered *weak evidence of difference* in sectors in a network if maximum measure is at least 10% greater than minimum measure, i.e. $\frac{maximum\ measure}{minumum\ measure} > 1.1$. We considered *evidence of difference* in sectors in a network if $\frac{maximum\ measure}{minumum\ measure} > 1.2$. When $\frac{maximum\ measure}{minumum\ measure} > 1.5$, we considered *strong evidence of difference*. We then looked through each measure in all networks to reach compelling observations about the differences of sectors through all networks. Instead of the counts of maximum incidences, we also performed a weighted count: when the evidence was weak, the sector received an add of 0.5 in the weighted count; then the evidence was strong, the sector received an add of 2 in the weighted count. The results were qualitatively the same, so we omit the tables with weighted counts. Also useful here is the definition of lower sectors (peripheral and intermediary), upper sectors (intermediary and hubs) and extreme sectors (peripheral and hubs). We should also state when measurements peak often at the intermediary sector, be it a maximum or minimum peak.

The tables that follow have terms for measurements that might be immediately inferred, but this depends on the background of the reader. Therefore the terms are explicitly defined in Appendix Appendix B.

3.4. Characters

Most often, peripheral and intermediary sectors use more digits, uppercase letters and punctuation. Hubs sectors were found to use more spaces in most cases. These results are illustrated in Table 6.

Table 6: Characters in each Erdös sector (**p.** for periphery, **i.** for intermediary, **h.** for hubs). TAG: 6

	g.	**p.**	**i.**	**h.**
$chars$	1485813	552986	554328	378499
$chars_{\%}$	100.00	37.22	37.31	25.47
$\frac{spaces}{chars}$	12.94	12.79	12.82	13.35
$\frac{punct}{chars-spaces}$	9.54	10.53	10.15	7.20
$\frac{digits}{chars-spaces}$	4.49	7.13	3.87	1.54
$\frac{letters}{chars-spaces}$	83.95	80.09	83.95	89.65
$\frac{vowels}{letters}$	36.94	36.10	36.98	38.00
$\frac{uppercase}{letters}$	4.49	4.60	4.68	4.07

3.5. Tokens and words

Hubs were found to use more contractions and stopwords, while the peripheral sectors exhibit a greater incidence of punctuations among tokens. Although the token diversity ($\frac{|tokens\neq|}{|tokens|}$) found in the peripheral sector is 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. These results are exemplified in Table 7 where mean and variance were taken with respect to the length in characters of tokens, known words and stopwords.

Table 7: Tokens in each Erdös sector (**p.** for periphery, **i.** for intermediary, **h.** for hubs). TAG: 1

	g.	**p.**	**i.**	**h.**
$tokens$	286232	146472	134852	4908
$tokens_{\%}$	100.00	51.17	47.11	1.71
$tokens \neq$	3.00	4.01	3.66	24.08
$\frac{knownw}{tokens}$ %	25.76	25.11	26.21	32.84
$\frac{knownw \neq}{knownw}\%$	4.51	6.22	5.92	42.80
$\frac{stopw}{knownw}\%$	42.46	38.92	43.60	98.33
$\frac{punct}{tokens}\%$	33.18	34.09	32.56	23.11
$\frac{contrac}{tokens}\%$	0.16	0.10	0.18	1.67
$\mu(\overline{tokens})$	3.19	3.10	3.26	3.65
$\sigma(\overline{tokens})$	2.53	2.54	2.52	2.60
$\mu(\overline{knownw})$	4.89	4.69	5.06	5.50
$\sigma(\overline{knownw})$	2.37	2.41	2.31	2.28
$\mu(\overline{knownw \neq})$	6.53	6.39	6.27	6.16
$\sigma(\overline{knownw \neq})$	2.53	2.50	2.46	2.42
$\mu(\overline{stopw})$	2.83	2.83	2.83	2.81
$\sigma(\overline{stopw})$	0.87	0.84	0.86	1.17

3.6. Sizes of sentences

Hubs present the lowest average sentence size in terms of characters, tokens, known words or punctuations. This result is illustrated in Table 8 and might be considered counterintuitive given that punctuation is more abundant in the texts of less connected participants.

Table 8: Sentences sizes in each Erdös sector (**p.** for periphery, **i.** for intermediary, **h.** for hubs). TAG: 16

	g.	**p.**	**i.**	**h.**
$sents$	10757	1252	4529	4978
$sents_{\%}$	99.98	11.64	42.09	46.27
$\mu_S(chars)$	113.88	143.37	120.21	100.65
$\sigma_S(chars)$	318.65	750.47	276.21	88.88
$\mu_S(tokens)$	24.78	28.83	26.72	21.98
$\sigma_S(tokens)$	40.56	77.72	42.08	20.23
$\mu_S(knownw)$	7.81	8.37	8.26	7.25
$\sigma_S(knownw)$	8.18	9.38	9.30	6.56
$\mu_S(stopw)$	7.78	7.61	7.92	7.70
$\sigma_S(stopw)$	6.88	6.94	7.36	6.39
$\mu_S(puncts)$	4.29	5.42	5.04	3.33
$\sigma_S(puncts)$	9.92	13.08	12.13	5.82

3.7. Messages

Connectivity was found correlated to smaller messages in terms of characters, tokens, known words and punctuations. Connectivity was also found correlated to smaller messages in terms of the number of sentences, but it was less consistent. Interestingly, the number of stopwords per message was found greater in all sectors (but in different networks). This result is exemplified in Table 9.

Table 9: Messages sizes in each Erdös sector (**p.** for periphery, **i.** for intermediary, **h.** for hubs). TAG: 0

	g.	**p.**	**i.**	**h.**
$msgs$	1992	286	841	865
$msgs_{\%}$	100.00	14.36	42.22	43.42
$\mu_M(sents)$	5.21	6.08	6.43	3.74
$\sigma_M(sents)$	6.78	4.03	9.40	3.26
$\mu_M(tokens)$	145.82	230.45	186.07	78.71
$\sigma_M(tokens)$	260.61	291.17	326.68	127.13
$\mu_M(knownw)$	38.83	56.29	48.87	23.29
$\sigma_M(knownw)$	50.54	58.28	58.67	31.16
$\mu_M(stopw)$	34.29	41.96	42.42	23.84
$\sigma_M(stopw)$	41.11	32.32	52.81	25.35
$\mu_M(puncts)$	36.34	66.11	47.66	15.49
$\sigma_M(puncts)$	103.42	114.84	135.49	39.61
$\mu_M(chars)$	637.40	977.77	811.14	355.94
$\sigma_M(chars)$	1054.36	1195.70	1290.46	566.92

3.8. POS tags

We found that lower connectivity yields more nouns and less verbs and adverbs. Also, the fraction of adjectives does not change consistently, but given that peripherals use more nouns, we can conclude that hubs use more adjectives per noun. This suggests that the networks gather issues through the peripheral sector. These issues are qualified and proposed to be acted upon by the more connected participants. This is a further indicative that peripheral sectors are related to diversity while hubs relate to specialization. These results are exemplified in Table 10. Weaker evidence was found that hubs use more *adpositions*, determinants and 'particles and other functional words' while peripherals use more numerals.

