Stability in human interaction networks: sector relative sizes, prominence of topological measures and time activity statistics.

Renato Fabbri,[1, a)] Vilson V. da Silva Jr.,[b)] Ricardo Fabbri,[c)] Deborah C. Antunes,[d)] and Marilia M. Pisani[e)]
São Carlos Institute of Physics, University of São Paulo (IFSC/USP)

(Dated: 3 May 2015)

This article reports interaction networks stability by means of three quantitative criteria: activity distribution in time and among participants; a sound classification of vertices in peripheral, intermediary and hub sectors; the combination of basic measures into principal components with greater variance. We analyzed the temporal activity and topology evolution of networks in four email lists by considering window sizes from 50 to 10,000 messages, which were made to slide to generate snapshots of the network along a timeline. Activity in terms of seconds and minutes exhibit an uniform pattern, while hours, days and months exhibit stable concentrations. Participant activity follows the expected distribution of scale-free networks. We compare these networks to Erdös-Rényi networks in order to assign members to three distinct sectors, namely hubs, intermediary and periphery. Approximately 5% of the vertices are hubs, 15-35% are intermediary and the remainder belongs periphery. The metrics that most contribute to the dispersion of participants in the topological measures space were found to be centrality-related (degree, strength and betweenness), followed by symmetry-related, and then clustering coefficient. The results also embrace a sketch of a physics-based typology of agents. Because the network properties reported did not depend on the email list and were stable over time, and because observed structure is in accordance with expectations driven from literature for human interaction networks, we believe that the properties observed are also present in general human interaction networks. Current unfoldings include governance and accountability proposals and implementations, anthropological physics experiments, audiovisual reconnaissance, and the report of quantitative differences of textual production as connectivity of agents changes.

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

'The reason for the persistent plausibility of the typological approach, however, is not a static biological one, but just the opposite: dynamic and social. The fact that human society has been up to now divided into classes affects more than the external relations of men. The marks of social repression are left within the individual soul.' - *Adorno et al, 1969, p. 747*

I. INTRODUCTION

Studies on human interaction networks have started long before modern computers, dating back to the nineteenth century, while the foundation of social network analysis is generally attributed to the psychiatrist Jacob Moreno[1]. With the increasing availability of data related to human interactions, research on these networks has grown continuously. Contributions can now be found in a variety of fields in the literature, from social sciences and humanities[2] to computer science[3] and physics[4,5], given the multidisciplinary nature of the topic. One of the approaches from an exact science perspective is to represent interaction networks as complex networks [gmane,barabasiHumanDyn,newmanFriendship], with which several features of human interaction have been revealed. For example, the topology of human interaction networks exhibits a scale-free trace, 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 [6,7], but only to a limited extent, since research is normally focused on a particular metric or functionality, such as accessibility or community detection[8,9].

In this paper we analyze the evolution of human interaction networks, by considering interaction in email lists as their representative. Using a timeline of activity snapshots with a constant number of contiguous messages in email lists, we found a remarkable stability for several of the network properties. Because this stability was shared by all email lists, we advocate that some of the conclusions can be valid for more general classes of interaction networks. In particular, this allows us to discuss typolo-

[a)] http://ifsc.usp.br/~fabbri/; Electronic mail: fabbri@usp.br
[b)] http://automata.cc/; Electronic mail: vilson@void.cc; Also at IFSC-USP
[c)] http://www.lems.brown.edu/~rfabbri/; Electronic mail: rfabbri@iprj.uerj.br; Instituto Politécnico, Universidade Estadual do Rio de Janeiro (IPRJ)
[d)] http://lattes.cnpq.br/1065956470701739; Electronic mail: deborahantunes@gmail.com; Curso de Psicologia, Universidade Federal do Cerá (UFC)
[e)] http://lattes.cnpq.br/6738980149860322; Electronic mail: marilia.m.pisani@gmail.com; Centro de Cincias Naturais e Humanas, Universidade Federal do ABC (CCNH/UFABC)

gies in the context of such networks, in an attempt to bridge the gap between approaches based solely on data analysis (i.e. from a hard sciences perspective) and those relevant to the social sciences. This is important insofar as typologies are the canon of scientific literature for classification of human agents[10].

The paper is organized as follows. Section I A describes related work, while details of the data and methods of analysis are given in Section II and Section III. Section IV brings the results and discussion, leading to Section V for conclusions and further work.

A. Related work

Works on network evolution often consider solely network growth, in which there is a monotonic increase in the number of events considered[6]. Exceptions are reported in this section, with emphasis on those more closely related to the present article.

Network types have been reported relating number of participants, intermittence of them and network longevity[6]. Such results are confluent and endorsed by results of Section IV D. Two topologically different networks are reported to emerge from human interaction networks, depending on the frequency of interactions, which can either be a generalized power law or an exponential connectivity distribution[11]. In email list networks, scale-free properties were reported with $\alpha = 1$[3] (as are web browsing and library loans[4]), and different linguistic traces were related to weak and strong ties[12].

Unreciprocated edges often exceed 50% in the networks analyzed, which matches empirical evidence from the literature[7]. No correlation of topological characteristics and geographical coordinates was found[13], therefore geographical positions were not considered in our study. Gender related behavior in mobile phone datasets has been reported[14], but this was not considered in the present article because email messages and addresses have no gender related metadata[15].

II. DATA DESCRIPTION: EMAIL LISTS AND MESSAGES

Email list messages were obtained from the GMANE email archive[15], which consists of more than 20,000 email lists and more than 130,000,000 messages[16]. These lists cover a variety of topics, mostly technology-related. The archive can be described as a corpus with metadata of its messages, including sent time, place, sender name, and sender email address. The GMANE usage in scientific research is reported in studies of isolated lists and of lexical innovations[3,12].

