

One in Four Is Enough – Strategies for Selecting Ego Mailboxes for a Group Network View

Antonio Zilli, Francesca Grippa, eBMS-ISUFI, University of Lecce, Italy

Peter Gloor, Robert Laubacher, MIT CCS

{antonio.zilli, francesca.grippa}@ebms.unile.it; {pgloor, rjl}@mit.edu

Abstract

Recently, researchers have started analyzing e-mail archives of individuals and groups as an approximation of social ties. It can be hard to obtain complete e-mail archives covering all exchanges between a group of individuals. Frequently, only e-mailboxes of a subset of the analyzed actors are available for analysis.

In this project we report on some experiments to find the best ego networks (i.e. mailboxes) to give a “reasonably” complete picture of the full social group network. We also report on the stability of social network metrics with respect to incomplete networks.

We have collected the complete individual mailboxes over a period of 20 weeks of 53 researchers working in the same lab, collaborating on different (research and educational) projects. We have done a series of simulations to identify the best strategies and metrics for analysis of incomplete e-mail networks. Applying snowball sampling and subsequently adding more members of the group, we have compared a globally optimal selection strategy, adding the next-best member with respect to the chosen metric, a locally best strategy, adding the next best member within the already known network, and a random selection strategy. As sampling metrics, we used individual and group betweenness centrality, group density, number of nodes and edges, and others. We have categorized ego networks by roles of individual actors as lab manager, project and subproject managers and project contributors. Lab managers and project managers are in the core, individual contributors are in the periphery of the group network. Results show that good approximations of group network structures are already obtained with 25% to 30% of the mailboxes of the community.

1. Introduction

Analyzing Internet-based communication flow opens up new chances to increase knowledge worker productivity for our networked society. Opportunities range from the speed with which knowledge is made available within a large community to the capability of monitoring patterns of how knowledge spreads. Web tools enable virtual teams to collaborate in a synchronous way, without being forced to be co-located: exchange of information can be realized via web applications that computerize the iterative tasks of distributing knowledge to others.

From a management point of view, studies on networks and on the networked society can use information about communication flow and interactions among actors of a community to improve team performance by improving processes and technological tools. Data stored in log files of applications or e-mail archives contains metadata such as the owner of information, sender and recipients of e-mail, users of knowledge, date of the exchange and the exchanged information itself. Many aspects of a team’s capability and performance are analyzed with a network approach: viruses or ideas spreading among connected PCs [1,2] formal and informal organizational structures [3, 12], or collaborative properties of teams [4, 10].

One of the main challenges of studying networks in organizations [13] is to obtain reasonably complete network data. In conventional network analysis, researcher had to manually collect information about who had communicated with whom by interviewing study subjects, or by convincing them to fill out a survey. Recently, electronic archives such as e-mail logs have

alleviated this task. They have introduced, however, new technical and organizational challenges. On the one hand, members of a network might all be using different, incompatible email systems. On the other hand, even if all email data would be available, researchers have to overcome large privacy and confidentiality concerns. Because of this reasons, social network researchers frequently have to accept incomplete electronic email archives. However, earlier work has already shown that incomplete data can indeed return a reasonably good view of the entire network structure [8, 9].

Our own work contributes to this field of study, by examining how large a subset of a group of actors is needed to get a reasonably close approximation of the entire group network. To put it in other words, we are exploring the question of how many ego networks of a group have to be combined to get an approximation of the group network.

2. Background

In our work we are studying the social network based on e-mail exchange. Clearly e-mail communication logs are not enough to capture all interaction between group members, as group members usually use many other communication channels, such as face-to-face meetings, phone, chat, etc that complement email communication. In a related project we are studying the correlation between different communication channels [6].

In interview and questionnaire based data collection for social network analysis, respondents can only report their immediate communication partners. In e-mail based analysis, ties between alters not involving ego can automatically be discovered, e.g. through “cc”-relationships (for an example see figure 2). In addition, respondents might give biased answers to survey questions, or might just forget some interactions they had with others [16]. In e-mail based analysis, all interaction is collected automatically. On the other hand, inclusion of non-task related communication, such as jokes or invitations to barbecues as well as spam e-mail might distort the picture. This means that also in this case data cleanup might be required.

Obtaining e-mail archives from all members of a community is not an easy task. Frequently researchers will have to settle for a fraction of all mailboxes of the members of a community. The goal of our study is to give guidelines for the collection and evaluation of partially complete electronic communication archives. Compared to earlier work [9], which was based on simulation, our work is based on actual data. We rely on insights by [8] about the stability of centrality measures in incomplete networks.

