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 integrants 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 the labMacambira.sf.net collective for numerous discussions and guidance with regards to free and digital culture.

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 document.

"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. 175p. Thesis (Doctor in Science) - Instituto de Física de São Carlos, Universidade de São Paulo, São Carlos, 2016.

This work reports on stable (or invariant) 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 Erdös-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 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.

Keywords: Complex networks. Text mining. Pattern recognition. Statistics. Social network analysis. Typology.

RESUMO

FABBRI, R. **Estabilidade topológica e diferenciação textual em redes de interação humana: análise estatística e dados ligados**. 2016. 175p. Tese (Doutorado em Ciências) - Instituto de Física de São Carlos, Universidade de São Paulo, São Carlos, 2016.

Este trabalho relata propriedades 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 Erdös-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.

Palavras-chave: Redes complexas. Mineração de texto. Reconhecimento de padrões. Estatística. Análise de redes sociais. Tipologia.

LIST OF FIGURES

LIST OF TABLES

LIST OF FRAMES

LIST OF ABBREVIATIONS AND ACRONYMS

ABNT	Associação Brasileira de Normas Técnicas
abnTeX	ABsurdas Normas para TeX
EESC	Escola de Engenharia de São Carlos
IAU	Instituto de Arquitetura e Urbanismo
IBGE	Instituto Brasileiro de Geografia e Estatística
ICMC	Instituto de Ciências Matemáticas e de Computação
IFSC	Instituto de Física de São Carlos
IQSC	Instituto de Química de São Carlos
USP	Universidade de São Paulo
USPSC	Campus USP de São Carlos

LIST OF SYMBOLS

Γ	Letra grega Gama
Λ	Lambda
ζ	Letra grega minúscula zeta
$\in$	Pertence

CONTENTS

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.[?,?] With the increasing availability of data related to human interactions, research about these networks has grown continuously. Contributions can now be found in a variety of fields, from social sciences and humanities[?] to computer science[?] and physics,[?,?] given the multidisciplinary nature of the topic. One of the approaches from an exact science perspective is to represent interaction networks as complex networks,[?,?] with which several features of human interaction have been revealed. 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,[?,?] but only to a limited extent, since research is normally focused on a particular metric or task, such as accessibility or community detection.[?,?]

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.[?] Network types have been discussed with regard to the number of participants, intermittence of their activity and network longevity.[?] 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.[?] In email list networks, scale-free properties were reported with $\alpha \approx 1.8$[?] (as in web browsing and library loans[?]), and different linguistic traces were related to weak and strong ties.[?]

The fact that unreciprocated edges often exceed 50% in human interaction networks[?] motivated the inclusion of symmetry metrics in our analysis. No correlation of topological characteristics and geographical coordinates was found,[?] therefore geographical positions were not considered in our study. Gender related behavior in mobile phone datasets was indeed reported[?] but it is not relevant for the present work because email messages and addresses have no gender related metadata.[?]

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". We might add to this definition that a complex network is also a large graph (even while there seems not to be a consensus to what *large* means in such context) and that it is a graph representation of a system found in nature or in real or empirical systems. 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, to reproduce.

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). Such 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 i represented as a point while each edge is represented by a line between correspondent nodes. The matrix representation is essential for algebraic reasoning and for deriving measures while the node-edge representation is important for illustration and intuitive guidance of the characterization of the systems.

1.1.1.1 A good justification for the complex networks theory

The estimated number of atoms in the universe is often used as a reference of largeness and is 10^80. Let us find the number of vertices needed to reach such number of possible networks. Let also us 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:

$$2^{\binom{n}{2}} > 10^{80} \Rightarrow log_2[2^{\binom{n}{2}}] > log_2(10^{80}) \Rightarrow \binom{n}{2} > \frac{log_{10}(10^{80})}{log_{10}2} \Rightarrow$$

$$\Rightarrow \frac{n.(n-1)}{2} > \frac{80}{log_{10}2} \Rightarrow N > 23,5988$$

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 measures for nodes and the network (and less often for the edges).

1.1.1.2 Basic measures

Section 2.2.3 gives a mathematical account of the following measures, which are here for pointing the characteristics of basic types of networks presented in the next section. Such measures 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.

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 ones:

- The Erdös-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 vertices), low clustering and low average distance between nodes.

* This name is also used for the model in which, for a fixed number of vertices and a fixed number of edges, all graphs are equally likely. This is the model originally introduces by Paul Erdös and Alfréd Rényi.[?] We choose the definition given inline, which is closely related to the one given in this footnote, because it is more commonly used nowadays.

- Spatial network, also called geographic network or geometric graph: nodes are located in a metric space and the probability that two vertices are connected is greater as the distance between them gets smaller. These networks present characteristic degrees, high clustering and large average distance between nodes.

- Small-word network: defined as a network where the typical distance between vertices grows with the logarithm of the number of nodes while the average clustering coefficient is not small (larger than e.g. in the Erdös-Rényi model). To construct a small-world network, start with a regular lattice in which each vertex is connected to k nearest neighbors. Each edge 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 vertices (as in the Erdös-Rényi model) and a high average clustering coefficient (as in the spatial network). This model presents also a characteristic degree.

- Scale-free networks: in which the degree distribution $p(k)$ follows a power law ($p(k) = C.k^{-\gamma}$ where C and k are constants). These networks are qualitatively characterized by the presence of a large number of poorly connected and of few highly connected vertices. 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 of 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 vertex and an edge, i.e. on how the system is modeled as a graph. Another noteworthy remark is that the Erdös-Rényi networks, i.e. graphs of the Erdös-Rényi model, are frequently pin-pointed as the networks with trivial topological properties. Even though, it is posed as a paradigmatic "complex network", concept often defined as graphs with non-trivial topological properties, which is a contradiction and exposes that complex networks is not a very well defined notion, as is the case with the complexity field 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.[?] 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.[?] Text mining tasks include document summarization, sentiment analysis and

natural language processing techniques such as part of speech tagging.[?] Among application are social media monitoring, automated ad placement, publishing and making tools for semantics, sentiment and general natural language.[?] It is believed that applying text mining to social media can yield interesting findings in human behavior.[?] Although there is no clear cut, text mining is sometimes divided into linguistic and non-linguistic.[?] In the first case, linguistic techniques 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 **??** 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. On 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 as illustrated in Figure **??**. As an independent field, graph drawing arose in the 1990s[?] while representing graphs as node-link diagrams has a long tradition which remotes at least to the works of Ramon Llull in the 13th century.[?] To glimpse at the theory involved in visualizing networks, we mention three aspects:

- criteria for the quality of layouts include the number of crossing edges, the area of the drawing relative to 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 well known literature. Be f_a the attraction force, f_r the repulsion force, d the distance between the vertices and k a constant. The model introduced by Fruchterman and Reingold[?] defines the forces as:

$$f_a = \frac{d^2}{k} \tag{1.1}$$

$$f_d = -\frac{k^2}{d} \tag{1.2}$$

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

The core difference of dynamic graphs to static graphs is that vertices and edges can be added and removed over time. In dynamic graph visualization most usually graphs are represented as animated diagrams or charts based on a timeline.[?] In this thesis we make use of node-link diagrams of both static and dynamic graphs.

1.1.4 Linked 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 in 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.

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

1.1.5 Social participation

1.1.6 Other

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 seem to be more traditional in relation to the achievement of node-edge network diagrams. Evolutionary graph or network visualization is an example variant of dynamic graph visualization. The nomenclature of vertices and edges vary widely among interested fields (mathematics, physics, biology, sociology, etc). A vertex might be called e.g. a node, a point, an agent, a 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 computer and 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,[?] the expression linked data

appeared in a 2006 design note about the Semantic Web project[?] and refers to structured data that can emphasizes interlinking and usefulness through semantic queries.

Social participation, *social envolvement* 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 the 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 the 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 memberships, in that it fits the literature prediction, in Section .
- Lace, loop, selfloop and autoloop are terms used to designate an edge from a vertex to itself.

1.3 A historical note

1.4 Considerations about the presented work

1.5 Structure of the thesis

2 MATERIALS AND METHODS

2.1 Data and scripts

Email list messages were obtained from the Gmane email archive, which consists of more than $20,000$ email lists (discussion groups) and more than 130×10^6 messages.[?] These lists cover a variety of topics, mostly technology-related. The archive can be described as a corpus along with message metadata, including sent time, place, sender name, and sender email address. The usage of the Gmane database in scientific research is reported in studies of isolated lists and of lexical innovations.[?,?]

We observed various email lists and selected five of them together with data from Twitter, Facebook and Participabr for a thorough analysis, from which general properties can be inferred. These lists are as follows:

- Linux Audio Users list[*],with participants from different countries with artistic and technological interests. English is the prevailing language. Abbreviated as LAU from now on.
- Linux Audio Developers list[†], with participants from different countries; a more technical and less active version of LAU. English is the prevailing language. Abbreviated as LAD from now on.
- Developer's list for the standard C++ library[‡], with computer programmers from different countries. English is the prevailing language. Abbreviated as CPP from now on.
- List of the MetaReciclagem project[§], a Brazilian email list for digital culture. Portuguese is the prevailing language, although some messages are written in Spanish and English. Abbreviated as MET from now on.
- List for de discussion of the election reform[¶]. English is the prevailing language. Abbreviated ELE from now on.

The first 20,000 messages of each list were considered, with basic attributes of total timespan, authors, threads and missing messages indicated in Table 1. We considered 140 additional email lists to report on the interdependence between the number of participants

* gmane.linux.audio.users is list ID in Gmane.
† gmane.linux.audio.devel is list ID in Gmane.
‡ gmane.comp.gcc.libstdc++.devel is list ID in Gmane.
§ gmane.politics.organizations.metareciclagem is list ID in Gmane.
¶ gmane.politics.election-methods is list ID in GMANE.

Table 1: Columns $date_1$ and $date_M$ have dates of first and last messages from the 20,000 messages 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).

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

and the number of discussion threads. Furthermore, 12 networks from Facebook (8), Twitter (2) and Participabr (2) were scrutinized, and their analysis is given in the Supporting Information document for the purpose of testing the generality of the results.

2.1.1 Availability

The data and scripts used to derive the results, figures and tables are publicly available. Email messages are downloadable from the Gmane public database.[?] Data annotated from Facebook and Twitter are in a public repository.[?] Data from Participabr was used from the linked data/semantic web RDF triples,[?] available in.[?] Computer scripts are delivered through public domain Python PyPI packages and open Git repositories.[?] This open approach to both data and scripts reinforces the scientific aspect of the contribution[?] and mitigates ethical and moral issues involved in researching systems constituted of human individuals.[?,?]

2.2 Methods

2.2.1 Temporal activity statistics

Messages were counted over time as histograms in the scales of seconds, minutes, hours, days of the week, days of the month, and months of the year. Most standard measures of location and dispersion, e.g. the usual mean and standard deviation, hold little meaning in a compact Riemannian manifold, such as the recurrent time periods that we are interested in. Similar measures were taken using circular statistics,[?] in which each measurement t is represented as a unit complex number, $z = e^{i\theta} = \cos(\theta) + i\sin(\theta)$, where $\theta = t\frac{2\pi}{T}$, and T is the period in which the counting is repeated. For example, $\theta = 12\frac{2\pi}{24} = \pi$ for a message sent at $t = 12h$ and given $T = 24h$ for days. The moments m_n, lengths of moments R_n, mean angles θ_μ, and rescaled mean angles θ'_μ are defined as:

$$
\begin{aligned}
m_n &= \frac{1}{N}\sum_{i=1}^{N} z_i^n \\
R_n &= |m_n| \\
\theta_\mu &= Arg(m_1) \\
\theta'_\mu &= \frac{T}{2\pi}\theta_\mu
\end{aligned}
\tag{2.1}
$$

θ'_μ is used as the measure of location. Dispersion is measured using the circular variance $Var(z)$, the circular standard deviation $S(z)$, and the circular dispersion $\delta(z)$:

$$
\begin{aligned}
Var(z) &= 1 - R_1 \\
S(z) &= \sqrt{-2\ln(R_1)} \\
\delta(z) &= \frac{1 - R_2}{2R_1^2}
\end{aligned}
\tag{2.2}
$$

Also, the ratio $r = \frac{b_l}{b_h}$ between the lowest b_l and the highest b_h incidences on the histograms served as a further clue of how close the distribution was to being uniform. As expected, a positive correlation was found in all $r, Var(z)$, $S(z)$ and $\delta(z)$ dispersion measures, which can be noticed in Section B.1.1 of the Supporting Information. The circular dispersion $\delta(z)$ was found more sensitive and therefore preferred in the discussion of results.

2.2.2 Interaction networks

Edges in interaction networks can be modeled both as weighted or unweighted, as directed or undirected.[?,?,?] Networks in this thesis are directed and weighted, the most informative of the possibilities. We did not investigate directed unweighted, undirected weighted, and undirected unweighted representations of the interaction networks.

The interaction networks were obtained as follows: a direct response from participant B to a message from participant A yields an edge from A to B, as information went from A to B. The reasoning is: if B wrote a response to a message from A, he/she read what A wrote and formulated a response, so B assimilated information from A, thus $A \rightarrow B$. Edges in both directions are allowed. Each time an interaction occurs, the value of one is added to the edge weight. Selfloops were regarded as non-informative and discarded. Inverting edge direction yields the status network: B read the message and considered what A wrote worth responding, giving status to A, thus $B \rightarrow A$. This thesis considers by convention the information network as described above ($A \rightarrow B$) and depicted in Figure 1. These interaction networks are reported in the literature as exhibiting scale-free and small-world properties, as expected for a number of social networks.[?,?]

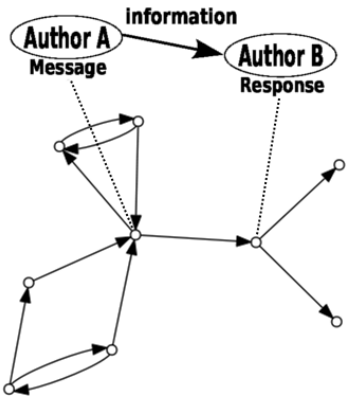


Figure 1: The formation of interaction networks from exchanged messages. Each vertex represents a participant. A reply message from author B to a message from author A is regarded as evidence that B received information from A and yields a directed edge. Multiple messages add "weight" to a directed edge. Further details are given in Section 2.2.2.

2.2.3 Topological metrics

The topology of the networks was characterized from a small selection of the most basic and fundamental measurements for each vertex,[?] as follows:

- 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.
- Clustering coefficient cc_i: fraction of pairs of neighbors of i that are linked, i.e. the standard clustering coefficient metric for undirected graphs.
- Betweenness centrality bt_i: fraction of geodesics that contain vertex i. The betweenness centrality index was computed for weighted digraphs as specified in.[?]

The non-standard metrics below were formulated to capture symmetries in the activity of participants:

- Asymmetry of vertex i: $asy_i = \frac{k_i^{in} - k_i^{out}}{k_i}$.

- Average asymmetry of edges at vertex i:
  $\mu_i^{asy} = \frac{\sum_{j \in J_i} e_{ji} - e_{ij}}{|J_i|}$, where e_{ij} is 1 if there is an edge from i to j, and 0 otherwise, and J_i is the set of neighbors of vertex i.

- Standard deviation of asymmetry of edges:
  $\sigma_i^{asy} = \sqrt{\frac{\sum_{j \in J_i} [\mu_i^{asy} - (e_{ji} - e_{ij})]^2}{|J_i|}}$.

- Disequilibrium: $dis_i = \frac{s_i^{in} - s_i^{out}}{s_i}$.

- Average disequilibrium of edges:
  $\mu_i^{dis} = \frac{\sum_{j \in J_i} \frac{w_{ji} - w_{ij}}{w_{ji} + w_{ij}}}{|J_i|}$, where w_{xy} is the weight of edge $x \to y$ and zero if there is no such edge.

- Standard deviation of disequilibrium of edges: $\sigma_i^{dis} = \sqrt{\frac{\sum_{j \in J_i} \left[\mu_i^{dis} - \frac{w_{ji} - w_{ij}}{w_{ji} + w_{ij}}\right]^2}{|J_i|}}$.

Both standard and non-standard metrics are used for the Erdös sectioning (described in Section 2.2.4) and for performing principal component analysis (PCA) (as described in Section 2.2.5).

2.2.4 Erdös sectioning

It is often useful to think of vertices as hubs, peripheral and intermediary. We have therefore derived the peripheral, intermediary and hub sectors of an empirical network from a comparison against an Erdös-Rényi network with the same number of edges and vertices, as depicted in Figure 2. We refer to this procedure as *Erdös sectioning*, with the resulting sectors being named as *Erdös sectors*. The Erdös sectioning was recognized as a theoretical possibility by M. O. Jackson in his video lectures,[?] but to our knowledge it has not as yet been applied to empirical data.

The degree distribution $\widetilde{P}(k)$ of a real network with a scale-free profile $\mathcal{N}_f(N, z)$ with N vertices and z edges has less average degree nodes than the distribution $P(k)$ of an Erdös-Rényi network with the same number of vertices and edges. Indeed, we define in this work the intermediary sector of a network to be the set of all the nodes whose degree is less abundant in the real network than on the Erdös-Rényi model:

$$\widetilde{P}(k) < P(k) \Rightarrow \text{k is intermediary degree} \tag{2.3}$$

If $\mathcal{N}_f(N, z)$ is directed and has no self-loops, the probability of the existence of an edge between two arbitrary vertices is $p_e = \frac{z}{N(N-1)}$. A vertex in the ideal Erdös-Rényi

digraph with the same number of vertices and edges, and thus the same probability p_e for the presence of an edge, will have degree k with probability

$$P(k) = \binom{2(N-1)}{k} p_e^k (1-p_e)^{2(N-1)-k} \tag{2.4}$$

The lower degree fat tail corresponds to the border vertices, i.e. the peripheral sector or periphery where $\widetilde{P}(k) > P(k)$ and k is lower than any value of k in the intermediary sector. The higher degree fat tail is the hub sector, i.e. $\widetilde{P}(k) > P(k)$ and k is higher than any value of k in the intermediary sector. The reasoning for this classification is as follows: vertices so connected that they are virtually nonexistent in the Erdös-Rényi model, are coherently associated to the hub sector. Vertices with very few connections, which are way more abundant than expected in the Erdös-Rényi model, are assigned to the periphery. Vertices with degree values predicted as the most abundant in the Erdös-Rényi model, near the average, and less frequent in the real network, are classified as intermediary.

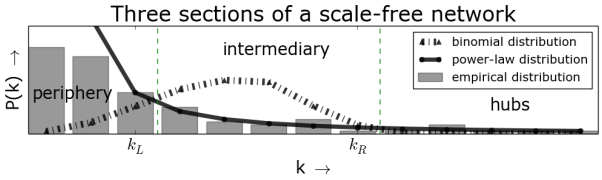


Figure 2: Classification of vertices by comparing degree distributions.[?] The binomial distribution of the Erdös-Rényi network model exhibits more intermediary vertices, while a scale-free network, associated with the power-law distribution, has more peripheral and hub vertices. The sector borders are 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 "Erdös 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.

To ensure statistical validity of the histograms, bins can be chosen to contain at least η vertices of the real network. The range Δ of incident values of degree k should be partitioned in m parts $\Delta = \cup_{i=1}^{m} \Delta_i$, with $\Delta_i \cap \Delta_j = \emptyset \; \forall \; i \neq j$ and:

$$\Delta_i = \Bigg\{ k \;\Big|\; \begin{aligned} & \overline{\Delta}_{i-1} < k \leq l \text{ and} \\ & \Bigg[\Bigg[N - \sum_{k=0}^{\overline{\Delta}_{i-1}} \eta_k < \eta \text{ and } l = \overline{\Delta} \Bigg] \text{ or} \\ & \Bigg[\sum_{k=\overline{\Delta}_{i-1}+1}^{l} \eta_k \geq \eta \text{ and} \\ & \Bigg(\sum_{k=\overline{\Delta}_{i-1}+1}^{l-1} \eta_k < \eta \text{ or } l = \overline{\Delta}_{i-1} + 1 \Bigg) \Bigg] \Bigg] \Bigg\} \end{aligned} \tag{2.5}$$

where η_k is the number of vertices with degree k, while $\overline{\Delta}_{(i)} = max(\Delta_{(i)})$, and $\overline{\Delta}_0 = -1$. Equation 2.3 can now be written in the form:

$$\begin{aligned} & \sum_{x=min(\Delta_i)}^{\overline{\Delta}_i} \widetilde{P}(x) < \sum_{x=min(\Delta_i)}^{\overline{\Delta}_i} P(x) \Leftrightarrow \\ & \Leftrightarrow \Delta_i \text{ spans intermediary degree values.} \end{aligned} \tag{2.6}$$

If the strength s is used for comparison of the real network against the Erdös-Rényi model, P remains the same, but $P(\kappa_i)$ with $\kappa_i = \frac{s_i}{\overline{w}}$ should be used, where $\overline{w} = 2\frac{z}{\sum_i s_i}$ is the average weight of an edge and s_i is the strength of vertex i. For in and out degrees

(k^{in}, k^{out}), the real network should be compared against

$$\hat{P}(k^{way}) = \binom{N-1}{k^{way}} p_e^k (1-p_e)^{N-1-k^{way}}, \tag{2.7}$$

where *way* can be *in* or *out*. In and out strengths (s^{in}, s^{out}) are divided by $\overline{w}$ and compared also using $\hat{P}$. Note that p_e remains the same, as each edge yields an incoming (or outgoing) edge, and there are at most $N(N-1)$ incoming (or outgoing) edges, thus $p_e = \frac{z}{N(N-1)}$, as with the total degree.

In other words, let γ and ϕ be integers in the intervals $1 \leq \gamma \leq 6$, $1 \leq \phi \leq 3$, and each of the basic six Erdös sectioning possibilities $\{E_\gamma\}$ have three Erdös sectors $E_\gamma = \{e_{\gamma,\phi}\}$ defined as

$$\begin{aligned} e_{\gamma,1} &= \{\ i \mid \overline{k}_{\gamma,L} \geq \overline{k}_{\gamma,i}\} \\ e_{\gamma,2} &= \{\ i \mid \overline{k}_{\gamma,L} < \overline{k}_{\gamma,i} \leq \overline{k}_{\gamma,R}\} \\ e_{\gamma,3} &= \{\ i \mid\ \ \overline{k}_{\gamma,i} > \overline{k}_{\gamma,R}\}, \end{aligned} \tag{2.8}$$

where $\{\overline{k}_{\gamma,i}\}$ is

$$\begin{aligned} \overline{k}_{1,i} &= k_i \\ \overline{k}_{2,i} &= k_i^{in} \\ \overline{k}_{3,i} &= k_i^{out} \\ \overline{k}_{4,i} &= \frac{s_i}{\overline{w}} \\ \overline{k}_{5,i} &= \frac{s_i^{in}}{\overline{w}} \\ \overline{k}_{6,i} &= \frac{s_i^{out}}{\overline{w}} \end{aligned} \tag{2.9}$$

and both $\overline{k}_{\gamma,L}$ and $\overline{k}_{\gamma,R}$ are found using $P(\overline{k})$ or $\hat{P}(\overline{k})$ as described above and illustrated in Figure 2.

Since different metrics can be used to identify the three types of vertices, more than one metric can be used simultaneously, which is convenient when analysing small networks, such as the cases where only 50 messages are considered in Section B.3 of the Supporting Information. After a careful consideration of possible combinations, these were reduced to six:

- Exclusivist criterion C_1: vertices are only classified if the class is the same according to all metrics. In this case, vertices classified do not usually reach N (or 100%), which is indicated by a black line in Figure 4.

- Inclusivist criterion C_2: a vertex has the class given by any of the metrics. Therefore, a vertex may belong to more than one class, and the total number of memberships may exceed N (or 100%), which is indicated by a black line in Figure 4.
- Exclusivist cascade C_3: vertices are only classified as hubs if they are hubs according to all metrics. Intermediary are the vertices classified either as intermediary or hubs with respect to all metrics. The remaining vertices are regarded as peripheral.
- Inclusivist cascade C_4: vertices are hubs if they are classified as such according to any of the metrics. The remaining vertices are intermediary if they belong to this category for any of the metrics. Peripheral vertices are those which are classified as such with respect to all metrics.
- Exclusivist externals C_5: vertices are hubs if they are classified as such according to all the metrics. Vertices are peripheral if they are peripheral or hubs for all metrics. The remaining nodes are intermediary.
- Inclusivist externals C_6: hubs are vertices classified as hubs according to any metric. The remaining vertices are peripheral if they are classified as such according to any metric. The rest of the vertices are intermediary.

Using Equations (2.8), these *compound criteria* C_δ, with δ integer in the interval $1 \leq \delta \leq 6$, can be specified as:

$$
\begin{aligned}
C_1 &= \{c_{1,\phi} = \{i \mid i \in e_{\gamma,\phi},\ \forall\, \gamma\}\} \\
C_2 &= \{c_{2,\phi} = \{i \mid \exists\ \gamma : i \in e_{\gamma,\phi}\}\} \\
C_3 &= \{c_{3,\phi} = \{i \mid i \in e_{\gamma,\phi'},\ \forall\, \gamma,\ \forall\, \phi' \geq \phi\}\} \\
C_4 &= \{c_{4,\phi} = \{i \mid i \in e_{\gamma,\phi'},\ \forall\, \gamma,\ \forall\, \phi' \leq \phi\}\} \\
C_5 &= \{c_{5,\phi} = \{i \mid i \in e_{\gamma,\phi'},\ \forall\, \gamma, \\
&\qquad\qquad \forall\, (\phi'+1)\%4 \leq (\phi+1)\%4\}\} \\
C_6 &= \{c_{6,\phi} = \{i \mid i \in e_{\gamma,\phi'},\ \forall\, \gamma, \\
&\qquad\qquad \forall\, (\phi'+1)\%4 \geq (\phi+1)\%4\}\}
\end{aligned}
\tag{2.10}
$$

Notice that the exclusivist cascade is the same sectioning of an inclusivist cascade from periphery to hubs, but with inverted order of sectors. The simplification of all possible compound possibilities to the small set listed above might be formalized in strict mathematical terms, but this was considered out of the scope for current interests.

2.2.5 Principal Component Analysis of topological metrics

Principal Component Analysis (PCA) is a well documented technique[?] and is used here to address the following questions: 1) which metrics contribute to each principal

component and in what proportion; 2) how much of the dispersion is concentrated in each component; 3) which are the expected values and dispersions for these quantities over various networks. This enables one to characterize human interaction networks in terms of the relative importance of network metrics and the way they combine.

Let $\mathbf{X} = \{X[i,j]\}$ be a matrix where each element is the value of the metric j at vertex i. Let $\mu_X[j] = \frac{\sum_i X[i,j]}{I}$ be the mean of metric j over all I vertices, $\sigma_X[j] = \sqrt{\frac{\sum_i (X[i,j]-\mu_X[j])^2}{I}}$ the standard deviation of metric j, and $\mathbf{X'} = \{X'[i,j]\} = \left\{\frac{X[i,j]-\mu_X[j]}{\sigma_X[j]}\right\}$ the matrix with the *z-score* of each metric. Let $\mathbf{V} = \{V[j,k]\}$ be the matrix $J \times J$ of eigenvectors of the covariance matrix $\mathbf{C}$ of $\mathbf{X'}$, one eigenvector per column. Each eigenvector combines the original metrics into one principal component, therefore $V'[j,k] = 100\frac{|V[j,k]|}{\sum_{j'}|V[j',k]|}$ is the percentage of the principal component k that is proportional to the metric j. Let $\mathbf{D} = \{D[k]\}$ be the eigenvalues associated with the eigenvectors $\mathbf{V}$, then $D'[k] = 100\frac{D[k]}{\sum_{k'} D[k']}$ is the percentage of total dispersion of the system that the principal component k is responsible for. We consider, in general, the three largest eigenvalues and the respective eigenvectors in percentages: $\{(D'[k],\ V'[j,k])\}$. These usually sum up between 60 and 95% of the dispersion and reveal patterns for a first analysis. In particular, given L snapshots l of the interaction network, we are interested in the mean $\mu_{V'}[j,k]$ and the standard deviation $\sigma_{V'}[j,k]$ of the contribution of metric j to the principal component k, and the mean $\mu_{D'}[k]$ and the standard deviation $\sigma_{D'}[k]$ of the contribution of the component k to the dispersion of the system:

$$\begin{aligned}
\mu_{V'}[j,k] &= \frac{\sum_{l=1}^{L} V'[j,k,l]}{L} \\
\sigma_{V'}[j,k] &= \sqrt{\frac{\sum_{l=1}^{L}(\mu_{V'} - V'[j,k,l])^2}{L}} \\
\mu_{D'}[k] &= \frac{\sum_{l=1}^{L} D'[k,l]}{L} \\
\sigma_{D'}[k] &= \sqrt{\frac{\sum_{l=1}^{L}(\mu_{D'} - D'[k,l])^2}{L}}
\end{aligned} \tag{2.11}$$

The covariance matrix $\mathbf{C}$ is the correlation matrix because $\mathbf{X'}$ is normalized. Therefore, $\mathbf{C}$ is also directly observed as a first clue for patterns by the most simple associations: low absolute values indicate low correlation (and a possible independence); high values indicate positive correlation; negative values with a high absolute value indicate negative correlation. Notice that in this case the variable k is not the degree value but a principal component. In the results the principal components are numbered according to the magnitude of associated eigenvalue and k is incorporated into the notation (e.g. PC2 for metrics of $\mu_{V'}[j,2]$).

2.2.6 Evolution and audiovisualization of the networks

The evolution of the networks was observed within sequences of snapshots. In each sequence, a fixed number of messages, i.e. the window size ws, was used for all snapshots. The snapshots were made disjoint in the message timeline, and were used to perform both PCA with topological metrics and Erdös sectioning. Figures and tables were usually inspected with $ws = \{50, 100, 200, 400, 500, 800, 1000, 2000, 2500, 5000, 10000\}$ messages. Variations in the number of vertices, edges and other network characteristics, within the same window size ws, are given in Section B.3 of the Supporting Information document.

2.2.7 The Versinus graph visualization method

Network structures were mapped to video animations, sound and musical structures developed for this research.[?]Such *audiovisualizations* were crucial in the initial steps and to guide the research into the most important features of network evolution. Versinus is a visualization method for dynamic graphs based on experimental observations. This method receives dedicated attention by recurrence of the suggestion, by fellow researchers, to write about it. In visualizing a network, the method consists of creating an animation, of a fixed-size message sliding window (e.g. 400 messages) and partitioning the network in two fixed-layout segments: a sinusoid for the most connected vertexes (hubs and intermediary) and a straight line for the less connected (peripheral). A vertex holds the same position throughout the animation. Also, visual cues of properties - such as color, height and width, and rank of vertex with degree criteria - play a central role. Numbers with individual measures for each vertex blink periodically. Versinus differs from the few works on the visualization of dynamic graphs because it is a simple method that has developed for practical needs and is the result of experimentations, although a number of criteria have guided its development.[?,?,?]

Let Δ be a fixed number of messages (e.g. $\Delta = 400$). Let also $s_i^{i+\Delta}$ be sets of Δ consecutive email messages along time. A sequence $S^{\Delta,M}$ of such sets, with the first message positioned in each the M messages (e.g. $M = 20000$), can be written as:

$$S^{\Delta,M} = \{s_i^{i+\Delta}\}_{i=0}^{M-\Delta} \tag{2.12}$$

Each set s_i yields an interaction network, as described in Section 2.2.2. Each of such sequence $S^{\Delta,M}$ of sets presents stable properties, while each participant exhibits a wide variation of characteristics. Understanding the mechanisms of this compatibility (unstable vertices and stable network) led to experimenting with a series of layouts and visualization techniques, from which Versinus emerged.

Taking advantage of the fact that vertices are roughly split into usual 80% of peripheral, 15% of intermediary and 5% of hubs, hubs are laid on the first half of a sinusoid,

intermediary on the second half, and peripheral on the straight line. This configuration can be improved in various forms, to which Section 3.2.3 is dedicated. Figure 3 has an image of such a layout. The fixed position of each vertex is defined by the overall structure, i.e. with respect to all M messages.

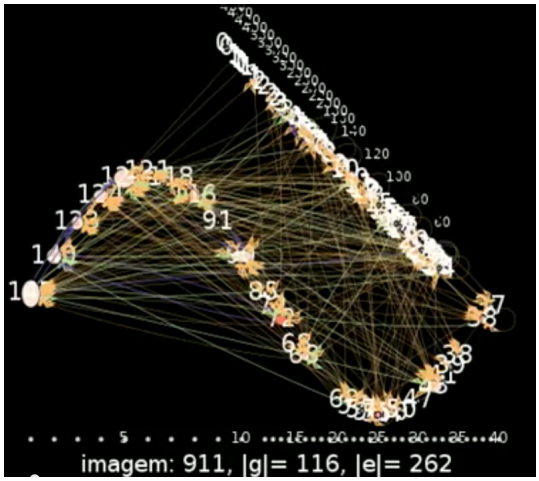


Figure 3: The versinus visualization method in use. 5% of the most connected vertexes (hubs) are on the left half-period of the sinusoid. 15% of the most connected remaining vertices are on the right half-period. 80% of the least connected vertices are on the straight line, above the sinusoidal shape. White dots with numbers keep track of node position in the overall degree ordering. Measures blink periodically near the vertices they are related to.

2.2.8 Text measures

This work focuses on very simple metrics derived from texts as they have been sufficient for current interests. Such metrics are:

- Frequency of characters: letters, vowels, punctuations and uppercase letters.
- Number of tokens, frequency of punctuations among tokens, frequency of known words, frequency of words that have wordnet synsets, frequency of tokens that are stopwords.
- Mean and standard deviation for word, token sentence and message sizes.
- Fraction of morphosyntactic classes, such as adverbs, adjectives, nouns and other POS (Part-Of-Speech) tags.
- Fraction of words in each wordnet[?] top-most hypernyms, such as abstraction and physical entities for nouns or act for verbs.

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

2.2.8.1 Relating text and topology

The topological and textual measures were related by:

1. textual measures in hub, intermediary and peripheral network sectors, which are delimited by topological criteria as described in Section 2.2.4.

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

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

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

$$D_{n,n'} > c(\alpha)\sqrt{\frac{n+n'}{nn'}} \tag{2.13}$$

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

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

We need to compare empirical distribution functions, therefore $D_{n,n'}$ is given, as are n and n'. All terms in equation 2.13 are positive and $c(\alpha)$ can be isolated:

$$c(\alpha) < \frac{D_{n,n'}}{\sqrt{\frac{n+n'}{nn'}}} = c' \tag{2.14}$$

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

2.2.9 Linked Open Social Database for Scientific Benchmarking

The data used here were obtained from Facebook, Twitter, IRC, Email and the detached instances of ParticipaBR, AA and Cidade Democrática. These were represented as linked data to homonenize access, complying with current best practices and facilitationg analyzes which integrate third party and provided instances.

Data was gathered from:

- public APIs (Twitter, Email);
- public logs (IRC and AA);
- Netvizz software[?] and subsequent donation by users (Facebook);
- donation by system administrators (AA, ParticipaBR, Cidade Democrática).

This section introduces the underlying data in a very concise fashion. Further information is available in the Appendix B and in an article.[?]

2.2.9.1 Snapshots

Of central importance to the database is the concept of a snapshot. A snapshot is herein a set of data gathered together, at a contiguous time span. For example: the first 20 thousand email messages of an email list comprises a snapshot; the tweets from the MAMA music event are a snapshot; the friendship, interaction and posts structures of a facebook group, prospected at the same time, are a snapshot.