We analyzed many email lists, but selected only four in order to make a thorough analysis, from which general properties can be inferred. These lists, selected as representing both a diverse set and ordinary lists, are:

TABLE I. 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 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. A relation holds for all lists carefully considered: as the number of participants increases, the number of threads decreases. This underpins a typology sketch of networks, as discussed in Section IV D.

list	$date_1$	$date_M$	N	Γ	$\overline{M}$
LAU	2003-06-29	2005-07-23	1181	3372	5
LAD	2003-06-30	2009-10-07	1268	3109	4
MET	2005-08-01	2008-03-07	492	4607	23
CPP	2002-03-12	2009-08-25	1052	4506	7

- Linux Audio Users list[17], with participants holding hybrid artistic and technological interests, from different countries. English is the language used the most. Abbreviated as LAU from now on.
- Linux Audio Developers list[18], with participants from different countries, and English is the language used the most. A more technical and less active version of LAU. Abbreviated LAD from now on.
- Development list for the standard C++ library[19], with computer programmers from different countries. English is the language used the most. Abbreviated as CPP from now on.
- List of the MetaReciclagem project[20], with Brazilian activists holding digital culture interests. Portuguese is the most used language, although Spanish and English are also incident. Abbreviated MET from now on.

The first 20,000 messages of each list were considered, with total timespan, authors, threads and missing messages indicated in Table I.

III. CHARACTERIZATION METHODS

The email lists and the networks generated from them were characterized by using five procedures, namely: 1) statistics of activity along time, from seconds to years; 2) sectioning of the networks in hubs, intermediary and peripheral vertices; 3) topological metrics and their dispersion; 4) iterative visualization and data mining; 5) typological speculation about networks and participants. Each of these procedures are described below.

A. Time activity statistics

Messages were counted along time with respect to seconds, minutes, hours, days of the week, days of the

month, and months of the year. This resulted in histograms from which patterns could be drawn. The ratio $\frac{b_h}{b_l}$ between the highest and lowest incidences on the histograms served as a hint of how the observed distribution is compared to a uniform distribution.

The average and the dispersion were taken using circular statistics, in which each *measurement* (data point) is represented as a complex number with modulus equal to one, $z = e^{i\theta} = \cos(\theta) + i\sin(\theta)$, where $\theta = measurement\frac{2\pi}{period}$. The moments m_n, lengths of moments R_n, mean angle θ_μ, and rescaled mean angle θ'_μ are defined as:

$$\begin{aligned} m_n &= \frac{1}{N}\sum_{i=1}^{N} z_i^n \\ R_n &= |m_n| \\ \theta_\mu &= Arg(m_1) \\ \theta'_\mu &= \frac{period}{2\pi}\theta_\mu \end{aligned} \tag{1}$$

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 $\delta(z)$:

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

B. Interaction networks

Interaction networks can be modeled both weighted or unweighted, both directed or undirected[3,21,22]. Networks in this article are directed and weighted, the more informative of trivial possibilities, i.e. we did not investigate directed unweighted, undirected weighted, and undirected unweighted representations of the interaction networks. The 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 $A \to B$. Inverting edge direction yields the status network: B read the message and considered what A wrote worth responding, giving status to A, thus $B \to A$. This article uses the information network as described above and depicted in Figure 1. Edges in both directions are allowed. Each time an interaction occurs, one is added to the edge weight. Self-loops were regarded as non-informative and discarded. These human social interaction networks are reported in the literature as exhibiting scale-free and small world properties, as expected for (some) social networks[1,3].

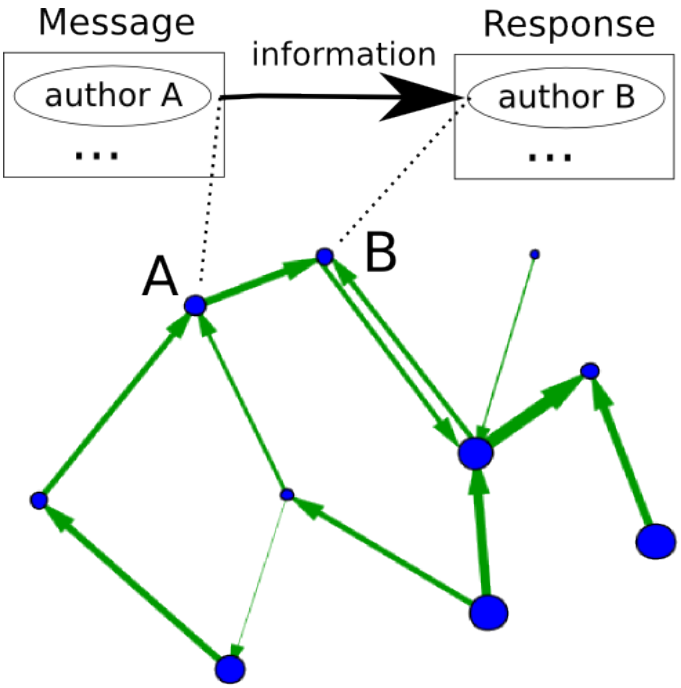

FIG. 1. Formation of interaction network from email messages. Each vertex represents a participant. A reply message from participant B to a message from participant A is regarded as evidence that B received information from A and yields a directed edge. Multiple messages add "weight" to a directed edge. Further details are given in Section III B.