Email traffic within the community of eBMS-ISUFI (eBusiness Management School - Institute for Advanced Multidisciplinary Studies, www.ebms.it) was collected as a basis for this work. eBMS-ISUFI is an advanced research centre on e-business, based in Lecce, Italy, where both research and educational projects are carried out. This community is characterized by a high level of technological skill, high collaboration among individuals and the existence of subgroups. The research activities are project oriented; educational programs (a one year Master and the Ph.D. program) are focused on research projects. In our project we monitored a community consisting of 53 people¹. People were categorized in 6 roles: “decision makers”, “decision makers and coordinators”, “coordinators”, “contributors”, “students” – master and Ph.D. students – and “project-oriented researchers” (see Table 1). eBMS is managed via a coordination group that takes final decisions on the most important questions regarding strategy, projects, partnerships and the general scientific focus of the activities. “Decision maker” is the person of the monitored community who belongs to this coordination group and who doesn’t have any other direct involvement in the project activities of contributors. At the same time she is the direct coordinator of the scientific activities of the Ph.D. students. Other people that belong to the coordination group are directly involved in the coordination of the project teams and tasks, they act like a communication channel in which knowledge flows from the top management to individual contributors. This is the reason why they are called “decision makers

¹ A few people didn’t give their authorization for using their communication in the analysis. These actors were excluded from the community.

and coordinators”. “Coordinators” are researchers who lead teams of “contributors” who work together on a task. In addition, eBMS is realizing a more complex project in collaboration with a university in Morocco, and a group of researchers is involved exclusively in this project: researchers in this group are called “project-oriented researchers”.

E-mail sent and received by these people was collected for 20 weeks, this way 53 complete e-mail archives were built.

Community description	
Decision makers	1
Decision maker and coordinators	7
Coordinators	5
Contributors	11
Students	22
Project oriented researchers	7
<i>Total actors in the community</i>	<i>53</i>

Table 1: Composition of the community.

These archives were analyzed using the social network analysis and visualization tool TeCFlow (Temporal Communication Flow Analyzer) [17].

We evaluated the social structure of the network and how network properties are affected by the incompleteness of data by subsequently adding mailboxes and then calculating the resulting network parameters. Three sampling strategies were tested to find the best one for a certain scenario: a globally optimal selection strategy, a locally best strategy and a random selection strategy.

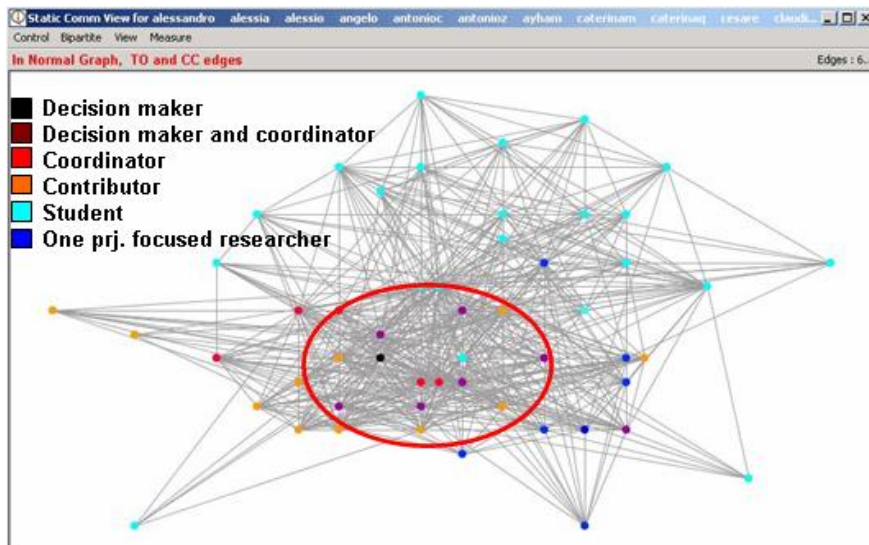


Figure 1: View of the structure of the network, nodes are coloured depending on the role of the actor. The red circle is the core of the network: all decision makers and coordinators are inside.

3. Structure of the Network

TeCFlow visualizes strength of the interactions among people by measuring the number of e-mails exchanged. The closer two nodes are together, the more e-mail the two actors have exchanged. Figure 1 shows the structure of the full eBMS e-mail network, nodes are coloured by the role of the actors.

This view shows that the network consists of two parts: one part is the staff members, the other is made up by the graduate students, only few of them are well-connected to the staff community. This is because most of the students joined eBMS less than three months before the

start of this project, they were therefore not yet well connected to staff members. The well-connected PhD students have been involved in long-term research projects.

Not surprisingly, the students attending the same classes communicated mostly among themselves (top area of Figure 1). Staff members exchanging higher numbers of e-mails are in the bottom area, they are mostly “decision makers and coordinators” (purple nodes). They surround the director of the school, coordinating the different activities of the school. Single-project focussed contributors are in the periphery of the network. There is just one exception, a “decision maker and coordinator” located outside of the core: he is involved in activities that make him more connected to external partners (not present in the analyzed community) than to eBMS people. In summary, there are no surprises here between an actor’s role at the school and her network position.