2.2.9.2 Facebook data

Friendship ego networks (networks whose constituents are friends of a user) were donated from individual users in 2013 and 2014. Friendship and interaction networks from groups were gathered from groups where the author was a participant. Additionally, some groups have post texts along some metadata, such as the number of likes.

2.2.9.3 Twitter data

Tweets were gathered through the Twitter streaming public API. Each snapshot is unified by a distinct hashtag. Edges are canonically yield by retweets but replies and user mentions are also kept in the database.

2.2.9.4 IRC data

Public IRC logs were used to render IRC snapshots. The database has records of users to which the message is directed or mentions.

2.2.9.5 Email data

Email snapshots refer to individual email lists. All messages were obtained from the Gmane public email database.[?] Each message has the original text and the text without some of the lines from previous messages or that are software code. Most importantly, each message instance holds the ID of the message it is a reply to, if any.

2.2.9.6 ParticipaBR data

The ParticipaBR is a Brazilian federal platform for social participation. Texts are derived from blog posts and networks are derived from friendship and interaction criteria.

2.2.9.7 Cidade Democrática data

Cidade Democrática is a Brazilian civil society social participation portal. Data gathered is complex with many types of instances and no intuitive criteria for deriving networks, such as friendships or replies.

2.2.9.8 AA data

The Algorithmic Autoregulation[?] is a software development methodology based on testifying and sharing ongoing work. The data was gathered from different versions of the system and from an IRC log.

2.2.10 Linked open 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 me 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 steps[?]:

- Use URIs to identify things.[?]
- 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 Data[?] 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.[?,?]

2.2.10.1 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.[?]

As an example of a 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 .
```

3 RESULTS AND DISCUSSION

3.0.1 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, $\frac{min(incidence)}{max(incidence)} \in (0.784, .794)$ while simulations reach these values but have on average more discrepant higher and lower peaks, i.e. if $\xi = \frac{min(incidence')}{max(incidence')}$ 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 thirds of activity are allocated in the second 12h of each day. Days of the week revealed a

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

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 B.1.1 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

Table 3: Activity percentages along the hours of the day. Nearly identical distributions were observed on other social systems as shown in Section B.1.2.1 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.

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

Table 4: 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

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.

3.0.2 Stable sizes of Erdös sectors

The distribution of vertices in the hub, intermediary and periphery Erdös sectors is remarkably stable along time if the snapshots hold 200 or more messages, as it is clear in Figure 4 and in Section B.3 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.[?] 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 B.3 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 Erdös sectors.

In order to support the generality of these findings, we list the Erdös sector sizes of 12 networks from Facebook, Twitter and Participabr in Table 75 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.

3.0.3 Stability of principal components

The principal components of the participants are very stable in the topological space, i.e. in the space of network measures. 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 not surprising since they are highly correlated, e.g. degree and strength have Spearman

Table 5: Activity along the days of the month cycle. Nearly identical distributions are found in all systems as indicated in Section B.1.2.3 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.

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

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 B.1.2.4 of the Supporting Information document.

<table>
<tr><th></th><th>m.</th><th>b.</th><th>t.</th><th>q.</th><th>s.</th></tr>
<tr><td>Jan</td><td>10.22</td><td rowspan="2">19.56</td><td rowspan="2">28.24</td><td rowspan="3">35.09</td><td rowspan="4">49.16</td></tr>
<tr><td>Fev</td><td>9.34</td></tr>
<tr><td>Mar</td><td>8.67</td><td rowspan="2">15.53</td></tr>
<tr><td>Apr</td><td>6.86</td><td rowspan="2">20.93</td></tr>
<tr><td>Mai</td><td>7.28</td><td rowspan="2">14.07</td><td rowspan="3">30.36</td></tr>
<tr><td>Jun</td><td>6.80</td></tr>
<tr><td>Jul</td><td>8.97</td><td rowspan="2">16.29</td><td rowspan="2">24.47</td><td rowspan="5">50.84</td></tr>
<tr><td>Ago</td><td>7.32</td></tr>
<tr><td>Set</td><td>8.18</td><td rowspan="2">16.25</td><td rowspan="3">26.36</td><td rowspan="3">34.55</td></tr>
<tr><td>Out</td><td>8.06</td></tr>
<tr><td>Nov</td><td>7.64</td><td rowspan="2">18.30</td></tr>
<tr><td>Dez</td><td>10.66</td></tr>
</table>

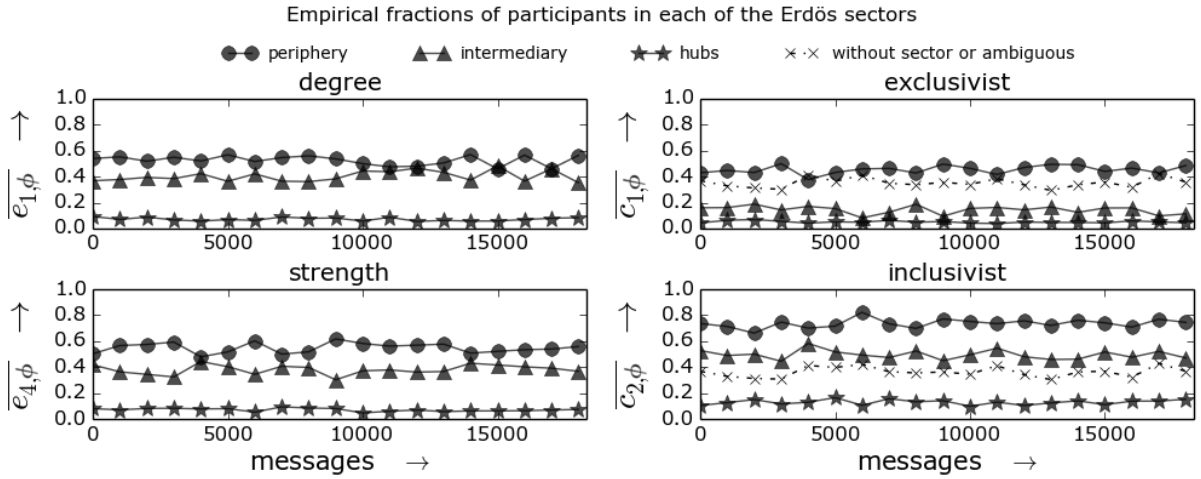


Figure 4: Stability of Erdös sector sizes. Fractions of participants derived from degree and strength criteria, E_1 and E_4 described in Section 2.2.4, are both on the left. Fractions derived from the exclusivist C_1 and the inclusivist C_2 compound criteria are shown in the plots to the right. The ordinates $\overline{e_{\gamma,\phi}} = \frac{|e_{\gamma,\phi}|}{N}$ 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 B.3 and B.4 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.

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.

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%)

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.

	PC1		PC2		PC3	
	μ	σ	μ	σ	μ	σ
cc	0.89	0.59	1.93	1.33	21.22	2.97
s	11.71	0.57	2.97	0.82	2.45	0.72
s^{in}	11.68	0.58	2.37	0.91	3.08	0.78
s^{out}	11.49	0.61	3.63	0.79	1.61	0.88
k	11.93	0.54	2.58	0.70	0.52	0.44
k^{in}	11.93	0.52	1.19	0.88	1.41	0.71
k^{out}	11.57	0.61	4.34	0.70	0.98	0.66
bt	11.37	0.55	2.44	0.84	1.37	0.77
asy	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
dis	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

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

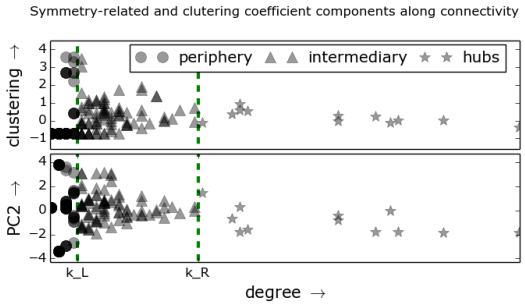


Figure 5: The first plot highlights the well-known pattern of degree versus clustering coefficient, characterized by the higher clustering coefficient of lower degree vertices. The second plot shows the greater dispersion of the symmetry-related ordinates dominant in the second principal component (PC2). 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.

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 B.2 and B.4 of the Supporting Information 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.

3.0.4 Types from Erdös 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 Erdös sector to which a participant belongs can be regarded as implying a social type for this participant. In this case, the type of a participant changes both along time and as different networks are considered, despite the stability of the network. Therefore, the potential for prejudice of such participant typology is attenuated.[?] 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.[?]

The importance of this issue can be grasped by the consideration of static types

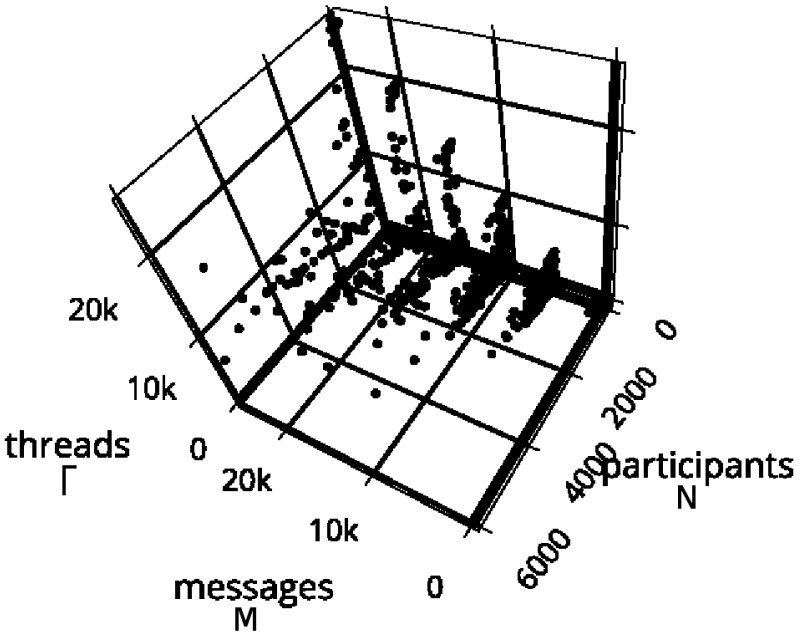


Figure 6: A scatter plot of number of messages M versus number of participants N versus number of threads Γ for 140 email lists. Highest Γ is associated with low N. The correlation between N and Γ is negative for low values of N but positive otherwise. This negative correlation between N and Γ can also 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.

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 Erdös sectors and the implicit participant types were made, which are consistent with the literature[?]: 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.

3.0.5 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 Erdös Sector and on the activity along different timescales. In fact, these findings confirm our initial hypothesis - based on the literature? - 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.

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 Erdös 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.? 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 analysis 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?

3.1 Text results and discussion

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

The next sections summarize results of immediate interest and further insights can be obtained by browsing through the tables and figures of the Appendix C.

3.1.1 General characteristics of activity distribution among participants

	g.	p.	i.	h.
N	17	7	6	4
$N_{\%}$	100.00	41.18	35.29	23.53
M	100.00	11.00	37.00	52.00
$M_{\%}$	100.00	11.00	37.00	52.00
Γ	18.00	4.00	8.00	6.00
$\Gamma_{\%}$	100.00	22.22	44.44	33.33
$\frac{\Gamma}{M}\%$	18.00	36.36	21.62	11.54
$\mu(\gamma)$	2.67	2.50	2.75	2.67
$\sigma(\gamma)$	0.47	0.50	0.43	0.47

Table 9: Distribution of participants, messages and threads among each Erdös sector (p. for periphery, i. for intermediary, h. for hubs) in a total time period of 0.71 years (from 2007-04-24T18:54:28 to 2008-01-10T19:17:26). 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.

Hubs and periphery swap fractions of participants and activity: while peripheral sector has $\approx 75\%$ of participants, it produces $\approx 10\%$ of all messages. Conversely, hubs sector present $\approx 10\%$ of participants and produces $\approx 75\%$ of all messages. Fewer threads are created by the hubs in proportion to total messages sent, while threads created by the periphery are twice as frequent as general messages. This suggests a complementarity

between peripheral diversity and hub specialization which, on its turn, deepens the understanding of the interaction network as a meaningful system, notably if yield by online activity. These assertions are condensed in Table 9.

3.1.2 Characters

	g.	p.	i.	h.
$chars$	82933	7162	28170	47601
$chars_{\%}$	100.00	8.64	33.97	57.40
$\frac{spaces}{chars}$	14.96	13.59	15.21	15.01
$\frac{punct}{chars-spaces}$	8.17	6.98	8.03	8.44
$\frac{digits}{chars-spaces}$	0.90	1.97	0.77	0.80
$\frac{letters}{chars-spaces}$	88.72	88.88	88.98	88.54
$\frac{vogals}{letters}$	40.47	39.17	40.72	40.53
$\frac{uppercase}{letters}$	5.27	6.22	5.39	5.05

Table 10: Characters in each Erdös sector (p. for periphery, i. for intermediary, h. for hubs).

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

3.1.3 Tokens and words

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

Punctuations among tokens are less abundant in hubs, and discrepancies here are larger than with character comparisons (subsection 3.1.2). Known words are used more frequently by hubs.

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

	g.	p.	i.	h.
$tokens$	17964	1539	6064	10361
$tokens_{\%}$	100.00	8.57	33.76	57.68
$tokens \neq$	15.21	32.16	25.89	18.02
$\frac{knownw}{tokens}$	36.48	35.74	38.03	35.69
$\frac{knownw \neq}{knownw}$	8.62	24.73	15.31	10.84
$\frac{stopw}{knownw}$	11.43	10.00	11.71	11.47
$\frac{punct}{tokens}$	23.73	22.35	22.82	24.47
$\frac{contrac}{tokens}$	0.01	0.00	0.00	0.02
$\mu(\overline{tokens})$	3.84	3.94	3.86	3.82
$\sigma(\overline{tokens})$	2.99	3.10	2.96	2.99
$\mu(\overline{knownw})$	3.28	3.27	3.32	3.26
$\sigma(\overline{knownw})$	1.81	1.86	1.80	1.81
$\mu(\overline{knownw \neq})$	5.12	4.35	4.92	4.98
$\sigma(\overline{knownw \neq})$	2.20	2.22	2.25	2.15
$\mu(\overline{stopw})$	1.77	1.84	1.77	1.76
$\sigma(\overline{stopw})$	0.74	0.76	0.73	0.75

Table 11: Token sizes in each Erdös sector (p. for periphery, i. for intermediary, h. for hubs).

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

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

3.1.4 Sizes of tokens and words

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

3.1.5 Sizes of sentences

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

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

	g.	p.	i.	h.
$sents$	558	45	211	304
$sents_{\%}$	99.64	8.04	37.68	54.29
$\mu_S(chars)$	147.51	158.07	132.44	155.44
$\sigma_S(chars)$	147.95	154.11	135.56	154.01
$\mu_S(tokens)$	32.21	34.22	28.75	34.09
$\sigma_S(tokens)$	32.08	32.64	29.17	33.60
$\mu_S(knownw)$	9.90	9.82	9.10	10.39
$\sigma_S(knownw)$	10.37	11.34	9.14	10.92
$\mu_S(stopw)$	1.18	1.11	1.15	1.20
$\sigma_S(stopw)$	1.57	1.22	1.52	1.64
$\mu_S(puncts)$	7.65	7.67	6.57	8.35
$\sigma_S(puncts)$	11.30	8.19	10.69	11.98

Table 12: Sentences sizes in each Erdös sector (p. for periphery, i. for intermediary, h. for hubs).

	g.	p.	i.	h.
$msgs$	100	11	37	52
$msgs_{\%}$	100.00	11.00	37.00	52.00
$\mu_M(sents)$	6.54	4.91	6.62	6.83
$\sigma_M(sents)$	5.09	3.26	6.28	4.35
$\mu_M(tokens)$	179.79	140.00	163.97	199.46
$\sigma_M(tokens)$	183.42	71.56	182.94	197.24
$\mu_M(knownw)$	55.24	40.18	51.92	60.79
$\sigma_M(knownw)$	61.67	26.49	59.82	67.33
$\mu_M(stopw)$	6.55	4.55	6.54	6.98
$\sigma_M(stopw)$	6.92	3.11	7.11	7.29
$\mu_M(puncts)$	42.77	31.36	37.46	48.96
$\sigma_M(puncts)$	48.91	14.92	48.61	52.78
$\mu_M(chars)$	829.31	651.09	761.35	915.37
$\sigma_M(chars)$	878.54	342.18	889.47	937.65

Table 13: Messages sizes in each Erdös sector (p. for periphery, i. for intermediary, h. for hubs).

3.1.6 Messages

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

	g.	p.	i.	h.
NOUN	56.40	58.68	55.67	56.50
X	16.46	16.20	16.58	16.42
ADP	7.22	6.78	7.38	7.18
DET	6.07	5.12	6.41	6.02
VERB	5.13	4.46	5.90	4.75
ADJ	3.29	3.97	2.87	3.44
ADV	2.07	2.15	1.82	2.21
PRT	1.79	1.32	1.51	2.04
PRON	1.18	0.83	1.48	1.04
NUM	0.33	0.50	0.27	0.34
CONJ	0.07	0.00	0.12	0.05
PUNC	0.00	0.00	0.00	0.00

Table 14: POS tags in each Erdös sector (p. for periphery, i. for intermediary, h. for hubs). Universal POS tags[?]: 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.

3.1.7 POS tags

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

3.1.8 Wordnet synsets

	g.	p.	i.	h.
N	88.79	91.15	87.41	89.26
ADJ	6.11	7.08	5.34	6.42
VERB	0.24	0.00	0.38	0.19
ADV	4.86	1.77	6.86	4.13
POS	21.05	22.01	21.61	20.58
POS!	72.54	74.83	71.48	72.85

Table 15: Percentage of synsets with each of the POS tags used by Wordnet. The last lines give the percentage of words considered from all of the tokens (POS) and from the words with synset (POS!). The tokens not considered are punctuations, unrecognized words, words without synsets, stopwords and words for which Wordnet has no synset tagged with POS tags . Values for each Erdös sectors are in the columns p. for periphery, i. for intermediary, h. for hubs.

	g.	p.	i.	h.
$\mu(min\,depth)$	6.52	6.41	6.52	6.53
$\sigma(min\,depth)$	1.94	1.77	2.01	1.93
$\mu(max\,depth)$	7.02	6.99	7.00	7.03
$\sigma(max\,depth)$	2.13	1.99	2.20	2.11
$\mu(holonyms)$	0.65	0.73	0.65	0.63
$\sigma(holonyms)$	1.41	1.48	1.39	1.40
$\mu(meronyms)$	1.08	0.83	0.86	1.25
$\sigma(meronyms)$	4.66	2.72	4.02	5.22
$\mu(domains)$	0.07	0.07	0.10	0.06
$\sigma(domains)$	0.27	0.26	0.30	0.25
$\mu(similar)$	0.00	0.00	0.00	0.00
$\sigma(similar)$	0.00	0.00	0.00	0.00
$\mu(verb\,groups)$	0.00	0.00	0.00	0.00
$\sigma(verb\,groups)$	0.00	0.00	0.00	0.00
$\mu(lemmas)$	3.07	2.80	3.19	3.04
$\sigma(lemmas)$	2.14	1.74	2.32	2.08
$\mu(entailments)$	0.00	0.00	0.00	0.00
$\sigma(entailments)$	0.00	0.00	0.00	0.00
$\mu(hyponyms)$	3.42	3.30	3.02	3.68
$\sigma(hyponyms)$	13.17	21.19	7.42	14.14
$\mu(hypernyms)$	1.09	1.09	1.08	1.09
$\sigma(hypernyms)$	0.30	0.34	0.28	0.31

Table 16: Measures of wordnet features in each Erdös sector (p. for periphery, i. for intermediary, h. for hubs).

	g.	p.	i.	h.
entity.n.01	100.00	100.00	100.00	100.00
total	100.00	100.00	100.00	100.00

Table 17: Counts for the most incident synsets at the semantic roots in each Erdös sector (p. for periphery, i. for intermediary, h. for hubs). Yes.

	g.	p.	i.	h.
abstraction.n.06	58.86	62.14	55.58	60.29
physical_entity.n.01	41.14	37.86	44.42	39.71
total	100.00	100.00	100.00	100.00

Table 18: Counts for the most incident synsets one step from the semantic roots in each Erdös sector (p. for periphery, i. for intermediary, h. for hubs).

3.1.9 Differentiation of the texts from Erdös sectors

Results from our adaptation of the Kolmogorov-Smirnov test suggest that the texts produced by each sector are extremely different. Intermediary sectors sometimes

	g.	p.	i.	h.
matter.n.03	24.03	22.33	24.96	23.74
communication.n.02	11.97	15.86	11.26	11.76
group.n.01	11.70	13.59	9.42	12.76
relation.n.01	10.12	9.39	9.51	10.61
object.n.01	8.99	4.85	11.08	8.40
psychological_feature.n.01	8.99	6.80	8.55	9.61
measure.n.02	8.43	7.77	9.42	7.93
attribute.n.02	7.65	8.74	7.42	7.62
causal_agent.n.01	5.06	7.12	5.15	4.67
thing.n.12	3.04	3.24	3.23	2.89
process.n.06	0.03	0.32	0.00	0.00
total	100.00	100.00	100.00	100.00

Table 19: Counts for the most incident synsets two step from the semantic roots in each Erdös sector (p. for periphery, i. for intermediary, h. for hubs).

	g.	p.	i.	h.
substance.n.01	26.49	26.15	27.39	26.01
position.n.07	10.58	9.62	10.37	10.86
definite_quantity.n.01	8.40	5.77	9.23	8.32
state.n.02	8.25	10.00	8.30	7.95
event.n.01	7.62	5.77	7.16	8.19
whole.n.02	7.37	2.69	9.23	7.01
social_group.n.01	7.05	9.23	5.71	7.51
written_communication.n.01	6.98	8.85	5.91	7.32
person.n.01	5.71	8.08	5.91	5.21
message.n.02	5.29	8.46	4.36	5.34
body_of_water.n.01	3.21	3.08	3.42	3.10
cognition.n.01	3.03	2.31	3.01	3.17
total	100.00	100.00	100.00	100.00

Table 20: Counts for the most incident synsets three step from the semantic roots in each Erdös sector (p. for periphery, i. for intermediary, h. for hubs).

exhibit greater differences from periphery and hubs than these extreme sectors themselves (Tables **??** and **??**). This differentiation of the three sectors is a strong indicative that the Erdös Sectioning described in[?] reveals meaningful sectors of the networks.

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

- Differences of textual production of the Erdös sectors are extreme. This can be noticed from the high values on these tables, beyond reference values used for the acceptance of the null hypothesis (see Section 2.2.8.1).
- Differences between sectors on the same network (Tables **??**, **??**, **??** and **??**) are

	g.	p.	i.	h.
$\mu(min\,depth)$	0.00	0.00	0.00	0.00
$\sigma(min\,depth)$	0.00	0.00	0.00	0.00
$\mu(max\,depth)$	0.00	0.00	0.00	0.00
$\sigma(max\,depth)$	0.00	0.00	0.00	0.00
$\mu(holonyms)$	0.00	0.00	0.00	0.00
$\sigma(holonyms)$	0.00	0.00	0.00	0.00
$\mu(meronyms)$	0.00	0.00	0.00	0.00
$\sigma(meronyms)$	0.00	0.00	0.00	0.00
$\mu(domains)$	0.02	0.00	0.03	0.02
$\sigma(domains)$	0.15	0.00	0.17	0.15
$\mu(similar)$	4.29	4.00	3.63	4.68
$\sigma(similar)$	4.41	3.50	3.25	4.99
$\mu(verb\,groups)$	0.00	0.00	0.00	0.00
$\sigma(verb\,groups)$	0.00	0.00	0.00	0.00
$\mu(lemmas)$	2.07	2.08	2.49	1.85
$\sigma(lemmas)$	1.79	1.50	2.06	1.64
$\mu(entailments)$	0.00	0.00	0.00	0.00
$\sigma(entailments)$	0.00	0.00	0.00	0.00
$\mu(hyponyms)$	0.00	0.00	0.00	0.00
$\sigma(hyponyms)$	0.00	0.00	0.00	0.00
$\mu(hypernyms)$	0.00	0.00	0.00	0.00
$\sigma(hypernyms)$	0.00	0.00	0.00	0.00

Table 21: Measures of wordnet features in each Erdös sector (p. for periphery, i. for intermediary, h. for hubs).

	g.	p.	i.	h.
public.a.01	40.34	57.14	33.33	40.59
temporal.s.01	16.48	0.00	11.11	22.77
chief.s.01	13.64	0.00	24.07	10.89
new.a.01	6.25	4.76	1.85	8.91
global.s.01	4.55	9.52	5.56	2.97
ocular.a.02	3.98	19.05	1.85	1.98
variable.a.01	3.41	0.00	1.85	4.95
impermanent.a.01	2.84	0.00	5.56	1.98
simple.a.01	2.27	4.76	3.70	0.99
alive.a.01	2.27	0.00	7.41	0.00
standard.a.01	2.27	0.00	0.00	3.96
virtual.s.01	1.70	4.76	3.70	0.00
total	100.00	100.00	100.00	100.00

Table 22: Counts for the most incident synsets at the semantic roots in each Erdös sector (p. for periphery, i. for intermediary, h. for hubs). Yes.

greater than differences between same sector from distinct lists (Tables **??**, **??**, **??**

	g.	p.	i.	h.
$\mu(min\,depth)$	1.72	2.83	1.70	1.66
$\sigma(min\,depth)$	1.38	1.77	1.17	1.51
$\mu(max\,depth)$	1.72	2.83	1.70	1.66
$\sigma(max\,depth)$	1.38	1.77	1.17	1.51
$\mu(holonyms)$	0.00	0.00	0.00	0.00
$\sigma(holonyms)$	0.00	0.00	0.00	0.00
$\mu(meronyms)$	0.00	0.00	0.00	0.00
$\sigma(meronyms)$	0.00	0.00	0.00	0.00
$\mu(domains)$	0.21	0.00	0.26	0.18
$\sigma(domains)$	0.41	0.00	0.44	0.39
$\mu(similar)$	0.00	0.00	0.00	0.00
$\sigma(similar)$	0.00	0.00	0.00	0.00
$\mu(verb\,groups)$	0.23	0.50	0.22	0.23
$\sigma(verb\,groups)$	0.48	0.50	0.49	0.47
$\mu(lemmas)$	3.42	1.67	3.49	3.48
$\sigma(lemmas)$	2.09	0.75	1.97	2.23
$\mu(entailments)$	0.12	0.00	0.16	0.09
$\sigma(entailments)$	0.32	0.00	0.36	0.29
$\mu(hyponyms)$	6.52	3.00	7.97	5.27
$\sigma(hyponyms)$	13.64	2.71	18.33	6.38
$\mu(hypernyms)$	0.78	0.83	0.82	0.73
$\sigma(hypernyms)$	0.42	0.37	0.38	0.45

Table 23: Measures of wordnet features in each Erdös sector (p. for periphery, i. for intermediary, h. for hubs).

	g.	p.	i.	h.
think.v.03	24.52	0.00	25.00	25.33
act.v.01	16.13	75.00	13.16	16.00
change.v.01	12.26	0.00	11.84	13.33
change.v.02	11.61	25.00	11.84	10.67
include.v.01	9.68	0.00	10.53	9.33
make.v.03	7.74	0.00	11.84	4.00
affect.v.05	3.23	0.00	2.63	4.00
work.v.01	3.23	0.00	0.00	6.67
be.v.01	3.23	0.00	1.32	5.33
travel.v.01	3.23	0.00	5.26	1.33
use.v.01	2.58	0.00	3.95	1.33
insist.v.01	2.58	0.00	2.63	2.67
total	100.00	100.00	100.00	100.00

Table 24: Counts for the most incident synsets at the semantic roots in each Erdös sector (p. for periphery, i. for intermediary, h. for hubs). Yes.

and **??**).

	g.	p.	i.	h.
reason.v.01	27.73	0.00	28.33	28.57
change_shape.v.01	13.45	0.00	13.33	14.29
end.v.02	13.45	0.00	13.33	14.29
interact.v.01	13.45	100.00	10.00	12.50
try.v.01	7.56	0.00	6.67	8.93
create_verbally.v.01	5.04	0.00	10.00	0.00
surprise.v.01	4.20	0.00	3.33	5.36
assert.v.01	3.36	0.00	3.33	3.57
evaluate.v.02	3.36	0.00	3.33	3.57
specify.v.03	3.36	0.00	1.67	5.36
return.v.01	2.52	0.00	3.33	1.79
discard.v.01	2.52	0.00	3.33	1.79
total	100.00	100.00	100.00	100.00

Table 25: Counts for the most incident synsets one step from the semantic roots in each Erdös sector (p. for periphery, i. for intermediary, h. for hubs).

	g.	p.	i.	h.
deduce.v.01	32.67	0.00	32.69	35.56
start.v.14	15.84	0.00	15.38	17.78
interrupt.v.04	15.84	0.00	15.38	17.78
relate.v.05	9.90	25.00	9.62	8.89
write.v.01	5.94	0.00	11.54	0.00
catch.v.01	4.95	0.00	3.85	6.67
communicate.v.02	3.96	50.00	0.00	4.44
dump.v.01	2.97	0.00	3.85	2.22
treat.v.01	1.98	0.00	1.92	2.22
name.v.01	1.98	0.00	3.85	0.00
correct.v.01	1.98	25.00	1.92	0.00
represent.v.09	1.98	0.00	0.00	4.44
total	100.00	100.00	100.00	100.00

Table 26: Counts for the most incident synsets two step from the semantic roots in each Erdös sector (p. for periphery, i. for intermediary, h. for hubs).

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

3.1.10 Correlation of topological and textual metrics

Correlation of degree and strength metrics is substantially smaller for intermediary sector. Also noteworthy is the negative correlation of degree and message size (number of characters, tokens or sentences) that intermediaries presented. This and other insights

	g.	p.	i.	h.
disrespect.v.01	35.71	25.00	45.45	30.77
inform.v.01	14.29	50.00	0.00	15.38
map.v.01	7.14	0.00	0.00	15.38
object.v.01	7.14	0.00	9.09	7.69
ignore.v.01	7.14	0.00	9.09	7.69
prefer.v.03	7.14	0.00	18.18	0.00
debug.v.01	7.14	25.00	9.09	0.00
double.v.01	3.57	0.00	0.00	7.69
adhere.v.06	3.57	0.00	0.00	7.69
program.v.01	3.57	0.00	0.00	7.69
roll_up.v.02	3.57	0.00	9.09	0.00
total	100.00	100.00	100.00	100.00

Table 27: Counts for the most incident synsets three step from the semantic roots in each Erdös sector (p. for periphery, i. for intermediary, h. for hubs).

	g.	p.	i.	h.
$\mu(min\,depth)$	0.00	nan	0.00	0.00
$\sigma(min\,depth)$	0.00	nan	0.00	0.00
$\mu(max\,depth)$	0.00	nan	0.00	0.00
$\sigma(max\,depth)$	0.00	nan	0.00	0.00
$\mu(holonyms)$	0.00	nan	0.00	0.00
$\sigma(holonyms)$	0.00	nan	0.00	0.00
$\mu(meronyms)$	0.00	nan	0.00	0.00
$\sigma(meronyms)$	0.00	nan	0.00	0.00
$\mu(domains)$	0.11	nan	0.20	0.00
$\sigma(domains)$	0.31	nan	0.40	0.00
$\mu(similar)$	0.00	nan	0.00	0.00
$\sigma(similar)$	0.00	nan	0.00	0.00
$\mu(verb\,groups)$	0.00	nan	0.00	0.00
$\sigma(verb\,groups)$	0.00	nan	0.00	0.00
$\mu(lemmas)$	2.00	nan	1.20	3.00
$\sigma(lemmas)$	1.33	nan	0.40	1.41
$\mu(entailments)$	0.00	nan	0.00	0.00
$\sigma(entailments)$	0.00	nan	0.00	0.00
$\mu(hyponyms)$	0.00	nan	0.00	0.00
$\sigma(hyponyms)$	0.00	nan	0.00	0.00
$\mu(hypernyms)$	0.00	nan	0.00	0.00
$\sigma(hypernyms)$	0.00	nan	0.00	0.00

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

can be drawn from Tables **??**, **??** and **??**. Overall, negligible correlation is found between textual and topological metrics.

	g.	p.	i.	h.
still.r.01	44.44	nan	80.00	0.00
promptly.r.02	22.22	nan	0.00	50.00
overhead.r.01	11.11	nan	0.00	25.00
well.r.01	11.11	nan	20.00	0.00
entirely.r.02	11.11	nan	0.00	25.00
total	100.00	nan	100.00	100.00

Table 29: Counts for the most incident synsets at the semantic roots in each Erdös sector (p. for periphery, i. for intermediary, h. for hubs). Yes.

	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

Table 30: KS distances on size of tokens. TAG: 6

	g.	p.	i.	h.
g.	0.000	2.920	7.301	4.728
	0.000	0.018	0.095	0.051
p.	2.920	0.000	8.522	5.895
	0.018	0.000	0.112	0.065
i.	7.301	8.522	0.000	6.307
	0.095	0.112	0.000	0.100
h.	4.728	5.895	6.307	0.000
	0.051	0.065	0.100	0.000

Table 31: KS distances on size of known words. TAG: 6

	g.	p.	i.	h.
g.	0.000	1.192	1.491	1.551
	0.000	0.026	0.073	0.047
p.	1.192	0.000	1.977	2.194
	0.026	0.000	0.098	0.070
i.	1.491	1.977	0.000	2.078
	0.073	0.098	0.000	0.113
h.	1.551	2.194	2.078	0.000
	0.047	0.070	0.113	0.000

Table 32: KS distances on size of sentences. TAG: 6

	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

Table 33: KS distances on use of adjectives on sentences. TAG: 3

	g.	p.	i.	h.
g.	0.000 0.000	1.334 0.033	1.124 0.020	1.538 0.024
p.	1.334 0.033	0.000 0.000	0.578 0.016	2.172 0.057
i.	1.124 0.020	0.578 0.016	0.000 0.000	2.206 0.044
h.	1.538 0.024	2.172 0.057	2.206 0.044	0.000 0.000

Table 34: KS distances on use of substantives on sentences. TAG: 3

	g.	p.	i.	h.
g.	0.000 0.000	1.484 0.036	0.978 0.017	1.277 0.020
p.	1.484 0.036	0.000 0.000	0.349 0.010	2.157 0.056
i.	0.978 0.017	0.349 0.010	0.000 0.000	1.739 0.035
h.	1.277 0.020	2.157 0.056	1.739 0.035	0.000 0.000

Table 35: KS distances on use of punctuations on sentences. TAG: 3

3.1.11 Formation of principal components

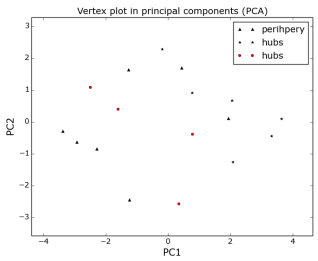


Figure 7: First two principal components.

	cc	d	s	$\mu_S(p)$	$\sigma_S(p)$	$\mu_S(kw)$	$\sigma_S(kw)$	$\mu_S(sw)$	$\sigma_S(sw)$
cc	1.06	0.18	0.11	-0.15	0.13	0.36	0.39	-0.03	0.09
(p.)	1.17	0.85	0.20	-0.12	-0.04	0.47	0.41	0.06	0.62
(i.)	1.20	-0.37	-0.04	-0.01	0.17	0.53	0.62	-0.28	-0.37
(h.)	1.33	-1.21	-0.83	0.33	-0.41	0.17	0.52	0.51	0.66
d	0.18	1.06	0.99	0.18	0.53	0.38	0.27	0.26	0.57
	0.85	1.17	0.93	-0.11	0.02	0.57	0.60	-0.17	0.75
	-0.37	1.20	0.87	0.61	0.44	0.32	0.21	1.02	1.04
	-1.21	1.33	0.84	-0.43	0.77	-0.30	-0.55	-0.55	-0.71
s	0.11	0.99	1.06	0.10	0.41	0.18	0.11	0.08	0.34
	0.20	0.93	1.17	-0.06	0.07	0.40	0.51	-0.30	0.54
	-0.04	0.87	1.20	0.27	0.21	0.08	-0.15	0.85	0.91
	-0.83	0.84	1.33	-1.22	-0.23	-1.14	-1.28	-1.28	-1.32
$\mu_S(p)$	-0.15	0.18	0.10	1.06	0.78	0.67	0.51	0.70	0.39
	-0.12	-0.11	-0.06	1.17	1.08	0.68	0.41	0.76	0.12
	-0.01	0.61	0.27	1.20	1.15	1.03	0.78	0.92	0.79
	0.33	-0.43	-1.22	1.33	0.45	1.32	1.31	1.32	1.29
$\sigma_S(p)$	0.13	0.53	0.41	0.78	1.06	0.82	0.70	0.54	0.57
	-0.04	0.02	0.07	1.08	1.17	0.86	0.71	0.51	0.47
	0.17	0.44	0.21	1.15	1.20	1.12	0.91	0.86	0.73
	-0.41	0.77	-0.23	0.45	1.33	0.47	0.45	0.44	0.29
$\mu_S(kw)$	0.36	0.38	0.18	0.67	0.82	1.06	0.97	0.64	0.67
	0.47	0.57	0.40	0.68	0.86	1.17	1.10	0.44	0.79
	0.53	0.32	0.08	1.03	1.12	1.20	1.11	0.66	0.52
	0.17	-0.30	-1.14	1.32	0.47	1.33	1.28	1.28	1.23
$\sigma_S(kw)$	0.39	0.27	0.11	0.51	0.70	0.97	1.06	0.36	0.48
	0.41	0.60	0.51	0.41	0.71	1.10	1.17	0.10	0.92
	0.62	0.21	-0.15	0.78	0.91	1.11	1.20	0.43	0.31
	0.52	-0.55	-1.28	1.31	0.45	1.28	1.33	1.33	1.32
$\mu_S(sw)$	-0.03	0.26	0.08	0.70	0.54	0.64	0.36	1.06	0.76
	0.06	-0.17	-0.30	0.76	0.51	0.44	0.10	1.17	-0.19
	-0.28	1.02	0.85	0.92	0.86	0.66	0.43	1.20	1.18
	0.51	-0.55	-1.28	1.32	0.44	1.28	1.33	1.33	1.32
$\sigma_S(sw)$	0.09	0.57	0.34	0.39	0.57	0.67	0.48	0.76	1.06
	0.62	0.75	0.54	0.12	0.47	0.79	0.92	-0.19	1.17
	-0.37	1.04	0.91	0.79	0.73	0.52	0.31	1.18	1.20
	0.66	-0.71	-1.32	1.29	0.29	1.23	1.32	1.32	1.33

Table 36: Pierson correlation coefficient for the topological and textual measures.

Principal components formation seem to be the less stable of all results reported in this study. First component, with $\approx 25\%$ of dispersion, relies heavily on POS tags, and slightly on sizes of tokens, sentences and messages. Second component, with almost 12% of dispersion, blends topology, POS tags and size of textual units. Third component, with about 8.5% of dispersion is mostly nouns frequency and size of textual units. Fourth and fifth components present less than 5% of total dispersion, but are included in the

	PC1	PC2	PC3	PC4	PC5
cc	-3.84	-5.10	-31.80	9.76	-22.06
(p.)	-9.03	-10.13	-30.55	-4.77	-11.17
(i.)	0.26	-15.63	37.08	-3.17	6.95
(h.)	-7.00	22.74	28.27	-5.22	-5.22
d	-10.20	-24.84	3.64	-0.15	1.34
	-11.87	-14.57	-6.71	14.22	-5.89
	11.61	12.06	5.21	27.91	20.47
	7.44	-25.19	0.51	7.60	7.60
s	-7.41	-27.67	4.95	-7.51	-3.76
	-9.18	-12.11	17.12	24.67	1.31
	8.41	13.63	29.19	-13.73	0.58
	13.89	-3.62	-5.85	-1.29	-1.29
$\mu_S(p)$	-11.78	13.03	12.10	-15.34	-16.62
	-8.50	18.81	2.44	8.13	-16.08
	14.32	-4.67	-14.13	-11.00	25.09
	-13.53	-6.65	-10.24	-3.61	-3.61
$\sigma_S(p)$	-14.56	1.35	1.43	-14.80	-3.29
	-11.61	15.37	9.50	-3.23	-14.81
	14.03	-7.98	-10.43	-16.02	-1.90
	-3.29	-24.50	35.74	-7.33	-7.33
$\mu_S(kw)$	-14.99	7.93	-9.72	-1.21	7.61
	-16.21	4.91	-1.11	-3.44	13.21
	12.73	-13.04	0.36	-1.11	-0.99
	-12.96	-9.16	-16.09	-25.59	-25.59
$\sigma_S(kw)$	-12.65	7.94	-16.56	-7.93	18.37
	-15.68	-0.14	7.38	-10.56	17.41
	10.06	-15.45	2.09	23.68	-14.15
	-13.92	-4.00	-0.65	6.10	6.10
$\mu_S(sw)$	-11.81	9.78	13.26	19.14	-14.16
	-3.80	16.64	-21.56	15.34	14.73
	14.78	7.50	-0.47	-2.87	-9.72
	-13.92	-3.93	-1.67	27.54	27.54
$\sigma_S(sw)$	-12.76	-2.35	6.54	24.17	12.80
	-14.11	-7.33	3.64	-15.64	-5.38
	13.80	10.05	1.04	-0.51	-20.16
	-14.05	0.21	-0.98	0.83	0.83
λ	49.30	19.02	14.48	8.44	4.75
	46.67	28.95	10.95	7.95	3.63
	57.01	28.08	10.20	3.34	1.37
	70.05	24.25	5.70	0.00	0.00

Table 37: PCA formation

Supporting Information document for completeness of exposition.

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

3.1.12 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. Appendix **??** and Figures **??**-**??** are dedicated to this histogram differences.

3.2 Results from visualization

Results from versinus are divided in two groups: observations on features that made it useful for the task, and the network properties it made possible to grasp.

3.2.1 Useful visualization features for dynamic networks

Among the numerous insights related to versinus, a few of them seem more fundamental, or plain useful. Such insights were incorporated to Versinus as the result of tests which presented clear benefits within the context of our research. The folowing list is an attempt to present them in an importance-first order:

1. Vertices need to remain static. Even if they move smoothly, one notices solely transient artifacts.

2. Very connected sectors (hubs and intermediary) need to be in a curve, otherwise the edges enclose each other and reasoning about the network becomes harsh.

3. Height and width of a vertex are very informative, specially if measures 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 is 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 (e.g. all M messages) is useful for understanding how much a vertex preserves the position in different scales of activity.

Many other insights were given by Versinus, such as possible visualization tools, other kind of convenient layouts and glyph elaborations. These receives dedicated attention in Section **??**.

3.2.2 Understanding of network properties through Versinus

A number of hypotheses were drawn about the networks for which Versinus was designed. Another number of hypothesis were driven from versinus use itself.

As suggested by Palla, Barabási and Vicsek,[?] stability of participant activity in social networks is more incident in smaller networks. In accordance 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 intermitence of hubs itself was one of the top hypotheses which motivated Versinus development. The stability of the network structure, concomitant with the instability of the activity of each participant, motivated a deeper analysis.[?] 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 vertex. Other hypotheses, such as discrepancies in the authority and the degree of a vertex, are numerous but need further research to be valuable.

3.2.3 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). Measures of each vertex can be exposed with a vertical displacement, to enable multiple measures, 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, specially geometric figures and iterative force-based methods for positioning vertex 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.[?]

3.3 Linked data results

Current results include data selection and preparation for knowledge discovery. In this respect, the main result is the data made available, which enables benchmarking of scientific results and easy experimentations. Secondary results include data outline through figures and tables, software support and example SparQL queries.

3.3.1 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`.

3.3.2 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 38: 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

3.3.3 Software tools

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

3.3.3.1 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 44.

Table 39: 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[?]
Cidade Democrática	MySQL database	donated by admins	Participation
Email	mbox	obtained through Gmane public database	Gmane[?]
Facebook	GDF, GML and TAB	obtained through Netvizz[?]	Social[?]
IRC	plain text log	obtained through Supybot logging	Social
ParticipaBR	PostgreSQL database	donated by admins	Participation
Twitter	JSON	obtained through Twitter streaming API	Social

3.3.3.2 Topological and textual analysis

Routines are available for taking topological and textual measures from the database. Auxiliary routines, such as performing principal component analysis and taking Kolmogorov-Smirnov measures, are available to ease pattern recognition. Single, timeline and multi-scale analyzes are automated.

3.3.3.3 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 toolbox[?] in association with the Music and Visuals toolboxes.[?,?]

3.3.3.4 Migration from deprecated toolboxes

Routines mentioned in Sections 3.3.3.2 and 3.3.3.3 are being migrated from deprecated toolboxes[?,?] into newly designed toolboxes.[?,?]

3.3.4 Diagrams of the data and auxiliary tables

The database exploration can be assisted through diagrams which expose the structure from each provenance. Such diagrams are exemplified in the Appendix **??** and fully available in a dedicated article[?] with some tables to make it easier to understand the data provided. A simplified example is given in Figure 8 where the friendship structure of the Facebook snapshots are exposed.

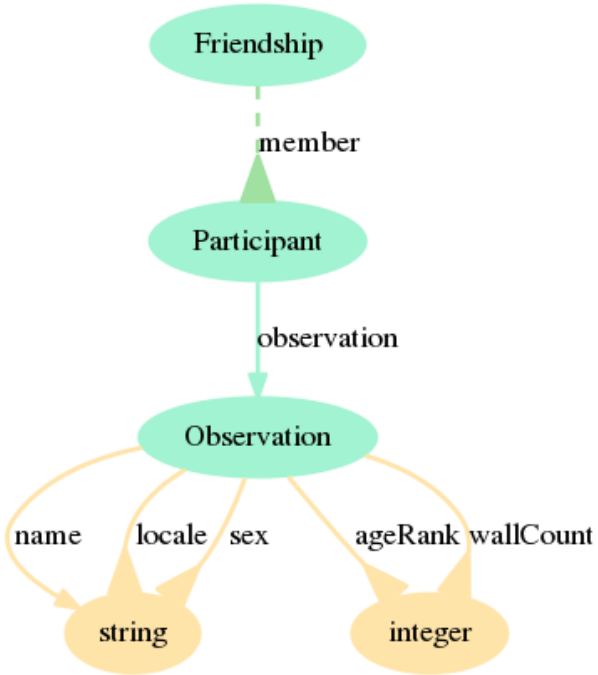


Figure 8: A diagram of the structure involved in the friendship networks of the Facebook snapshots. A green edge denotes an OWL existential class restriction; an inverted nip denotes an OWL universal class restriction; a full (non-dashed) edge denotes an OWL functional property axiom. Further information and complete diagrams for each provenance are in the dedicated article.[?]

3.3.5 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 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 an 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 .
}
```