Edges can be created from all antecedent message authors on the message-response thread to each message author. We only linked the immediate antecedent to the new message author, both for simplicity and for the valid objection that in adding two edges, $x \to y$ and $y \to z$, there is also a weaker connection between x and z. Potential interpretations for this weaker connection are: double length, half weight or with one more "obstacles". This suggests the adequacy of centrality measurements to account for the connectivity with all nodes, such as betweenness centrality and accessibility[8,23].

C. Erdös sectioning

In scale-free networks, the peripheral, intermediary and hubs sectors can be derived from a comparison with an Erdös-Rényi network with the same number of edges and vertices[24], as depicted in Figure 2. We shall refer to this procedure as *Erdös sectioning*, with the resulting sectors being referred to as *Erdös sectors* or *primitive sectors*.

The degree distribution $\widetilde{P}(k)$ of an ideal scale-free network $\mathcal{N}_f$ 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:

$$\widetilde{P}(k) < P(k) \Rightarrow \text{k is intermediary degree} \quad (3)$$

If $\mathcal{N}_f$ is directed and has no self-loops, the probability 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:

$$P(k) = \binom{2(N-1)}{k} p_e^k (1-p_e)^{2(N-1)-k} \quad (4)$$

The lower degree fat tail represents the border vertices, i.e. the peripheral sector or periphery where $\widetilde{P}(k) > P(k)$ and k is lower than any intermediary sector value of k. The higher degree fat tail is the hub sector, i.e. $\widetilde{P}(k) > P(k)$ and k is higher than any intermediary sector value of k. The reasoning for this classification is: 1) vertices so connected that they are virtually inexistent in networks connected at pure chance (e.g. without preferential attachment) are correctly associated to the hubs sector. Vertices with very few connections, which are way more abundant than expected by pure chance, are assigned to the periphery. Vertices with degree values predicted as the most abundant if connections are created by pure chance, near the average, and less frequent in scale-free phenomena, are classified as intermediary.

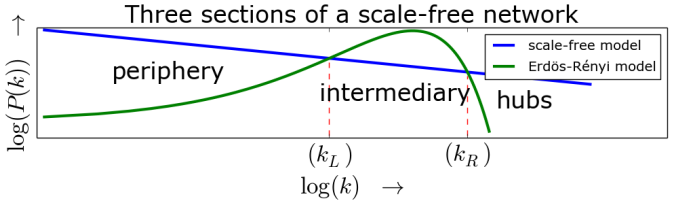

FIG. 2. Degree distribution of scale-free and Erdös-Rényi ideal networks. The latter has more intermediary vertices, while the former has more peripheral and hub vertices. The sector borders are defined by the two intersections $k_{\Leftarrow}$ and $k_{\Rightarrow}$ of the connectivity distributions. Characteristic degrees are in compact intervals of degree: $[0, k_{\Leftarrow}]$, $(k_{\Leftarrow}, k_{\Rightarrow}]$, $(k_{\Rightarrow}, k_{max}]$ for the Erdös sectors (periphery, intermediary and hubs).

To ensure statistical validity of the histograms, bins can be chosen to contain at least η vertices of the real network. Thus, each bin, starting at degree k_i, spans $\Delta_i = [k_i, k_j]$ degree values, where j is the smallest integer with which there are at least η vertices with degree larger than or equal k_i, and less than or equal k_j. This changes equation 3 to:

$$\sum_{x=k_i}^{k_j} \widetilde{P}(x) < \sum_{x=k_i}^{k_j} P(x) \Rightarrow \text{i is intermediary} \quad (5)$$

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

$$\hat{P}(k^{way}) = \binom{N-1}{k^{way}} p_e^k (1-p_e)^{N-1-k^{way}} \quad (6)$$

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 $1 \le \gamma \le 6$, $1 \le \phi \le 3$, and the basic six Erdös sectioning possibilities $\{E_\gamma\}$ have three Erdös sectors $E_\gamma = \{e_{\gamma,\phi}\}$ defined as:

$$\begin{aligned} e_{\gamma,1} &= \{\, i \mid \overline{k}_{\gamma,L} \ge \overline{k}_{\gamma,i}\} \\ e_{\gamma,2} &= \{\, i \mid \overline{k}_{\gamma,L} < \overline{k}_{\gamma,i} \le \overline{k}_{\gamma,R}\} \\ e_{\gamma,3} &= \{\, i \mid \quad\quad \overline{k}_{\gamma,i} < \overline{k}_{\gamma,R}\} \end{aligned} \quad (7)$$

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

$$\begin{aligned} \overline{k}_{1,i} &= k_i \\ \overline{k}_{2,i} &= k_i^{in} \\ \overline{k}_{3,i} &= k_i^{out} \\ \overline{k}_{4,i} &= \frac{s_i}{\overline{w}} \\ \overline{k}_{5,i} &= \frac{s_i^{in}}{\overline{w}} \\ \overline{k}_{6,i} &= \frac{s_i^{out}}{\overline{w}} \end{aligned} \quad (8)$$

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.

Since different metrics can be used to identify the three types of vertices, compound criteria can be defined. For example, a very stringent criterion can be used, according to which a vertex is only regarded as pertaining to a sector if it is so for all the metrics. 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 (usually) do not reach 100%, which is indicated by a black line in Figures 3.
- 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 total members may add more than 100%, which is indicated by a black line in Figure 3.

- 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 so according to any of the metrics. The remaining vertices are classified as intermediary if they belong to this category for any of the metrics. Peripheral vertices will then be those which were not classified as hub or intermediary with any of the metrics.
- Exclusivist externals C_5: vertices are only hubs if they are classified as such according to all the metrics. The remaining vertices are classified as peripheral if they fall into the periphery or hub classes by any metric. The rest of the nodes are classified as intermediary.
- Inclusivist externals C_6: hubs are vertices classified as hubs according to any metric. The remaining vertices will be peripheral if they are classified as such according to any metric. The rest of the vertices will be intermediary vertices.

