

**UNIVERSIDADE DE SÃO PAULO
INSTITUTO DE FÍSICA DE SÃO CARLOS**

Renato Fabbri

**Topological stability and textual differentiation in human
interaction networks: statistical analysis and linked data**

São Carlos

2016

Renato Fabbri

**Topological stability and textual differentiation in human
interaction networks: statistical analysis and linked data**

Thesis presented to the Graduate Program
in Physics at the Instituto de Física de São
Carlos, Universidade de São Paulo, to obtain
the degree of Doctor in Science.

Concentration area: Applied Physics
Option: Computational Physics

Supervisor: Prof. Dr. Osvaldo Novais de
Oliveira Junior

Original version

**São Carlos
2016**

É possível elaborar a ficha catalográfica em LaTeX ou incluir a fornecida pela Biblioteca. Para tanto observe a programação contida nos arquivos USPSC-modelo.tex e fichacatalografica.tex e/ou gere o arquivo fichacatalografica.pdf.

A biblioteca da sua Unidade lhe fornecerá um arquivo PDF com a ficha catalográfica definitiva, que deverá ser salvo como fichacatalografica.pdf no diretório do seu projeto.

Folha de aprovação em conformidade
com o padrão definido
pela Unidade.

No presente modelo consta como
folhadeaprovacao.pdf

This work is dedicated to God and my family, whose constant support made it possible.

ACKNOWLEDGEMENTS

I thank Prof. Dr. Osvaldo Novais de Oliveira Junior and Prof. Dr. Luciano da Fontoura Costa whose research inspired this thesis.

I thank the members of the São Carlos Physics Institute, including the instructors, the administration and the secretariat for the mindfulness and patience whenever I needed to reach them.

I thank the Brazilian Presidency and the United Nations Development Program for the partnership established with this research in the year of 2014. I thank Ricardo Poppi for his time and experience, which was of great help in this endeavor.

I thank the labMacambira.sf.net collective for numerous discussions and guidance with regards to free and digital culture. Namely, I thank Ricardo Fabbri, Vilson Vieira, Daniel Penalva, Caleb Luporini, Danilo Shiga, Antonio Pessotti and Carlos Lobo for insights and fruitful discussions.

I thank my family for the invaluable support in every step.

I thank the open source and free software communities for sharing their work, which made possible the developments presented in this thesis.

*“Call to me and I will answer you and tell you
great and unsearchable things you do not know.”*

Jeremiah 33:3

ABSTRACT

FABBRI, R. **Topological stability and textual differentiation in human interaction networks: statistical analysis and linked data.** 2016. 101p. Thesis (Doctor in Science) - Instituto de Física de São Carlos, Universidade de São Paulo, São Carlos, 2016.

This work reports on stable (or invariant) topological properties and textual differentiation in human interaction networks, with benchmarks derived from public email lists. Activity along time and topology were observed in snapshots in a timeline, and at different scales. Our analysis shows that activity is practically the same for all networks across timescales ranging from seconds to months. The principal components of the participants in the topological metrics space remain practically unchanged as different sets of messages are considered. The activity of participants follows the expected scale-free outline, thus yielding the hub, intermediary and peripheral classes of vertices by comparison against the Erdős-Rényi model. The relative sizes of these three sectors are essentially the same for all email lists and the same along time. Typically, 3-12% of the vertices are hubs, 15-45% are intermediary and 44-81% are peripheral vertices. Texts from each of such sectors are shown to be very different through direct measurements and through an adaptation of the Kolmogorov-Smirnov tests. These properties are consistent with the literature and may be general for human interaction networks, which has important implications for establishing a typology of participants based on quantitative criteria. For guiding and supporting this research, we also developed a visualization method of dynamic networks through animations. To facilitate verification and further steps in the analyses, we supply a linked data representation of data related to our results.

Keywords: Complex networks. Text mining. Pattern recognition. Statistics. Social network analysis. Typology. Data visualization. Linked data. Semantic web.

RESUMO

FABBRI, R. **Estabilidade topológica e diferenciação textual em redes de interação humana: análise estatística e dados ligados**. 2016. 101p. Tese (Doutorado em Ciências) - Instituto de Física de São Carlos, Universidade de São Paulo, São Carlos, 2016.

Este trabalho relata propriedades topológicas estáveis (ou invariantes) e diferenciação textual em redes de interação humana, com referências derivadas de listas públicas de e-mail. A atividade ao longo do tempo e a topologia foram observadas em instantâneos ao longo de uma linha do tempo e em diferentes escalas. A análise mostra que a atividade é praticamente a mesma para todas as redes em escalas temporais de segundos a meses. As componentes principais dos participantes no espaço das métricas topológicas mantêm-se praticamente inalteradas quando diferentes conjuntos de mensagens são considerados. A atividade dos participantes segue o esperado perfil livre de escala, produzindo, assim, as classes de vértices dos hubs, dos intermediários e dos periféricos em comparação com o modelo Erdős-Rényi. Os tamanhos relativos destes três sectores são essencialmente os mesmos para todas as listas de e-mail e ao longo do tempo. Normalmente, 3-12% dos vértices são hubs, 15-45% são intermediário e 44-81% são vértices periféricos. Os textos de cada um destes setores são considerados muito diferentes através de testes de Kolmogorov-Smirnov. Estas propriedades são consistentes com a literatura e podem ser gerais para redes de interação humana, o que tem implicações importantes para o estabelecimento de uma tipologia dos participantes com base em critérios quantitativos. De modo a guiar e apoiar esta pesquisa, nós também desenvolvemos um método de visualização para redes dinâmicas através de animações. Para facilitar a verificação e passos seguintes nas análises, nós fornecemos uma representação em dados ligados dos dados relacionados aos nossos resultados.

Palavras-chave: Redes complexas. Mineração de texto. Reconhecimento de padrões. Estatística. Análise de redes sociais. Tipologia. Visualização de dados. Dados ligados. Web semântica.

LIST OF FIGURES

- Figure 1 – The formation of interaction networks from exchanged messages. Each vertex represents a participant. A reply message from author B to a message from author A is regarded as evidence that B received information from A and yields a directed edge. Multiple messages add “weight” to a directed edge. Further details are given in Section 2.2.2. . 41
- Figure 2 – Classification of vertices by comparing degree distributions.¹ The binomial distribution of the Erdős-Rényi network model exhibits more intermediary vertices, while a scale-free network, associated with the power-law distribution, has more peripheral and hub vertices. 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 periphery, intermediary and hub sectors, the “Erdős sectors”. The connectivity distribution of empirical interaction networks, e.g. derived from email lists, can be sectioned by comparison against the associated binomial distribution with the same number of vertices and edges. In this figure, a snapshot of 1000 messages from CPP list yields the degree distribution of an interaction network of 98 nodes and 235 edges. A thorough explanation of the method is provided in Section 2.2.4. 44
- Figure 3 – The Versinus visualization method in use. 5% of the most connected vertices (hubs) are on the left half-period of the sinusoid. 15% of the most connected remaining vertices are on the right half-period. 80% of the least connected vertices are on the straight line, above the sinusoidal shape. White dots in the bottom with numbers keep track of node position in the overall degree ordering. Metrics blink periodically near the vertices they are related to. 49
- Figure 4 – Stability of Erdős sector sizes. Fractions of participants derived from degree and strength criteria, E_1 and E_4 described in Section 2.2.4, are both on the left. Fractions derived from the exclusivist C_1 and the inclusivist C_2 compound criteria are shown in the plots to the right. The ordinates $\overline{e_{\gamma, \phi}} = \frac{|e_{\gamma, \phi}|}{N}$ denote the fraction of participants in sector ϕ through criterion E_γ and, similarly, $\overline{c_{\delta, \phi}} = \frac{|c_{\delta, \phi}|}{N}$ denotes the fraction of participants in sector ϕ through criterion C_δ . Sections ?? and ?? of the Supporting Information bring a systematic collection of such timeline figures with all simple and compound criteria specified in Section 2.2.4, with results for networks from Facebook, Twitter and Participabr. . . . 57

Figure 5 – The first plot highlights the well-known pattern of degree versus clustering coefficient, characterized by the higher clustering coefficient of lower degree vertices. The second plot shows the greater dispersion of the symmetry-related ordinates dominant in the second principal component (PC2). This larger dispersion suggests that symmetry-related metrics are more powerful, for characterizing interaction networks than the clustering coefficient, especially for hubs and intermediary vertices. This figure reflects a snapshot of the LAU list with 1000 contiguous messages.	59
Figure 6 – A scatter plot of number of messages M versus number of participants N versus number of threads Γ for 140 email lists. Highest Γ is associated with low N . The correlation between N and Γ is negative for low values of N but positive otherwise. This negative correlation between N and Γ can also be observed in Table 1. Accordingly, for $M = 20000$ messages, this inflection of correlation was found around $N = 1500$, while CPP, LAU, LAD, MET lists present smaller networks.	61
Figure 7 – A diagram of the structure involved in the friendship networks of the Facebook snapshots. A green edge denotes an OWL existential class restriction; an inverted nip denotes an OWL universal class restriction; a full (non-dashed) edge denotes an OWL functional property axiom. Further information and complete diagrams for each provenance are given in . ²	89

LIST OF TABLES

Table 1 – Columns $date_1$ and $date_M$ have dates of first and last messages from the 20,000 messages considered 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).	38
Table 2 – The rescaled circular mean θ'_μ and the circular dispersion $\delta(z)$, described in Section 2.2.1, for different timescales. This example table was constructed using all LAD messages, and the results are the same for other lists, as shown in Section ?? of the Supporting Information document. The most uniform distribution of activity was found in seconds and minutes. Hours of the day exhibited the most concentrated activity (lowest $\delta(z)$), with mean between 2 p.m. and 3 p.m. ($\theta' = -9.61$). Weekdays, days of the month and months have mean near zero (i.e. near the beginning of the week, month and year) and high dispersion. Note that θ'_u has the dimensional unit of the corresponding time period while $\delta(z)$ is dimensionless.	53
Table 3 – Activity percentages along the hours of the day. Nearly identical distributions were observed on other social systems as shown in Section ?? of the Supporting Information document. Highest activity was observed between noon and 6pm (with 1/3 of total day activity), followed by the time period between 6pm and midnight. Around 2/3 of the activity takes place from noon to midnight but the activity peak occurs between 11 a.m. and 12 p.m. This table shows results for the activity in CPP. . . .	54
Table 4 – Activity percentages along weekdays. Higher activity was observed during workweek days, with a decrease of activity on weekend days of at least one third and at most two thirds.	54
Table 5 – Activity along the days of the month cycle. Nearly identical distributions are found in all systems as indicated in Section ?? of the Supporting Information. Although slightly higher activity rates are found in the beginning of the month, the most important feature seems to be the homogeneity made explicit by the high circular dispersion in Table 2. This specific example and empirical table correspond to the activity of the MET email list.	56

Table 6 – Activity percentages on months along the year. Activity is usually concentrated in Jun-Aug and/or in Dec-Mar, potentially due to academic calendars, vacations and end-of-year holidays. This table corresponds to activity in LAU. Similar results are shown for other lists in Section ?? of the Supporting Information document.	57
Table 7 – Distribution of activity among participants. The first column shows the percentage of messages sent by the most active participant. The column for the first quartile (Q_1) gives the minimum percentage of participants responsible for at least 25% of total messages with the actual percentage in parentheses. Similarly, the column for the first three quartiles Q_3 gives the minimum percentage of participants responsible for 75% of total messages. The last decile D_{-1} column shows the maximum percentage of participants responsible for 10% of messages.	58
Table 8 – Loadings for the 14 metrics into the principal components for the MET list, 1000 messages in 20 disjoint positions. The clustering coefficient (cc) appears as the first metric in the table, followed by 7 centrality metrics and 6 symmetry-related metrics. Note that the centrality measurements, including degrees, strength and betweenness centrality, are the most important contributors for the first principal component, while the second component is dominated by symmetry metrics. The clustering coefficient is only relevant for the third principal component. The three components have in average more than 85% of the variance. The low standard deviation σ implies that the principal components are considerably stable.	58
Table 9 – Distribution of participants, messages and threads among each Erdős 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 respecto to the total quantity (100% for g.) while μ and σ denote mean and standard deviation. TAG of list in?: 10	63
Table 10 – KS distances on size of tokens. TAG: 6	64
Table 11 – KS distances on size of known words. TAG: 1	65
Table 12 – KS distances on size of sentences. TAG: 2	65
Table 13 – KS distances on use of adjectives on sentences. TAG: 3	65
Table 14 – KS distances on use of substantives on sentences. TAG: 1	66
Table 15 – KS distances on use of punctuations on sentences. TAG: 8	66

Table 16 – c' values for substantives. Comparison of the same sector between lists, each author is an observation. See subsection 3.1.2 for discussion and directions.	66
Table 17 – c' values for adjectives. Comparison of the same sector between lists, each author is an observation. See subsection 3.1.2 for discussion and directions.	66
Table 18 – c' values for stopwords. Comparison of the same sector between lists, each author is an observation. See subsection 3.1.2 for discussion and directions.	67
Table 19 – c' values for punctuations/char. Comparison of the same sector between lists, each author is an observation. See subsection 3.1.2 for discussion and directions.	67
Table 20 – Characters in each Erdős sector (p. for periphery, i. for intermediary, h. for hubs). TAG: 6	69
Table 21 – Tokens in each Erdős sector (p. for periphery, i. for intermediary, h. for hubs). TAG: 1	70
Table 22 – Sentences sizes in each Erdős sector (p. for periphery, i. for intermediary, h. for hubs). TAG: 16	71
Table 23 – Messages sizes in each Erdős sector (p. for periphery, i. for intermediary, h. for hubs). TAG: 0	71
Table 24 – POS tags in each Erdős sector (p. for periphery, i. for intermediary, h. for hubs). Universal POS tags ³ : 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	72
Table 25 – 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 Erdős sectors are in the columns p. for periphery, i. for intermediary, h. for hubs. TAG: 12	74
Table 26 – Measures of wordnet features of nouns in each Erdős sector (p. for periphery, i. for intermediary, h. for hubs). TAG: 13	76
Table 27 – Measures of wordnet features of adjectives in each Erdős sector (p. for periphery, i. for intermediary, h. for hubs). TAG: 9	77