3.3.6 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 yield the data. This rises the the understanding of the right to study such data as the right to access the self, in parity with anthropological endaviors.[?,?]

3.3.7 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 fully available in a dedicated article.[?] The method can be extended to comprise further OWL axioms and restrictions, but is currently performed to fit present needs with maximum simplicity. 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 an 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 exposed in the next section and in the Supporting Information file.

4 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 Erdös 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 participants in each Erdös 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 Supporting Information might ease hypothesizing about the generality of these characteristics.

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.[?] Also, we plan to further explore and publish the audiovisualizations used for this research[?,?] and the linguistic differences found in each of the Erdös sectors.[?]

4.1 Text final remarks

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

4.2 Linked data final remarks

The database presented in this article 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> 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 Appendices and of the articles produced in this research[?,?,?,?] for further directions on the available structures and for an overview complement.

4.2.1 Further work

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

- existence of information "ducts", observable through similarity measures. These might coincide with asymmetries of edges between vertice 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 morphosintactic 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 features distribution which is likely to reveal multimodal outlines and other non-trivial characteristics.
- Extend analysis to the windowed approach along the timelines used in this work, where hub, peripheral and intermediary sectors where topologically characterized.[?]
- For ELE list, the more connected the sector, the longer the messages are. This is the inverse of what was found in the other lists, and was considered a peculiarity of the culture bonded with the political subject of ELE list. This hypothesis should be further verified.
- Tackle the same analysis on networks with languages other than English. This is especially important for easing applications[?] and should rely on dedicated implementation of tokenization, lemmatization and attribution of POS tags.
- Observe a broader set of human interaction networks and the resulting types of networks and participants with respect to topological and textual features.
- Analyse interaction networks from other platforms such as LinkedIn, etc.
- Sentiment analysis was not approached in this work, but might be a good endavior since the subject has received considerable attention from the scientific literature but has not included topological features.

Appendix

APPENDIX A – LINKED OPEN SOCIAL DATA FOR SCIENTIFIC BENCHMARKING (DIAGRAMS)

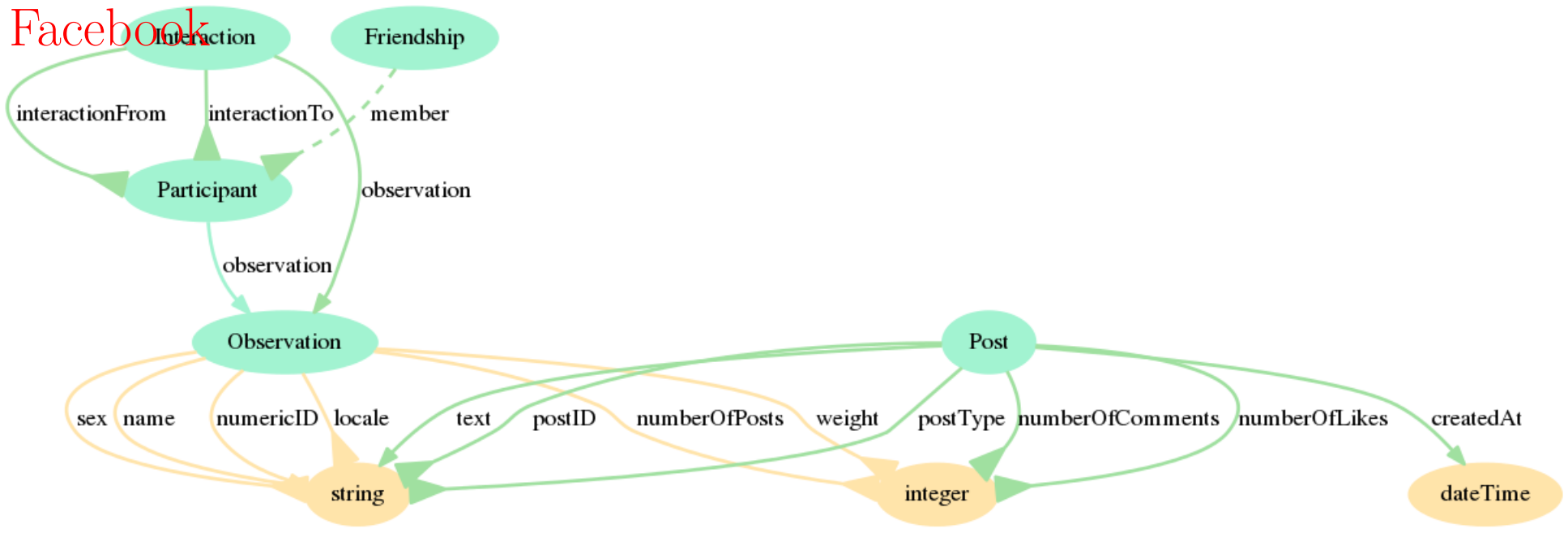
In this document we provide diagrams for the social protocols in the Linked Open Social Database: Facebook, Twitter, IRC, Email, ParticipaBR, Cidade Democrática and AA. Each social protocol diagram was broken in two, one presents the relations among main classes (blue nodes) and data types (orange nodes), the other presents metadata for the snapshots. Every class instance is related to the snapshot instance by the triple `class_uri po:snapshot snapshot_uri`. Such triples are omitted for simplicity. Due to the large number of relations, the rendering of diagrams are automatized and displays some overlaps. Even so, the images are useful for grasping what is in the current database and for conducting explorations. Edges in the diagrams have:

- green color if representing an OWL existential class restriction (all individuals from the class present at least one triple with the property as predicate);
- inverted nip if representing an OWL universal class restriction (all individuals presenting triples with the property as predicate are from the class);
- full edges (non-dashed) if representing a functional property axiom (there is at most one triple with the property as the predicate for each individual).

Furthermore, this document ends with two sets of tables, one with references for snapshot groups, such as wikipedia or contact links, the other with counts of triples, participants, edges/interactions/relations and characters.

A.1 Facebook data

Each Facebook snapshot is yield by either an user, from which the friends constitute a friendship network, or a group, which participants can yield friendship and interaction networks and posts information with text and some metadata. Further information is found on the following diagrams, the tables on the end of this document or in the main document of this article.[?]

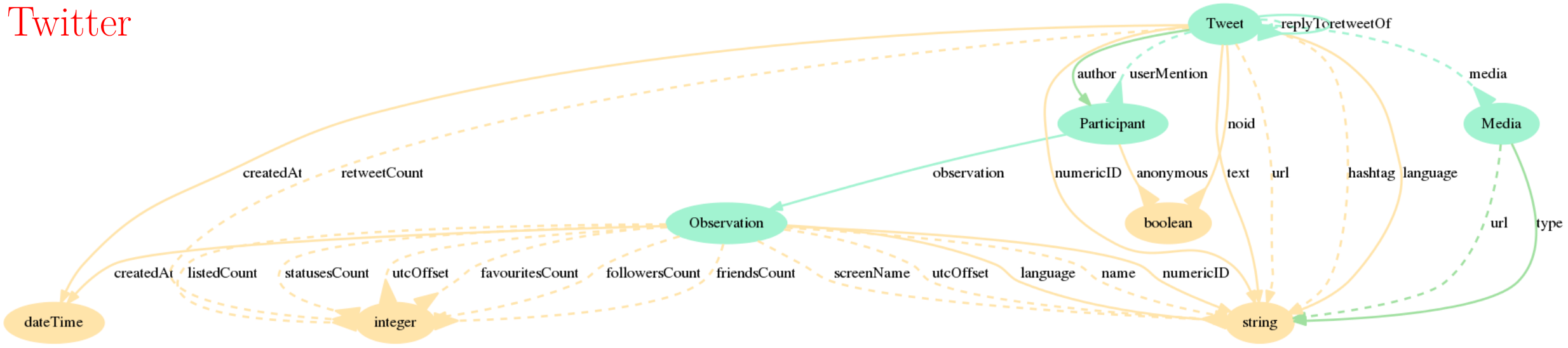

Facebook
Interaction
Friendship
interactionFrom
interactionTo
member
Participant
observation
observation
Observation
Post
sex
name
numericID
locale
text
postID
numberOfPosts
weight
postType
numberOfComments
numberOfLikes
createdAt
string
integer
dateTime

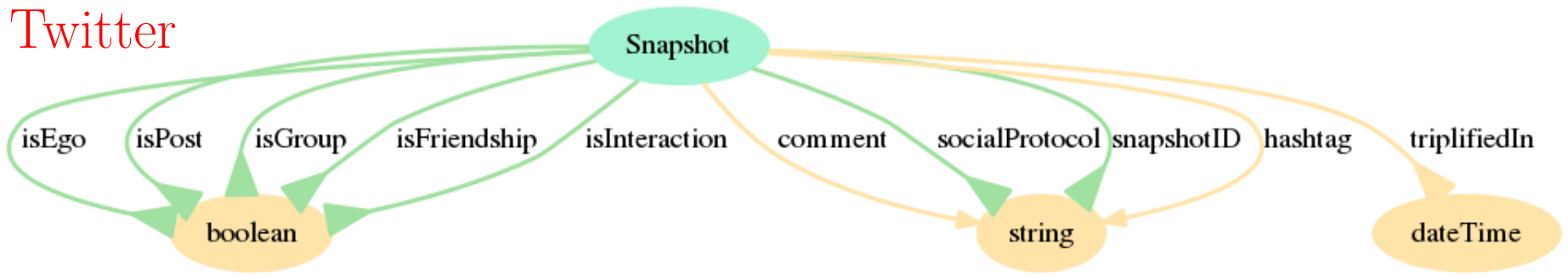
Facebook


Snapshot
snapshotID
numericID
comment
socialProtocol
acquiredThrough
name
interactionsAnonymized
isFriendship
isInteraction
isGroup
friendshipsAnonymized
isPost
isEgo
triplifiedIn
dateObtained
string
boolean
dateTime
date

A.2 Twitter data

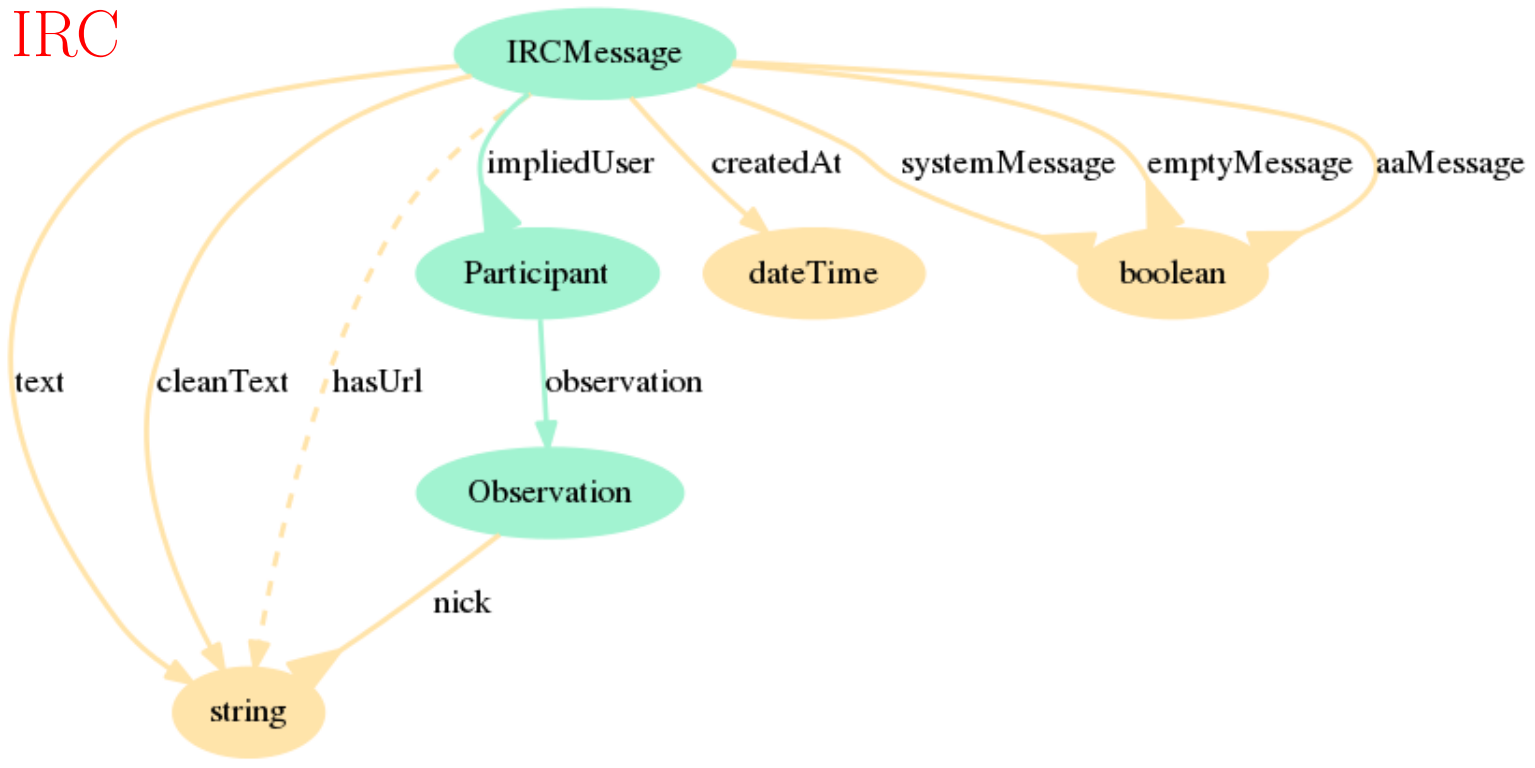
Each Twitter snapshot is yield by a hashtag. Retweets (`po:retweetOf` are usually considered the interactions between users. The database present also `po:replyTo` and `po:userMention` which might be useful in understanding the networking. Further information is found on the following diagrams, the tables on the end of this document or in the main document of this article.[?]

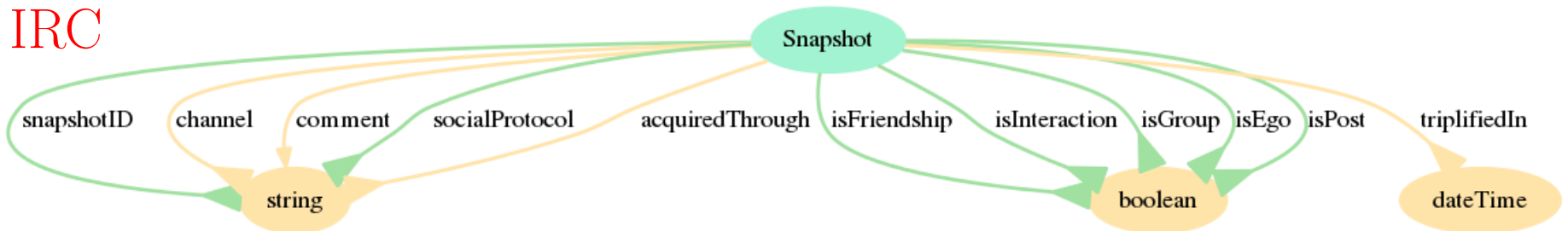

Twitter
Tweet
reply1oretweetOf
author
userMention
media
Participant
noid
Media
createdAt
retweetCount
observation
numericID
anonymous
text
url
hashtag
language
Observation
boolean
url
type
createdAt
listedCount
statusesCount
utcOffset
favouritesCount
followersCount
friendsCount
screenName
utcOffset
language
name
numericID
dateTime
integer
string


Twitter
Snapshot
isEgo
isPost
isGroup
isFriendship
isInteraction
comment
socialProtocol
snapshotID
hashtag
triplifiedIn
boolean
string
dateTime

A.3 IRC data

Each IRC snapshot is yield by an IRC channel. IRC messages are either server messages (e.g. join and exit) marked with `po:systemMessage true` and having an `po:impliedUser user_uri`, or user messages, which yield interactions through `po:directedTo` and `po:mentions` properties. Text messages without the user names are delivered through the `po:cleanText` property. Further information is found on the following diagrams, the tables on the end of this document or in the main document of this article.[?]


IRC
IRCMessage
impliedUser
createdAt
systemMessage
emptyMessage
aaMessage
Participant
dateTime
boolean
text
cleanText
hasUrl
observation
Observation
nick
string

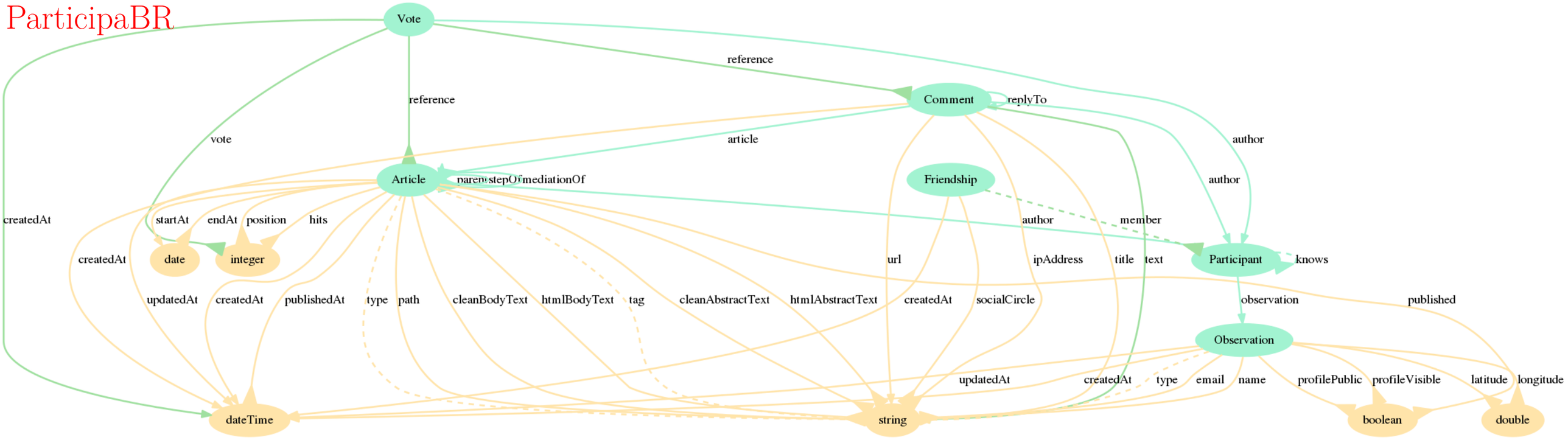

IRC
Snapshot
snapshotID
channel
comment
socialProtocol
acquiredThrough
isFriendship
isInteraction
isGroup
isEgo
isPost
triplifiedIn
string
boolean
dateTime

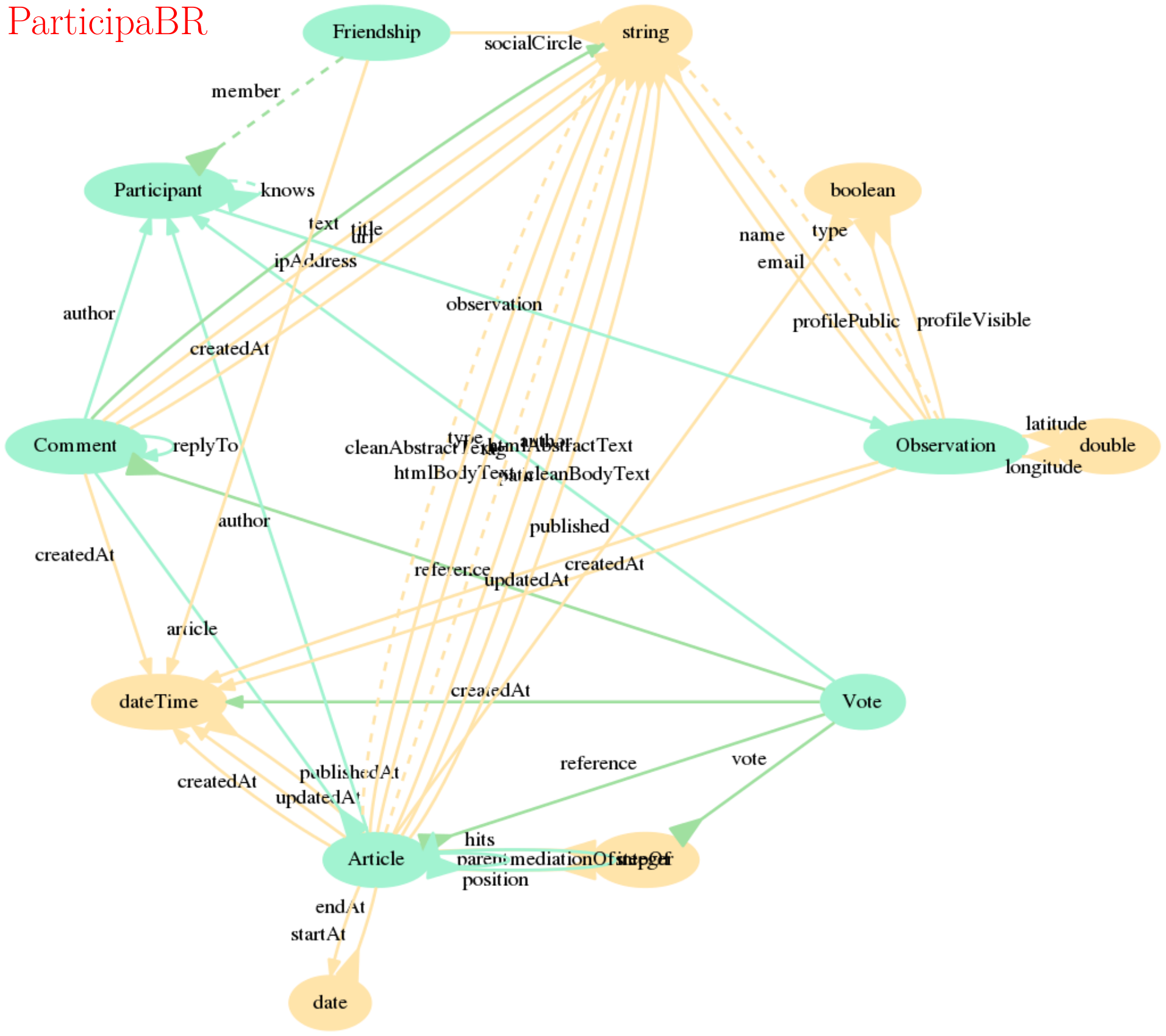
A.4 Email data

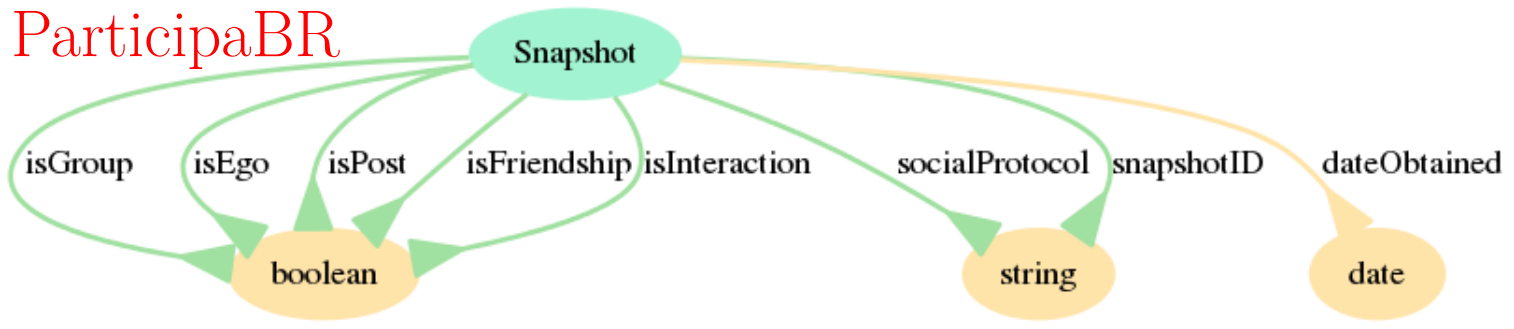
Each email snapshot is yield by an Email list. Interactions are yield through `po:replyTo` relations although `po:to` and `po:cc` might also be considered. The email body is given by `po:text` relations while `po:cleanText` links to text with lines removed where they are trivially from previous messages or computer code. Further information is found on the following diagrams, the tables on the end of this document or in the main document of this article.[?]

A.5 ParticipaBR data

The ParticipaBR snapshot is yield by a data dump donated by the system administrators of the federal portal of social participation ParticipaBR. Articles can have parent articles (`po:parent`), be step of a collection of articles (`po:stepOf`) and be a mediation of other articles (`po:mediationOf`). Interactions are yield by comments which are `po:replyTo` other comments or which are made directly to an article. This snapshot holds also friendship structures. The language used is mainly Brazilian Portuguese, but English and Spanish are also incident. Due to the higher complexity of the diagram, an additional figure is given rendered with another layout algorithm. Further information is found on the following diagrams, the tables on the end of this document or in the main document of this article.[?]