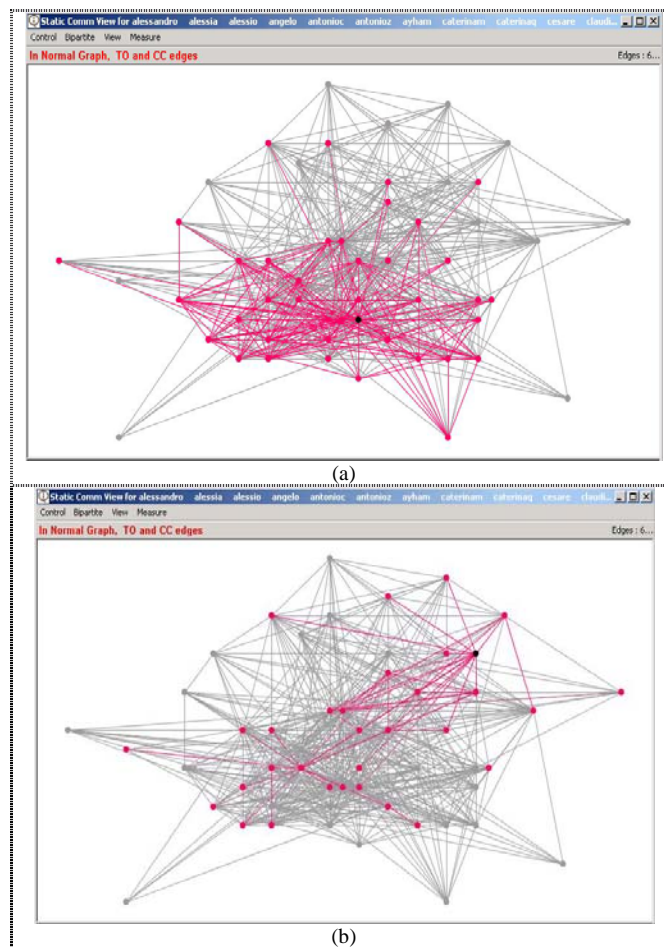


Figure 2: Some ego networks embedded (red) in the complete network (gray). The black dot is the owner of the mailbox that produced the ego network.

In Figure 2, two sample ego networks embedded in the group network are presented. Figure 2.a is the ego network of a “decision maker and coordinator” who was project coordinator of a few projects. He is mostly communicating with permanent eBMS staff, and with a few Ph.D. and masters students whose research activities are related to his projects. In Figure 2.b the ego network of a Ph.D. student is shown. She is connected mainly with other Ph.D. students, and with some researchers involved in her research activities, and with the “decision maker” (director of the Ph.D. program).

4. Experimental Setup

In order to study the impact of sampling, the benchmarking strategy was based on the availability of detailed knowledge about the complete network and about the ego-networks’

properties. All archives were merged to create one dataset (duplicated e-mails were automatically eliminated by the TeCFlow application). This way a detailed and accurate evaluation of the most important network properties was obtained (Table 2).

Network properties of the community	
Number of actors	53
Time interval	01/07/05 - 23/11/05
Number of messages ²	6826
Group Betweenness Centrality	0,0959
Group Degree Centrality	0,5637
Density	0,2192
Total Number of edge:	604

Table 2: Network properties evaluated using all archives.

In a second phase each ego-network was analyzed and its characteristics were tabulated (see Appendix 1).

To study the changes in the global network properties we added one ego archive after the other. We used different variables to determine the merging order:

1. local betweenness centrality of the mailbox owner obtained from the complete group network, from higher to lower values (Figure 3.a);
2. local betweenness centrality of the mailbox owner obtained from the complete network, from lower to higher values (Figure 3.b);
3. global betweenness centrality of ego-networks, from higher to lower values (Figure 3.f);
4. density of ego-networks, from higher to lower values (Figure 3.d);
5. density of ego-networks, from lower to higher values (Figure 3.e);
6. number of edges in ego-networks, from higher to lower values (Figure 3.c);
7. size (number of actors) of ego-networks, from higher to lower values (Figure 3.h);
8. size (number of actors) of ego-networks, from lower to higher values (Figure 3.i);
9. number of received e-mail, from higher to lower values (Figure 3.g).

For each of these metrics, the mailbox of the “next best member not already added” was added to the previous ones and each time the new network was analyzed and new values of the monitored parameters (density, betweenness and degree centrality) were computed. We call this the “optimal” selection strategy. This strategy is interesting for comparing results but not effective as a sampling strategy because if all mailboxes are available no sampling is required. Instead if a social network analysis has to be realized on a large team we face the task of connecting the right data (that is mailboxes).

The evolution of the network parameters was analyzed to look for the minimum number of e-mail archives needed for group density, betweenness and degree centrality values to differ not more than 25% and 10%, respectively, from the group network parameters. The values of the analyzed parameters are, by definition, between 0 and 1. Among the three analyzed parameters the value of group betweenness centrality is the smallest (see Table 3). The implication of this is that to be in the 25% and the 10% range gets more difficult. The result is that more mailboxes are needed for reaching the threshold for group betweenness centrality than for the other two parameters.

Curves obtained with this strategy doesn’t always show flat shapes while sample increases. On the contrary some of the presented plots show very abrupt steps. Many reasons could explain this feature. First of all, usually 5 to 6 mailboxes are needed to connect all members of the community, because not all members are directly connected with all people within the community (on the contrary, only one actor is fully connected with all other 52 people). There are some people who have exceptional networks: their ego-network exhibits high global betweenness centrality and even in the group network they have a high value for local betweenness centrality. They are like hubs of a scale-free network [14]. Their addition to the

² Messages with many recipients are counted once per recipient.

group network leads to a steep change in the slope of the curve. As example see Figure 3.c: the 16th mailbox introduces a spike in global betweenness and degree centrality. This means that just adding a few edges can have a strong impact on the connectedness of the network.

