

# Time stability in human interaction networks

Renato Fabbri<sup>1, a)</sup>*São Carlos Institute of Physics, University of São Paulo (IFSC/USP)*

(Dated: 3 June 2015)

In this study, we demonstrate a remarkably stable activity in interaction networks. The activity along time and topology evolution were investigated in four e-mail lists by considering window sizes from 50 to 10,000 messages, which were made to slide and generate snapshots of the network in a timeline. Furthermore, the activity in terms of seconds, minutes, hours, days and months, is practically the same for all lists. The activity of participants followed the expected scale-free behavior, thus allowing us to establish three classes of vertices by comparing with the Erdős-Rényi model, namely hubs, intermediary and peripheral vertices. The relative size of these three sectors did not vary with time and was essentially the same for all e-mail lists. Typically, 3-12% of the vertices are hubs, 15-45% are intermediary and the remainder are peripheral vertices. The metrics that contribute most to the dispersion of participants in the topological measures space were centrality measurements (degree, strength and betweenness), followed by symmetry-related metrics and then clustering coefficient. Similar results for the distribution of participants in the three categories and for the relative importance of the topological metrics were obtained for 12 additional networks from Facebook, Twitter and Participa.br. Consistent with expectations driven from the literature, these properties may be general for human interaction networks, which has important implications in establishing a typology based on objective, quantitative criteria.

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.’ - 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 in mid twentieth century<sup>1</sup>. 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, from social sciences and humanities<sup>2</sup> to computer science<sup>3</sup> and physics<sup>4,5</sup>, given the multidisciplinary nature of the topic. One of the approaches from an exact science perspective is to represent interaction networks as complex networks<sup>4-6</sup>, 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<sup>7,8</sup>, but only to a limited extent, since research is normally focused

on a particular metric or task, such as accessibility or community detection<sup>9,10</sup>.

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, is coherent with literature, and has been observed in some networks from Twitter, Facebook and Participa.br, we advocate that some of the conclusions might be valid for more general classes of interaction networks. In particular, this allows us to discuss typologies 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<sup>11</sup>.

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. Appendix has directions to access data and computer scripts to reproduce the analysis. Support Information document has a throughout exposition of measurements in networks from four email lists and a brief analysis of networks from Twitter, Facebook and Participa.br.

<sup>a)</sup> <http://ifsc.usp.br/~fabbri/>; Electronic mail: [fabbri@usp.br](mailto:fabbri@usp.br)

### A. Related work

Research on network evolution often considers solely network growth, in which there is a monotonic increase in the number of events considered<sup>7</sup>. Exceptions are reported in this section, with emphasis on those more closely related to the present article.

Network types have been discussed with regard to the number of participants, intermittence of their activity and network longevity<sup>7</sup>. Two topologically different networks emerged from human interaction networks, depending on the frequency of interactions, which can either be a generalized power law or an exponential connectivity distribution<sup>12</sup>. In email list networks, scale-free properties were reported with  $\alpha = 1$ <sup>3</sup> (as are web browsing and library loans<sup>4</sup>), and different linguistic traces were related to weak and strong ties<sup>13</sup>.

Unreciprocated edges often exceed 50% in the networks analyzed, which matches empirical evidence from the literature<sup>8</sup> and motivated the inclusion of symmetry metrics on our analysis. No correlation of topological characteristics and geographical coordinates was found<sup>14</sup>, therefore geographical positions were not considered in our study. Gender related behavior in mobile phone datasets was indeed reported<sup>15</sup>, but this was not considered in the present article because email messages and addresses have no gender related metadata<sup>16</sup>.

## II. DATA DESCRIPTION: EMAIL LISTS AND MESSAGES

Email list messages were obtained from the GMANE email archive<sup>16</sup>, which consists of more than 20,000 email lists and more than 130,000,000 messages<sup>17</sup>. 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<sup>3,13</sup>.

We analyzed many email lists (and data from Twitter, Facebook and Participa.br) 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:

- Linux Audio Users list<sup>18</sup>, 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<sup>19</sup>, 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.