ParticipaBR
Vote
reference
reference
vote
Comment
replyTo
author
article
author
Article
parentStepOfmediationOf
Friendship
author
member
createdAt
startAt
endAt
position
hits
createdAt
date
integer
url
ipAddress
title
text
Participant
knows
updatedAt
createdAt
publishedAt
type
path
cleanBodyText
htmlBodyText
tag
cleanAbstractText
htmlAbstractText
createdAt
socialCircle
observation
published
Observation
updatedAt
createdAt
type
email
name
profilePublic
profileVisible
latitude
longitude
dateTime
string
boolean
double

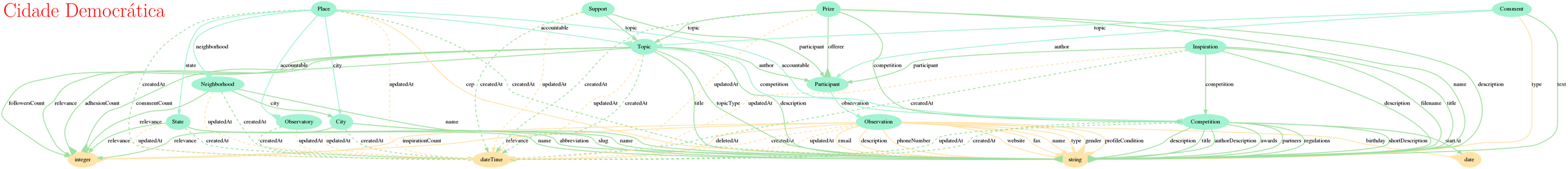

ParticipaBR
Friendship
string
socialCircle
member
Participant
knows
boolean
text
title
url
name
type
email
ipAddress
author
observation
profilePublic
profileVisible
createdAt
Comment
replyTo
cleanAbstractText
htmlAbstractText
htmlBodyText
cleanBodyText
Observation
latitude
double
longitude
author
published
createdAt
reference
updatedAt
createdAt
article
dateTime
createdAt
Vote
createdAt
publishedAt
updatedAt
reference
vote
Article
hits
position
endAt
startAt
date


ParticipaBR
Snapshot
isGroup
isEgo
isPost
isFriendship
isInteraction
socialProtocol
snapshotID
dateObtained
boolean
string
date

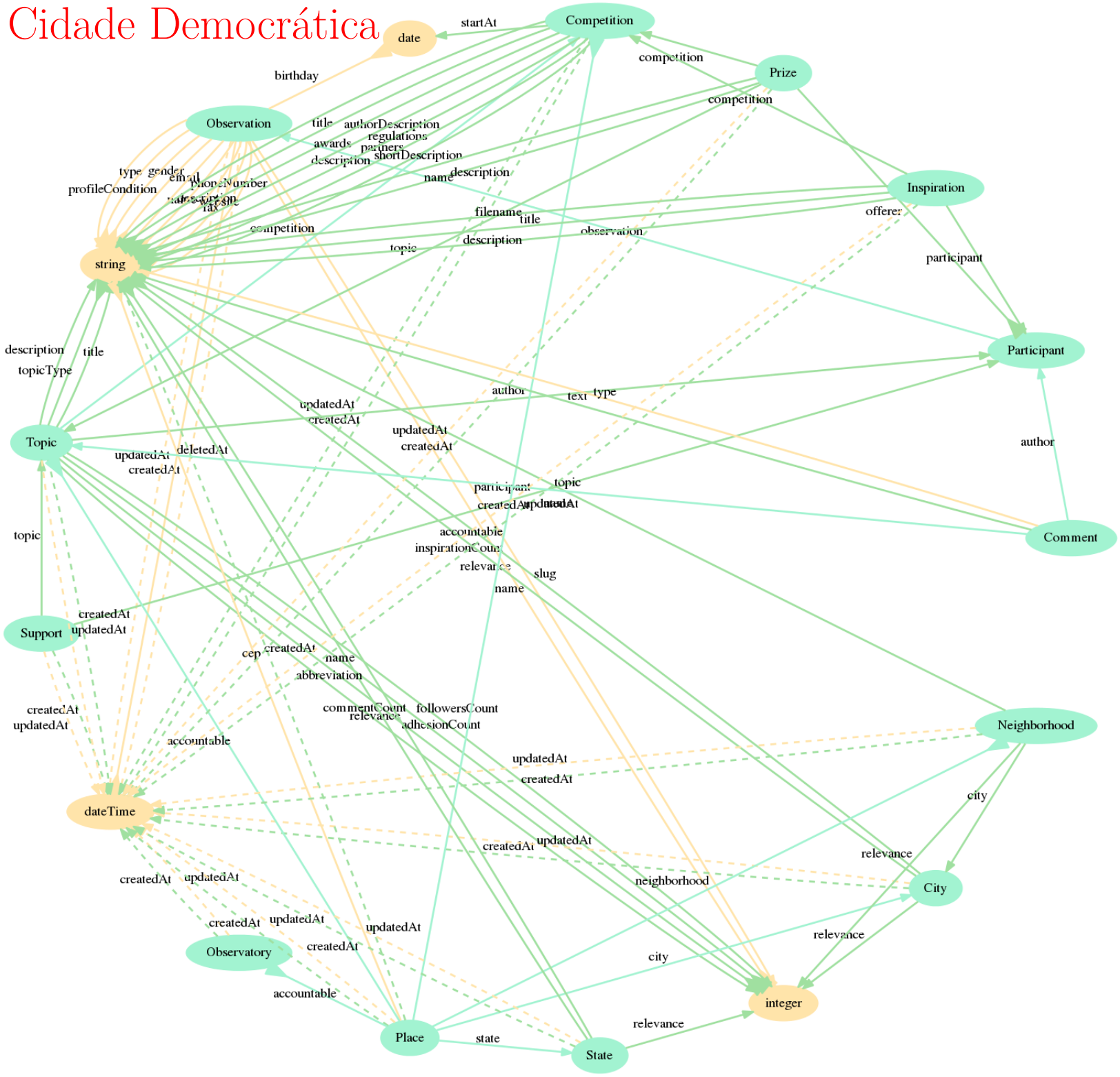
A.6 Cidade Democrática data

The Cidade Democrática snapshot is yield by a data dump donated by the system administrators of the civil society social participation portal Cidade Democrática. This snapshot holds a complex structure of both Topics/Inspirati-ons/Observatories/Supports/Competitions/Pri and of State/City/Neighbor-hood/Place. The language used is mainly Brazilian Portuguese. Due to the higher complexity of the diagram, an additional figure is given rendered with another layout algorithm. Further information is found on the following diagrams, the tables on the end of this document or in the main document of this article.[?]

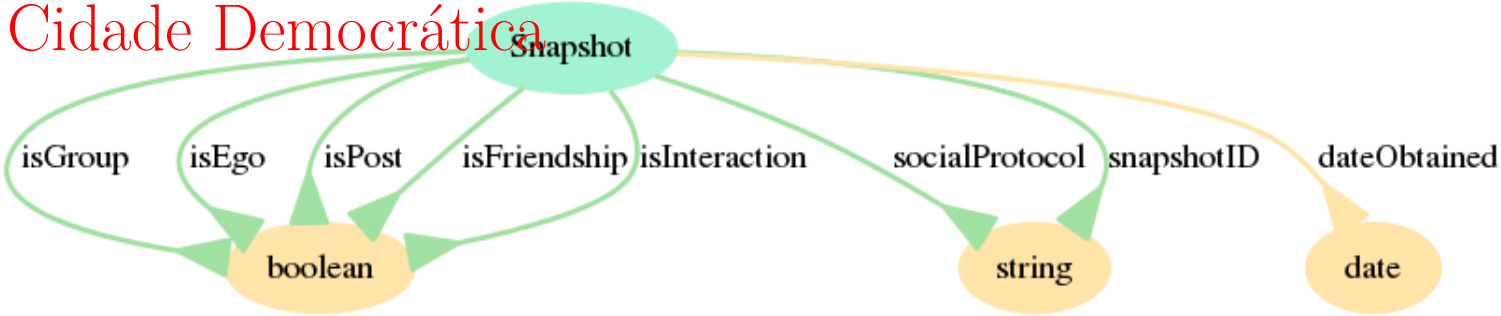
Cidade Democrática



Cidade Democrática

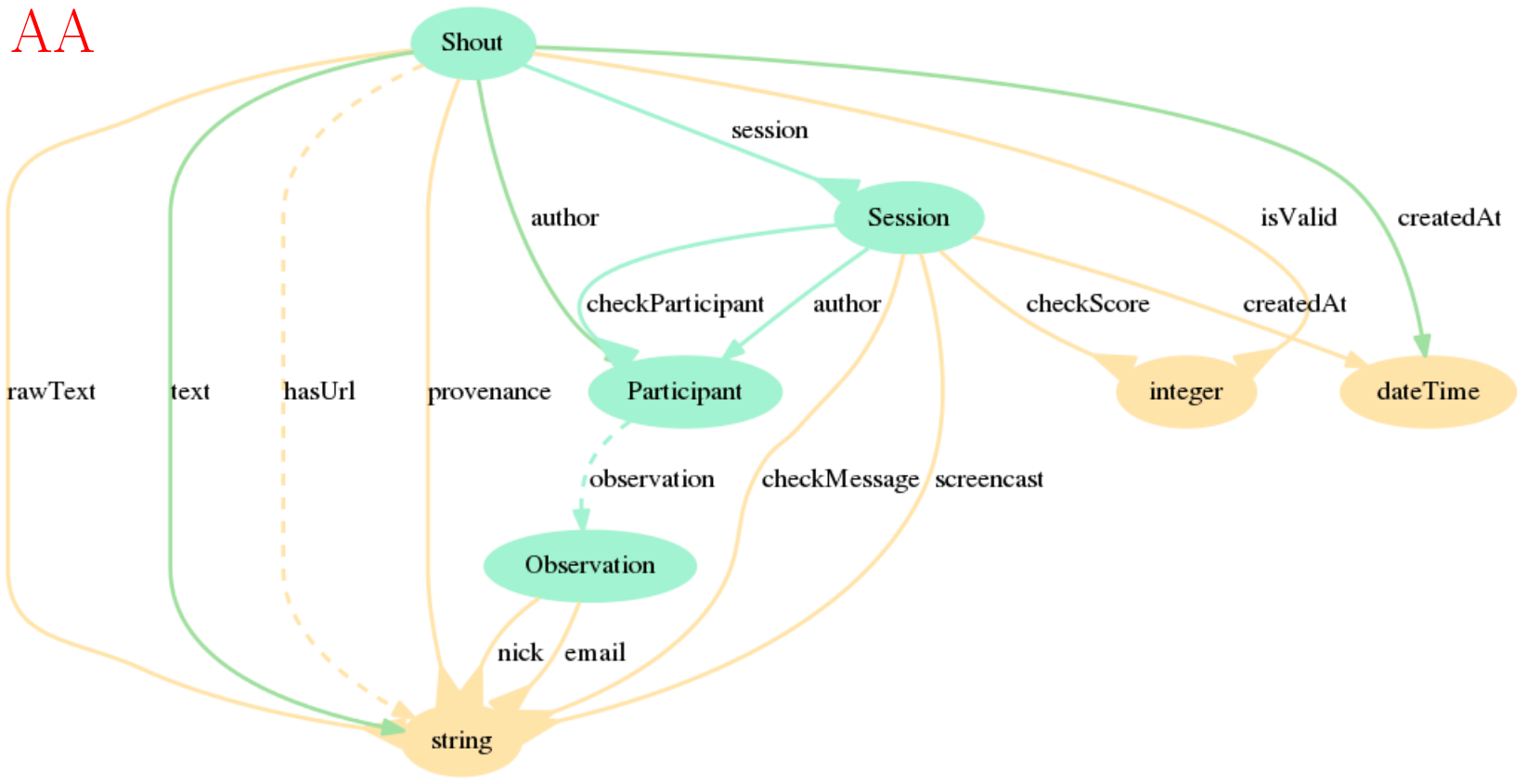


Cidade Democrática

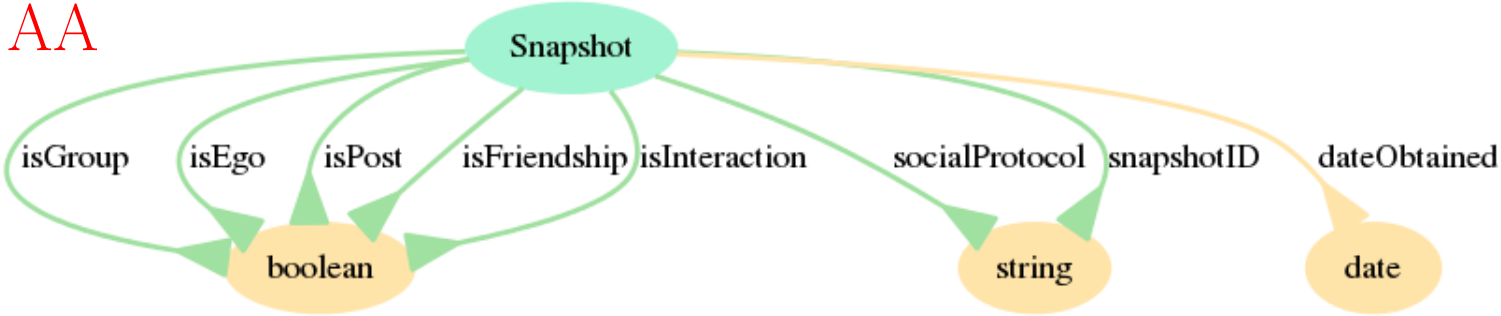


A.7 AA data

The AA (Algorithmic Autoregulation) snapshots are yield by a data dump donated by the system administrators and by a mined IRC log. The system pursue simplicity and most of data consists of detached shouts with `po:text` and `po:author`. Further information is found on the following diagrams, the tables on the end of this document or in the main document of this article.[?]


AA
Shout
session
author
Session
isValid
createdAt
checkParticipant
author
checkScore
createdAt
rawText
text
hasUrl
provenance
Participant
integer
dateTime
observation
checkMessage
screencast
Observation
nick
email
string

AA


Snapshot
isGroup
isEgo
isPost
isFriendship
isInteraction
socialProtocol
snapshotID
dateObtained
boolean
string
date

A.8 Snapshot references

Table 40: All the Facebook snapshots are either the result of individuals who downloaded their data (and donated to the first author) or data downloaded from groups. In the first case, it is senseless to present references. In the second case, we present the group name and a link to a post in the group where data and figures were delivered back to the group.

group name	**url(s)**
Adorno Nao Eh Enfeite	<https://www.facebook.com/groups/265217103529531/permalink/525654127485826/>
Ativistas Da Inclusao Digital	<https://www.facebook.com/groups/423602557691243/permalink/525201037531394/>
Ciencias Com Fronteiras	<https://www.facebook.com/groups/contraaexclusao/permalink/269103356558439/>
Computer Art	<https://www.facebook.com/groups/computerart/permalink/259389137529870/>
Coolmeia	<https://www.facebook.com/groups/coolmeia/permalink/380091142098291/> , <https://www.facebook.com/groups/coolmeia/permalink/489757754464962/>
Democracia Direta Ja	<https://www.facebook.com/groups/ddjbrasil/permalink/347023325397298/>
Democracia Pura	<https://www.facebook.com/groups/democraciapura/permalink/310907215704321/>
Economia	<https://www.facebook.com/groups/economa1/permalink/586007714743535/>
Economia Criativa Digital	<https://www.facebook.com/groups/economiacriativadigital/permalink/438313682916103/>
Educacoes E Aprendizagens XXI	<https://www.facebook.com/groups/geaxxi/permalink/433229973421182/>
Latesfip	<https://www.facebook.com/groups/183557128478424/permalink/266610616839741/>
Living Bridges Planet	<https://www.facebook.com/groups/livingbridgesplanet/permalink/352950408144951/>
Mobilizacoes Culturais Interior SP	<https://www.facebook.com/groups/131639147005593/permalink/144204529082388/>
Partido Pirata	<https://www.facebook.com/groups/partidopiratabrasil/permalink/10151409024509317/>
Politicas Culturas Brasileiras	<https://www.facebook.com/groups/pcult/permalink/519626544747423/>
Praca Popular	<https://www.facebook.com/groups/215924991863921/permalink/319279541528465/>
Rede Tranzmidias	<https://www.facebook.com/groups/318333384951196/permalink/346658712118663/>
Silicon Valley Global Network	<https://www.facebook.com/groups/109971182359978/permalink/589326757757749/>
Solidarity Economy	<https://www.facebook.com/groups/9149038282/permalink/10151461945623283/>
Study Group SNA	<https://www.facebook.com/groups/140630009439814/permalink/151470598355755/>
Tecnoxamanismo	<https://www.facebook.com/groups/505090906188661/permalink/733144993383250/> , <https://www.facebook.com/groups/505090906188661/permalink/733157380048678/>

Table 41: Different Twitter snapshots are yield by different hashtags. In this table we present each snapshot with the respective hashtag and a reference to the subject.

snapshot hashtag	**observation**	**reference**
#arenaNETmundial	a Brazilian discussion hub about free culture, democracy and the internet	<http://www.participa.br/netmundial>
#art	tweets with the generic hashtag #art	<https://en.wikipedia.org/wiki/Art>
#ChennaiFloods	heavy rainfall generated by the annual northeast monsoon in November–December 2015	<https://en.wikipedia.org/wiki/2015_South_Indian_floods>
#dilma	the 36th President of Brazil	<https://en.wikipedia.org/wiki/Dilma_Rousseff>
#ForaDilma	2015-16 anti-government protests in Brazil	<https://en.wikipedia.org/wiki/2015-16_protests_in_Brazil>
#ForaCunha	2015-16 anti-corruption protests in Brazil	<https://en.wikipedia.org/wiki/2015-16_protests_in_Brazil>
#fuck	tweets with the generic hashtag #fuck	<https://en.wikipedia.org/wiki/Fuck>
#game	tweets with the generic hashtag #game	<https://en.wikipedia.org/wiki/Game>
#god	tweets with the generic hashtag #god	<https://en.wikipedia.org/wiki/God>
#MAMA2015	the grand 2015 Mnet Asian Music Awards	<https://en.wikipedia.org/wiki/2015_Mnet_Asian_Music_Awards>
#music	tweets with the generic hashtag #music	<https://en.wikipedia.org/wiki/Music>
#obama	the 44th President of the United States	<https://en.wikipedia.org/wiki/Barack_Obama>
#python	the Python programming language	<https://en.wikipedia.org/wiki/Python_(programming_language)>
#QuartaSemRacismoClubeSDV	an anti-racism netweaving	<https://twitter.com/hashtag/quartasemracismoclubesdv>
#science	tweets with the generic hashtag #science	<https://en.wikipedia.org/wiki/Science>
#SnapDetremura	reference for Snapchat about a celebrated person	<https://twitter.com/detremura>

Table 42: Different IRC snapshots are yield by different channels. In this table we present each snapshot with the respective channel and a reference to the subject.

snapshot channel	observation	reference
#foradoeixo	a Brazilian network of culture related collectives	<https://pt.wikipedia.org/wiki/Fora_do_Eixo>
#hackerspace-cps	a hackerspace in Campinas, Brazil	<https://lhc.net.br/wiki/P%C3%A1gina_principal>
#hackerspaces-br	Brazilian hackerspaces channel	<https://garoa.net.br/wiki/Hackerspaces_Brasileiros>
#labmacambira	Brazilian channel for the labMacambira collective	<http://labmacambira.sourceforge.net/>

Table 43: Different Email snapshots are yield by different email lists. In this table we present each snapshot with the respective list and a reference to the subject.

Gmane ID	observation	reference
gmane.linux.audio.users	the Linux Audio Users	<http://linuxaudio.org>
gmane.politics.organizations.metareciclagem	a network about technology and social transformation	<https://metareciclagem.github.io>
gmane.linux.audio.devel	the Linux Audio Developers	<http://lists.linuxaudio.org/listinfo/linux-audio-dev>
gmane.comp.gcc.libstdc++.devel	the C++ standard library	<https://gcc.gnu.org/libstdc++/>

Table 44: References for the snapshots of the detached instances ParticipaBR, Cidade Democrática and AA.

social protocol	observations	reference
ParticipaBR	a Brazilian federal portal of social participation	<http://www.participa.br/>
Cidade Demorática	a Brazilian civil society portal of social participation	<http://www.cidadedemocratica.org.br/>
AA	the Algorithmic Autoregulation software development methodology	?

A.9 Trivial counts in each snapshot

Table 45: Number of triples (ntriples), number of relations/interactions/edges (nedges), number of participants (nparticipants) and number of characters (nchars) in each snapshot.

snapshot id	ntriples	nedges	nparticipants	nchars
aa-irc-legacy-labmacambira_lalenia.txt	53,140	0	117	558,466
aa-mongo-legacy	22,773	0	37	240,172
aa-mysql-legacy	790,796	0	157	2,753,354
cidadedemocratica-legacy	915,852	0	23,079	6,871,848
facebook-legacy-AdornoNaoEhEnfeite29032013	8,459	1,292	293	26,113
facebook-legacy-AntonioAnzoategui18022013	1,676	328	52	0
facebook-legacy-AtivistasDaInclusaoDigital09032013	25,642	5,592	306	0
facebook-legacy-Auricultura10042013	19,515	3,898	412	14,015
facebook-legacy-BrunoMialich31012013	40,794	9,320	502	0
facebook-legacy-CalebLuporini13042013	105,962	24,653	1,050	0
facebook-legacy-CalebLuporini19022013	104,746	24,391	1,026	0
facebook-legacy-CienciasComFronteiras29032013	110,734	23,302	2,921	0
facebook-legacy-ComputerArt10032013	260,050	62,819	1,342	0
facebook-legacy-Coolmeia06032013	76,063	16,534	1,202	0
facebook-legacy-DanielPenalva18022013	3,519	682	113	0
facebook-legacy-DemocraciaDiretaJa14032013	258,679	59,323	3,053	54,443
facebook-legacy-DemocraciaDiretaJa14072013	257,490	59,781	3,607	58,035
facebook-legacy-DemocraciaPura06042013	32,227	6,730	627	65,062
facebook-legacy-Economia14042013	238,790	54,001	3,587	52,664
facebook-legacy-EconomiaCriativaDigital03032013	185,905	43,128	1,684	0
facebook-legacy-EducacoesEAprendizagensXXI02032013	106,918	24,802	1,285	0
facebook-legacy-GabrielaThume19022013	18,581	4,108	307	0
facebook-legacy-GrahamForrest28012013	1,370	185	90	0
facebook-legacy-LailaManuelle17012013	201,071	48,572	969	0
facebook-legacy-LarissaAnzoategui20022013	24,824	5,191	580	0
facebook-legacy-Latesfip08032014	11,554	2,009	306	0
facebook-legacy-LivingBridgesPlanet29032013	149,708	32,494	3,077	52,808
facebook-legacy-LuisCirne07032013	16,619	3,390	437	0
facebook-legacy-MariliaMelloPisani10042013	84,770	19,040	1,230	0
facebook-legacy-Mirtes16052013	39,415	9,075	445	0

Continued on next page

Table 45 – *Continued from previous page*

snapshot id	ntriples	nedges	nparticipants	nchars
facebook-legacy-MobilizacoesCulturaisInteriorSP13032013	26,518	6,096	298	0
facebook-legacy-PartidoPirata23032013	45,495	8,537	1,943	36,313
facebook-legacy-PedroPauloRocha10032013	215,888	50,591	1,932	0
facebook-legacy-PeterForrest28012013	8,156	1,829	120	0
facebook-legacy-PoliticasCulturasBrasileiras08032013	178,289	41,690	1,278	69,756
facebook-legacy-PracaPopular16032013	4,539	932	65	4,249
facebook-legacy-RafaelReinehr09042013	174,423	39,586	2,297	0
facebook-legacy-RamiroGiroldo20022013	9,928	2,020	264	0
facebook-legacy-RedeTranzmidias02032013	25,111	4,940	391	54,907
facebook-legacy-RenatoFabbri02032013	93,134	21,579	974	0
facebook-legacy-RenatoFabbri03032013	93,690	21,711	978	0
facebook-legacy-RenatoFabbri11072013	114,552	26,440	1,256	0
facebook-legacy-RenatoFabbri18042013	104,156	24,072	1,124	0
facebook-legacy-RenatoFabbri20012013	86,416	20,085	868	0
facebook-legacy-RenatoFabbri29112012	82,093	19,083	823	0
facebook-legacy-RicardoFabbri18022013	11,372	2,327	344	0
facebook-legacy-RitaWu08042013	83,895	18,935	1,165	0
facebook-legacy-RonaldCosta12062013	31,338	6,557	730	0
facebook-legacy-SiliconValleyGlobalNetwork27042013	77,158	15,740	2,130	50,251
facebook-legacy-SolidarityEconomy12042013	14,230	2,404	525	67,774
facebook-legacy-StudyGroupSNA05042013	5,604	480	448	25,474
facebook-legacy-THackDay26032013	1,844	420	41	0
facebook-legacy-Tecnoxamanismo08032014	11,106	2,069	318	0
facebook-legacy-Tecnoxamanismo15032014	14,589	2,702	450	0
facebook-legacy-ThaisTeixeira19022013	26,424	6,088	296	0
facebook-legacy-VilsonVieira18022013	19,688	4,334	336	0
facebook-legacy-ViniciusSampaio18022013	90,515	21,360	725	0
facebook-legacy-avlab_BarthorLaZule22022014	16,005	3,513	279	0
facebook-legacy-avlab_CalebLuporini25022014	125,577	29,268	1,215	0
facebook-legacy-avlab_CamilaBatista23022014	21,138	4,476	462	0
facebook-legacy-avlab_CarlosDiego25022014	171,744	39,401	2,020	0
facebook-legacy-avlab_CristinaMekitarian23022014	24,647	5,572	337	0
facebook-legacy-avlab_DanielGonzales23022014	196,406	45,318	2,162	0
facebook-legacy-avlab_FelipeBrait23022014	1,228,605	299,082	4,611	0
facebook-legacy-avlab_FelipeVillela22022014	2,475	477	81	0
facebook-legacy-avlab_JoaoMeirelles25022014	52,371	11,649	825	0

Continued on next page

Table 45 – *Continued from previous page*

snapshot id	ntriples	nedges	nparticipants	nchars
facebook-legacy-avlab_JoaoMekitarian23022014	88,765	20,821	783	0
facebook-legacy-avlab_JulianaSouza23022014	129,757	29,942	1,427	0
facebook-legacy-avlab_KarinaGomes22022014	9,073	1,906	207	0
facebook-legacy-avlab_LucasOliveira26022014	62,871	14,764	545	0
facebook-legacy-avlab_MarcelaLucatelli25022014	138,733	31,647	1,735	0
facebook-legacy-avlab_MariliaPisani25022014	114,765	25,830	1,635	0
facebook-legacy-avlab_NatachaRena22022014	642,769	154,758	3,391	0
facebook-legacy-avlab_OrlandoCoelho22022014	5,149	848	251	0
facebook-legacy-avlab_PalomaKliss25022014	493,774	119,520	2,242	0
facebook-legacy-avlab_PedroRocha25022014	346,883	81,910	2,749	0
facebook-legacy-avlab_RenatoFabbri22022014	124,703	28,780	1,369	0
facebook-legacy-avlab_SarahLuporini25022014	505,853	121,502	2,835	0
facebook-legacy-avlab_SatoBrasil25022014	1,519,394	371,249	4,914	0
facebook-legacy-ego_MarceloSaldanha19112014	130,556	29,440	1,828	0
facebook-legacy-ego_MariliaPisani06052014	122,231	27,581	1,701	0
facebook-legacy-ego_MassimoCanevacci19062013	273,328	59,995	4,764	0
facebook-legacy-ego_RenatoFabbri06022014	123,993	28,606	1,367	0
facebook-legacy-ego_RenatoFabbri19112014	153,410	35,514	1,622	0
facebook-legacy-ego_VJPixel23052014	231,608	54,752	1,800	0
facebook-legacy-posavlab_AnaCelia18032014	53,939	12,167	753	0
facebook-legacy-posavlab_ElenaGarnelo04032014	93,472	21,723	940	0
facebook-legacy-posavlab_FabiBorges08032014	159,584	36,592	1,888	0
facebook-legacy-posavlab_GeorgeSanders08032014	108,071	24,706	1,321	0
facebook-legacy-posavlab_GrazielleMachado18032014	24,264	5,254	464	0
facebook-legacy-posavlab_RenatoFabbri19032014	129,360	29,890	1,400	0
facebook-legacy-posavlab_RicardoPoppi18032014	76,104	17,234	1,024	0
gmane-legacy-comp.gcc.libstdcpp.devel1-20000	324,051	14,786	1,036	30,126,252
gmane-legacy-linux.audio.devel1-20000	377,307	17,076	1,232	26,969,596
gmane-legacy-linux.audio.users1-20000	349,304	16,362	1,147	25,065,928
gmane-legacy-politics.organizations.metareciclagem1-20000	338,679	15,230	477	54,260,954
irc-legacy-foradoeixo	685,623	4,308	3,318	3,777,424
irc-legacy-hackerspace-cps	253,614	1,860	607	1,059,675
irc-legacy-hackerspaces-br	980,556	210	347	8,420,840
irc-legacy-labmacambira	1,535,463	58,525	1,561	8,358,970
participabr-legacy	159,602	2,207	3,825	2,045,617
twitter-legacy-ChennaiFloods	6,793,705	101,824	46,493	23,237,802

Continued on next page

Table 45 – *Continued from previous page*

snapshot id	ntriples	nedges	nparticipants	nchars
twitter-legacy-ForaCunha	88,963	1,656	2,747	372,131
twitter-legacy-ForaDilma	26,818	668	659	113,810
twitter-legacy-MAMA2015	20,356,960	426,558	33,080	75,358,785
twitter-legacy-QuartaSemRacismoClubeSDV	367,460	5,000	5,785	1,635,867
twitter-legacy-SnapDetremura	26,834	461	621	124,448
twitter-legacy-arenaNETmundial	388,134	15,797	5,898	2,825,121
twitter-legacy-art	2,814,803	32,655	30,486	9,539,413
twitter-legacy-dilma	78,424	1,692	2,274	332,005
twitter-legacy-fuck	3,631	40	93	14,727
twitter-legacy-game	229,992	1,548	4,682	1,229,910
twitter-legacy-god	1,514,365	20,132	22,117	5,560,140
twitter-legacy-music	8,150,863	39,456	54,006	21,116,573
twitter-legacy-obama	1,143,873	20,623	20,330	4,481,840
twitter-legacy-python	5,786	4	17	1,758
twitter-legacy-science	369,013	3,673	7,156	1,910,216

B HUMAN INTERACTION NETWORKS STABILITY SUPPORTING INFORMATION

This Appendix holds circular statistics and histograms of activity along time in Section B.1, the fraction of vertices in the peripheral, intermediary and hub sectors in Section B.3 and the combination of basic topological measures into principal components with greater variance in Section B.2. There is a focus on email list interaction networks for benchmarking and Section B.4 reinforces the results with the analysis of networks from Facebook, Twitter and Participabr. More context (e.g. methods, discussion, data and scripts) is given in the main document[?] to which this current document supplies supporting information.

B.1 Temporal activity in different scales

This section presents information derived from the theory presented in Section 2.2.1 for supporting the results in Section 3.0.1.

B.1.1 Temporal circular measures

The metrics with which we report measurements and results of activity along time are the rescaled circular mean θ'_μ, the standard deviation $S(z)$, the variance $Var(z)$, the circular dispersion $\delta(z)$ and the ratio between the lowest b_l and the highest b_h incidences $\frac{b_l}{b_h}$ at each time scale. Also, the mean $\mu(\frac{b'_l}{b'_h})$ and the standard deviation $\sigma(\frac{b'_l}{b'_h})$ of the relation between the minimum b'_l and the maximum b'_h incidences are given for 1000 uniform distribution simulations within the same number of bins and with the same number of samples [*]. Greater dispersion is found on the scales of seconds and minutes, followed by days of the month. Greater localization is found in the scale of hours of the day, followed by weekdays and months.

Table 46: LAU circular measurements.

scale	θ'_μ	$S(z)$	$Var(z)$	$\delta(z)$	$\frac{b_l}{b_h}$	$\mu(\frac{b'_l}{b'_h})$	$\sigma(\frac{b'_l}{b'_h})$
seconds	–//–	3.31	1.00	29337.65	0.78	0.78	0.03
minutes	–//–	3.13	0.99	8879.19	0.76	0.78	0.03
hours	-8.76	1.56	0.71	4.92	0.12	0.87	0.02
weekdays	-0.21	2.14	0.90	45.41	0.62	0.95	0.01
month days	-0.64	2.76	0.98	1001.75	0.67	0.85	0.02
months	3.55	2.30	0.93	94.53	0.64	0.92	0.02

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

Table 47: LAD circular measurements.

scale	θ'_μ	$S(z)$	$Var(z)$	$\delta(z)$	$\frac{b_l}{b_h}$	$\mu(\frac{b'_l}{b'_h})$	$\sigma(\frac{b'_l}{b'_h})$
seconds	–//–	3.13	0.99	9070.17	0.78	0.78	0.03
minutes	–//–	3.60	1.00	205489.40	0.82	0.78	0.03
hours	-9.61	1.52	0.68	4.36	0.10	0.87	0.02
weekdays	-0.03	2.03	0.87	29.28	0.58	0.95	0.01
month days	-2.65	2.93	0.99	2657.77	0.67	0.85	0.02
months	-0.56	2.14	0.90	44.00	0.44	0.92	0.02

Table 48: MET circular measurements.

scale	θ'_μ	$S(z)$	$Var(z)$	$\delta(z)$	$\frac{b_l}{b_h}$	$\mu(\frac{b'_l}{b'_h})$	$\sigma(\frac{b'_l}{b'_h})$
seconds	–//–	3.06	0.99	5910.47	0.79	0.77	0.03
minutes	–//–	3.14	0.99	9696.29	0.75	0.77	0.03
hours	-9.20	1.35	0.60	2.76	0.05	0.87	0.02
weekdays	-0.27	1.86	0.82	13.82	0.35	0.95	0.02
month days	3.58	2.49	0.95	237.30	0.64	0.85	0.02
months	-2.92	1.73	0.78	9.20	0.33	0.92	0.02

Table 49: CPP circular measurements.

scale	θ'_μ	$S(z)$	$Var(z)$	$\delta(z)$	$\frac{b_l}{b_h}$	$\mu(\frac{b'_l}{b'_h})$	$\sigma(\frac{b'_l}{b'_h})$
seconds	–//–	3.31	1.00	28205.46	0.79	0.78	0.03
minutes	–//–	3.18	0.99	12275.59	0.79	0.78	0.03
hours	-9.39	1.48	0.67	3.91	0.09	0.87	0.02
weekdays	-0.17	1.83	0.81	12.66	0.39	0.95	0.01
month days	-10.12	3.16	0.99	10789.17	0.65	0.85	0.02
months	0.15	2.34	0.93	115.49	0.67	0.92	0.02

B.1.2 Temporal histograms

B.1.2.1 Histograms of activity along the hours of the day

Higher activity was observed between noon and 6pm, followed by the time period between 6pm and midnight. Around 2/3 of the whole activity takes place from noon to midnight. The activity peak occurs around midday, with a slight skew toward one hour before noon.

Table 50: LAU activity along the hours of the day.