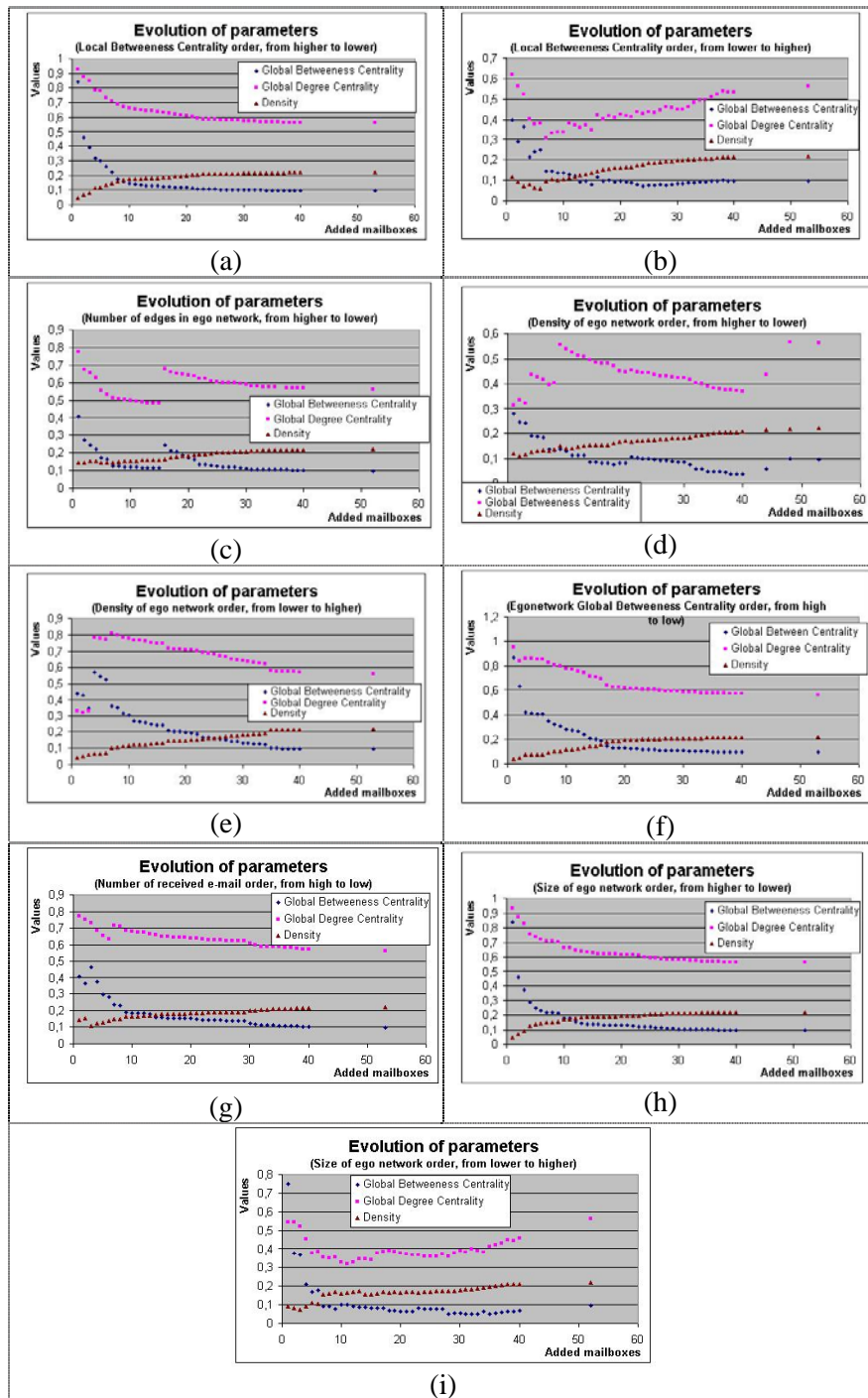


Figure 3: Evolution of network parameters obtained with optimal selection strategy.

The experimental conclusion is that a few people, the hubs that connect the different parts of the community, determine the global characteristic of the community. While an individual contributor tends to collaborate with a limited number of other members, the coordinators and decision makers interact with many subgroups (although with small frequency), increasing the values of group betweenness and degree centrality (and therefore reducing distances between actors).

5. Experiment 1: Locally Best Selection Strategy

Our experimental strategy was based on the hypothesis that best convergence will be obtained by analyzing the emergent group network. Our algorithm started by choosing a random ego network (mailbox). The next ego network to be merged was selected by looking at actor betweenness and degree centrality values within the emergent group network: the mailbox of the most central member by betweenness (or degree) of the emergent group network not yet included was added. Then the procedure was applied again and the evolution of the global parameters was studied.

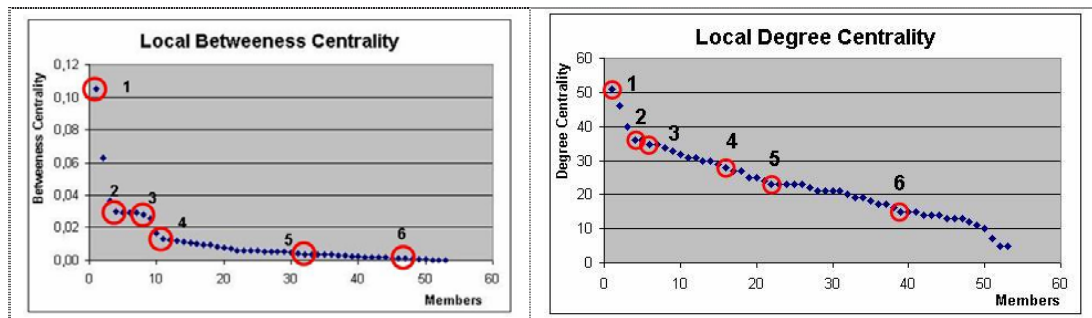


Figure 4: Dots in the circles represent different actors selected as “random” starting point for locally best strategy (53 actors).