Table 28 – Measures of wordnet features of verbs in each Erdős sector (p. for periphery, i. for intermediary, h. for hubs). TAG: 3	78
Table 29 – Measures of wordnet features of adverbs in each Erdős sector (p. for periphery, i. for intermediary, h. for hubs). TAG: 17	78
Table 30 – Wordnet synset hypernyms from nouns in each Erdős sector.	80
Table 31 – Wordnet synset hypernyms from adjectives in each Erdős sector.	80
Table 32 – Wordnet synset hypernyms from verbs in each Erdős sector.	81
Table 33 – Wordnet synset hypernyms from adverbs in each Erdős sector.	81
Table 34 – Pierson correlation coefficient for the topological and textual measures. TAG: 9	82
Table 35 – PCA formation TAG: 11	84
Table 36 – Number of snapshots from each provenance.	87
Table 37 – Social platforms, original formats and further observations for the database.	88

LIST OF ABBREVIATIONS AND ACRONYMS

HTML	HyperText Markup Language
HTTP	HyperText Transfer Protocol
URI	Uniform Resource Identifier
RDF	Resource Description Framework
SPARQL	SPARQL Protocol and RDF Query Language
W3C	World Wide Web Consortium
LOSD	Linked Open Social Database
IRC	Internet Relay Chat
AA	Algorithmic Autoregulation
POS	Part-Of-Speech

CONTENTS

1	INTRODUCTION	27
1.1	Related knowledge	28
1.1.1	Complex networks	28
1.1.1.1	A good justification for applying the complex networks theory	29
1.1.1.2	Basic metrics	29
1.1.1.3	Basic types of networks	30
1.1.2	Text mining of social data	31
1.1.3	Visualization of static and dynamic graphs	31
1.1.4	Linked (open) data	32
1.1.4.1	RDF	33
1.1.5	Social participation	33
1.1.6	Other	33
1.2	Polysemy and synonyms	34
1.2.1	More specific terminology problems in the complex networks field	35
1.3	Historical note	35
1.4	Further considerations	35
2	MATERIALS AND METHODS	37
2.1	Core data and scripts	37
2.1.1	Linked Open Social Database for scientific benchmarking	38
2.1.1.1	Snapshots	38
2.1.1.2	Facebook data	39
2.1.1.3	Twitter data	39
2.1.1.4	IRC data	39
2.1.1.5	Email data	39
2.1.1.6	ParticipaBR data	39
2.1.1.7	Cidade Democrática data	39
2.1.1.8	AA data	39
2.1.2	Availability	39
2.2	Methods	40
2.2.1	Temporal activity statistics	40
2.2.2	Interaction networks	41
2.2.3	Topological metrics	42
2.2.4	Erdős sectioning	42
2.2.5	Principal Component Analysis of topological metrics	47
2.2.6	Evolution and audiovisualization of the networks	48

2.2.7	The Versinus graph visualization method	48
2.2.8	Textual metrics	50
2.2.8.1	About Wordnet	50
2.2.8.2	Relating text and topology	51
3	RESULTS AND DISCUSSION	53
3.0.1	Activity along time	53
3.0.2	Stable sizes of Erdös sectors	55
3.0.3	Stability of principal components	55
3.0.4	Types from Erdös sectors	59
3.0.5	Implications of the main findings	60
3.1	Text-related results and discussion	62
3.1.1	General characteristics of activity distribution among sectors	63
3.1.2	Evidence that the texts from Erdös sectors differ	63
3.1.3	What and how the texts from the sectors differs	68
3.1.4	Characters	68
3.1.5	Tokens and words	69
3.1.6	Sizes of sentences	70
3.1.7	Messages	71
3.1.8	POS tags	72
3.1.9	Wordnet-related results	72
3.1.9.1	Wordnet POS tags	73
3.1.9.2	Wordnet synsets characteristics	74
3.1.9.3	Wordnet synset hypernyms	79
3.1.10	Correlation of topological and textual metrics	82
3.1.11	Formation of principal components	83
3.1.12	Results still to be interpreted	83
3.2	Results from visualization	83
3.2.1	Useful visualization features for dynamic networks	83
3.2.2	Understanding of network properties through Versinus	85
3.2.3	Refinement of Versinus	86
3.3	Linked data results	86
3.3.1	Standardization	86
3.3.2	Data outline	87
3.3.3	Software tools	87
3.3.3.1	Triplification routines	87
3.3.3.2	Topological and textual analysis	88
3.3.3.3	Multimedia rendering	88
3.3.3.4	Migration from deprecated toolboxes	88
3.3.4	Diagrams of the data and auxiliary tables	88

3.3.5	SPARQL queries	89
3.3.6	License issues	90
3.3.7	Data-driven ontology synthesis	91
4	CONCLUSION AND FUTURE WORK	93
4.1	Textual analysis final remarks	93
4.2	Linked data final remarks	93
4.3	Future work	94
	BIBLIOGRAPHY	97

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.^{4,5} 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 humanities⁶ to computer science⁷ and physics,^{8,9} given the multidisciplinary nature of the topic. One of the approaches from an exact science perspective is to represent interaction networks as complex networks,^{8,9} with which several features of human interaction have been revealed. For example, the topology of human interaction networks exhibits a scale-free outline, which points to the existence of a small number of highly connected hubs and a large number of poorly connected nodes. The dynamics of complex networks representing human interaction has also been addressed,^{10,11} but only to a limited extent, since research is normally focused on a particular metric or task, such as accessibility or community detection.^{12,13}

There are numerous articles, books, websites and software tools about complex and social networks and about text mining in social media. There are fewer endeavours to characterize these networks beyond general features such as the scale-free aspect or to deal with text produced by social networks from the complex networks background. Research on network evolution is often restricted to network growth, in which there is a monotonic increase in the number of events.¹⁰ Network types have been discussed with regard to the number of participants, intermittence of their activity and network longevity.¹⁰ Two topologically different networks emerged from human interaction networks, depending on whether the frequency of interactions follows a generalized power law or an exponential connectivity distribution.¹⁴ In email list networks, scale-free properties were reported with $\alpha \approx 1.8$ ⁷ (as in web browsing and library loans⁸), and different linguistic traces were related to weak and strong ties.¹⁵

In this thesis, e-mail lists were chosen to represent human interaction networks, which were analyzed according to two broad aspects: the possible stability of the topology of the networks and the linguistic usage of distinct types of participants in the network. In the analysis, the topology of the networks was studied with regard to well-established topological metrics and well-established networks models. The text of the networks was studied with regard to statistics derived from the strings and from syntactics and semantics. The fact that unreciprocated edges often exceed 50% in human interaction networks¹¹ motivated the inclusion of symmetry metrics in our analysis. No correlation of topological characteristics and geographical coordinates was found,¹⁶ therefore geographical positions

were not considered in our study. Gender related behavior in mobile phone datasets was indeed reported¹⁷ but it is not relevant for the present work because email messages and addresses have no gender related metadata.¹⁸ As for the language usage in the networks, emphasis was given to statistical features of the text such as number of tokens and known words, sizes of tokens, known words, sentences and messages. Statistics were also derived from syntactics through POS tags (e.g. nouns, adjectives, verbs) and semantics through Wordnet synsets.

The remainder of this chapter provides some background knowledge relevant to the topics discussed in the thesis, especially complex networks, text mining, graph visualization and linked data. In Chapter 2, the data and methods were described while the results and discussion are in Chapter 3. Conclusions and further work are stated in Chapter 4. The appendices bring further results from this research.

1.1 Related knowledge

1.1.1 Complex networks

Although not universally accepted, it is commonplace to define a complex network to be a “graph with non-trivial topological features”.⁵ We might add to this definition that a complex network is also a large graph (even though no consensus appears to exist as to what *large* means in this context) and that it is a graph representation of a system found in natural, real or empirical system. Another way to approach the definition of “complex networks” is to define it as complex systems modeled as networks. This second definition is also useful but is even more problematic as there is no consensus of what a *complex system* is. Even so, one should keep in mind that authors often define a complex system to be a system composed with many parts in which “the whole is more than the sum of its parts”. Authors also often consider complex systems to have capabilities to “process information”, to adapt and to reproduce.¹⁹

A graph is a structure that consists of a set of objects (called vertices) and a set of binary/dual relations of the objects (called edges). A graph might be unweighted and undirected (the simplest possibility), weighted and undirected, unweighted and directed, or weighted and directed.

The most usual representations of graphs (and networks) are the matrix, list and node-edge representations. In the matrix representation, each entry a_{ij} is non-zero if i is linked to j ; entries might be other than 0 and 1 in weighted graphs; undirected graphs yield symmetric matrices. There are two common list representations of graphs, one lists each pair of vertices that are connected, the other holds a list for each vertex in which are all the vertices connected to it (a list of lists). In the node-edge representation, each node is represented as a point while each edge is represented by a line between corresponding nodes. The matrix representation is essential for algebraic reasoning and for deriving metrics

while the node-edge representation is important for illustration and intuitive guidance in characterizing the systems.

1.1.1.1 A good justification for applying the complex networks theory

The estimated number of atoms in the universe, often used as a reference of largeness, is $\approx 10^{80}$. Let us find the number of vertices needed to reach such number of possible networks. Consider the simplest case of the unweighted and undirected networks. Each edge can exist or not (i.e. it is a Bernoulli variable) and with n vertices there are at most $\binom{n}{2}$ edges. Therefore:

$$\begin{aligned} 2^{\binom{n}{2}} > 10^{80} &\Rightarrow \log_2[2^{\binom{n}{2}}] > \log_2(10^{80}) \Rightarrow \binom{n}{2} > \frac{\log_{10}(10^{80})}{\log_{10}2} \Rightarrow \\ &\Rightarrow \frac{n \cdot (n-1)}{2} > \frac{80}{\log_{10}2} \Rightarrow N > 23,5988 \end{aligned}$$

That is, with only 24 vertices we have more possible networks than the estimated number of atoms in the universe. We should also add that the number of possible networks grows very fast with the number of vertices. This is a good reason for characterizing such systems by means of paradigmatic networks and generic metrics for nodes and the network (and edges, but it is less often).

1.1.1.2 Basic metrics

Section 2.2.3 gives a mathematical account of the following metrics, which are used here for characterizing basic types of networks in the next section. Such metrics are:

- Degree k_i : number of edges linked to vertex i .
- In-degree k_i^{in} : number of edges ending at vertex i .
- Out-degree k_i^{out} : number of edges departing from vertex i .
- Strength s_i : sum of weights of all edges linked to vertex i .
- In-strength s_i^{in} : sum of weights of all edges ending at vertex i .
- Out-strength s_i^{out} : sum of weights of all edges departing from vertex i .
- Betweenness centrality bt_i : fraction of geodesics that contain vertex i .
- Clustering coefficient cc_i : fraction of pairs of neighbors of i that are linked, i.e. the standard clustering coefficient metric for undirected graphs.

In the following discussion, we also use the concept of distance between a pair of nodes, which is the number of edges between the nodes.

1.1.1.3 Basic types of networks

Complex networks are often characterized in terms of paradigmatic models. There are diverse models, but we can glimpse the background theory with the following ones²⁰:

- The Erdős-Rényi model^{*}: each pair of nodes is connected with a fixed random probability p . This model presents a characteristic degree ($n.p$ where n is the number of nodes), low clustering and low average distance between nodes.
- Spatial network, also called geographic network or geometric graph: nodes are located in a metric space and the probability that two nodes are connected is greater as the distance between nodes gets smaller. These networks present characteristic degrees, high clustering and large average distance between nodes.
- Small-world network: defined as a network where the typical distance between nodes grows with the logarithm of the number of nodes while the average clustering coefficient is not small (larger than e.g. in the Erdős-Rényi model). One method for constructing a small-world network is to start with a regular lattice in which each node is connected to k nearest neighbors. Each link is then rewired with probability p . With intermediate values of p such as $0.01 < p < 0.1$, we obtain a network with both short average distance between nodes (as in the Erdős-Rényi model) and a high average clustering coefficient (as in the spatial network). This model presents a characteristic degree.
- Scale-free networks: in which the degree distribution $p(k)$ follows a power law ($p(k) = C.k^{-\alpha}$ where C and α are constants). These networks are qualitatively characterized by the presence of a large number of poorly connected and of few highly connected hubs. Important is the absence of a characteristic degree, thus the name 'scale-free network'.
- Other networks: among important models of networks are exponential networks, networks with community structure and hybrid models.

Real networks most often exhibit scale-free and small-world properties. This is the case of most e.g. social, gene and food networks. However, one should be cautious about such statement because the networks derived from the real systems depend heavily in what is considered a node and a link, i.e. on how the system is modeled as a graph. Another noteworthy remark is that the Erdős-Rényi networks, i.e. graphs of the Erdős-Rényi model, are frequently pin-pointed as the networks with trivial topological properties. This model

^{*} This name is also used for the model in which, for a fixed number of nodes and a fixed number of links, all networks are equally likely. This is the model originally introduced by Paul Erdős and Alfréd Rényi.²¹ We choose the definition given in the main text, which is closely related to the one given in this footnote, because it is more commonly used nowadays.

is considered as a paradigmatic “complex network”, concept often defined as graphs with non-trivial topological properties, which is a contradiction. Therefore, the term complex networks is not a very well defined notion, as it occurs for the *complexity theory* in general.