Table 10: POS tags in each Erdös sector (**p.** for periphery, **i.** for intermediary, **h.** for hubs). Universal POS tags [32]: VERB - verbs (all tenses and modes); NOUN - nouns (common and proper); PRON - pronouns; ADJ - adjectives; ADV - adverbs; ADP - adpositions (prepositions and postpositions); CONJ - conjunctions; DET - determiners; NUM - cardinal numbers; PRT - particles or other function words; X - other: foreign words, typos, abbreviations; PUNCT - punctuation. TAG: 13

	g.	**p.**	**i.**	**h.**
NOUN	51.86	63.77	48.31	37.37
X	0.08	0.14	0.02	0.07
ADP	7.25	5.23	7.86	9.69
DET	7.48	6.47	7.43	9.28
VERB	16.93	11.93	20.01	20.54
ADJ	3.97	3.37	3.83	5.18
ADV	4.02	2.41	4.45	6.05
PRT	3.17	3.98	2.37	3.05
PRON	3.16	1.29	3.64	5.55
NUM	0.43	0.30	0.38	0.74
CONJ	1.65	1.11	1.70	2.49

3.9. Wordnet-related results

For correctly analyzing text production in terms of the Wordnet lexical database, we only considered words that had synsets[7] and that had at least one synset with the POS tag obtained with the POS tagger. There are often more than one synset with the same POS tag for each word, thus we chose the most frequent synset as ranked by Wordnet. This resulted in portions of tokens considered of $\approx 30\%$, but of more than 90% of all tokens with Wordnet synsets. This yields less strong results, but which we found still relevant as no similar outcome was found in the scientific literature by the authors. Moreover, obser-

[7] Examples of categories of tokens without synsets are stopwords, punctuations, numerals, acronyms and typos.

vations seem consistent and meaningful, Measures regarding Wordnet synsets often reach an extreme value (maximum or minimum) in the intermediary sector, which we understood as evidence that:

- the Erdös sectors are in fact relevant for human social structures, at least to the ones analyzed in this thesis.
- Human social networks present relations between connectivity and semantics.
- The intermediary sector might hold a deeper identity than that of a sector bounded by hubs and periphery sectors.
- The analysis of the language employed in social networksm, using Wordnet, reveals aspects of the structures which are not clear though the non-semantic analysis we performed.

Appendix Appendix C presents Wordnet-related results derived from POS tags and from synset relations (such as hypernyms and meronyms).

3.10. Correlation of topological and textual metrics

Overall, the correlation between textual and topological measurements was small. An exception is strength which was very often negatively correlated to the mean and variance of the number of punctuations (and sometimes with the number of known words or stopwords) with values below -0.4, but a few positive and high values (above 0.5) were also found. Interestingly, the number of punctuations per sentence was most often correlated to the number of stopwords while most often *not* correlated to the number of known words. Noteworthy is that degree is negatively correlated to clustering coefficient in intermediary and hubs sectors, which is consistent with the literature, but it is positively correlated for peripheral sectors. Other strong correlation associations of textual and topological measures were found but not systematically and might be indicatives of style from the different lists analyzed. These results are exemplified in Table 11.

Table 11: Pearson correlation coefficient for the topological and textual measures. TAG: 9

	cc	d	s	$\mu_S(p)$	$\sigma_S(p)$	$\mu_S(kw)$	$\sigma_S(kw)$	$\mu_S(sw)$	$\sigma_S(sw)$
cc	**1.00**	-0.03	-0.08	0.04	0.10	0.05	0.10	0.09	0.21
(p.)	**1.00**	**0.64**	0.42	0.12	0.19	0.09	0.22	0.08	0.22
(i.)	**1.00**	-0.58	-0.51	-0.10	-0.08	-0.26	-0.11	-0.24	-0.19
(h.)	**1.00**	**-0.86**	**-0.85**	0.33	0.09	0.14	0.21	0.14	0.11
d	-0.03	**1.00**	**0.98**	-0.05	0.00	0.04	0.05	0.09	0.12
	0.64	**1.00**	**0.78**	0.11	0.16	-0.00	0.16	0.06	0.22
	-0.58	**1.00**	**0.86**	0.10	0.14	0.29	0.18	0.30	0.28
	-0.86	**1.00**	**1.00**	-0.51	-0.25	-0.42	-0.34	-0.47	-0.35
s	-0.08	**0.98**	**1.00**	-0.05	-0.01	0.02	0.02	0.05	0.09
	0.42	**0.78**	**1.00**	0.10	0.15	0.10	0.19	0.21	0.35
	-0.51	**0.86**	**1.00**	0.13	0.07	0.29	0.10	0.32	0.35
	-0.85	**1.00**	**1.00**	-0.50	-0.25	-0.40	-0.32	-0.47	-0.34
$\mu_S(p)$	0.04	-0.05	-0.05	**1.00**	**0.82**	**0.65**	**0.61**	0.19	0.52
	0.12	0.11	0.10	**1.00**	**0.96**	**0.65**	**0.86**	0.18	0.59
	-0.10	0.10	0.13	**1.00**	**0.84**	**0.77**	**0.76**	0.34	0.50
	0.33	-0.51	-0.50	**1.00**	**0.78**	**0.93**	**0.96**	**0.92**	**0.97**
$\sigma_S(p)$	0.10	0.00	-0.01	**0.82**	**1.00**	0.58	**0.92**	0.16	0.52
	0.19	0.16	0.15	**0.96**	**1.00**	0.54	**0.89**	0.11	**0.62**
	-0.08	0.14	0.07	**0.84**	**1.00**	**0.73**	**0.98**	0.26	0.44
	0.09	-0.25	-0.25	**0.78**	**1.00**	**0.89**	**0.84**	**0.85**	**0.76**
$\mu_S(kw)$	0.05	0.04	0.02	**0.65**	0.58	**1.00**	**0.64**	**0.73**	**0.67**
	0.09	-0.00	0.10	**0.65**	0.54	**1.00**	**0.73**	**0.71**	**0.65**
	-0.26	0.29	0.29	**0.77**	**0.73**	**1.00**	**0.76**	**0.74**	**0.72**
	0.14	-0.42	-0.40	**0.93**	**0.89**	**1.00**	**0.94**	**0.97**	**0.95**
$\sigma_S(kw)$	0.10	0.05	0.02	**0.61**	**0.92**	**0.64**	**1.00**	0.27	0.56
	0.22	0.16	0.19	**0.86**	**0.89**	**0.73**	**1.00**	0.30	**0.79**
	-0.11	0.18	0.10	**0.76**	**0.98**	**0.76**	**1.00**	0.31	0.48
	0.21	-0.34	-0.32	**0.96**	**0.84**	**0.94**	**1.00**	**0.88**	**0.96**
$\mu_S(sw)$	0.09	0.09	0.05	0.19	0.16	**0.73**	0.27	**1.00**	**0.61**
	0.08	0.06	0.21	0.18	0.11	**0.71**	0.30	**1.00**	0.53
	-0.24	0.30	0.32	0.34	0.26	**0.74**	0.31	**1.00**	**0.74**
	0.14	-0.47	-0.47	**0.92**	**0.85**	**0.97**	**0.88**	**1.00**	**0.94**
$\sigma_S(sw)$	0.21	0.12	0.09	0.52	0.52	**0.67**	0.56	**0.61**	**1.00**
	0.22	0.22	0.35	0.59	**0.62**	**0.65**	**0.79**	0.53	**1.00**
	-0.19	0.28	0.35	0.50	0.44	**0.72**	0.48	**0.74**	**1.00**
	0.11	-0.35	-0.34	**0.97**	**0.76**	**0.95**	**0.96**	**0.94**	**1.00**

