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. However 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. 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. Results show that good approximations of group network structures are already obtained with 25% to 30% of the mailboxes of the community.

1. Introduction

One of the main challenges of studying networks in organizations [1] 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 [5,6].

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.

This work is based on the email traffic within the community of eBMS-ISUFI (eBusiness Management School - Institute for Advanced Multidisciplinary Studies, www.ebms.it). It is an advanced research centre on e-business, based in Lecce, Italy, where both research and educational projects are carried out. Research activities are project oriented; educational programs (a one year Master and the Ph.D. program) are focused on research projects.

A community consisting of 53 people was monitored for 20 weeks. The community was structured in 6 roles (see **Table 1**), including decision maker, coordinators of project teams, individual contributors, and students.

The analysis was done using the social network analysis and visualization tool TeCFlow (Temporal Communication Flow Analyzer) [8]. We evaluated the structure of the network (evaluating some global network metrics) and how network properties are affected by the incompleteness of data by sequentially adding mailboxes. Three sampling strategies were tested: a globally optimal selection strategy, a locally best strategy and a random selection strategy.

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

Table 1: Roles of members of the community.

Network properties of the community	
Number of actors	53
Time interval	01/07/05 - 23/11/05
Number of messages[1]	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.

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.

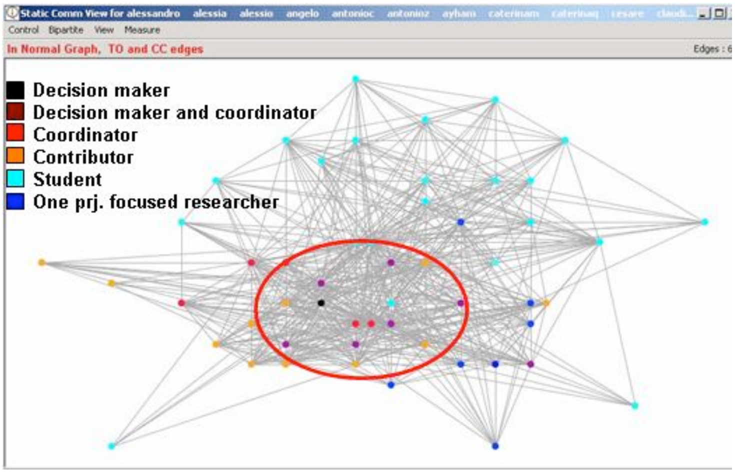

Static Comm View for alessandro alessia alessio angelo antonioc antonioz ayham caterinam caterinaq cesare claudi...
Control Bipartite View Measure
In Normal Graph, TO and CC edges
Edges : 6
■ Decision maker
■ Decision maker and coordinator
■ Coordinator
■ Contributor
■ Student
■ One prj. focused researcher

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.

In Figure 2, two ego networks embedded in the complete 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. Figure 2.b displays the ego network of a Ph.D. student: 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).

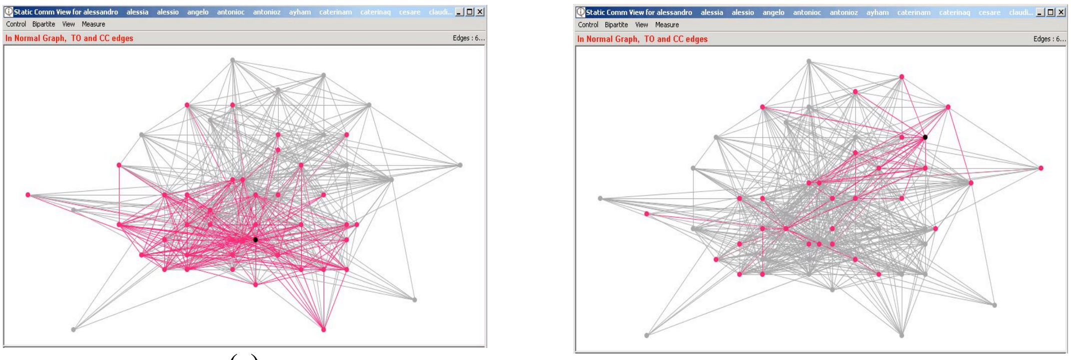
4. Experimental Setup

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: (see also **Table 2**)

1. local betweenness centrality of the mailbox's owner obtained from the complete group network, from higher to lower (Figure 3.a) and from lower to higher values (Figure 3.b);
2. global betweenness centrality of ego-networks, from higher to lower values (Figure 3.f);
3. density of ego-networks, from higher to lower (Figure 3.d) and from lower to higher values (Figure 3.e);
4. number of edges in ego-networks, from higher to lower values (Figure 3.c);
5. size (number of actors) of ego-networks, from higher to lower values (Figure 3.h) and from lower to higher values (Figure 3.i);
6. 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 We call this the "optimal" selection strategy. The evolution of the network parameters was analyzed to look for the minimum number of mailboxes needed for network parameters values to differ less than 25% and 10% from the complete network values. More mailboxes are needed for reaching the threshold for group betweenness centrality than for the other two parameters, as its value is very small (see **Table 2**).

This analysis is interesting for comparing results but it isn't effective as sampling strategy: it is based on the availability of all mailboxes but in this case no sampling is required.


Static Comm View for alessandro alessia alessio angelo antonioc antonioz ayham caterinam caterinaq cesare claudi...
Control Bipartite View Measure
In Normal Graph, TO and CC edges
Edges : 6...
Static Comm View for alessandro alessia alessio angelo antonioc antonioz ayham caterinam caterinaq cesare claudi...
Control Bipartite View Measure
In Normal Graph, TO and CC edges
Edges : 6...

