Stability in human interaction networks: primitive typology of vertex, prominence of measures and activity statistics

Renato Fabbri,[a] Vilson Vieira da Silva Junior,[b] Ricardo Fabbri,[c] Deborah Christina Antunes,[d] and Marilia Mello Pisani[e]
São Carlos Institute of Physics, University of São Paulo (IFSC/USP)

(Dated: 21 January 2015)

This article reports a characterization of interaction networks and its temporal stability. Such a task involves a selection of aspects to investigate, which lead to: 1) activity distribution in time and among participants; 2) a sound classification of vertex: peripheral, intermediary and hub sectors; 3) combination of basic measures into components with greater dispersion (PCA). While time patterns of activity are not obvious, participant activity follows concentrations expected in scale-free networks. Comparison of incident networks with ideal Erdös-Rényi networks bearing the same number of edges and vertexes reveals a sound criterion for distinguishing sectors on the networks. Principal components in the basic measures space revealed interesting and regular patterns of independence and dispersion. This includes a ranking of measures that most contribute to dispersion: 1) degree and strength measures, 2) symmetry related quantization, and 3) clusterization. Results suggest typologies for these networks and participants. Further work include considerations of text production, psychoanalysis inspired typologies, participatory democracy exploitation of observed properties, and better visualization support for network evolution.

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

'The conception of personality structure is the best safeguard against the inclination to attribute persistent trends in the individual to something "innate" or "basic" or "racial" within him. The Nazi allegation that natural, biological traits decide the total being of a person would not have been such a successful political device had it not been possible to point to numerous instances of relative fixity in human behavior and to challenge those who thought to explain them on any basis other than a biological one.'

- Adorno et al, 1969, p. 747

I. INTRODUCTION

The present work is aimed at finding common characteristics among (email) interaction networks. This includes observations along time, which imply network evolution, a field that has received dedicated attention from the research community for more than a decade[1,2].

While significant measures will depend on the model and system characteristics[3,4], this work considers only directed, weighted and human interaction networks. Undirected and unweighted representation of such networks is also found in the literature and can be obtained by simplification[5].

Text mining and typologies of online participants benefit from the results here presented[6,7]. Although all networks considered originated from email lists, coherence with literature suggests that results hold for a more general class of interaction networks, such as observed in online platforms (e.g. LinkedIn, Facebook, Twitter).

A. Related work

This work approaches network stability through temporal evolution observations. This evolution is characterized by a constant number of contiguous messages, of which snapshots are considered to yield a timeline of the network, which is not explored to date, as far as authors know[8]. Works on network evolution often consider solely network growth, in which there is a monotonic increase in the number of events considered[1]. Moreover, selected exceptions are reported in this section.

The evolution of interaction networks was addressed with a community focus, in a work that ignores the direction of edges[1]. Two topologically different networks are reported to emerge, depending on the frequency of interactions, presenting a generalized power law or an exponential connectivity distribution[9]. In email networks, free-scale properties were verified[10], and different linguis-

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

tic traces were related to weak and strong ties[5].

Such results are in accordance with phenomena observed in this work. Linguistic characterization is being described in a deriving article[6]. See Appendix B for further considerations of related work, with emphasis on the results from research herein reported.

II. DATA DESCRIPTION

A. Email lists and messages

Email list messages were obtained from the GMANE email archive[11], which consists of more than 20,000 email lists and more than 130,000,000 messages[12]. These lists cover a variety of topics, mostly technology-related. It can be described as a corpus with metadata of its messages, such as time, place, sender name, sender email address. GMANE usage in scientific research is reported in studies of isolated lists and of lexical innovations[5,10]. Appendix C is dedicated scripts for gathering and processing GMANE email messages.

B. Chosen dataset

The four lists below were selected for the resulting diversity, easing initial observance of natural and general properties.

- Linux Audio Users list[13]. Dominated by participants with hybrid artistic and technological interests. Participants are from different countries, and English is the language used the most. Abbreviated as LAU from now on.
- Linux Audio Developers list[14]. Participants are 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[15]. Dominated by specialized computer programmers. Participants are from different countries, and English is the language used the most. Abbreviated as CPP from now on.
- List of the MetaReciclagem project[16]. Dominated by Brazilian activists and digital culture interests. Participants are mostly Brazilians, and 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 exposed in Table I.

TABLE I. Columns $date_1$ and $date_M$ have first and last message dates from the 20,000 messages considered in each email list. N is the number of participants (number of different email addresses). Γ is the number of threads (count of messages without antecedent). $\overline{M}$ is messages missing in the 20,000 collection, $100\frac{23}{20000} = 0.115$ percent in the worst case. MET notably has the fewer participants and the larger number of threads. This relation holds for each pair of the lists considered: as the number of participants increases, the number of threads decreases.

list	$date_1$	$date_M$	N	Γ	$\overline{M}$
LAU	Jun/29/2003	Jul/23/2005	1183	3373	5
LAD	Jun/30/2003	Oct/07/2009	1268	3113	4
MET	Ago/01/2005	Mar/07/2008	492	4607	23
CPP	Mar/13/2002	Aug/25/2009	1052	4506	7

III. CHARACTERIZATION METHODS

After immersion with the data and appropriated literature, the following methods for network characterization were chosen: 1) statistics of activity along time; 2) division of network in hubs, intermediary and peripheral vertex; 3) prominence of topological measures; 4) evolutive visualizations and quantitative observations; 5) typological elaborations of networks and participants. Each of these methods are described bellow with supporting structures.

A. Temporal activity statistics

Number of messages along time with respect to seconds, minutes, hours, days of the week, days of the month, and months of the year. These are exhibited as tables in Appendix D 2. Results are outlined in Section IV A.

B. Interaction network

Regarding literature[10,17,18], interaction networks can be modeled both weighted or unweighted, both directed or undirected. Networks in this article are directed and weighted, considered as more informative among possibilities (directed unweighted, undirected weighted, and undirected unweighted). More precisely, the networks reported are erected 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 read what A wrote and formulated a response, so B assimilated information from A, thus $A \to B$. Inverting edge direction yields the status network, as B read the message and considered what A wrote worth responding, giving status to A, thus $B \to A$. This article uses the information network described above and depicted in Figure 1. Edges in both directions are al-

lowed. Each time an interaction occurs, one is added to edge weight. Self-loops were regarded as non-informative and discarded. This networks are described as exhibiting free-scale and small world properties, as expected for a social network[10].

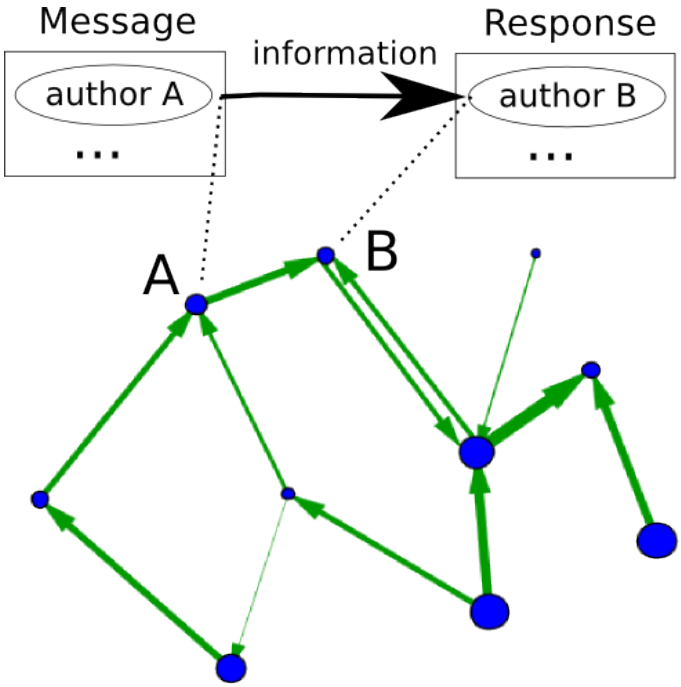

FIG. 1. Formation of interaction network from email messages. Each vertex represents a participant. A reply message from participant B to a message from participant A is considered as evidence that B received information from A. Multiple messages add "weight" to directed edge. Further details are in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

Previous messages on the thread create directed edges from their author to the observed message's author. Edges can be created from all antecedent messages on the message-response thread. In this work, only the immediate predecessor are linked to new message's author, both for simplicity and for the valid objection that in adding two edges, $x \to y$ and $y \to z$, there is also a connection between $x \to z$. Potential interpretations for this weaker connection are usually common sense, such as: double length, half weight or with one more "obstacle". This suggests the adoption of other centrality measures that account for the connectivity with all nodes, such as betweenness centrality and accessibility[19,20].

1. Sectioning network in periphery, intermediary and hubs classes

Because of social networks tendency to have a scale-free distribution of connectivity, one can compare it to an Erdös-Rényi random graph and consider peripheral, intermediary and hub sectors[21], as depicted in Figure 2.

The degree distribution $\widetilde{P}(k)$ of an ideal scale-free network $\mathcal{N}_f$ with N vertexes and z edges, has less average degree nodes when compared with the distribution $P(k)$ of an Erdös-Rényi random graph with the same number of vertexes and edges:

$$\widetilde{P}(k) < P(k) \Rightarrow \text{k is intermediary degree} \qquad (1)$$

If $\mathcal{N}_f$ is directed and has no self-loops, the probability of the presence of an edge between two arbitrary vertexes is $p_e = \frac{z}{N(N-1)}$ (see Appendix A). A vertex in the ideal Erdös-Rényi digraph with the same number of vertexes and edges, and thus the same probability p_e for the presence of an edge, will have degree k with probability:

$$P(k) = \binom{2(N-1)}{k} p^k (1-p)^{2(N-1)-k} \qquad (2)$$

The lower degree fat tail constitute the border vertexes or peripheral sector. The higher degree fat tail is the hub sector.

The arguments behind this classification are: 1) vertexes so connected that they are virtually inexistent in networks whose edges are incident regardless of vertex properties, are correctly associated to the hubs sector. Vertexes with very few connections, which are way more abundant than expected in networks whose edges are incident regardless of vertex properties, are correctly associated to periphery. Degree values near average, predicted as the most abundant if connections are created without criteria, and less frequent in free-scale phenomena, are correctly associated to intermediary vertexes.

To assure statistical validity, bins can be chosen to contain at least η vertexes. Thus, each bin, starting at degree k_i, spans $\Delta_i = [k_i, k_j]$ degree values, where j is the smallest integer in which degrees k_i-k_j contain at least η values. This changes equation 1 to:

$$\sum_{x=k_i}^{k_j} \widetilde{P}(x) < \sum_{x=k_i}^{k_j} P(x) \Rightarrow \text{i is intermediary} \qquad (3)$$

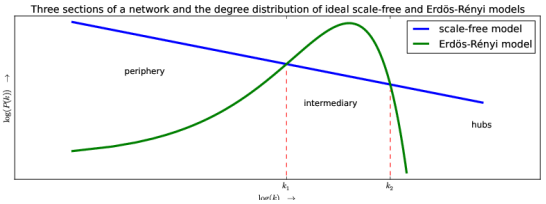

FIG. 2. Degree distribution on scale-free and Erdös-Rényi ideal networks. The later has more intermediary, as the former has more peripheral and hub vertexes. Sections are given by the two intersections k_1 and k_2 of the connectivity distributions. Characteristic degrees are in compact intervals of degree: $[0, k_1]$, $(k_1, k_2]$, $(k_2, k_{max}]$ for the three sections considered (periphery, intermediary and hubs).

If instead 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 $\overline{w}$ the average weight of an edge and s_i the vertex strength. For in and out degrees and strengths, comparisons should be made with $\kappa_i = 2k_i^{in}$, $\kappa_i = 2k_i^{out}$, $\kappa_i = 2\frac{s_i^{in}}{\overline{w}}$ and $\kappa_i = 2\frac{s_i^{out}}{\overline{w}}$. Results of these criteria for network segmentation are discussed in subsection IV B.

As a further refinement of the network segmentation, compound criteria is used for classification of vertex, considering all measures: total, in and out degree and strength. After a careful inspection of possible combinations, these were abbreviated to six:

- Exclusivist criteria: vertex are only classified if the class is the same with respect to all measures. In this case, total vertex classified (usually) does not reach 100%, which is depicted by a black line in Appendix E.
- Inclusivist criteria: vertex has class given by any of the measures. In this case, a vertex can have more than one class and the fraction of class attribution beyond total number of vertexes is also depicted by a black line in Appendix E.
- Exclusivist cascade: hubs are only hubs if classified as hub with respect to all measures. Intermediary are the vertexes classified as intermediary or hub with respect to all measures. Vertexes left are regarded as peripheral vertex.
- Inclusivist cascade: hubs are vertexes classified as hubs by any of the measures. From the vertexes left, if any is classified as intermediary by any measure, than it is intermediary. The rest of the vertexes are peripheral.
- Exclusivist externals: hubs have unanimous classification with respect to all measures. Of vertexes left, peripheral vertexes are the ones classified as hub or peripheral by simple criterion. The rest represent intermediary sector.
- Inclusivist externals: hubs are vertexes classified as hubs with respect to any measure. From vertexes left, if a vertex is classified as peripheral with respect to any measure, than it is peripheral. The rest is regarded intermediary sector.

These compound criteria, and reduction of possibilities to them, can be formalized in strict mathematical terms. This was considered out of the scope of the present article.

Results from applying this classification method is further reported in Section IV B.

2. Topological measures

This article restricts topological analysis to a small selection of the most basic measures of each vertex:

- Degree d_i: number edges linked to node i.
- In-degree d_i^{in}: number of edges ending at node i.
- Out-degree d_i^{out}: number of edges departing from node i.
- Strength s: sum of weights of all edges linked to node i.
- In-strength s_i^{in}: sum of weights of all edges ending at node i.
- Out-strength s_i^{out}: sum of weights of all edges departing from node i.
- Clustering coefficient cc_i: fraction of pairs of neighbors of i that are linked. This is the standard clustering coefficient for undirected graphs. Usage of other clustering coefficients (e.g. for directed graphs) was considering out of scope, as such clustering coefficients are usually employed in special contexts.
- Betweenness centrality bt_i: fraction of geodesics that contain the node i. Betweenness centrality index considered directions and weight, as specified in[22].

In order to capture asymmetries in the activity of participants, the following metrics were introduced (see subsection IV C):

- asymmetry of vertex i: $asy_i = \frac{d_i^{in} - d_i^{out}}{d_i}$.
- mean of edge asymmetry of vertex i: $\mu_i^{asy} = \frac{\sum_{j \in J_i} e_{ji} - e_{ij}}{|J_i|}$. Where e_{xy} is 1 if there is and edge from x to y, 0 otherwise. $|J_i|$ is the number of neighbors of vertex i.
- standard deviation of edge asymmetry of vertex i: $\sigma_i^{asy} = \sqrt{\frac{\sum_{j \in J_i} [\mu_{asy} - (e_{ji} - e_{ij})]^2}{|J_i|}}$
- disequilibrium of vertex i: $dis_i = \frac{s_i^{in} - s_i^{out}}{s_i}$.
- mean of edge disequilibrium of vertex i: $\mu_i^{dis} = \frac{\sum_{j \in J_i} \frac{w_{ji} - w_{ij}}{s_i}}{|J_i|}$, where w_{xy} is weight of edge $x \to y$ and zero if there is no such edge.
- standard deviation of edge disequilibrium of vertex i: $\sigma_i^{dis} = \sqrt{\frac{\sum_{j \in J_i} [\mu_{dis} - \frac{(w_{ji} - w_{ij})}{s_i}]^2}{|J_i|}}$

C. Evolutive observations

Evolution of network is observed within a fixed number of messages (window size: ws) that shifts in the message timeline. All ws = 50, 100, 200, 400, 500, 800, 1000, 2000, 2500, 5000 and 10000 were used. Within a same ws, the number vertexes and edges vary in time, as do other network characteristics. Further work should deepen inspection of measure interdependence, this article holds to measures in Section III B 2.

Visualization of network evolution

In refining hypotheses, visualization of the network was crucial. Animations, image galleries and online gadgets were made[23–25]. Mapping of various topological measures to glyphs and layouts are being further explored as a parallel research. Furthermore, stable aspects of measures prominence along time are captured through mean and standard deviation (see Section III B 2 and Appendix D 1). Constant sector sizes along time are observed in a timeline fashion in Appendix E.

D. Typological deepening

There are other ways to split a network. To point a common example, the center of the network is defined as all the nodes whose maximum distance to any other node is the radius[26]. In the same framework, the periphery (as opposed to the center) consists of the nodes whose maximum distance to any node is the diameter[27]. Accordingly, the intermediary sector can be defined as the nodes that are not in the center or in the periphery. Interestingly, in the email networks analyzed, with these criteria, the center can often be a factor of 4 times larger than the periphery and the intermediary group often exceed 93% of the nodes[28].

Models of human dynamics can be used to predict and classify activity. In this case, agent activity is commonly considered a Poisson process, as a consequence of the randomly distributed events in time. Even so, evidence-based models suggests that human activity patterns follow non-Poisson statistics, characterized by a long tail of inactivity with of bursts of rapidly occurring events[29,30]. Emails are reported as having a heavy tailed distribution with $\alpha = 1$, together with web browsing and library loans[29].

Typologies can also be conveniently adapted from psychiatric, psychological and psychoanalytic theories. Concerning empirical research, Theodor Adorno was a core conceiver of an one-of-a-kind typology that resulted from observing authoritarian personality traces[31], sometimes depicted as an authoritarian syndrome. Other typologies include Jung's extroversion-introversion trait with four modes of orientation. This four modes are divided in two perceiving functions (sensation and intuition) and two judging functions (thinking and feeling)[32]. Myers-Briggs Type Indicator extrapolated Jungian theories into a questionnaire and added perceiving and judging as a fourth dipole[33]. Even plain Freudian criteria, such as neurosis, psychosis, perversity and denegation, can be used directly for such categorization, as they have verbal and behavioral typical traces[34,35].

It was considered central to benefit from key human typologies, both by adding descriptions to a type and by further characterizing classes in the terms encountered.

IV. RESULTS AND DISCUSSION

A. Constancy and discrepancy of activity along time

1. Seconds and minutes

The incidence of messages at each second in a minute and at each minute in an hour is compatible with uniform distribution tests. If compared to simulations using an uniform distribution[36], messages were slightly more evenly distributed in all lists: for both seconds and minutes $\frac{max(incidence)}{min(incidence)} \in (1.26, 1.275]$. Simulations reach these values, but have in average more discrepant higher and lower peaks $\xi = \frac{max(incidence')}{min(incidence')} \Rightarrow \mu_\xi = 1.2918$ and $\sigma_\xi = 0.04619$.

2. Hours of the day

Table V shows how the four lists distribute activity along the day. Afternoon was the most active 6h period of the day. Second 12h more active than first 12h. Even so, activity peak occurs around midday, with a skew towards earlier hours.

3. Days of the week

Weekdays also exhibit an interesting pattern, to which is dedicated Table VI: a decrease of activity on weekends of at least one third, reaching two thirds in extreme cases.

4. Days along the month

Table VII shows activity along the month, in which no prevalent pattern was observed. Variation of activity in the days along the month is less prominent, one cannot point much more than a - probably not statistically relevant - tendency of first and second weeks to be more active. The most important trait might be their homogeneity with respect to activity. Last days of the month (29, 30 and 31) are not present in every month, and observed activity is proportional to incidence rates.

5. Months and larger divisions of the year

Table VIII is dedicated to activity in months and larger divisions of the year. Observation points two periods of more prominent activities: Jun-Aug (MET and LAD), and Dec-Mar (CPP, LAU and LAD), which fit academic calendars, vacations and end-of-year holidays.

B. Scalable fat-tail structure

There is a concentration of hub activity and of vertex with few connections. Table IX is dedicated to exposing this distribution of activity among participants.

As specified in Section III B 1, in order to classify nodes as hubs, intermediary or peripheral, the incident connectivity distribution $\widetilde{P}(k)$ is compared to a connectivity distribution of an ideal Erdös Rènyi network $P(k)$ with the same number of vertexes and edges. In the networks inspected for this article (see Section III B), if degree distribution is used for classification, hubs sector size reaches peaks with 10% of all vertexes. If strength is used for comparison ($\widetilde{P}(k_i = \frac{s_i}{\overline{w}})$) is compared to $P(k_i)$), hubs account for approximately 5% of all vertex, i.e. strength classification yields half the number of hubs as plain degree. This results hold for in and out degrees and strengths,

Classification criteria exposed in Section III B 1 was used efficiently with windows of at least 200 messages. Specially with 1000 or more messages, criteria yields stable fractions of $\approx 5\%$ of hubs, $\approx [15-20]\%$ of intermediary and $\approx [75-80]\%$ peripheral vertexes, which match literature expectations[37]. A compound criteria, also described in Section III B 1, can be used as a classification refinement. This is specially useful in dealing with fewer messages, in which case the structure degenerates with respect to some of the degree and strength measures, but not all. A minimum window size for observation of more general properties can be inferred by monitoring the giant component and the degeneration of hub, intermediary and peripheral sections. This degeneration is critical in the span of 50-100 messages for window sizes. With compound criteria, such as exclusive cascade of Figure 15, the networks seem to hold basic structure even with as few as 20-50 messages. This indicates that concentration of activity and low-activity participants occurs even with very few messages.

For the histograms used in the classification process, the usage of at least η vertexes for each bin did not yield significant differences. That was understood as a consequence of the observation scale: *There are between 20 and 1200 participants in the message window sizes used to derive most results ($ws \in [200, 1500]$ messages). As peripheral vertexes are abundant and span few degrees, there are more than η vertexes with each low degree value. Higher degrees cases requires a bit more reasoning. Specified ws implicates that each participant is $p \in [0.1\%, 0.5\%]$ of all participants. Therefore, if incident connectivity is very improbable in an Edös Rènyi network (less than p, the probability that a single participant represent when histogram is normalized to density function), than it is not an intermediary connectivity, but a hub.* This concludes the explanation of why using at least η vertexes for each bin did not impact results.