3.11. Formation of principal components

Using principal component analysis [37], the principal components formation of textual and topological metrics seems to be the less stable of all results reported in this study. The concentration of dispersion often peaked in the intermediary sector. Components are most often composed of both topological or textual features. Other than that, we observe that PCA is sensitive to the

measurements included and should reveal other insights in other settings. These results are exemplified in Table 12.

Table 12: PCA formation. TAG: 11

	PC1	**PC2**	**PC3**	**PC4**	**PC5**
cc	1.56	5.55	5.27	66.36	4.43
(p.)	2.07	21.29	-3.09	-16.58	-30.47
(i.)	3.53	-5.95	-6.72	51.70	8.40
(h.)	9.65	-14.73	3.26	14.54	27.15
d	3.28	39.23	-2.42	-3.75	1.71
	2.31	29.69	-4.90	7.49	10.16
	6.23	18.94	16.55	7.89	-1.65
	-10.34	13.75	-14.11	5.08	9.43
s	2.89	39.14	-2.73	-6.53	2.19
	2.30	29.95	-4.12	6.71	9.55
	6.37	19.06	16.00	10.14	-1.42
	-9.80	13.47	-14.84	6.58	13.35
$\mu_S(p)$	11.87	-6.88	-23.79	0.56	19.99
	-12.99	-3.75	-21.38	18.20	-12.48
	8.32	-18.13	13.89	6.33	-6.38
	10.80	-6.99	-19.67	0.26	6.55
$\sigma_S(p)$	11.31	-0.30	-25.29	10.55	-14.13
	-10.90	-0.79	-27.98	-9.77	8.49
	10.21	-14.05	15.44	-6.84	12.95
	8.72	-6.76	-21.47	2.17	-16.46
$\mu_S(kw)$	19.19	-6.59	-2.36	-5.99	11.85
	-19.64	0.08	-0.64	7.86	-7.62
	17.95	-8.73	0.93	-2.42	-16.55
	14.15	7.50	-7.33	-20.52	6.85
$\sigma_S(kw)$	17.26	-1.70	1.16	0.48	-24.22
	-17.26	1.60	-1.20	-18.84	10.50
	16.92	2.12	-4.56	-13.03	22.97
	13.31	10.62	-1.67	19.42	-11.12
$\mu_S(sw)$	15.36	-0.55	19.90	-4.40	14.27
	-14.96	6.19	21.48	9.07	-5.17
	15.77	4.76	-12.12	0.29	-20.25
	12.34	12.79	7.28	-15.30	8.15
$\sigma_S(sw)$	17.31	-0.06	17.07	-1.37	-7.20
	-17.57	6.65	15.22	-5.48	5.55
	14.70	8.27	-13.80	-1.36	9.42
	10.89	13.38	10.37	16.12	-0.94
λ	36.91	22.07	16.02	11.02	7.75
	37.77	25.85	14.81	8.29	7.11
	33.74	23.23	20.12	10.67	6.77
	40.07	27.24	20.10	6.53	3.68

3.12. Notes about the typology derived from the Erdös sectors

In compliance to [10], the Erdös sector to which a participant belongs can be regarded as implying a social type for this participant. In this case, the type of a participant changes both along time and as different networks are considered, despite the stability of the network. The association of such typology to the results reported in this present article is remarkable: the language used in each of the sectors are different, but each participant is related to all the sectors as different networks are considered. This suggests that language is modulated with respect to the network in which the participant is interacting. Moreover, hubs and intermediary participants usually have intermittent activity, and stable activity was found only in very small communities [38], which further suggests that the sensitivity to context is high for the choice of linguistic features by the speaker (more specifically, in our case, the writer).