This strategy was repeated for different “starting” actors in order to understand how ego network properties of the starting actor impacted the construction of the group network. In particular we looked at the betweenness centrality distribution of the group network and experimented with different starting egos (“1”, “2”, “3”, “4”, “5”, “6” in Figure 4).

The convergence curves of the local best strategy are flatter than for the optimal strategy discussed in the previous section and shown in figure 3. The first few points in the curve now display a more ordered behaviour than with the globally optimal strategy because now a connection exists in the succeeding mailboxes: the egos with locally highest betweenness/degree centrality not already merged will be added. Generally, values of the three parameters converge toward the final values more quickly.

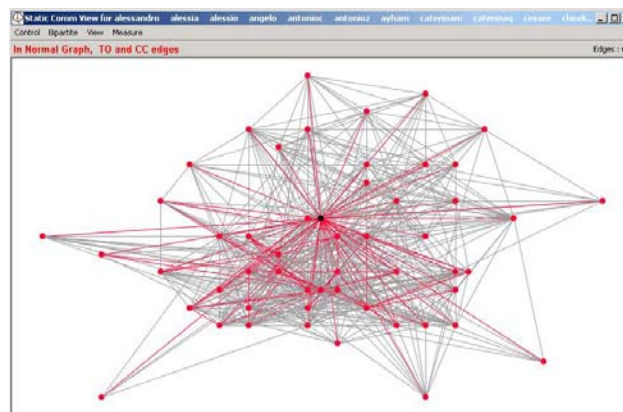


Figure 5: The ego network of the exceptional actor, she had to send (and to receive) many direct e-mails.

Another interesting feature is that the abrupt steps in the curves do not appear with this sampling strategy. This is because the mailboxes that introduce those steps are integrated in the sample in the initial phase of the procedure: those mailboxes belong to the hub connecting the other actors.

One of the ego networks deserves further discussion (it is shown in Figure 5): it is very centralized, it does not have a high number of edges (120 in the ego network out a total of 604 edges in the group network) and many of them connect ego directly with other people. These

structure is the result of a particular task where this member of the community had to send (and receive) many individual e-mails during a short period of time. The impact of this particular ego network was evaluated by deleting this ego and reproducing the same analysis with 52 actors. Results did not change much (see Figure 6 and Table 3).

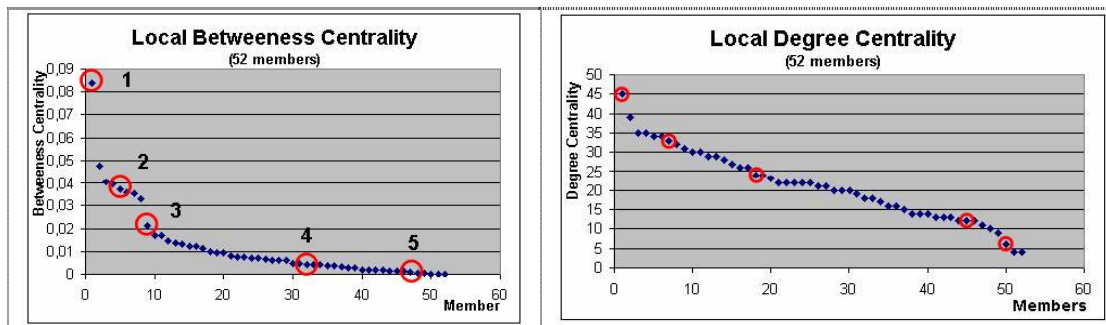


Figure 6: The most centralized ego network was purged from the community and the second strategy evaluated again (52 actors)

6. Experiment 2: Random Selection Strategy

The final sampling strategy tried to find out the minimal number of mailboxes necessary to have a good view of the complete network by random sampling. E-mail archives were merged one after the other in no specific order and each time global network properties were extracted. Three different random samples were studied.

With the random sampling abrupt changes in the values of the parameters occur (fig. 6). They're more radical than in the global optimal strategy, they appear when the mailbox of a hub is merged into the sample. In fig. 2.a and fig. 5 we had introduced two hubs, when one of them is merged to the sample, the network changes radically.