Appendix E is dedicated to figures of these networks and their evolution.

C. Prevalence of centrality over asymmetries and asymmetries over clusterization

The principal component (PCA[38]) exhibit ponderation of centrality measures: degrees, strengths and betweenness centrality. Clustering coefficient is presented in almost perfect orthogonality. Dispersion is more prevalent in symmetry related measures than clustering coefficient. This holds for all network snapshots observed, even with as few messages as to degenerate structure. Symmetric and asymmetric edges have been reported as bonded to different roles played by participants and relations[2]. Principal components formation from original measures can be observed in Tables II, III and IV. Individual vertexes relation with top two principal components can be appreciated in Figures 3 and 4. This peculiar first component that consists of the averaged sum of degree, strength and betweenness measures was verified to be incident in virtually all networks with 500 or more messages and most smaller networks (degeneration of basic structure is critical with $ws \approx 50 - 100$ messages). This composition of principal component suggests that all six degree and strength measures are equally important for system characterization, although it is known that they do not relate to the same participation characteristics.

As expected, degree and strength are highly correlated, with Spearman correlation coefficient $\in [0.95, 1]$ and Pearson coefficient $\in [0.85, 1)$ for larger window sizes ($ws > 1000$). Also trivial: high degree is associated with low clustering coefficient, as can be observed in Figure 5.

When symmetry of node connectivity is considered, the first component remains mostly the same, but clustering coefficient is only relevant to third and fourth components. A snapshot of vertexes with respect to these first two principal components are in Figure 4. This asymmetry and disequilibrium measures revealed as more proper measures to characterization of hubs and intermediaries, as seen in greater spreading of second PCA plot. Both symmetry of node overall activity and of individual relations play important role in second component, as can be observed in Table IV.

D. Primitive typology

This work aimed at finding common characteristics among (email) interaction networks. Analysis involved primary measures observance and formal criteria for co-

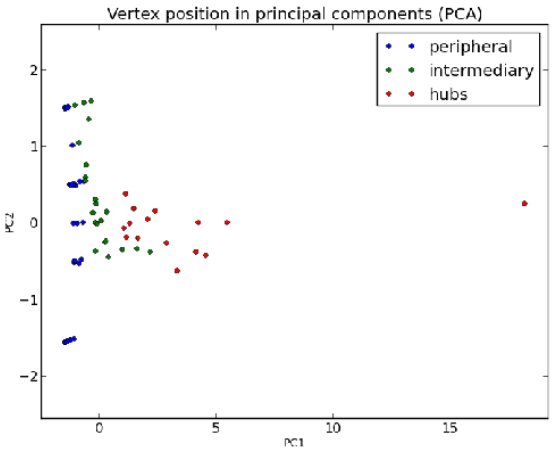

FIG. 3. PCA of in and out degree and strength, betweenness centrality and clustering coefficient, as specified in Section III B 2. Table III has the composition of principle components from the original measures. First principle component is a pondered sum of centrality measures: degrees, strengths and betweenness centrality. Second component is mostly clustering coefficient in this figure, but asymmetries holds second component if also considered. Similarity to plot in Figure 5 was verified with all window sizes considered ($ws \in [100, 10000]$), which exposes a common relation is held by degree, strength and betweenness measures to clustering coefficient.

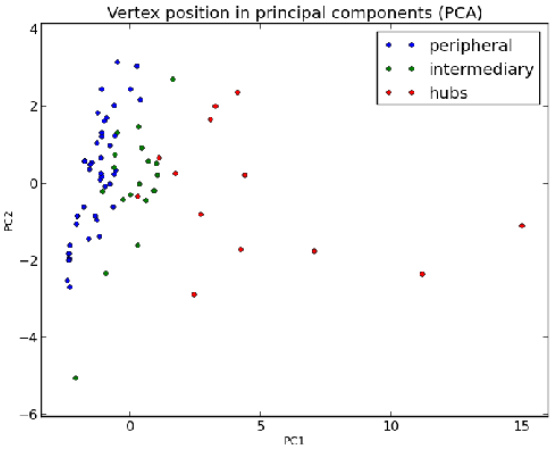

FIG. 4. Degree and strength, clustering coefficient, betweenness centrality and symmetry related measures are used for this scatter plot of principal components. Compositions of first three components are in table IV and measure details in subsection III B 2. Most importantly, clustering coefficient is only relevant for third component, being second component representative of symmetry measurements of vertex interactions.

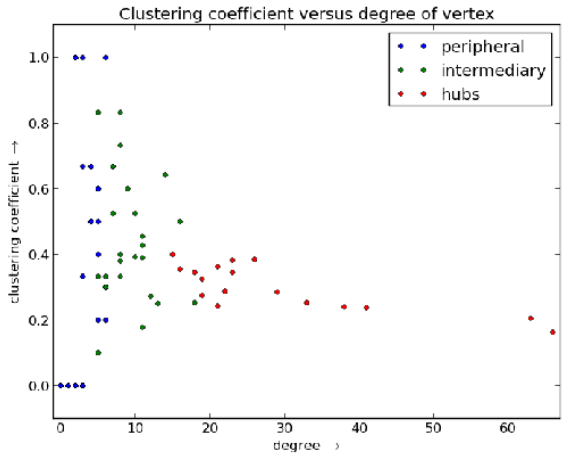

FIG. 5. Clustering versus degree of vertex with a window size of $ws = 1000$ email messages (LAU list). General layout is in accordance with literature: connected vertexes have low clusterization while higher clusterization is gradually more incident as number of connections is lowered.

herent ratios of hub, intermediary and hub sectors. Nevertheless, inspection done by visualizations and raw data manipulations suggests agents peculiarities of typological character. These are initial observations, which should be further developed as posed in Section IV A:

- Core hubs usually have intermittent activity. Very stable activity was found on MET hubs, which motivated its integration to this work. Literature reports greater stability of participation in smaller communities[1], reason why smaller number of participants in MET was considered coherent with the stable activity of hubs.

- Typically, hub activity is trivial: they interact as much as possible, in every occasion with everyone. Peripheral vertex activity also follows a simple pattern: they interact very rarely, in very few occasions. Intermediary vertexes seem responsible for network structure.

- Network operation exhibit modes that seem dictated by intermediary vertexes behavior. For example, these can exhibit preferential communication to peripheral, intermediary, or hub vertexes; 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 way more than receive responses.

- Most obvious community structure, as observed by

high clustering coefficient, is found only in peripheral and intermediary sectors.

This "primitive typology", characterized by peripheral, intermediary and hub types, can be further scrutinized using concepts involved in other typologies, such Meyer-Briggs, Pavlov or F-Scale. This has no pretension of being a direct result from numeric analysis, it is a description refinement of the found structure, in typological terms. Although initial, this bridges human and exact sciences in the most pertinent way authors were able to, as is herein considered a result.

V. CONCLUSIONS AND FUTURE WORK

Characterization of interaction networks resulted from stability observations. Along temporal activity statistics, this work reports the stability of the principal components (in the concentration of dispersion and composition) and of the ternary partitioning (periphery, intermediary, hubs) relative sizes, evident in the comparison with the Erdös-Rényi model.

A. Further work

The task of delivering a first and general characterization of chosen interaction networks involved starting a larger effort. The different aspects covered requires not only different analytical background, but also considerations about textual production and social psychology. These are receiving attention within dedicated works and are summarized in this section.

1. Constancy of general characteristics eases tipologization

Regarding topological aspects of interaction networks, further work should inspect other measures (e.g. closeness centrality, accessibility), and statistics in each of the three connective sectors: hubs, intermediary and peripheral.

Observance of attributes with greater contribution to principal components of LDA should reveal best chances to present these three sections as clusters in the network measurements space. Another possibility, specially for a brute-force characterization of such sections, is to remove vertexes with degree close to k_1 or k_2 depicted in figure 2. The subtraction $\widetilde{P}(k)-P(k)$ should result in two positive clusters for periphery and hubs, and a negative cluster for intermediary vertexes. This might support classification of the three sectors by clustering, a more traditional approach to classification.

Observed networks were coherent with literature in different aspects, such as concentration of activity, and clusterization versus connectivity patterns. Even so, analysis of data from other virtual environments, such as Facebook, Twitter and LinkedIn, should verify the generality of this report.

A related work observed textual production of network sectors[6]. Resulting knowledge purposes networks and participants tipologization, and both topological and textual analysis should foster characterization of interaction networks and participation incidences in a dedicated study. Stability reported in this article eases tipologization of outliers and more usual participation patterns.

2. Results exploitation

Usage of such characteristics are taking place in linked data and electronic government technologies[39–41]. Further steps involve elaboration and tests of social dynamics that takes advantages of these results.

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[42]. Authors thanks GMANE creators and maintainers, specifically: GMANE is run by Lars Magne Ingebrigtsen, and the administrators are Tom Koelman, Jason R. Mastaler, Steinar Bang, Jon Ericson, Wolfgang Schnerring, Sebastian D.B. Krause, Nicolas Bareil, Raymond Scholz, and Adam Sjgren. Authors thank referred email lists communities and welcome feedback as core contribution to this, and similar, research.

Appendix A: Deduction of edge existence probability in a directed network without self-loops

Be $\mathcal{N}$ a directed network without self-loops with z edges and N vertexes. The probability that an edge exists between two arbitrary vertex is $p_e = \frac{z}{max(\text{number of edges}|\text{ N vertexes})}$, where $max(\text{number of edges}|\text{ N vertexes}) = 2[(N-1)+(N-2)...1] = 2[\sum_1^{N-1} i] = 2[\frac{N(N-1)}{2}]$ is the maximum number of edges for a network with N vertexes. Therefore:

$$p_e = \frac{z}{max(\text{number of edges}|\text{ N vertexes})} = \\ = \frac{z}{2[(N-1)+(N-2)+...+1]} = \frac{z}{2\frac{N(N-1)}{2}} = \\ p_e = \frac{z}{N(N-1)} \quad \text{(A1)}$$

Appendix B: Further consideration of related work

Unreciprocated edges often exceed 50%, which matches empirical evidence reported in[2]. Although no correlation of topological characteristics and geographical position was found in a pertinent study[43], geographical incidences should be present in further refinement of the analysis.

The seminal Nature Letter by Palla, Barabási and Vicsek[1] has strong confluence with this work, suggesting that smaller size of MET community is responsible for the stronger hubs observed.

Controllability of these networks is also an uncovered issue. These has unintuitive properties and might bring into forefront crucial differences between email interaction networks and interaction networks in Facebook or Twitter[44–46].

Gender related behavior in mobile phone datasets has been reported[47]. This can be further investigated to hold in email lists and in evolving terms as community oriented, non-private interactions takes are drawn from public email groups with hundreds or thousands of participants.

Considered years altogether, tenths of thousands of participants can post on a list. The most active lists usually reaches a few thousands of participants. Analysis of resulting data might lead to deeper insights in community-related network evolution[11].

Appendix C: Data and scripts

Messages are downloaded from GMANE database by RSS in the mbox email text format. They are requested one by one to avoid reaching maximum size of the requests accepted by GMANE API.

Every message has about 30 fields, from which the following are crucial for the present work:

- "From" field, as it specifies the sender of the message, in the usual format of "First_name Last_Name $< email >$".
- "Date" field, which is given with the resolution of a second.
- "Message-ID", important to state antecedent/consequent relation between messages and therefore from an author to a replier.
- "References", has the ID of the message it is an answer to, if any, and earlier messages in the thread.

Field "In-Reply-To" has only the ID of the message it replies and can be sometimes a shortcut or an alternative to "References". Also, the textual content of the messages, accessed through "payload" method of the mbox message object, is of central interest and the authors dedicated an article to include the textual content of the messages to the analysis[6].

1. Third party libraries and software

The programming framework used is mainly Python-based, with emphasis on usual scientific tools. More specifically, scripts where written for 2.7.3 version of Python, with the following third party libraries: Numpy, Pylab/Matplotlib, NetworkX, IGraph. Behind the scenes, Graphviz is accessed via PyGraphviz to make network drawings.

2. Python scripts

All results were obtained with scripts written in the Python programming language. These are kept in a public git repository for backup and sharing with research community[48]. Core scripts, for deriving structures and results exhibited in this article, are in the LEIAME file.

Appendix D: Tables

1. PCA tables

TABLE II. Principal components composition in the simplest case: with degree, clustering coefficient and betweenness centrality. LAU list, $ws = 1000$ messages in 20 disjoint positioning was used for statistics. First component is a pondered sum of degree and betweenness centrality measures. Second component is mostly clustering coefficient. First and second components sum more than 95% of total dispersion.

	PC1		PC2		PC3	
	μ	σ	μ	σ	μ	σ
d	**48.02**	1.39	2.82	1.74	48.09	0.32
cc	4.12	2.94	**90.45**	3.98	3.98	0.77
bt	**47.87**	1.55	6.74	4.08	47.93	0.46
λ	64.67	0.52	33.26	0.23	2.08	0.40

TABLE III. Principal components composition in percentages. LAU list, $ws = 1000$ messages in 20 disjoint positioning was used for statistics. First component is a pondered sum of degree and strength and betweenness centrality measures. Second component is mostly clustering coefficient. First and second components sum more than 90% of dispersion.

	PC1		PC2		PC3	
	μ	σ	μ	σ	μ	σ
d	**14.58**	0.14	0.43	0.35	1.51	1.08
d^{in}	**14.12**	0.14	1.71	1.22	17.80	6.20
d^{out}	**13.95**	0.12	2.80	1.83	21.15	5.62
s	**14.48**	0.13	0.78	0.65	5.51	4.71
s^{in}	**14.10**	0.14	2.17	1.28	17.32	6.11
s^{out}	**14.05**	0.13	2.08	1.14	19.31	4.86
cc	0.99	0.70	**83.38**	4.83	2.75	1.62
bt	**13.73**	0.19	6.65	1.31	14.66	10.14
λ	81.80	0.83	12.53	0.09	3.24	0.62

TABLE IV. Principal components formation with measures of symmetry described (see Section III B 2). LAU list, $ws = 1000$ messages in 20 disjoint positioning was used for statistics. In this case, clusterization is pushed to third components. Second component is primarily symmetry measures, but also out degree and strength, and disequilibrium standard deviation. Betweenness centrality again has a role similar to degree, but weaker. Clusterization component combines with disequilibrium, while asymmetry is combined to out degree and strength. Three components has in average 80.36% of dispersion.

	PC1		PC2		PC3	
	μ	σ	μ	σ	μ	σ
d	**11.51**	0.42	2.00	0.76	2.39	0.49
d^{in}	**11.45**	0.34	2.86	0.91	1.68	0.67
d^{out}	**10.68**	0.60	**7.43**	1.00	3.00	1.02
s	**11.37**	0.42	1.75	0.71	4.31	0.63
s^{in}	**11.33**	0.35	2.39	1.10	3.69	0.86
s^{out}	**10.74**	0.55	**6.14**	1.05	4.75	0.98
cc	0.91	0.64	2.68	1.67	**22.27**	6.43
bt	**10.87**	0.38	1.17	0.93	4.03	1.42
asy	3.99	1.45	**18.13**	1.67	2.55	1.77
μ_{asy}	4.15	1.40	**17.07**	1.78	2.49	1.67
σ_{asy}	1.21	0.67	**17.49**	0.79	3.29	2.33
dis	5.78	0.51	1.94	1.28	**24.75**	3.73
μ_{dis}	0.79	0.49	**14.00**	1.14	3.73	3.13
σ_{dis}	5.18	0.72	4.93	2.48	**17.04**	4.78
λ	51.09	1.07	20.04	1.31	9.23	6.63

2. Tables for activity along time

TABLE V. Percentage of activity ($\frac{\text{counted messages}}{\text{total messages}}$) in each hour, 6 hours and 12 hours. Maximum activity rates are in bold. In 1h columns, minimum activity is also bold. The less active period of the day is around 4-6h. Maximum activity is between 10-13h. Afternoon is most active in 6h division of the day. The noon has $\approx \frac{2}{3}$ of 24h activity.

	CPP			MET			LAU			LAD		
	1h	6h	12h	1h	6h	12h	1h	6h	12h	1h	6h	12h
0h	3.66			2.87			3.58			4.00		
1h	2.76			1.77			2.22			2.52		
2h	1.79	10.67		1.04	7.15		1.63	10.14		1.79	10.77	
3h	1.10			0.64			1.06			1.06		
4h	**0.68**			0.47			0.84			0.75		
5h	0.69		33.76	**0.38**		29.33	**0.82**		36.88	**0.66**		33.13
6h	0.83			0.72			1.17			0.85		
7h	1.24			1.33			2.37			1.56		
8h	2.28	23.09		2.67	22.18		3.54	26.74		2.96	22.36	
9h	4.52			4.40			6.04			4.68		
10h	6.62			6.29			**6.83**			5.93		
11h	**7.61**			6.78			6.79			6.40		
12h	6.44			**7.33**			6.11			**6.41**		
13h	6.04			7.08			6.26			6.12		
14h	6.47	**37.63**		7.09	**42.22**		6.38	**35.65**		6.33	**37.25**	
15h	6.10			7.14			5.93			5.98		
16h	6.22			6.68			5.52			6.40		
17h	6.36		**66.24**	6.89		**70.66**	5.46		**63.12**	6.02		**66.87**
18h	6.01			5.99			5.24			5.99		
19h	5.02			5.23			4.52			5.03		
20h	4.85	28.61		4.98	28.44		4.55	27.46		4.63	29.63	
21h	4.38			4.37			4.42			4.59		
22h	4.06			4.24			4.51			4.88		
23h	4.30			3.64			4.23			4.53		

TABLE VI. Concentration of activity on days along the week. Weekend days are at least $\frac{1}{3}$ less active and can reach $\frac{1}{3}$ of activity. MET concentrates activity in weekdays the most, leaving only 13.98% of total activity to Saturday and Sunday. LAU is the one that less concentrates activity in weekdays, reaching 20.94% of total activity in weekends. These might suggest professional relation of CPP and MET participants to the topics of interest, or a hobby relation of LAU and LAD participants.

	Mon	Tue	Wed	Thu	Fri	Sat	Sun
CPP	17.06	17.43	17.61	17.13	16.30	6.81	7.67
MET	17.53	17.54	16.43	17.06	17.46	7.92	6.06
LAU	15.71	15.80	15.88	16.43	15.13	10.13	10.91
LAD	14.91	17.73	17.01	15.40	14.25	10.39	10.30

TABLE VII. Activity along the days of the month. The pattern is to have no clear prevalent period. One might point a slight tendency for the first two weeks to be more active, although this table does not present statistical foundation for such an assumption. For the scope of this study, differences of activity along the month is assumed to be inexistent.

	CPP			MET			LAU			LAD		
day	1 day	7 days	14 days	1 day	7 days	14 days	1 day	7 days	14 days	1 day	7 days	14 days
1	3.19			3.01			3.34			3.22		
2	3.07			3.38			3.38			3.42		
3	3.20			3.55			3.20			2.87		
4	3.63	23.05		4.34	25.16		3.52	23.06		2.91	21.96	
5	2.85			3.93			2.68			3.30		
6	3.67			3.76			3.18			3.52		
7	3.45			3.18			3.77			2.27		
8	3.12		45.63	3.36		48.08	3.62		47.31	3.72		46.70
9	2.57			3.44			3.82			3.97		
10	2.92			3.17			3.06			3.77		
11	3.54	22.57		3.88	22.92		3.11	24.25		3.27	24.73	
12	3.23			2.94			3.40			2.75		
13	3.39			3.29			3.55			3.34		
14	3.81			2.83			3.69			3.93		
15	3.35			2.72			3.23			3.37		
16	3.77			2.96			2.94			3.37		
17	3.45			3.01			3.02			2.95		
18	3.47	23.02		3.39	21.87		3.63	22.84		3.22	22.82	
19	2.90			3.42			3.16			3.59		
20	2.80			3.09			3.25			3.21		
21	3.29			3.27			3.61			3.13		
22	2.88		46.31	2.92		43.56	3.80		44.01	3.07		46.00
23	4.01			3.27			3.03			3.06		
24	3.13			2.92			2.31			2.72		
25	3.57	23.29		2.83	21.69		2.38	21.17		3.16	23.18	
26	3.27			2.97			3.49			3.57		
27	3.27			3.41			2.92			3.92		
28	3.17			3.36			3.26			3.69		
29	3.68			2.93			3.34			3.15		
30	2.76	8.06	8.06	3.14	8.36	8.36	3.75	8.68	8.68	2.71	7.30	7.30
31	1.63			2.29			1.60			1.45		

TABLE VIII. Activity along the year, in months, trimesters, quadrimesters and semesters. Engagement in list participation seem to concentrate in two periods: middle of the year (Jun-Aug, lists MET and LAD), and transition from years (Dec-Mar, lists CPP, LAU and LAD). Messages were considered as to complete 12 months slots, so every month has the same time of occurrences.

