Dialog in e-Mail Traffic

Jean-Pierre Eckmann,[1,2] Elisha Moses,[3] and Danilo Sergi[1]
[1] *Département de Physique Théorique, Université de Genève, 1211 Genève 4, Switzerland*
[2] *Section de Mathématiques, Université de Genève, 1211 Genève 4, Switzerland*
[3] *Department of Physics of Complex Systems, Weizmann Institute of Science, Rehovot 76100, Israel*
(Dated: April 22, 2003)

Connectivity and topology are known to yield information about networks, whose origin is self-organized, but the impact of *temporal dynamics* in a network is still mostly unexplored. Using an information theoretic approach to e-mail exchange, we show that an e-mail network allows for a separation of static and dynamic structures within it. The static structures are related to organizational units such as departments. The temporally linked structures turn out to be more goal-oriented, functional units such as committees and user groups.

PACS numbers: 89.75.Hc, 89.75.Fb, 89.70.+c, 89.20.Hh
Keywords: e-mail network; synchronization; thematic organization

The theory of complex networks has developed tools to investigate quantitatively the properties of a number of new systems, such as the World Wide Web (WWW), the protein-protein interaction database and others (see [1, 2] and references therein for a review). The challenging difficulty shared by such systems is the interplay between their constituents, and the resulting collective effects. When the system is a fixed graph whose links describe interaction, concepts such as the clustering coefficient (or curvature) have been introduced and applied successfully. In e-mail networks *temporal dynamics* appears as a new ingredient, and therefore it is possible to ask about the synchronization of e-mail traffic between communicating users, and to determine the correlation (cohesion) between them.

In this Letter, we show how to obtain an objective measure of the interaction between the activity of users by employing tools provided by information theory, or the general theory of entropy [3]. We find that the temporal structure of the e-mail exchange reveals a new form of organization that is different from what can be captured by the more static notion of curvature, or any other study which neglects temporal aspects. It is intriguing that the question of how an organization communicates internally is similar to those that arise in the study of how an organ like the brain organizes its neuronal activity. We find analogies in our analysis to the approach which studies the appearance of correlations and synchronization between spike trains in the creation of a neural code [4].

The experiment Our data are extracted from the log files of one of the main mail servers at one of our universities, and consist of over $2 \cdot 10^6$ e-mail messages sent during a period of 83 days, connecting about ten thousand users. The content of the messages is of course never accessible, and the only data taken from the log file are the 'to', 'from', and 'time' fields. The data are first reduced to the internal mail within the institution, since external links are necessarily incomplete. Once aliases are resolved, we are left with a set of 3,188 users interchanging 309,125 messages.

A directed graph is then constructed by designating users as nodes and connecting any two of them with a directed link if an e-mail message has gone between them during the 83 days. This procedure defines a *static* graph. Statistical properties of the degree of this graph have been reported before [5, 6]. Connectivity of this *static* graph will reveal structures within the organization [5, 7, 8]. We have previously shown [1] that a powerful tool for identifying such structural organization is the number t of triangles (triplets in which all pairs communicate) that a node of valence v (total number of partners) participates in, normalized by the number of triangles $v(v-1)/2$ that it could potentially belong to. This defines the clustering coefficient $c = 2t/v(v-1)$, which as we showed induces a curvature on the graph [1].

One marked difference between the graphs of e-mail and of the WWW should be noted at this point. In the WWW, the central organizing role of 'hubs' (nodes with many outgoing links) that confer importance to 'authorities' (nodes with many ingoing links) has been noted [8] and utilized very successfully (e.g., by Google). The contribution of authorities and hubs is, however, *not* to the creation of communities and interest groups. This is evident since the high valence of both hubs and authorities tends to reduce their curvature considerably. High curvature nodes, in contrast, are usually the specialists of their community, that are highly connected in bi-directional links to others in the group. In the e-mail graph, hubs tend to be machines, mass mailers or users that transfer general messages (e.g., seminar notifications), going out to many users, while authorities are more like service desks. Thus the importance of hubs and authorities is small if we consider the core use of the e-mail structure as dealing with thematic rather than organizational issues. They do, however, play a role in such questions as diffusion of viruses, or more generally, how many people are being reached [5]. But most mass mailings do not solicit an answer, and therefore do not contribute to interaction ('dialog') as we define and study it in this Letter. In our analysis we discard mass mailings (more than 18 recipients) altogether. There remain 202,695 links.