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).  $\Gamma$  is the number of threads (count of messages without antecedent).  $\bar{M}$  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: within a same number of messages, as the number of participants increases, the number of threads decreases.

list	$date_1$	$date_M$	$N$	$\Gamma$	$\bar{M}$
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

- Development list for the standard C++ library<sup>20</sup>, with computer programmers from different countries. English is the language used the most. Abbreviated as CPP from now on.
- List of the MetaReciclagem project<sup>21</sup>, 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. In subsidiary experiments we considered 140 additional email lists, also retrieved from the Gmane public database, to analyze the interdependence between the number of participants and the number of threads. Furthermore, we used 12 additional networks from Facebook (8), Twitter (2) and Participa.br (2) to grasp the generality of the results driven from email lists.

## III. CHARACTERIZATION METHODS

The email lists and the networks generated from them were characterized using five procedures, namely: 1) statistics of activity along time, in scales from seconds to years; 2) sectioning of the networks in hubs, intermediary and peripheral vertices; 3) dispersion of basic topological metrics; 4) iterative visualization and data inspection. 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 close the observed distribution is 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  $\theta = \text{measurement} \frac{2\pi}{\text{period}}$ . The moments  $m_n$ , lengths of moments  $R_n$ , mean angle  $\theta_\mu$ , and rescaled mean angle  $\theta'_\mu$  are defined as:

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

$\theta'_\mu$  is used as the measure of location. Dispersion is measured using the circular variance  $\text{Var}(z)$ , the circular standard deviation  $S(z)$ , and the circular dispersion  $\delta(z)$ :

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

As expected, a positive correlation was found in all  $\text{Var}(z)$ ,  $S(z)$  and  $\delta(z)$  dispersion measures (as can be noticed in Supporting Information) and  $\delta(z)$  was preferred in the discussion of results.

## B. Interaction networks

Interaction networks can be modeled both weighted or unweighted, both directed or undirected<sup>3,22,23</sup>. 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  $A \rightarrow B$ . Inverting edge direction yields the status network: B read the message and considered what A wrote worth responding, giving status to A, thus  $B \rightarrow A$ . 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. Selfloops 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<sup>1,3</sup>.

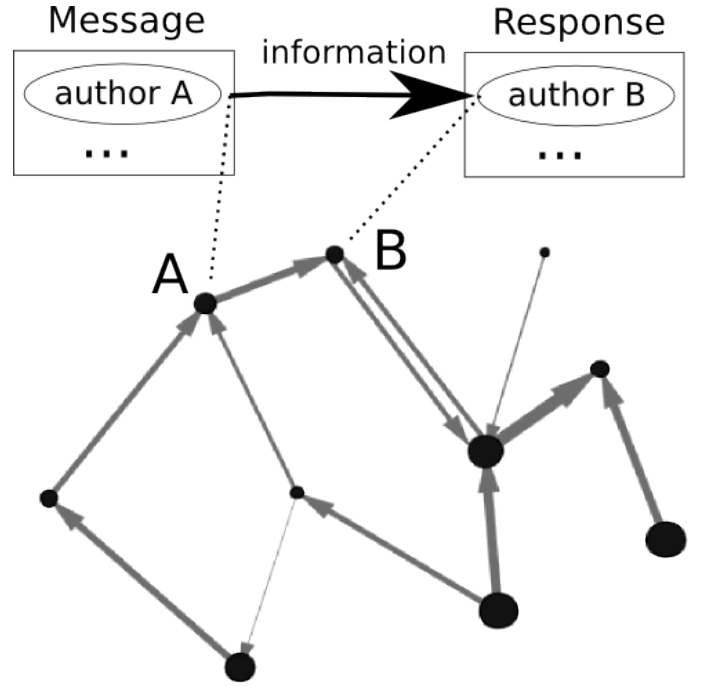


FIG. 1. Formation of interaction network from email messages. Each vertex represents a participant. A reply message from participant B to a message from participant A is regarded as evidence that B received information from A and yields a directed edge. Multiple messages add “weight” to a directed edge. 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,  $x \rightarrow y$  and  $y \rightarrow z$ , 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, eigenvector centrality and accessibility<sup>9,24</sup>.

## C. Erdős sectioning

In scale-free networks, the peripheral, intermediary and hubs sectors can be derived from a comparison with an Erdős-Rényi network with the same number of edges and vertices<sup>25</sup>, as depicted in Figure 2. We shall refer to this procedure as *Erdős sectioning*, with the resulting sectors being referred to as *Erdős sectors* or *primitive sectors*.