4. Conclusions and further work

This is a first systematic exploration of the relation between topological and textual measures in human interaction networks, as far the author knows. Different textual features were scrutinized and the main result found is that the language employed in each of the Erds sectors are very distinct. Furthermore, prominent patterns were found, e.g. the peripheral participants use more nouns while hubs use more verbs, adverbs and adjectives per noun, which suggests that less connected participants contribute with content and concepts, while hubs propose actions and qualifications on them. There should be exceptions to the findings and it is a fact that we left out of the analysis more subtle linguistic features e.g. those related to low percentages (¡5%) or to small differences. Further work should address these issues and expand the analysis to include more types of networks, more topological and linguistic measures, sets of vertices (e.g. using Ratio Cut), and comparison methods. We also envision the development of interactive visualization tools to analyze the networks and their embedded texts. Other potential next steps are:

- The observation of most incident words and word types, such as words related to polarity (e.g. good and bad), food or body parts.
- Interpretation of the constant difference found from incident and existent tokens histograms, found at the end of the Supplementary document of this paper.
- Extend word class observations, e.g. to include plurals, gender, common prefixes and suffixes.
- The observation of date and time in relation to the language employed in interaction networks and to activity characteristics (e.g. dispersion of sent time along the day or weekdays).
- A careful analysis of each linguistic feature distribution which is likely to reveal multimodal outlines and other non-trivial characteristics.
- Extend the analysis of language-related measurements to the windowed approach along the timelines employed in [10].
- Tackle the same analysis on networks with languages other than English.
- The significant differences found from the texts of the sectors raises the question as to why we are not conscious of these differences. One possibility is that they are part of our instinctive and unconscious social coordination and this hypothesis might be the goal of future efforts.
- The inclusion of more sophisticated analyses of text, such as those used in the Coh-Metrix framework for text complexity [39].
- Using machine learning to quantify the separability of the Erds sectors with respect to their respective texts.

Acknowledgements

The authors acknowledge the financial support of the So Paulo State Research Foundation (FAPESP grant 2017/05838-3) and the National Council for Scientific and Technological Development (CNPq, grant 140860/2013-4).

Appendix A. Additional tables of the textual differences found in all networks

In the following tables the counting of differences of textual features among the analyzed networks are provided. These results are auxiliary for the discussion in Section 3.

Table A.13: Counts of evidence of character-related differences in the Erdös sectors in each of the analyzed networks.

synset	p.	i.	h	peaks
$\frac{spaces}{chars}$	2	0	8	2
$\frac{punct}{chars-spaces}$	11	4	1	5
$\frac{digits}{chars-spaces}$	9	7	2	10
$\frac{letters}{chars-spaces}$	0	0	3	0
$\frac{vowels}{letters}$	0	1	1	1
$\frac{uppercase}{letters}$	13	3	1	6

Appendix B. Definition of the terms of text-related measurements

The following terms are used in Sections 3.4-3.11 and Appendix Appendix A. The terms related to part-of-speech (POS) were defined directly at Tables 10 and C.24.

Appendix C. Preliminary description of Wordnet-related results

Appendix C.1. Wordnet POS tags

The observations here are somewhat consistent with those in Section 3.8: peripherals use more nouns and less verbs and adverbs. The variation is related

Table A.14: Counts of evidence of token-related differences in the Erdös sectors in each of the analyzed networks.

synset	p.	i.	h	peaks
$\frac{knownw}{tokens}$	1	0	5	1
$\frac{knownw\neq}{knownw}$	13	1	4	9
$\frac{stopw}{knownw}$	0	0	14	2
$\frac{punct}{tokens}$	10	3	1	3
$\frac{contrac}{tokens}$	0	2	15	4
$\mu(\overline{tokens})$	0	1	2	1
$\sigma(\overline{tokens})$	7	1	0	2
$\mu(\overline{knownw})$	0	0	2	0
$\sigma(\overline{knownw})$	0	0	1	1
$\mu(\overline{knownw\neq})$	0	0	0	0
$\sigma(\overline{knownw\neq})$	0	0	0	0
$\mu(\overline{stopw})$	0	0	0	0
$\sigma(\overline{stopw})$	0	0	1	0

to adjectives, which was found more frequent in the hubs. These results are illustrated in Table C.24.

Table A.15: Counts of evidence of sentence-related differences in the Erdös sectors in each of the analyzed networks.

synset	**p.**	**i.**	**h**	**peaks**
$\mu_S(chars)$	9	3	1	6
$\sigma_S(chars)$	11	6	1	9
$\mu_S(tokens)$	10	2	1	5
$\sigma_S(tokens)$	9	7	1	9
$\mu_S(knownw)$	9	3	2	6
$\sigma_S(knownw)$	11	5	2	8
$\mu_S(stopw)$	2	3	7	7
$\sigma_S(stopw)$	6	7	4	10
$\mu_S(puncts)$	13	2	1	2
$\sigma_S(puncts)$	7	8	1	8

Table A.16: Counts of evidence of message-related differences in the Erdös sectors in each of the analyzed networks.

synset	**p.**	**i.**	**h**	**peaks**
$\mu_M(sents)$	5	7	2	10
$\sigma_M(sents)$	6	8	3	14
$\mu_M(tokens)$	10	5	2	6
$\sigma_M(tokens)$	8	8	2	9
$\mu_M(knownw)$	8	5	3	7
$\sigma_M(knownw)$	10	5	3	9
$\mu_M(stopw)$	5	6	6	8
$\sigma_M(stopw)$	7	6	3	11
$\mu_M(puncts)$	12	4	2	5
$\sigma_M(puncts)$	8	9	1	10
$\mu_M(chars)$	10	5	2	6
$\sigma_M(chars)$	9	7	2	8

Table A.17: Counts of evidence of differences related to POS tags in the Erdös sectors in each of the analyzed networks.

synset	**p.**	**i.**	**h**	**peaks**
NOUN	13	1	0	1
X	4	9	5	14
ADP	0	1	4	1
DET	1	0	9	2
VERB	0	0	6	1
ADJ	1	2	6	2
ADV	0	0	17	1
PRT	1	1	9	4
PRON	0	1	11	3
NUM	8	5	3	7
CONJ	2	6	4	8

Table A.18: Counts of evidence of differences related to Wordnet POS tags in the Erdös sectors in each of the analyzed networks.

synset	**p.**	**i.**	**h**	**peaks**
N	8	1	0	1
ADJ	0	2	12	6
VERB	0	1	16	2
ADV	0	0	9	1
POS	0	0	3	1
POS!	0	1	0	1

Table A.19: Counts of evidence of differences related to Wordnet noun synset characteristics in the Erdös sectors in each of the analyzed networks.