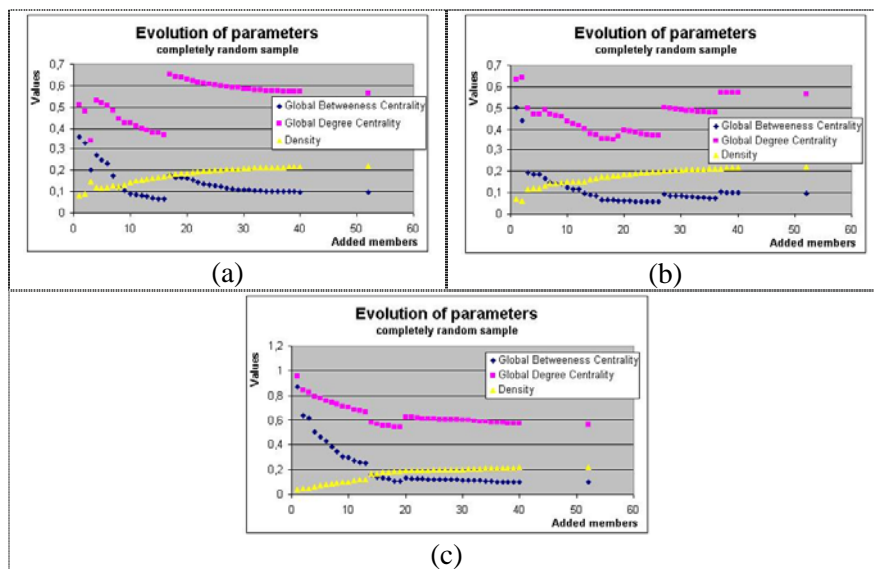


Figure 6: Evolution of network parameters obtained with random selection strategy.

7. Discussion

Table 3 shows the summary of the research results. Adding mailboxes one after the other to the group network moves the emerging network toward the final group network, although with different speed depending on the chosen strategy. The last column stands out: the average of the number of mailboxes necessary for reaching a defined limit for each global parameter has a very high standard deviation, about 30% of the value most of the time. Among the three examined

network parameters, betweenness centrality is the most difficult to approximate (except in one case: fig. 3.b [10]). On the other hand it is the most stable with respect to the sampling strategy, at the same time it needs a higher number of mailboxes to get within the range. This seems to suggest that this property is more deeply connected with the complete network structure while the other two are more dependent on subsets of the community.

Looking at the numbers, choosing the best strategy (Local Betweenness Centrality - low to high), with 12 of 53 mailboxes, that is a sample of 22%, two of the network parameters are as close as 25% of the final value; after merging 19 of the 53 mailboxes, that is a sample of 36%, all three parameters are in a range of 25% of the final value, while two of them are in the 10% range.

Community with 53 people										
Global Optimal Strategies										
Merging parameter	GBC		GDC		Density		Average	Stand. Dev	Average	Stand. Dev
	25%	10%	25%	10%	25%	10%	25%		10%	
LBC - Higher to lower	19	25	8	19	9	21	12	6,1	22	3,1
LBC - Lower to higher	13	32	22	36	22	31	19	5,2	33	2,6
Ego network GBC - Higher to lower	23	32	16	21	17	25	19	3,8	26	5,6
Density - High to Low	>40	>40	>40	>40	19	35	/	/	/	/
Density - Lower to higher	35	35	23	35	27	36	28	6,1	35	0,6
Edges in ego - Higher to lower	27	31	2	24	17	25	15	12,6	27	3,8
Egonetwork size - Higher to lower	24	30	9	20	10	24	14	8,4	25	5,0
Egonetwork size - Lower to higher	>40	>40	>40	37	17	36	/	/	/	/
Received e-mail - Higher to lower	31	38	9	30	11	31	17	12,2	33	4,4
Core members	-	-	8	-	9	-				

Community with 53 people										
Locally Best Strategy										
Position in the local betweenness centrality list	GBC		GDC		Density		Average	Stand. Dev	Average	Stand. Dev
	25%	10%	25%	10%	25%	10%	25%		10%	
1	23	31	10	22	14	24	16	6,7	26	4,7
4	23	31	2	22	14	24	13	10,5	26	4,7
8	23	31	2	22	14	24	13	10,5	26	4,7
11	23	31	4	22	14	24	14	9,5	26	4,7
32	24	30	4	21	15	24	14	10,0	25	4,6
47	24	32	5	23	15	25	15	9,5	27	4,7
Average Number of Mailboxes	23	31	5	22	14	24				
Standard Deviation	0,5	0,6	2,9	0,6	0,5	0,4				

Community with 52 people											
Locally Best Strategy											
		GBC		GDC		Density		Average	Stand. Dev	Average	Stand. Dev
		25%	10%	25%	10%	25%	10%				
Position in the local betweenness centrality list											
1		30	32	15	27	16	28	20	8,4	29	2,6
5		28	32	10	24	13	24	17	9,6	27	4,6
9		28	32	8	23	10	24	15	11,0	26	4,9
32		29	33	14	23	14	25	19	8,7	27	5,3
47		30	34	14	23	14	25	19	9,2	27	5,9
Average Number of Mailboxes		29	33	12	24	13	25				
Standard Deviation		1,0	0,9	3,0	1,7	2,2	1,6				
Community with 53 people											
Random Strategy											
		BC		DC		Density		Average	Stand. Dev	Average	Stand. Dev
		25%	10%	25%	10%	25%	10%				
random 1		28	32	17	22	15	24	20	7,0	26	5,3
random 2		27	37	27	37	15	27	23	6,9	34	5,8
random 3		26	35	11	22	14	27	17	7,9	28	6,6
Average Number of Mailboxes		27	35	27	18	26	15				
Standard Deviation		1,0	2,5	8,7	8,1	1,7	0,6				

Table 3: Number of mailboxes necessary to obtain a value of global parameters in 25% and 10% range of the final group value. Average values on the right concern the number of mailboxes for reaching goal for three parameters, average values under each strategy concern different implementations of the particular strategy.

Under everyday circumstances complete datasets frequently are not available and a method to identify the best egos to add is useful. This first result shows that it is possible to reproduce network characteristics of a complex community with a good approximation with much less than half of the total mailboxes. With locally optimal selection strategy, 25% of the mailboxes can give a reasonable approximation of the entire group network.