Using equations 7, these compound criteria C_δ, with δ integer in the interval $1 < \delta < 6$ can be described as:

$$\begin{aligned}
C_1 &= \{c_{1,\phi} = \{i \mid i \in e_{\gamma,\phi},\ \forall\ \gamma\}\} \\
C_2 &= \{c_{2,\phi} = \{i \mid \exists\ \gamma : i \in e_{\gamma,\phi}\}\} \\
C_3 &= \{c_{3,\phi} = \{i \mid i \in e_{\gamma,\phi'},\ \forall\ \gamma,\ \forall\ \phi' \geq \phi\}\} \\
C_4 &= \{c_{4,\phi} = \{i \mid i \in e_{\gamma,\phi'},\ \forall\ \gamma,\ \forall\ \phi' \leq \phi\}\} \\
C_5 &= \{c_{5,\phi} = \{i \mid i \in e_{\gamma,\phi'},\ \forall\ \gamma, \\
&\qquad \forall\ (\phi'+1)\%4 \leq (\phi+1)\%4\}\} \\
C_6 &= \{c_{6,\phi} = \{i \mid i \in e_{\gamma,\phi'},\ \forall\ \gamma, \\
&\qquad \forall\ (\phi'+1)\%4 \geq (\phi+1)\%4\}\}
\end{aligned} \tag{9}$$

The simplification of all the compound possibilities to the small set listed above can be formalized in strict mathematical terms, but this was considered out of the scope for current interests. It is worth noting that the exclusivist cascade is the same sectioning of an inclusivist cascade from periphery to hubs, but with inverted order of sectors precedence. These compound criteria can be used to examine network sections in the case of a low number of messages, such as in the last figures of Support Information.

D. Topological metrics for Principal Component Analysis

The topology of the networks was studied using Principal Component Analysis (PCA[25]) with 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: 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. The standard clustering coefficient for undirected graphs was used.
- Betweenness centrality bt_i: fraction of geodesics that contain vertex i. The betweenness centrality index considered directions and weight, as specified in[26].

In order to capture symmetries in the activity of participants, the following metrics were introduced for a vertex i:

- Asymmetry: $asy_i = \frac{k_i^{in} - k_i^{out}}{k_i}$.
- Mean of asymmetry of edges: $\mu_i^{asy} = \frac{\sum_{j \in J_i} e_{ji} - e_{ij}}{|J_i| = k_i}$. Where e_{xy} is 1 if there is and edge from x to y, 0 otherwise. J_i is the set of neighbors of vertex i, and $|J_i| = k_i$ is the number of neighbors of vertex i.
- Standard deviation of asymmetry of edges: $\sigma_i^{asy} = \sqrt{\frac{\sum_{j \in J_i} [\mu_{asy} - (e_{ji} - e_{ij})]^2}{k_i}}$.
- Disequilibrium: $dis_i = \frac{s_i^{in} - s_i^{out}}{s_i}$.
- Mean of disequilibrium of edges: $\mu_i^{dis} = \frac{\sum_{j \in J_i} \frac{w_{ji} - w_{ij}}{s_i}}{k_i}$, 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: $\sigma_i^{dis} = \sqrt{\frac{\sum_{j \in J_i} [\mu_{dis} - \frac{(w_{ji} - w_{ij})}{s_i}]^2}{k_i}}$.

E. Evolution and visualization of the networks

The evolution of the networks was observed within a fixed number of messages, which we refer to as the window size ws. This same number of contiguous messages ws was considered with different shifts in the message timeline to obtain snapshots. Each snapshot was used both to perform the Erdös sectioning and apply PCA for the topological metrics. The ws used were 50, 100, 200, 400, 500, 800, 1000, 2000, 2500, 5000 and 10000. Within

a same ws, the number of vertices and edges vary in time, as do other network characteristics. Such changes could be visualized with the tools described below.

Networks were visualized with animations, image galleries and online gadgets developed specifically for this research[27–29]. Such visualizations were crucial to guide research into the most important features of network evolution, and prompted us to capture the prominence of topological metrics along time using mean and standard deviations. Furthermore, the size of three sectors could be visualized in a timeline fashion (Figure 3). Visualization of network structure was especially useful as part of the email lists data mining, from which parts of relevant structures and results were driven.

IV. RESULTS AND DISCUSSION

A. Activity along time

The activity along time, in terms of seconds, minutes, hours, days and months, is practically the same for all lists. Using circular statistics we calculated average values and their dispersion for activity in all time scales, and found that both average and dispersion were very similar in all the lists. Table II shows the .. We chose to give detailed values in Table II-VI because these numbers can actually be used for characterizing nodes (participants) in other networks, as they are independent of the network under analysis. For example, they may serve for identification of outliers in a community.

Values are exemplified with Tables and Support information. The pattern is the homogeneity for seconds and minutes. Hours of the day revealed an abrupt peak around 11am with most active period being the afternoon. Days of the week revealed a decrease of at least one third and at most two thirds of activity on weekends. Days of the month were regarded as homogeneous with an inconclusive slight tendency of 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.

Messages were slightly more evenly distributed in all lists than in simulations[30] using uniform distribution for both seconds and minutes: $\frac{max(incidence)}{min(incidence)} \in (1.26, 1.275]$. Simulations reach these values but have in average more discrepant higher and lower peaks $\xi = \frac{max(incidence')}{min(incidence')} \Rightarrow \mu_\xi = 1.2918$ and $\sigma_\xi = 0.04619$. Therefore, the incidence of messages at each second of a minute and at each minute of an hour was considered uniform, i.e. no trend was detected and the pattern is the uniformity. Circular dispersion is maximized and the mean has little meaning as exposed in Table II. Patterns in each of the other time scales are detailed in histogram Tables III-VI and further exemplified in Supporting Information.