synset	**p.**	**i.**	**h**	**peaks**
$\mu(min\,depth)$	0	0	0	0
$\sigma(min\,depth)$	1	1	2	1
$\mu(max\,depth)$	0	0	0	0
$\sigma(max\,depth)$	0	1	3	1
$\mu(holonyms)$	7	4	4	6
$\sigma(holonyms)$	3	4	7	6
$\mu(meronyms)$	8	5	3	7
$\sigma(meronyms)$	12	4	2	9
$\mu(domains)$	6	4	5	8
$\sigma(domains)$	3	1	4	3
$\mu(lemmas)$	6	0	1	2
$\sigma(lemmas)$	6	2	2	4
$\mu(hyponyms)$	1	6	6	9
$\sigma(hyponyms)$	4	6	6	11
$\mu(hypernyms)$	0	0	0	0
$\sigma(hypernyms)$	4	4	4	6

Table A.20: Counts of evidence of differences related to Wordnet adjective synset characteristics in the Erdös sectors in each of the analyzed networks.

synset	**p.**	**i.**	**h**	**peaks**
$\mu(domains)$	2	6	8	10
$\sigma(domains)$	2	4	5	7
$\mu(similar)$	1	0	7	4
$\sigma(similar)$	4	0	5	3
$\mu(lemmas)$	1	2	1	2
$\sigma(lemmas)$	6	3	4	6

Table A.21: Counts of evidence of differences related to Wordnet verb synset characteristics in the Erdös sectors in each of the analyzed networks.

synset	**p.**	**i.**	**h**	**peaks**
$\mu(min\,depth)$	2	1	1	3
$\sigma(min\,depth)$	2	1	1	2
$\mu(max\,depth)$	2	1	0	1
$\sigma(max\,depth)$	3	1	1	2
$\mu(domains)$	7	3	4	4
$\sigma(domains)$	8	3	3	5
$\mu(verb\,groups)$	0	2	3	2
$\sigma(verb\,groups)$	0	0	0	0
$\mu(lemmas)$	0	0	2	0
$\sigma(lemmas)$	1	0	3	0
$\mu(entailments)$	7	1	7	3
$\sigma(entailments)$	4	1	5	3
$\mu(hyponyms)$	1	2	6	3
$\sigma(hyponyms)$	2	3	8	6
$\mu(hypernyms)$	2	2	0	2
$\sigma(hypernyms)$	1	0	1	1

Table A.22: Counts of evidence of differences related to Wordnet adverb synset characteristics in the Erdös sectors in each of the analyzed networks.

synset	**p.**	**i.**	**h**	**peaks**
$\mu(domains)$	3	3	10	10
$\sigma(domains)$	1	4	7	7
$\mu(lemmas)$	0	0	1	1
$\sigma(lemmas)$	3	1	2	4

Table B.23: Definition of text-related measurement terms.

term	**definition**
chars	characters
punct	punctuations
digits	numerical digits
letters	alphabetical letters
uppercase	uppercase letters
knownw	known words
knownw $\neq$	different known words
contrac	contractions
stopw	stopwords
μ	mean
σ	standard deviation
$\overline{something}$	the size of *something*
$\mu_S(something)$	mean of sizes of sentences measured by counting *something*
$\sigma_S(something)$	standard deviation of sentences measured by counting *something*
$\mu_M(something)$	mean of sizes of messages measured by counting *something*
$\sigma_M(something)$	standard deviation of messages measured by counting *something*

Table C.24: Percentage of synsets with each of the POS tags used by Wordnet. The last lines give the percentage of words considered from all of the tokens (POS) and from the words with synset (POS!). The tokens not considered are punctuations, unrecognized words, words without synsets, stopwords and words for which Wordnet has no synset tagged with POS tags. Values for each Erdös sectors are in the columns **p.** for periphery, **i.** for intermediary, **h.** for hubs. TAG: 12

	g.	**p.**	**i.**	**h.**
N	58.82	59.32	61.81	49.90
ADJ	10.62	10.44	10.17	12.06
VERB	5.06	4.85	4.38	7.16
ADV	25.50	25.39	23.64	30.89
POS	33.10	32.91	32.94	33.74
POS!	92.51	93.21	91.83	93.94

Appendix C.2. Wordnet synsets characteristics

Wordnet synsets with different POS tags have different relations (to other synsets). Therefore, we made separate observations about each POS tag. In each synset we performed a count of the number of the relations (e.g. max depth, hyponyms), thus yielding a mean and variance of each of the number of relations.

- Nouns, exemplified in Table C.25.
  - Minimum and maximum depth: differences were found in the mean of minimum and maximum depth of a synset between email lists, but not once among sectors of a network. Differences between the variance of minimum and maximum depth of synsets of sectors were found mostly nonexistent or weak.
  - Holonyms: differences in the number of holonyms per word were present in $\approx$ 85% of the networks and were more incident in the lower sectors in $\approx$ 90% of the observations in which we found such differences. Differences in the variance in the number of holonyms were also found with the same regularity, but were greater in the upper sectors in $\approx$ 80% of the networks. Both mean and variance of the number of holonyms peaked in the intermediary sector in $\approx$ 50% of the observations.
  - Meronyms: differences in the number of meronyms of nouns were present in $\approx$ 90% of the networks and were more incident in the lower sectors in $\approx$ 80% of the observations in which we found such differences. Differences in the variance in the number of meronyms were found in 100% of the networks and was often strong. The variance was greater in the periphery in 66.66% and in the lower sectors in $\approx$ 90% of the observations.
  - Domain: differences in the mean and variance of the number of domains of words were found respectively in 90% and 50% of the net-

works and maximum values were found evenly distributed across sectors. Peaks were found in the intermediary sector in $\approx 50\%$ of the networks.

- Lemmas: differences in the mean and variance of the number of lemmas of words were found respectively in 40% and 55% of the networks. In $\approx 90\%$ of the cases where there was difference in the mean, the maximum number of lemmas was found in the periphery. Peaks in the intermediary sector were less often, occurring only in $\approx 35\%$ of the observations.
- Hyponyms: differences in the mean and variance of the number of hyponyms of words were found respectively in 77.77% and 88.88% of the networks. In $\approx 93\%$ of the cases where there was difference in the mean, the maximum number of hyponyms was found indistinctly in the upper sectors. In 75% of the cases where there was difference in the variance, the maximum variance was found indistinctly in the upper sectors. Peaks occurred for both mean and variance in the intermediary sector in $\approx 75\%$ of the observations.
- Hypernyms: between the sectors of all networks analyzed, we found no differences in the mean of the number of hypernyms. There were differences in the variance of the number of hypernyms of the words used by the sectors in $\approx 72\%$ of the networks. Greatest values occurred indistinctly in all sectors and peaked in the intermediary sector in $\approx 50\%$ of the observations.