Even with an optimal selection strategy as described in section 4, global betweenness centrality is the most difficult parameter to approximate, and generally more mailboxes are needed than for the other two parameters. In the best case (53 people) with 14 mailboxes (26%) two parameters are in the 25% range, and with 24 (45%) two parameters are in the 10% range and the last one (betweenness centrality) is in the 25% range. These results do not change significantly when the exceptional actor is removed: for reaching the same quantitative results 10 (19%) and 23 (44%) of the mailboxes are needed.

At the same time it is clear that an uneven sample of mailboxes can give a completely false view of the network. In our experiments, density of individual ego network was not useful as a sampling strategy. An ego network with a small density value can exhibit a high value of centrality. When a low density and high centrality mailbox is added to a large sample, it introduces few edges but these could be influential in reconfiguring the network, therefore changing the global characteristics of the network.

One interesting result is obtained with the “locally best” strategy is that the sampling strategy is quite stable regardless of what starting ego is chosen. The standard deviation of the average number of mailboxes for different starting actors (rows under each sub tables) are the smallest ones, and among these the standard deviation relative to betweenness centrality is the smallest. This suggests that this measure is more stable with respect to the actor with which the procedure starts.

As was expected, random sampling is the worst method to rapidly create a group network. Figure 6 illustrates that when the most distinctive mailboxes, the mailboxes of hubs, are merged into the sample, the global network properties change abruptly and it is very difficult to extract any satisfying conclusions until an high fraction of data is used.

8. Conclusions

Problems arise when “real world” networks are analyzed. Online communities are growing, but applications for automatic social network analysis are only now slowly emerging. This work contributes by making analysis of large online networks more manageable. We have analyzed a fully connected e-mail network using TeCFlow. We first identified the structure of the community and the roles of key members. This permitted us to better understand the influence of adding egos to the group network.

This study shows that using a good sampling strategy 25% to 36% of mailboxes produce a reasonable value for global network parameters (global betweenness and degree centrality and density). Starting from a random ego network and extracting the person with highest “betweenness centrality” and merging his/her mailbox to the previous one and repeating this procedure, leads to a good approximation of the group network. With 26% of mailboxes two parameters are in the 25% range, and with 45% two parameters are in the 10% range and the last one (betweenness centrality) in the 25% range. This result is influenced by the small value of betweenness centrality in our fully connected network that makes it difficult to reach the 10% range.

An important result of our work is that the building of the sample by a “locally best selection strategy” is independent of the first mailbox analyzed. A reasonably good approximation of the network is obtained whether the starting ego is the most important decision maker of the organization or an individual contributor.

Acknowledgements

Our thanks to the technical staff of eBMS who helped us to set up a hardware and software system to collect such a huge amount of data ready for the analysis without affecting the network performance of the laboratory. Special thanks go to E. Rizzo and to Eng. M. Franza.

References

- [1] Pastor-Satorras, R., Vespignani, A.; Epidemic dynamic and endemic states in complex networks; *Phys. Rev. E*, Vol. 63, 066117
- [2] Pastor-Satorras, R., Epidemic dynamic in finite size scale-free networks, *Phys. Rev. E*, Vol 65, 035108(R)
- [3] Tyler, J. R., Wilkinson, D. M., Huberman, B. A.; Email as spectroscopy: automated discovery of community structure within organization; arXiv.org:cond-mat/0303264
- [4] Kidane, Y., Gloor, P.; Correlating Temporal Communication Patterns of the Eclipse Open Source Community with Performance and Creativity; NAACSOS Conference, June 26 - 28, Notre Dame IN, North American Association for Computational Social and Organizational Science, 2005
- [5] Gloor, P., Laubacher, R., Dynes, S. B.C., Zhao, Y., Visualization of Communication Patterns in Collaborative Innovation Networks: Analysis of some W3C working groups; ACM CKIM International Conference on Information and Knowledge Management, New Orleans, Nov 3-8, 2003.
- [6] Grippa F., Zilli A., Laubacher R., Gloor P., E-mail may not reflect the social network, International Sunbelt Social Network Conference 2006, 2006
- [7] Burt, R. S.; Structure holes: the social structure of competition, Harvard University Press, Cambridge, MA
- [8] Costenbader, E., Valente, T. W.; The stability of centrality measures when network are sampled, *Social network* 25 (2003) 283-307
- [9] Borgatti, S. P., Carley, K., and Krackhardt, D.; (in press) On the robustness of centrality measures under conditions of imperfect data. *Social Networks* this article has now been published, here is the citation information: *Volume 28, Issue 2, May 2006, Pages 124-136*
- [10] Gloor P.; Swarm Creativity: Competitive Advantage through Collaborative Innovation Networks; Oxford University Press, 2006.
- [11] Ebel, H., Mielsch, L., Bornholdt, S.; Scale-free topology of e-mail networks. arXiv:cond-mat/0201476v2 12 Feb 2002.
- [12] Guimera, R., Danon, L., Diaz-Guilera, A. Giralt, F., Arenas, A.; Self-similar community structure in organizations. ArXiv:cond-mat/0211498 v1, 22 Nov 2002.
- [13] Mintzberg H.; Mintzberg on Management, inside our strange world of organization, The free press, 1989.
- [14] Barabási, A.-L., Bonabeau, E.; Scale-free networks, *Scientific American* 288, 60-69 (2003).
- [15] Kogovsek, T., Ferligo, A.; Effects on reliability and validity of egocentered network measurements. *Social Networks*, 27, 3. 205-229. 2005.
- [16] Marin, A.; Are respondents more likely to list alters with certain characteristics?: Implications for name generator data. *Social Networks*, 26, 4. 289-307. 2004
- [17] Gloor, P., Zhao, Y.; TeCFlow - A Temporal Communication Flow Visualizer for Social Networks Analysis, ACM CSCW Workshop on Social Networks. ACM CSCW Conference, Chicago, Nov. 6. 2004