TABLE II. The rescaled circular mean θ'_μ and the circular dispersion $\delta(z)$ described in Section III A. This typical table was made using all LAD list messages, and the results are the same for other lists, as exposed in Supporting Information. Most uniform distribution of activity was found in seconds and minutes, where the mean has little meaning. Hours of the day exhibited the most concentrated activity (lowest $\delta(z)$), with mean between 14h and 15h ($\theta' = -9.61$). Weekdays, month days and months have mean near zero (i.e. near the beginning of the week, month and year) and high dispersion.

	θ'_μ	$\delta(z)$
seconds	–//–	9070.17
minutes	–//–	205489.40
hours	-9.61	4.36
weekdays	-0.03	29.28
month days	-0.07	2754.16
months	-0.56	44.00

TABLE III. Example of activity percentages 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. Nevertheless, the activity peak occurs around midday, with a slight skew toward one hour before noon.

<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="2">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="2">2.47</td></tr>
<tr><td>4h</td><td>0.68</td><td rowspan="2">1.37</td><td rowspan="3">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="2">4.35</td></tr>
<tr><td>7h</td><td>1.24</td><td rowspan="6">23.09</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="2">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="2">18.95</td><td rowspan="2">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><td rowspan="3">18.68</td></tr>
<tr><td>15h</td><td>6.10</td></tr>
<tr><td>16h</td><td>6.22</td><td rowspan="2">12.58</td><td rowspan="3">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></tr>
<tr><td>19h</td><td>5.02</td><td rowspan="6">28.61</td></tr>
<tr><td>20h</td><td>4.85</td><td rowspan="2">9.23</td></tr>
<tr><td>21h</td><td>4.38</td><td rowspan="2">12.73</td><td rowspan="3">17.59</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. Scalable fat-tail structure: constancy of membership fractions in the Erdös sectors

There is a concentration of hub activity and an abundance of vertex with few connections. Table VII shows this expected distribution of activity among participants in a scale-free context.

TABLE IV. Example of activity percentages along the days of the week. Higher activity was observed during weekdays, with a decrease of activity on weekends of at least one third and two thirds in extreme cases.

	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

TABLE V. No significant variation of activity in the days along the month was observed. One cannot point much more than a - probably not statistically relevant - tendency of first and second weeks to be more active. The most important trait seems to be homogeneity.

	1 day	5	10	15 days
1	5.00	18.21	33.84	51.29
2	3.02			
3	3.65			
4	3.18			
5	3.37			
6	3.22	15.63		
7	3.54			
8	2.59			
9	2.87			
10	3.41			
11	3.42	17.46	33.00	
12	3.50			
13	3.30			
14	3.48			
15	3.76			
16	3.37	15.54		48.71
17	3.45			
18	3.01			
19	2.97			
20	2.75			
21	3.26	16.97	33.17	
22	3.70			
23	3.20			
24	3.22			
25	3.59			
26	3.51	16.20		
27	3.12			
28	3.26			
29	3.70			
30	2.62			

The distribution of vertices in the hubs, intermediary, periphery Erdös sectors defined in Section III C is remarkably stable along time, provided that a sufficiently large sample of 200 or more messages is considered. Moreover, the same distribution applies to the networks of all the four email lists, as demonstrated in Figure 3 and the Support Information. Typically, $\approx [3-12]\%$ of the vertices are found to be hubs, $\approx [15-44]\%$ are intermediary and $\approx [44-82]\%$ are peripheral, which is consistent with the literature[31]. These results hold for the total, in and out degrees and strengths. Stable distributions can also be obtained for 100 or less messages if

TABLE VI. Example of activity percentages of the months along the year from LAD list messages. Activity is concentrated in Jun-Aug for MET and LAD, and in Dec-Mar for CPP, LAU and LAD (see Support Information). These observations fit academic calendars, vacations and end-of-year holidays.

	m.	b.	t.	q.	s.
Jan	11.24	18.51	26.46	36.07	57.96
Fev	7.26				
Mar	7.95	17.56			
Apr	9.61		31.50		
Mai	8.94	21.89		37.56	
Jun	12.95				
Jul	9.03	15.67	22.30		42.04
Ago	6.64				
Set	6.63	12.38		26.37	
Out	5.75		19.74		
Nov	7.61	13.99			
Dez	6.38				

TABLE VII. Distribution of activity among participants. The first column presents the percentage of messages sent by the most active participant. The column for the first quartile ($1Q$) shows the minimum percentage of participants responsible for at least 25% of total messages. Similarly, the column for the first three quartiles $1-3Q$ gives the minimum percentage of participants responsible for 75% of total messages. The last decile $-10D$ column brings the maximum percentage of participants responsible for 10% of messages.

list	hub	$1Q$	$1-3Q$	$-10D$
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%)

classification of the three sectors is performed with one of the compound criteria established in Section III C. The networks hold their basic structure with as few as 10-50 messages; concentration of activity and the abundance of low-activity participants take place even with very few messages, which is highlighted in the last figures of the Support Information. A minimum window size for observation of more general properties might be inferred by monitoring the giant component and the degeneration of the Erdös sectors.