1.1.2 Text mining of social data

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.²² 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.²³ Text mining tasks include document summarization, sentiment analysis and natural language processing techniques such as part of speech tagging.²⁴ Among the applications one may include social media monitoring, automated ad placement, and development of tools for semantics.²³ It is believed that applying text mining to social media can yield interesting findings in human behavior.²² Although there is no clear cut, text mining is sometimes divided into linguistic and non-linguistic.²² 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 (see Section 1.2 for a coherent distinction of the fields). 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 thesis we use both perspectives.

1.1.3 Visualization of static and dynamic graphs

Static graph visualization is achieved in many ways, most usually through the node-link (often called network diagram) and matrix representations. Representing graphs as node-link diagrams has a long tradition which goes back at least to the works of Ramon Llull in the 13th century.²⁵ To glimpse at the theory involved in visualizing networks,²⁶ we mention three aspects:

- criteria for the quality of layouts might include the number of crossing edges or the area of the drawing relative to the closest distance between two vertices.
- Layout methods are derived e.g. by placing vertices in a circular fashion, by using the eigen vectors from a worked out variant of the adjacency matrix as coordinates, or by force-based methods. For large graphs, including a number of social networks, the force-based networks are reported as useful. Therefore, we illustrate this method with the simplest model we could find in the specialized literature. Let f_a be the attraction force, f_r is the repulsion force, d is the distance between the vertices and

k is a constant. The model introduced by Fruchterman and Reingold²⁷ defines the forces as:

$$f_a = \frac{d^2}{k} \quad (1.1)$$

$$f_d = -\frac{k^2}{d} \quad (1.2)$$

On a computer software, one usually starts with a random layout and performs a number of iterations updating the position of nodes using these forces to obtain the intended force-directed layout.

- Graph drawings are often developed for specific applications e.g. in biology (as for protein and gene interactions), in social networks, in tree diagrams.

The core difference of dynamic graphs to static graphs is that vertices and edges can be added and removed over time. If we define the static graph G as $G := (V, E)$ where V are the vertices as E are the edges in G , a dynamic graph might be defined as $\Gamma := (G_1, G_2, \dots, G_n)$ where $G_i := (V_i, E_i)$ are static graphs and indices refer to a sequence of time steps (t_1, t_2, \dots, t_n) . In dynamic graph visualization most usually graphs are represented as animated diagrams or charts based on a timeline.²⁸ In this thesis we make use of node-link diagrams of both static and dynamic graphs.

1.1.4 Linked (open) data

The fields of social network analysis and complex networks are widely researched. However, there is a lack of open datasets for benchmarking results, especially associated with the complex networks field, yielding diverse results from poorly related sources. Recently, a myriad of results have been reported which are based on diverse datasets most often not accessible to researchers other than the publishing authors. In this thesis we present resources for having open databases to provide the scientific community with a friendly and common repertoire. We chose to use the linked data technology and follow W3C best practices for publishing data.

Linked data refers to data published in the web in such a way that it is machine readable and complies with a set of best practices. The web of data is constructed with documents on the web such as the web of HTML documents. In practice, the idea of linked data can be summarized by 1) the use of RDF to publish data on the web and 2) the use of RDF links to interlink data from different sources. The web is expected to be interconnected and to grow by the systematic application of four steps²⁹:

- Use URIs to identify things.³⁰
- Use HTTP URIs.

- Provide useful information when an URI is accessed via HTTP.
- Provide other URIs in the description of resources so human and machine agents can perform discovery.

The Linked Open Data³¹ builds an ever growing cloud of data, the global data space, which is usually conceived as centered around the DBPedia, a linked data representation of data from Wikipedia.^{32,33}

1.1.4.1 RDF

The Resource Description Framework (RDF), a W3C recommendation, is a model for data interchange. It is based on the idea of making statements about resources in the form of triples, i.e. expressions in the form “subject - predicate - object”. RDF can be serialized in several file formats, including RDF/XML, Turtle and Manchester, all of which, in essence, represent a labeled and directed multi-graph. RDF may be stored in a type of database referred to as a triplestore.³⁴

As an example of an RDF statement, the following triple in the Turtle format asserts that “the paper has color white”:

```
http://example.org/Things#Paper http://example.org/hasColor
http://example.org/Colors#White .
```

Integration and uniformity of access is obtained through linked data representation, as explained in Section 3.3.5.

1.1.5 Social participation

A significant share of our endeavor was oriented towards social participation, i.e. to facilitate civil engagement in a community, most significantly in State affairs. More concretely, we published data from a social participation federal portal,² applied complex networks and text mining criteria for resources recommendation^{35–37} and proposed a ranking algorithm for voted proposals in another federal participation portal.³⁸ Such works were performed within a United Nations Development Program consulting contract, in partnership with the Brazilian Presidency of the Republic and published publicly mainly as technical reports.^{35–37} This aspect of our research was important for maturing topics and understanding the extent to which they are applied in pragmatic contexts. It is left mostly as subsidiary documents in this thesis to allow for an easy public access.

1.1.6 Other

Given the multidisciplinary condition of our work and of the implied topics, many other fields of knowledge could be further explored in this introduction or the methods chapter. To name just a few of the most directly related fields: statistics, principal

component analysis, big data, social network analysis, social media mining, mathematical sociology, datasets, free culture, open source software, computer programming.

Of particular relevance are the typologies for human personality, such as the ones derived from the Myers-Briggs type indicator³⁹ and from authoritarian personality⁴⁰ theories, because we present a new typology of human participants in social networks in Section 3.0.4. Another topic we should highlight is what we called “anthropological physics”^{41,42}: the observation of natural/physical laws in human social systems. The term should not be confused with physical anthropology, which is a synonym for biological anthropology, a subfield of anthropology concerned with the evolution of humans.⁴³

1.2 Polysemy and synonyms

In the context of complex networks, the words *network* and *graph* are often used interchangeably, although the word *graph* might refer to the mathematical structure of vertices and edges and the word *network* might refer to the real system being represented as a graph or to the graph obtained by means of representing a real system. Furthermore, the word *graph* can be used to refer to a *graph of a function* (mathematics) or to an abstract datatype (computer science). This parallelism between network and graphs also apply to network visualization and graph visualization. One might add here the term *graph drawing*, another synonym for the visualization of graphs, although the term seems to be more traditional in relation to the achievement of node-edge network diagrams. Evolutionary graph visualization or evolutionary network visualization are examples of variants of *dynamic graph visualization*. The nomenclature of vertices and edges varies widely among interested fields (mathematics, physics, biology, sociology, etc). A vertex might be called e.g. a node, a point, an agent, an actor, a participant. An edge might be called e.g. a link, a bond, a relation, a tie, a connection.

The terms *text mining*, *natural language processing* and *computational linguistics* are often used for similar endeavors. A distinction might be made in that text mining refers to data mining of text, natural language processing is concerned with the interactions between the computer and the human natural languages, and computational linguistics aims for statistical or rule-based modeling of natural language from a computational perspective. Such fields are multidisciplinary and there is no sharp distinction between them.

Examined as fields of knowledge, the *linked data* and the *semantic web* terms are often used without distinction. Tim Berners-Lee coined both terms: the semantic web was conceived as a web of data that can be processed by machines,⁴⁴ the expression linked data appeared in a 2006 design note about the Semantic Web project²⁹ and refers to structured data that emphasizes interlinking and usefulness through semantic queries.

Social participation, social involvement and *social engagement* are synonyms that refer to the participation of an individual or group in a community or society. In Brazilian Portuguese, *controle social* can refer to the antagonist concepts of social participation or of a social control (played by the State or companies in the civil society).

1.2.1 More specific terminology problems in the complex networks field

Given that this thesis involves multidisciplinary and new knowledge, it might be of no surprise that the nomenclature is not very well defined. Here we pin-point some more specific conflicts that arise in the literature of complex networks to both exemplify this issue and to avoid some problems in interpreting the methods and results in this thesis:

- The *hubs* are, by the usual definition, the more connected vertices. In the context of the HITS (Hyperlink-Induced Topic Search) algorithm, for attributing centrality to vertices, most traditionally to web pages, the hubs are the vertices with greater out-degree (greater in-degree yield *authorities*).
- In some contexts, the center of network is the collection of vertices whose maximum distance to other vertices is the radius (i.e. the minimum maximum difference between vertices). In the same framework, the periphery of a network is the collection of vertices whose maximum distance to other vertices is the diameter (i.e. the maximum distance between vertices). By extension, the intermediary might be regarded as the set of vertices that are not in the center or the periphery. These definitions yield fractions of members that do not agree with the literature with respect to hubs, intermediary and periphery. We present a suitable method for deriving such classification, in the sense that it fits the literature prediction, in Section 2.2.4.
- Lace, loop, selfloop and autoloop are terms used to designate an edge from a vertex to itself.

1.3 Historical note

The knowledge fields involved in this thesis are very recent. To point just the main areas, complex networks has emerged in the final years of the 1990s decade⁵; text mining first workshops were held in 1999⁴⁵; as an independent field, graph drawing arose in the 1990s²⁸; the term linked data was coined in 2006.²⁹

1.4 Further considerations

An initial proposal of this research was to enable the use of complex and social networks scientific knowledge by the participant of the networks. This motivated the open software, data and texts produced, conferences attended, and the endeavors with the

United Nations, Brazilian Presidency and civil parties. As this was a practical goal, we found by hands-on processes that many fields are deeply related to the subject, which reflected in the number of fields tackled in this thesis and related documents.^{2,35-38,41,42,46,47} Furthermore, we understand that the open software, texts, videos and processes provided by our work contributes for the popularization of the knowledge and technologies implied by the empowerment of civil individuals and groups through the management of the networks in which they exist.

2 RESULTS AND DISCUSSION

2.0.1 Activity along time

Regular patterns of activity were observed along time in the scales of seconds, minutes, hours, days and months. Histograms in each of the time scales were computed as were circular average and dispersion values, and the results are given in Tables 2-6. For example, uniform activity is found with respect to seconds, minutes and days of the months. Weekend days exhibit about half the activity of regular weekdays, and there is a peak of activity between 11am and noon.

In the scales of seconds and minutes, activity is uniform, with the messages being slightly more evenly distributed in all lists than in simulations with the uniform distribution*. In the networks, $\frac{\min(\text{incidence})}{\max(\text{incidence})} \in (0.784, .794)$ while simulations reach these values but have on average more discrepant higher and lower peaks, i.e. if $\xi = \frac{\min(\text{incidence}')}{\max(\text{incidence}')}$ than $\mu_\xi = 0.7741$ and $\sigma_\xi = 0.02619$. Therefore, the incidence of messages at each second of a minute and at each minute of an hour was considered uniform. In these cases, the circular dispersion is maximized and the mean has little meaning as indicated in Table 2. As for the hours of the day, an abrupt peak is found between 11am and 12pm with the most active period being the afternoon, with one third of total daily activity, and two

* Numpy version 1.8.2, “random.randint” function, was used for simulations, algorithms in <https://github.com/ttm/percolation>.

Table 1: The rescaled circular mean θ'_μ and the circular dispersion $\delta(z)$, described in Section 2.2.1, for different timescales. This example table was constructed using all LAD messages, and the results are the same for other lists, as shown in Section ?? of the Supporting Information document. The most uniform distribution of activity was found in seconds and minutes. Hours of the day exhibited the most concentrated activity (lowest $\delta(z)$), with mean between 2 p.m. and 3 p.m. ($\theta' = -9.61$). Weekdays, days of the month and months have mean near zero (i.e. near the beginning of the week, month and year) and high dispersion. Note that θ'_u has the dimensional unit of the corresponding time period while $\delta(z)$ is dimensionless.

scale	mean θ'_μ	dispersion $\delta(z)$
seconds	-//-	9070.17
minutes	-//-	205489.40
hours	-9.61	4.36
weekdays	-0.03	29.28
month days	-2.65	2657.77
months	-0.56	44.00

Source: Prepared by the authors.

Table 2: Activity percentages along the hours of the day. Nearly identical distributions were observed on other social systems as shown in Section ?? of the Supporting Information document. Highest activity was observed between noon and 6pm (with 1/3 of total day activity), followed by the time period between 6pm and midnight. Around 2/3 of the activity takes place from noon to midnight but the activity peak occurs between 11 a.m. and 12 p.m. This table shows results for the activity in CPP.

	1h	2h	3h	4h	6h	12h
0h	3.66	6.42	8.20	9.30	10.67	33.76
1h	2.76					
2h	1.79	2.88	2.47	3.44	23.09	
3h	1.10					
4h	0.68	1.37	4.35	21.03		
5h	0.69					
6h	0.83	2.07	18.75	25.05		
7h	1.24					
8h	2.28	6.80	21.03	23.60	66.24	
9h	4.52					
10h	6.62	14.23	18.95	37.63		
11h	7.61					
12h	6.44	12.48	18.68	28.61		
13h	6.04					
14h	6.47	12.57	15.88	17.59		
15h	6.10					
16h	6.22	12.58	12.73	17.59		
17h	6.36					
18h	6.01	11.02	12.73	17.59		
19h	5.02					
20h	4.85	9.23	12.73	17.59		
21h	4.38					
22h	4.06	8.36	12.73	17.59		
23h	4.30					

Source: Prepared by the authors.

Table 3: Activity percentages along weekdays. Higher activity was observed during work-week days, with a decrease of activity on weekend days of at least one third and at most two thirds.