	CPP					MET					LAU					LAD				
	m.	b.	t.	q.	s.	m.	b.	t.	q.	s.	m.	b.	t.	q.	s.	m.	b.	t.	q.	s.
Jan	8.70	17.00	**27.23**	**36.48**	**54.26**	4.88	11.01	16.90	23.32	47.74	10.22	**19.56**	**28.23**	**35.09**	49.17	11.23	18.49	26.43	36.04	**57.95**
Fev	8.29					6.13					9.34					7.26				
Mar	**10.23**	**19.49**				5.89	12.31				8.67	15.52				7.94	17.55			
Apr	9.26		27.03			6.42		30.84			6.85		20.94			9.61		**31.51**		
Mai	9.41	17.78		33.46		10.46	**24.42**		**47.83**		7.27	14.09		30.37		8.94	**21.91**		**37.56**	
Jun	8.37					**13.96**					6.81					**12.97**				
Jul	8.70	15.68	22.94		45.73	13.23	23.41	**31.16**		**52.26**	8.96	16.28	24.47		**50.82**	9.02	15.65	22.29		42.05
Ago	6.98					10.28					7.31					6.63				
Set	7.26	15.36		30.06		7.75	16.80		28.86		8.18	16.24		34.54		6.63	12.38		26.40	
Oct	8.10		22.80			9.05		21.10			8.06		26.36			5.74		19.77		
Nov	7.86	14.69				7.46	12.06				7.63	18.30				7.63	14.02			
Dec	6.81					4.59					**10.66**					6.39				

TABLE IX. Distribution of activity among agents. First column is dedicated to percentage of messages sent by the most active participant. Column for the first quartile ($1Q$) exhibits minimum percentage of participants responsible for at least 25% of total messages. Similarly, the column for the first three quartiles $1-3Q$ exhibits minimum percentage of participants responsible for 75% of total messages. The last decile $10D$ column has maximum percentage of participants responsible for 10% of activity (messages).

list	hub	$1Q$	$1-3Q$	$10D$
CPP	14.41	0.19 (27.8%)	4.09 (75.13%)	83.65 (-10.04%)
MET	11.14	0.81 (30.61%)	8.33 (75,11%)	80.49 (-10.02%)
LAU	2.78	1.10 (25.16%)	13.02 (75,04%)	67.37 (-10.03%)
LAD	4.00	0.95 (25.50%)	11.83 (75,07%)	71.13 (-10.03%)

Appendix E: Figures of vertex classification fractions as the network evolves

Two lists are exhibited in this section, CPP and LAD. These structures are very similar in all four lists and laying extensively all figures is redundant. Window sizes of $ws = 10000$, 5000, 1000, 500, 250, 100 and 50 messages were used.

[1] Gergely Palla, Albert-László Barabási, and Tamás Vicsek, "Quantifying social group evolution," Nature **446**, 664–667 (2007).

[2] Elizabeth A Leicht, Gavin Clarkson, Kerby Shedden, and Mark EJ Newman, "Large-scale structure of time evolving citation networks," The European Physical Journal B **59**, 75–83 (2007).

[3] M. E. J. Newman, "The structure and function of complex networks," SIAM REVIEW **45**, 167–256 (2003).

[4] Mark EJ Newman, "Analysis of weighted networks," Physical Review E **70**, 056131 (2004).

[5] Kyle Marek-Spartz, Paula Chesley, and Hannah Sande, "Construction of the gmane corpus for examining the diffusion of lexical innovations," (2012).

[6] Renato Fabbri, "A connective differentiation of textual production in interaction networks," (2013), http://arxiv.org/abs/1412.7309.

[7] Renato Fabbri, "Participant typologies derived from textual and topological features in interaction networks," (2013).

[8] Patrick Doreian and Frans Stokman, *Evolution of social networks* (Routledge, 2013).

[9] Réka Albert and Albert-László Barabási, "Topology of evolving networks: local events and universality," Physical review letters **85**, 5234 (2000).

[10] Christian Bird, Alex Gourley, Prem Devanbu, Michael Gertz, and Anand Swaminathan, "Mining email social networks," in *Proceedings of the 2006 international workshop on Mining software repositories* (ACM, 2006) pp. 137–143.

[11] Lars Magne Ingebrigtsen, "Gmane," (2008).

[12] Wikipedia, "Gmane — Wikipedia, the free encyclopedi," .

[13] Gmane.linux.audio.users is list ID in GMANE.

[14] Gmane.linux.audio.devel is list ID in GMANE.

[15] Gmane.comp.gcc.libstdc++.devel is list ID in GMANE.

[16] Gmane.politics.organizations.metareciclagem is list ID in GMANE.

[17] Elizabeth A Leicht and Mark EJ Newman, "Community structure in directed networks," Physical review letters **100**, 118703 (2008).

[18] MEJ Newman, "Community detection and graph partitioning," arXiv preprint arXiv:1305.4974(2013).

[19] L da F Costa, Francisco A Rodrigues, Gonzalo Travieso, and PR Villas Boas, "Characterization of complex networks: A survey of measurements," Advances in Physics **56**, 167–242 (2007).

[20] BAN Travençolo and L da F Costa, "Accessibility in complex networks," Physics Letters A **373**, 89–95 (2008).

[21] Matthew O. Jackson.

[22] Ulrik Brandes, "A faster algorithm for betweenness centrality*," Journal of Mathematical Sociology **25**, 163–177 (2001).

[23] Renato Fabbri, Luciano da F. Costa, and Osvaldo N. de Oliveira jr, "Video visualizations of email interaction network evolution," (2013), http://www.youtube.com/watch?v=-t5jxQ8cKxM&list=PLf_EtaMqu3jU-1j4jiIUiyMqyVSzIYeh6.

[24] Renato Fabbri, Luciano da F. Costa, and Osvaldo N. de 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/.

[25] Renato Fabbri, Luciano da F. Costa, and Osvaldo N. de Oliveira jr, "Online gadget for making email interaction network images, gml files and measurements.." (2013), http://hera.ethymos.com.br:1080/redes/python/autoRede/escolheRedes.php.

[26] Radius is the minimum maximum distance to all nodes. Equivalently, the radius is the minimum eccentricity.

[27] Diameter is the maximum geodesic on the network.

[28] NetworkX Developers, "Networkx," (2010).

[29] Alexei Vázquez, João Gama Oliveira, Zoltán Dezsö, Kwang-Il Goh, Imre Kondor, and Albert-László Barabási, "Modeling bursts and heavy tails in human dynamics," Physical Review E **73**, 036127 (2006).

[30] Julián Candia, Marta C González, Pu Wang, Timothy Schoenharl, Greg Madey, and Albert-László Barabási, "Uncovering individual and collective human dynamics from mobile phone records," Journal of Physics A: Mathematical and Theoretical **41**, 224015 (2008).

[31] Some of them related to Nazism adoption, antisemitism and potential fascists.

[32] Carl Gustav Jung, HG Baynes, and RFC Hull, *Psychological types*, Vol. 4 (Routledge London, UK, 1991).

[33] Naomi L Quenk, *Essentials of Myers-Briggs type indicator assessment*, Vol. 66 (Wiley. com, 2009).

[34] Sigmund Freud, "Libidinal types.." The Psychoanalytic Quarterly(1932).

[35] Hans J Eysenck, "Types of personality: a factorial study of seven hundred neurotics," The British Journal of Psychiatry **90**, 851–861 (1944).

[36] Numpy version 1.6.1, "random.randint" function.

[37] Stefano Boccaletti, Vito Latora, Yamir Moreno, Martin Chavez, and D-U Hwang, "Complex networks: Structure and dynamics," Physics reports **424**, 175–308 (2006).

[38] Ian Jolliffe, *Principal component analysis* (Wiley Online Library, 2005).

[39] Renato Fabbri, Rodrigo Bandeira de Luna, Ricardo Augusto Poppi Martins, *et al.*, "Social participation ontology: community documentation, enhancements and use examples," arXiv preprint arXiv:1501.02662(2015).

[40] *Produto 5 da consultoria PNUD/ONU de Renato Fabbri*, https://github.com/ttm/pnud4/blob/master/latex/produto.pdf?raw=true% BibitemShutNoStop

[41] Renato Fabbri, "Ensaio sobre o auto-aproveitamento: um relato de investidas naturais na participa\ c {c}\~ ao social," arXiv preprint arXiv:1412.6868(2014).

[42] Theodor W Adorno, Else Frenkel-Brunswik, Daniel J Levinson, and R Nevitt Sanford, "The authoritarian personality.." (1950).

[43] Jukka-Pekka Onnela, Samuel Arbesman, Marta C González, Albert-László Barabási, and Nicholas A Christakis, "Geographic constraints on social network groups," PLoS one **6**, e16939 (2011).

[44] Tao Jia and Albert-László Barabási, "Control capacity and a random sampling method in exploring controllability of complex networks," Scientific reports **3** (2013).

[45] Yang-Yu Liu, Jean-Jacques Slotine, and Albert-László Barabási, "Control centrality and hierarchical structure in complex networks," Plos one **7**, e44459 (2012).

[46] Yang-Yu Liu, Jean-Jacques Slotine, and Albert-László Barabási, "Controllability of complex networks," Nature **473**, 167–173 (2011).

[47] Vasyl Palchykov, Kimmo Kaski, Janos Kertész, Albert-László Barabási, and Robin IM Dunbar, "Sex differences in intimate relationships," Scientific reports **2** (2012).

[48] Renato Fabbri, Luciano da F. Costa, and Osvaldo N. de Oliveira jr, "Scripts used for obtaining results used in this article .." (2013), sourceforge.net/p/labmacambira/fimDoMundo/ci/master/tree/python/toolkitGMANE/.

[49] Renato Fabbri, "Complex networks and natural language processing collection and diffusion of information and goods.." (2014), wiki.nosdigitais.teia.org.br/ARS.

[50] James P Bagrow, Dashun Wang, and Albert-Laszlo Barabasi, "Collective response of human populations to large-scale emergencies," PloS one **6**, e17680 (2011).

[51]Gourab Ghoshal, Nicholas Blumm, Zalan Forro, Maximilian Schich, Ginestra Bianconi, Jean-Philippe Bouchaud, and Albert-Laszlo Barabasi, "Dynamics of ranking processes in complex systems," (2012).
[52]Soon-Hyung Yook, Hawoong Jeong, A-L Barabási, and Yuhai Tu, "Weighted evolving networks," Physical Review Letters **86**, 5835 (2001).
[53]Dashun Wang, Zhen Wen, Hanghang Tong, Ching-Yung Lin, Chaoming Song, and Albert-László Barabási, "Information spreading in context," in *Proceedings of the 20th international conference on World wide web* (ACM, 2011) pp. 735–744.
[54]Nicholas Blumm, Gourab Ghoshal, Zalán Forró, Maximilian Schich, Ginestra Bianconi, Jean-Philippe Bouchaud, and Albert-László Barabási, "Dynamics of ranking processes in complex systems," Physical Review Letters **109**, 128701 (2012).
[55]Mark EJ Newman, Steven H Strogatz, and Duncan J Watts, "Random graphs with arbitrary degree distributions and their applications," Physical Review E **64**, 026118 (2001).
[56]Mark EJ Newman, "Random graphs with clustering," Physical review letters **103**, 058701 (2009).
[57]Aaron Clauset, Cosma Rohilla Shalizi, and Mark EJ Newman, "Power-law distributions in empirical data," SIAM review **51**, 661–703 (2009).
[58]Mark EJ Newman, "Assortative mixing in networks," Physical review letters **89**, 208701 (2002).
[59]Mark EJ Newman, "Modularity and community structure in networks," Proceedings of the National Academy of Sciences **103**, 8577–8582 (2006).
[60]MEJ Newman, "Communities, modules and large-scale structure in networks," Nature Physics **8**, 25–31 (2011).
[61]Aaron Clauset, Cristopher Moore, and Mark EJ Newman, "Hierarchical structure and the prediction of missing links in networks," Nature **453**, 98–101 (2008).
[62]MEJ Newman, "Complex systems: A survey," arXiv preprint arXiv:1112.1440(2011).
[63]Brian Ball and Mark EJ Newman, "Friendship networks and social status," arXiv preprint arXiv:1205.6822(2012).
[64]G. Deleuze, *Difference and Repetition* (Continuum, 1968).
[65]F. de Saussure, *Course in General Linguistics* (Books LLC, 1916).
[66]A. Papoulis S. U. Pillai, *Probability, Random Variables and Stochastic Processes* (McGraw Hill Higher Education, 2002).
[67]R. A. Johnson D. W. Wichern, *Applied Multivariate Statistical Analysis* (Prentice Hall, 2007).
[68]C. W. Therrien, *Discrete Random Signals and Statistical Signal Processing* (Prentice Hall, 1992).
[69]R. O. Duda P. E. Hart D. G. Stork, *Pattern Classification* (Wiley-Interscience, 2000).
[70]L. da F. Costa R. M. C. Jr., *Shape Analysis and Classification: Theory and Practice (Image Processing Series)* (CRC Press, 2000).
[71]D. Papineau, *Philosophy* (Oxford University Press, 2009).
[72]B. Russel, *A History of Western Philosophy* (Simon and Schuster Touchstone, 1967).
[73]F. G. G. Deleuze, *What Is Philosophy?* (Simon and Schuster Touchstone, 1991).

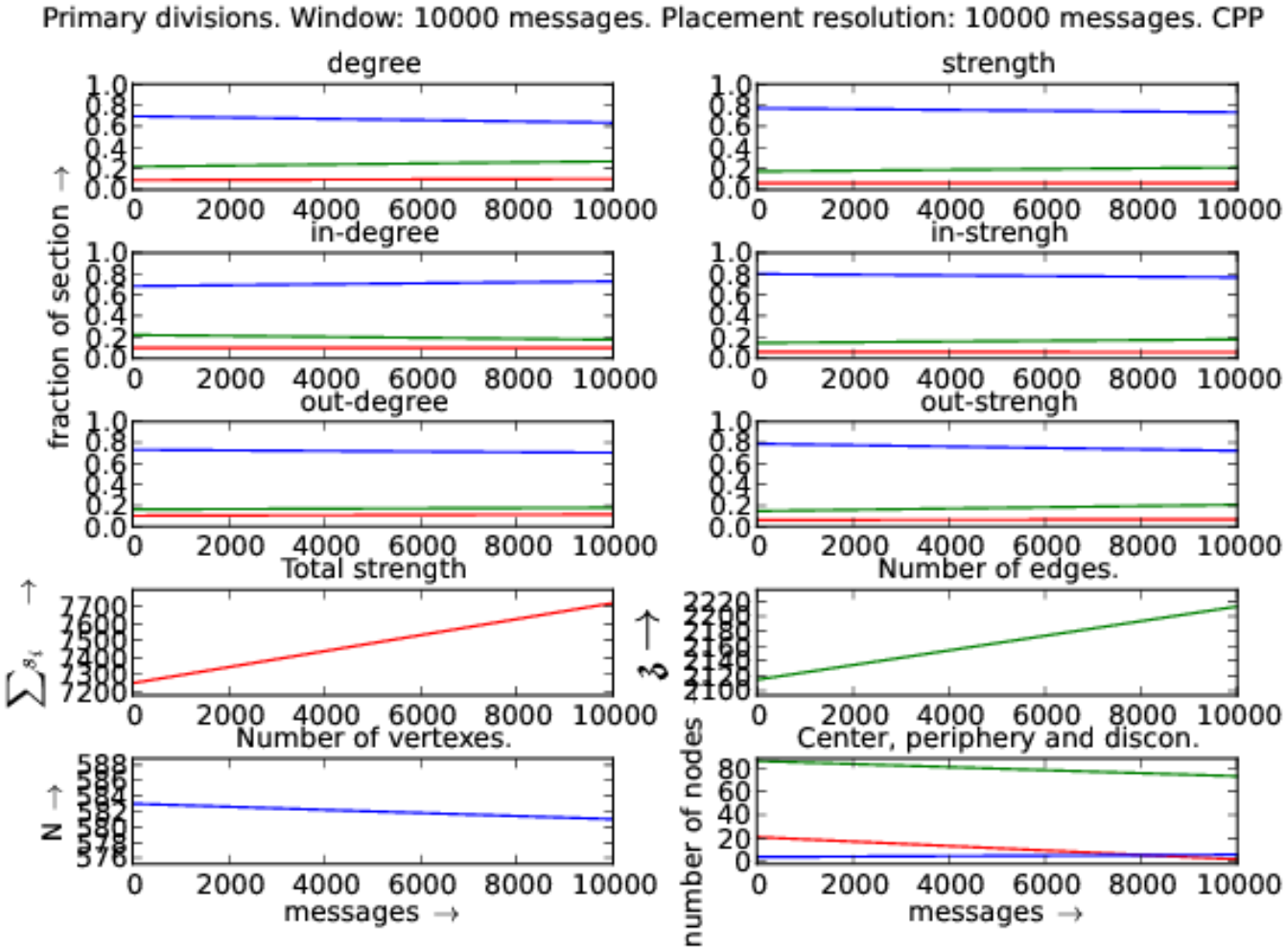

FIG. 6. Distribution of vertex with respect to each measure of activity: in and out degrees and strengths. CPP Std library official mailing list. In the first six plots, red is fraction of hubs, green is the fraction of intermediary and blue is for peripheral fraction. On the last plot, red is the center (maximum distance to another vertex is equal to radius), blue is periphery (maximum distance equals to diameter) of the giant component. On the same graph, green counts the disconnected vertexes.

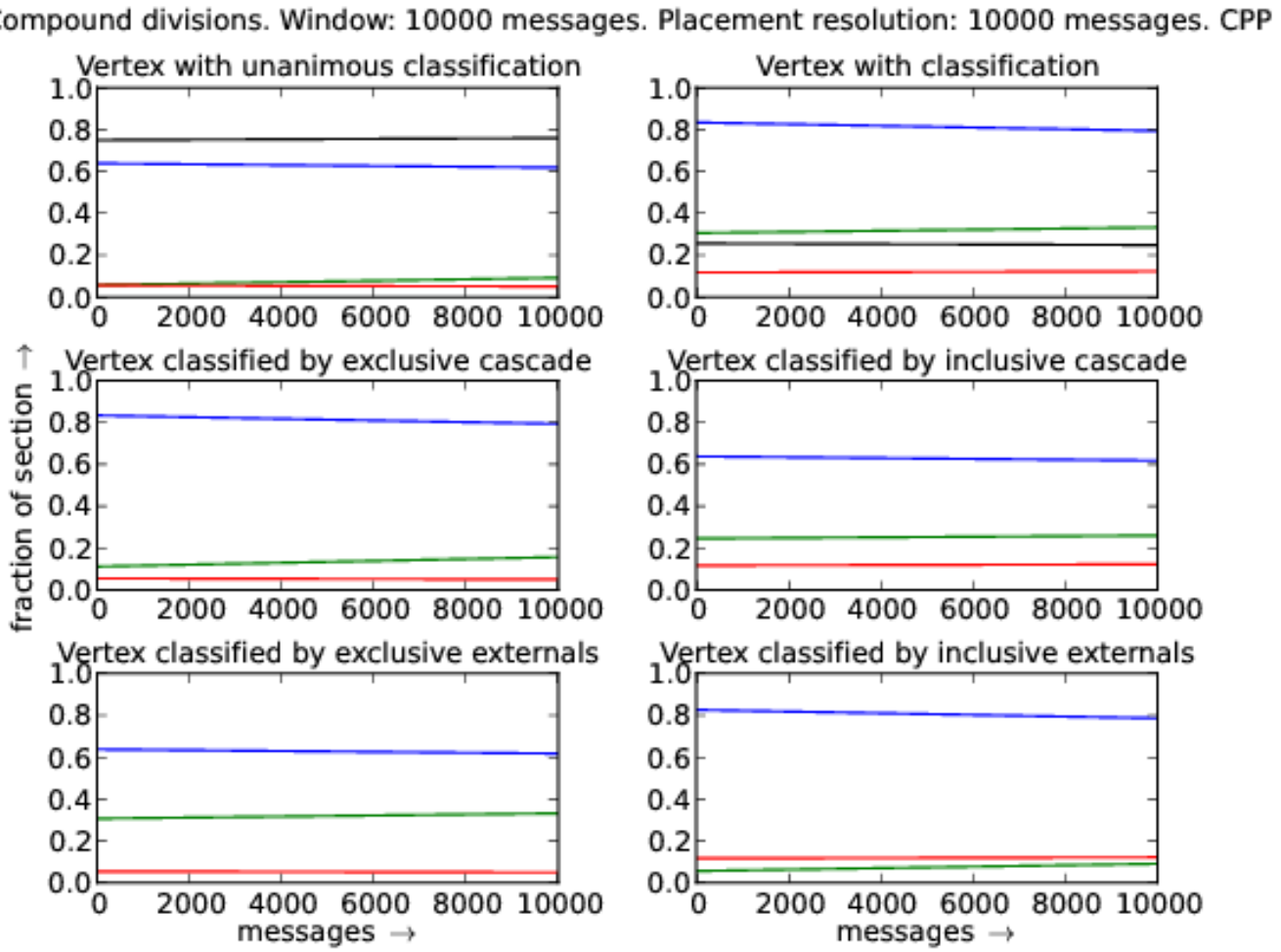

FIG. 7. Distribution of vertex with respect to each measure of activity: compound criteria. Red, green and blue are for hubs, intermediary and border (peripheral) vertex fractions. The first two plots picture classifications that are not functions. Thus, in the first plot, the fraction of vertexes with unique classification in plotted in black. On the second plot, black is used to depict how much a vertex was classified in more than section: $\frac{\text{number of classifications} - \text{number of nodes}}{\text{number of nodes}}$. Compound criteria is described in Section III B 1.