The different manner in which triangles and transitivity interplay in the WWW and in the e-mail graphs is also illuminating. The notion of curvature is a local one, based on the more basic concept of a 'co-link'. This is a link between two nodes that point to each other, establishing a 'friendly' connection based on mutual recognition. Building from the single pair, the fundamental unit of connectivity is the triangle [9]. In the WWW transitivity is natural, and we have shown previously that if node A is 'friendly' with nodes B and C, it is often correct to assume that B and C are friends. On the contrary, e-mailing is so prolific that A's having a dialog with B and with C usually does not imply that B and C carry out a dialog, and even if they do then the three communications determining the edges of the triangle could be independent; as a consequence, transitivity breaks down. We will see that static structures (such as departments) emerge as high-curvature ones, while dialog between members of a group implies a more functional, and perhaps goal-oriented structure.

Our analysis expands the notion of a mutual, or 'co-link' to the e-mail network by designating a link between nodes A and B only if A has sent a message to B and B *has sent a message back* during the whole period under investigation. We find 7,087 such pairs, sending 105,349 messages to each other, among 20,879 directed pairs who sent perhaps mail just one way (and out of the $3{,}188 \cdot 3{,}187/2$ possible connections in the graph).

The model To analyze the behavior of this reduced network, we view any pair of 'conversing' users as exchanging signals on a transmission line on which information can be propagated in both directions. We completely disregard the fact that there is internal information in the messages, discarding even information that is in principle available in the log files such as the size of the messages. The data for each pair is a spike train whose horizontal axis is time, with upward ticks for a mail sent $A \to B$ and a downward tick for $B \to A$ (some samples are shown in Fig. 2). We now define that A **and B conduct a dialog on a given day** if A sends mail on that day to B and B answers *on the same day*.

The temporal dynamics of the e-mail network immediately reveals new statistical properties, shown in Fig. 1. We define ΔT as the time delay between a message going from $A \to B$ and a response going from $B \to A$. While no clear power law is evident in Fig. 1, the behavior can be approximated by $P(\Delta t) \approx \Delta t^{-1}$. The appearance of a peak ranging from $\Delta t = 16$ to $\Delta t = 24$ can be explained by sociological behavior involving the time (usually 16 hours) between when people leave work and when they come back to their offices. This (already very weak) peak disappears when considering in the inset the basic time unit as a 'tick' of the system (= a message sent). We suspect that the approximate power law is caused by random communications between two users, while the flat incipient part implies actual correlation between two users (when the answer comes before 10 hours have passed, i.e. on the same day).

Choosing the basic 'tick' of the clock (the sending of a message in the network) as a *variable* time unit smoothens many features (as in Fig. 1, inset). In particular, the slowing down of the network over nights and weekends is eliminated. But the mathematics of 'correlation' becomes much more involved, and we have also checked that the interaction is very well captured by sticking to the more intuitive notion of 'same day'. In light of these considerations, we choose 24 hours as the natural time unit within which B is required to answer. In principle, some multiple of this unit could serve as well, but the results of Fig. 1 show that most interactions take place within 10 hours. We further checked that extending the choice of a day to an answer on the next day gives similar results.

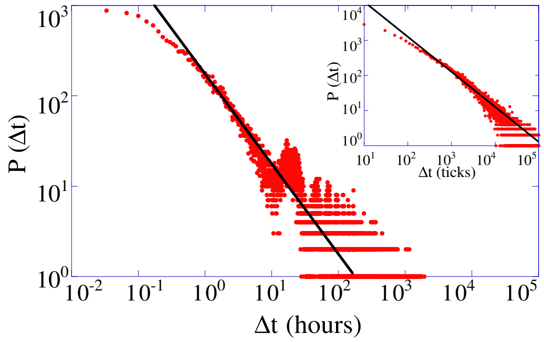

FIG. 1: The probability distribution of the response time till a message is 'answered' (see text for definitions). Inset: same but measured in 'ticks', i.e. units of messages sent in the system. Solid lines follow $\sim \Delta t^{-1}$ and are meant as a guide to the eye.