The degree distribution  $\tilde{P}(k)$  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 Erdős-Rényi network with the same number of vertices and edges. In-

deed, 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 Erdős-Rényi model:

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

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 Erdős-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  $\tilde{P}(k) > P(k)$  and  $k$  is lower than any intermediary sector value of  $k$ . The higher degree fat tail is the hub sector, i.e.  $\tilde{P}(k) > P(k)$  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 the real network, are classified as intermediary.

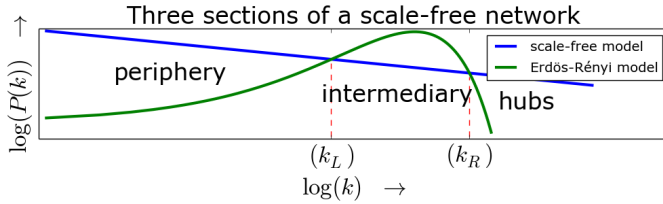


FIG. 2. Degree distribution of scale-free and Erdős-Rényi ideal networks. The latter has more intermediary vertices, while the former has more peripheral and hub vertices. The sector borders are defined by the two intersections  $k_L$  and  $k_R$  of the connectivity distributions. Characteristic degrees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

i To ensure statistical validity of the histograms, bins can be chosen to contain at least  $\eta$  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  $\eta$  vertices with degree larger than or equal  $k_i$ , and less than or equal  $k_j$ . This changes equation 3 to:

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

If strength  $s$  is used for comparison,  $P$  remains the same, but  $P(\kappa_i)$  with  $\kappa_i = \frac{s_i}{\bar{w}}$  should be used for comparison, with  $\bar{w} = 2 \frac{z}{\sum_i s_i}$  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  $\bar{w}$  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  $\gamma$  and  $\phi$  be integers in the intervals  $1 \leq \gamma \leq 6$ ,  $1 \leq \phi \leq 3$ , and each of the basic six Erdős sectioning possibilities  $\{E_\gamma\}$  have three Erdős sectors  $E_\gamma = \{e_{\gamma,\phi}\}$  defined as:

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

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

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

and both  $\bar{k}_{\gamma,L}$  and  $\bar{k}_{\gamma,R}$  are found using  $P(\bar{k})$  or  $\hat{P}(\bar{k})$  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 Figure 3.

- Inclusive 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.
- Inclusive 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.
- Inclusive 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  $C_\delta$ , with  $\delta$  integer in the interval  $1 < \delta < 6$  can be described as:

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

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 inclusive cascade from periphery to hubs, but with inverted order of sectors precedence. These compound criteria can be used to examine network sections considering all degrees and strengths and are specially useful 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<sup>26</sup>) 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<sup>27</sup>.

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 an 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  $x \rightarrow y$  and zero if there is no such edge.
- Standard deviation of disequilibrium of edges:  $\sigma_i^{dis} = \sqrt{\frac{\sum_{j \in J_i} [\mu_{dis} - \frac{(w_{ji} - w_{ij})}{s_i}]^2}{k_i}}$ .



TABLE II. The rescaled circular mean  $\theta'_\mu$  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 shown in the 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.

	$\theta'_\mu$	$\delta(z)$
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

### 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 Erdős sectioning and to 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, which is exhibited in the figures of Support Information.

Networks were visualized with animations, image galleries and online gadgets developed specifically for this research<sup>28–30</sup>. Such visualizations were crucial to guide research into the most important features of network evolution. Furthermore, the size of the three Erdős sectors could be visualized in a timeline fashion. Visualization of network structure was especially useful in the inspection of data and derived structures from the email lists.

## 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. Histograms in each time scale were calculated as were circular average values and their dispersion. We chose to give detailed values in Table II–VI because these numbers can actually be used for characterizing nodes (participants) in other networks, and networks themselves, as they are independent of the network under analysis. For example, they may serve for identification of outliers in a community.

TABLE III. Activity percentages along the hours of the day for the CPP list. Nearly identical distributions are found on other lists as shown in the Supporting Information. 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.

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

TABLE IV. Activity percentages along the days of the week for the four email lists. 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

In the scale of seconds and minutes, activity obeys a homogeneous pattern, with the messages being slightly more evenly distributed in all lists than in simulations using uniform distribution<sup>31</sup>. In the networks,  $\frac{\max(\text{incidence})}{\min(\text{incidence})} \in (1.26, 1.275]$  while simulations reach these values but have in average more discrepant higher and lower peaks  $\xi = \frac{\max(\text{incidence}')}{\min(\text{incidence}')} \Rightarrow \mu_\xi = 1.2918$  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. Circular dispersion is maximized and the mean has little meaning as indicated in Table II. As for the hours of the day, an abrupt peak appeared around 11am

with the 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 the first week being 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 III-VI.