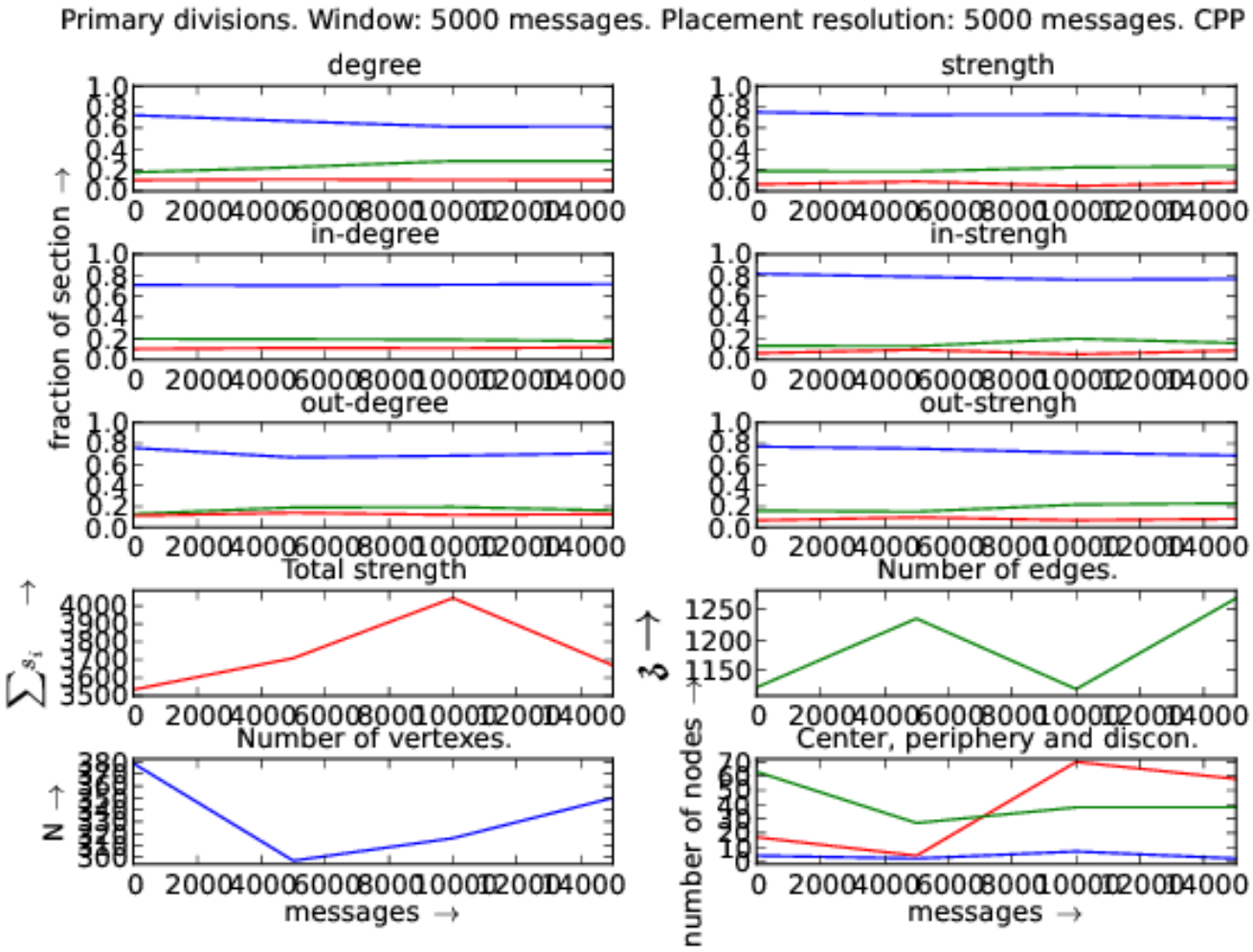

FIG. 8. Distribution of vertex with respect to each measure of activity: in and out degrees and strengths. CPP Std library official mailing list. In the first six plots, red is fraction of hubs, green is the fraction of intermediary and blue is for peripheral fraction. On the last plot, red is the center (maximum distance to another vertex is equal to radius), blue is periphery (maximum distance equals to diameter) of the giant component. On the same graph, green counts the disconnected vertexes.

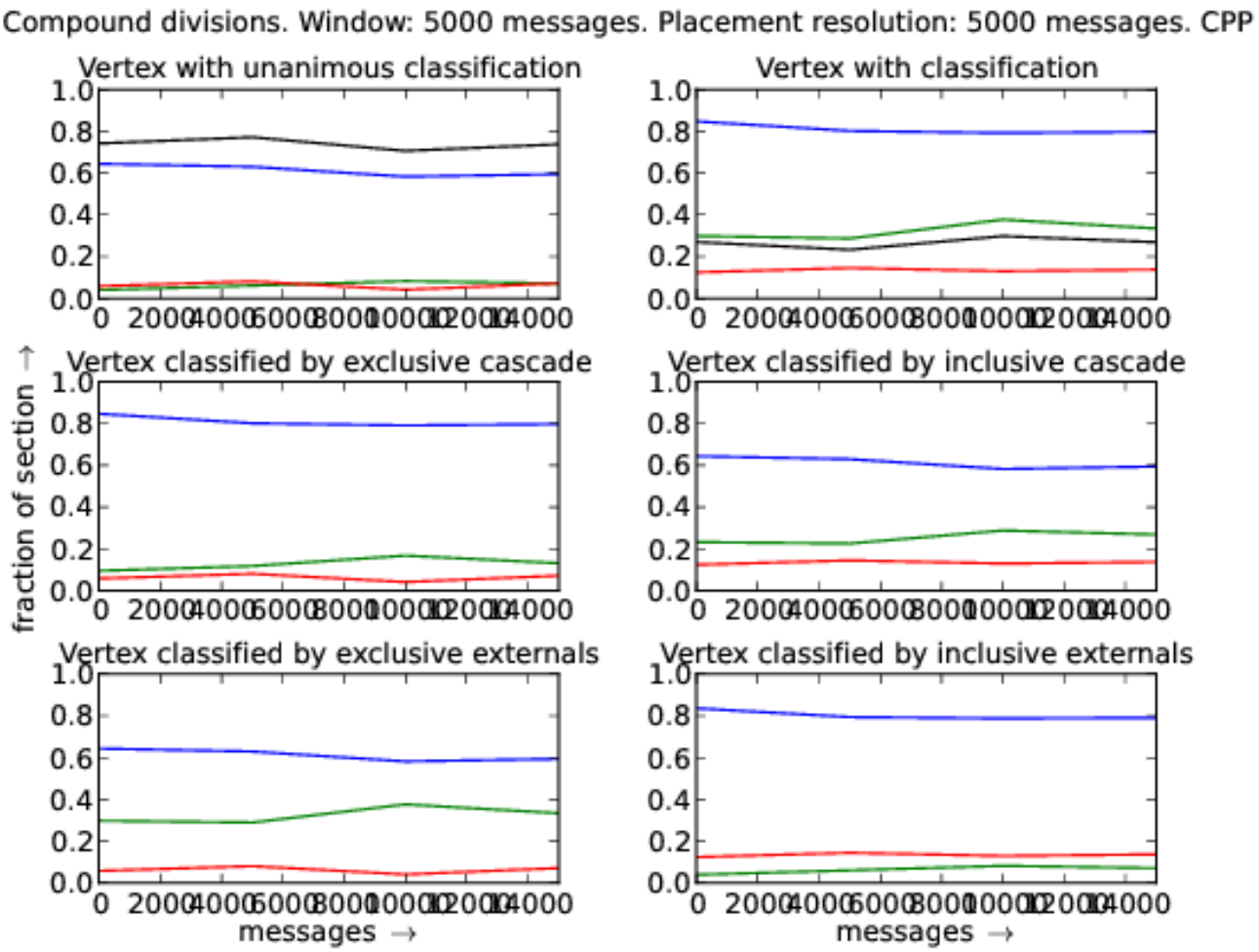

FIG. 9. Distribution of vertex with respect to each measure of activity: compound criteria. Red, green and blue are for hubs, intermediary and border (peripheral) vertex fractions. The first two plots picture classifications that are not functions. Thus, in the first plot, the fraction of vertexes with unique classification in plotted in black. On the second plot, black is used to depict how much a vertex was classified in more than section: $\frac{\text{number of classifications} - \text{number of nodes}}{\text{number of nodes}}$. Compound criteria is described in Section III B 1.

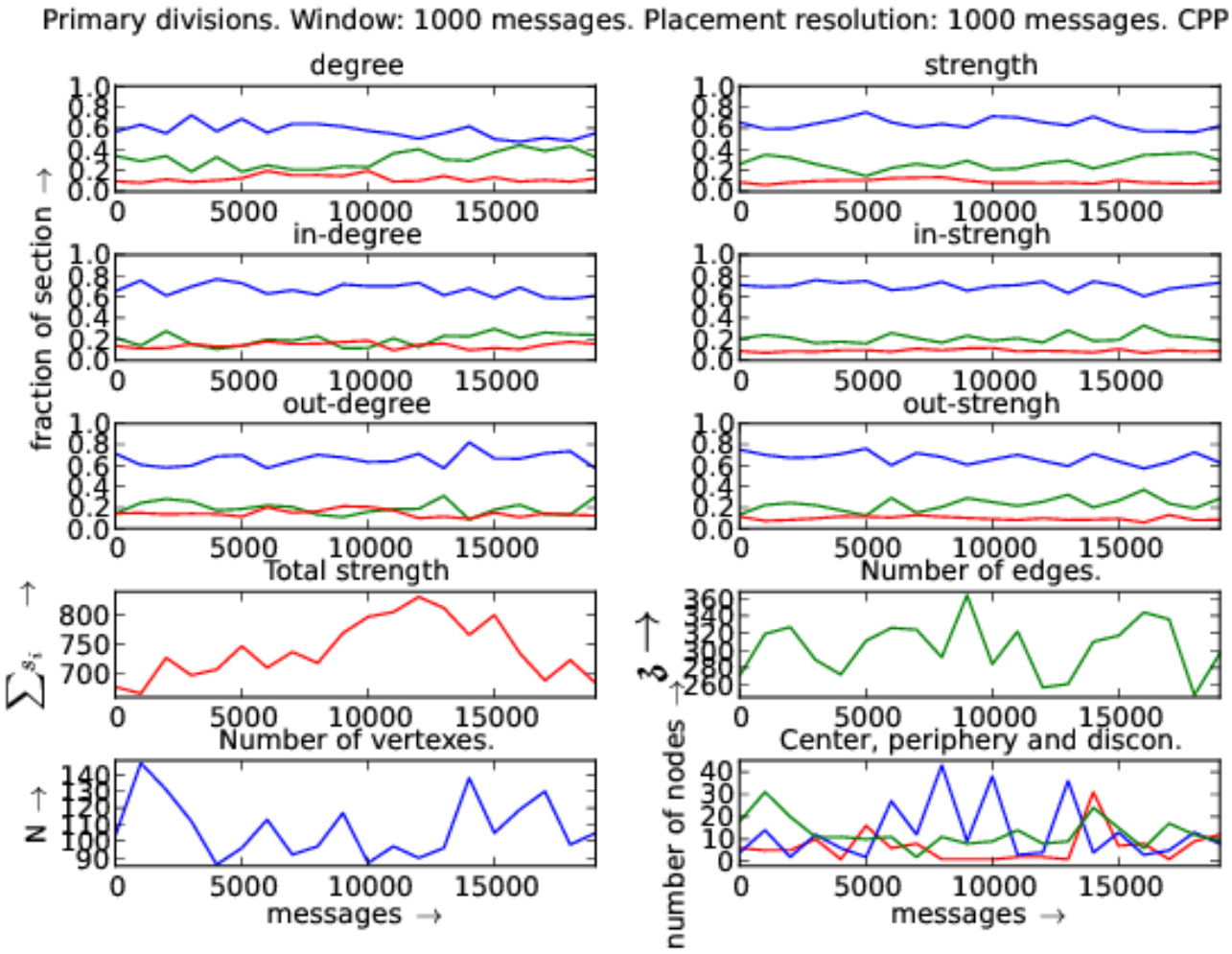

FIG. 10. Distribution of vertex with respect to each measure of activity: in and out degrees and strengths. CPP Std library official mailing list. In the first six plots, red is fraction of hubs, green is the fraction of intermediary and blue is for peripheral fraction. On the last plot, red is the center (maximum distance to another vertex is equal to radius), blue is periphery (maximum distance equals to diameter) of the giant component. On the same graph, green counts the disconnected vertexes.

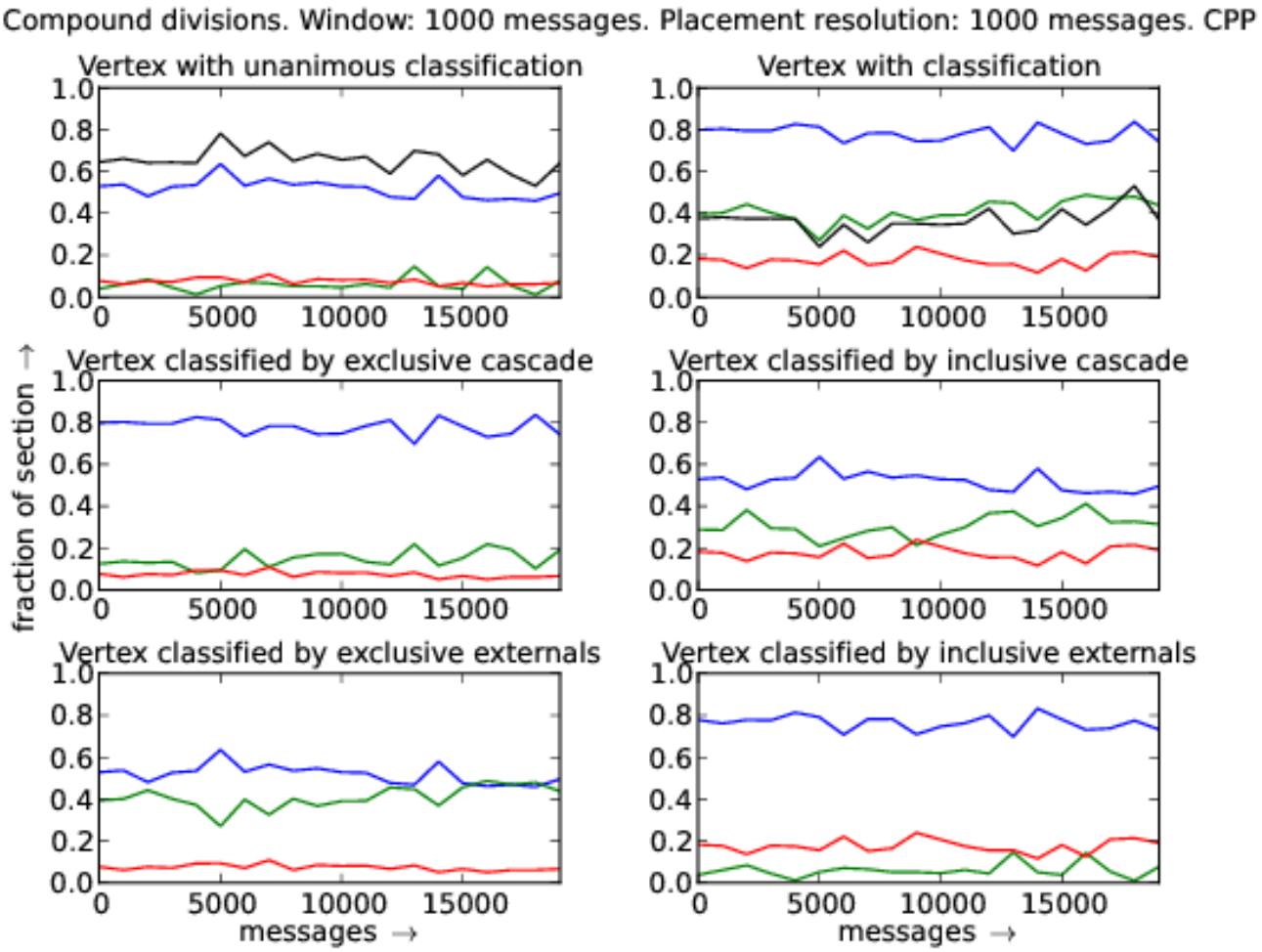

FIG. 11. Distribution of vertex with respect to each measure of activity: compound criteria. Red, green and blue are for hubs, intermediary and border (peripheral) vertex fractions. The first two plots picture classifications that are not functions. Thus, in the first plot, the fraction of vertexes with unique classification in plotted in black. On the second plot, black is used to depict how much a vertex was classified in more than section: $\frac{\text{number of classifications} - \text{number of nodes}}{\text{number of nodes}}$. Compound criteria is described in Section III B 1.

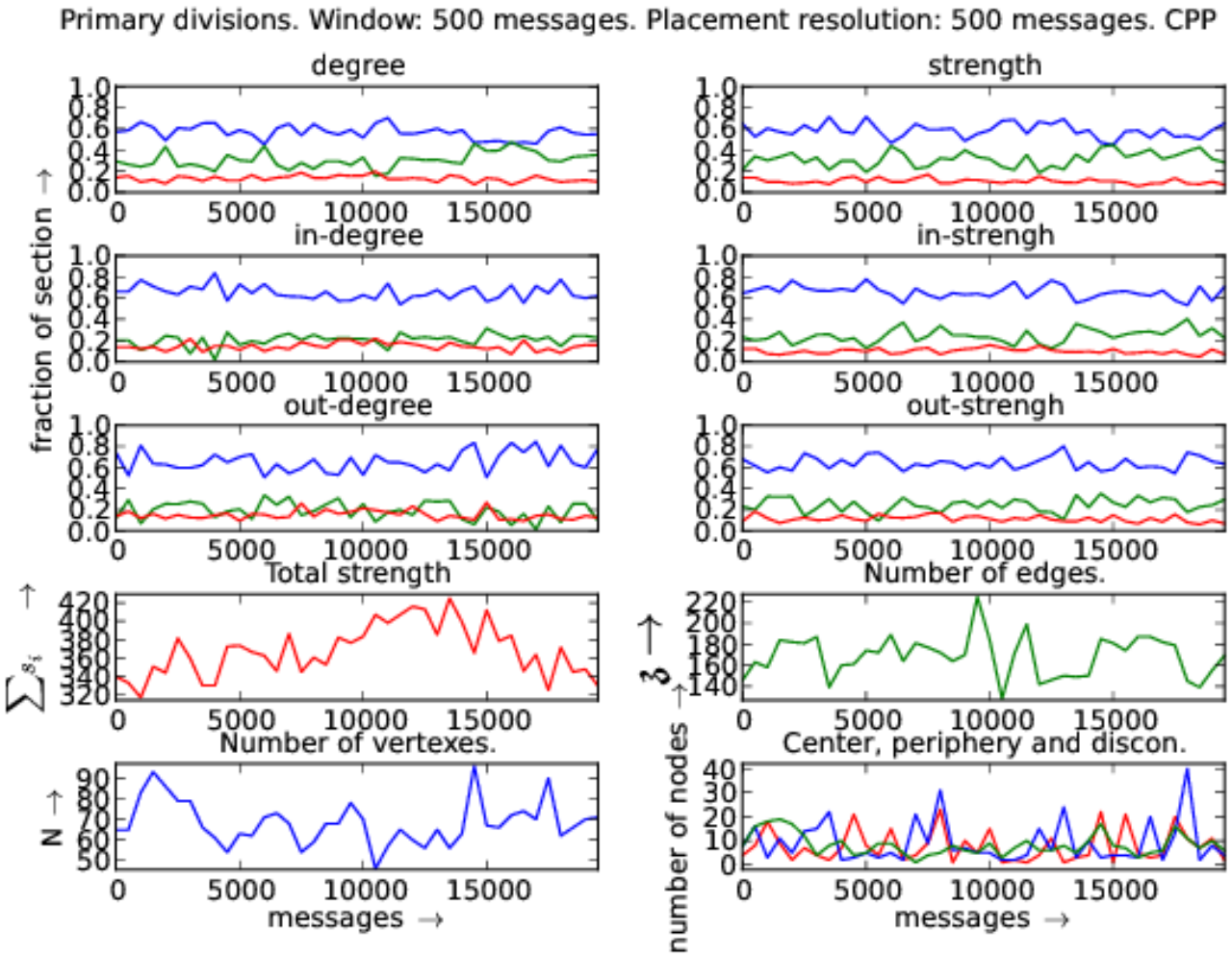

FIG. 12. Distribution of vertex with respect to each measure of activity: in and out degrees and strengths. CPP Std library official mailing list. In the first six plots, red is fraction of hubs, green is the fraction of intermediary and blue is for peripheral fraction. On the last plot, red is the center (maximum distance to another vertex is equal to radius), blue is periphery (maximum distance equals to diameter) of the giant component. On the same graph, green counts the disconnected vertexes.