Table C.25: Measures of wordnet features of nouns in each Erdös sector (**p.** for periphery, **i.** for intermediary, **h.** for hubs). TAG: 13

	g.	**p.**	**i.**	**h.**
$\mu(min\,depth)$	6.60	6.40	6.72	6.68
$\sigma(min\,depth)$	1.99	1.96	2.11	1.76
$\mu(max\,depth)$	6.95	6.64	7.15	7.09
$\sigma(max\,depth)$	2.28	2.19	2.42	2.10
$\mu(holonyms)$	0.17	0.10	0.24	0.17
$\sigma(holonyms)$	0.43	0.36	0.49	0.38
$\mu(meronyms)$	0.41	0.25	0.53	0.44
$\sigma(meronyms)$	1.38	1.12	1.47	1.57
$\mu(domains)$	0.12	0.13	0.12	0.10
$\sigma(domains)$	0.33	0.34	0.33	0.31
$\mu(lemmas)$	2.63	2.51	2.46	3.16
$\sigma(lemmas)$	2.30	2.30	2.02	2.66
$\mu(hyponyms)$	6.67	5.95	7.21	6.85
$\sigma(hyponyms)$	21.70	19.05	22.92	23.36
$\mu(hypernyms)$	1.01	1.01	1.01	1.01
$\sigma(hypernyms)$	0.10	0.11	0.10	0.10

- Adjectives, exemplified in Table C.26.
  - Domain: differences in the mean and variance of the number of domains of words were found respectively in 88.88% and 61.11% of the networks. In 87.5% of the cases where there was difference in the mean, the maximum number of domains was found indistinctly in the upper sectors. In $\approx$ 82% of the cases where there was difference in the variance, the maximum variance was found indistinctly in the upper sectors. Peaks occurred in the intermediary sector in 68.75% of the observations for the mean and in $\approx$ 54.55% of the observations for the variance.

- Similar: differences in the mean and variance of the number of similar synsets relations of adjectives were found respectively only in 44.45% and 61.11% of the networks. In ≈ 90% of the cases where there was difference in the mean, the maximum number of domains was found in the hubs sector. In ≈ 90% of the cases where there was difference in the variance, the maximum number of domains was found indistinctly in the extreme sectors. Peaks occurred in the intermediary sector in 50% of the observations for the mean and in ≈ 36.37% of the observations for the variance.
- Lemmas: differences in the mean and variance of the number of lemmas of adjectives were found respectively only in 27.78% and 72.22% of the networks. Maximum values occurred indistinctly in all sectors and peaks were found in the intermediary sector in ≈ 50% of the observed cases.

Table C.26: Measures of wordnet features of adjectives in each Erdös sector (**p.** for periphery, **i.** for intermediary, **h.** for hubs). TAG: 9

	g.	**p.**	**i.**	**h.**
$\mu(domains)$	0.05	0.06	0.05	0.06
$\sigma(domains)$	0.22	0.24	0.21	0.23
$\mu(similar)$	5.78	5.49	5.46	6.09
$\sigma(similar)$	6.78	6.45	6.55	7.00
$\mu(lemmas)$	1.65	1.71	1.63	1.66
$\sigma(lemmas)$	1.29	1.38	1.24	1.31

- Verbs, illustrated in Table C.27.
  - No significant differences were found in the mean and variance verb synset relations of minimum and maximum depth, verb groups, lemmas and hypernyms.
  - Domains and entailments: differences were often strong (i.e. > 1.5)

in both mean and variance. Due to the reduced number of verbs and the small values of mean and variance, we considered these measures as not significant.

- Hyponyms: differences in the mean and variance of the number of hyponyms of verbs were found respectively in 50% and 72.23% of the networks. In $\approx$ 90% of the cases where there was difference in the mean, the maximum number of hyponyms was found in the upper sectors (66.67% in the hubs sector). In $\approx$ 85% of the cases where there was difference in the variance, the maximum number of domains was found indistinctly in the upper sectors (61.54% in the hubs sector). Peaks occurred in the intermediary sector in $\approx$ 35% with respect to both mean and variance.

Table C.27: Measures of wordnet features of verbs in each Erdös sector (**p.** for periphery, **i.** for intermediary, **h.** for hubs). TAG: 3

	g.	**p.**	**i.**	**h.**
$\mu(min\,depth)$	1.41	1.48	1.46	1.35
$\sigma(min\,depth)$	1.41	1.42	1.46	1.37
$\mu(max\,depth)$	1.42	1.50	1.46	1.35
$\sigma(max\,depth)$	1.42	1.44	1.47	1.37
$\mu(domains)$	0.03	0.04	0.04	0.03
$\sigma(domains)$	0.18	0.19	0.19	0.16
$\mu(verb\,groups)$	0.45	0.46	0.47	0.44
$\sigma(verb\,groups)$	0.62	0.62	0.62	0.62
$\mu(lemmas)$	3.18	2.95	3.17	3.33
$\sigma(lemmas)$	2.15	2.06	2.17	2.18
$\mu(entailments)$	0.05	0.04	0.04	0.06
$\sigma(entailments)$	0.22	0.20	0.20	0.23
$\mu(hyponyms)$	14.39	11.45	15.42	15.47
$\sigma(hyponyms)$	42.12	31.44	46.49	44.58
$\mu(hypernyms)$	0.71	0.73	0.71	0.70
$\sigma(hypernyms)$	0.46	0.45	0.46	0.46

- Adverbs, exemplified in Table C.28.
  - Domains: differences in the mean and variance of the number of domains of adverbs were found respectively in $\approx 95.45\%$ and $\approx 66.67\%$ of the networks. In $\approx 82.35\%$ of the cases where there was difference in the mean, the maximum number of domains was found in the upper sectors (58.82% in the hubs sector). In $\approx 92\%$ of the cases where there was difference in the variance, the maximum number of domains was found indistinctly in the upper sectors (50% in the intermediary sector). Peaks occurred in the intermediary sector in $\approx 64.71\%$ and 75% in the mean and variance respectively.

– Lemmas: no systematic difference was found in the mean and variance of the number of lemmas of adverbs.