	Mon	Tue	Wed	Thu	Fri	Sat	Sun
LAU	15.71	15.81	15.88	16.43	15.14	10.13	10.91
LAD	14.92	17.75	17.01	15.41	14.21	10.40	10.31
MET	17.53	17.54	16.43	17.06	17.46	7.92	6.06
CPP	17.06	17.43	17.61	17.13	16.30	6.81	7.67

Source: Prepared by the authors.

thirds of activity are allocated in the second 12h of each day. Days of the week revealed a decrease between one third and two thirds of activity on weekends. Days of the month were regarded as homogeneous with an inconclusive slight tendency of the first week to be more active. Months of the year revealed patterns matching usual work and academic calendars. The time period examined here was not sufficient for the analysis of activity along the years. These patterns are exemplified in Tables 3-6.

2.0.2 Stable sizes of Erdős sectors

The distribution of vertices in the hub, intermediary and periphery Erdős sectors is remarkably stable along time if the snapshots hold 200 or more messages, as it is clear in Figure 4 and in Section ?? of the Appendix. Activity is highly concentrated on the hubs, while a very large number of peripheral vertices contribute to only a fraction of the activity. This is expected for a system with a scale-free profile, as confirmed with the distribution of activity among participants in Table 7.

Typically, $[3\% - 12\%]$ of the vertices are hubs, $[15\% - 45\%]$ are intermediary and $[44\% - 81\%]$ are peripheral, which is consistent with other studies.⁶⁷ These results hold for the total, in and out degrees and strengths. Stable sizes are also observed for 100 or less messages if the classification of the three sectors is performed with one of the compound criteria established in Section 2.2.4. The networks often hold this basic structure with as few as 10-50 messages, i.e. concentration of activity and the abundance of low-activity participants take place even with very few messages, which is highlighted in Section ?? of the Appendix. A minimum window size for the observation of more general properties might be inferred by monitoring both the giant component and the degeneration of the Erdős sectors.

In order to support the generality of these findings, we list the Erdős sector sizes of 12 networks from Facebook, Twitter and Participabr in Table ?? of the Appendix. The fractions of hubs, intermediary and peripheral nodes are essentially the same as for the email list networks but with exceptions and a greater variability.

2.0.3 Stability of principal components

The principal components of the participants are very stable in the topological space, i.e. in the space of network metrics. Table 8 exemplifies the formation of principal components by providing the averages over non-overlapped activity snapshots of a network. The most important result of this application of PCA, the stability of principal components, is underpinned by the very small dispersion of the contribution of each metric to each principal component.

The first principal component is an average of centrality metrics: degrees, strengths and betweenness centrality. On one hand, the similar relevance of all centrality metrics is

Table 4: Activity along the days of the month cycle. Nearly identical distributions are found in all systems as indicated in Section ?? of the Supporting Information. Although slightly higher activity rates are found in the beginning of the month, the most important feature seems to be the homogeneity made explicit by the high circular dispersion in Table 2. This specific example and empirical table correspond to the activity of the MET email list.

	1 day	5	10	15 days
1	3.05	18.25	35.24	50.96
2	3.38			
3	3.62			
4	4.25			
5	3.94			
6	3.73	16.98		
7	3.17			
8	3.26			
9	3.56			
10	3.26			
11	3.81	15.73	31.98	49.04
12	2.91			
13	3.30			
14	2.75			
15	2.95			
16	3.36	16.25		
17	3.16			
18	3.44			
19	3.36			
20	2.93			
21	3.20	15.79	32.78	
22	3.11			
23	3.60			
24	2.74			
25	3.13			
26	3.13	16.99		
27	3.07			
28	3.61			
29	3.60			
30	3.57			

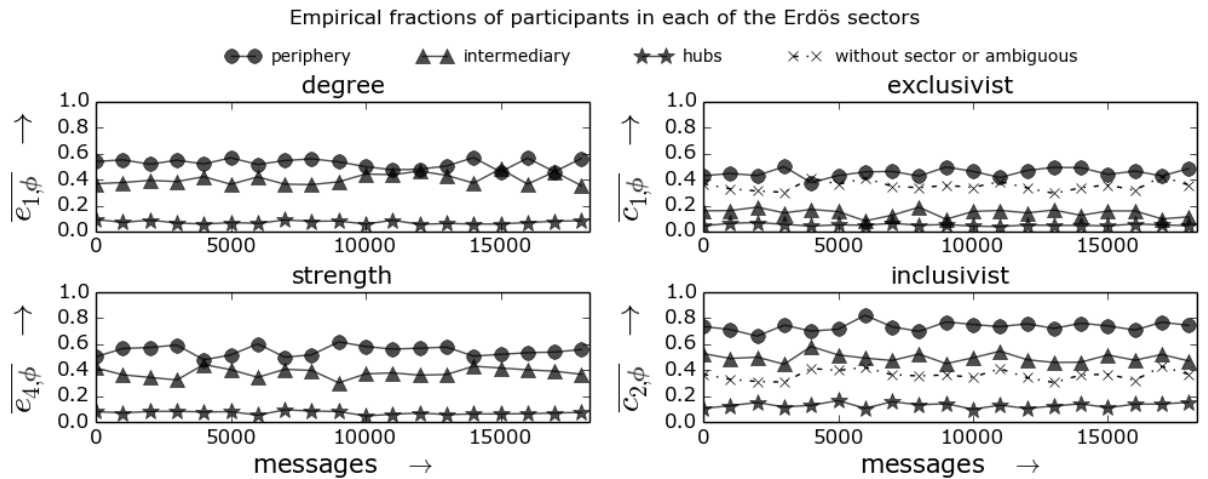
Source: Prepared by the authors.

Table 5: Activity percentages on months along the year. Activity is usually concentrated in Jun-Aug and/or in Dec-Mar, potentially due to academic calendars, vacations and end-of-year holidays. This table corresponds to activity in LAU. Similar results are shown for other lists in Section ?? of the Supporting Information document.

	m.	b.	t.	q.	s.
Jan	10.22	19.56	28.24	35.09	49.16
Fev	9.34				
Mar	8.67	15.53	20.93	30.36	
Apr	6.86				
Mai	7.28	14.07	24.47	34.55	50.84
Jun	6.80				
Jul	8.97	16.29	26.36		
Ago	7.32				
Set	8.18	16.25			
Out	8.06				
Nov	7.64	18.30			
Dez	10.66				

Source: Prepared by the authors.

Figure 1: Stability of Erdős sector sizes. Fractions of participants derived from degree and strength criteria, E_1 and E_4 described in Section 2.2.4, are both on the left. Fractions derived from the exclusivist C_1 and the inclusivist C_2 compound criteria are shown in the plots to the right. The ordinates $\overline{e}_{\gamma,\phi} = \frac{|e_{\gamma,\phi}|}{N}$ denote the fraction of participants in sector ϕ through criterion E_γ and, similarly, $\overline{c}_{\delta,\phi} = \frac{|c_{\delta,\phi}|}{N}$ denotes the fraction of participants in sector ϕ through criterion C_δ . Sections ?? and ?? of the Supporting Information bring a systematic collection of such timeline figures with all simple and compound criteria specified in Section 2.2.4, with results for networks from Facebook, Twitter and Participabr.



Source: Prepared by the authors.

Table 6: Distribution of activity among participants. The first column shows the percentage of messages sent by the most active participant. The column for the first quartile (Q_1) gives the minimum percentage of participants responsible for at least 25% of total messages with the actual percentage in parentheses. Similarly, the column for the first three quartiles Q_3 gives the minimum percentage of participants responsible for 75% of total messages. The last decile D_{-1} column shows the maximum percentage of participants responsible for 10% of messages.

list	hub	Q_1	Q_3	D_{-1}
LAU	2.78	1.19 (26.35%)	13.12 (75.17%)	67.32 (-10.02%)
LAD	4.00	1.03 (26.64%)	11.91 (75.18%)	71.14 (-10.03%)
MET	11.14	1.02 (34.07%)	8.54 (75.64%)	80.49 (-10.02%)
CPP	14.41	0.29 (33.24%)	4.18 (75.46%)	83.65 (-10.04%)

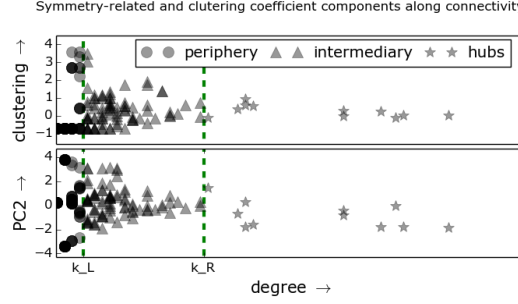
Source: Prepared by the authors.

Table 7: Loadings for the 14 metrics into the principal components for the MET list, 1000 messages in 20 disjoint positions. The clustering coefficient (cc) appears as the first metric in the table, followed by 7 centrality metrics and 6 symmetry-related metrics. Note that the centrality measurements, including degrees, strength and betweenness centrality, are the most important contributors for the first principal component, while the second component is dominated by symmetry metrics. The clustering coefficient is only relevant for the third principal component. The three components have in average more than 85% of the variance. The low standard deviation σ implies that the principal components are considerably stable.

	PC1		PC2		PC3	
	μ	σ	μ	σ	μ	σ
<i>cc</i>	0.89	0.59	1.93	1.33	21.22	2.97
<i>s</i>	11.71	0.57	2.97	0.82	2.45	0.72
<i>sⁱⁿ</i>	11.68	0.58	2.37	0.91	3.08	0.78
<i>s^{out}</i>	11.49	0.61	3.63	0.79	1.61	0.88
<i>k</i>	11.93	0.54	2.58	0.70	0.52	0.44
<i>kⁱⁿ</i>	11.93	0.52	1.19	0.88	1.41	0.71
<i>k^{out}</i>	11.57	0.61	4.34	0.70	0.98	0.66
<i>bt</i>	11.37	0.55	2.44	0.84	1.37	0.77
<i>asy</i>	3.14	0.98	18.52	1.97	2.46	1.69
μ^{asy}	3.32	0.99	18.23	2.01	2.80	1.82
σ^{asy}	4.91	0.59	2.44	1.47	26.84	3.06
<i>dis</i>	2.94	0.88	18.50	1.92	3.06	1.98
μ^{dis}	2.55	0.89	18.12	1.85	1.57	1.32
σ^{dis}	0.57	0.33	2.74	1.63	30.61	2.66
λ	49.56	1.16	27.14	0.54	13.25	0.95

Source: Prepared by the authors.

Figure 2: The first plot highlights the well-known pattern of degree versus clustering coefficient, characterized by the higher clustering coefficient of lower degree vertices. The second plot shows the greater dispersion of the symmetry-related ordinates dominant in the second principal component (PC2). This larger dispersion suggests that symmetry-related metrics are more powerful, for characterizing interaction networks than the clustering coefficient, especially for hubs and intermediary vertices. This figure reflects a snapshot of the LAU list with 1000 contiguous messages.



Source: Prepared by the authors.

not surprising since they are highly correlated, e.g. degree and strength have Spearman correlation coefficient $\in [0.95, 1]$ and Pearson coefficient $\in [0.85, 1]$ for window sizes greater than a thousand messages. On the other hand, each of these metrics is related to a different participation characteristic, and their equal relevance for variability, as measured by the principal component, is noticeable. Also, this suggests that these centrality metrics are equally adequate for characterizing the networks and the participants.

According to Table 8 and Figure 5, dispersion is larger in symmetry-related metrics than in clustering coefficient. We conclude that the symmetry metrics are more powerful, in terms of dispersion in the topological metrics space, in characterizing interaction networks and their participants, than the clustering coefficient, especially for hubs and intermediary vertices (peripheral vertices have larger dispersion with regard to the clustering coefficient). Interestingly, the clustering coefficient is always combined with the standard deviation of the asymmetry and disequilibrium of edges σ^{asy} and σ^{dis} in the third principal component.

Similar results are presented in Sections ?? and ?? of the Appendix for other email lists and interaction networks. A larger variability was found for the latter networks, which motivated the use of interaction networks derived from email lists for benchmarking.

2.0.4 Types from Erdős sectors

Assigning a type to a participant raises important issues about the scientific cannon for human types and the potential for stigmatization and prejudice. The Erdős 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. Therefore, the potential for prejudice of such participant typology is attenuated.⁴⁰ In other words, an individual is a hub in a number of networks and peripheral in other networks, and even within the same network he/she most probably changes type along time.⁶¹

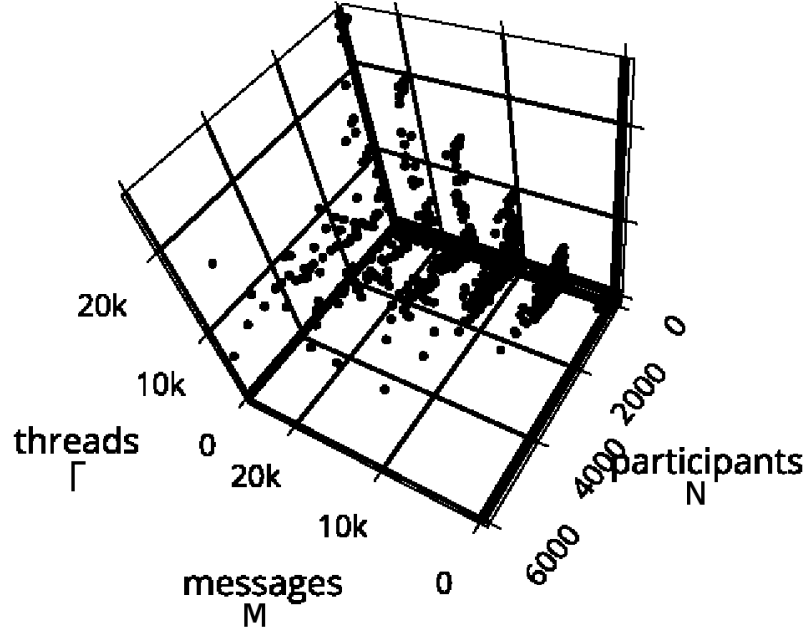
The importance of this issue can be grasped by the consideration of static types derived from quantitative criteria. For example, in email lists with a small number of participants, the number of threads has a negative correlation with the number of participants. When the number of participants exceeds a threshold, the number of threads has a positive correlation with the number of participants. This finding is illustrated in Figure 6 and can also be observed in Table 1. The assignment of types to individuals, in this latter case, has more potential for prejudice because the derived participant type is static and one fails to acknowledge that human individuals are not immutable entities.