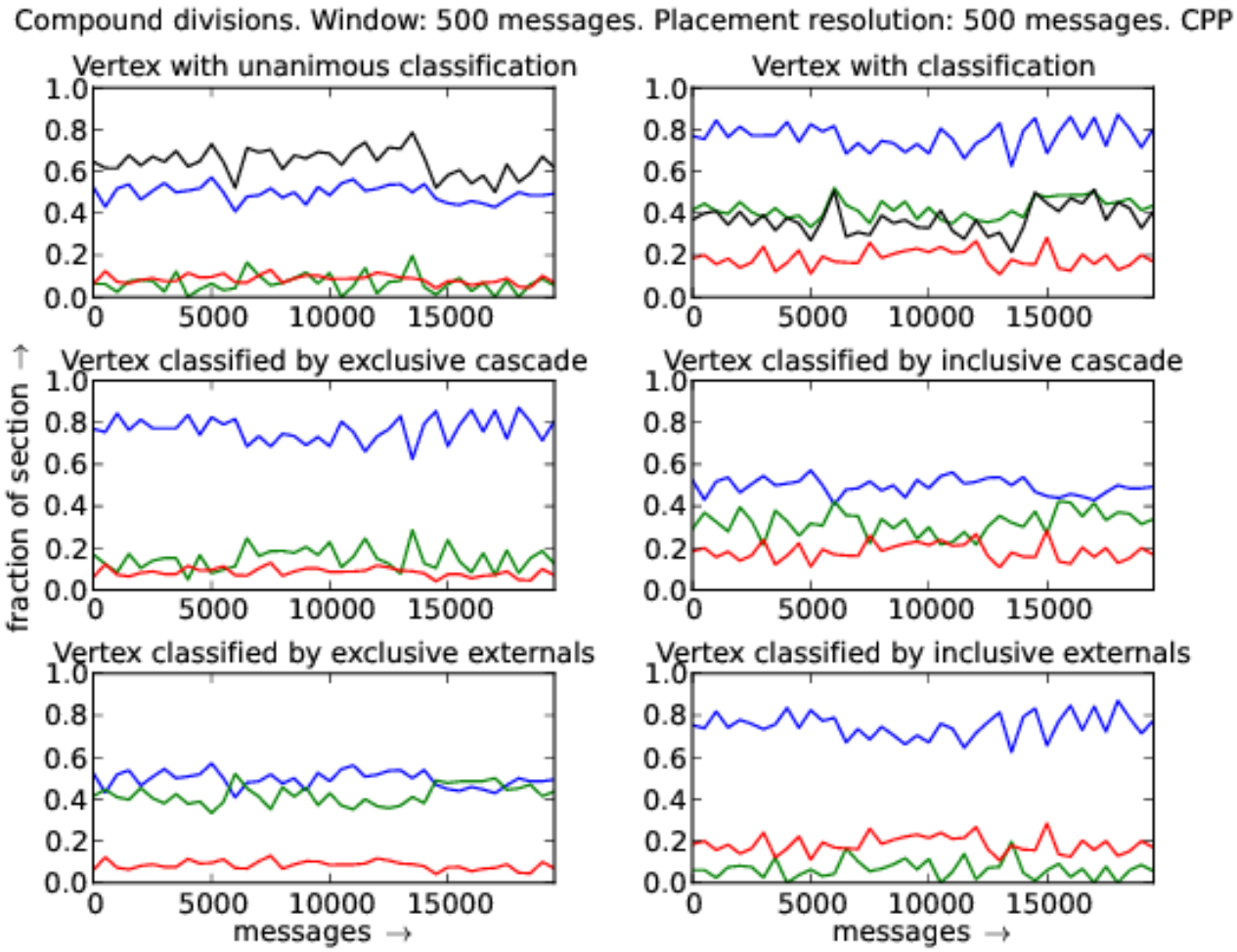

FIG. 13. Distribution of vertex with respect to each measure of activity: compound criteria. Red, green and blue are for hubs, intermediary and border (peripheral) vertex fractions. The first two plots picture classifications that are not functions. Thus, in the first plot, the fraction of vertexes with unique classification in plotted in black. On the second plot, black is used to depict how much a vertex was classified in more than section: $\frac{\text{number of classifications} - \text{number of nodes}}{\text{number of nodes}}$. Compound criteria is described in Section III B 1.

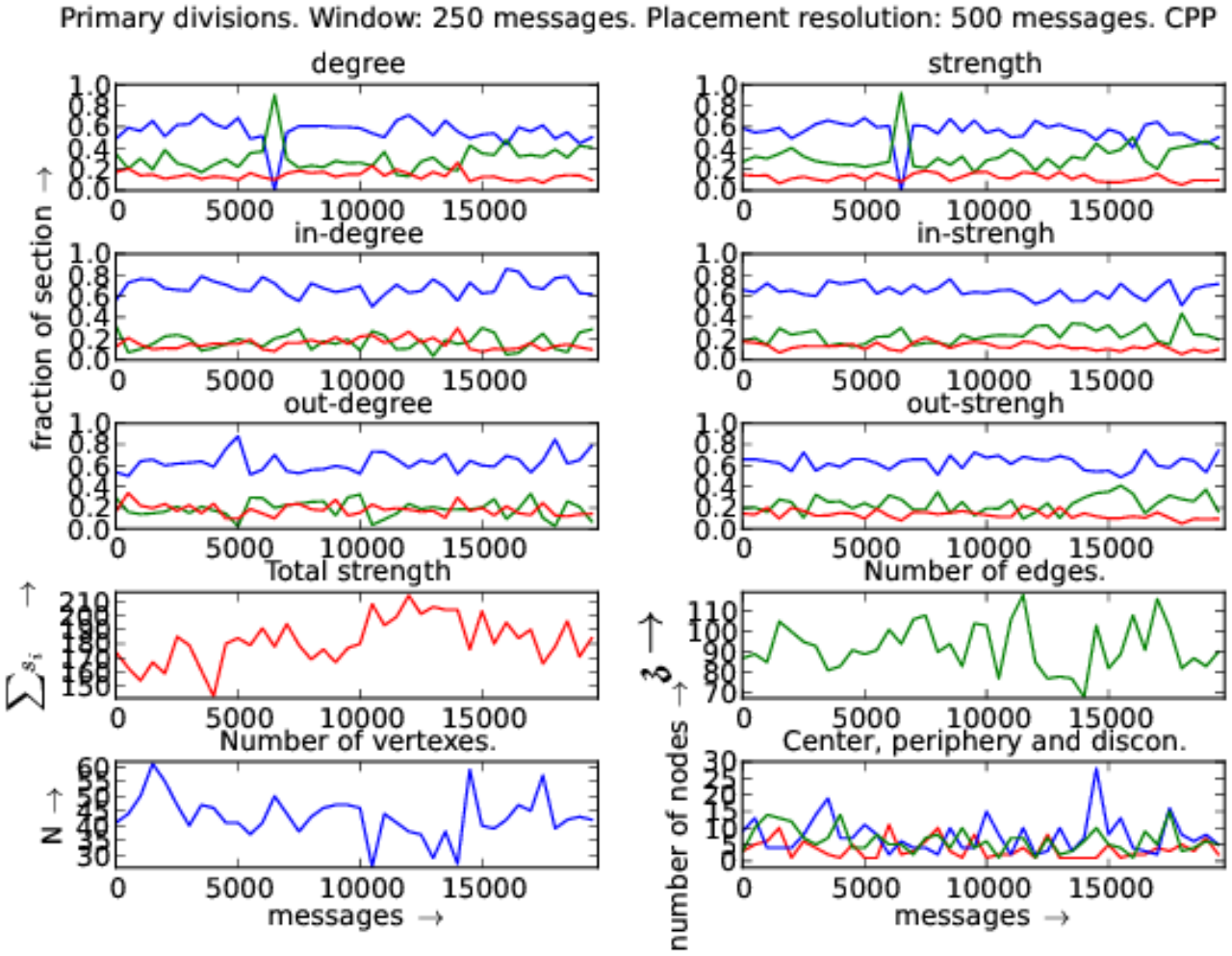

FIG. 14. Distribution of vertex with respect to each measure of activity: in and out degrees and strengths. CPP Std library official mailing list. In the first six plots, red is fraction of hubs, green is the fraction of intermediary and blue is for peripheral fraction. On the last plot, red is the center (maximum distance to another vertex is equal to radius), blue is periphery (maximum distance equals to diameter) of the giant component. On the same graph, green counts the disconnected vertexes.

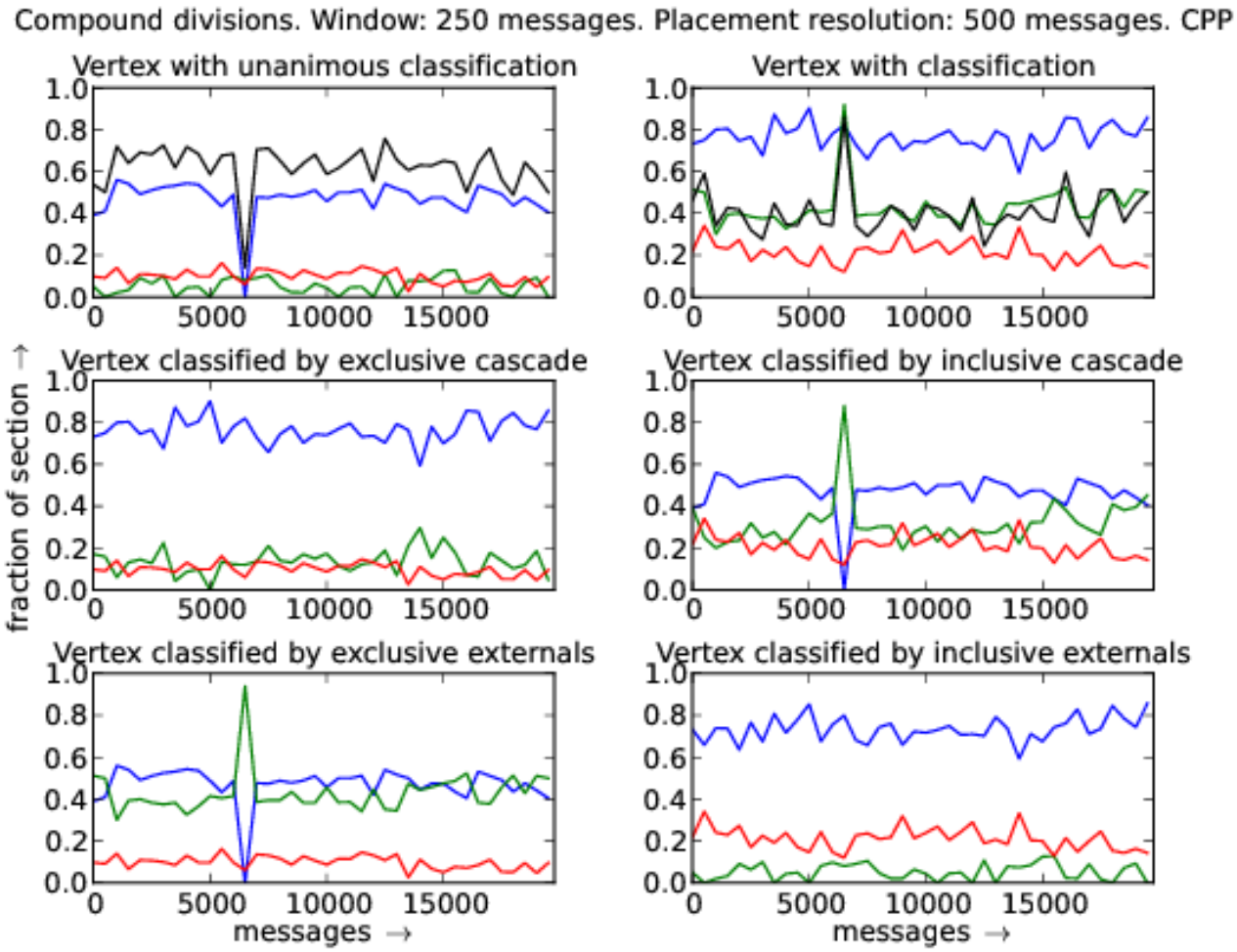

FIG. 15. Distribution of vertex with respect to each measure of activity: compound criteria. Red, green and blue are for hubs, intermediary and border (peripheral) vertex fractions. The first two plots picture classifications that are not functions. Thus, in the first plot, the fraction of vertexes with unique classification in plotted in black. On the second plot, black is used to depict how much a vertex was classified in more than section: $\frac{\text{number of classifications} - \text{number of nodes}}{\text{number of nodes}}$. Compound criteria is described in Section III B 1.

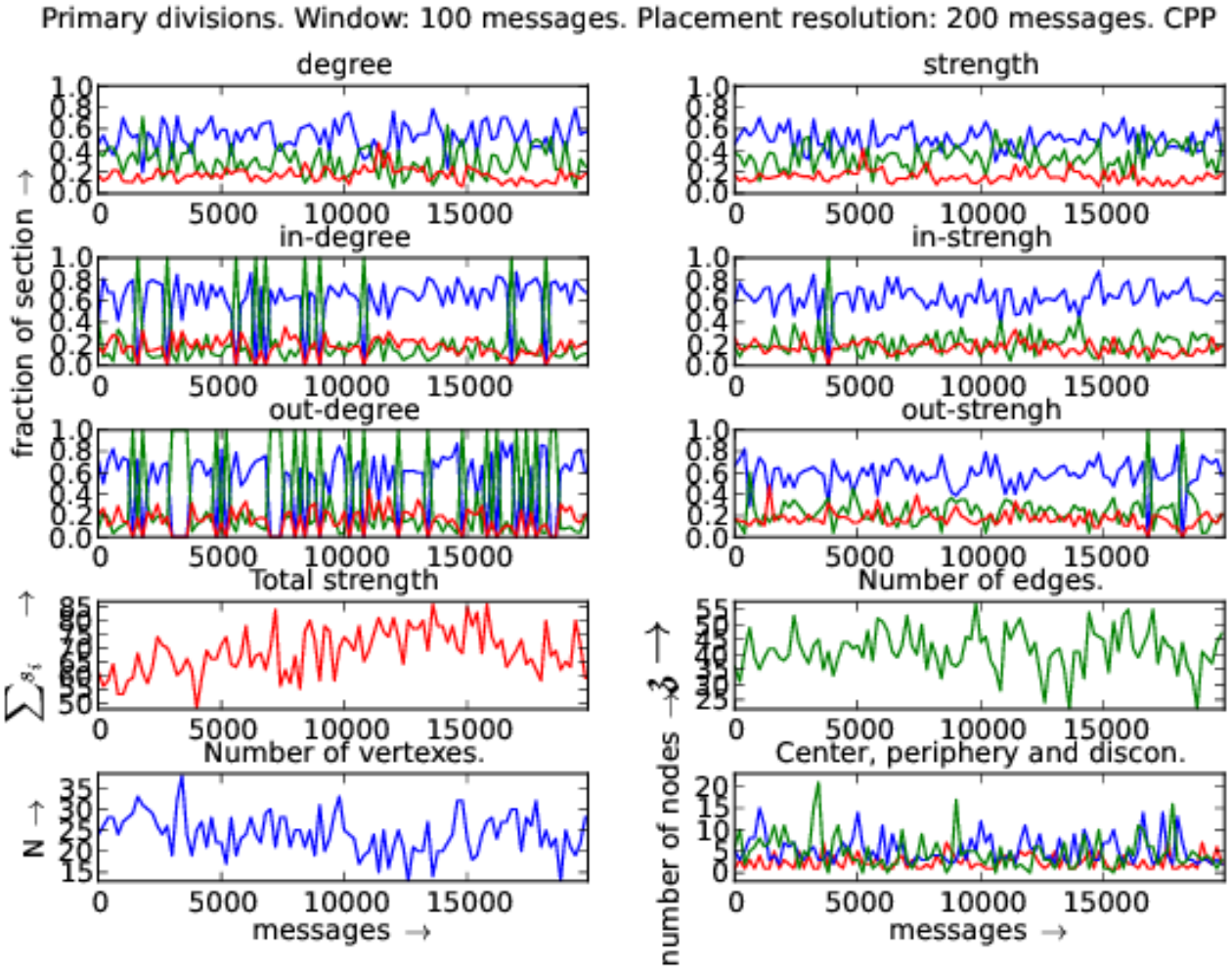

FIG. 16. Distribution of vertex with respect to each measure of activity: in and out degrees and strengths. CPP Std library official mailing list. In the first six plots, red is fraction of hubs, green is the fraction of intermediary and blue is for peripheral fraction. On the last plot, red is the center (maximum distance to another vertex is equal to radius), blue is periphery (maximum distance equals to diameter) of the giant component. On the same graph, green counts the disconnected vertexes.

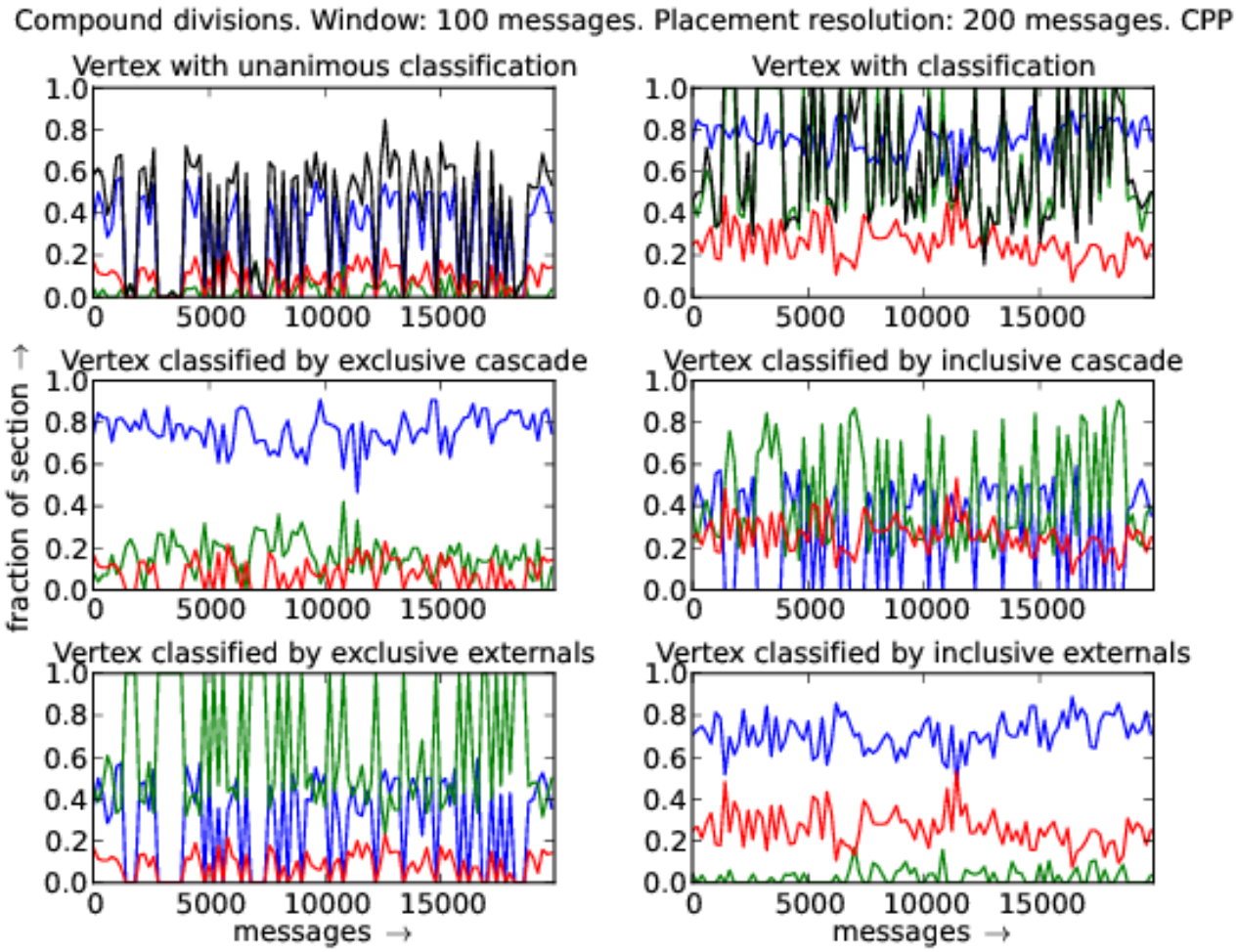

FIG. 17. Distribution of vertex with respect to each measure of activity: compound criteria. Red, green and blue are for hubs, intermediary and border (peripheral) vertex fractions. The first two plots picture classifications that are not functions. Thus, in the first plot, the fraction of vertexes with unique classification in plotted in black. On the second plot, black is used to depict how much a vertex was classified in more than section: $\frac{\text{number of classifications} - \text{number of nodes}}{\text{number of nodes}}$. Compound criteria is described in Section III B 1.