### B. Scalable fat-tail structure: constancy of membership fractions in each Erdős sector

The distribution of vertices in the hubs, intermediary, periphery Erdős sectors 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 Supporting Information. Activity is highly concentrated on the hubs, while a very large number of peripheral vertices contribute to

TABLE V. Activity in the days along the month for MET list. Nearly identical distributions are found on other lists as indicated in 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 II.

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

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

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

only a fraction of the activity. This is expected for a system with a scale-free trace, as confirmed with the data in Table VII for the distribution of activity among participants.

Typically,  $\approx [3 - 12]\%$  of the vertices are found to be hubs,  $\approx [15 - 45]\%$  are intermediary and  $\approx [44 - 81]\%$  are peripheral, which is consistent with the literature<sup>32</sup>. These results hold for the total, in and out degrees and strengths. Stable distributions can also be obtained for 100 or less messages if classification of the three sectors is performed with one of the compound criteria established in Section IIIC. 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 Supporting Information. A minimum window size for observation of more general properties might be inferred by monitoring the giant component and the degeneration of the Erdős sectors.

In order to verify the possible generality of these findings, we obtained the Erdős sectors of 12 networks from Facebook, Twitter and Participa.br. The results are given in Tables XXX in the Supporting Information, which indicate that the percentages of hubs, intermediary and peripheral nodes are essentially the same as for the email lists.

### 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<sup>8</sup>. The contribution of each metric to the variance is very similar for all the networks, and did not vary with time. In applying PCA to the snapshots, the contribution of each metric to the prin-



FIG. 3. Fractions of agents in each Erdős sector, where the fractions for hubs, intermediary and peripheral vertices are represented in red, green and blue, respectively. We used two simple criteria, namely degree and strength, for the graphics on the left. For the graphics 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 Supporting Information for a collection of such timeline figures with all simple and compound criteria and metrics.

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%)

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 Supporting Information for the other lists, and considering only a few metrics.

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 of these metrics is related to a different participation characteristic, and their equal relevance is noticeable. The clustering coefficient is presented in almost

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% of the variance.

	PC1		PC2		PC3	
	$\mu$	$\sigma$	$\mu$	$\sigma$	$\mu$	$\sigma$
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
$\mu_{asy}$	3.32	0.99	18.23	2.01	2.80	1.82
$\sigma_{asy}$	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
$\mu_{dis}$	2.55	0.89	18.12	1.85	1.57	1.32
$\sigma_{dis}$	0.57	0.33	2.74	1.63	30.61	2.66
$\lambda$	49.56	1.16	27.14	0.54	13.25	0.95

perfect orthogonality to centrality metrics.

Dispersion was more prevalent in symmetry-related metrics than for the clustering coefficient, as indicated in Table VIII. This is also illustrated 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





FIG. 4. The first plot shows degree versus clustering coefficient. This typical pattern is well known, since high clustering is more incident in vertices with lower degrees. The second plot is analogous but the first component is an average of centrality metrics. The second component remains related to the clustering coefficient. The third plot exhibits the greater dispersion in the symmetry-related second component. In this case, the clustering coefficient is only relevant for the third component. This greater 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 was obtained 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 metrics of degrees, strengths and betweenness centrality, the symmetry-related metrics and clustering coefficient.

dispersion in the third component (clustering related). The PCA plot in the third system of Figure 4, where all metrics are considered, reflects the relevance of the symmetry-related metrics for the variance. We conclude that the latter metrics can be more meaningful in characterizing interaction networks (and their participants) than the clustering coefficient, especially for hubs and intermediary vertices.

The relative importance of the topological metrics was also observed for the additional 12 networks from Facebook, Twitter and Participa.br. With the exception of two of these networks, the overall behavior was maintained in that centrality measurements were found to be the most relevant to explain network topology, followed by symmetry-related metrics and then clustering coefficient. The results are given in Tables XXXI, XXXII, XXXIII, XXXIV in the Supporting Information. There are larger differences between two of these networks than between two (Gmane) email networks, as the latter were much more regular.

