Temporal stability in human interaction networks

Renato Fabbri[1, a)]
São Carlos Institute of Physics, University of São Paulo (IFSC/USP)

(Dated: 13 August 2015)

In this study, we demonstrate a remarkably stable activity in human interaction networks. The activity along time and topology evolution was investigated 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. Notably, the activity in timescales ranging from seconds to 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 against the Erdös-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 are 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 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[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, 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[4,5], 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 task, 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 these properties were shared by networks from Twitter, Facebook and Participa.br, and are consistent with the literature, 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[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. Subsidiary results from the email lists and of networks from Twitter, Facebook and Participa.br are given in the Supporting Information.

A. Related work

Research on network evolution is often restricted to network growth, in which there is a monotonic increase in the number of events[6]. Exceptions are reported in this

a) http://ifsc.usp.br/~fabbri/; Electronic mail: fabbri@usp.br

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[6]. 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[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 analyzed networks, which matches empirical evidence from the literature[7] and motivated the inclusion of symmetry metrics in our analysis. 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 was indeed 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 (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 representative of both a diverse set and ordinary lists, are:

- 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.

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 discussion threads (count of messages without antecedent). $\overline{M}$ is the number of messages missing in the 20,000 collection. $100\frac{23}{20000} = 0.115$ percent in the worst case.

list	$date_1$	$date_M$	N	Γ	$\overline{M}$
LAU	2003-06-29	2005-07-23	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

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 discussion 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. 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 over 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 clue 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 unit complex number, $z = e^{i\theta} = \cos(\theta) + i\sin(\theta)$, where $\theta = measurement\frac{2\pi}{T}$, where T is the period in which the counting is repeated. E.g. $\theta = 12\frac{2\pi}{24} = \pi$ for a message sent at $12h$ and given $T = 24h$ for days. The moments m_n, lengths of moments R_n, mean angle θ_μ, and rescaled mean angle θ'_μ are defined as:

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

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 $\delta(z)$:

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

As expected, a positive correlation was found in all $Var(z)$, $S(z)$ and $\delta(z)$ dispersion measures, as can be noticed in Section I A of the 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[3,21,22]. Networks in this paper are directed and weighted, the most informative of trivial possibilities. 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 paper uses the information network as described above and depicted in Figure 1. Edges in both directions are allowed. Each time an interaction occurs, the value of one is added to the edge weight. Selfloops were regarded as non-informative and discarded. These human social interaction networks are reported in the literature as exhibiting scale-free and small world properties, as expected for (some) social networks[3,23].

C. Erdös sectioning

In a scale-free network, the peripheral, intermediary and hubs sectors can be derived from a comparison against an Erdös-Rényi network with the same number of edges and vertices[25], 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*.

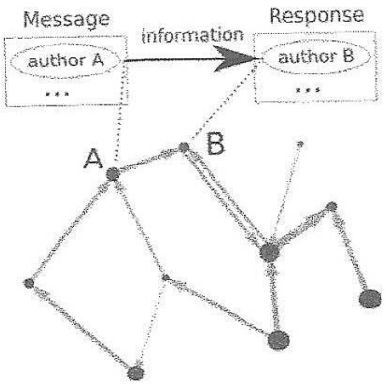

FIG. 1. The formation of interaction networks from email messages. Each vertex represents a participant. A reply message from B to a message from A is regarded as evidence that B has received information from A and yields a directed edge. Multiple messages add "weight" to a directed edge. Further details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

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

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

If $\mathcal{N}_f(N, z)$ 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

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

The lower degree fat tail corresponds to the border vertices, i.e. the peripheral sector or periphery where $\widetilde{P}(k) > P(k)$ and k is lower than any 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 as follows: 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.