Further observations regarding the Erdős sectors and the implicit participant types were made, which are consistent with the literature¹⁰: 1) hubs and intermediary participants usually have intermittent activity, and stable activity was found only in smaller communities. For instance, the MET list had stable hubs while LAU, LAD and CPP exhibited intermittent hubs. 2) Network structure seems to be most influenced by the activity of intermediary participants as they have less extreme roles than hubs and peripheral participants and can therefore connect to the sectors and other participants in a more selective and explicit manner.

2.0.5 Implications of the main findings

The findings reported in this thesis arose from an exploratory procedure to visually inspect the networks and to analyze considerable amounts of interaction networks data. While this procedure has certainly an ad hoc nature, the statistics in the data are sufficiently robust for important features from these interaction networks to be extracted. Temporal stability, in the sense that interaction networks could be considered as stationary time series, is the most important feature. Also relevant is the significant stability found on the principal components, on the fraction of participants in each Erdős Sector and on the activity along different timescales. In fact, these findings confirm our initial hypothesis - based on the literature⁵ - that interaction networks should exhibit some stability traces. The potential generality of these findings is suggested by the analysis of networks derived from diverse systems, with interaction networks from public email lists serving as proper benchmarks. Indeed, with such benchmarks one can compare any social network system. Furthermore, this analysis enables us to establish an outline of human interaction networks. It takes the hub, intermediary and periphery sectors out of the scientific folklore and into classes drawn from quantitative criteria. It enables the conception of non-static human types derived from natural properties.

Figure 3: A scatter plot of number of messages M versus number of participants N versus number of threads Γ for 140 email lists. Highest Γ is associated with low N . The correlation between N and Γ is negative for low values of N but positive otherwise. This negative correlation between N and Γ can also be observed in Table 1. Accordingly, for $M = 20000$ messages, this inflection of correlation was found around $N = 1500$, while CPP, LAU, LAD, MET lists present smaller networks.



Source: Prepared by the authors.

We envisage that the knowledge generated in the analysis may be exploited in applications where the type of each participant and the relative proportion of participants in each sector can be useful metadata. Just by way of illustration, this could be applied in semantic web initiatives, given that the Erdős sectorialization is static in a given snapshot. These results are also useful for classifying resources, e.g. in social media, and for resources recommendation to users.³⁷ Finally, the knowledge acquired with a quantitative treatment of the whole data may help guide the creation through collective processes of documents to assist in participatory democracy.

Perhaps the most outreaching implications are related to sociological consequences. The results expose a classification of human individuals which is directly related to the concentration of wealth and based on natural laws. The derived human typology changes over different systems and over time in the same system, which implies a negation of the absolute concentration of wealth. Such concentration exists but changes across different wealth criteria and with time. Also, the hubs stand out as dedicated, sometimes enslaved, components of the social system. The peripheral participants have very limited interaction with the network. This suggests that intermediary participants tend to dictate structure,

legitimate the hubs and stand out as authorities.

With regard to the limitations of our study, one should emphasize that not all types of human interaction networks were analyzed. Therefore, the plausible generalization of properties has to be treated with caution, as a natural tendency of such systems and not as a rule. Also, the stable properties in the networks were not explored to the limit, which leaves many open questions. For example, what are the maximum and minimum sizes of the networks for which they hold? What is the outcome of PCA when more metrics are considered? What is the granularity in which the activity along the timescales is preserved? Do the findings reported also apply to other systems, beyond human networks?

2.1 Text-related results and discussion

The most important result of including textual metrics in our analysis is the statistical evidence of extreme differentiation of each Erdős sector with respect to the texts produced. This conclusion can be reached by observing the differences in the measurements of textual features in each sector, but is reached with greater theoretical background from the adaptation of the Kolmogorov-Smirnov test we presented in Section 2.2.8.2. Other relevant results are:

- the achievement of references for the amount of nouns, adverbs, sizes of words, depth of Wordnet synsets and other linguistic traces, used in social networks. We did not find in literature any indication for such values and understood useful to acknowledge e.g. that about 15% of the characters are spaces and more than 50% of tokens are nouns. These values are available in⁷ and not in the body of the text, since we focus in this thesis in the evidence that the texts from distinct sectors differ.
- Indicatives of what is different in the texts produced by each of the Erdős sectors. 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 reaches 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 and figures of^{7,78} and the Conclusions chapter. We illustrate with just one table of each kind, and from networks obtained with 2000 messages. In⁷ we display tables for various networks. In the same document, we relate the email lists to each numerical TAG in the tables of the following sections. The scale of 1000 and 2000 messages was chosen for deriving results, as the networks in this scale are found with stable topological structure, as exposed in Section 3.0.2. The findings

with 1000 messages are the same as with 2000 messages, and there are many tables. This motivated the exclusion of the tables with measurements in networks of 1000 messages.

2.1.1 General characteristics of activity distribution among sectors

Before we dig into the findings derived from text-related measures, let us look once more at the general structure of these networks, now giving emphasis on the activity of the sectors. In almost all our observations, the peripheral sector is responsible for starting most of the discussion threads, i.e. messages to the list which are not replies. This is surprising since the peripheral sector is responsible for fewer 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. These assertions are condensed in Table 9. Less often, the intermediary sector is responsible for the greatest number of messages and of threads. Also meaningful is that the hubs sector is responsible for most of the messages, which is not completely obvious: hub participants are far more active but way less numbered. 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.

Table 8: Distribution of participants, messages and threads among each Erdős 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 respecto to the total quantity (100% for g.) while μ and σ denote mean and standard deviation. TAG of list in?: 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

Source: Prepared by the authors.

2.1.2 Evidence that the texts from Erdős sectors differ

Results from our adaptation of the Kolmogorov-Smirnov test (see Section 2.2.8.2) present substantial evidence that the texts produced by each sector are different. Tables 10-

19 illustrate three results:

- There is statistical evidence that the textual production of the Erdős sectors are different. This can be noticed from the high values of c' on these tables, beyond reference values used for the acceptance of the null hypothesis. Also, we regarded as non-negligible the values often above 0.1 for the Kolmogorov-Smirnov statistic (the maximum difference between the cumulative distributions), which we recognized as relevant for assuming differences in the underlying distributions in our study (see Section 2.2.8.2).
- Intermediary sectors sometimes exhibit greater differences from periphery and hubs than these extreme sectors between themselves (Tables 10 and 14). This differentiation of the three sectors is an indicative that the Erdős Sectioning described in Section 2.2.4 reveals meaningful sectors of the networks.
- Evidence of differences between sectors on the same network (Tables 10-??) is often greater than differences between the same sector from distinct lists (Tables 16-19).

Table 9: KS distances on size of tokens. TAG: 6

	g.	p.	i.	h.
g.	0.000	4.327	17.168	7.851
a	0.000	0.014	0.115	0.044
p.	4.327	0.000	18.907	7.833
	0.014	0.000	0.129	0.045
i.	17.168	18.907	0.000	15.540
	0.115	0.129	0.000	0.129
h.	7.851	7.833	15.540	0.000
	0.044	0.045	0.129	0.000

Source: Prepared by the authors.

Table 10: KS distances on size of known words. TAG: 1

	g.	p.	i.	h.
g.	0.000 0.000	5.904 0.043	5.264 0.040	5.549 0.150
p.	5.904 0.043	0.000 0.000	9.547 0.083	7.073 0.193
i.	5.264 0.040	9.547 0.083	0.000 0.000	4.058 0.111
h.	5.549 0.150	7.073 0.193	4.058 0.111	0.000 0.000

Source: Prepared by the authors.

Table 11: KS distances on size of sentences. TAG: 2

	g.	p.	i.	h.
g.	0.000 0.000	0.733 0.020	2.077 0.038	2.834 0.059
p.	0.733 0.020	0.000 0.000	1.642 0.048	2.589 0.080
i.	2.077 0.038	1.642 0.048	0.000 0.000	4.139 0.097
h.	2.834 0.059	2.589 0.080	4.139 0.097	0.000 0.000

Source: Prepared by the authors.

Table 12: KS distances on use of adjectives on sentences. TAG: 3

	g.	p.	i.	h.
g.	0.000 0.000	0.461 0.011	0.564 0.010	0.617 0.010
p.	0.461 0.011	0.000 0.000	0.385 0.011	0.800 0.021
i.	0.564 0.010	0.385 0.011	0.000 0.000	0.986 0.020
h.	0.617 0.010	0.800 0.021	0.986 0.020	0.000 0.000

Source: Prepared by the authors.

Table 13: KS distances on use of substantives on sentences. TAG: 1

	g.	p.	i.	h.
g.	0.000 0.000	0.642 0.023	1.791 0.067	6.936 0.537
p.	0.642 0.023	0.000 0.000	1.007 0.044	6.970 0.560
i.	1.791 0.067	1.007 0.044	0.000 0.000	7.510 0.607
h.	6.936 0.537	6.970 0.560	7.510 0.607	0.000 0.000

Source: Prepared by the authors.

Table 14: KS distances on use of punctuations on sentences. TAG: 8

	g.	p.	i.	h.
g.	0.000 0.000	1.380 0.039	3.583 0.069	2.894 0.046
p.	1.380 0.039	0.000 0.000	1.718 0.054	2.871 0.085
i.	3.583 0.069	1.718 0.054	0.000 0.000	5.398 0.114
h.	2.894 0.046	2.871 0.085	5.398 0.114	0.000 0.000

Source: Prepared by the authors.

Table 15: c' values for substantives. Comparison of the same sector between lists, each author is an observation. See subsection 3.1.2 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

Source: Prepared by the authors.

Table 16: c' values for adjectives. Comparison of the same sector between lists, each author is an observation. See subsection 3.1.2 for discussion and directions.

	CPP-LAD	CPP-LAU	CPP-ELE	LAD-LAU	LAD-ELE	LAU-ELE
P	0.44	0.34	2.57	0.20	2.32	2.37
I	0.74	0.99	3.72	0.32	3.37	3.10
H	0.26	0.32	3.72	0.29	4.36	4.24

Source: Prepared by the authors.

Table 17: c' values for stopwords. Comparison of the same sector between lists, each author is an observation. See subsection 3.1.2 for discussion and directions.

	CPP-LAD	CPP-LAU	CPP-ELE	LAD-LAU	LAD-ELE	LAU-ELE
P	3.31	3.26	6.68	0.57	5.36	5.41
I	1.45	1.08	5.16	0.91	5.00	4.92
H	0.98	0.68	4.35	1.05	4.73	5.01

Source: Prepared by the authors.

Table 18: c' values for punctuations/char. Comparison of the same sector between lists, each author is an observation. See subsection 3.1.2 for discussion and directions.

	CPP-LAD	CPP-LAU	CPP-ELE	LAD-LAU	LAD-ELE	LAU-ELE
P	5.74	4.88	8.28	2.23	5.37	6.60
I	3.23	2.49	4.16	0.96	3.40	3.51
H	2.49	1.87	4.02	1.36	3.05	3.71

Source: Prepared by the authors.

2.1.3 What and how the texts from the sectors differs

In the next sections we will look through text-related measures 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 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:

- Peripherals use more nouns while hubs use more verbs and adverbs. The fraction of adjectives did not change systematically with respect to connectivity, but given that the nouns are more numerous in the periphery, there are more adjectives per noun in the hubs sector texts.
- The size of tokens and words were found greater in the more connected sectors, which might be related to the use of a more specialized vocabulary. Sentences and messages were found smaller in the more connected sectors.
- The differences found with respect to Wordnet synsets 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 incidences are more systematic and this analysis assisted by Wordnet is semantic, reasons that motivated our inclusion of these results by means not of example tables but by explicit counts of greater incidence in each sector throughout the networks.

2.1.4 Characters

Most often, peripheral and intermediary sectors use more digits and upper case letters. Hubs use more letters and vowels among letters. The use of white spaces, for example, does not seem to have any relation to connectivity. These results are illustrated in Table [20](#).

Table 19: Characters in each Erdős sector (p. for periphery, i. for intermediary, h. for hubs).
TAG: 6

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

Source: Prepared by the authors.

2.1.5 Tokens and words

In most of the networks analyzed, hubs use longer words, which might be related to the use of a specialized vocabulary. Hubs use more contractions and known words, while peripheral sector exhibit a greater incidence of punctuations among tokens. Although the token diversity ($\frac{|\text{tokens} \neq|}{|\text{tokens}|}$) 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. The same observation apply to the lexical diversity ($\frac{|\text{kw} \neq|}{\text{kw}}$). These results are exemplified in Table 21 where mean and variance were taken with respect to the length in characters of tokens, known words and stopwords.

Table 20: Tokens in each Erdős sector (p. for periphery, i. for intermediary, h. for hubs).
TAG: 1

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

Source: Prepared by the authors.

2.1.6 Sizes of sentences

Hubs present the lowest average sentence size, in characters, tokens or known words. This result is illustrated in Table 22 and might be considered counterintuitive given that punctuation is more abundant in the texts of less connected participants.

Table 21: Sentences sizes in each Erdős sector (p. for periphery, i. for intermediary, h. for hubs). TAG: 16

	g.	p.	i.	h.
<i>sents</i>	10757	1252	4529	4978
<i>sents%</i>	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

Source: Prepared by the authors.