C. Stability of principal components and the prevalence of symmetry over clusterization for dispersion

The topology was analyzed using standard, well-established metrics of centrality and clustering. We also introduced symmetry metrics because of evidence of their importance in social contexts[7]. The contribution of each metric to the variance is very similar for all the networks,

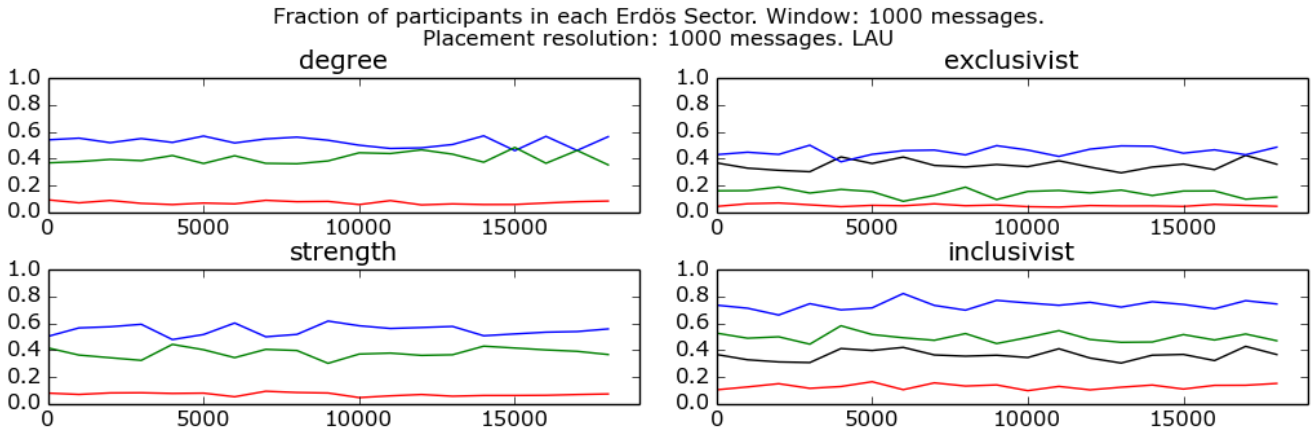

FIG. 3. Fraction of agents in each Erdös sector. Hubs, intermediary and periphery fractions are represented in red, green and blue. For this figure, we used two simple criteria, namely degree and strength, for the graphics on the left. For the graphs on the right we employed the Exclusivist and inclusivist compound criteria, with black lines representing the fraction of vertices without class and with more than one class, respectively. See Support Information for a collection of such timeline figures with all simple and compound criteria and metrics.

and did not vary with time. In applying PCA to the snapshots, the contribution of each metric to the principal components resulted in very small standard deviation. Table VIII exemplifies the principal components formation with all the metrics considered for the MET email list. Similar results are presented in the Support Information for the other lists, and considering only a few metrics.

TABLE VIII. Loadings for the 14 metrics into the principal components for the MET list, $ws = 1000$ messages in 20 disjoint positioning. 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 80.36% of the variance.

	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

The first principal component is an average of centrality metrics: degrees, strengths and betweenness centrality. Therefore, all of these centrality measurements are equally important for characterizing the networks. On one hand, the relevance of all centrality metrics is not surprising since they may be highly correlated. The degree and strength, for instance, are highly correlated, with Spearman correlation coefficient $\in [0.95, 1]$ and Pearson coefficient $\in [0.85, 1)$ for $ws > 1000$. On the other hand, each measure relates to a different participation characteristic, and their equal relevance is noticeable. The clustering coefficient is presented in almost perfect orthogonality to centrality measures.

Dispersion was more prevalent in symmetry-related metrics than the clustering coefficient. These relations are presented in Figure 4 and Table VIII. Plots of vertices in the components exhibited in Table VIII are shown in Figure 4, where each vertex is colored according to the sector they belong to. As expected, peripheral vertices have very low values in the first component (centrality related) and greater dispersion in the third component (clustering related). The PCA plot in the third system of Figure 4, where all metrics are considered, reflects symmetry metrics relevance for the variance. We concluded that the symmetry-related measurements can be more meaningful in characterizing interaction networks (and their participants) than the clustering coefficient, especially for hubs and intermediary vertices.

D. Primitive types from Erdös sectors

Let a sector to which a vertex belongs yield a type for the corresponding participant. Accordingly, a participant may be considered peripheral, intermediary or a hub. This trivial consideration raises important ques-