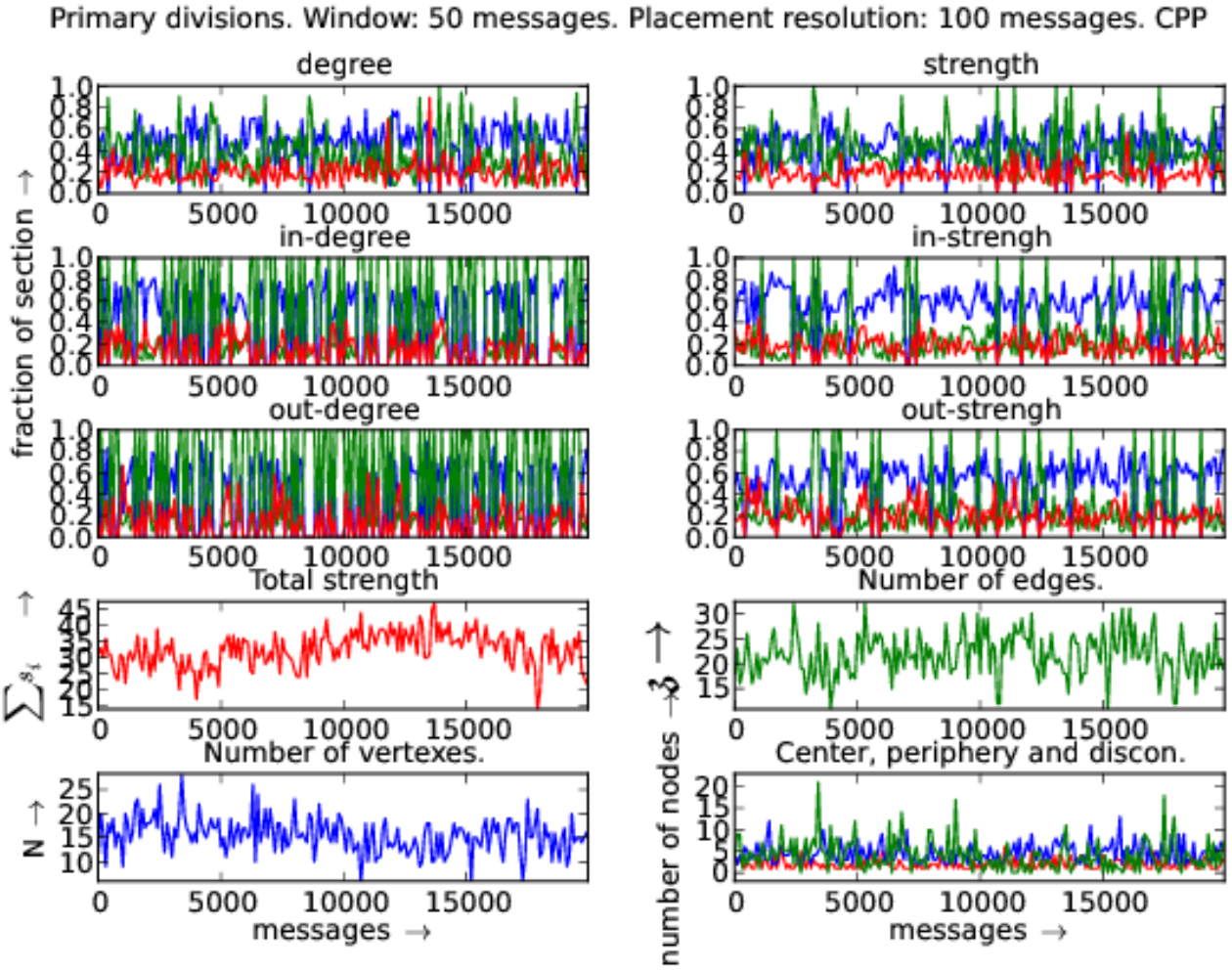

FIG. 18. Distribution of vertex with respect to each measure of activity: in and out degrees and strengths. CPP Std library official mailing list. In the first six plots, red is fraction of hubs, green is the fraction of intermediary and blue is for peripheral fraction. On the last plot, red is the center (maximum distance to another vertex is equal to radius), blue is periphery (maximum distance equals to diameter) of the giant component. On the same graph, green counts the disconnected vertexes.

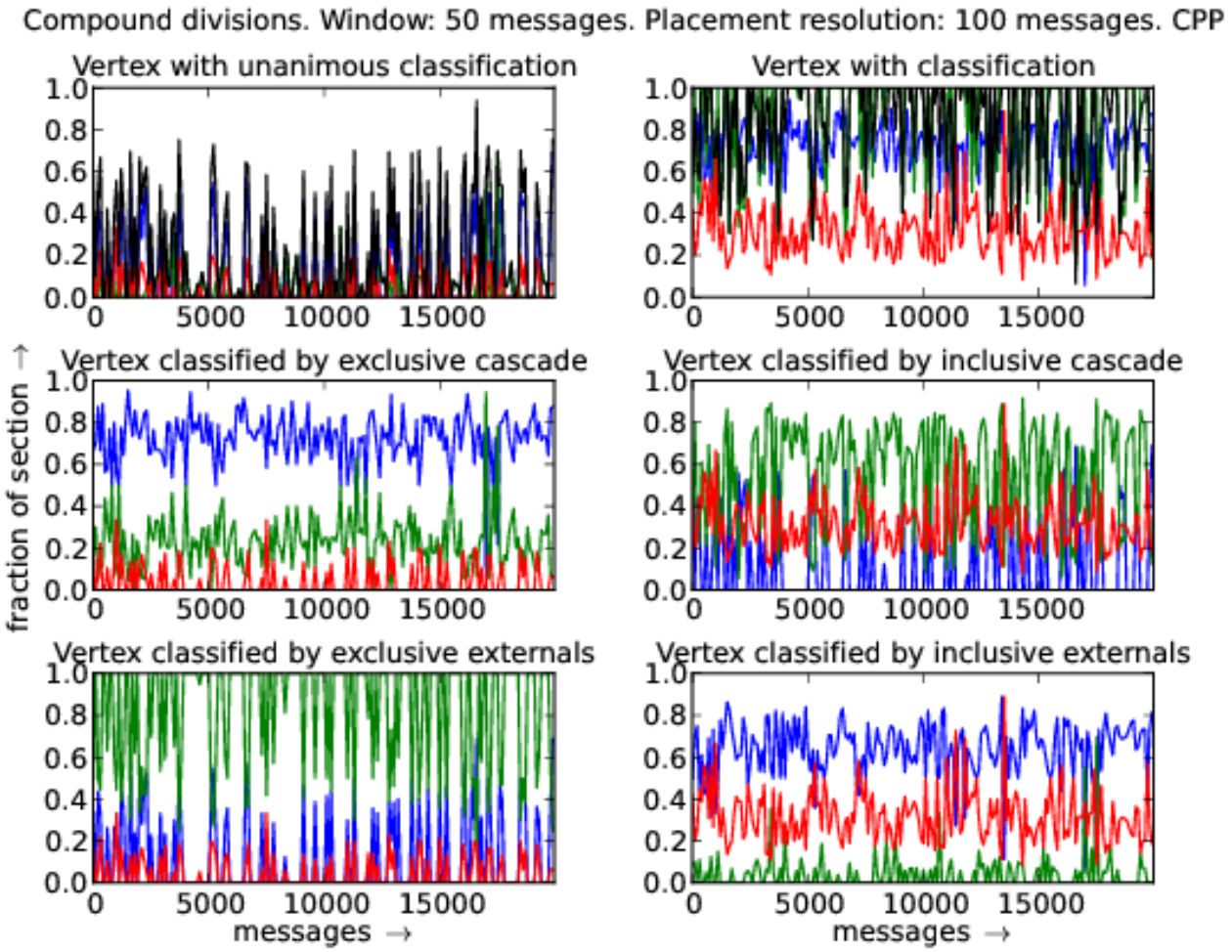

FIG. 19. Distribution of vertex with respect to each measure of activity: compound criteria. Red, green and blue are for hubs, intermediary and border (peripheral) vertex fractions. The first two plots picture classifications that are not functions. Thus, in the first plot, the fraction of vertexes with unique classification in plotted in black. On the second plot, black is used to depict how much a vertex was classified in more than section: $\frac{\text{number of classifications} - \text{number of nodes}}{\text{number of nodes}}$. Compound criteria is described in Section III B 1.

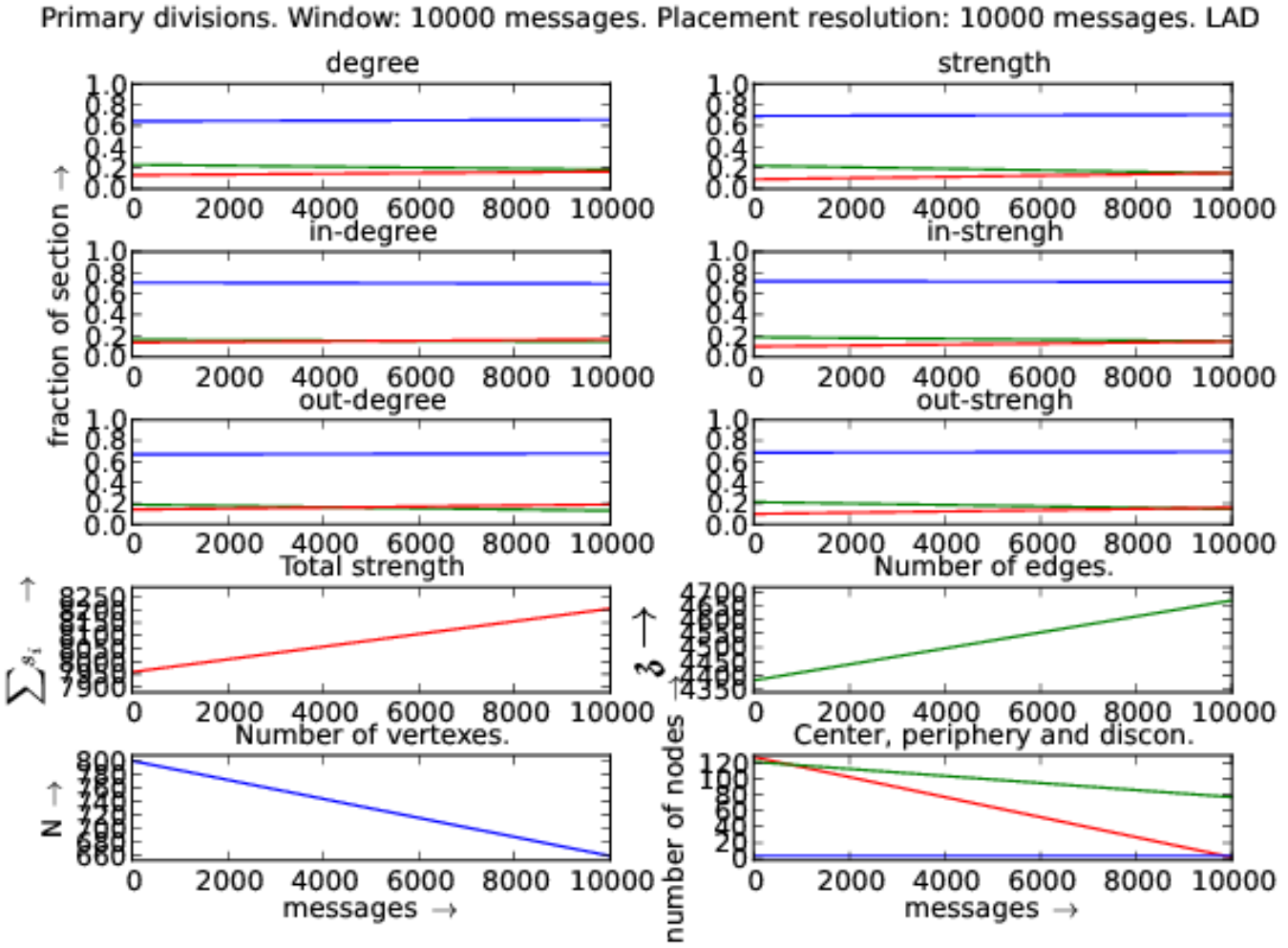

FIG. 20. Distribution of vertex with respect to each measure of activity: in and out degrees and strengths. CPP Std library official mailing list. In the first six plots, red is fraction of hubs, green is the fraction of intermediary and blue is for peripheral fraction. On the last plot, red is the center (maximum distance to another vertex is equal to radius), blue is periphery (maximum distance equals to diameter) of the giant component. On the same graph, green counts the disconnected vertexes.

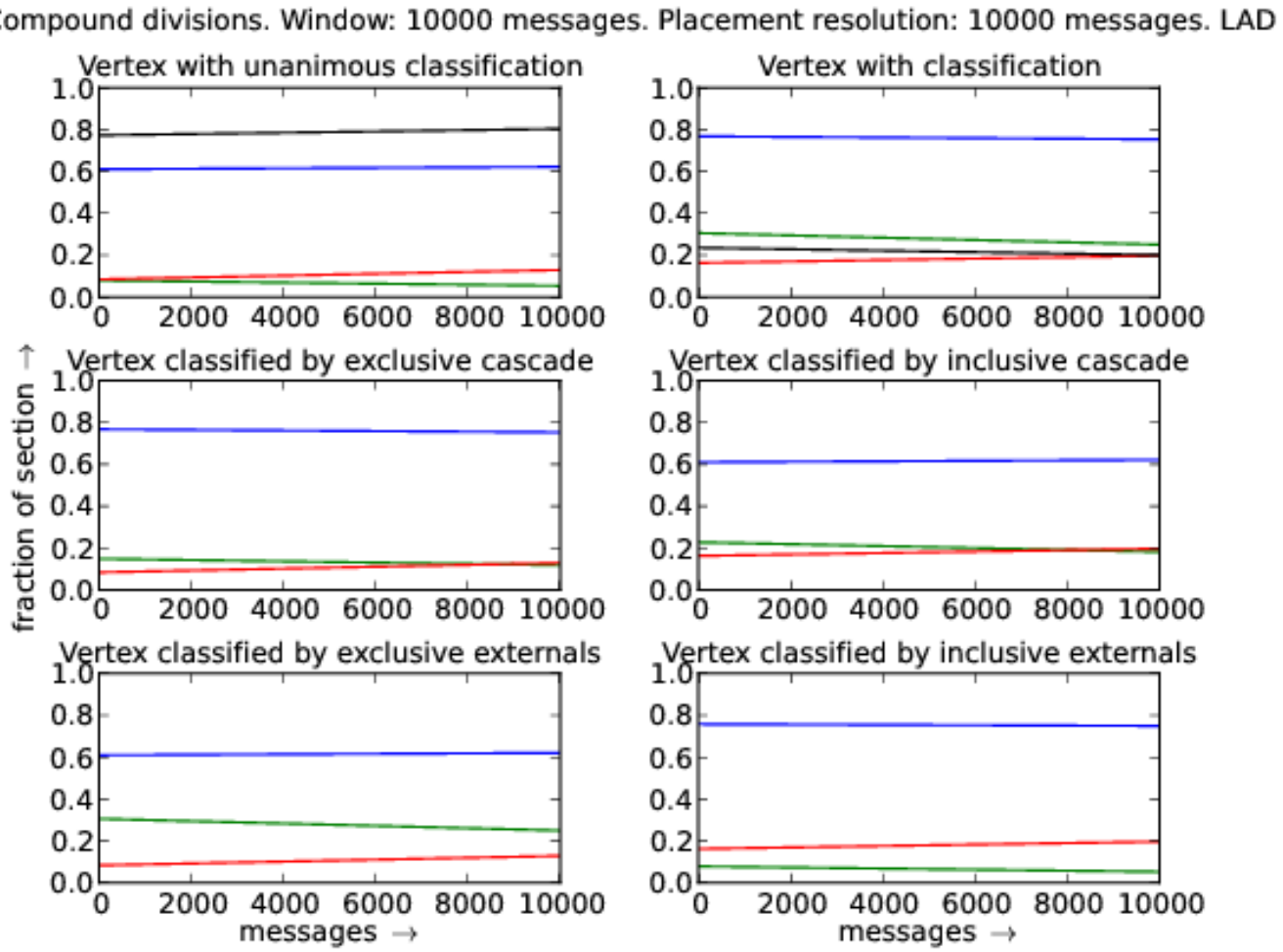

FIG. 21. Distribution of vertex with respect to each measure of activity: compound criteria. Red, green and blue are for hubs, intermediary and border (peripheral) vertex fractions. The first two plots picture classifications that are not functions. Thus, in the first plot, the fraction of vertexes with unique classification in plotted in black. On the second plot, black is used to depict how much a vertex was classified in more than section: $\frac{\text{number of classifications} - \text{number of nodes}}{\text{number of nodes}}$. Compound criteria is described in Section III B 1.

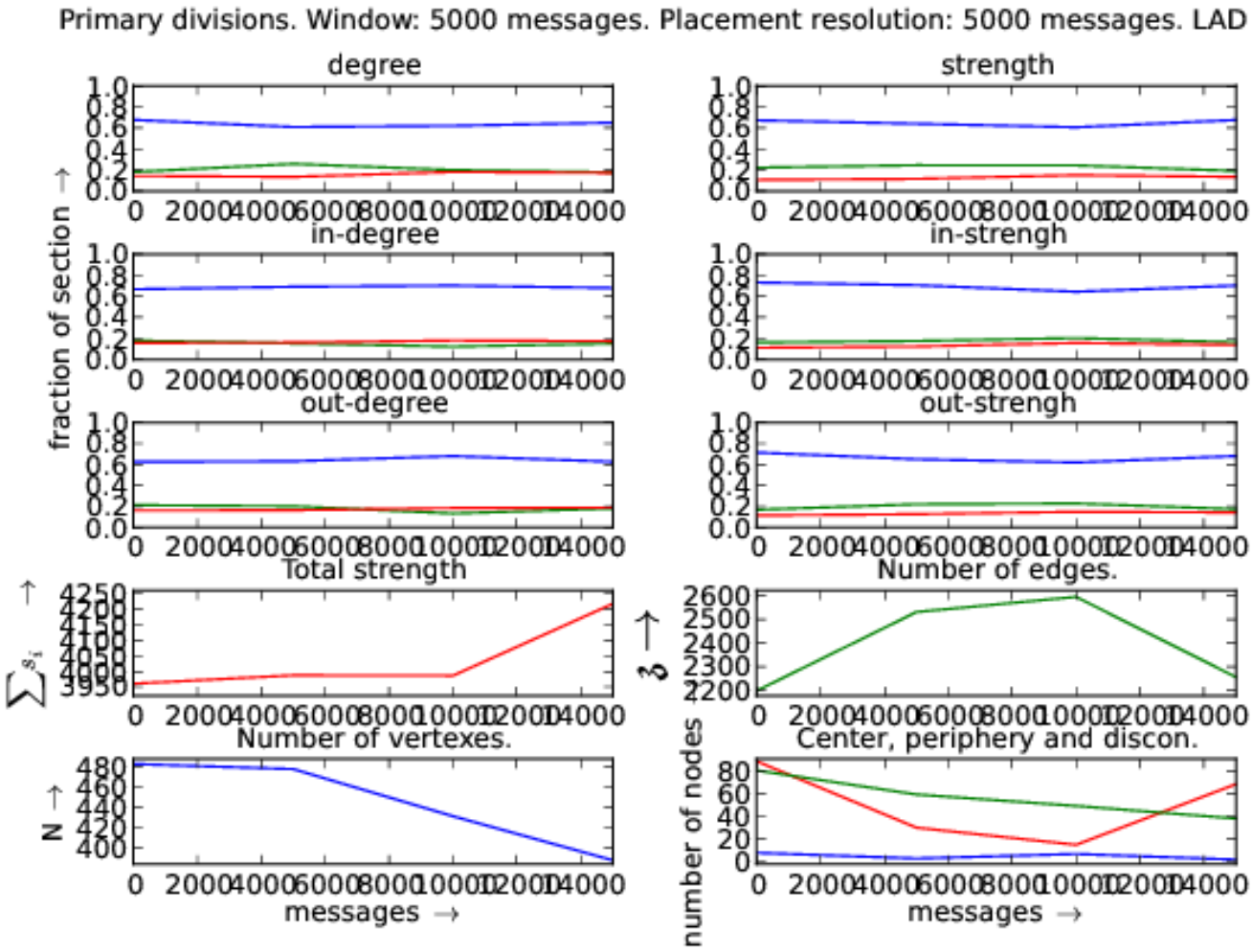

FIG. 22. Distribution of vertex with respect to each measure of activity: in and out degrees and strengths. CPP Std library official mailing list. In the first six plots, red is fraction of hubs, green is the fraction of intermediary and blue is for peripheral fraction. On the last plot, red is the center (maximum distance to another vertex is equal to radius), blue is periphery (maximum distance equals to diameter) of the giant component. On the same graph, green counts the disconnected vertexes.

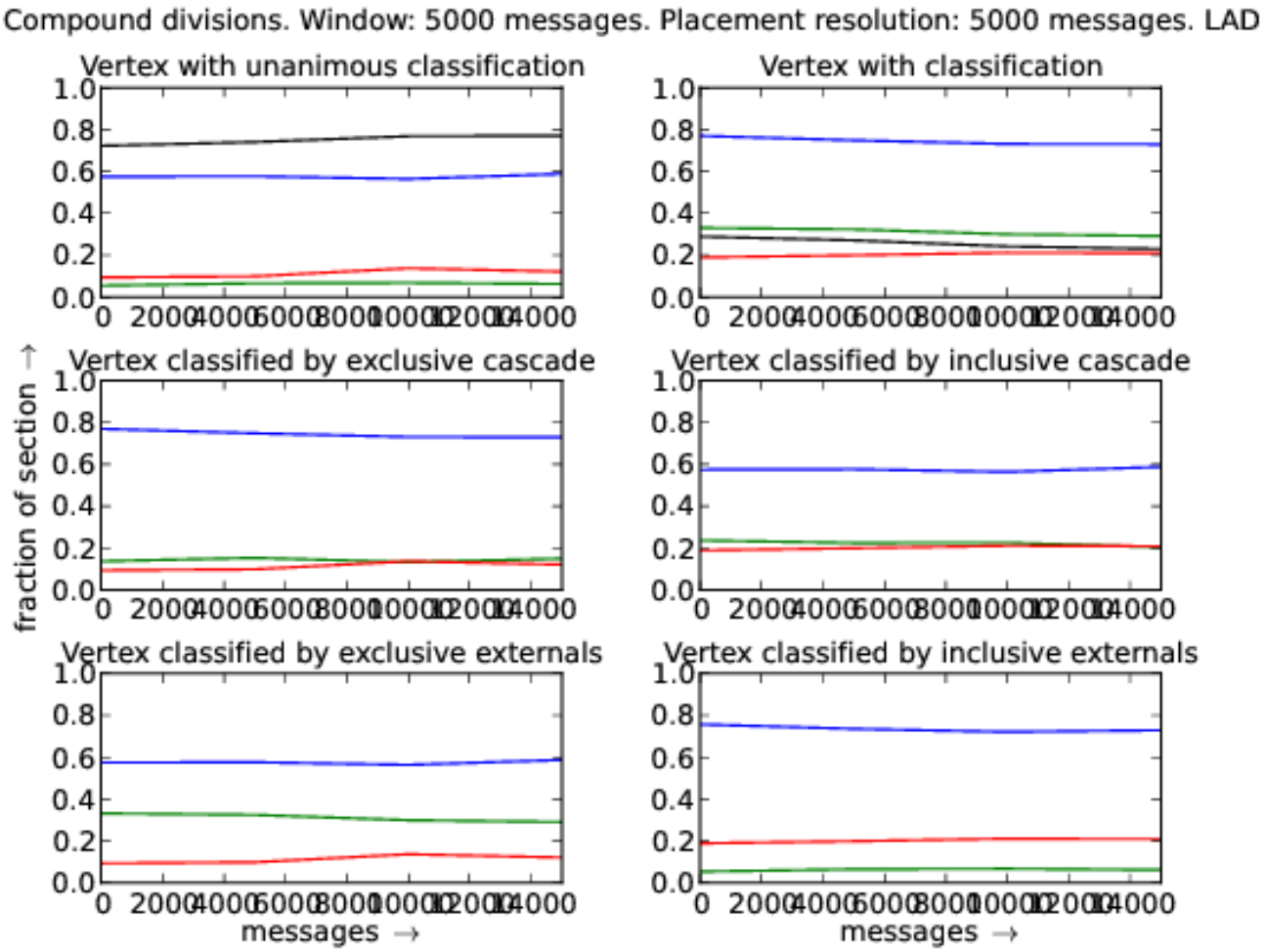

FIG. 23. Distribution of vertex with respect to each measure of activity: compound criteria. Red, green and blue are for hubs, intermediary and border (peripheral) vertex fractions. The first two plots picture classifications that are not functions. Thus, in the first plot, the fraction of vertexes with unique classification in plotted in black. On the second plot, black is used to depict how much a vertex was classified in more than section: $\frac{\text{number of classifications} - \text{number of nodes}}{\text{number of nodes}}$. Compound criteria is described in Section III B 1.