Table C.28: Measures of wordnet features of adverbs in each Erdös sector (**p.** for periphery, **i.** for intermediary, **h.** for hubs). TAG: 17

	g.	**p.**	**i.**	**h.**
$\mu(domains)$	0.09	0.09	0.08	0.09
$\sigma(domains)$	0.28	0.28	0.27	0.28
$\mu(lemmas)$	3.23	3.07	3.29	3.23
$\sigma(lemmas)$	2.23	2.20	2.33	2.22

Appendix C.3. Wordnet synset hypernyms

In measuring the incidence of hypernyms, significant differences were often found, but greater values occurred in all sectors. This motivated the inclusion of the tables below in which incidences are observed in all the networks at once. The drawback is that the individual values, found in the Supplementary document, may have to be investigated. The advantage is the more immediate observation of the findings with respect to all the networks analyzed. Each line holds the count of the greater incidence of each hypernym in each sector, the count of peaks in the intermediary sector, the total number of networks in which the synset was found with incidence greater than 10% in at least one sector, and the depth of the hypernym. All the words with synsets were taken into account but when measures were below 10% in all sectors, the differences were considered negligible in the corresponding network. For each POS tag, we consider synsets for which there is such significant incidence in a minimum number of networks, i.e. *total* should be above a certain threshold. This threshold was chosen to yield few synsets and some structure: total ≥ 16 for nouns; total ≥ 11 for adjectives; total ≥ 14 for verbs; total ≥ 13 for adverbs. This is surely an *ad hoc* procedure, but suited our purposes of making sense of many measurements and for a first semantic consideration of the language used by the Erdös sectors.

The analysis comprises hundreds of thousands of synsets in each POS tag. Few synsets are listed below as most of them are discarded because of low incidence.

- Noun synsets: differences in the use of nouns with physical entity hypernyms were found indistinctly in all sectors. In deeper layers, more systematic differences arise. With depth 2, hubs use more nouns related to attribute and psychological features. Lower sectors use more nouns related to measure, with 62.5% of the lists where this difference was found having greater values in the peripheral sector. Communication related nouns were found mostly in extreme sectors. With depth 3, hubs presented more nouns related to written communication, event and cognition. Peripherals showed greater use of nouns related to definite quantity. Message-related nouns often peaked at the intermediary sector. These results are shown in Table C.29.

Table C.29: Wordnet synset hypernyms from nouns in each Erdös sector.

synset	p.	i.	h	peaks	total	depth
abstraction.n.06	2	0	1	1	18	1
physical_entity.n.01	3	3	4	4	18	1
attribute.n.02	4	2	11	6	18	2
communication.n.02	7	2	5	5	18	2
causal_agent.n.01	5	2	7	4	16	2
psychological_feature.n.01	2	1	11	6	18	2
object.n.01	5	3	4	4	18	2
measure.n.02	10	5	1	6	18	2
written_communication.n.01	1	3	8	6	13	3
definite_quantity.n.01	12	4	2	6	18	3
event.n.01	2	1	11	7	17	3
person.n.01	4	2	6	5	16	3
message.n.02	7	4	7	10	18	3
whole.n.02	6	2	9	6	18	3
cognition.n.01	3	0	12	6	17	3

- Adjective synsets: the use of adjectives was found less systematic. The synsets varied greatly among lists and differences were not strong. We observed weak evidence that hubs use more adjectives related to certainty, and that the use of such adjectives peaked at the intermediary sector. Even weaker evidence was found that hubs use more adjectives related to newness. These results are shown in Table C.30.

Table C.30: Wordnet synset hypernyms from adjectives in each Erdös sector.

synset	p.	i.	h	peaks	total	depth
certain.a.02	0	3	8	4	11	1
new.a.01	2	1	4	4	9	1

- Verbs synset hypernyms of move and travel were more numerous in the peripheral sector. Verbs related to change were more common in the hubs sector. Verbs related to making had differences in the frequency of use among sectors, but had greatest incidence in all sectors (but in distinct networks). With depth 2, hubs exhibited greater use of verbs related to state and evaluate while peripherals exhibited greater use of verbs related to keeping and putting. With depth 3, in the upper sectors was found a greater use of verbs related to thinking. Hubs used more increase-related verbs. Periphery presented more verbs related to running and communication. With depth 4, lower sectors used more verbs related to informing, peripherals might be regarded as using more verbs related to recording (set in a permanent form), and hubs as using more verbs related to adding. These results are shown in Table C.31.

Table C.31: Wordnet synset hypernyms from verbs in each Erdös sector.

synset	p.	i.	h	peaks	total	depth
move.v.02	9	2	2	4	14	1
travel.v.01	10	0	1	5	15	1
change.v.02	3	1	9	5	13	1
make.v.03	5	5	4	8	16	1
use.v.01	4	0	6	2	16	1
change.v.01	1	4	8	6	15	1
state.v.01	0	3	13	5	16	2
keep.v.03	9	3	2	4	14	2
interact.v.01	8	5	4	8	18	2
evaluate.v.02	1	3	13	5	18	2
put.v.01	10	1	1	3	14	2
think.v.01	1	6	10	7	17	3
run.v.01	9	0	3	5	14	3
increase.v.01	3	3	10	6	16	3
communicate.v.02	10	4	4	8	18	3
inform.v.01	8	7	3	12	18	4
record.v.01	8	3	4	6	15	4
add.v.01	2	4	10	6	17	4

- Adverb synsets were found with particularly interesting patterns as greater use of adverbs related to possibility and stillness was found in the intermediary sector. Adverbs related to "however" and "even" were more frequent in the peripheral sector while adverbs related to "well" (good way to perform) was more used by hubs. These results are shown in Table C.32.

Table C.32: Wordnet synset hypernyms from adverbs in each Erdös sector.

synset	p.	i.	h	peaks	total	depth
however.r.01	7	2	4	8	13	1
even.r.01	7	3	5	9	16	1
possibly.r.01	1	8	5	12	14	1
well.r.01	2	2	7	6	13	1
still.r.01	2	9	1	11	13	1

References

[1] J. L. Moreno, Who shall survive?: A new approach to the problem of human interrelations., The Journal of Social Psychology 6 (1935) 388–393.

[2] M. Newman, Networks: an introduction, Oxford University Press, 2010.

[3] B. Latour, Reassembling the social. an introduction to actor-network-theory, Journal of Economic Sociology 14 (2) (2013) 73–87.