<table>
<tr><th></th><th>1h</th><th>2h</th><th>3h</th><th>4h</th><th>6h</th><th>12h</th></tr>
<tr><td>0h</td><td>3.58</td><td rowspan="2">5.80</td><td rowspan="3">7.43</td><td rowspan="4">8.49</td><td rowspan="6">10.14</td><td rowspan="12">36.88</td></tr>
<tr><td>1h</td><td>2.22</td></tr>
<tr><td>2h</td><td>1.63</td><td rowspan="2">2.69</td></tr>
<tr><td>3h</td><td>1.06</td><td rowspan="4">2.72</td></tr>
<tr><td>4h</td><td>0.84</td><td rowspan="2">1.66</td><td rowspan="4">5.20</td></tr>
<tr><td>5h</td><td>0.82</td></tr>
<tr><td>6h</td><td>1.17</td><td rowspan="2">3.54</td><td rowspan="4">7.07</td><td rowspan="6">26.74</td></tr>
<tr><td>7h</td><td>2.37</td></tr>
<tr><td>8h</td><td>3.53</td><td rowspan="2">9.57</td><td rowspan="5">23.20</td></tr>
<tr><td>9h</td><td>6.04</td><td rowspan="4">19.67</td></tr>
<tr><td>10h</td><td>6.83</td><td rowspan="2">13.62</td></tr>
<tr><td>11h</td><td>6.79</td></tr>
<tr><td>12h</td><td>6.11</td><td rowspan="2">12.36</td><td rowspan="3">18.75</td><td rowspan="4">24.68</td><td rowspan="6">35.66</td><td rowspan="12">63.12</td></tr>
<tr><td>13h</td><td>6.26</td></tr>
<tr><td>14h</td><td>6.38</td><td rowspan="2">12.31</td></tr>
<tr><td>15h</td><td>5.93</td><td rowspan="4">16.91</td></tr>
<tr><td>16h</td><td>5.52</td><td rowspan="2">10.98</td><td rowspan="4">20.73</td></tr>
<tr><td>17h</td><td>5.46</td></tr>
<tr><td>18h</td><td>5.23</td><td rowspan="2">9.75</td><td rowspan="4">14.30</td><td rowspan="6">27.46</td></tr>
<tr><td>19h</td><td>4.52</td></tr>
<tr><td>20h</td><td>4.55</td><td rowspan="2">8.97</td><td rowspan="5">17.71</td></tr>
<tr><td>21h</td><td>4.42</td><td rowspan="4">13.16</td></tr>
<tr><td>22h</td><td>4.51</td><td rowspan="2">8.74</td></tr>
<tr><td>23h</td><td>4.23</td></tr>
</table>

Table 51: LAD activity along the hours of the day.

<table>
<tr><th></th><th>1h</th><th>2h</th><th>3h</th><th>4h</th><th>6h</th><th>12h</th></tr>
<tr><td>0h</td><td>4.01</td><td rowspan="2">6.53</td><td rowspan="2">8.32</td><td rowspan="3">9.37</td><td rowspan="4">10.78</td><td rowspan="12">33.11</td></tr>
<tr><td>1h</td><td>2.52</td></tr>
<tr><td>2h</td><td>1.79</td><td rowspan="2">2.84</td></tr>
<tr><td>3h</td><td>1.06</td><td rowspan="3">2.46</td><td rowspan="4">3.81</td></tr>
<tr><td>4h</td><td>0.75</td><td rowspan="2">1.40</td></tr>
<tr><td>5h</td><td>0.66</td><td rowspan="5">22.33</td></tr>
<tr><td>6h</td><td>0.85</td><td rowspan="2">2.41</td><td rowspan="3">5.36</td></tr>
<tr><td>7h</td><td>1.56</td><td rowspan="4">19.93</td></tr>
<tr><td>8h</td><td>2.95</td><td rowspan="2">7.61</td></tr>
<tr><td>9h</td><td>4.66</td><td rowspan="3">16.98</td></tr>
<tr><td>10h</td><td>5.92</td><td rowspan="2">12.32</td></tr>
<tr><td>11h</td><td>6.40</td><td rowspan="13">66.89</td></tr>
<tr><td>12h</td><td>6.41</td><td rowspan="2">12.53</td><td rowspan="3">18.85</td><td rowspan="3">24.82</td><td rowspan="5">37.24</td></tr>
<tr><td>13h</td><td>6.12</td></tr>
<tr><td>14h</td><td>6.32</td><td rowspan="2">12.29</td></tr>
<tr><td>15h</td><td>5.97</td><td rowspan="3">18.39</td><td rowspan="4">23.44</td></tr>
<tr><td>16h</td><td>6.40</td><td rowspan="2">12.42</td></tr>
<tr><td>17h</td><td>6.02</td><td rowspan="6">29.65</td></tr>
<tr><td>18h</td><td>5.99</td><td rowspan="2">11.02</td><td rowspan="3">15.65</td></tr>
<tr><td>19h</td><td>5.03</td></tr>
<tr><td>20h</td><td>4.63</td><td rowspan="2">9.22</td><td rowspan="4">18.63</td></tr>
<tr><td>21h</td><td>4.59</td><td rowspan="3">14.00</td></tr>
<tr><td>22h</td><td>4.88</td><td rowspan="2">9.41</td></tr>
<tr><td>23h</td><td>4.53</td></tr>
</table>

Table 52: MET activity along the hours of the day.

<table>
<tr><th></th><th>1h</th><th>2h</th><th>3h</th><th>4h</th><th>6h</th><th>12h</th></tr>
<tr><td>0h</td><td>2.87</td><td rowspan="2">4.64</td><td rowspan="3">5.67</td><td rowspan="4">6.31</td><td rowspan="5">7.15</td><td rowspan="12">29.33</td></tr>
<tr><td>1h</td><td>1.77</td></tr>
<tr><td>2h</td><td>1.04</td><td rowspan="2">1.67</td></tr>
<tr><td>3h</td><td>0.64</td><td rowspan="4">1.48</td></tr>
<tr><td>4h</td><td>0.47</td><td rowspan="2">0.85</td><td rowspan="4">2.89</td></tr>
<tr><td>5h</td><td>0.38</td></tr>
<tr><td>6h</td><td>0.72</td><td rowspan="2">2.04</td><td rowspan="3">4.71</td><td rowspan="7">22.18</td></tr>
<tr><td>7h</td><td>1.33</td></tr>
<tr><td>8h</td><td>2.67</td><td rowspan="2">7.07</td><td rowspan="5">20.14</td></tr>
<tr><td>9h</td><td>4.40</td><td rowspan="3">17.47</td></tr>
<tr><td>10h</td><td>6.29</td><td rowspan="2">13.07</td></tr>
<tr><td>11h</td><td>6.78</td><td rowspan="13">70.67</td></tr>
<tr><td>12h</td><td>7.33</td><td rowspan="2">14.41</td><td rowspan="3">21.50</td><td rowspan="4">28.65</td><td rowspan="5">42.22</td></tr>
<tr><td>13h</td><td>7.08</td></tr>
<tr><td>14h</td><td>7.09</td><td rowspan="2">14.24</td></tr>
<tr><td>15h</td><td>7.14</td><td rowspan="4">20.72</td></tr>
<tr><td>16h</td><td>6.68</td><td rowspan="2">13.58</td><td rowspan="5">24.79</td></tr>
<tr><td>17h</td><td>6.89</td></tr>
<tr><td>18h</td><td>5.99</td><td rowspan="2">11.22</td><td rowspan="4">16.19</td><td rowspan="7">28.44</td></tr>
<tr><td>19h</td><td>5.23</td></tr>
<tr><td>20h</td><td>4.98</td><td rowspan="2">9.34</td><td rowspan="5">17.22</td></tr>
<tr><td>21h</td><td>4.37</td></tr>
<tr><td>22h</td><td>4.24</td><td rowspan="2">7.88</td><td rowspan="4">12.25</td></tr>
<tr><td>23h</td><td>3.64</td></tr>
</table>

Table 53: CPP activity along the hours of the day.

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

B.1.2.2 Histograms of activity along weekdays.

Most notably, the pattern of activity along weekdays presents a decrease of activity on weekend days of at least one third and at most two thirds compared against workweek days.

	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

B.1.2.3 Histograms of activity along the days of the month

The most important feature seems to be the homogeneity made explicit by the high circular dispersion in the tables of Section B.1.1. Slightly higher activity rates are found in the beginning of the month, although not statistically significant.

Table 54: LAU activity along the days of the month.

	1 day	5	10	15 days
1	3.36			
2	3.43			
3	3.31	16.21		
4	3.37			
5	2.75		33.71	
6	3.03			
7	3.93			
8	3.62	17.50		50.82
9	3.84			
10	3.09			
11	3.20			
12	3.40			
13	3.67	17.11		
14	3.71			
15	3.14		34.02	
16	3.08			49.18
17	3.13			
18	3.43	16.91		
19	3.61			
20	3.67			
21	3.60			
22	3.42			
23	2.80	15.43		
24	2.64			
25	2.97		32.27	
26	3.06			
27	2.69			
28	3.79	16.85		
29	3.75			
30	3.57			

Table 55: LAD activity along the days of the month.

	1 day	5	10	15 days
1	3.29			
2	3.38			
3	2.85	15.77		
4	2.94			
5	3.31		33.63	
6	3.60			
7	2.68			
8	3.78	17.85		50.50
9	3.88			
10	3.91			
11	3.22			
12	2.79			
13	3.50	16.87		
14	3.95			
15	3.40		33.41	
16	3.32			
17	2.95			
18	3.50	16.54		
19	3.69			
20	3.07			
21	2.76			
22	3.35			
23	3.32	15.71		49.50
24	3.15			
25	3.13		32.96	
26	3.68			
27	4.02			
28	3.49	17.25		
29	3.34			
30	2.72			

Table 56: MET activity along the days of the month.

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

Table 57: CPP activity along the days of the month.

	1 day	5	10	15 days
1	3.22			
2	3.08			
3	3.19	15.98		
4	3.65			
5	2.84		31.82	
6	3.65			
7	3.53			
8	3.10	15.84		49.62
9	2.49			
10	3.07			
11	3.47			
12	3.26			
13	3.55	17.80		
14	3.84			
15	3.68		34.22	
16	3.74			50.38
17	3.40			
18	3.41	16.42		
19	2.95			
20	2.93			
21	3.15			
22	3.64			
23	3.51	17.13		
24	3.32			
25	3.51		33.96	
26	3.54			
27	3.21			
28	3.40	16.84		
29	3.83			
30	2.86			

B.1.2.4 Histograms of activity along months of the year

Activity is concentrated in Jun-Aug and/or in Dec-Mar. These observations mostly fit academic calendars, vacations and end-of-year holidays.

Table 58: LAU activity along the months of the year.

<table>
<tr><th></th><th>m.</th><th>b.</th><th>t.</th><th>q.</th><th>s.</th></tr>
<tr><td>Jan</td><td>10.22</td><td rowspan="2">19.56</td><td rowspan="3">28.24</td><td rowspan="4">35.09</td><td rowspan="6">49.16</td></tr>
<tr><td>Fev</td><td>9.34</td></tr>
<tr><td>Mar</td><td>8.67</td><td rowspan="2">15.53</td></tr>
<tr><td>Apr</td><td>6.86</td><td rowspan="3">20.93</td></tr>
<tr><td>Mai</td><td>7.28</td><td rowspan="2">14.07</td><td rowspan="4">30.36</td></tr>
<tr><td>Jun</td><td>6.80</td></tr>
<tr><td>Jul</td><td>8.97</td><td rowspan="2">16.29</td><td rowspan="3">24.47</td><td rowspan="6">50.84</td></tr>
<tr><td>Ago</td><td>7.32</td></tr>
<tr><td>Set</td><td>8.18</td><td rowspan="2">16.25</td><td rowspan="4">34.55</td></tr>
<tr><td>Out</td><td>8.06</td><td rowspan="3">26.36</td></tr>
<tr><td>Nov</td><td>7.64</td><td rowspan="2">18.30</td></tr>
<tr><td>Dez</td><td>10.66</td></tr>
</table>

Table 59: LAD activity along the months of the year.

<table>
<tr><th></th><th>m.</th><th>b.</th><th>t.</th><th>q.</th><th>s.</th></tr>
<tr><td>Jan</td><td>11.24</td><td rowspan="2">18.51</td><td rowspan="3">26.46</td><td rowspan="4">36.07</td><td rowspan="6">57.96</td></tr>
<tr><td>Fev</td><td>7.26</td></tr>
<tr><td>Mar</td><td>7.95</td><td rowspan="2">17.56</td></tr>
<tr><td>Apr</td><td>9.61</td><td rowspan="3">31.50</td></tr>
<tr><td>Mai</td><td>8.94</td><td rowspan="2">21.89</td><td rowspan="4">37.56</td></tr>
<tr><td>Jun</td><td>12.95</td></tr>
<tr><td>Jul</td><td>9.03</td><td rowspan="2">15.67</td><td rowspan="3">22.30</td><td rowspan="6">42.04</td></tr>
<tr><td>Ago</td><td>6.64</td></tr>
<tr><td>Set</td><td>6.63</td><td rowspan="2">12.38</td><td rowspan="4">26.37</td></tr>
<tr><td>Out</td><td>5.75</td><td rowspan="3">19.74</td></tr>
<tr><td>Nov</td><td>7.61</td><td rowspan="2">13.99</td></tr>
<tr><td>Dez</td><td>6.38</td></tr>
</table>

Table 60: MET activity along the months of the year.

<table>
<tr><th></th><th>m.</th><th>b.</th><th>t.</th><th>q.</th><th>s.</th></tr>
<tr><td>Jan</td><td>4.87</td><td rowspan="2">11.00</td><td rowspan="3">16.89</td><td rowspan="4">23.30</td><td rowspan="6">47.70</td></tr>
<tr><td>Fev</td><td>6.13</td></tr>
<tr><td>Mar</td><td>5.89</td><td rowspan="2">12.30</td></tr>
<tr><td>Apr</td><td>6.41</td><td rowspan="3">30.81</td></tr>
<tr><td>Mai</td><td>10.45</td><td rowspan="2">24.40</td><td rowspan="4">47.87</td></tr>
<tr><td>Jun</td><td>13.95</td></tr>
<tr><td>Jul</td><td>13.24</td><td rowspan="2">23.47</td><td rowspan="3">31.21</td><td rowspan="6">52.30</td></tr>
<tr><td>Ago</td><td>10.22</td></tr>
<tr><td>Set</td><td>7.75</td><td rowspan="2">16.79</td><td rowspan="4">28.83</td></tr>
<tr><td>Out</td><td>9.04</td><td rowspan="3">21.09</td></tr>
<tr><td>Nov</td><td>7.45</td><td rowspan="2">12.05</td></tr>
<tr><td>Dez</td><td>4.59</td></tr>
</table>

Table 61: CPP activity along the months of the year.

<table>
<tr><th></th><th>m.</th><th>b.</th><th>t.</th><th>q.</th><th>s.</th></tr>
<tr><td>Jan</td><td>8.70</td><td rowspan="2">17.00</td><td rowspan="3">27.23</td><td rowspan="4">36.49</td><td rowspan="6">54.27</td></tr>
<tr><td>Fev</td><td>8.29</td></tr>
<tr><td>Mar</td><td>10.23</td><td rowspan="2">19.49</td></tr>
<tr><td>Apr</td><td>9.26</td><td rowspan="3">27.03</td></tr>
<tr><td>Mai</td><td>9.41</td><td rowspan="2">17.78</td><td rowspan="4">33.46</td></tr>
<tr><td>Jun</td><td>8.37</td></tr>
<tr><td>Jul</td><td>8.70</td><td rowspan="2">15.68</td><td rowspan="3">22.94</td><td rowspan="6">45.73</td></tr>
<tr><td>Ago</td><td>6.98</td></tr>
<tr><td>Set</td><td>7.26</td><td rowspan="2">15.36</td><td rowspan="4">30.06</td></tr>
<tr><td>Out</td><td>8.10</td><td rowspan="3">22.80</td></tr>
<tr><td>Nov</td><td>7.89</td><td rowspan="2">14.69</td></tr>
<tr><td>Dez</td><td>6.81</td></tr>
</table>

B.2 PCA of measures along the timeline

Loadings for topological metrics in principal components for LAD, LAU, MET, CPP, lists with the window size of $ws = 1000$ messages 20 disjoint positioning. Further details are given in Sections 2.2.3 and 3.0.3 of the main document.[?]

B.2.1 Betweenness, clustering and degree

This simplest PCA is characterized by the coupling of the centrality measures of degree k and betweenness bt in the first component. Second component is mostly the clustering coefficient cc. These three measures contribute almost equally to the dispersion of the system: first component holds two thirds of the dispersion and is resultant of two measures while the second component holds one third of the dispersion and results from one measure.

Table 62: LAU principal components formation and concentration of dispersion.

	PC1		PC2		PC3	
	μ	σ	μ	σ	μ	σ
cc	6.03	3.73	87.60	5.25	4.52	0.93
k	47.13	1.76	3.01	1.98	47.90	0.38
bt	46.84	1.97	9.39	4.31	47.58	0.57
λ	64.99	0.60	33.08	0.41	1.93	0.36

Table 63: LAD principal components formation and concentration of dispersion.

	PC1		PC2		PC3	
	μ	σ	μ	σ	μ	σ
cc	6.42	4.05	86.60	5.50	5.19	1.45
k	46.98	1.86	2.95	1.65	47.61	0.57
bt	46.59	2.18	10.45	4.72	47.20	0.90
λ	64.96	0.71	33.08	0.41	1.96	0.52

Table 64: MET principal components formation and concentration of dispersion.

	PC1		PC2		PC3	
	μ	σ	μ	σ	μ	σ
cc	5.82	3.76	87.26	5.12	4.93	1.19
k	47.18	1.82	4.35	4.01	47.63	0.57
bt	47.01	1.96	8.40	4.22	47.44	0.67
λ	64.94	0.76	33.13	0.45	1.93	0.62

Table 65: CPP principal components formation and concentration of dispersion.

	PC1		PC2		PC3	
	μ	σ	μ	σ	μ	σ
cc	3.61	2.13	91.86	3.24	3.59	0.98
k	48.24	0.99	2.96	2.25	48.25	0.43
bt	48.15	1.14	5.18	3.89	48.16	0.56
λ	65.24	0.51	33.30	0.17	1.46	0.49

B.2.2 Betweenness, clustering, degrees and strengths

In this extension of the previous plot, the set of centrality metrics are extended to include strength s and the in- and out- degrees (k^{in}, k^{out}) and strengths (s^{in}, s^{out}). First component holds an average of the centrality metrics (k, k^{in}, k^{out}, s, s^{in}, s^{out}) while second component is mostly clustering coefficient *cc*. All metrics contribute the about same for total dispersion ($\approx \frac{94}{8} = 11.75\%$).

Table 66: LAU principal components formation and concentration of dispersion.

	PC1		PC2		PC3	
	μ	σ	μ	σ	μ	σ
cc	1.59	0.81	80.37	5.18	3.09	1.89
s	14.40	0.15	0.81	0.68	4.75	4.43
s^{in}	14.00	0.14	2.32	1.49	18.98	4.93
s^{out}	13.96	0.14	2.72	1.44	18.25	6.36
k	14.49	0.15	0.54	0.35	1.37	0.98
k^{in}	14.01	0.13	2.72	1.35	18.69	5.01
k^{out}	13.85	0.13	2.37	1.73	22.63	3.79
bt	13.69	0.22	8.16	1.62	12.23	8.33
λ	81.87	0.88	12.48	0.15	3.33	0.70

Table 67: LAD principal components formation and concentration of dispersion.

	PC1		PC2		PC3	
	μ	σ	μ	σ	μ	σ
cc	1.83	1.11	80.38	11.45	3.78	4.43
s	14.25	0.17	1.34	1.81	9.88	5.76
s^{in}	13.99	0.19	2.06	1.70	17.62	6.15
s^{out}	14.03	0.22	1.81	1.98	15.44	6.68
k	14.38	0.13	0.95	1.64	3.45	3.15
k^{in}	14.05	0.14	2.26	1.66	13.44	7.26
k^{out}	13.96	0.15	1.72	1.53	16.14	6.37
bt	13.51	0.35	9.48	2.86	20.26	9.87
λ	82.32	1.61	12.52	0.26	2.97	1.21

Table 68: MET principal components formation and concentration of dispersion.

	PC1		PC2		PC3	
	μ	σ	μ	σ	μ	σ
cc	1.16	0.76	81.72	3.00	1.61	1.78
s	14.32	0.16	1.76	1.12	11.39	5.50
s^{in}	14.17	0.11	2.29	1.29	14.46	3.72
s^{out}	14.09	0.17	1.72	1.18	17.54	5.37
k	14.39	0.16	1.73	0.63	4.76	2.82
k^{in}	14.12	0.13	1.02	0.71	11.69	6.93
k^{out}	14.06	0.13	3.11	1.58	12.18	9.24
bt	13.69	0.26	6.64	2.01	26.37	12.37
λ	83.41	1.53	12.53	0.11	2.34	1.16

Table 69: CPP principal components formation and concentration of dispersion.

	PC1		PC2		PC3	
	μ	σ	μ	σ	μ	σ
cc	0.84	0.61	80.59	6.89	2.30	2.19
s	14.28	0.07	0.97	1.03	15.89	1.15
s^{in}	14.18	0.12	2.89	1.71	13.50	5.19
s^{out}	14.07	0.23	2.83	1.63	18.80	4.94
k	14.42	0.08	0.78	0.67	7.48	2.71
k^{in}	14.29	0.10	2.36	1.41	7.21	4.49
k^{out}	14.16	0.12	3.62	1.83	8.79	4.58
bt	13.76	0.22	5.96	1.88	26.03	7.94
λ	83.32	1.42	12.60	0.08	2.61	1.15

B.2.3 Betweenness, clustering, degrees, strengths and symmetry measures

Loadings for 14 topological metrics in the first three principal components are given for LAD, LAU, MET, CPP, list. The clustering coefficient *cc* appears as the first metric in the tables, followed by 7 centrality metrics (k, k^{in}, k^{out}, s, s^{in}, s^{ou}, bt) and 6 symmetry-related metrics (asy, μ^{asy}, σ^{asy}, dis, μ^{dis} and σ^{dis}). The centrality metrics are most important for the first principal component, while the second component is predominantly the result of symmetry metrics. The clustering coefficient is only relevant for the third principal component, coupled with standard deviations of asymmetry σ^{asy} and disequilibrium σ^{dis}. The three components sum up to $\approx 90\%$ of the variance. In the PCA measures of the CPP list, the last table of these PCA-related tables, this general pattern is depicted in boldface. Further details are given in Sections 2.2.3 and 3.0.3 of the main document.[?]

Table 70: LAU principal components formation and concentration of dispersion.

	PC1		PC2		PC3	
	μ	σ	μ	σ	μ	σ
cc	1.64	0.77	2.42	1.71	19.20	3.96
s	12.80	0.46	0.89	0.82	2.53	0.63
s^{in}	12.47	0.42	2.30	0.97	2.29	0.81
s^{out}	12.37	0.46	2.89	1.24	2.64	0.58
k	12.93	0.44	0.82	0.73	1.32	0.45
k^{in}	12.54	0.37	2.88	1.13	1.02	0.56
k^{out}	12.32	0.46	3.82	1.14	1.57	0.68
bt	12.19	0.46	1.06	0.62	2.64	0.89
asy	0.93	0.81	20.38	0.82	1.66	1.09
μ^{asy}	0.96	0.83	20.26	0.82	1.66	1.04
σ^{asy}	6.18	0.71	1.24	0.92	27.98	1.74
dis	0.90	0.79	20.36	0.82	1.54	1.07
μ^{dis}	0.92	0.61	19.02	0.84	1.45	1.12
σ^{dis}	0.86	0.51	1.64	1.10	32.51	1.90
λ	48.41	0.52	27.95	0.36	12.81	0.79

Table 71: LAD principal components formation and concentration of dispersion.

	PC1		PC2		PC3	
	μ	σ	μ	σ	μ	σ
cc	1.96	0.95	3.07	1.46	17.94	5.38
s	12.34	0.57	1.72	0.99	2.43	0.93
s^{in}	12.06	0.64	3.18	0.98	1.98	1.09
s^{out}	12.22	0.48	1.14	0.78	2.83	0.79
k	12.54	0.56	1.43	0.87	0.92	0.44
k^{in}	12.15	0.61	3.81	0.79	0.61	0.42
k^{out}	12.27	0.45	1.51	1.08	1.56	0.39
bt	11.73	0.64	1.80	0.88	2.28	1.00
asy	1.51	0.97	19.66	1.63	3.02	1.66
μ^{asy}	1.41	0.99	19.53	1.62	3.00	1.69
σ^{asy}	5.62	0.68	2.01	1.23	27.46	3.31
dis	1.58	0.98	19.57	1.65	3.21	1.71
μ^{dis}	1.84	1.00	18.62	1.52	2.08	1.13
σ^{dis}	0.77	0.59	2.94	1.60	30.68	3.34
λ	48.65	1.03	27.84	0.31	13.00	0.77

Table 72: MET principal components formation and concentration of dispersion.

	PC1		PC2		PC3	
	μ	σ	μ	σ	μ	σ
cc	1.18	0.71	3.00	2.35	22.39	2.71
s	12.34	0.66	1.74	1.17	1.55	0.75
s^{in}	12.25	0.62	1.74	0.96	1.45	0.77
s^{out}	12.11	0.72	2.42	1.35	1.78	0.78
k	12.48	0.63	1.46	0.91	0.54	0.48
k^{in}	12.32	0.56	1.54	1.22	0.65	0.62
k^{out}	12.12	0.67	3.10	1.15	0.87	0.74
bt	11.85	0.62	1.46	0.87	1.16	0.70
asy	1.79	1.22	19.35	2.15	3.29	2.15
μ^{asy}	1.84	1.22	19.17	2.16	3.31	2.23
σ^{asy}	4.17	0.79	3.91	2.35	27.79	3.96
dis	1.78	1.18	19.26	2.15	3.38	2.29
μ^{dis}	1.53	1.10	18.23	2.12	3.32	1.71
σ^{dis}	2.23	0.93	3.61	2.38	28.54	3.23
λ	49.05	1.01	27.79	0.30	13.30	1.35

Table 73: CPP principal components formation and concentration of dispersion.

	PC1		PC2		PC3	
	μ	σ	μ	σ	μ	σ
cc	0.89	0.59	1.93	1.33	21.22	2.97
s	11.71	0.57	2.97	0.82	2.45	0.72
s^{in}	11.68	0.58	2.37	0.91	3.08	0.78
s^{out}	11.49	0.61	3.63	0.79	1.61	0.88
k	11.93	0.54	2.58	0.70	0.52	0.44
k^{in}	11.93	0.52	1.19	0.88	1.41	0.71
k^{out}	11.57	0.61	4.34	0.70	0.98	0.66
bt	11.37	0.55	2.44	0.84	1.37	0.77
asy	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
dis	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

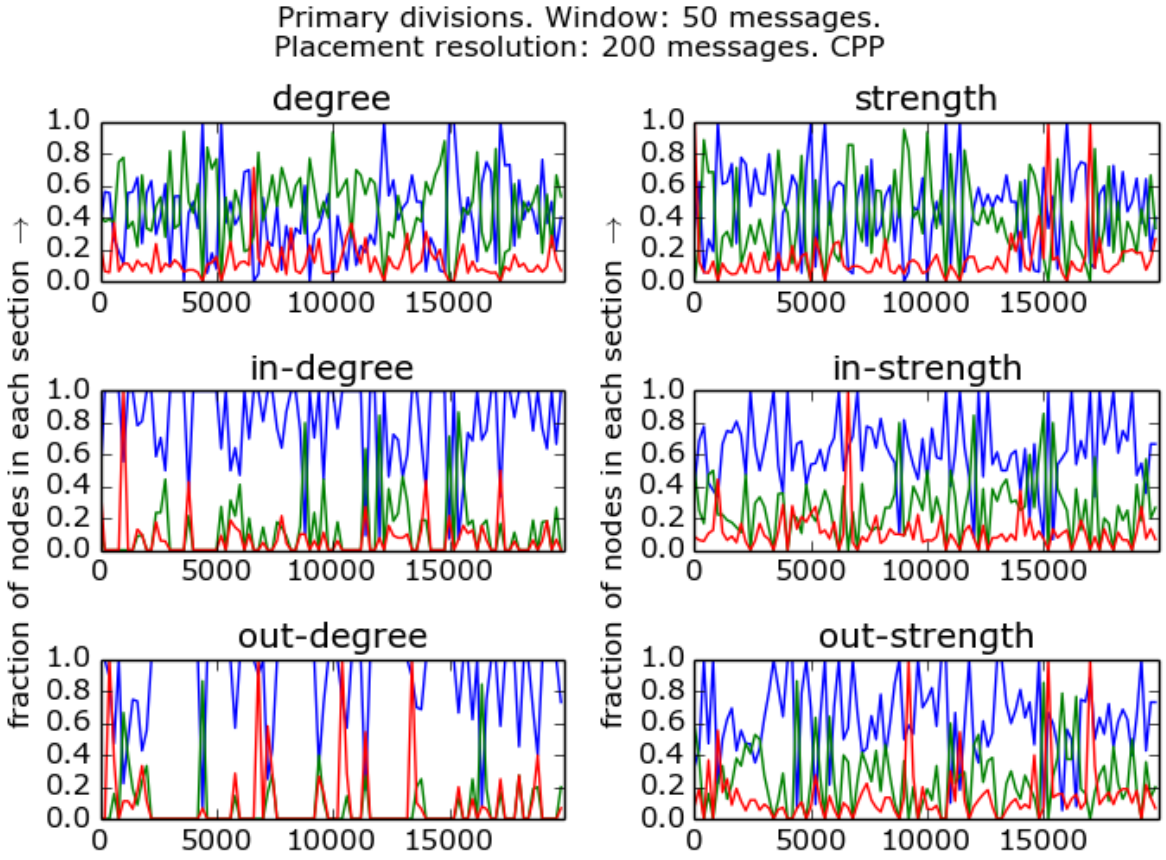
B.3 Fraction of participants in each Erdös Sector along the timeline

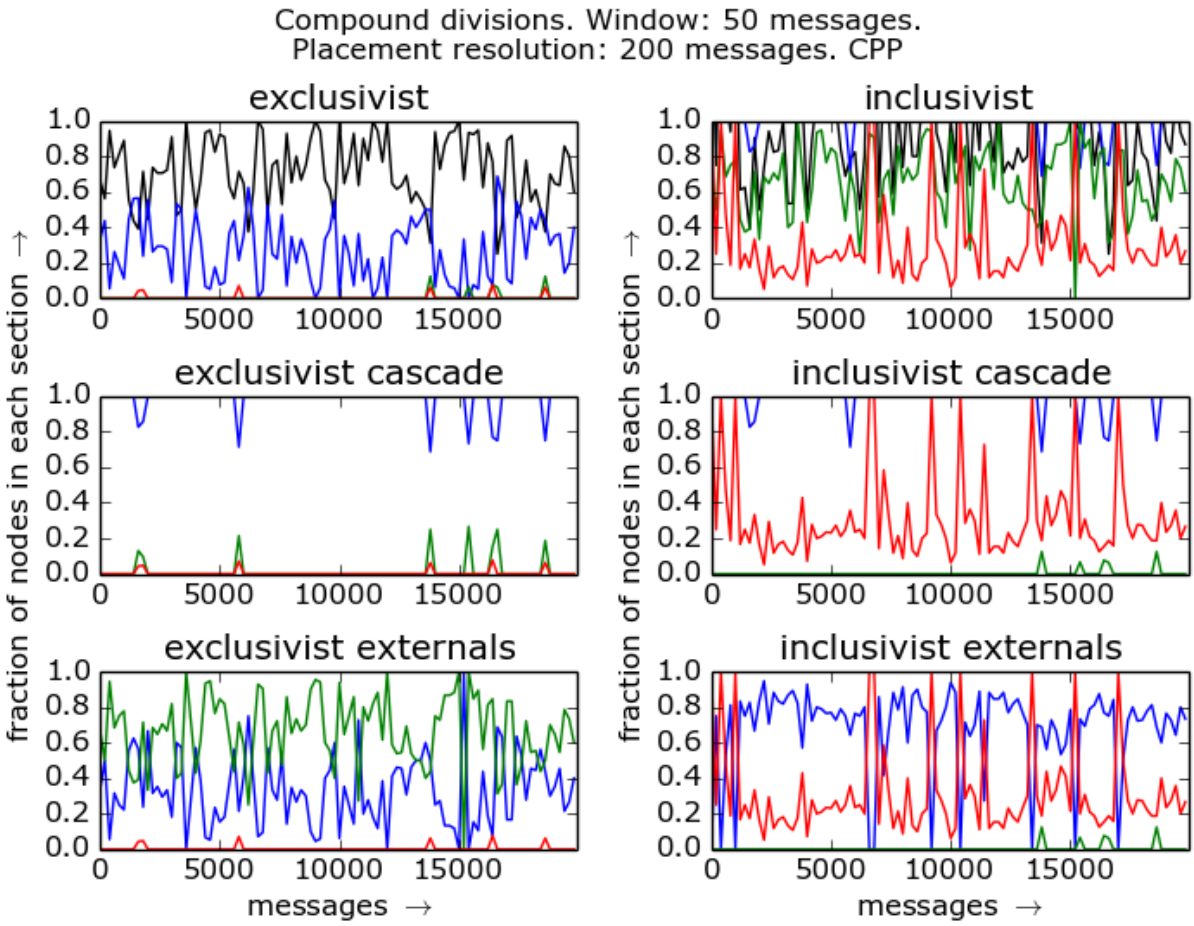
In this section, the figures present timelines with the fractions of participants in each Erdös sector with respect to each criterion defined in Section 2.2.4 of the main document.[?] Step and window sizes of 50, 100, 250, 500, 1000 and 5000 messages are shown below, first for CPP, then for LAD list.

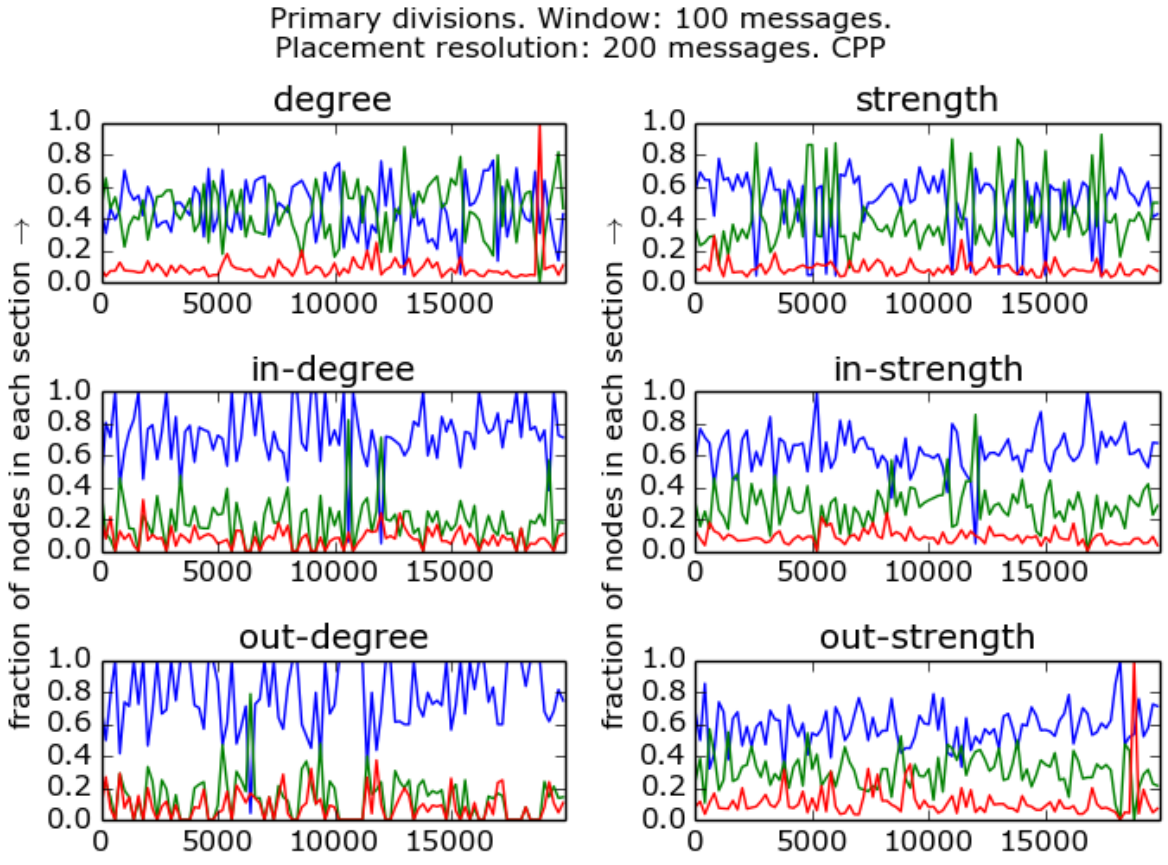
Each step size takes two pages of plot. On the first page, the criterion is based on each centrality metric observed separately: total, in and out degrees (k, k^{in}, k^{out}) and strengths (s, s^{in}, s^{out}). In the first six plots of every page, the colors have meanings as follows: red for hubs, green for the intermediary and blue for the periphery.