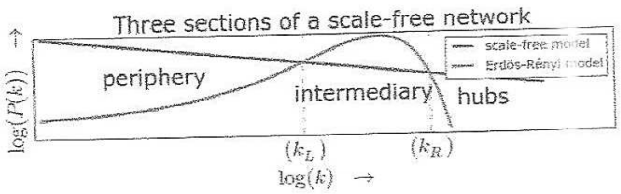

FIG. 2. Classification of the scale-free network vertices by comparing the degree distribution against that of an Erdös-Rényi ideal network. The latter has more intermediary vertices, while the former has more peripheral and hub vertices. The sector borders are defined by the two intersections k_L and 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 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. The range Δ of incident values should be partitioned in m parts $\Delta = \cup_{i=1}^{m} \Delta_i$, with $\Delta_i \cap \Delta_j \; \forall \; i \neq j$. Thus, $\Delta_i = \{max(\Delta_{i-1}) < k \leq j \;\; | \;\; \sum_{\Delta_{i-1}+1}^{j} \eta_k > \eta\}$, with η_k the number of vertices with degree k and $max(\Delta_0) = -1$.This changes equation 3 to

$$\sum_{x=min(\Delta_i)}^{max(\Delta_i)} \tilde{P}(x) < \sum_{x=min(\Delta_i)}^{max(\Delta_i)} P(x) \Rightarrow \quad (5)$$

$\Rightarrow \Delta_i$ hods intermediary degree values.

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

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

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 do not usually reach 100%, which is indicated by a black line in Figure 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 the total number of members may exceed 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.

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, and 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 to apply PCA for the topological metrics. The values of ws employed 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 Section II of the Supporting Information.

Networks were visualized with animations, image galleries and online gadgets developed specifically for this research[28–30]. Such visualizations were crucial to guide research into the most important features of network evolution. Furthermore, the size of the three Erdös 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.

F. Availability of data and scripts

In order to share routines with the scientific community and the whole of the society, all the required software to achieve the results reported in this article, including tables and figures in the Supporting Information, are available through a public domain Python package and an open Git repository[15].

Data from social networks used in this study were gathered and used within the anthropological physics framework[31]. All data used are also publicly accessible, either because they were already in public domain or because we published our own annotations. Messages were downloaded from the GMANE public database[16]. Data annotated from Facebook and Twitter are in a public repository[32]. Data from Participa.br was used from the linked data/semantic web RDF triples reported in[33] and available in[34].

This open approach is a way to enhance the reliability of the methods, of algorithmic routines, of data consistency, and of the results themselves. Also, by handing not only the framework and results, but the exact data and processes that render them, this enriches the scientific nature of our contribution[35].

TABLE II. The rescaled circular mean θ'_μ and the circular dispersion $\delta(z)$ described in Section III A. This typical table was constructed using all LAD list messages, and the results are the same for other lists, as shown in Section I A of the Supporting Information. The 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.

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

IV. RESULTS AND DISCUSSION

A. Activity along time

The observed 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 computed as were circular average values and their dispersion. We chose to provide 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.

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 simula-

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 Section IB1 of 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	2.88				
3h	1.10		2.47			
4h	0.68	1.37		3.44		
5h	0.69					
6h	0.83	2.07	4.35		23.09	
7h	1.24					
8h	2.28	6.80		21.03		
9h	4.52		18.75			
10h	6.62	**14.23**				
11h	**7.61**					
12h	6.44	12.48	**18.95**	**25.05**	**37.63**	**66.24**
13h	6.04					
14h	6.47	12.57				
15h	6.10		18.68			
16h	6.22	12.58		23.60		
17h	6.36					
18h	6.01	11.02	15.88		28.61	
19h	5.02			17.59		
20h	4.85	9.23				
21h	4.38		12.73			
22h	4.06	8.36				
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

tions using uniform distribution[36]. In the networks, $\frac{max(incidence)}{min(incidence)} \in (1.26, 1.275]$ while simulations reach these values but have on 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. 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 between one third and two thirds of activity on weekends. Days of the month were regarded as homogeneous with an inconclusive slight tendency of the first week 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 Erdös sector

The distribution of vertices in the hubs, intermediary, periphery Erdös 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 email lists analyzed, as demonstrated in Figure 3 and in Section II of the Supporting Information. Activity is highly concentrated on the hubs, while a very large number of peripheral vertices contribute to only a fraction of the activity. This is expected for a system with a scale-free profile, as confirmed by the data in Table VII for the distribution of activity