#### D. Types from Erdős sectors

A sector to which a vertex belongs can be regarded as yielding a type to the corresponding participant. Assigning a type to a participant inevitably raises an important question regarding the possible stigmatization. We take the view that the participation typology inherent in the Erdős sectors is not stigmatizing because the type of an individual changes constantly<sup>33</sup>. That is to say, an individual is a hub in a number of networks and peripheral in other networks, and even within a network he/she probably changes type along time. Indeed, we did observe often transitions of participants from one sector to another. The typology proposed here bridges exact and human sciences and may be enriched with concepts from other typologies, such as Meyer-Briggs, Pavlov or the authoritarian types of the F-Scale<sup>33</sup>.

We analyzed the time evolution of the networks using visualization tools developed for this research<sup>34,35</sup> and inspected the raw data to infer the main characteristics of each type. Our main observations may be summarized as follows:

- Core hubs usually have intermittent activity. Very stable activity was found on MET hubs, which is consistent with the literature where greater stability occurs in smaller communities<sup>7</sup>.
- 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. 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.

With regard to the networks as the whole objects of analysis, we were able to observe a negative correlation between the number of threads and the number of participants. When the number of participants exceeds a threshold, the number of threads displays a positive correlation with the number of participants. This finding is illustrated in Figure 5 and can also be observed in Table I. Network types can be derived from such results, which is left for the reader and future work.

## V. CONCLUSIONS

The most important result from the analysis of time evolution of the four email lists is certainly the time-independence observed not only for the activity but also for the properties of the networks themselves. For example, the relative fractions of participants classified as hubs, intermediary and peripheral vertices remained practically constant along time, and this applied to all the email lists studied. Furthermore, the PCA analysis of the topological metrics characterizing the networks also indicated that the contribution of each metric did not vary in time. Centrality metrics were found to be the most relevant to characterize the network topology, followed by symmetry-related metrics, which were more relevant, with respect to variance, than clustering.

A systematic study of the activity of participants belonging to the three distinct Erdős sectors indicated simple patterns for hubs and peripheral vertices, while the network structure was governed by the intermediary vertices. Because these properties were shared by all email

## Messages x Participants x Threads

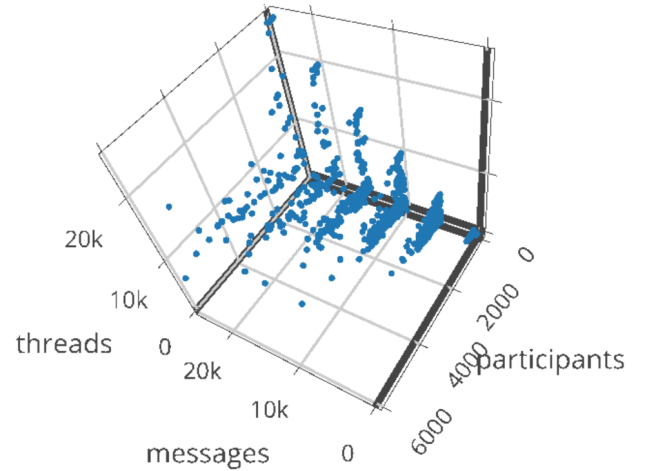


FIG. 5. A scatter plot of messages ( $M$ ) versus participants ( $N$ ) versus threads ( $\Gamma$ ) of 140 email lists. The highest number of threads are found in lists with few participants. The correlation between  $N$  and  $\Gamma$  is negative for low values of  $N$  but positive otherwise. This negative correlation between  $N$  and  $\Gamma$  can also be observed in Table I. For  $M = 20000$  messages, positive correlation of  $N$  and  $\Gamma$  is present mostly above 1500 participants. All LAU, LAD, MET are smaller networks.

lists and were time-independent, are coherent with literature, and was also found in networks from Facebook, Twitter and Participa.br, we may consider the classification of agents into Erdős sectors as leading to a human typology which bridges between exact sciences, with objective procedures for the classification, and human sciences, where there is a legacy in the observation of human types.

## 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<sup>33</sup>. 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: DATA AND SCRIPTS

Messages were downloaded from the Gmane public database<sup>17</sup>. 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<sup>16</sup>.