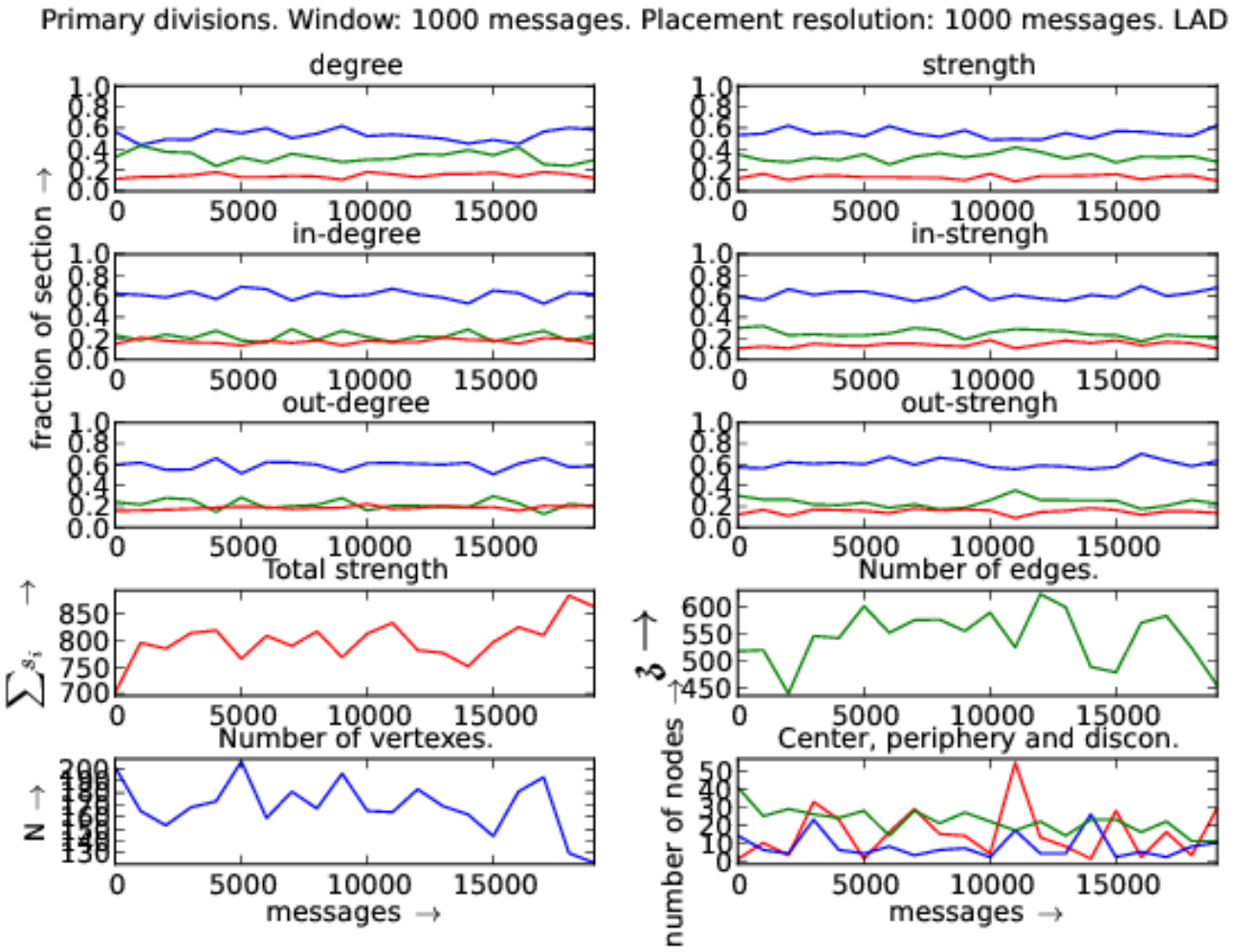

FIG. 24. Distribution of vertex with respect to each measure of activity: in and out degrees and strengths. CPP Std library official mailing list. In the first six plots, red is fraction of hubs, green is the fraction of intermediary and blue is for peripheral fraction. On the last plot, red is the center (maximum distance to another vertex is equal to radius), blue is periphery (maximum distance equals to diameter) of the giant component. On the same graph, green counts the disconnected vertexes.

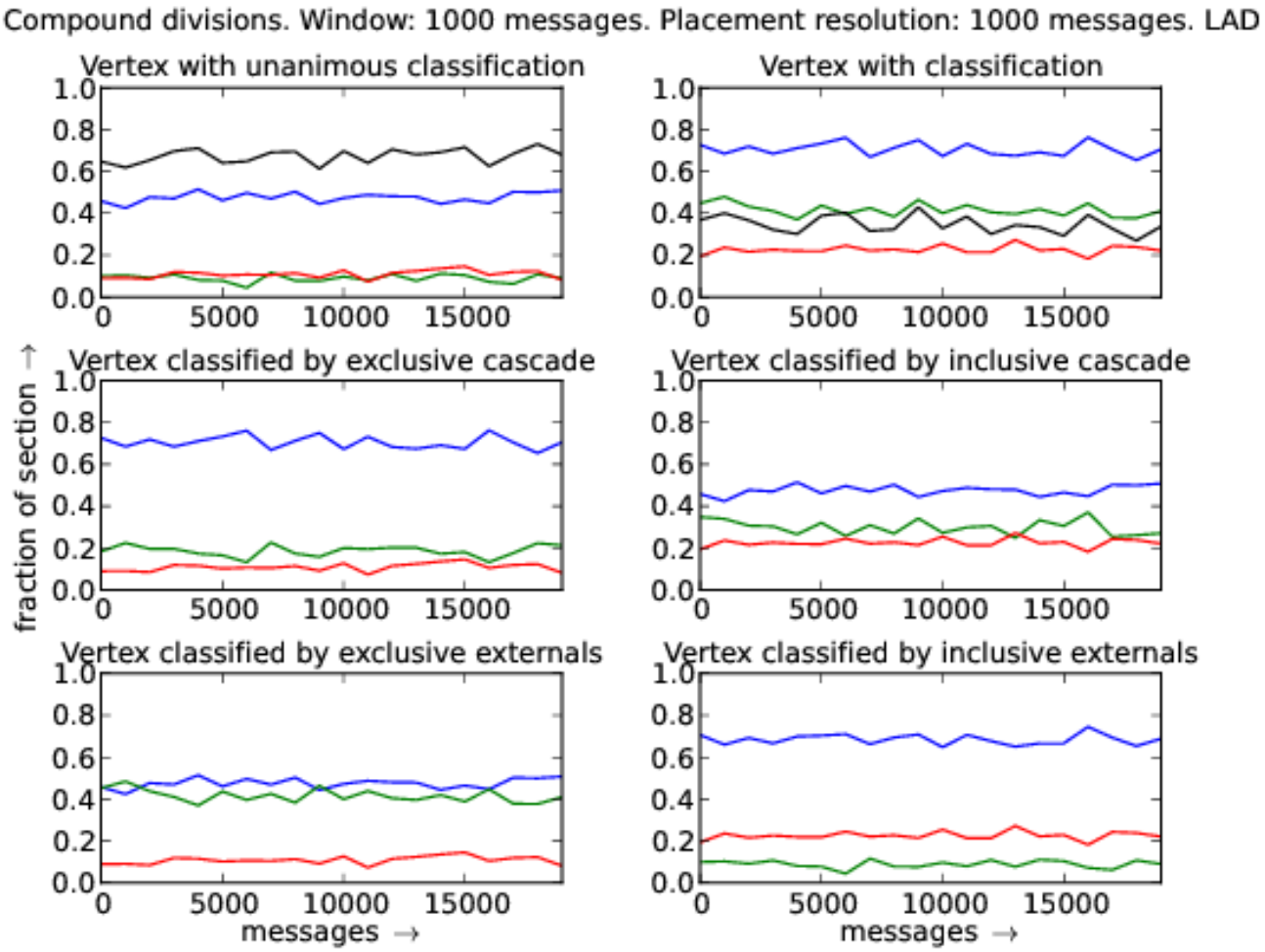

FIG. 25. Distribution of vertex with respect to each measure of activity: compound criteria. Red, green and blue are for hubs, intermediary and border (peripheral) vertex fractions. The first two plots picture classifications that are not functions. Thus, in the first plot, the fraction of vertexes with unique classification in plotted in black. On the second plot, black is used to depict how much a vertex was classified in more than section: $\frac{\text{number of classifications} - \text{number of nodes}}{\text{number of nodes}}$. Compound criteria is described in Section III B 1.

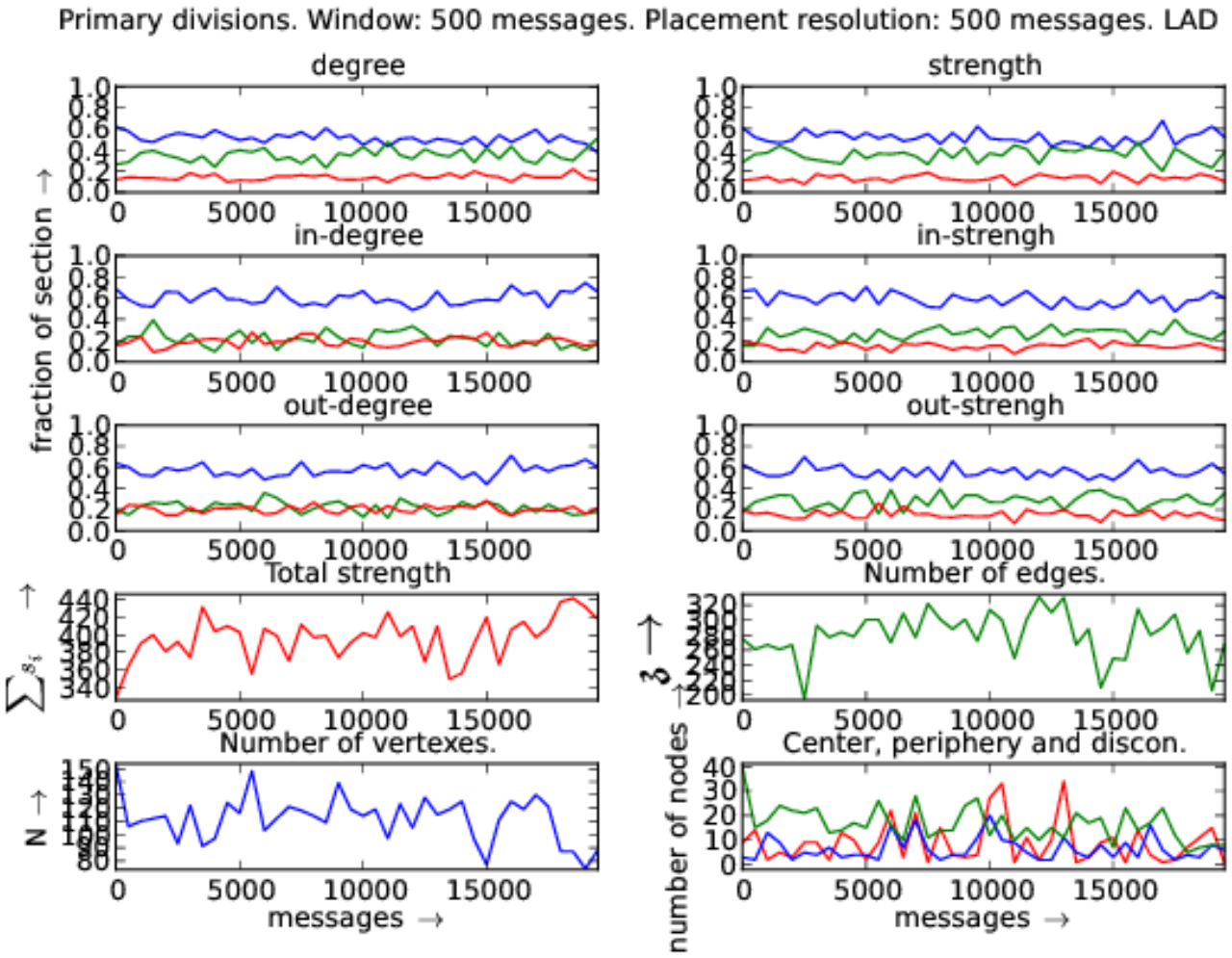

FIG. 26. Distribution of vertex with respect to each measure of activity: in and out degrees and strengths. CPP Std library official mailing list. In the first six plots, red is fraction of hubs, green is the fraction of intermediary and blue is for peripheral fraction. On the last plot, red is the center (maximum distance to another vertex is equal to radius), blue is periphery (maximum distance equals to diameter) of the giant component. On the same graph, green counts the disconnected vertexes.

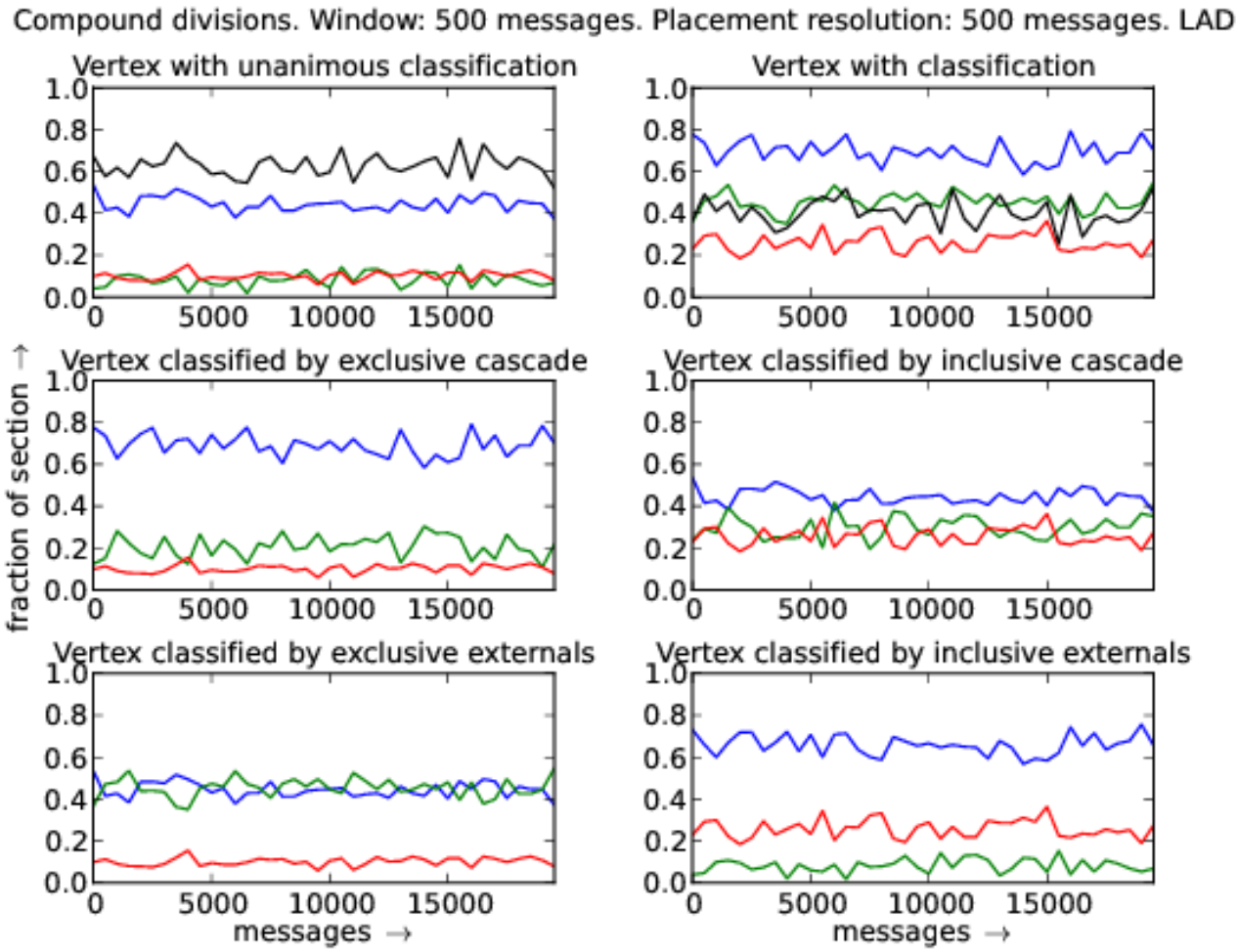

FIG. 27. Distribution of vertex with respect to each measure of activity: compound criteria. Red, green and blue are for hubs, intermediary and border (peripheral) vertex fractions. The first two plots picture classifications that are not functions. Thus, in the first plot, the fraction of vertexes with unique classification in plotted in black. On the second plot, black is used to depict how much a vertex was classified in more than section: $\frac{\text{number of classifications} - \text{number of nodes}}{\text{number of nodes}}$. Compound criteria is described in Section III B 1.

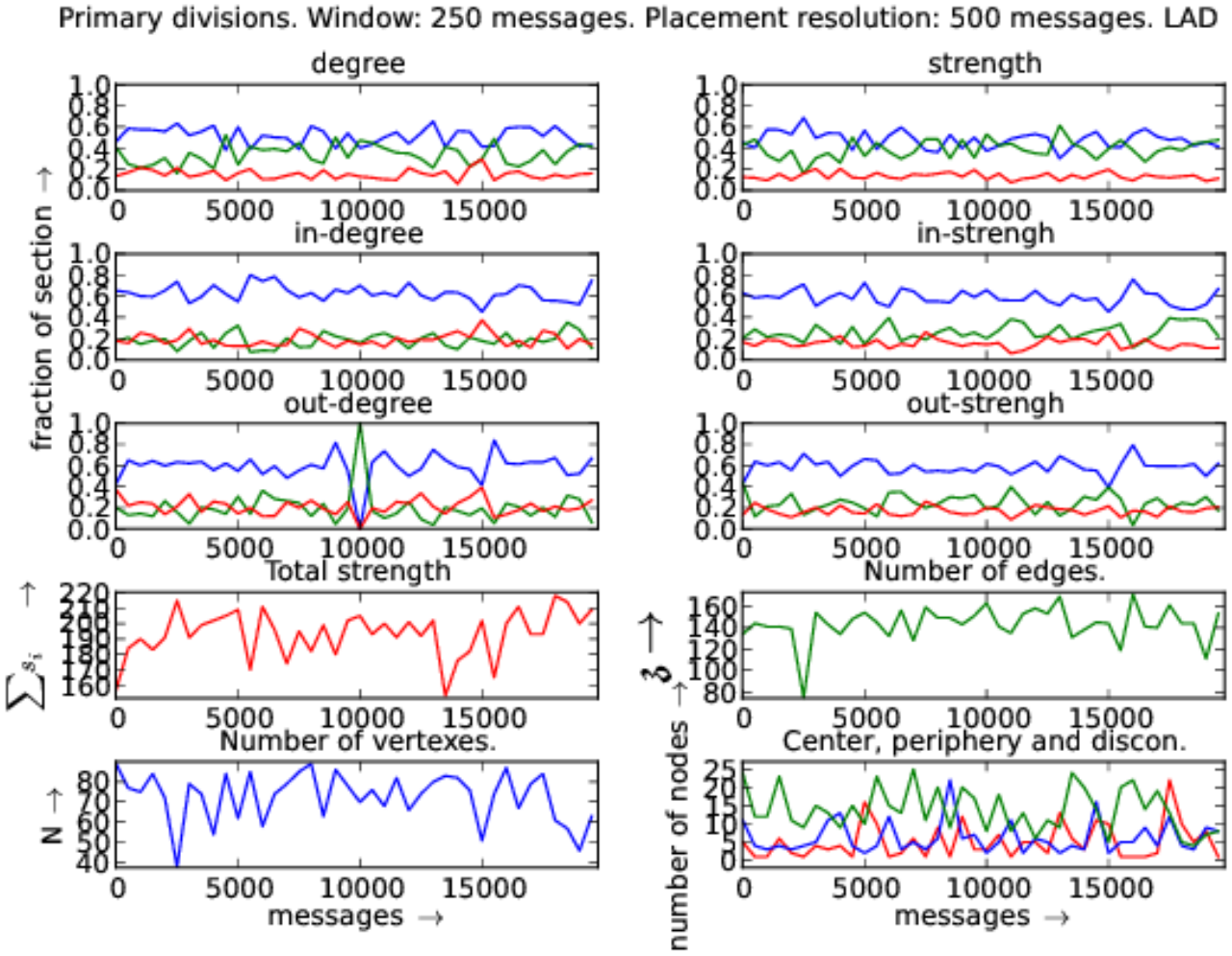

FIG. 28. Distribution of vertex with respect to each measure of activity: in and out degrees and strengths. CPP Std library official mailing list. In the first six plots, red is fraction of hubs, green is the fraction of intermediary and blue is for peripheral fraction. On the last plot, red is the center (maximum distance to another vertex is equal to radius), blue is periphery (maximum distance equals to diameter) of the giant component. On the same graph, green counts the disconnected vertexes.