The mathematical description of the experiment proceeds through two steps, cf. Fig. 2. First, at a more local level, we consider a pair of communicating users, that we shall denote by A and B. We introduce the probabilities $p_A(i)$ and $p_B(i)$, where $i = 0, 1$. The value 1 corresponds to the event that at least one e-mail has been sent to the partner on a given day, while the value 0 corresponds to having sent none on that day. The measured values of these probabilities are given by

$$p_A(i) = N_A(i)/d \ ,$$

where $N_A(i)$ is the number of days for which the event i occurred for A, and similarly for B (and d is the total number of days $d = 83$). We then characterize the *joint* activity of A and B by considering the probabilities $p_{AB}(i, j)$ defined as

$$p_{AB}(i, j) = N_{AB}(i, j)/d \ ,$$

where N_{AB} is the number of days where A was in state i and B in state j (i.e., sending mail to the partner or not) and $i, j \in \{0, 1\}$. It is now possible to determine to which extent the activity of A influences the activity of B by means of the *mutual information* $I_p(A,B)$ (the subscript p stands for pair):

$$I_p(A,B) = \sum_{i,j=0,1} p_{AB}(i, j) \cdot \log \left(\frac{p_{AB}(i, j)}{p_A(i) \cdot p_B(j)} \right).$$

I_p measures in what way knowing what A does will predict what B does and vice versa (note that $I_p(A,B) = I_p(B,A)$).

The next step consists in considering every triangle of communicating users; to be specific we designate them by A, B, and C. In order to capture their joint activity we introduce the probabilities $p_{ABC}(i_1, i_2; i_3, i_4; i_5, i_6) = p_{ABC}(\mathbf{i})$,

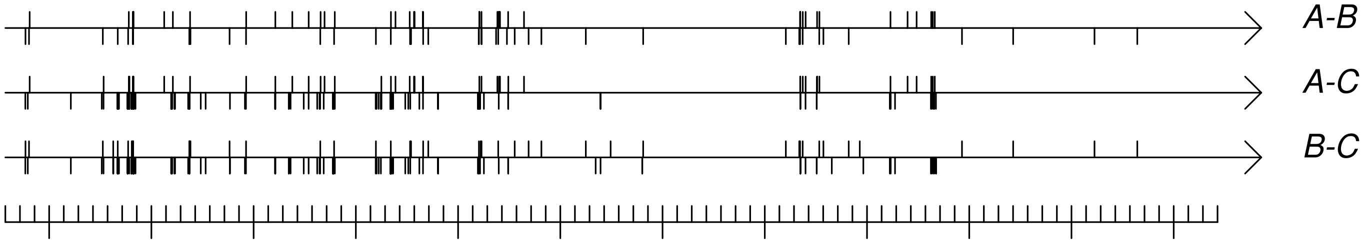

FIG. 2: Spike trains for the three communication channels determining the edges of a triangle formed by the users A, B and C. We have $I_p(A,B) = 0.095$, $I_p(A,C) = 0.394$, $I_p(B,C) = 0.172$ and $I_t = 1.606$. It is important to note that I_p and I_t capture the synchronization of the e-mail exchange at two different levels. I_t measures the coherence of the triangle as a whole, and can take on high values even though some of the I_p's are relatively small. The fourth horizontal line represents the whole period under analysis divided into days (and weekends), introduced with the purpose to help to visualize the events determining the probabilities entering I_p and I_t (cf. text).

where $i_1,\dots,i_6 \in \{0,1\}$. The pair (i_1,i_2) refers to the communication $A \leftrightarrow B$, (i_3,i_4) to $A \leftrightarrow C$, and (i_5,i_6) to $B \leftrightarrow C$. For example the pair $(i_1 = 1, i_2 = 0)$ has to be interpreted as the occurrence where on a given day A sends mail to B, but B does not send mail to A. An equivalent, evident interpretation holds for all other pairs. In formulas the above probabilities read

$$p_{ABC}(i_1,i_2;i_3,i_4;i_5,i_6) = N_{ABC}(i_1,i_2;i_3,i_4;i_5,i_6)/d\ ,$$

$N_{ABC}(\mathbf{i})$ being the number of days where the pattern (event) $\mathbf{i}$ occurred.

We now define the *temporal cohesion* of a triangle as the degree of synchronization between the activity of the three users. This is achieved by looking at a form of the *mutual information* $I_t(A,B,C)$ (in this case the subscript t stands for triangle) defined as

$$\begin{aligned} I_t(A,B,C) &= \sum_{i_1,\dots,i_6=0,1} p_{ABC}(\mathbf{i}) \\ &\times \log\left(\frac{p_{ABC}(i_1,i_2;i_3,i_4;i_5,i_6)}{p_{AB}(i_1,i_2)\cdot p_{AC}(i_3,i_4)\cdot p_{BC}(i_5,i_6)}\right). \end{aligned}$$

Note that the temporal cohesion $I_t(A,B,C)$ is invariant under any permutation of A, B and C. Also, $I_t \leq \log(16)$ and the maximum is attained when the four possible patterns for each edge are equiprobable and fully correlated. More insight into I_p and I_t can be gained by looking at Fig. 2, showing the three communications determining a triangle. A statistical quantity of interest, shown in Fig. 3 is the number t of triangles that a user participates in. The distribution of both the static and the dynamic (temporal cohesion ≥ 0.1) triangles follow a power law over two decades, with exponent -1.2.