Appendix 1

In the Table 4 characterization of ego networks and of actors in the complete network are reported. Each actor is defined by his/her role (first column). Roles are described in the Table 5.

role in the community	size of ego network (number of actors)	number of edges in the ego network	density of ego network	actor's BC in ego network	actor's BC in complete network	GBC of the ego network	actor DC in complete network ³	# received mail (to+cc)
5	38	116	0,0825	0,3827	0,0059	0,3614	24	162
4	15	19	0,0905	0,7857	0,0002	0,7508	11	24
4	28	143	0,1892	0,2455	0,0036	0,2227	23	152
2	39	161	0,1086	0,5382	0,0282	0,5286	35	330
3	29	97	0,1195	0,3695	0,0128	0,342	28	81
3	32	164	0,1653	0,2704	0,0057	0,2525	27	193
5	29	40	0,0493	0,4841	0,0037	0,6023	15	58
5	38	61	0,0434	0,3163	0,0033	0,4219	15	28
5	37	59	0,0443	0,1775	0,0005	0,5075	10	15
3	37	174	0,1306	0,5436	0,0299	0,5369	36	337
2	33	88	0,0833	0,6122	0,0103	0,6	31	304
4	30	152	0,1747	0,2655	0,0052	0,2472	23	134
5	32	50	0,0504	0,5462	0,0035	0,5969	16	38
4	33	140	0,1326	0,3608	0,0117	0,3421	30	107
5	20	20	0,05	0,2632	0	0,7122	5	13
6	31	58	0,0624	0,349	0,0015	0,4418	14	136
5	53	120	0,0435	0,8425	0,1053	0,8398	51	277
2	40	226	0,1449	0,414	0,0293	0,4046	40	754
2	32	175	0,1764	0,2525	0,0084	0,2367	31	395
3	36	178	0,1413	0,4643	0,0114	0,4525	32	197
5	38	57	0,0405	0,3302	0,0021	0,4685	14	32
5	40	61	0,0391	0,2673	0,0015	0,6281	13	25
5	38	60	0,0427	0,3048	0,0016	0,4408	14	17
5	27	36	0,0513	0,3918	0,0034	0,587	13	54
5	39	71	0,0479	0,3828	0,006	0,4267	21	76
1	35	54	0,0454	0,8725	0,0294	0,8671	35	130
2	47	92	0,0426	0,8748	0,0628	0,8716	46	369
5	23	23	0,0455	0,4762	0,0006	0,8058	7	16
6	26	49	0,0754	0,3145	0,0004	0,49	13	29
5	35	59	0,0496	0,5053	0,0031	0,56	19	80
5	42	77	0,0447	0,4585	0,0165	0,4308	25	23
5	42	83	0,0506	0,344	0,0074	0,3938	21	27
4	31	68	0,0731	0,3686	0,0023	0,335	18	98
5	41	83	0,0506	0,3311	0,0075	0,3643	23	35
5	29	65	0,08	0,2631	0,0013	0,3766	12	21
3	27	123	0,1752	0,1848	0,0013	0,2189	19	84
4	33	153	0,1449	0,4	0,0122	0,3853	30	198
2	32	146	0,1472	0,3588	0,0107	0,3431	29	242
6	23	69	0,1344	0,386	0,0051	0,3736	23	58
5	42	67	0,0389	0,3026	0,0029	0,4404	15	29
5	42	78	0,0453	0,255	0,005	0,4458	17	30
5	34	68	0,0606	0,354	0,0056	0,4957	17	53
4	28	137	0,1812	0,1981	0,0016	0,2337	21	81

³ The local degree centrality evaluated in ego network and in complete network is the same as it is defined by the incoming and outgoing e-mail, both in personal mailbox.

6	31	81	0,0871	0,4128	0,0059	0,3844	22	146
4	35	93	0,0782	0,5644	0,0096	0,5483	27	129
6	32	79	0,0796	0,5508	0,0054	0,5276	23	93
2	22	55	0,119	0,6033	0,0042	0,589	21	96
4	40	97	0,0622	0,7044	0,0361	0,6963	36	183
4	20	44	0,1158	0,0054	0	0,3976	5	12
4	25	133	0,2217	0,1188	0,001	0,1831	20	135
5	43	165	0,0914	0,5189	0,0258	0,5042	33	81
6	45	116	0,05	0,5968	0,0293	0,5781	34	216
6	35	81	0,0681	0,5273	0,0093	0,5035	25	193

Table 4: For each actor of the community the set of network parameter's value are reported.

1	decision maker
2	decision maker and coordinator
3	coordinator
4	contributors
5	students (master and PhD)
6	project-oriented researchers

Table 5: The connection between numbers used in the first column of the table 1 and the role in the community.