2.1.7 Messages

Connectivity was found correlated to smaller messages in terms of characters and tokens. Connectivity was also found correlated to smaller messages in terms of the number of sentences, but it was less consistent. This result is exemplified in Table 23.

Table 22: Messages sizes in each Erdős sector (p. for periphery, i. for intermediary, h. for hubs). TAG: 0

	g.	p.	i.	h.
<i>msgs</i>	1992	286	841	865
<i>msgs%</i>	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

Source: Prepared by the authors.

2.1.8 POS tags

We found that lower connectivity yields more nouns and less verbs and adverbs. Also, the fraction of adjectives does not change, 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 24. Weaker evidence was found that hubs use more *adpositions*, determinants and 'particles and other functional words' while peripherals use more numerals.

Table 23: POS tags in each Erdős sector (p. for periphery, i. for intermediary, h. for hubs).

Universal POS tags³: 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
PUNC	0.00	0.00	0.00	0.00

Source: Prepared by the authors.

2.1.9 Wordnet-related results

For correctly analyzing text production in terms of the Wordnet lexical database, we only considered words that had synsets of with the POS tag obtained with the POS tagger. 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. Measures regarding Wordnet synsets often reach an extreme value (maximum or minimum) in the intermediary sector, which we understood as evidence that:

- the Erdős sectioning of the networks into peripheral, intermediary and hubs 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 social networks texts using Wordnet reveals aspects of the structures which are not clear with the non-semantic analysis we performed.

Furthermore, the analysis of the measures we obtained by means of the Wordnet is not trivial because of the number of different measures and because differences in measures are not obviously relevant. 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. $\frac{\text{maximum measure}}{\text{minimum measure}} > 1.1$. We considered *evidence of difference* in sectors in a network if $\frac{\text{maximum measure}}{\text{minimum measure}} > 1.2$. When $\frac{\text{maximum measure}}{\text{minimum measure}} > 1.5$, 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. 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 point when measurements peak at the intermediary sector, be it a maximum or minimum peak.

Noteworthy is that extra skepticism should be kept in mind about these results because of the unquantified noise in the measurements. Observations seem consistent and meaningful, but only about a third of total tokens were considered. This is because tokens were discarded if not having a Wordnet synset or when not having a synset with a POS tag true to the POS tag attributed by the POS tagger. Besides that, there are often more than one synset with the same POS tag for each word, and we chose the most frequent synset as ranked by Wordnet. In the positive side, we observe that $\approx 95\%$ of tokens which had synsets were considered. Examples of types tokens without synsets are stopwords, punctuations, numerals, acronyms and typos.

2.1.9.1 Wordnet POS tags

The observations here are somewhat consistent with those in Section 3.1.8: peripherals use more nouns and less verbs and adverbs. The variation here is regarding adjectives, which was found more frequent in hubs texts in this reduced set of tokens. These results are illustrated in Table 25.

Table 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 Erdős 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

Source: Prepared by the authors.

2.1.9.2 Wordnet synsets characteristics

Wordnet synsets with different POS tags have different relations. 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 relations.

- Nouns, exemplified in Table 26.
 - 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 was 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 was 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: words with more meronyms 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 was 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 networks 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 25: Measures of wordnet features of nouns in each Erdős sector (p. for periphery, i. for intermediary, h. for hubs). TAG: 13

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

Source: Prepared by the authors.

- Adjectives, exemplified in Table 27.
 - 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 $\approx 90\%$ of the cases where there was difference in the mean, the maximum number of domains was found in the hubs sector. In $\approx 90\%$ of the cases where there was difference in the variance, the maximum number of domains was found indistinctly the extreme sectors. Peaks occurred in the intermediary sector in 50% of the observations for the mean and in $\approx 36.37\%$ 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 $\approx 50\%$ of the observed cases.

Table 26: Measures of wordnet features of adjectives in each Erdős 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

Source: Prepared by the authors.

- Verbs, illustrated in Table 28.
 - 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 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 27: Measures of wordnet features of verbs in each Erdős sector (p. for periphery, i. for intermediary, h. for hubs). TAG: 3

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

Source: Prepared by the authors.

- Adverbs, exemplified in Table 29.
 - 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 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 28: Measures of wordnet features of adverbs in each Erdős sector (p. for periphery, i. for intermediary, h. for hubs). TAG: 17

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

Source: Prepared by the authors.

2.1.9.3 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 have to be reviewed in.⁷ 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: $\text{total} \geq 16$ for nouns; $\text{total} \geq 11$ for adjectives; $\text{total} \geq 14$ for verbs; $\text{total} \geq 13$ 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 texts from the sectors.

- Noun synsets: differences in the use of nouns with physical entity hypernyms was 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 30.

Table 29: Wordnet synset hypernyms from nouns in each Erdős 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

Source: Prepared by the authors.

- 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 always peaked at the intermediary sector. Even weaker evidence was found that hubs use more adjectives related to newness. These results are shown in Table 31.

Table 30: Wordnet synset hypernyms from adjectives in each Erdős 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

Source: Prepared by the authors.

- Verbs synset hypernyms of move and travel was found more numerous in the peripheral sector. Verbs related to change was found more common in the hubs sector. Verbs related to making was found with 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 32.

Table 31: Wordnet synset hypernyms from verbs in each Erdős 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

Source: Prepared by the authors.

- Adverb synsets was 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 33.

Table 32: Wordnet synset hypernyms from adverbs in each Erdős 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

Source: Prepared by the authors.

Table 33: Pierson 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

Source: Prepared by the authors.

2.1.10 Correlation of topological and textual metrics

Overall, small correlation is found between textual and topological metrics. An exception is that, in the hubs sector, strength 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 indicative of style from the different lists analyzed. These results are exemplified in Table 34.

2.1.11 Formation of principal components

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 topological or textual features. Other than that, we observe that PCA is sensitive to metrics included and should reveal other insights in other settings. These results are exemplified in Table 35.

2.1.12 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 . We considered this result a number of times in the past years and presented it to other researchers, but reached no conclusions about its meaning. Appendix ?? are dedicated to this histogram differences.

2.2 Results from visualization

The results associated with Versinus are divided into two groups: observations on features that made it useful for the task of analyzing the general properties of human interaction networks, and the network properties it made possible to grasp.

2.2.1 Useful visualization features for dynamic networks

Among the numerous insights related to Versinus, a few seem more fundamental, while others were simply useful. These insights were incorporated to Versinus as the result

Table 34: 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

Source: Prepared by the authors.

of tests which presented clear benefits within the context of our research. The following

list is an attempt to present the observations and insights in an importance-first order:

1. Vertices need to remain static. Even if they move smoothly, one should notice solely transient artifacts from the structure.
2. Very connected sectors (hubs and intermediary) need to be in a curve, otherwise the edges enclose each other and it is difficult to infer information from the network.
3. The height and width of a vertex are very informative, especially if metrics mapped to them have a strong relation, such as out-degree (mapped to height in Versinus) and in-degree (mapped to width).
4. The color of nodes is also informative although less than height and weight, as differences in the latter are more noticeable.
5. An ordering of nodes, related to their fixed position, is very useful. Among all tests, ordering of vertices by degree was considered the most informative, which led to the hub, intermediary and peripheral sectioning of the network delineated in Section 2.2.4. As node position in the layout is fixed throughout an animation which comprises consecutive but distinct network activity, such ordering is done with respect to the resulting network of all the activity. Numbering these positions with respect to the order of the vertices in the larger structure (i.e. all M messages) is useful for understanding to which extent a vertex preserves the position in different scales of activity.

Many other insights were derived from Versinus, such as possible visualization tools, other kinds of convenient layouts and glyph elaborations. These received attention in Section 3.2.3.

2.2.2 Understanding of network properties through Versinus

A number of hypotheses were drawn about the networks for which Versinus was designed. As suggested by Palla, Barabási and Vicsek,¹⁰ stability of participant activity in social networks is more incident in smaller networks. Consistent with this result, all hubs have intermittent activity in the settings analyzed, except for the email list with the smallest number of participants (the Metareciclagem email list). The intermittence of hubs was one of the top hypotheses which motivated the development of Versinus. The stability of the network structure, concomitant with the instability of the activity of each participant, motivated a deeper analysis.⁷⁸ In doing so, we also found evidence for another hypothesis drawn from Versinus: that in- and out-degree differences in each vertex are important for network characterization. Furthermore, the visualization suggests that there are modes of operation of the network. As an example, the intermediary sector often

communicates mostly with the hubs or with the peripheral vertices. Other hypotheses, such as discrepancies in the authority and the degree of a vertex, are numerous but need further research to be validated.

2.2.3 Refinement of Versinus

Versinus was convenient for obtaining insights about how to enhance its layout and use. It was immediate to think of a tool for using Versinus in real-time, but less obvious are some ideas about the layout and visual guides. To further enable visualization of hubs and intermediary vertex, the sinusoid can have many periods with a decaying frequency. The upper straight line can also have an oscillating outline. The two halves of the sinusoidal period could be moved independently. The waveform need not to be a sinusoid. One can think of many ways to make more informative glyphs. Also, visual and auditory signals for specific occurrences can be interesting (e.g. when a new vertex appears, when one vanishes, when an ordering of vertices changes). Measurements of each vertex can be shown with a vertical displacement, to enable multiple measurements, to avoid the need to blink the numbers and to keep network visualization free from occlusion. Working with Versinus has also suggested other kinds of layout for vertices, especially geometric figures and iterative force-based methods for positioning vertices in a fixed layout. The traditional matrix representation of the graphs has been gazed upon as support to Versinus as has been some recent approaches to network visualization.⁶²

2.3 Linked data results

The current results include data selection and preparation for knowledge discovery. In this respect, the main result lies in the fact that data were made available, which enables benchmarking of scientific results and experimentations. Secondary results include data outline through figures and tables, software support and example SparQL queries.

2.3.1 Standardization

The data is embedded into standard URIs and triples, i.e. translated to RDF. URIs are built in the namespace <http://purl.org/socialparticipation/participationontology/> which are identified herein with the prefix `po:`. Classes and properties are built by adding a suffix to the root, as in the class `po:Participant` or in the property `po:text`. Classes have “UpperCamelCase” suffixes while properties have “lowerCamelCase” suffixes. All class instances, such as participants, messages, friendships and interactions, are linked to snapshots through the triple `<instance> po:snapshot <snapshot_uri>`. Message texts, including comments, are objects in the triple: `<message_id> po:text <message_text>`. Pre-processed texts are objects of triples: `<message_id> po:cleanText <message_text>`. More specialized predicates are used for delivering text when necessary, such as `po:htmlBodyText`

and `po:cleanBodyText` used for ParticipaBR articles (instances of the class `po:Article`). A participant URI is unique throughout the provenance (e.g. the same for the same participant in all Twitter snapshots). To enable annotations which differ when the snapshot changes, `po:Observation` class instances are used in the triple `<participant_uri> po:observation <observation_uri>`. The observation instances are then linked to the snapshot and the data.

Instances are built on top of the class they derive from plus a hashtag character, a provenance string (e.g. `facebook-legacy` or `participabr-legacy`) of the snapshot they refer to, and an identifier; i.e. `po:Participant#<provenance-legacy>-<id>`. All snapshot URIs follow the formation rule: `po:<SnapshotProvenance>#<snapshot_id>`. All snapshot ids follow the formation rule: `<platform>-legacy-<further_identifier>`; e.g. `irc-legacy-labmacambira` or `email-legacy-linux.audio.devel1-20000`.

2.3.2 Data outline

The database consists of 34,120,026 triples, 3,172,927 edges yield by interactions or relations, 382,568 participants and 253,155,020 characters. Among all snapshots, 63 are ego snapshots, 54 are group snapshots; 49 have interaction edges, 89 have friendship edges; 43 have text content from messages.

Table 35: Number of snapshots from each provenance.

social protocol	number of snapshots
Algorithmic Autoregulation	3
Cidade Democrática	1
Email	4
Facebook	88
IRC	4
ParticipaBR	1
Twitter	16
all	117

2.3.3 Software tools

The database is released with software for rendering itself, analyses and multimedia artifacts.

2.3.3.1 Triplification routines

For each social platform there is a *triplification* routine, i.e. a script for translating data to RDF. Original formats and further observations are presented in Table 37.

Table 36: Social platforms, original formats and further observations for the database.

social platform	original format	further observations	toolbox
AA	MySQL and MongoDB databases; IRC text logs	donated by AA users	Participation ⁶⁸
Cidade Democrática	MySQL database	donated by admins	Participation
Email	mbox	obtained through Gmane public database	Gmane ⁵⁰
Facebook	GDF, GML and TAB	obtained through Netvizz ⁴⁹	Social ⁶⁹
IRC	plain text log	obtained through Supybot logging	Social
ParticipaBR	PostgreSQL database	donated by admins	Participation
Twitter	JSON	obtained through Twitter streaming API	Social

Source: Prepared by the author.

2.3.3.2 Topological and textual analysis

Routines are available for taking the topological and textual measurements from the database. Auxiliary routines, such as performing principal component analysis and taking Kolmogorov-Smirnov measurements, are available to facilitate pattern recognition. All the analysis routines used for this thesis are in these publicly accessible scripts.