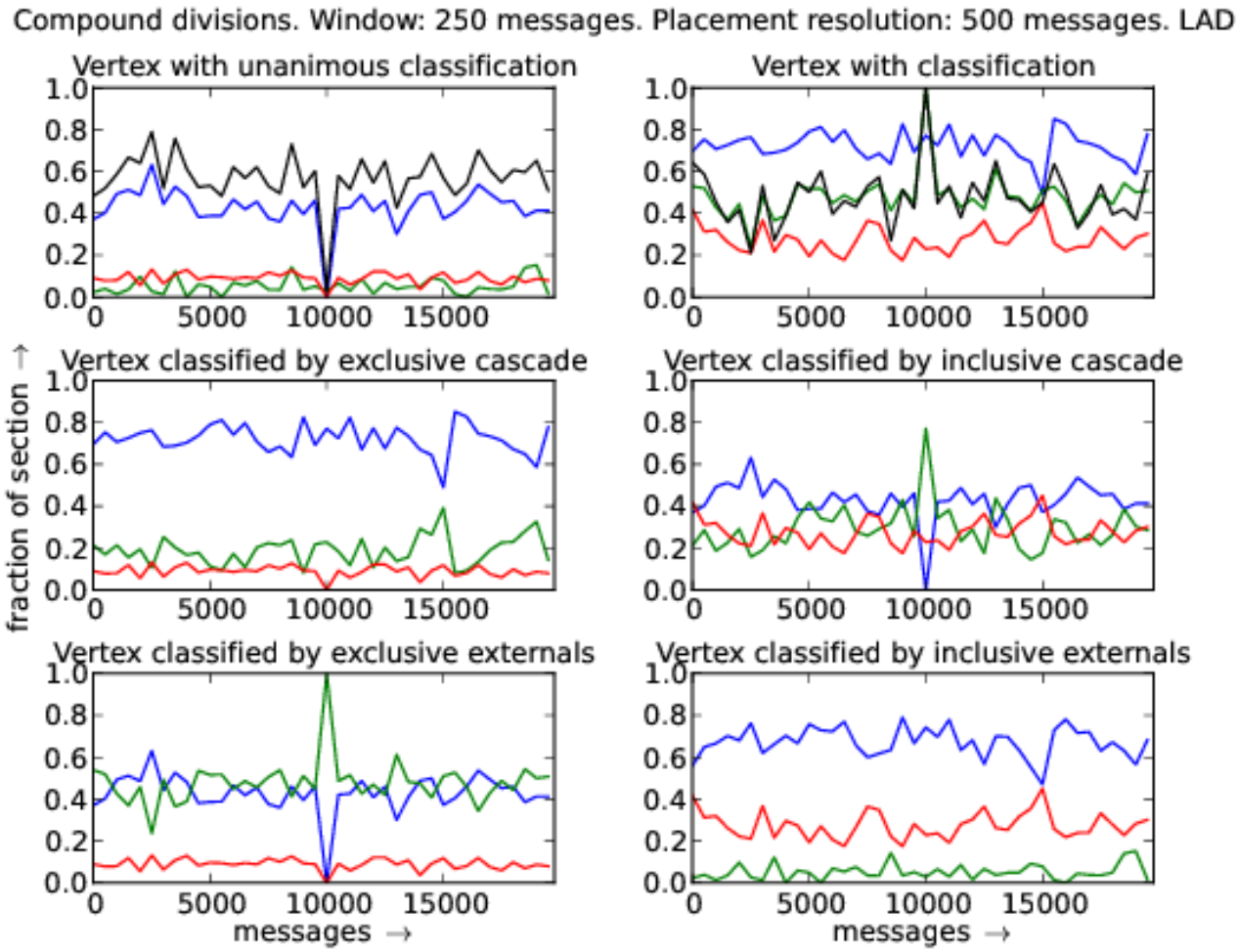

FIG. 29. Distribution of vertex with respect to each measure of activity: compound criteria. Red, green and blue are for hubs, intermediary and border (peripheral) vertex fractions. The first two plots picture classifications that are not functions. Thus, in the first plot, the fraction of vertexes with unique classification in plotted in black. On the second plot, black is used to depict how much a vertex was classified in more than section: $\frac{\text{number of classifications} - \text{number of nodes}}{\text{number of nodes}}$. Compound criteria is described in Section III B 1.

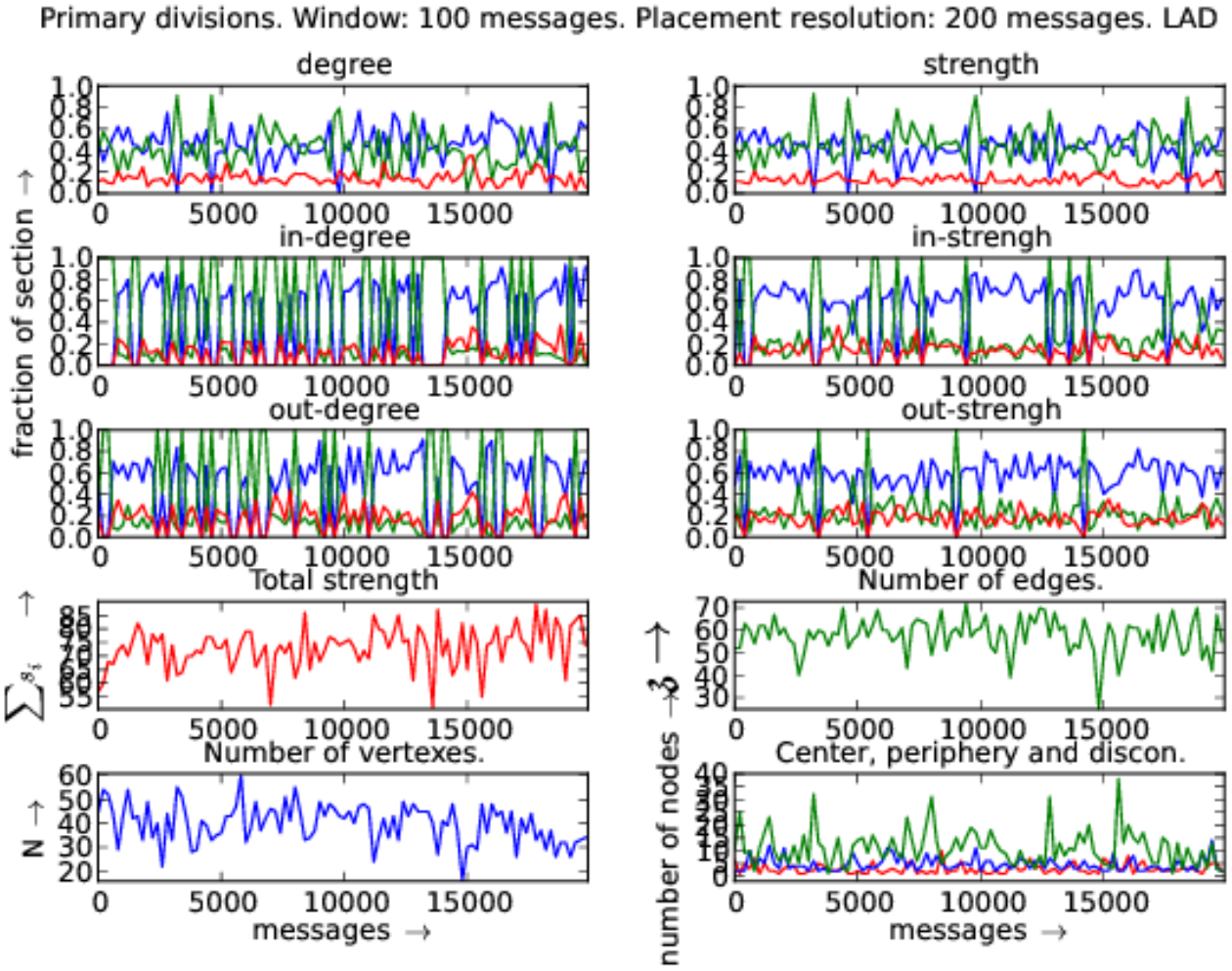

FIG. 30. Distribution of vertex with respect to each measure of activity: in and out degrees and strengths. CPP Std library official mailing list. In the first six plots, red is fraction of hubs, green is the fraction of intermediary and blue is for peripheral fraction. On the last plot, red is the center (maximum distance to another vertex is equal to radius), blue is periphery (maximum distance equals to diameter) of the giant component. On the same graph, green counts the disconnected vertexes.

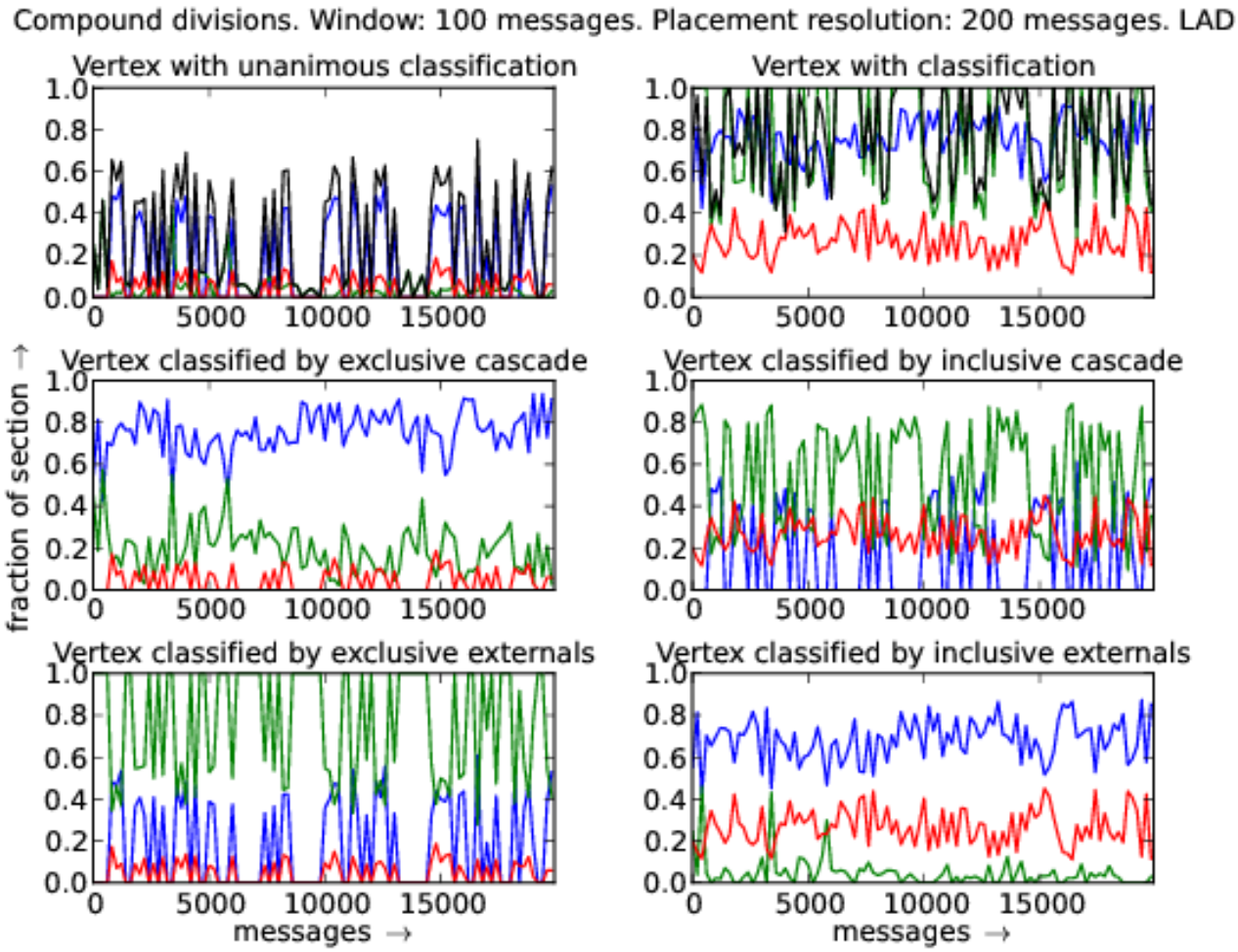

FIG. 31. Distribution of vertex with respect to each measure of activity: compound criteria. Red, green and blue are for hubs, intermediary and border (peripheral) vertex fractions. The first two plots picture classifications that are not functions. Thus, in the first plot, the fraction of vertexes with unique classification in plotted in black. On the second plot, black is used to depict how much a vertex was classified in more than section: $\frac{\text{number of classifications} - \text{number of nodes}}{\text{number of nodes}}$. Compound criteria is described in Section III B 1.

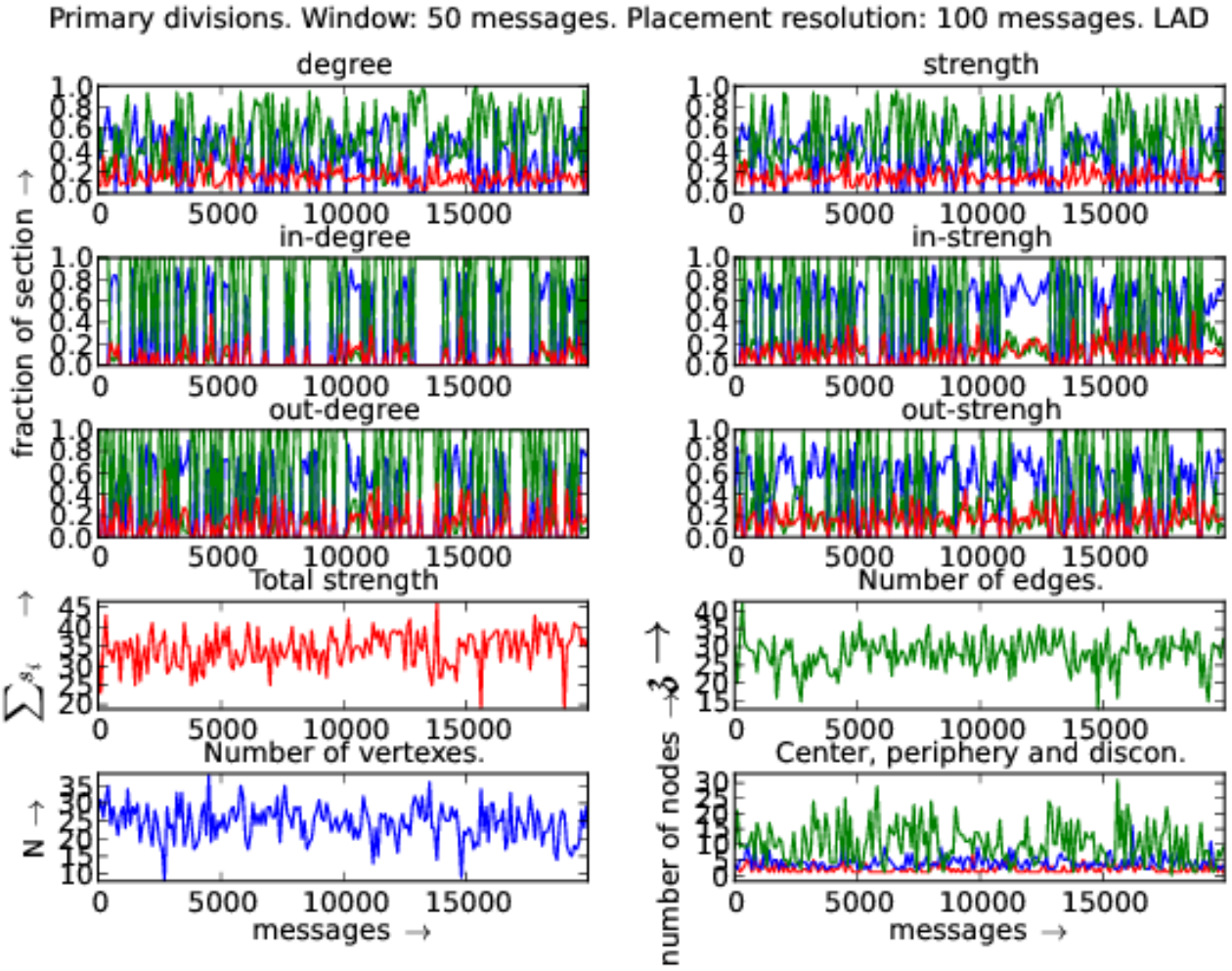

FIG. 32. Distribution of vertex with respect to each measure of activity: in and out degrees and strengths. CPP Std library official mailing list. In the first six plots, red is fraction of hubs, green is the fraction of intermediary and blue is for peripheral fraction. On the last plot, red is the center (maximum distance to another vertex is equal to radius), blue is periphery (maximum distance equals to diameter) of the giant component. On the same graph, green counts the disconnected vertexes.

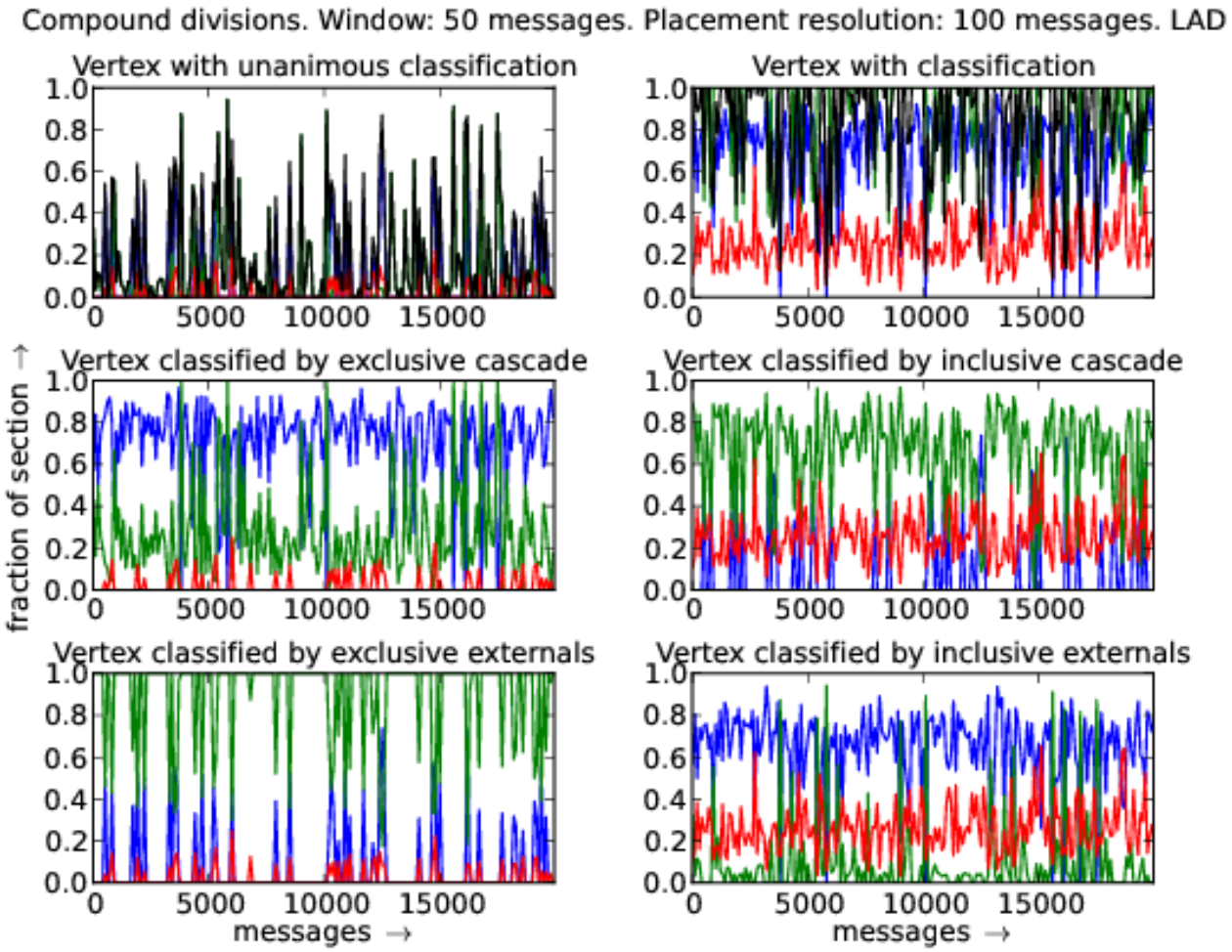

FIG. 33. Distribution of vertex with respect to each measure of activity: compound criteria. Red, green and blue are for hubs, intermediary and border (peripheral) vertex fractions. The first two plots picture classifications that are not functions. Thus, in the first plot, the fraction of vertexes with unique classification in plotted in black. On the second plot, black is used to depict how much a vertex was classified in more than section: $\frac{\text{number of classifications} - \text{number of nodes}}{\text{number of nodes}}$. Compound criteria is described in Section III B 1.