[4] C. Bird, A. Gourley, P. Devanbu, M. Gertz, A. Swaminathan, Mining email social networks, in: Proceedings of the 2006 international workshop on Mining software repositories, ACM, 2006, pp. 137–143.

[5] A. Vázquez, J. G. Oliveira, Z. Dezsö, K.-I. Goh, I. Kondor, A.-L. Barabási, Modeling bursts and heavy tails in human dynamics, Physical Review E 73 (3) (2006) 036127.

[6] B. Ball, M. E. Newman, Friendship networks and social status, arXiv preprint arXiv:1205.6822.

[7] F. V. Ordenes, B. Theodoulidis, J. Burton, T. Gruber, M. Zaki, Analyzing customer experience feedback using text mining: A linguistics-based approach, Journal of Service Research 17 (3) (2014) 278–295.

[8] V. Gupta, G. S. Lehal, et al., A survey of text mining techniques and

applications, Journal of emerging technologies in web intelligence 1 (1) (2009) 60–76.

[9] J. Perkins, Python 3 text processing with NLTK 3 cookbook, Packt Publishing Ltd, 2014.

[10] R. Fabbri, R. Fabbri, D. C. Antunes, M. M. Pisani, O. N. de Oliveira Junior, Temporal stability in human interaction networks, Physica A: Statistical Mechanics and its Applications 486 (2017) 92–105.

[11] S. Kwon, M. Cha, K. Jung, Rumor detection over varying time windows, PloS one 12 (1) (2017) e0168344.

[12] F. Jin, E. Dougherty, P. Saraf, Y. Cao, N. Ramakrishnan, Epidemiological modeling of news and rumors on twitter, in: Proceedings of the 7th Workshop on Social Network Mining and Analysis, ACM, 2013, p. 8.

[13] J. Wei, B. Bu, X. Guo, M. Gollagher, The process of crisis information dissemination: impacts of the strength of ties in social networks, Kybernetes 43 (2) (2014) 178–191.

[14] Y. Yang, K. Niu, Z. He, Exploiting the topology property of social network for rumor detection, in: 2015 12th International Joint Conference on Computer Science and Software Engineering (JCSSE), IEEE, 2015, pp. 41–46.

[15] A. Baronchelli, V. Loreto, L. Dall'Asta, A. Barrat, Bootstrapping communication in language games: Strategy, topology and all that, in: The evolution of language, World Scientific, 2006, pp. 11–18.

[16] T. Gong, J. Ke, J. W. Minett, W. S. Wang, A computational framework to simulate the coevolution of language and social structure, in: Artificial Life IX: Proceedings of the 9th International Conference on the Simulation and Synthesis of Living Systems, 2004, pp. 158–64.

[17] V. Loreto, L. Steels, Social dynamics: Emergence of language, Nature Physics 3 (11) (2007) 758.

[18] A. Baronchelli, V. Loreto, F. Tria, Language dynamics (2012).

[19] C. Castellano, S. Fortunato, V. Loreto, Statistical physics of social dynamics, Reviews of modern physics 81 (2) (2009) 591.

[20] Z. Dörnyei, A. Henry, P. D. MacIntyre, Motivational dynamics in language learning, Vol. 81, Multilingual Matters, 2014.

[21] T. Raducha, T. Gubiec, Predicting language diversity with complex networks, PloS one 13 (4) (2018) e0196593.

[22] S. Joksimović, N. Dowell, O. Poquet, V. Kovanović, D. Gašević, S. Dawson, A. C. Graesser, Exploring development of social capital in a cmooc through language and discourse, The Internet and Higher Education 36 (2018) 54–64.

[23] J. Ke, T. Gong, W. S. Wang, Language change and social networks, Communications in Computational Physics 3 (4) (2008) 935–949.

[24] D. R. Amancio, R. Fabbri, O. N. Oliveira, M. G. Nunes, L. da Fontoura Costa, Opinion discrimination using complex network features, in: Complex Networks, Springer, 2011, pp. 154–162.

[25] N. Aletras, B. P. Chamberlain, Predicting twitter user socioeconomic attributes with network and language information, in: Proceedings of the 29th on Hypertext and Social Media, ACM, 2018, pp. 20–24.

[26] A. F. Colladon, P. A. Gloor, Measuring the impact of spammers on e-mail and twitter networks, International Journal of Information Management.

[27] R. Fabbri, Topological stability and textual differentiation in human interaction networks: statistical analysis, visualization and linked data, Ph.D. thesis, Universidade de São Paulo (2017).

[28] Wikipedia contributors, Gmane — Wikipedia, the free encyclopedia, https://en.wikipedia.org/w/index.php?title=Gmane&oldid=839255062, [Online; accessed 7-July-2019] (2018).

[29] K. Marek-Spartz, P. Chesley, H. Sande, Construction of the gmane corpus for examining the diffusion of lexical innovations, WebSci.

[30] E. A. Leicht, M. E. Newman, Community structure in directed networks, Physical review letters 100 (11) (2008) 118703.

[31] M. Newman, Community detection and graph partitioning, arXiv preprint arXiv:1305.4974.

[32] S. Petrov, D. Das, R. McDonald, A universal part-of-speech tagset, arXiv preprint arXiv:1104.2086.

[33] R. Fabbri, Resumo introdutrio ao natural language toolkit (nltk), Sourceforge, https://sourceforge.net/p/labmacambira/rcpln/ci/master/tree/pln/trabNLTK/resumoNLTK.pdf?format=raw (2013).

[34] G. A. Miller, Wordnet: a lexical database for english, Communications of the ACM 38 (11) (1995) 39–41.

[35] R. Fabbri, Incidncia de letras, palavras e sentenas na obra de machado de assis, Sourceforge, http://sourceforge.net/p/labmacambira/rcpln/ci/master/tree/pln/trabLetras/resumoLetras.pdf?format=raw (2013).

[36] R. Fabbri, F. G. De León, A statistical distance derived from the kolmogorov-smirnov test: specification, reference measures (benchmarks) and example uses, arXiv preprint arXiv:1711.00761.

[37] I. Jolliffe, Principal component analysis, Springer, 2011.

[38] G. Palla, A.-L. Barabási, T. Vicsek, Quantifying social group evolution, Nature 446 (7136) (2007) 664–667.

[39] A. C. Graesser, D. S. McNamara, J. M. Kulikowich, Coh-metrix: Providing multilevel analyses of text characteristics, Educational researcher 40 (5) (2011) 223–234.