2.3.3.3 Multimedia rendering

It is a core purpose of the framework to provide routines for rendering audiovisualizations of the data. Social structures are rendered into music, images and video animations through the Percolation toolbox⁷⁰ in association with the Music and Visuals toolboxes.^{71,72}

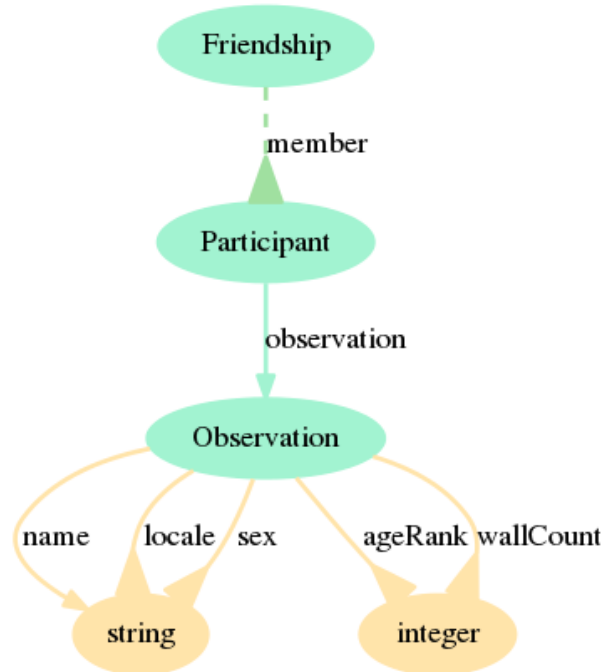
2.3.3.4 Migration from deprecated toolboxes

Routines mentioned in Sections 3.3.3.2 and 3.3.3.3 are being migrated from deprecated toolboxes^{73,74} into newly designed toolboxes.^{70,72}

2.3.4 Diagrams of the data and auxiliary tables

The database exploration can be assisted through diagrams which shows the structure from each provenance. Such diagrams are exemplified in the Appendix ?? and are fully available in an article² with some tables to make it easier to understand the data provided. A simplified example is given in Figure 7 where the friendship structure of the Facebook snapshots is illustrated.

Figure 4: A diagram of the structure involved in the friendship networks of the Facebook snapshots. A green edge denotes an OWL existential class restriction; an inverted nip denotes an OWL universal class restriction; a full (non-dashed) edge denotes an OWL functional property axiom. Further information and complete diagrams for each provenance are given in .²



Source: Prepared by the author.

2.3.5 SPARQL queries

There are numerous useful and general purpose SPARQL queries to be performed against the database. Here we write some of such queries selected by their simplicity and potential to be varied. All queries assume the use of the preamble `PREFIX po: <http://purl.org/socialparticipation/po/>`.

1. Retrieve the number of participants:

```
SELECT (COUNT(DISTINCT ?author) as ?c) WHERE { ?author a po:Participant . }
```

2. Retrieve the number of relations, be them interactions or friendships:

```
SELECT (COUNT(?interaction) as ?c) WHERE {
  { ?interaction a po:Friendship } UNION { ?interaction a po:Interaction } UNION
  { ?interaction po:retweetOf ?message } UNION { ?interaction po:replyTo ?message }
}
UNION { ?interaction po:directedTo ?participant }
}
```

3. Retrieve all text produced by a specific user:

```
SELECT (CONCAT(?text) as ?texts) WHERE {  
  ?activity po:author <user_uri> . ?activity po:text ?text .  
}
```

4. List 1000 users (URIs and names) with the most friendships and the number of friendships in descending order by the number of friendships:

```
SELECT DISTINCT ?participant (COUNT(?friendship) as ?c) WHERE {  
  ?friendship a po:Friendship . ?friendship po:member ?participant .  
} ORDER BY DESC(?c) LIMIT 1000
```

5. Retrieve text messages with the word “pineapple” (case insensitive):

```
SELECT ?text WHERE {  
  ?activity po:text ?text . FILTER regex(?text, 'pineapple', 'i')  
}
```

6. List participants and respective full names whose name has the substring “Amanda”:

```
SELECT DISTINCT ?participant ?name WHERE {  
  ?participant po:observation ?obs . ?obs po:name ?name .  
  FILTER regex(?name, 'Amanda', 'i')  
}
```

7. Return all pairs of friends of a participant which are friends themselves:

```
SELECT DISTINCT ?friend1 ?friend2 WHERE {  
  ?friendship1 po:member <participant_uri> . ?friendship1 po:member ?friend1 .  
  ?friendship2 po:member <participant_uri> . ?friendship2 po:member ?friend2 .  
  ?friendship3 po:member ?friend1 . ?friendship3 po:member ?friend2 .  
}
```

8. Return all interactions from replies in a snapshot:

```
SELECT ?from ?to WHERE {  
  ?message1 po:snapshot <snapshot_uri> . ?message2 po:replyTo ?message1 .  
  ?message1 po:author ?from . ?message2 po:author ?to .  
}
```

2.3.6 License issues

The database presented in this thesis is released under public domain. Computer scripts are in git repositories and PyPI Python packages, also under public domain. Although most data is already in open licenses (Twitter, Email, Participabr, Cidade Democrática, and AA data), IRC and Facebook data was collected and donated by the individuals which yielded the data. This rises the understanding of the right to study such data as the right to access the self, in line with anthropological endeavors.^{42,75}

2.3.7 Data-driven ontology synthesis

OWL Ontologies are critical tools to describe taxonomies and the structure of knowledge. Most ontologies are created by domain experts even though there often is data they organize that is given by a software system and which has a predefined structure.

We developed a simple ontology synthesis method that probes the ontological structure in data with SPARQL queries and post-processing. The results are OWL code and diagrams which are exemplified in the Appendix ?? and available in an article.² The method can be extended to comprise further OWL axioms and restrictions, but is currently performed to fit present needs with maximum simplicity. The present needs are limited to informative figures and the steps implemented are as follows:

1. Obtain all distinct classes with the query:

```
SELECT DISTINCT ?class_uri WHERE { ?s a ?class_uri }
```

2. For each class, obtain the properties that occur as predicates in triples where the subject is an instance of the class:

```
SELECT DISTINCT ?property_uri WHERE { ?s a <class_uri> . ?s ?property_uri ?o . }
```

Such properties are used to assert existential and universal restrictions for the class.

3. Compare the total number of individuals (?cs1) of the class (class_uri) with the number of such individuals (?cs2) that are subjects of at least one triple where the predicate is the property (property_uri). If the numbers match, there is an existential restriction for the class. The queries are:

```
SELECT (COUNT(DISTINCT ?s) as ?cs1) WHERE { ?s a <class_uri> }
```

```
SELECT (COUNT(DISTINCT ?s) as ?cs) WHERE {
```

```
  ?s a <class_uri>. ?s <property_uri> ?o .
```

```
}
```

4. Find the number of instances which are subjects of triples where the predicate is the property but are not instances of the class. If there is zero of such instances, there is a universal restriction:

```
SELECT (COUNT(DISTINCT ?s)=0 as ?cs) WHERE {
```

```
  ?s <property_uri> ?o . ?s a ?ca . FILTER(str(?ca) != 'class_uri')
```

```
}
```

5. To keep a record of the restrictions (and occurring triples), get all object classes or datatypes where the subject is an instance of the class and the predicate is the property:

```
SELECT DISTINCT ?co (datatype(?o) as ?do) WHERE {
```

```
  ?s a <class_uri>. ?s <property_uri> ?o . OPTIONAL { ?o a ?co . }
```

```
}
```

6. Obtain all distinct properties:

```
SELECT DISTINCT ?p WHERE { ?s ?p ?o }
```

7. Check if each property is functional, i.e. if it occurs at most once with each subject. This is performed by counting the objects and further verifying that they are at most one. The query is:

```
SELECT DISTINCT (COUNT(?o) as ?co) WHERE { ?s <property_uri> ?o } GROUP BY ?s
```

8. For each property, find the incident range and domain with the queries:

```
SELECT DISTINCT ?co (datatype(?o) as ?do) WHERE {  
  ?s <property_uri> ?o . OPTIONAL { ?o a ?co . }  
}
```

and

```
SELECT DISTINCT ?cs WHERE { ?s <property_uri> ?o . ?s a ?cs . }
```

9. Render diagrams as explained in the next section and in the Supporting Information file.

3 CONCLUSION AND FUTURE WORK

The very small standard deviations of principal components formation (see Sections 2.2.5 and 3.0.3), the presence of the Erdős sectors even in networks with few participants (see Sections 2.2.4 and 3.0.2), and the recurrent activity patterns along different timescales (see Sections 2.2.1 and 3.0.1), go a step further in characterizing scale-free networks in the context of the interaction of human individuals. Furthermore, the importance of symmetry-related metrics, which surpassed that of clustering coefficient, with respect to dispersion of the system in the topological measures space, might add to the current understanding of key-differences between digraphs and undirected graphs in complex networks. Noteworthy is also the very stable fraction of participants in each Erdős sector when the network reaches more than 200 participants. Benchmarks were derived from email list networks and the supplied analysis of networks from Facebook, Twitter and ParticipaBR in the Appendix might ease hypothesizing about the generality of these characteristics.

3.1 Textual analysis final remarks

This is a first systematic exploration of the relation between topological and textual metrics in human interaction networks, as far the authors know. Different textual features were scrutinized and were found to present evident patterns, specially in relation to topological measures and the Erdős sectors. Furthermore, results show that peripheral participants use more nouns while hubs use more verbs, which suggests that less connected participants bring content and concepts, while hubs propose action on them. 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.^{35–37, 76–78}

Most importantly, we understand reasonable to conclude, from all the distinct textual characteristics found between the Erdős sectors, that, as a rule of thumb, the texts from each of the sectors differ. Surely there should be exceptions and it is a fact that we left out of the analysis more subtle textual aspects e.g. those related to low percentages (<5%) or to small differences. These might be the subject of future contributions.

3.2 Linked data final remarks

The database presented in this thesis constitutes a large database with diverse provenance. Even so, the database should be expanded in upon need or requests from feedback. All data should be available online in the [<http://linkedopensocialdata.org>](http://linkedopensocialdata.org)

address in near future to fulfill the purpose of being a common repertoire in current research. One should reach the diagrams and tables of the articles produced in this research^{2,46,78,79} for further directions on the available structures and for an overview complement.

3.3 Future work

Further work should expand the analysis to include more types of networks and more metrics. The data and software needed to attain these results received dedicated and in-depth documentation as they enable a greater level of transparency and work share, which is adequate for both benchmarking and specifically for the study of systems constituted by human individuals (see Section 2.1). The derived typology of hub, intermediary and peripheral participants has been applied for semantic web and participatory democracy efforts, and these developments might be enhanced to yield scientific knowledge.^{2,36,37} Also, we plan to further explore and publish the audiovisualizations used for this research^{46,61} and the linguistic differences found in each of the Erdős sectors.⁷⁹

Similarity measures of texts in message-response threads has been thought about by us, and some results should be organized in near future. In this respect, there are two core hypotheses obtained from recent experiments:

- the existence of information “ducts”, observable through similarity measures. These might coincide with asymmetries of edges between vertex pairs, with homophily or with message-response threads, to point just a few possibilities.
- Valuable insights might be obtained 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 morphosyntactic classes, incidences of particular Wordnet synset characteristics and distances.

Current results suggest 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 (e.g. tokens, sentence sizes).

Other potential next steps are:

- The observation of most incident words and word types, such as words related to cursing, food or body parts.
- Interpretation of the constant difference found from incident and existent tokens histograms, exposed in Section 3.1.12.
- 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 weekdays).
 - A careful analysis of each textual feature distribution which is likely to reveal multimodal outlines and other non-trivial characteristics.
 - Extend analysis of textual measures to the windowed approach along the timelines, where hub, peripheral and intermediary sectors were 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 applications⁷⁶ 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.
 - Analyze interaction networks from other platforms such as LinkedIn, Diáspora, etc.
 - Sentiment analysis was not used in this work, but might be a good endeavor since the subject has received considerable attention from the scientific literature but has not included topological features as far as we know.
 - This thesis focused on differences of texts among Erdős sectors but we envisage that comparison of texts from social networks with canonical texts (e.g. Shakespeare or the King James Bible) might yield other powerful insights.

BIBLIOGRAPHY

- 1 JACKSON, M. O. **Social and Economic Networks: Models and Analysis**. 2013. <https://class.coursera.org/networksonline-001>.
- 2 FABBRI, R.; JUNIOR, O. N. O. **Linked Open Social Data for Scientific Benchmarking**. [S.l.]: GitHub, 2016. <https://github.com/ttm/linkedOpenSocialData/raw/master/paper.pdf>.
- 3 PETROV, S.; DAS, D.; MCDONALD, R. A universal part-of-speech tagset. **arXiv preprint arXiv:1104.2086**, 2011.
- 4 MORENO, J. L. Who shall survive?: A new approach to the problem of human interrelations. **The Journal of Social Psychology**, Nervous and Mental Disease Publishing Co, v. 6, p. 388–393, 1935.
- 5 NEWMAN, M. **Networks: an introduction**. [S.l.]: Oxford University Press, 2010.
- 6 LATOUR, B. Reassembling the social. an introduction to actor-network-theory. **Journal of Economic Sociology**, National Research University Higher School of Economics, v. 14, n. 2, p. 73–87, 2013.
- 7 BIRD, C.; GOURLEY, A.; DEVANBU, P.; GERTZ, M.; SWAMINATHAN, A. Mining email social networks. In: ACM. **Proceedings of the 2006 international workshop on Mining software repositories**. [S.l.], 2006. p. 137–143.
- 8 VÁZQUEZ, A.; OLIVEIRA, J. G.; DEZSÖ, Z.; GOH, K.-I.; KONDOR, I.; BARABÁSI, A.-L. Modeling bursts and heavy tails in human dynamics. **Physical Review E**, APS, v. 73, n. 3, p. 036127, 2006.
- 9 BALL, B.; NEWMAN, M. E. Friendship networks and social status. **arXiv preprint arXiv:1205.6822**, 2012.
- 10 PALLA, G.; BARABÁSI, A.-L.; VICSEK, T. Quantifying social group evolution. **Nature**, Nature Publishing Group, v. 446, n. 7136, p. 664–667, 2007.
- 11 LEICHT, E. A.; CLARKSON, G.; SHEDDEN, K.; NEWMAN, M. E. Large-scale structure of time evolving citation networks. **The European Physical Journal B**, Springer, v. 59, n. 1, p. 75–83, 2007.
- 12 TRAVENÇOLO, B.; COSTA, L. d. F. Accessibility in complex networks. **Physics Letters A**, Elsevier, v. 373, n. 1, p. 89–95, 2008.
- 13 NEWMAN, M. E. Modularity and community structure in networks. **Proceedings of the National Academy of Sciences**, National Acad Sciences, v. 103, n. 23, p. 8577–8582, 2006.
- 14 ALBERT, R.; BARABÁSI, A.-L. Topology of evolving networks: local events and universality. **Physical review letters**, APS, v. 85, n. 24, p. 5234, 2000.
- 15 MAREK-SPARTZ, K.; CHESLEY, P.; SANDE, H. Construction of the gmane corpus for examining the diffusion of lexical innovations. 2012.