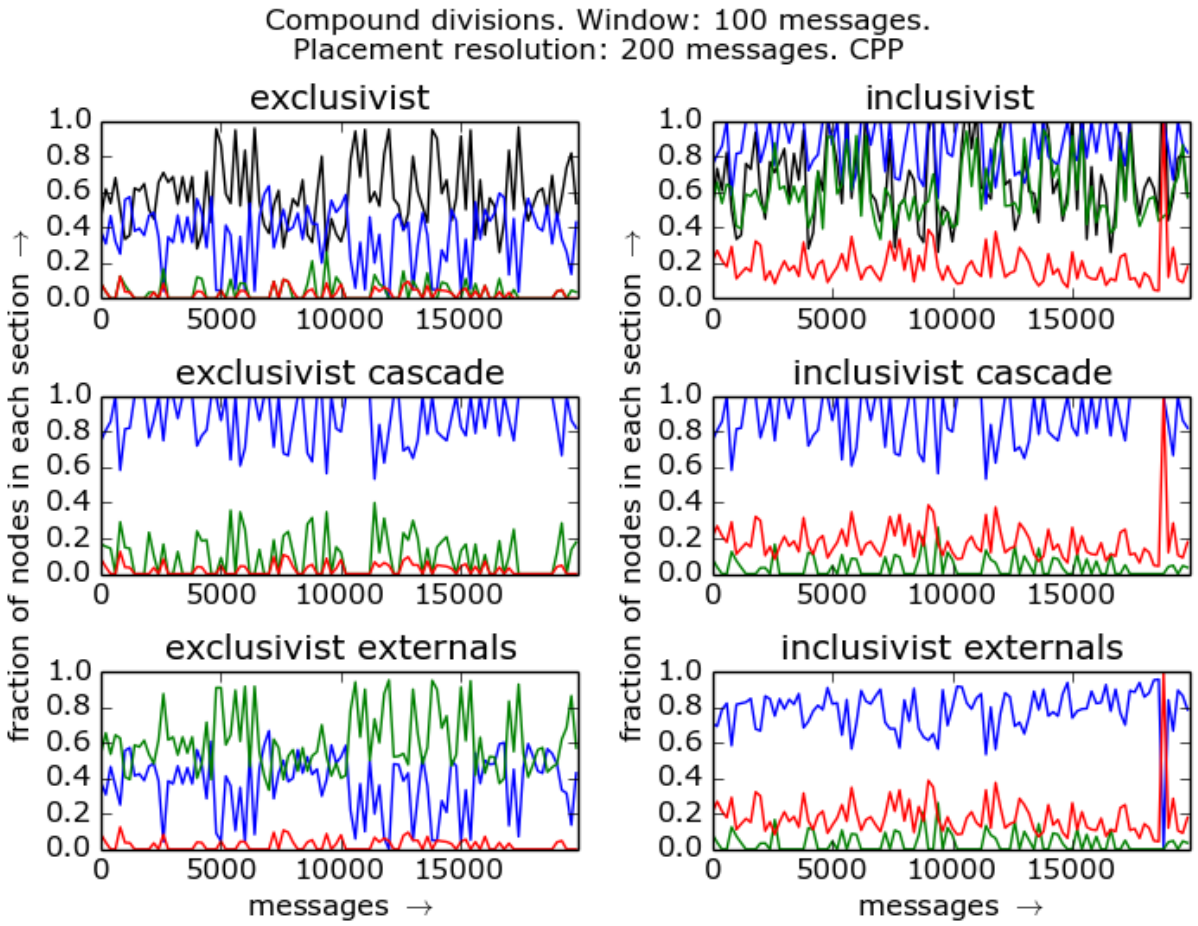
On the second page we show the fractions of participants with respect to each compound criterion for the Erdös sectioning. In the first plot, the fraction of vertices with unique classification is in black: $\frac{\text{number of nodes uniquely classified}}{\text{number of nodes}}$. On the second plot, black represents the fraction of classifications beyond the number of vertices: $\frac{\text{number of classifications} - \text{number of nodes}}{\text{number of nodes}}$.

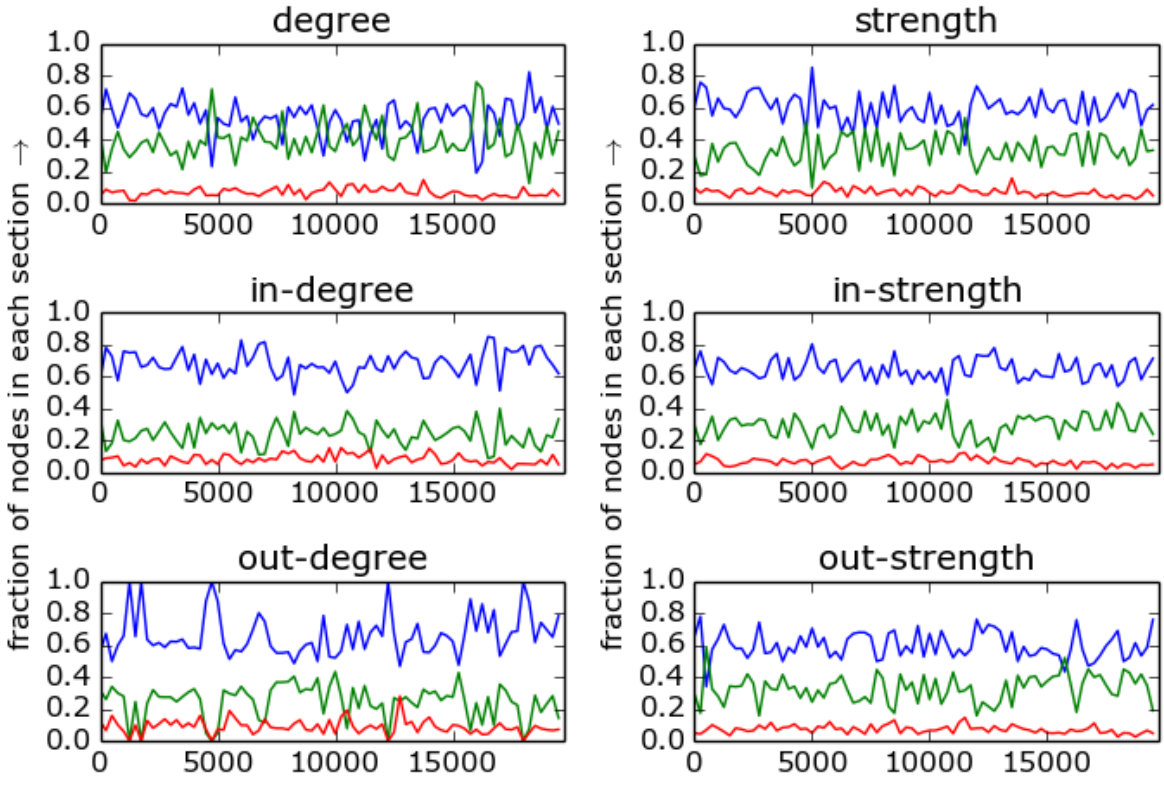
B.3.1 CPP list

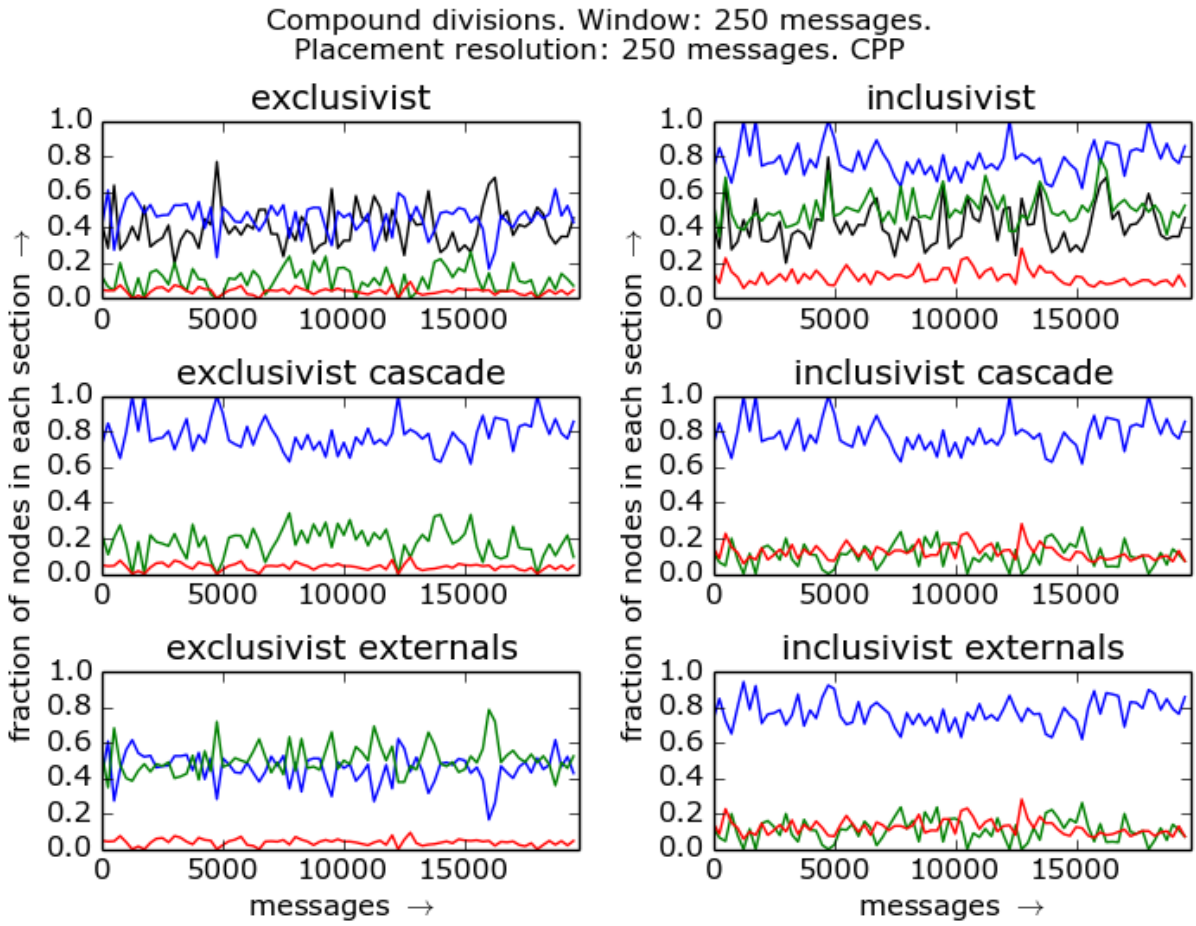


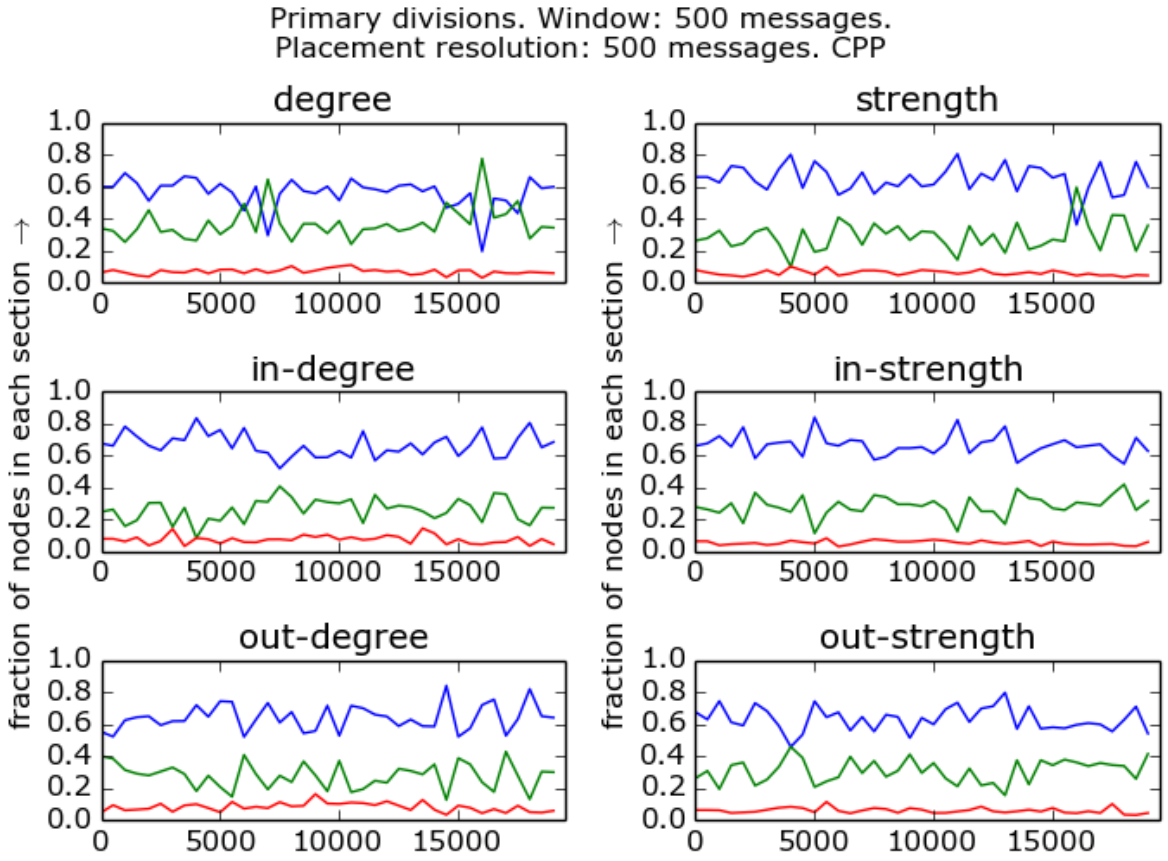

Compound divisions. Window: 50 messages.
Placement resolution: 200 messages. CPP
exclusivist
inclusivist
exclusivist cascade
inclusivist cascade
exclusivist externals
inclusivist externals
fraction of nodes in each section →
messages →
0.0
0.2
0.4
0.6
0.8
1.0
0
5000
10000
15000

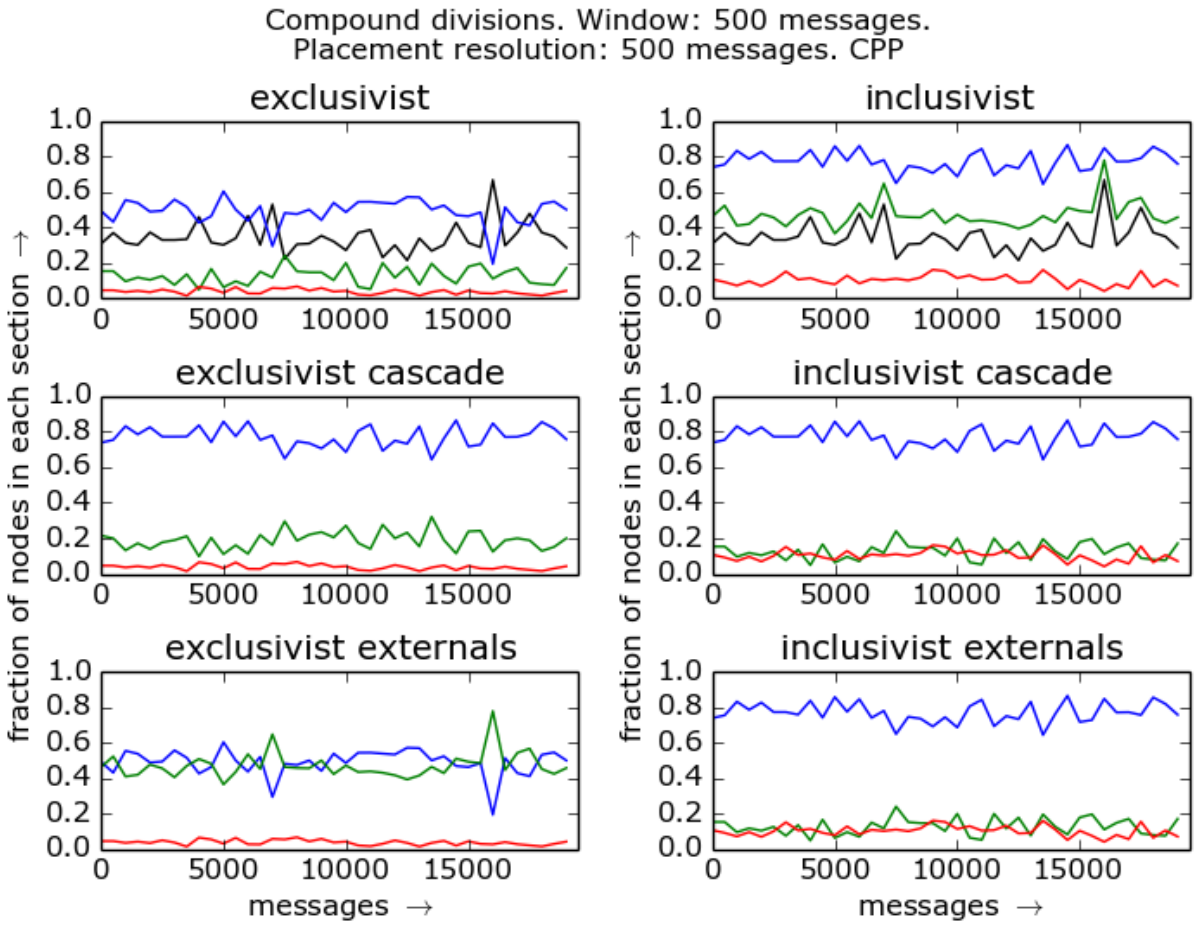

Primary divisions. Window: 100 messages.
Placement resolution: 200 messages. CPP
degree
strength
in-degree
in-strength
out-degree
out-strength
fraction of nodes in each section →
0.0
0.2
0.4
0.6
0.8
1.0
0
5000
10000
15000

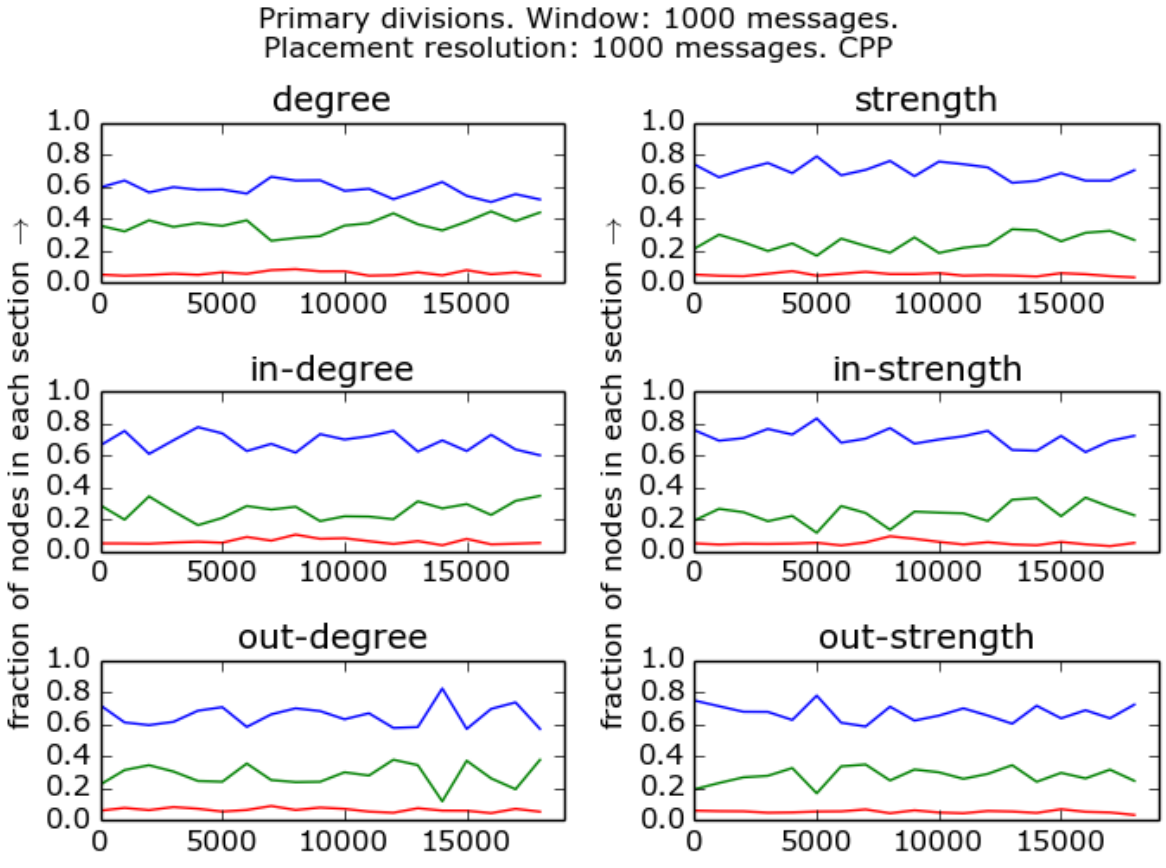

Compound divisions. Window: 100 messages.
Placement resolution: 200 messages. CPP
exclusivist
inclusivist
exclusivist cascade
inclusivist cascade
exclusivist externals
inclusivist externals
fraction of nodes in each section →
messages →
0.0
0.2
0.4
0.6
0.8
1.0
0
5000
10000
15000

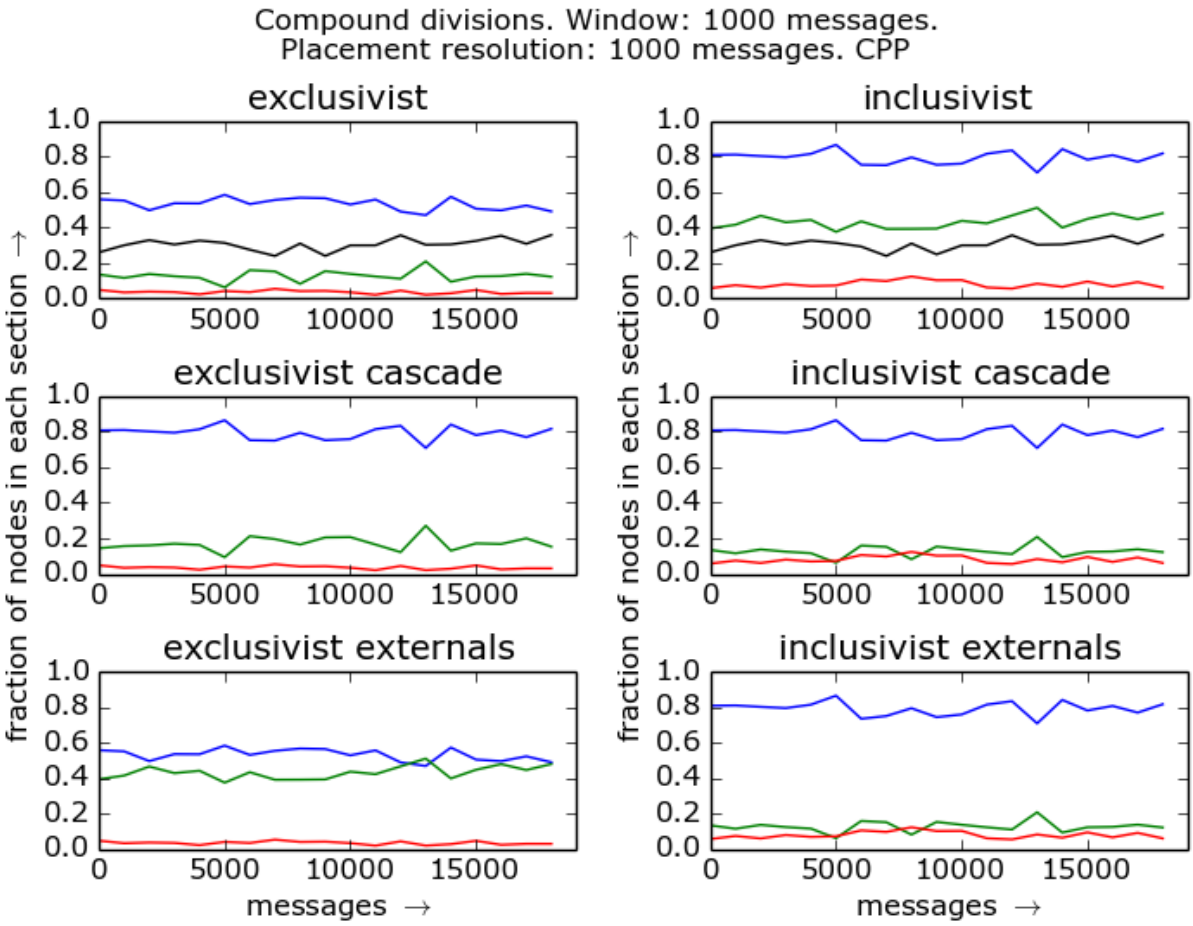

Primary divisions. Window: 250 messages.
Placement resolution: 250 messages. CPP
degree
strength
in-degree
in-strength
out-degree
out-strength
fraction of nodes in each section →
0.0
0.2
0.4
0.6
0.8
1.0
0
5000
10000
15000

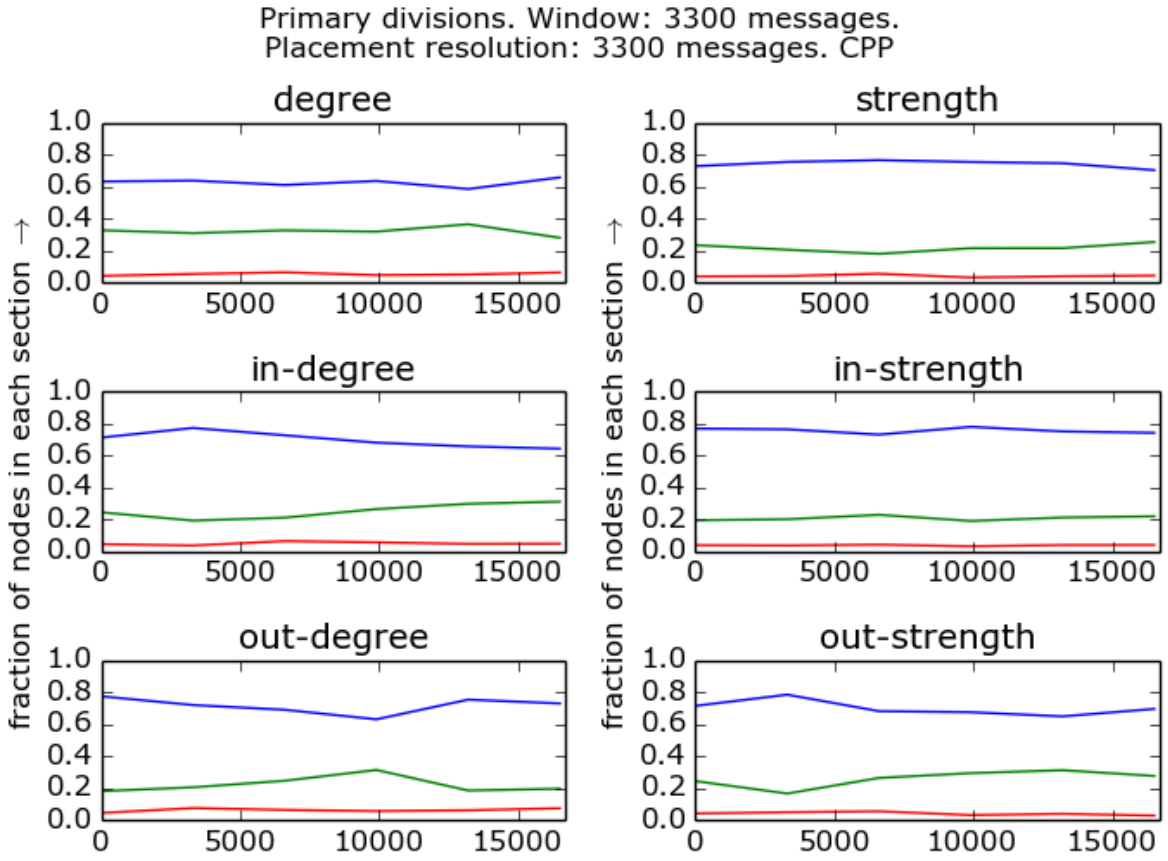

Compound divisions. Window: 250 messages.
Placement resolution: 250 messages. CPP
exclusivist
inclusivist
exclusivist cascade
inclusivist cascade
exclusivist externals
inclusivist externals
fraction of nodes in each section →
messages →
0.0
0.2
0.4
0.6
0.8
1.0
0
5000
10000
15000

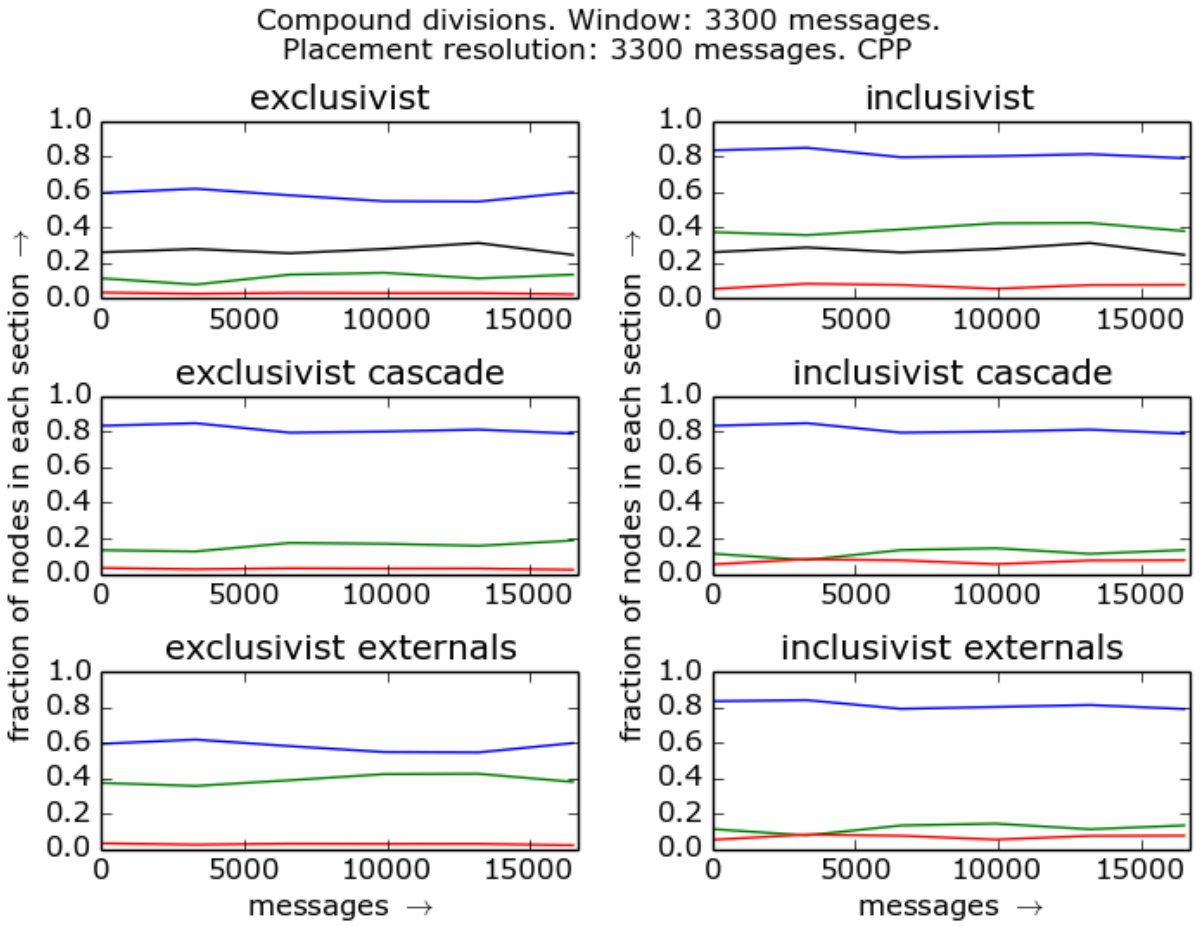

Primary divisions. Window: 500 messages.
Placement resolution: 500 messages. CPP
degree
strength
in-degree
in-strength
out-degree
out-strength
fraction of nodes in each section →
0.0
0.2
0.4
0.6
0.8
1.0
0
5000
10000
15000