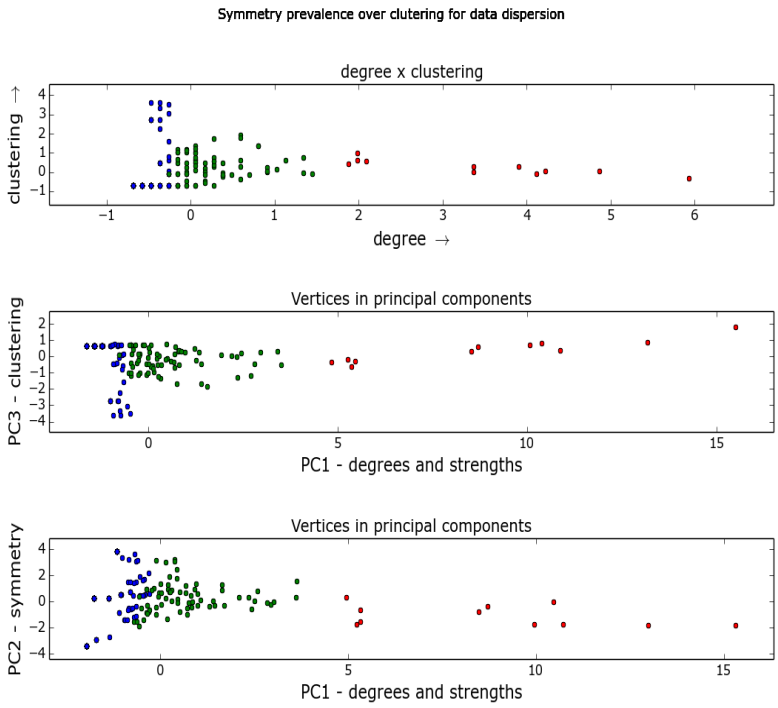

FIG. 4. The first plot shows degree versus clustering coefficient. In the second plots, the first component is an average of centrality metrics, but the second component is still clustering coefficient. The third plot exhibits the greater dispersion in the symmetry-related second component. In this case, clustering coefficient is only relevant for the third component. Greater dispersion suggests that symmetry-related metrics are more powerful for characterizing interaction networks than clustering coefficient, especially for hubs and intermediary vertices. This general layout is recurrent and consistent with the literature: most connected vertices have low clusterization while higher clusterization is gradually more incident as the number of connections is lowered. This prevalence of the symmetry-related measures over clustering coefficient for the dispersion of participants in the space of topological measures is an important result from the PCA analysis, together with temporal stability of component formation, and can be observed in Table VIII. This figure was done with a snapshot of the LAU list in a window size of $ws = 1000$ messages. Similar structures were observed in all window sizes $ws \in [500, 10000]$ and for networks of other email lists, which points to a common relationship between the measures of degree, strength and betweenness centrality, the measures of symmetry and the measure of clustering coefficient described in Section III D.

tions: is this typology stigmatizing or prejudice inclined in any way? What else do we know about each type and about the typology itself? These general considerations about this primitive typology will help us answer the first questions:

- A participant belongs to many networks (his family, family of a friend, the email list to which he belongs, work friends, etc.).
- A participant might belong to all three sectors at the same time. Actually, it is reasonable to assume that almost every person belongs to all three sectors for some of the networks.
- The participant often transitions from one sector to another within a network.
- Reported stability of network structure is concomitant with the continuous change of participants and the type of each participant.

From these, we can assume that the typology is not stigmatizing because the type of an individual changes constantly[32]. That is to say, an individual is a hub in a number of network and peripheral in other networks, and even within a network he/she probably changes type along time.

Our approach to the second question was the inspection of network evolution from the raw data and using the tools for visualization[33,34]. This allowed us to infer relevant characteristics of each type, summarized as follows.

- Core hubs usually have intermittent activity. Very stable activity was found on MET hubs, which motivated its integration to this work. There are reports in the literature of greater stability of participation in smaller communities[6], which is the reason why the smaller number of participants in MET was considered coherent with the stable activity of hubs.

- Typically, the activity of hubs is trivial: they interact as much as possible, in every occasion with everyone. The activity of peripheral vertices also follows a simple pattern: they interact very rarely, in very few occasions. Therefore, intermediary vertices seem responsible for the network structure. For example, intermediary vertices may exhibit preferential communication to peripheral, intermediary, or hub vertices; can be marked by stable communication partners; can involve stable or intermittent patterns of activity.

- Some of the most active participants receive many responses with relative few messages sent, and rarely are top hubs. These seem as authorities and contrast with participants that respond much more than receive responses.

- The most obvious community structure, as observed by a high clustering coefficient, is found only in peripheral and intermediary sectors.

This typology is a bridge from exact to human sciences and may be enriched with concepts from other typologies, such as Meyer-Briggs, Pavlov or the authoritarian types of the F-Scale[32].

With regard to the networks as the whole objects of analysis, we observed that there are two contrasting types of network: i) those with few and stable agents, where the number of threads per message is large; ii) others with many agents displaying intermittent activity, for which the number of threads per message is considerably lower than for the other type of network. This result is exemplified in Table I and is also consistent with the literature[6], which reports that the smaller size of communities is responsible for the stronger hubs observed.

V. CONCLUSIONS AND FURTHER WORK

The characterization of interaction networks resulted from stability observations. Along temporal activity statistics, this work reports the stability of the formation of principal components and of the relative sizes of periphery, intermediary and hubs Erdös sectors. These results suggests and eases the observance of types in both agents and the networks.

Observed systems were coherent with literature in different aspects, such as concentration of activity or such as the clusterization versus connectivity pattern. Even so, analysis of data from other virtual environments, such as Facebook, Twitter and LinkedIn, might help understanding how general are these structures and what are proper uses.

Using the results here presented, significant differences were found in the textual production of each Edös sector[33]. Another article describes the network time evolution visualization method used here[34]. Implementations and consequences of these results are taking place in software packages, linked data, electronic government technologies, and anthropological physics experiments[35–39].

ACKNOWLEDGMENTS

Renato Fabbri is grateful to CNPq (process: 140860/2013-4, project 870336/1997-5), United Nations Development Program (PNUD/ONU, contract: 2013/000566; project BRA/12/018) and the Postgraduate Committee of the IFSC/USP. This author is also grateful for the American Jewish Committee for maintaining an online copy of the Adorno book used on the epigraph[32]. Authors thank Gmane creators and maintainers. Authors thank referred email lists communities and welcome feedback as core contribution to this, and similar, research. Finally, authors thank the developers and users of Python scientific tools.

Appendix A: Data and scripts

Messages were downloaded from the Gmane public database[16]. All routines necessary to achieve the results reported in this article, including tables and figures of Supporting information, are available through a public domain Python package and an open Git repository[35].