- 16 ONNELA, J.-P.; ARBESMAN, S.; GONZÁLEZ, M. C.; BARABÁSI, A.-L.; CHRISTAKIS, N. A. Geographic constraints on social network groups. **PLoS one**, Public Library of Science, v. 6, n. 4, p. e16939, 2011.
- 17 PALCHYKOV, V.; KASKI, K.; KERTÉSZ, J.; BARABÁSI, A.-L.; DUNBAR, R. I. Sex differences in intimate relationships. **Scientific reports**, Nature Publishing Group, v. 2, 2012.
- 18 FABBRI, R. **Python package to observe temporal stability in the GMANE database**. 2015. <<https://github.com/ttm/percolation>>.
- 19 HOLLAND, J. H. **Complexity: A very short introduction**. [S.l.]: OUP Oxford, 2014.
- 20 COSTA, L. d. F.; RODRIGUES, F. A.; TRAVIESO, G.; BOAS, P. V. Characterization of complex networks: A survey of measurements. **Advances in Physics**, Taylor & Francis, v. 56, n. 1, p. 167–242, 2007.
- 21 ERDÖS, P.; RÉNYI, A. On random graphs, i. **Publicationes Mathematicae (Debrecen)**, v. 6, p. 290–297, 1959.
- 22 ORDENES, F. V.; THEODOULIDIS, B.; BURTON, J.; GRUBER, T.; ZAKI, M. Analyzing customer experience feedback using text mining: A linguistics-based approach. **Journal of Service Research**, SAGE Publications Sage CA: Los Angeles, CA, v. 17, n. 3, p. 278–295, 2014.
- 23 GUPTA, V.; LEHAL, G. S. et al. A survey of text mining techniques and applications. **Journal of emerging technologies in web intelligence**, Academy Publisher, PO Box 40 Oulu 90571 Finland, v. 1, n. 1, p. 60–76, 2009.
- 24 BIRD, S.; KLEIN, E.; LOPER, E. **Natural language processing with Python: analyzing text with the natural language toolkit**. [S.l.]: " O'Reilly Media, Inc.", 2009.
- 25 WILSON, R.; WATKINS, J. J. **Combinatorics: ancient & modern**. [S.l.]: OUP Oxford, 2013.
- 26 EADES, P.; KLEIN, K. Graph visualization. **EDBT School**, 2015.
- 27 FRUCHTERMAN, T. M.; REINGOLD, E. M. Graph drawing by force-directed placement. **Software: Practice and experience**, Wiley Online Library, v. 21, n. 11, p. 1129–1164, 1991.
- 28 BECK, F.; BURCH, M.; DIEHL, S.; WEISKOPF, D. A taxonomy and survey of dynamic graph visualization. In: WILEY ONLINE LIBRARY. **Computer Graphics Forum**. [S.l.], 2016.
- 29 BERNERS-LEE, T. **Linked data, 2006**. 2006.
- 30 MASINTER, L.; BERNERS-LEE, T.; FIELDING, R. T. Uniform resource identifier (uri): Generic syntax. 2005.
- 31 UMBRICH, J.; DECKER, S.; HAUSENBLAS, M.; POLLERES, A.; HOGAN, A. Towards dataset dynamics: Change frequency of linked open data sources. CEUR, 2010.

- 32 BIZER, C.; LEHMANN, J.; KOBILAROV, G.; AUER, S.; BECKER, C.; CYGANIAK, R.; HELLMANN, S. Dbpedia-a crystallization point for the web of data. **Web Semantics: science, services and agents on the world wide web**, Elsevier, v. 7, n. 3, p. 154–165, 2009.
- 33 AUER, S.; BIZER, C.; KOBILAROV, G.; LEHMANN, J.; CYGANIAK, R.; IVES, Z. Dbpedia: A nucleus for a web of open data. In: **The semantic web**. [S.l.]: Springer, 2007. p. 722–735.
- 34 CYGANIAK, R.; WOOD, D.; LANTHALER, M. Rdf 1.1 concepts and abstract syntax. **W3C Recommendation**, v. 25, p. 1–8, 2014.
- 35 FABBRI, R. **United Nations Development Programme: Tools for content classification in the ParticipaBR Brazilian federal portal of social participation**. [S.l.]. <<https://github.com/ttm/pnud3/blob/master/latex/produto.pdf?raw=true>>.
- 36 _____. **United Nations Development Programme: Adaptations and increments for the ParticipaBR Brazilian federal portal of social participation**. [S.l.]. <<https://github.com/ttm/pnud4/blob/master/latex/produto.pdf?raw=true>>.
- 37 _____. **Content extraction through API from the Brazilian Federal Portal of Social Participation and its tools to a social participation cloud**. [S.l.], 2014. (Desenvolvimento de Metodologias de Articulação e Gestão de Políticas Públicas para Promoção da Democracia Participativa). <<https://github.com/ttm/pnud5/blob/master/latex/produto.pdf?raw=true>>.
- 38 FABBRI, R.; POPPI, R. Continuous voting by approval and participation. **arXiv preprint arXiv:1505.06640**, 2015.
- 39 MACDAID, G. P.; MCCAULLEY, M. H.; KAINZ, R. I. **Atlas of type tables: Myers-Briggs Type indicator**. [S.l.]: Center for Applications of Psychological Type, 2005.
- 40 ADORNO, T. W.; FRENKEL-BRUNSWIK, E.; LEVINSON, D. J.; SANFORD, R. N. The authoritarian personality. Harpers, 1950.
- 41 FABBRI, R. **What are you and I? [Anthropological physics fundamentals]**. 2015. <https://www.academia.edu/10356773/What_are_you_and_I_anthropological_physics_fundamentals_>.
- 42 ANTUNES, D.; FABBRI, R.; PISANI, M. M. **Anthropological physics and social psychology in the critical research of networks**. **International Conference on Complex Systems (2015)**, 2015.
- 43 STANFORD, C.; ALLEN, J. S.; ANTÓN, S. C. **Biological anthropology: the natural history of humankind**. [S.l.]: Pearson, 2016.
- 44 BERNERS-LEE, T.; HENDLER, J.; LASSILA, O. et al. The semantic web. **Scientific american**, New York, NY, USA:, v. 284, n. 5, p. 28–37, 2001.
- 45 WITTEN, I. H. **Text Mining**. 2004.
- 46 FABBRI, R. Versinus: a visualization method for graphs in evolution. **arXiv preprint arXiv:1412.7311**, 2013. <<http://arxiv.org/abs/1412.7311>>.

- 47 _____. **A distance metric between histograms derived from the Kolmogorov-Smirnov test statistic: specification, measures reference and example uses.** [S.l.]. <<https://github.com/ttm/kolmogorov-smirnov/raw/master/paper.pdf>>.
- 48 WIKIPEDIA. **Gmane — Wikipedia, The Free Encyclopedia.** <<http://en.wikipedia.org/wiki/Gmane>>.
- 49 RIEDER, B. Studying facebook via data extraction: the netvizz application. In: **ACM. Proceedings of the 5th Annual ACM Web Science Conference.** [S.l.], 2013. p. 346–355.
- 50 INGEBRIGTSEN, L. M. **Gmane.** 2008.
- 51 FABBRI, R.; FABBRI, R.; VIEIRA, V.; PENALVA, D.; SHIGA, D.; MENDONÇA, M.; NEGRÃO, A.; ZAMBIANCHI, L.; THUMÉ, G. S. The algorithmic autoregulation software development methodology/a metodologia de desenvolvimento de software autorregulação algorítmica. **Revista Electronica de Sistemas de Informação**, Faculdade Cenecista de Campo Largo-FACECLA, v. 13, n. 2, p. 1, 2014.
- 52 FABBRI, R. **A Python package to deliver social linked data.** 2015. <<https://github.com/ttm/social>>.
- 53 _____. **Data from Participa.br, Cidade Democrática and AA, in XML/RDF and Turtle/RDF.** [S.l.]: Datahub, 2014. (Desenvolvimento de Metodologias de Articulação e Gestão de Políticas Públicas para Promoção da Democracia Participativa). <<http://datahub.io/organization/socialparticipation>>.
- 54 WOELFLE, M.; OLLIARO, P.; TODD, M. H. Open science is a research accelerator. **Nature Chemistry**, Nature Publishing Group, v. 3, n. 10, p. 745–748, 2011.
- 55 ANTUNES, D. C.; FABBRI, R.; PISANI, M. M. **Anthropological physics and social psychology in the critical research of networks.** 2015. CSDC'15 online conference, Conference on Complex Systems. <<https://www.youtube.com/watch?v=oeOKYc3-nbM>>.
- 56 MARDIA, K. V.; JUPP, P. E. **Directional statistics.** [S.l.]: John Wiley & Sons, 2009. v. 494.
- 57 LEICHT, E. A.; NEWMAN, M. E. Community structure in directed networks. **Physical review letters**, APS, v. 100, n. 11, p. 118703, 2008.
- 58 NEWMAN, M. Community detection and graph partitioning. **arXiv preprint arXiv:1305.4974**, 2013.
- 59 BRANDES, U. A faster algorithm for betweenness centrality*. **Journal of Mathematical Sociology**, Taylor & Francis, v. 25, n. 2, p. 163–177, 2001.
- 60 JOLLIFFE, I. **Principal component analysis.** [S.l.]: Wiley Online Library, 2005.
- 61 FABBRI, R. **Video visualizations of email interaction network evolution.** 2013–5. <https://www.youtube.com/playlist?list=PLf_EtaMqu3jVodaqDjN7yaSgsQx2Xna3d>.

- 62 ELZEN, S. van den; HOLTEN, D.; BLAAS, J.; WIJK, J. J. van. Reordering massive sequence views: Enabling temporal and structural analysis of dynamic networks. In: IEEE. **Visualization Symposium (PacificVis), 2013 IEEE Pacific**. [S.l.], 2013. p. 33–40.
- 63 _____. Reordering massive sequence views: Enabling temporal and structural analysis of dynamic networks. In: IEEE. **Visualization Symposium (PacificVis), 2013 IEEE Pacific**. [S.l.], 2013. p. 33–40.
- 64 KOOP, D.; FREIRE, J.; SILVA, C. T. Visual summaries for graph collections. In: IEEE. **Visualization Symposium (PacificVis), 2013 IEEE Pacific**. [S.l.], 2013. p. 57–64.
- 65 MILLER, G. A. Wordnet: a lexical database for english. **Communications of the ACM**, ACM, v. 38, n. 11, p. 39–41, 1995.
- 66 WIKIPEDIA. **Kolmogorov–Smirnov test** — **Wikipedia, The Free Encyclopedia**. 2015. [Online; accessed 26-September-2015]. Disponível em: https://en.wikipedia.org/w/index.php?title=Kolmogorov%E2%80%93Smirnov_test&oldid=682456076.
- 67 BOCCALETTI, S.; LATORA, V.; MORENO, Y.; CHAVEZ, M.; HWANG, D.-U. Complex networks: Structure and dynamics. **Physics reports**, Elsevier, v. 424, n. 4, p. 175–308, 2006.
- 68 FABBRI, R. **participation toolbox**. [S.l.]: GitHub, 2015. <https://github.com/ttm/participation>.
- 69 _____. **social toolbox**. [S.l.]: GitHub, 2015. <https://github.com/ttm/social>.
- 70 _____. **Percolation toolbox**. [S.l.]: GitHub, 2015. <https://github.com/ttm/percolation>.
- 71 _____. **Music toolbox**. [S.l.]: GitHub, 2015. <https://github.com/ttm/music>.
- 72 _____. **Visuals toolbox**. [S.l.]: GitHub, 2015. <https://github.com/ttm/visuals>.
- 73 _____. **Gmane legacy repository**. [S.l.]: GitHub, 2015. <https://github.com/ttm/gmaneLegacy>.
- 74 _____. **Percolation legacy repository**. [S.l.]: GitHub, 2015. <https://github.com/ttm/percolationLegacy>.
- 75 _____. **What are you and I? [Anthropological physics fundamentals]**. 2015.
- 76 _____. Ensaio sobre o auto-aproveitamento: um relato de investidas naturais na participação social. **arXiv preprint arXiv:1412.6868**, 2014.
- 77 FABBRI, R.; LUNA, R. B. de; MARTINS, R. A. P. et al. Social participation ontology: community documentation, enhancements and use examples. **arXiv preprint arXiv:1501.02662**, 2015.
- 78 FABBRI, R.; FABBRI, R.; ANTUNES, D. C.; PISANI, M. M.; JR, O. N. O. Temporal stability in human interaction networks. **arXiv preprint arXiv:1310.7769**, 2013.
- 79 FABBRI, R. A connective differentiation of textual production in interaction networks. 2013. <http://arxiv.org/abs/1412.7309>.