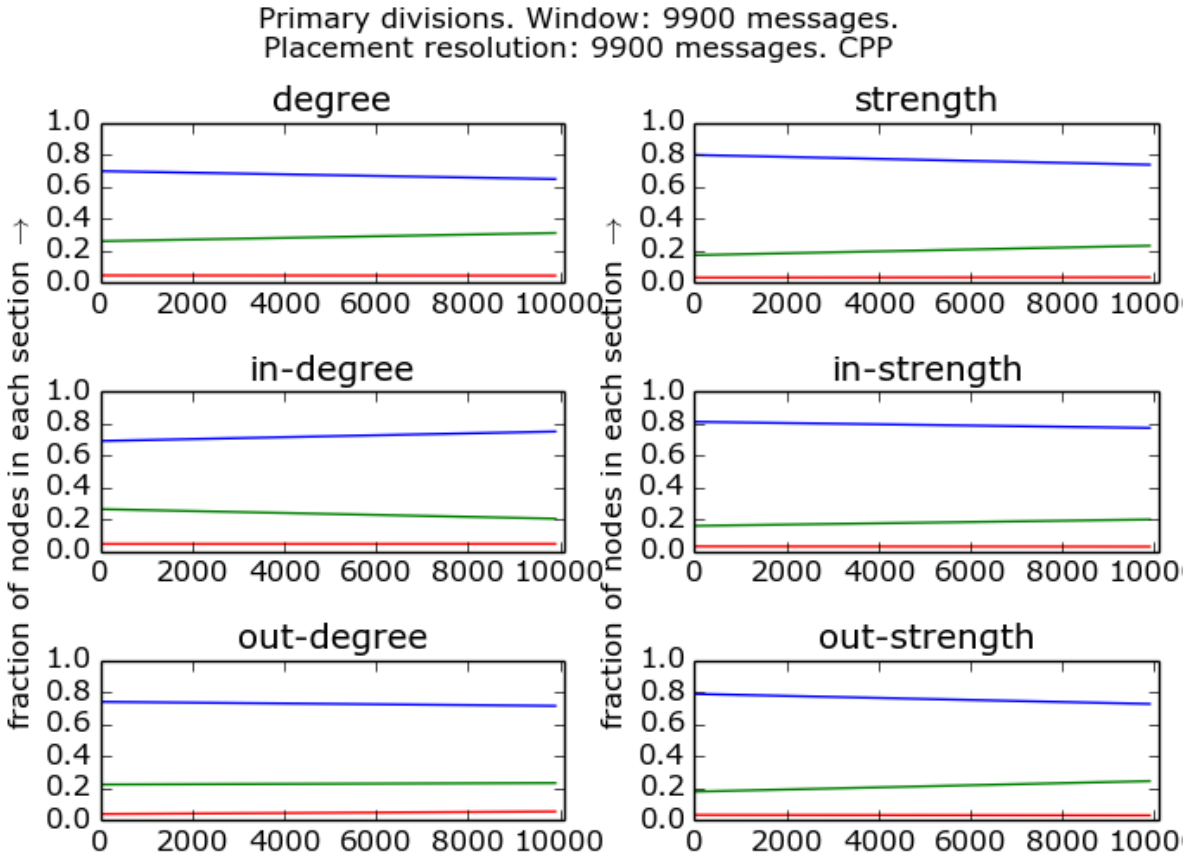

Compound divisions. Window: 500 messages.
Placement resolution: 500 messages. CPP
exclusivist
inclusivist
exclusivist cascade
inclusivist cascade
exclusivist externals
inclusivist externals
fraction of nodes in each section →
messages →
0.0
0.2
0.4
0.6
0.8
1.0
0
5000
10000
15000

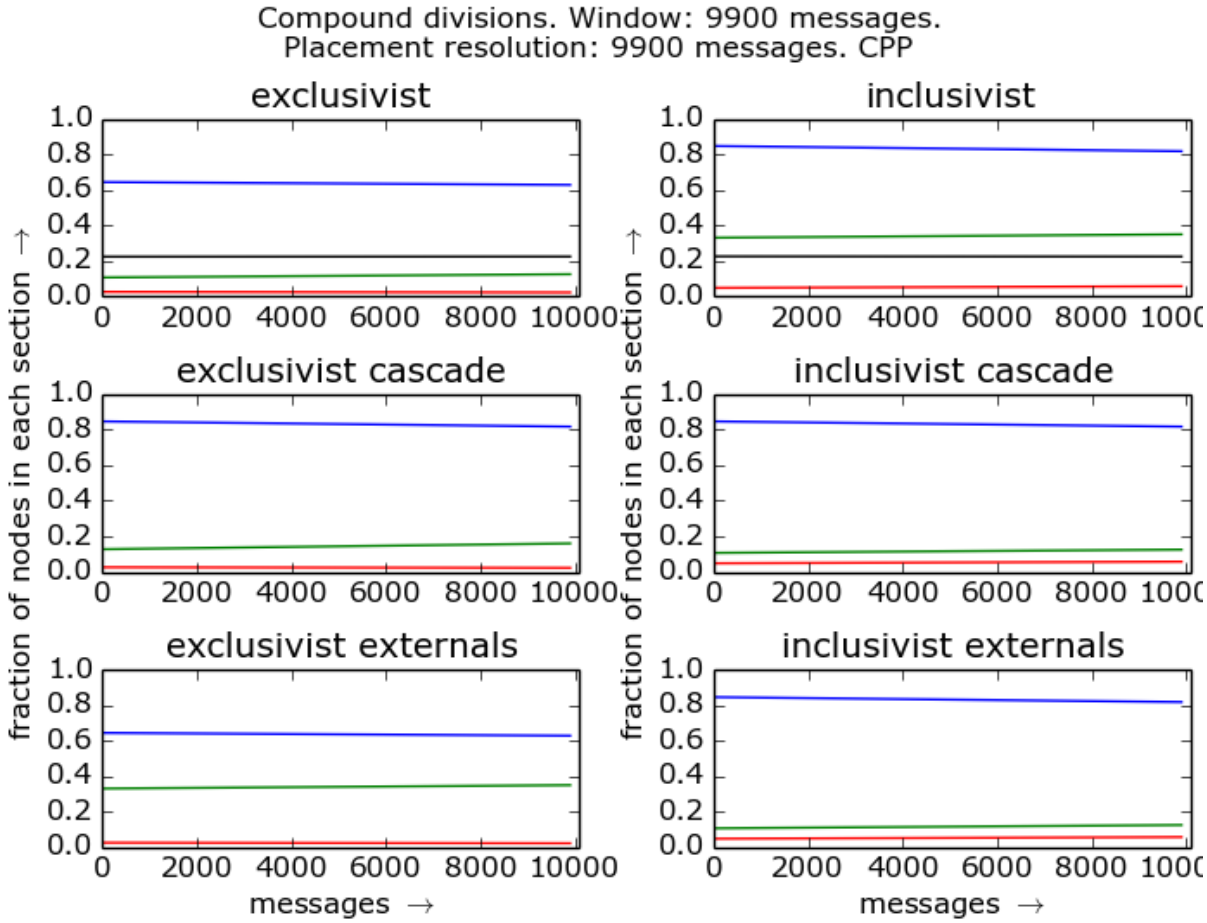

Primary divisions. Window: 1000 messages.
Placement resolution: 1000 messages. CPP
degree
strength
in-degree
in-strength
out-degree
out-strength
fraction of nodes in each section →
0.0
0.2
0.4
0.6
0.8
1.0
0
5000
10000
15000


Compound divisions. Window: 1000 messages.
Placement resolution: 1000 messages. CPP
exclusivist
inclusivist
exclusivist cascade
inclusivist cascade
exclusivist externals
inclusivist externals
fraction of nodes in each section →
messages →
0.0
0.2
0.4
0.6
0.8
1.0
0
5000
10000
15000

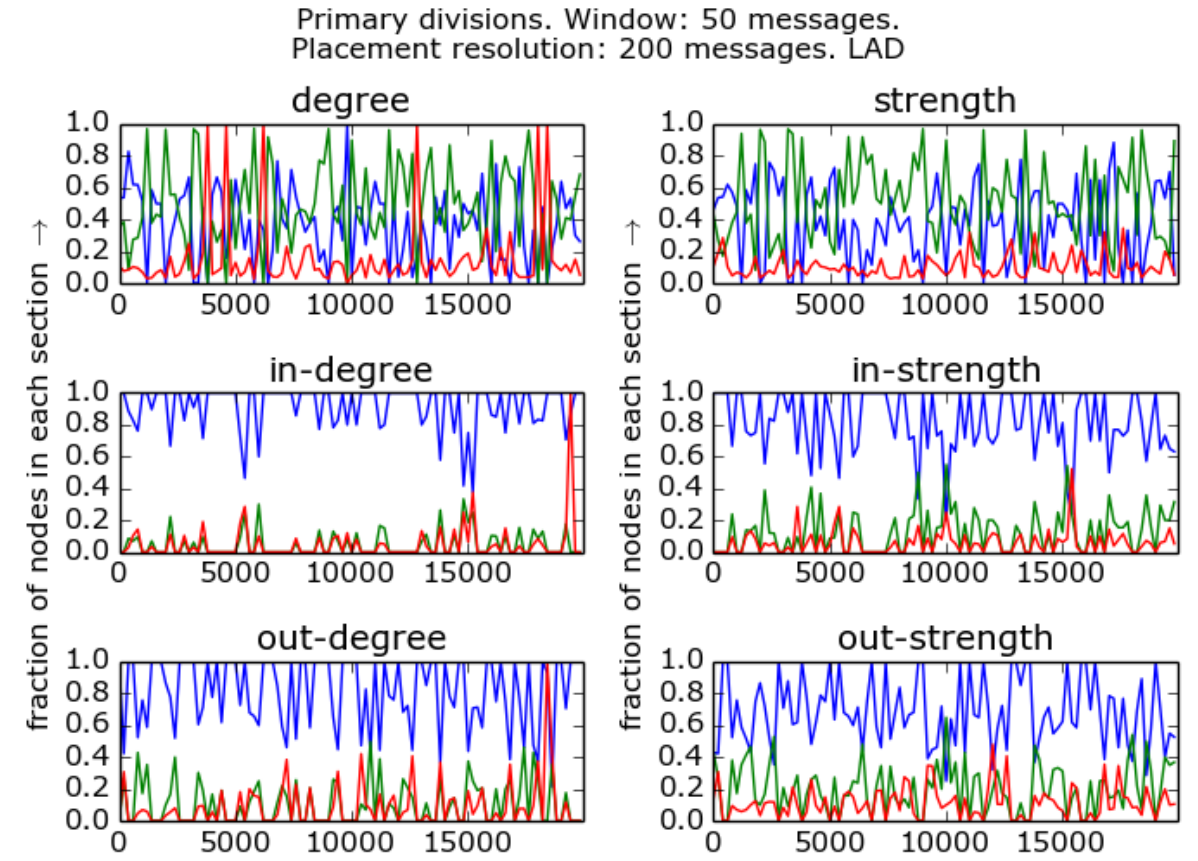

Primary divisions. Window: 3300 messages.
Placement resolution: 3300 messages. CPP
degree
strength
in-degree
in-strength
out-degree
out-strength
fraction of nodes in each section →
0.0
0.2
0.4
0.6
0.8
1.0
0
5000
10000
15000

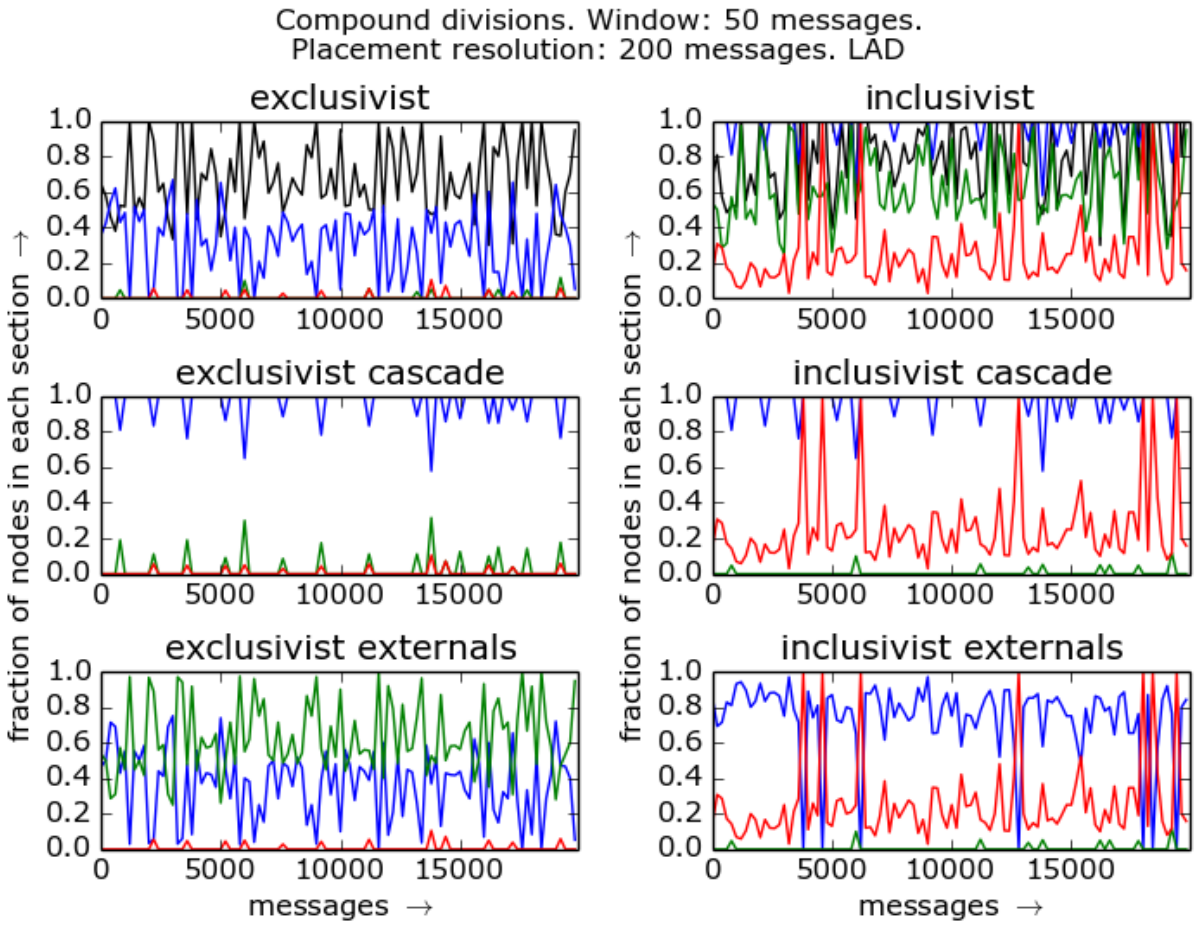

Compound divisions. Window: 3300 messages.
Placement resolution: 3300 messages. CPP
exclusivist
inclusivist
exclusivist cascade
inclusivist cascade
exclusivist externals
inclusivist externals
fraction of nodes in each section →
messages →
0.0
0.2
0.4
0.6
0.8
1.0
0
5000
10000
15000

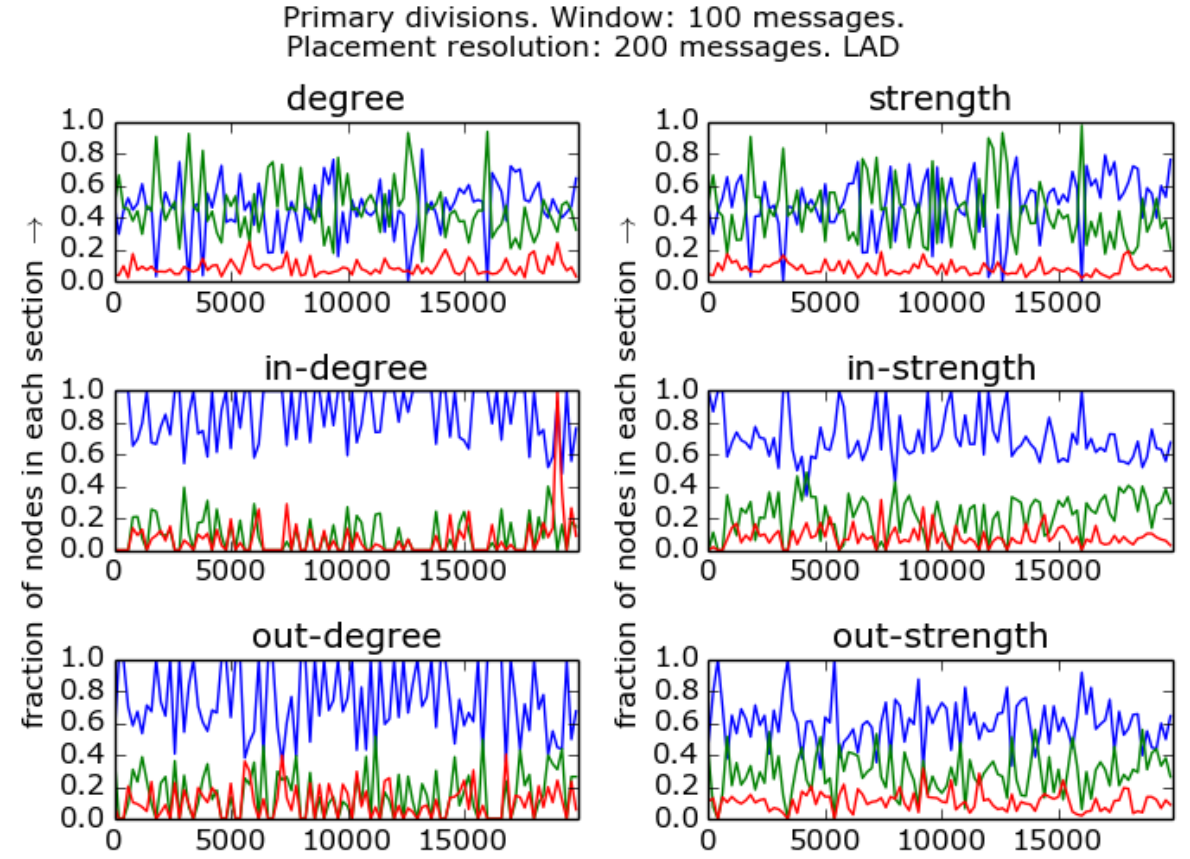

Primary divisions. Window: 9900 messages.
Placement resolution: 9900 messages. CPP
degree
strength
in-degree
in-strength
out-degree
out-strength
fraction of nodes in each section →
0.0
0.2
0.4
0.6
0.8
1.0
0
2000
4000
6000
8000
10000

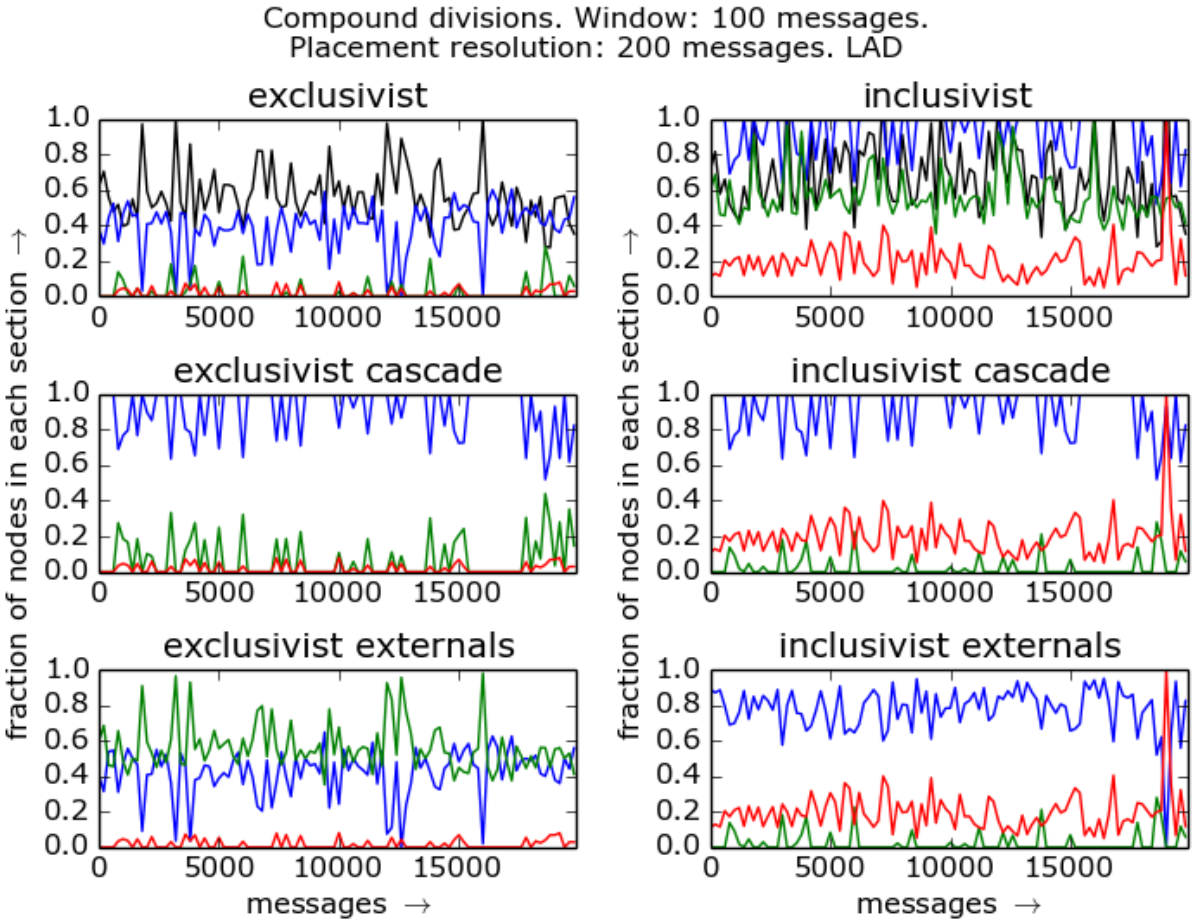

Compound divisions. Window: 9900 messages.
Placement resolution: 9900 messages. CPP
exclusivist
inclusivist
exclusivist cascade
inclusivist cascade
exclusivist externals
inclusivist externals
fraction of nodes in each section →
messages →
0.0
0.2
0.4
0.6
0.8
1.0
0
2000
4000
6000
8000
10000

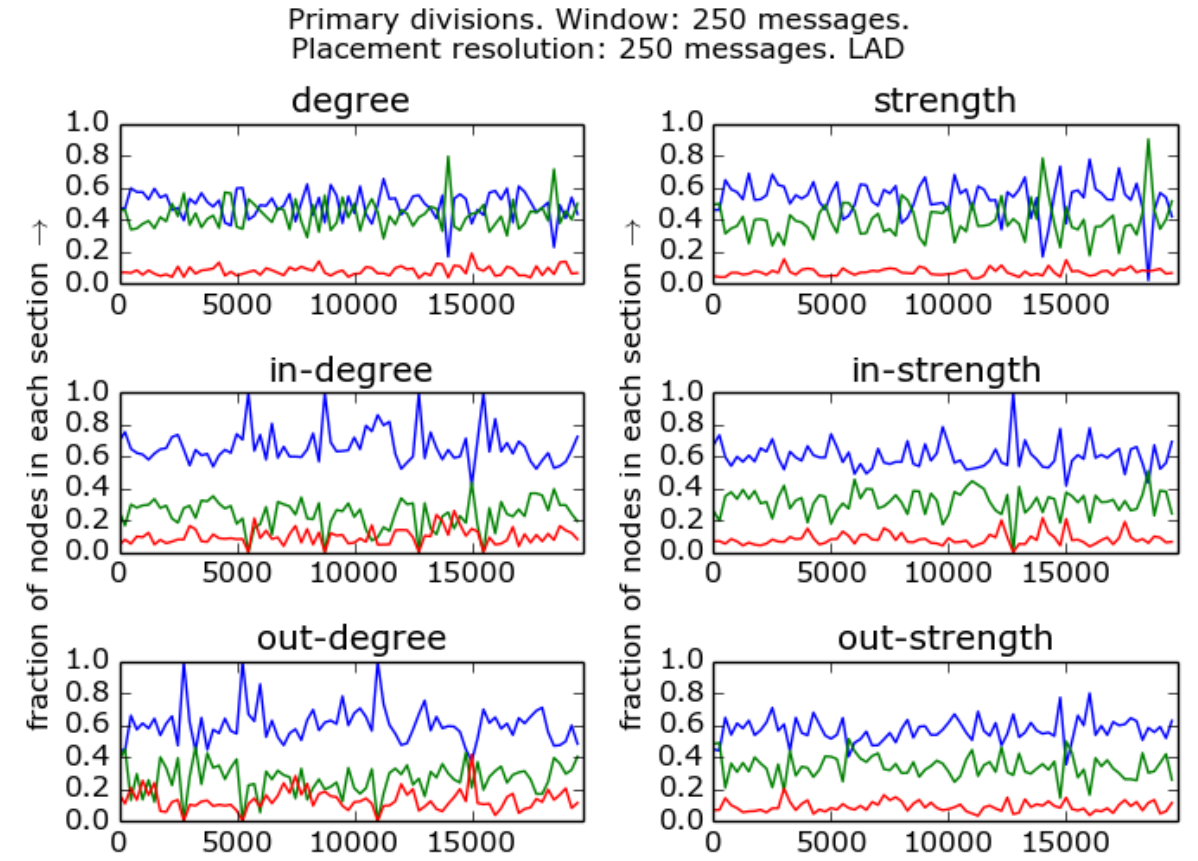
B.3.2 LAD list

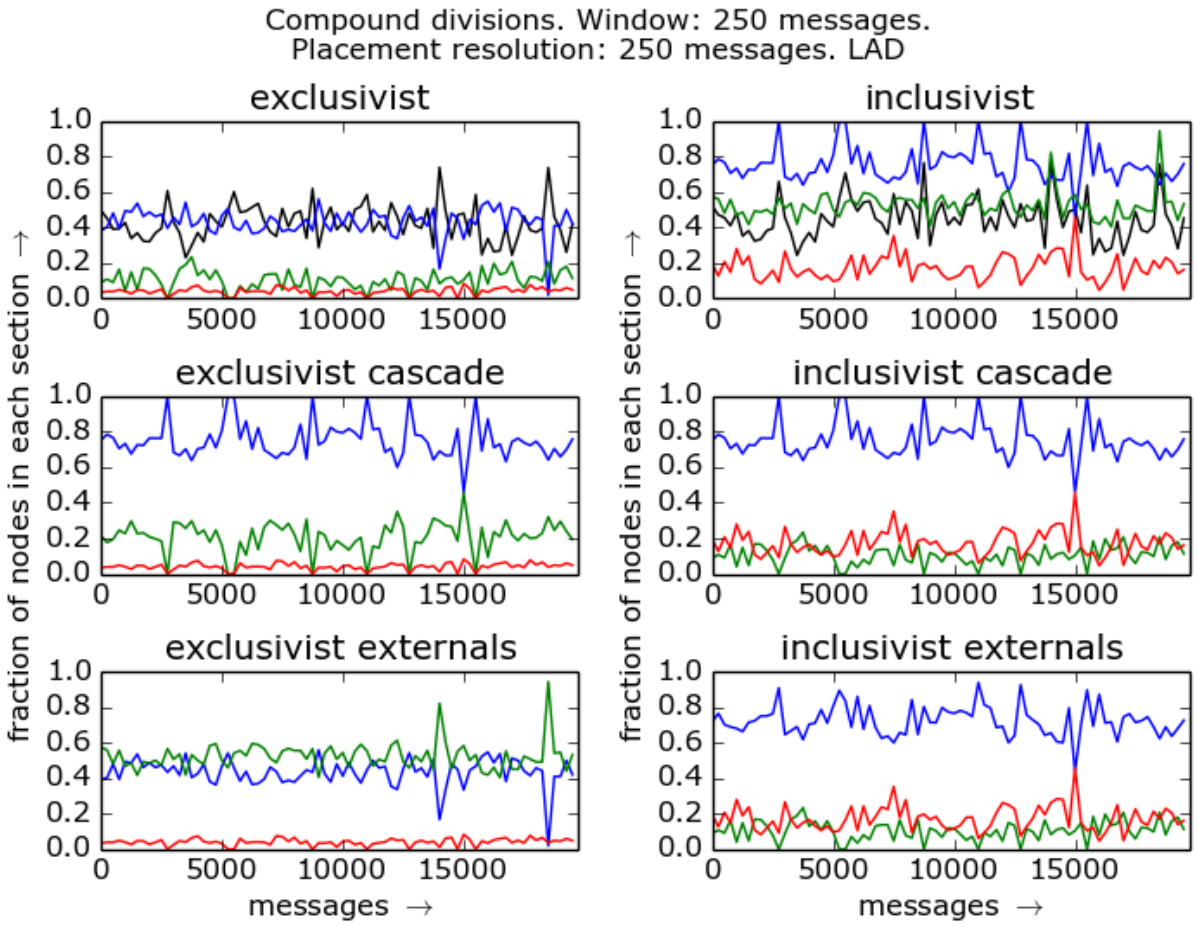

Primary divisions. Window: 50 messages.
Placement resolution: 200 messages. LAD
degree
strength
in-degree
in-strength
out-degree
out-strength
fraction of nodes in each section →
0.0
0.2
0.4
0.6
0.8
1.0
0
5000
10000
15000

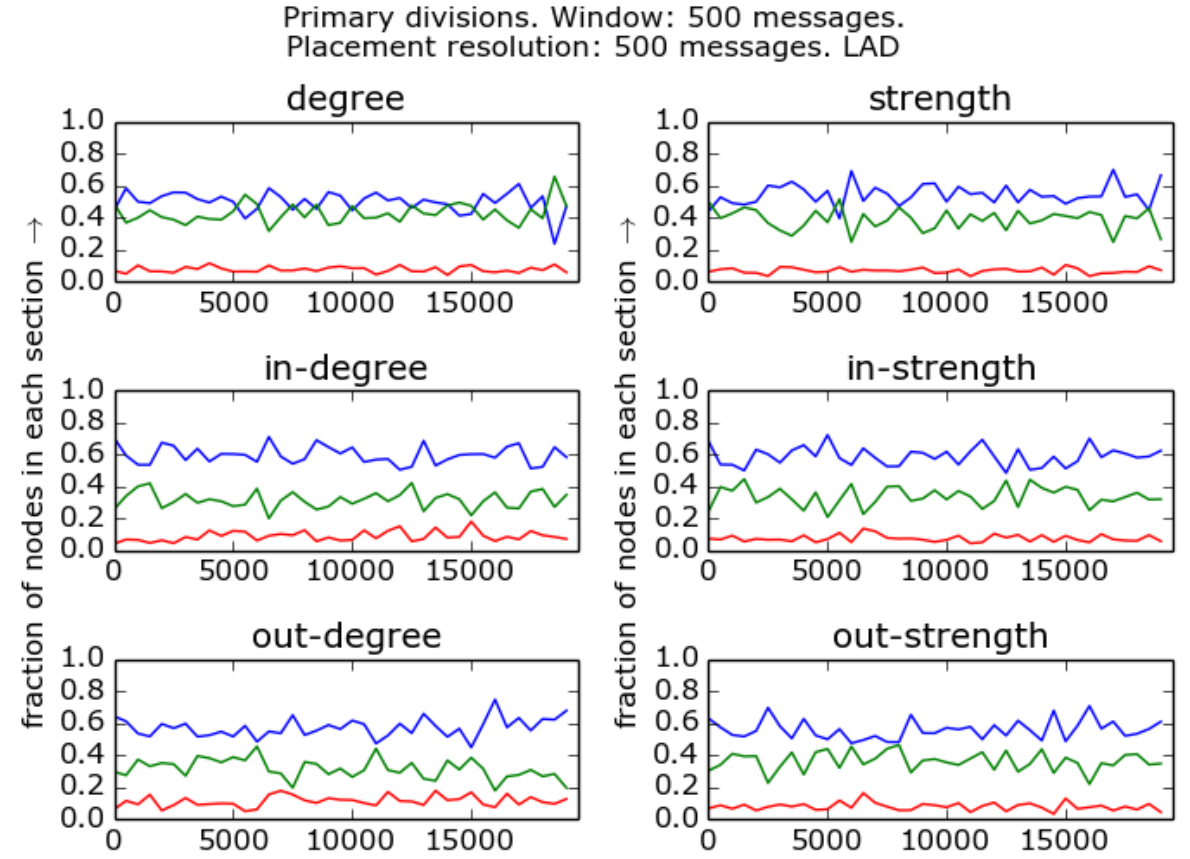

Compound divisions. Window: 50 messages.
Placement resolution: 200 messages. LAD
exclusivist
inclusivist
exclusivist cascade
inclusivist cascade
exclusivist externals
inclusivist externals
fraction of nodes in each section →
messages →
0.0
0.2
0.4
0.6
0.8
1.0
0
5000
10000
15000

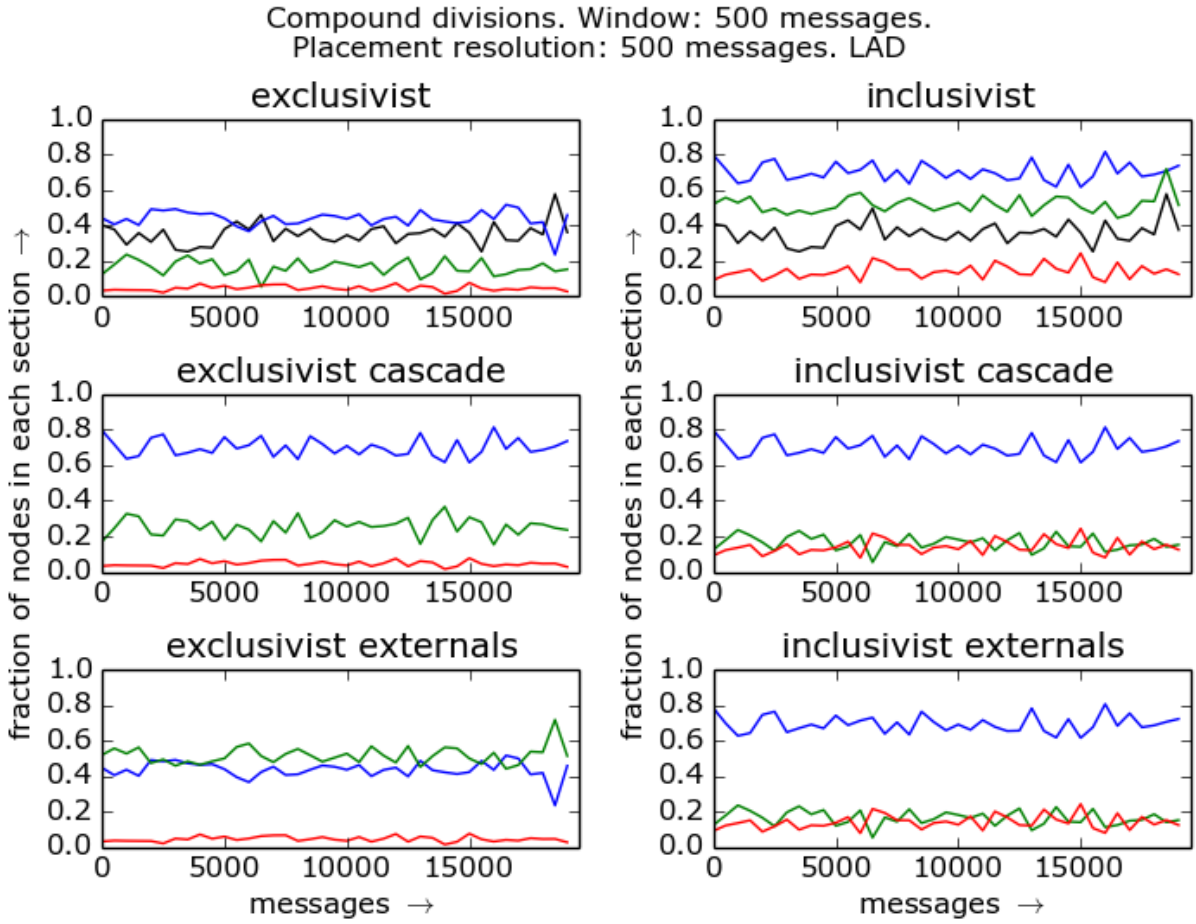

Primary divisions. Window: 100 messages.
Placement resolution: 200 messages. LAD
degree
strength
in-degree
in-strength
out-degree
out-strength
fraction of nodes in each section →
0.0
0.2
0.4
0.6
0.8
1.0
0
5000
10000
15000