TABLE V. Activity in the days along the month for the MET list. Nearly identical distributions are found on other lists as indicated in Section IB3 of the Supporting Information. Although slightly higher activity rates are found in the beginning of the month, the most important feature seems to be the homogeneity made explicit by the high circular dispersion in Table 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			
6	3.73	16.98		
7	3.17			
8	3.26			
9	3.56			
10	3.26			
11	3.81	15.73	31.98	49.04
12	2.91			
13	3.30			
14	2.75			
15	2.95			
16	3.36	16.25		
17	3.16			
18	3.44			
19	3.36			
20	2.93			
21	3.20	15.79	32.78	
22	3.11			
23	3.60			
24	2.74			
25	3.13			
26	3.13	16.99		
27	3.07			
28	3.61			
29	3.60			
30	3.57			

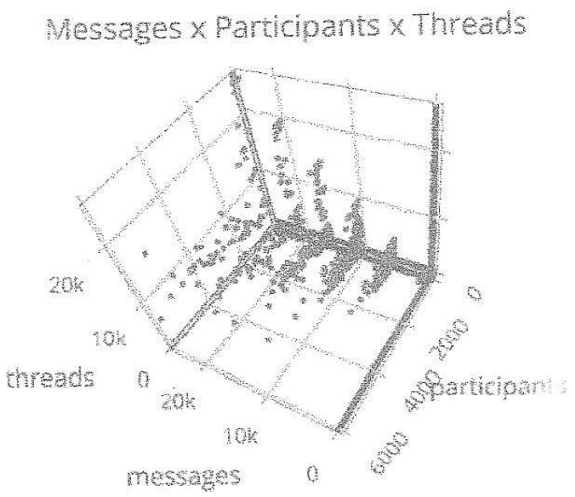

FIG. 5. A scatter plot of number of messages (M) versus number of participants (N) versus number of threads (Γ) for 140 email lists. Highest number of threads are found in lists with few participants. The correlation between N and Γ is negative for low values of N but positive otherwise. This negative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

ple patterns for hubs and peripheral vertices, while the network structure was governed by the intermediary vertices. These properties were shared by all email lists and were time-independent, being consistent with the literature. Moreover, both the distribution of Erdös sectors and the contribution from the metrics to the PCA were found to apply to networks from Facebook, Twitter and Participa.br. We may theref:e consider the classification of agents into Erdös sectors as a first step leading to a human typology which bridges exact sciences, with objective procedures for the classification, and human sciences, where there is a legacy in the observation of human types.

ACKNOWLEDGMENTS

Financial support was obtained from CNPq (140860/2013-4, project 870336/1997-5), United Nations Development Program (contract: 2013/000566; project BRA/12/018) and FAPESP. The authors are grateful to the American Jewish Committee for maintaining an online copy of the Adorno book used on the epigraph[38], to GMANE creators and maintainers, and to the communities of the email lists and other groups used in the analysis, and to the Brazilian Presidency of the Republic for keeping Participa.br code and data open. We are also grateful to developers and users of Python scientific tools.

[1] J. L. Moreno, "Who shall survive?: A new approach to the problem of human interrelations." (1934).
[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] R. Fabbri, "Python package to observe time stability in the gmane database," (2015), https://pypi.python.org/pypi/gmane.
[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] M. Newman, *Networks: an introduction* (Oxford University Press, 2010).
[24] 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).
[25] M. O. Jackson, "Social and economic networks: Models and analysis," (2013). https://class.coursera.org/networksonline-001.
[26] I. Jolliffe, *Principal component analysis* (Wiley Online Library, 2005).
[27] U. Brandes, "A faster algorithm for betweenness centrality*," Journal of Mathematical Sociology **25**, 163–177 (2001).
[28] R. Fabbri, "Video visualizations of email interaction network evolution," (2013-5), https://www.youtube.com/playlist?list=PLf_EtaMqu3jVodaqDjN7yaSgsQx2Xna3d, https://www.youtube.com/playlist?list=PLf_