Restoring transitivity With the help of the temporal cohesion I_t it is now possible to replace the static transitivity by a novel notion of temporal transitivity. The assumption is that if the e-mail exchange in a triangle is highly synchronized, the three users are indeed involved in a common dialog. This transitive relationship between users can be extended naturally to adjacent triangles. This idea relies on the observation that in the presence of two highly synchronized triangles with a common edge, the four users are supposed to influence each other's activity. In this way it is possible to extract the groups of users carrying out a dialog. We thus construct a new, conjugate, graph where we first draw a node for each triangle for which I_p is larger than a given cutoff. Two of these nodes will be connected by a link if the corresponding triangles have a common edge, that is, if 4 people A, B, C, D are involved in these 2 triangles (say A, B, C, and B, C, D). Such a construction (called the conjugate graph) will offer a perspective on the appearance of circles of users sharing a common interest, defining thematic groups.

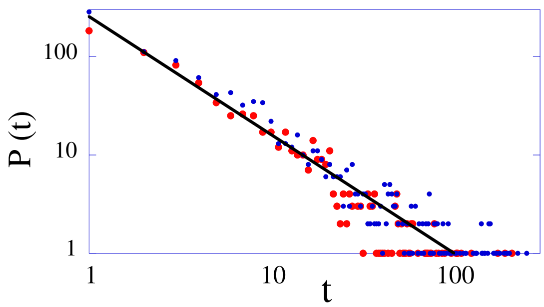

FIG. 3: Histogram of static and temporal statistical quantities. Probability distribution of the number t of triangles that a user participates in. Blue circles indicate static triangles, while red ones indicate 'temporally cohesive' triangles (i.e. mutual information $I_t \geq 0.1$). Both lines are well fit by $\sim \Delta t^{-1.2}$ and the black line is a guide to the eye with this slope.

Discussion of the results For the purpose of comparison, we first consider in Fig. 4 the *static* graph resulting from our e-mail network. For the sake of clarity only nodes, i.e. users, with a curvature larger than 0.1 are present; in addition, every pair of users must have exchanged at least 10 e-mails. The temporal dynamics intrinsic to the e-mail exchange is here neglected and triangles represent a sign of static transitive recognition, carrying no information about temporal cohesion between the individual communications. In this case we see the clear appearance of departmental communities. Our findings on the organizational aspects of the e-mail traffic are thus in agreement with the findings of [8], but are based here on the quantitative concept of curvature.

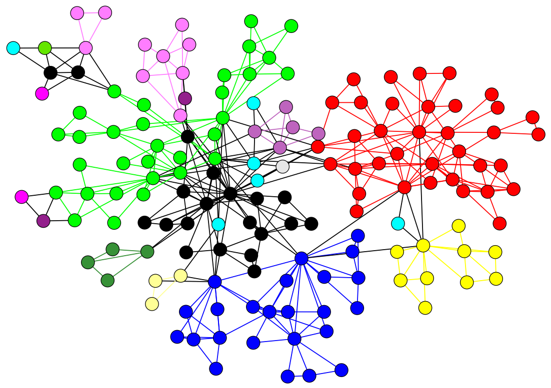

FIG. 4: The static structure of the graph of e-mail traffic, arranged according to curvature, based on triangles of mutual recognition. Time is thus *not* taken into account, and the graph of users arranges itself primarily according to departments, shown in various colors.

Fig. 5, on the other hand, shows the conjugate graph associated with a cutoff of $I_t \geq 0.5$. We recognize several highly connected, totally separated clusters, indicating different thematic groups. Some of the clusters identified in Fig. 2 survive and are lifted in part to the graph of temporal dynamics, indicating that within some departments there are dialogs; furthermore some departments split into different interest groups. However, we find many clusters that are new, and do not appear in the high curvature graph. These are typically users that are not in the same department, as shown by the multiple colors of the disks. Very few users appear in more than one cluster, so that the spreading of functional information is restricted within the thematic communities, in contrast to spreading of computer viruses for example, which propagate easily through the entire graph [5].