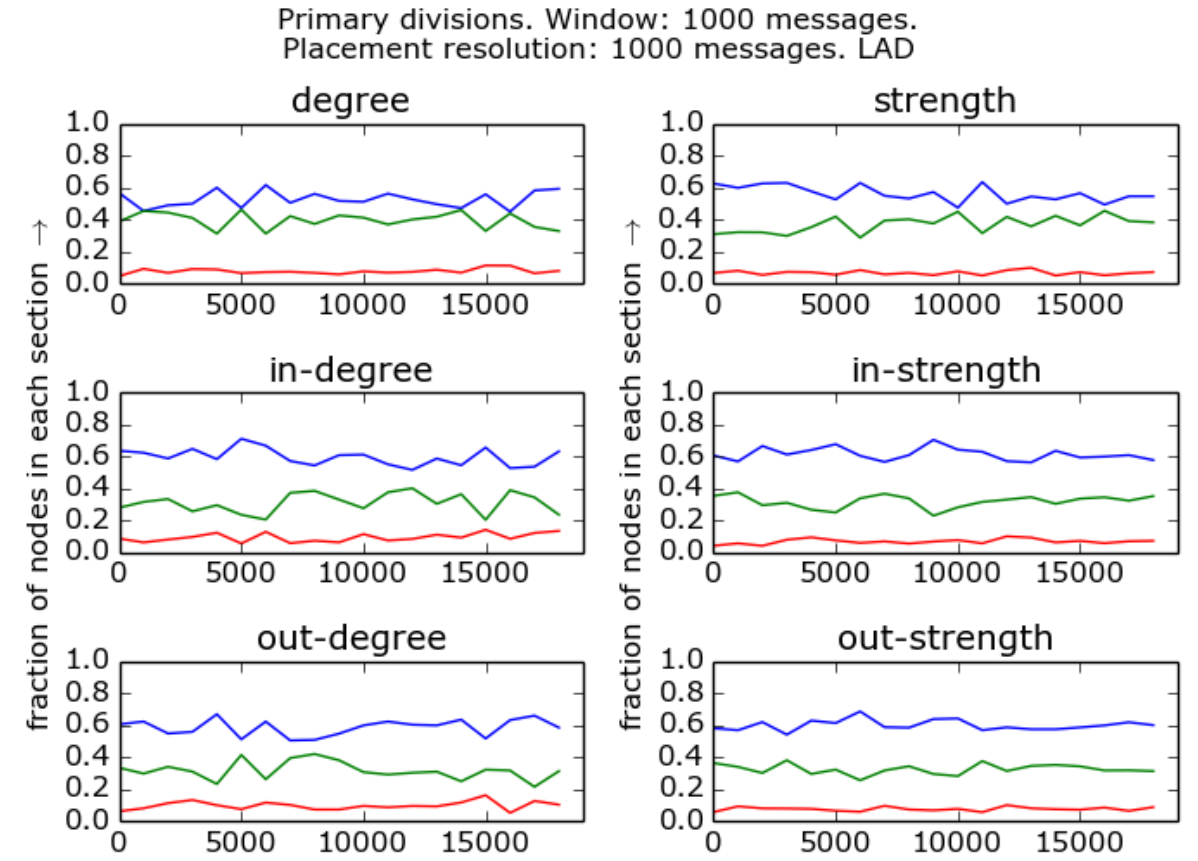

Compound divisions. Window: 100 messages.
Placement resolution: 200 messages. LAD
exclusivist
inclusivist
exclusivist cascade
inclusivist cascade
exclusivist externals
inclusivist externals
fraction of nodes in each section →
messages →
0.0
0.2
0.4
0.6
0.8
1.0
0
5000
10000
15000

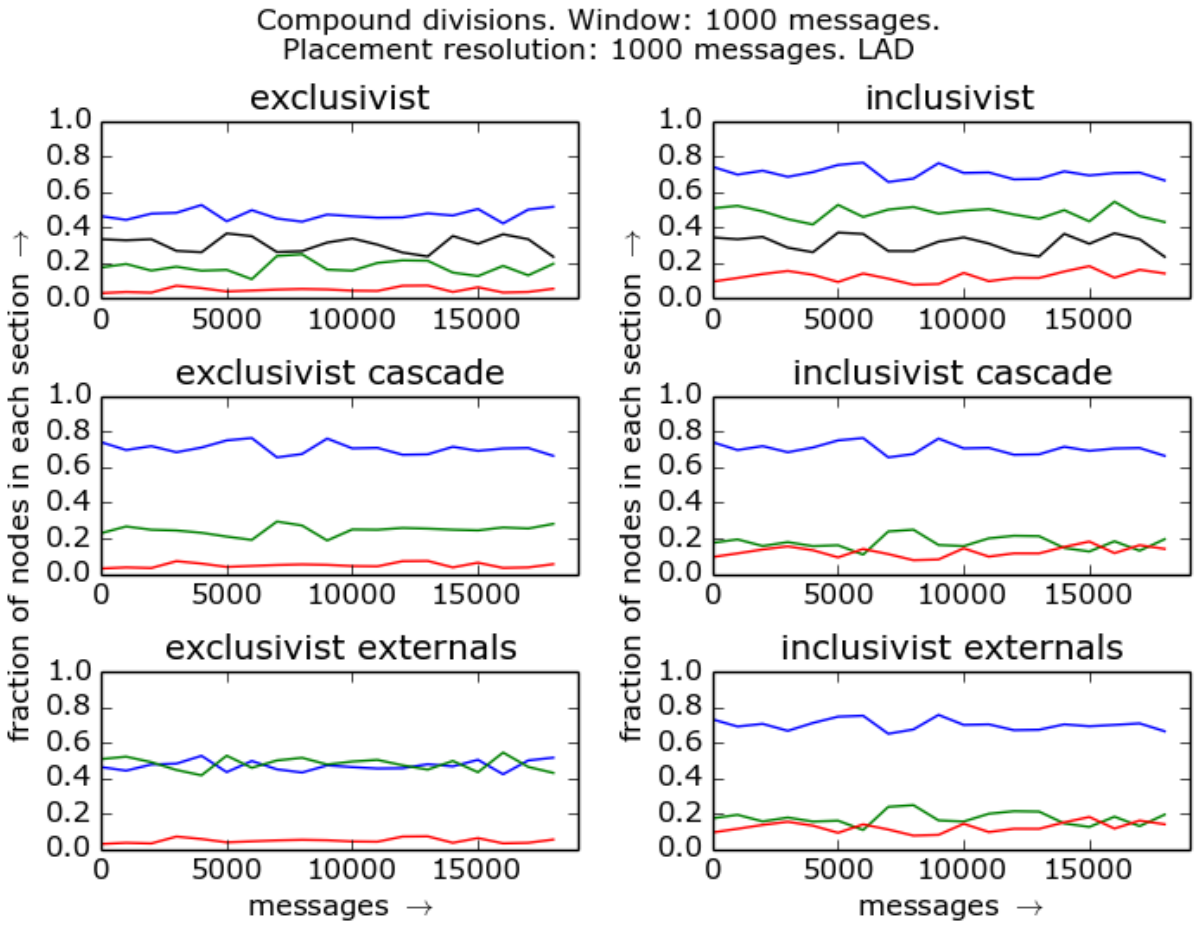

Primary divisions. Window: 250 messages.
Placement resolution: 250 messages. LAD
degree
strength
in-degree
in-strength
out-degree
out-strength
fraction of nodes in each section →
1.0
0.8
0.6
0.4
0.2
0.0
0
5000
10000
15000

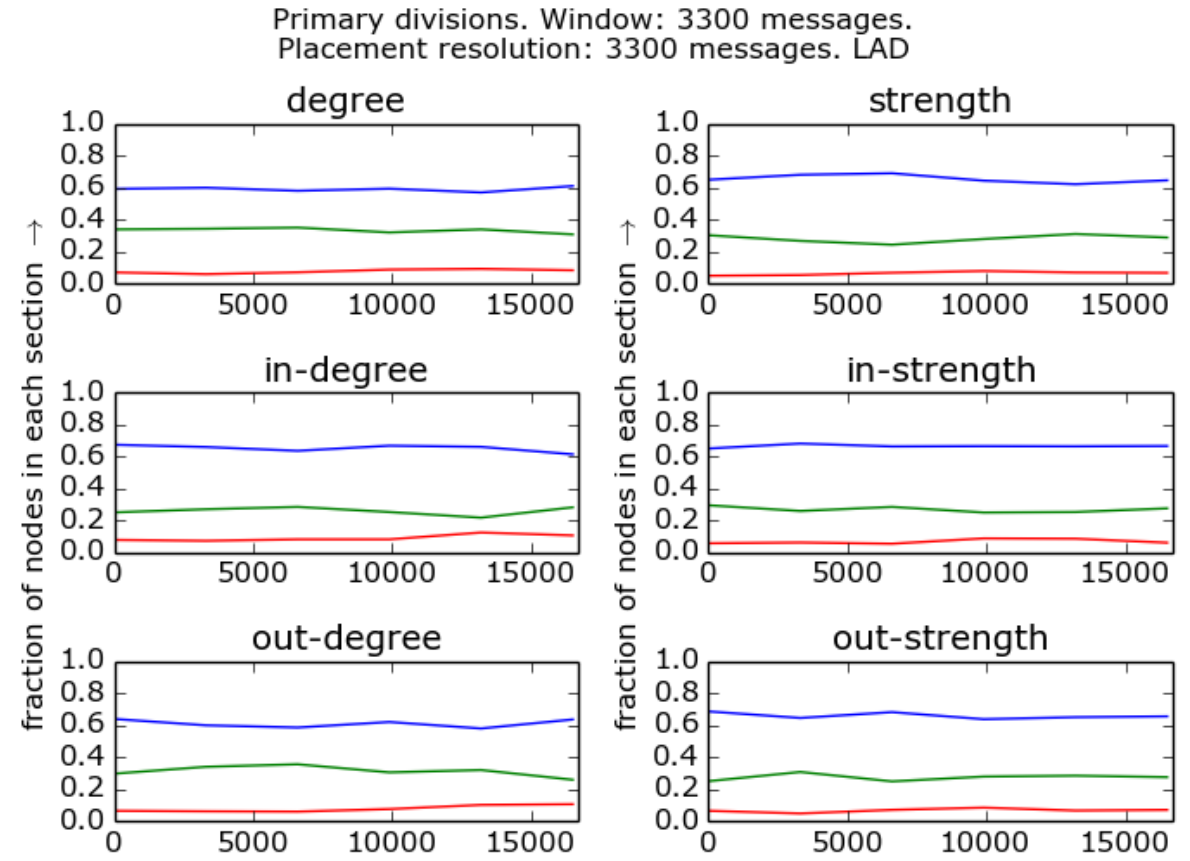

Compound divisions. Window: 250 messages.
Placement resolution: 250 messages. LAD
exclusivist
inclusivist
exclusivist cascade
inclusivist cascade
exclusivist externals
inclusivist externals
fraction of nodes in each section →
messages →
1.0
0.8
0.6
0.4
0.2
0.0
0
5000
10000
15000

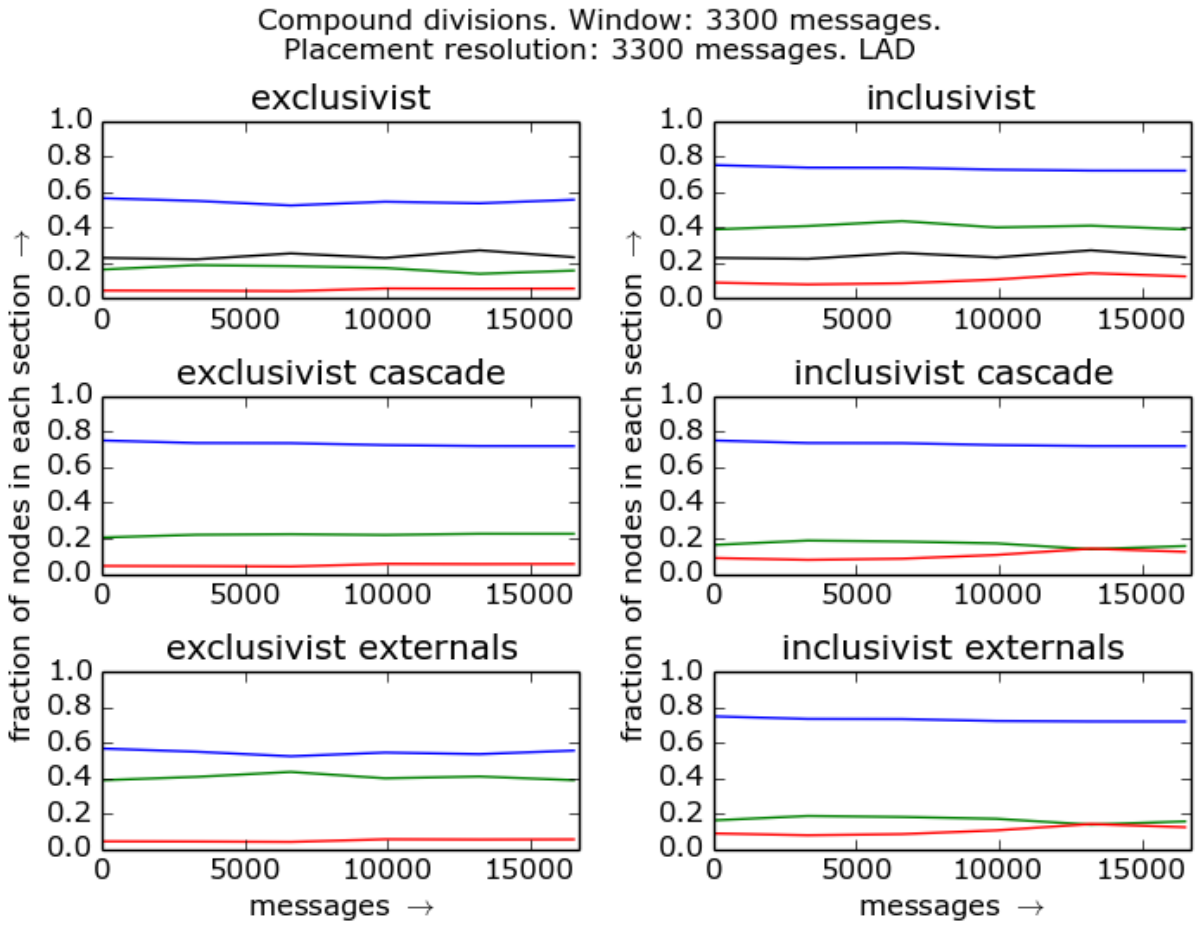

Primary divisions. Window: 500 messages.
Placement resolution: 500 messages. LAD
degree
strength
in-degree
in-strength
out-degree
out-strength
fraction of nodes in each section →
0.0
0.2
0.4
0.6
0.8
1.0
0
5000
10000
15000

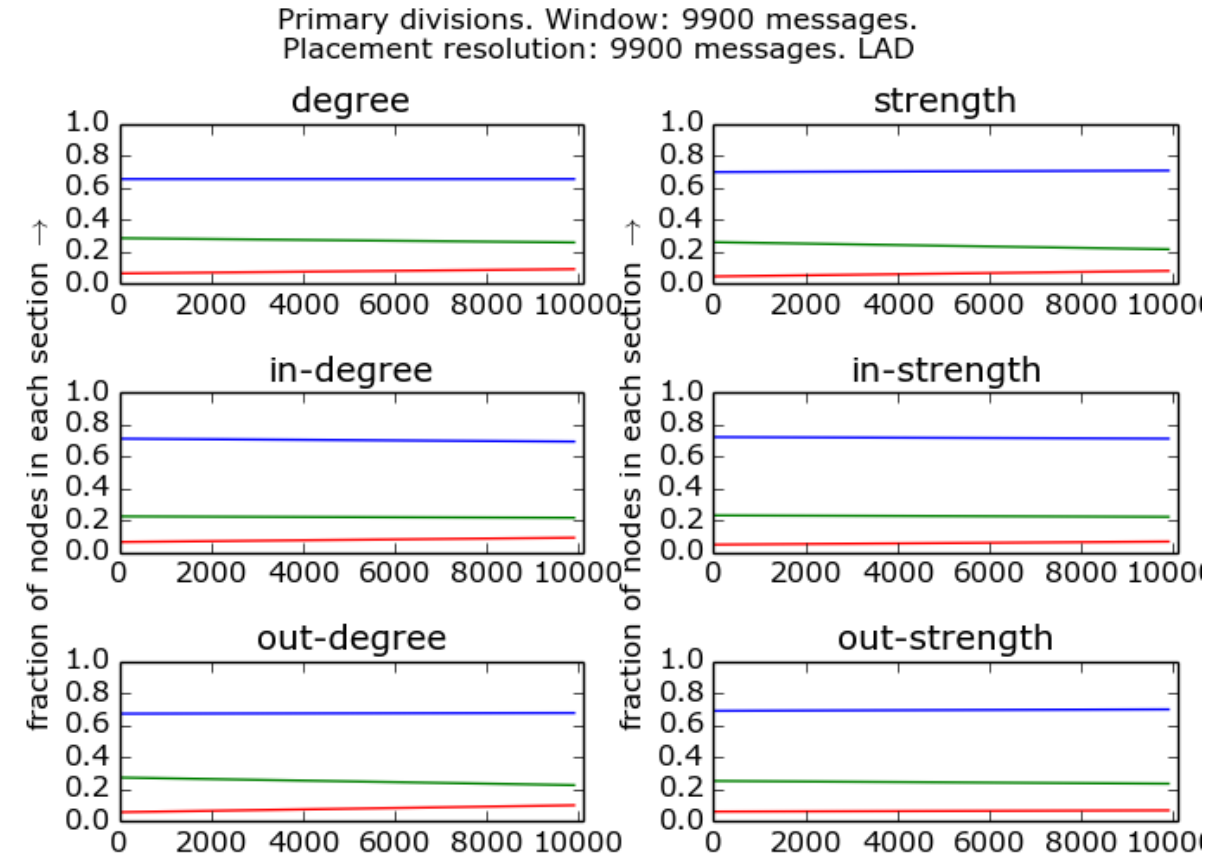

Compound divisions. Window: 500 messages.
Placement resolution: 500 messages. LAD
exclusivist
inclusivist
exclusivist cascade
inclusivist cascade
exclusivist externals
inclusivist externals
fraction of nodes in each section →
messages →
0.0
0.2
0.4
0.6
0.8
1.0
0
5000
10000
15000

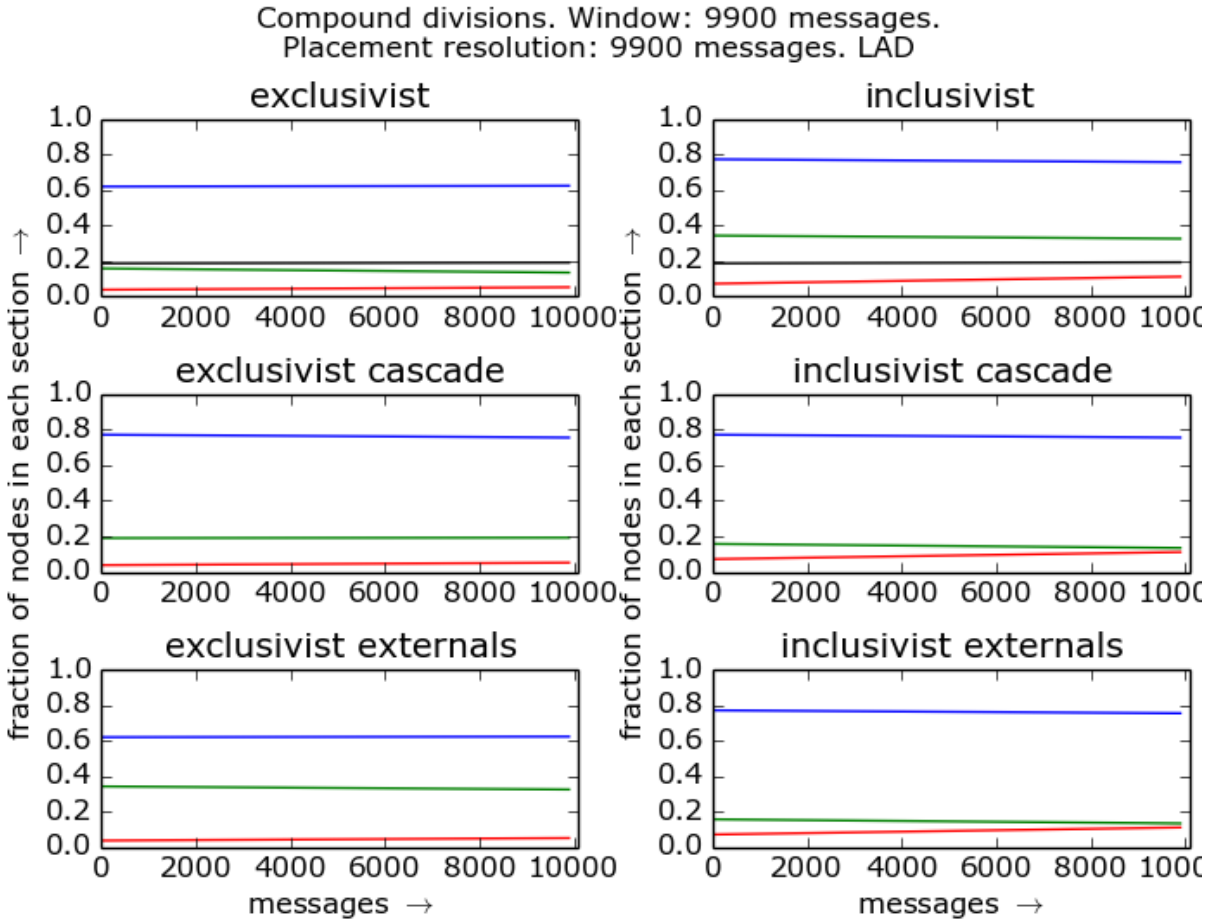

Primary divisions. Window: 1000 messages.
Placement resolution: 1000 messages. LAD
degree
strength
in-degree
in-strength
out-degree
out-strength
fraction of nodes in each section →
0.0
0.2
0.4
0.6
0.8
1.0
0
5000
10000
15000


Compound divisions. Window: 1000 messages.
Placement resolution: 1000 messages. LAD
exclusivist
inclusivist
exclusivist cascade
inclusivist cascade
exclusivist externals
inclusivist externals
fraction of nodes in each section →
messages →
0.0
0.2
0.4
0.6
0.8
1.0
0
5000
10000
15000


Primary divisions. Window: 3300 messages.
Placement resolution: 3300 messages. LAD
degree
strength
in-degree
in-strength
out-degree
out-strength
fraction of nodes in each section →
0.0
0.2
0.4
0.6
0.8
1.0
0
5000
10000
15000


Compound divisions. Window: 3300 messages.
Placement resolution: 3300 messages. LAD
exclusivist
inclusivist
exclusivist cascade
inclusivist cascade
exclusivist externals
inclusivist externals
fraction of nodes in each section →
messages →
0.0
0.2
0.4
0.6
0.8
1.0
0
5000
10000
15000


Primary divisions. Window: 9900 messages.
Placement resolution: 9900 messages. LAD
degree
strength
in-degree
in-strength
out-degree
out-strength
fraction of nodes in each section →
0.0
0.2
0.4
0.6
0.8
1.0
0
2000
4000
6000
8000
10000


Compound divisions. Window: 9900 messages.
Placement resolution: 9900 messages. LAD
exclusivist
inclusivist
exclusivist cascade
inclusivist cascade
exclusivist externals
inclusivist externals
fraction of nodes in each section →
messages →
0.0
0.2
0.4
0.6
0.8
1.0
0
2000
4000
6000
8000
10000

B.4 Stability in networks from Twitter, Facebook, Participabr

To ease hypothesizing about the generality of the reported stability of human interaction networks, this section presents the topological analysis of networks from Twitter, Facebook and Participabr. Selected networks are summarized in Table 74. Their Erdös sector relative sizes are given in Table 75. The formation of Principal components are given in Tables 76, 77, 78 and 79. The friendship networks considered are undirected and unweighted, therefore all measurements of strength, in- and out- centralities, asymmetry and disequilibrium have little or no meaning, which is why F1, F2, F3, F4 and F5 are only present in Table 76. The most important results from this analysis are:

- a further indicative that the stability reported with a focus on email interaction networks is valid for a broader class of phenomena.
- The stability in email interaction networks is higher than for the other networks, considering the same number of participants. This was especially important in choosing email networks for benchmarking and probing general properties.

Table 74: Selected networks from three social platforms: Facebook, Twitter and Participabr. Both friendship and interaction networks were observed, yielding undirected and directed networks, respectively. The number of agents N and the number of edges z are given on the last columns. The acronyms, one for each network, are used throughout Tables 75, 77, 76, 78 and 79. Data was collected in 2013 and 2014 within the anthropological physics framework.[?]

acronym	provenance	edge	directed	description
F1	Facebook	friendship	no	the friendship network of Renato Fa
F2	Facebook	friendship	no	the friendship network of Massimo Canevacci
F3	Facebook	friendship	no	the friendship network of a brazilian direc
F4	Facebook	friendship	no	the friendship network of the Silicon Valley C
F5	Participa.br	friendship	no	the friendship network of a brazilian federal so
I1	Facebook	interaction	yes	the interaction network of the Silicon Valley C
I2	Facebook	interaction	yes	the interaction network of a Solidarity
I3	Facebook	interaction	yes	the interaction network of a brazilian direc
I4	Facebook	interaction	yes	the interaction network of the 'Cience wit
I5	Participa.br	interaction	yes	the interaction network of a brazilian federal so
TT1	Twitter	retweet	yes	the retweet network of $\approx 22k$ tweets with the has
TT2	Twitter	retweet	yes	same as TT1, but disconnected agents a

Table 75: Percentage of agents in each Erdös sector in the friendship and interaction networks of Table 74. The ratios found in email networks are preserved. I1 and I4 are outliers, probably because they should be better characterized as a superposition of networks, rather than one coherent network. The degree was used for establishing the sectors.

	periphery	intermediary	hubs
F1	53.11	43.31	3.58
F2	58.98	39.29	1.72
F3	65.41	31.87	2.72
F4	66.49	32.03	1.48
F5	62.98	36.12	0.90
I1	4.81	94.23	0.96
I2	53.12	45.31	1.56
I3	58.41	40.19	1.40
I4	39.06	59.43	1.51
I5	54.95	43.69	1.35
TT1	74.86	24.49	0.65
TT2	76.57	22.86	0.57

Table 76: Formation of first three principal components for each of the five friendship networks of Table 74 in the simplest case: dimensions correspond to degree k, clustering coefficient cc and betweenness centrality bt. Participabr yields the networks that most resemble the email networks. Overall, the general characteristic is preserved: first component is an average of degree and betweenness, while clustering is the most relevant for the second component. The friendship network of Renato Fabbri (F1) is the only network whose first component has more than 20% of clustering coefficient and second component has more than 20% of degree centrality.

	PC1					PC2					PC3			
	F1	F2	F3	F4	F5	F1	F2	F3	F4	F5	F1	F2	F3	F4
cc	25.80	12.22	11.54	5.04	0.94	58.87	78.22	68.86	90.39	88.86	6.95	1.13	18.20	3.02
k	36.43	43.96	45.61	47.40	49.50	25.66	9.98	6.10	7.63	6.42	44.94	49.52	42.00	48.4
bt	37.77	43.82	42.85	47.56	49.56	15.46	11.80	25.04	1.98	4.72	48.11	49.35	39.80	48.5
λ	53.15	53.06	46.26	55.36	63.80	28.69	32.57	34.27	33.25	33.57	18.16	14.37	19.47	11.3

Table 77: Formation of the first three principal components for each of the seven interaction networks of Table 74 in the simplest case: dimensions correspond to degree k, clustering coefficient cc and betweenness centrality bt. Twitter yields the networks that most resemble the email networks. Overall, the general characteristic is preserved: first component is an average of degree and betweenness, while clustering is the most relevant for the second component.

	PC1							PC2					
	I1	I2	I3	I4	I5	TT1	TT2	I1	I2	I3	I4	I5	TT1
cc	14.43	17.12	11.54	0.69	13.26	2.17	2.72	74.78	70.72	79.30	96.63	76.59	95.75
k	42.68	41.77	44.37	49.65	43.41	48.94	48.67	13.85	11.48	8.31	2.35	11.26	0.14
bt	42.89	41.11	44.09	49.66	43.34	48.89	48.61	11.37	17.80	12.39	1.02	12.15	4.12
λ	64.58	61.97	56.95	62.01	50.92	64.82	64.83	31.57	30.98	32.56	33.35	32.51	33.33

Table 78: Formation of the first three principal components for each of the seven interaction networks of Table 74 considering total, in- and out- degrees (k, k^{in}, k^{out}) and strengths (s, s^{in}, s^{out}), clustering coefficient cc and betweenness centrality bt. Twitter yields the networks that most resemble email networks. The general characteristic is preserved: first component is an average of degree and betweenness, while clustering coefficient is the most relevant for the second component. Important differences are: the clustering coefficient was only important to the third component for two of the networks ($I2$, $I3$) and does not contribute significantly to any of the first three principal components in $I5$; in the first component, I5 exhibited less contribution from in-strength, in-degree and betweenness, I4 exhibited less contribution from out-degree.

	PC1							PC2					
	I1	I2	I3	I4	I5	TT1	TT2	I1	I2	I3	I4	I5	TT1
cc	2.79	4.34	2.57	0.82	1.29	0.66	0.76	28.44	9.46	3.29	21.95	6.95	29.82
s	15.28	15.84	16.46	16.01	16.70	15.49	15.47	3.78	4.95	2.90	3.26	17.78	1.95
s^{in}	14.48	12.81	13.62	14.63	4.50	11.85	11.84	11.77	18.29	17.41	12.44	16.19	19.03
s^{out}	12.13	12.12	12.59	12.91	19.02	13.87	13.85	17.19	16.79	20.12	18.81	8.90	13.42
k	15.32	16.22	16.12	16.20	21.12	15.48	15.46	3.13	4.18	6.25	2.88	9.26	3.32
k^{in}	14.49	13.56	12.90	15.34	7.29	12.99	12.98	10.45	16.50	19.68	11.13	20.75	17.89
k^{out}	11.70	11.25	11.80	9.24	21.09	14.20	14.19	19.14	20.50	21.19	26.13	0.19	12.36
bt	13.82	13.86	13.93	14.86	8.99	15.47	15.45	6.10	9.32	9.16	3.41	19.99	2.20
λ	71.73	60.58	60.35	64.53	41.28	70.06	70.08	15.23	21.53	20.13	16.42	22.83	13.83

Table 79: Formation of the first three principal components for each of the seven interaction networks of Table 74 considering total, in- and out- degrees (k, k^{in}, k^{out}) and strengths (s, s^{in}, s^{out}), clustering coefficient cc, betweenness centrality bt and symmetry related metrics (asy, μ^{asy}, σ^{asy}, dis, μ^{dis} and σ^{dis} defined in Section 2.2.3). The characteristics found in email interaction networks are preserved: the first component is an average of degree and betweenness, the second component is mostly governed by symmetry related metrics, and clustering coefficient is mostly relevant for the third component. Standard deviation of asymmetry and disequilibrium metrics are again coupled to clustering coefficient in the third component. Important differences are: the first component is a less regular average of centrality measures and has a greater contribution of symmetry metrics; the first component of I5 is formed mostly from symmetry, not centrality, metrics.

	PC1							PC2						
	I1	I2	I3	I4	I5	TT1	TT2	I1	I2	I3	I4	I5	TT1	T
cc	3.46	4.19	2.44	0.36	2.18	1.28	1.17	3.06	1.61	1.23	1.19	2.57	3.03	2.
s	10.05	9.21	9.60	9.31	3.54	10.27	10.59	5.81	5.74	7.33	8.47	9.24	6.26	5.
s^{in}	9.57	8.03	9.21	8.74	0.78	7.75	7.99	4.63	0.59	1.27	6.69	2.77	5.38	5.
s^{out}	7.88	6.21	5.45	6.97	5.76	9.25	9.54	6.78	10.23	12.76	9.20	10.26	5.27	4.
k	10.44	10.02	9.88	10.39	5.80	10.80	11.05	4.62	5.13	5.66	5.54	14.08	4.48	4.
k^{in}	10.12	9.30	9.50	9.98	4.43	8.64	8.86	2.69	0.70	0.88	4.49	9.61	5.40	5.
k^{out}	7.27	5.29	4.43	5.43	9.11	10.10	10.33	8.36	12.52	13.63	5.65	11.61	3.38	3.
bt	9.62	7.97	7.53	8.93	2.25	10.47	10.78	3.77	8.42	9.14	6.95	8.12	5.60	5.
asy	5.42	7.05	7.97	8.48	15.47	6.16	5.79	14.17	12.88	11.78	11.02	4.67	12.48	13
μ^{asy}	5.48	6.99	7.99	8.47	15.44	6.18	5.80	14.12	13.04	11.78	11.01	4.72	12.46	13
σ^{asy}	6.53	7.39	7.63	7.15	2.37	5.59	5.48	1.69	3.80	1.75	8.46	7.49	5.94	5.
dis	5.02	6.67	7.78	8.08	15.41	5.98	5.59	14.12	13.41	11.92	11.53	4.80	12.45	13
μ^{dis}	5.33	7.01	7.24	6.92	14.34	5.49	5.14	13.33	10.15	9.47	8.02	5.05	11.86	12
σ^{dis}	3.82	4.68	3.34	0.81	3.12	2.03	1.88	2.85	1.77	1.39	1.77	5.00	6.01	5.
λ	46.11	43.48	44.29	46.95	30.34	44.12	43.52	26.42	24.97	24.76	19.99	23.91	25.98	26

C TABLES RELATING TEXT AND TOPOOGY IN EMAIL NETWORKS

Annex

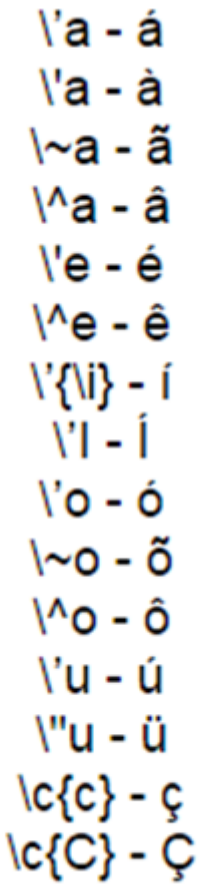
ANNEX A – EXEMPLO DE ANEXO

Elemento opcional, que consiste em um texto ou documento não elaborado pelo autor, que serve de fundamentação, comprovação e ilustração, conforme a ABNT NBR 14724.[?]

O **ANEXO B** exemplifica como incluir um anexo em pdf.

ANNEX B – ACENTUAÇÃO (MODO TEXTO - LaTeX)

Figure 9: Acentuação (modo texto - LaTeX)



Fonte: **?**

ANNEX C – SÍMBOLOS ÚTEIS EM LaTeX

Figure 10: Símbolos úteis em LaTeX

$\hbar$	=	$\hbar$
$\vec{k}$	=	$\vec{k}$
$\AA$	=	$\AA$
$\%$	=	$\%$
$\int$	=	$\int$
$\int_{a}^{b}$	=	$\int_{a}^{b}$
∂	=	∂
$\frac{\partial}{\partial x}$	=	$\frac{\partial}{\partial x}$
∇	=	∇
$\sum_{i=1}^{n}$	=	$\sum_{i=1}^{n}$
$\prod_{i=1}^{n} $	=	$\prod_{i=1}^{n}$

Fonte: ?

ANNEX D – LETRAS GREGAS EM LaTeX

Figure 11: Letras gregas em LaTeX

α	=	α			
β	=	β			
γ	=	γ	Γ	=	Γ
δ	=	δ	Δ	=	Δ
ϵ	=	ϵ			
ζ	=	ζ			
η	=	η			
θ	=	θ	Θ	=	Θ
ι	=	ι			
κ	=	κ			
λ	=	λ	Λ	=	Λ
μ	=	μ			
ν	=	ν			
ξ	=	ξ	Ξ	=	Ξ
o	=	o			
π	=	π	Π	=	Π
ρ	=	ρ			
σ	=	σ	Σ	=	Σ
τ	=	τ			
υ	=	υ	Υ	=	Υ
ϕ	=	ϕ	Φ	=	Φ
χ	=	χ			
ψ	=	ψ	Ψ	=	Ψ
ω	=	ω	Ω	=	Ω

Fonte: **?**