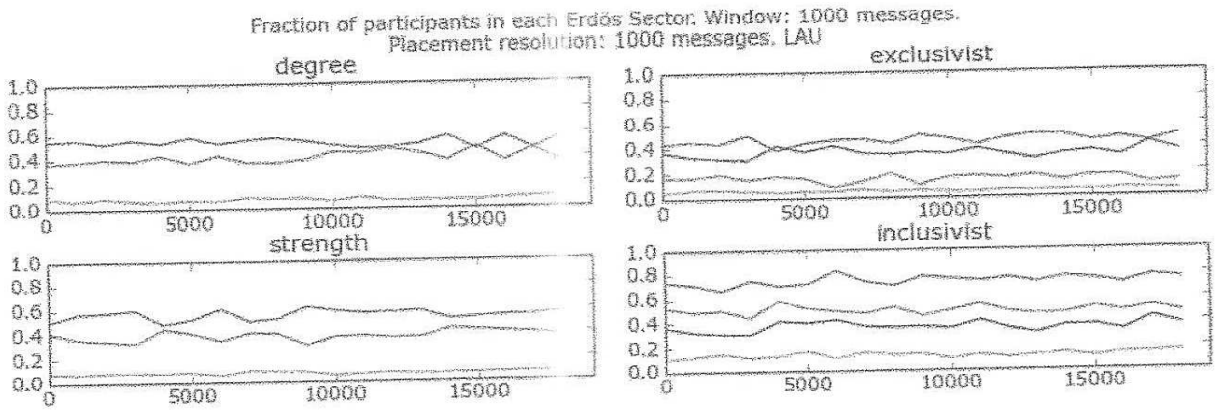

FIG. 3. Fractions of agents in each Erdös sector, where the fractions for hubs, intermediary and peripheral vertices are represented in red, green and blue, respectively. We used two simple criteria, namely degree and strength, for the graphics on the left. For the graphics on the right we employed the Exclusivist and Inclusivist compound criteria, with black lines representing the fraction of vertices without class and with more than one class, respectively. See Section II of Supporting Information for a collection of such timeline figures with all simple and compound criteria and metrics. Table S30, also from Supporting Information, presents these fractions of agents in snapshots of networks from Facebook, Twitter and Participa.br.

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 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 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 S31, S32, S33, S34 of 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 Erdös 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 Erdös sectors is not stigmatizing because the type of an individual changes along time and as different networks are considered[38]. 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[38].

We analyzed the time evolution of the networks using visualization tools developed for this research[39,40] 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[6].
- 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 com-

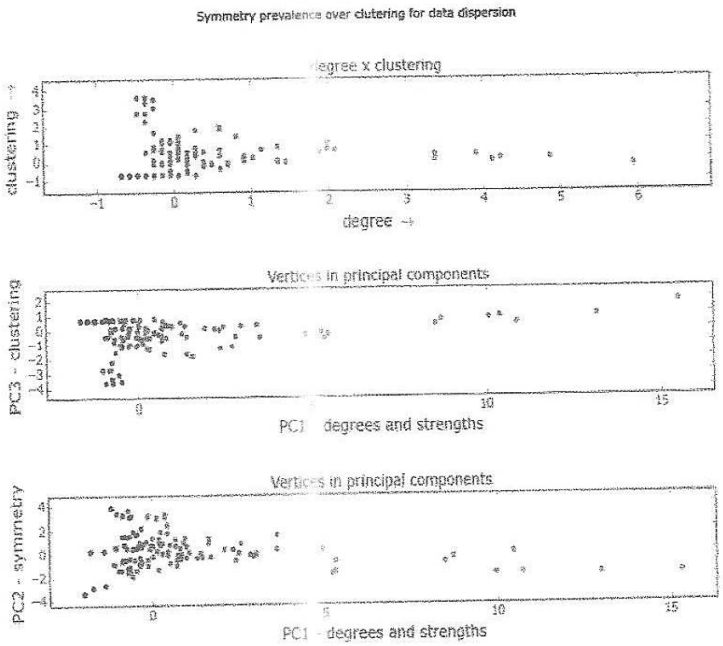

FIG. 4. The first plot shows degree versus clustering coefficient. This typical pattern is well known, since high clustering is more incident in vertices with lower degrees. The second plot is analogous but the first component is an average of centrality metrics. The second component remains related 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.

munication 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, i.e. members known each other often, is found mostly 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. Obviously, network types can be derived from such results, which was not attempted here but 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 across all 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 Erdös sectors indicated sim-