Some conclusions may be drawn regarding the nature of communities that emerge by conducting a dialog in the internet network. Two people engaged in a project can, if necessary, pick up the phone and tie all loose ends efficiently. However, a group with three or more participants may find it hard to coordinate conference calls, and in general will benefit from the lower time constraints that allow each participant to formulate his views and present them to a forum by e-mail. This makes e-mail an ideal medium for discussion groups involved in a given project, or a committee involved in a functional activity. Indeed, we have identified two committees in the clusters of Fig. 5 that are involved in non-academic activities within the university. A third group can be identified as visiting scientists (e.g. post-docs etc.) from a common foreign nationality.

The choice of a university's e-mail network is perhaps not ideal for identifying such 'groups of dialog'. This is because the major activity in a university is research, which usually involves few individuals, and is almost never advanced by committee. We thus speculate that the role of dialog in defining functional communities will be greater in large organizations such as companies [8] or government offices.

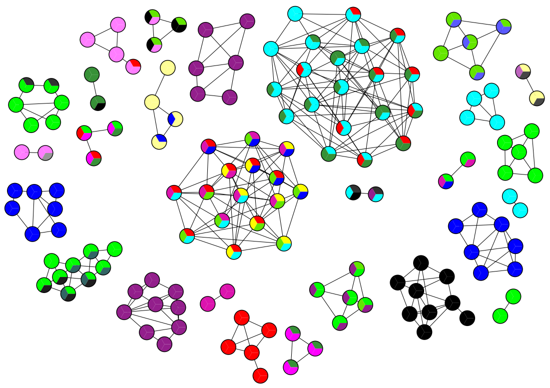

FIG. 5: The conjugate graph, for a cutoff of 0.5. Each node is a triangle of 3 people conferring with temporal cohesion $I_t \geq 0.5$, and each link connects two adjacent such triangles. The 3 colors of each node are the departments of the 3 people (same color code as in Fig. 4). Note the strong clustering of the graph into very compact groups of people. The users *cross* department boundaries (their interests and connections are not shown out of considerations of privacy).

Summary We have studied e-mail communications over 83 days and quantified the synchronization of groups of coherently communicating users. This synchronization reveals the existence of common interests within those groups. This form of organization cannot emerge by applying the analysis based on static concepts such as curvature, since those detect only *structural* rather than *thematic* organization. The reason is that in the context of static e-mail networks a triangle does not automatically imply transitivity from a thematic point of view. But we have demonstrated that transitivity can be recaptured by taking into account the temporal dynamics of the e-mail traffic.

We thank P. Collet for many helpful discussions and P. Choukroun and A. Malaspinas for help with the e-mail data. This work was supported by the Fonds National Suisse, and by the Minerva Foundation (Munich).

[1] J.-P. Eckmann and E. Moses, Proc. Natl. Acad. Sci. USA **99**, 5825 (2002).
[2] R. Albert and A.-L. Barabási, Rev. Mod. Phys. **74**, 47 (2002).
[3] T.M. Cover and J.A. Thomas, *Elements of Information Theory*, (John Wiley & Sons, New York, 1991).
[4] F. Rieke, D. Warland, R.R. van Steveninck, and W. Bialek, *Spikes: Exploring the Neural Code*, (MIT Press, Cambridge Massachusetts, 1997).
[5] H. Ebel, L.-I. Mielsch, and S. Bornholdt, Phys. Rev. E **66**, 35103 (2002).

[6] We have confirmed a scaling law for the number M_S of messages sent and received (M_R) per user, and find that both scale as $M_{S,R} \sim M_{S,R}^{-1.0\pm 0.1}$. However, the degree of the graph, i.e. the valence or number of links going in $v_{\rm in}$ and out $v_{\rm out}$ of a node, does not have a simple scaling behavior. Instead, one power law of $v^{-0.93}$ characterizes the behavior of both $v_{\rm in}$ and $v_{\rm out}$ till $v = 12$, where a sharp transition to a second power law with $v^{-1.84}$ fits the behavior till $v \approx 200$ where our data end.

[7] R. Guimera, *et al*, cond-mat/0211498. (2002).

[8] J.R. Tyler, D.M. Wilkinson, B.A. Huberman, cond-mat/030364 (2003).

[8] J.M. Kleinberg, Jour. ACM **46**, 604 (1999).

[9] R. Milo, *et al*, Science **2**98, 824 (2002).