[1] M. Newman, *Networks: an introduction* (Oxford University Press, 2010).
[2] B. Latour, "Reassembling the social. an introduction to actor-network-theory," Journal of Economic Sociology **14**, 73–87 (2013).
[3] C. Bird, A. Gourley, P. Devanbu, M. Gertz, and A. Swaminathan, "Mining email social networks," in *Proceedings of the 2006 international workshop on Mining software repositories* (ACM, 2006) pp. 137–143.
[4] A. Vázquez, J. G. Oliveira, Z. Dezsö, K.-I. Goh, I. Kondor, and A.-L. Barabási, "Modeling bursts and heavy tails in human dynamics," Physical Review E **73**, 036127 (2006).
[5] B. Ball and M. E. Newman, "Friendship networks and social status," arXiv preprint arXiv:1205.6822 (2012).
[6] G. Palla, A.-L. Barabási, and T. Vicsek, "Quantifying social group evolution," Nature **446**, 664–667 (2007).
[7] E. A. Leicht, G. Clarkson, K. Shedden, and M. E. Newman, "Large-scale structure of time evolving citation networks," The European Physical Journal B **59**, 75–83 (2007).
[8] B. Travençolo and L. d. F. Costa, "Accessibility in complex networks," Physics Letters A **373**, 89–95 (2008).
[9] M. E. Newman, "Modularity and community structure in networks," Proceedings of the National Academy of Sciences **103**, 8577–8582 (2006).
[10] K. Gergen and M. Gergen, *Historical social psychology* (Psychology Press, 2014).
[11] R. Albert and A.-L. Barabási, "Topology of evolving networks: local events and universality," Physical review letters **85**, 5234 (2000).
[12] K. Marek-Spartz, P. Chesley, and H. Sande, "Construction of the gmane corpus for examining the diffusion of lexical innovations," (2012).
[13] J.-P. Onnela, S. Arbesman, M. C. González, A.-L. Barabási, and N. A. Christakis, "Geographic constraints on social network groups," PLoS one **6**, e16939 (2011).

[14]V. Palchykov, K. Kaski, J. Kertész, A.-L. Barabási, and R. I. Dunbar, "Sex differences in intimate relationships," Scientific reports **2** (2012).
[15]L. M. Ingebrigtsen, "Gmane," (2008).
[16]Wikipedia, "Gmane — Wikipedia, the free encyclopedi," (2013), online; accessed 27-October-2013.
[17]Gmane.linux.audio.users is list ID in GMANE.
[18]Gmane.linux.audio.devel is list ID in GMANE.
[19]Gmane.comp.gcc.libstdc++.devel is list ID in GMANE.
[20]Gmane.politics.organizations.metareciclagem is list ID in GMANE.
[21]E. A. Leicht and M. E. Newman, "Community structure in directed networks," Physical review letters **100**, 118703 (2008).
[22]M. Newman, "Community detection and graph partitioning," arXiv preprint arXiv:1305.4974 (2013).
[23]L. d. F. Costa, F. A. Rodrigues, G. Travieso, and P. Villas Boas, "Characterization of complex networks: A survey of measurements," Advances in Physics **56**, 167–242 (2007).
[24]M. O. Jackson, "Social and economic networks: Models and analysis," (2013), https://class.coursera.org/networksonline-001.
[25]I. Jolliffe, *Principal component analysis* (Wiley Online Library, 2005).
[26]U. Brandes, "A faster algorithm for betweenness centrality*," Journal of Mathematical Sociology **25**, 163–177 (2001).
[27]R. Fabbri, L. d. F. Costa, and O. N. d. Oliveira jr, "Video visualizations of email interaction network evolution," (2013), http://www.youtube.com/watch?v=-t5jxQ8cKxM&list=PLf_EtaMqu3jU-1j4jiIUiyMqyVSzIYeh6.
[28]R. Fabbri, L. d. F. Costa, and O. N. d. Oliveira jr, "Image gallery of email interaction networks." (2013), http://hera.ethymos.com.br:1080/redes/python/autoRede/gmane.linux.audio.devel_3000-4200-280/.
[29]R. Fabbri, L. d. F. Costa, and O. N. d. Oliveira jr, "Online gadget for making email interaction network images, gml files and measurements." (2013), http://hera.ethymos.com.br:1080/redes/python/autoRede/escolheRedes.php.
[30]Numpy version 1.6.1, "random.randint" function, was used for simulations, algorithms in https://pypi.python.org/pypi/gmane.
[31]S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D.-U. Hwang, "Complex networks: Structure and dynamics," Physics reports **424**, 175–308 (2006).
[32]T. W. Adorno, E. Frenkel-Brunswik, D. J. Levinson, and R. N. Sanford, "The authoritarian personality." (1950).
[33]R. Fabbri, "A connective differentiation of textual production in interaction networks," (2013), http://arxiv.org/abs/1412.7309.
[34]R. Fabbri, "Versinus: a visualization method for graphs in evolution," arXiv preprint arXiv:1412.7311 (2014).
[35]R. Fabbri, "Python package to analyze the gmane database," (2015), https://pypi.python.org/pypi/gmane.
[36]R. Fabbri, R. B. de Luna, R. A. P. Martins, *et al.*, "Social participation ontology: community documentation, enhancements and use examples," arXiv preprint arXiv:1501.02662 (2015).
[37]*Produto 5 da consultoria PNUD/ONU de Renato Fabbri*, https://github.com/ttm/pnud4/blob/master/latex/produto.pdf?raw=true.
[38]R. Fabbri, "Ensaio sobre o auto-aproveitamento: um relato de investidas naturais na participa\ c {c}\~ ao social," arXiv preprint arXiv:1412.6868 (2014).
[39]R. Fabbri, "What are you and i? [anthropological physics fundamentals]," academia.edu (2015), https://www.academia.edu/10356773/What_are_you_and_I_anthropological_physics_fundamentals_.