- <sup>1</sup>M. Newman, *Networks: an introduction* (Oxford University Press, 2010).
- <sup>2</sup>B. Latour, “Reassembling the social. an introduction to actor-network-theory,” *Journal of Economic Sociology* **14**, 73–87 (2013).
- <sup>3</sup>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.
- <sup>4</sup>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).
- <sup>5</sup>B. Ball and M. E. Newman, “Friendship networks and social status,” arXiv preprint arXiv:1205.6822 (2012).
- <sup>6</sup>L. M. Ingebrigtsen, “Gmane,” (2008).
- <sup>7</sup>G. Palla, A.-L. Barabási, and T. Vicsek, “Quantifying social group evolution,” *Nature* **446**, 664–667 (2007).
- <sup>8</sup>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).
- <sup>9</sup>B. Travençolo and L. d. F. Costa, “Accessibility in complex networks,” *Physics Letters A* **373**, 89–95 (2008).
- <sup>10</sup>M. E. Newman, “Modularity and community structure in networks,” *Proceedings of the National Academy of Sciences* **103**, 8577–8582 (2006).
- <sup>11</sup>K. Gergen and M. Gergen, *Historical social psychology* (Psychology Press, 2014).
- <sup>12</sup>R. Albert and A.-L. Barabási, “Topology of evolving networks: local events and universality,” *Physical review letters* **85**, 5234 (2000).
- <sup>13</sup>K. Marek-Spartz, P. Chesley, and H. Sande, “Construction of the gmane corpus for examining the diffusion of lexical innovations,” (2012).
- <sup>14</sup>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).
- <sup>15</sup>V. Palchykov, K. Kaski, J. Kertész, A.-L. Barabási, and R. I. Dunbar, “Sex differences in intimate relationships,” *Scientific reports* **2** (2012).
- <sup>16</sup>R. Fabbri, “Python package to analyze the gmane database,” (2015), <https://pypi.python.org/pypi/gmane>.
- <sup>17</sup>Wikipedia, “Gmane — Wikipedia, the free encyclopedia,” (2013), online; accessed 27-October-2013.
- <sup>18</sup>Gmane.linux.audio.users is list ID in GMANE.
- <sup>19</sup>Gmane.linux.audio.devel is list ID in GMANE.
- <sup>20</sup>Gmane.comp.gcc.libstdc++.devel is list ID in GMANE.
- <sup>21</sup>Gmane.politics.organizations.metareciclagem is list ID in GMANE.
- <sup>22</sup>E. A. Leicht and M. E. Newman, “Community structure in directed networks,” *Physical review letters* **100**, 118703 (2008).
- <sup>23</sup>M. Newman, “Community detection and graph partitioning,” arXiv preprint arXiv:1305.4974 (2013).
- <sup>24</sup>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).
- <sup>25</sup>M. O. Jackson, “Social and economic networks: Models and analysis,” (2013), <https://class.coursera.org/networksonline-001>.
- <sup>26</sup>I. Jolliffe, *Principal component analysis* (Wiley Online Library, 2005).
- <sup>27</sup>U. Brandes, “A faster algorithm for betweenness centrality\*,” *Journal of Mathematical Sociology* **25**, 163–177 (2001).
- <sup>28</sup>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](http://www.youtube.com/watch?v=-t5jxQ8cKxM&list=PLf_EtaMqu3jU-1j4jiIUiYMqyVSzIYeh6).
- <sup>29</sup>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/](http://hera.ethymos.com.br:1080/redes/python/autoRede/gmane.linux.audio.devel_3000-4200-280/).
- <sup>30</sup>R. Fabbri, L. d. F. Costa, and O. N. d. Oliveira jr, “On-line gadget for making email interaction network images, gml files and measurements.” (2013), <http://hera.ethymos.com.br:1080/redes/python/autoRede/escolheRedes.php>.
- <sup>31</sup>Numpy version 1.6.1, “random.randint” function, was used for simulations, algorithms in <https://pypi.python.org/pypi/gmane>.
- <sup>32</sup>S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D.-U. Hwang, “Complex networks: Structure and dynamics,” *Physics reports* **424**, 175–308 (2006).
- <sup>33</sup>T. W. Adorno, E. Frenkel-Brunswick, D. J. Levinson, and R. N. Sanford, “The authoritarian personality.” (1950).
- <sup>34</sup>R. Fabbri, “A connective differentiation of textual production in interaction networks,” (2013), <http://arxiv.org/abs/1412.7309>.
- <sup>35</sup>R. Fabbri, “Versinus: a visualization method for graphs in evolution,” arXiv preprint arXiv:1412.7311 (2014).