(a) (b)

Figure 2: Two ego networks embedded (red) in the complete network (gray). The black dot is the owner of the mailbox that produced the ego network. In (a) the ego network of a "decision maker and coordinator", in (b) the ego network of a Ph.D. student.

Some of the presented plots in Figure 3 show very abrupt steps. In the starting merging phase, 5 to 6 mailboxes are needed to connect all members of the community, as only one actor is connected with every other members, in this phase values of the parameters change significantly at each step. Moreover 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. The merging of their mailbox to the network leads to a steep change in the slope of the curve (As an 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. While contributors tend to collaborate with few others, coordinators and decision makers interact with members of many subgroups, thus reducing the distance between nodes and changing the values of our network metrics.

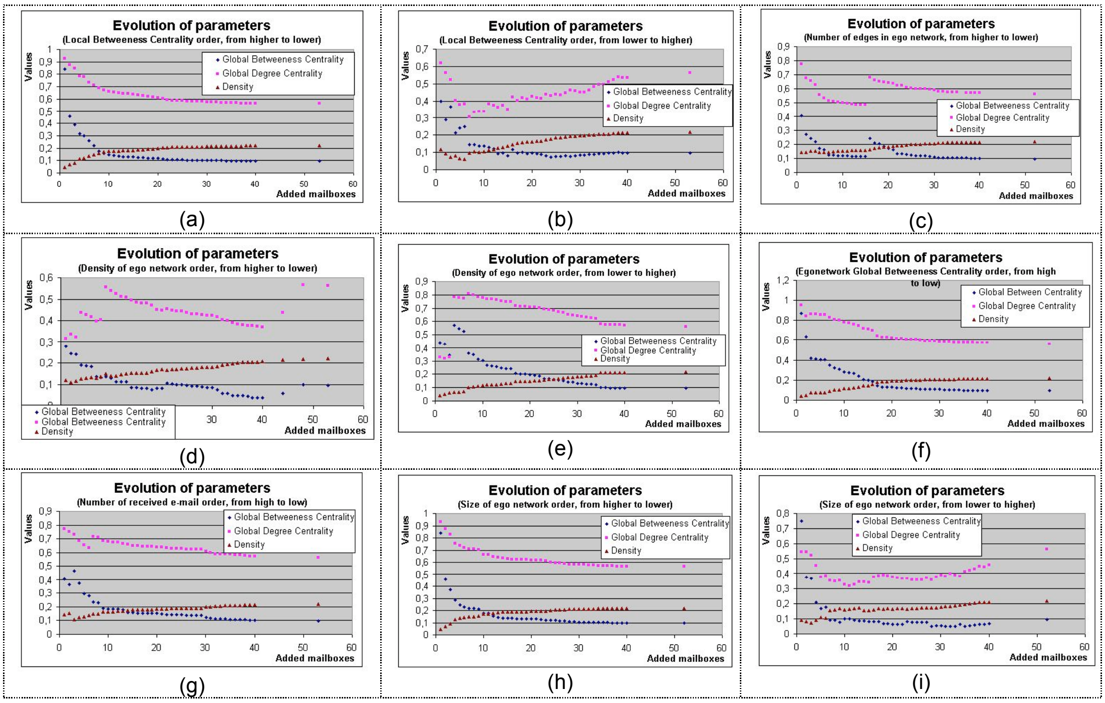
5. Experiment 1: Locally Best Selection Strategy

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

The convergence curves of the local best strategy are flatter than for the optimal strategy discussed in the previous section and shown in figure 2. 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.

6. Experiment 2: Random Selection Strategy

As last strategy e-mail archives were merged in random order. In this case many abrupt changes in the values of the parameters occur (Figure 4) and they're stronger than in the global optimal strategy and appear whenever the mailbox of a hub is merged into the sample.



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

Figure 3: Evolution of network parameters obtained with optimal 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. Betweenness centrality is the most difficult to approximate (it needs a higher number of mailboxes to get within the chosen range), but it is the most stable with respect to the sampling strategy. This property seems to be 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.

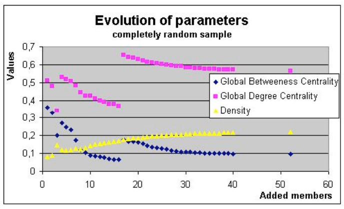


Figure 4: Evolution of network parameters obtained with random selection strategy.

Even with the "locally best strategy", global betweenness centrality is the most difficult metric to approximate. In the best case 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 most exceptional actor is removed (ego with the highest global betweeness centrality): 10 (19%) and 23 (44%) mailboxes are needed to reach the same quantitative results.

At the same time 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. 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.

Global Optimal Strategies

	GBC		GDC		Density		Ave.	Stand. Dev	Ave.	Stand. Dev
Merging parameter	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	-				

Locally Best Strategy

	GBC		GDC		Density		Ave.	Stand. Dev	Ave.	Stand. Dev
Position in the local betweenness centrality list	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				

Random Strategy

	BC		DC		Density		Average	Stand. Dev	Average	Stand. Dev
	25%	10%	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.

One interesting result 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, 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 sampling process starts.

As was expected, random sampling is the worst method to approximate a group network.

8. Conclusions

Problems arise when "real world" networks are analyzed. This work contributes to automatic social network analysis 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.

Using an appropriate sampling strategy, 25% to 36% of the mailboxes of a community produce a reasonable value for global network metrics (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 (global betweenness centrality) in the 25% range. This result is influenced by the small value of betweenness centrality in our complete network that makes it difficult to reach the 10% range.

An other important result 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 independent of whether the starting ego is the most important decision maker of the organization or a more peripheral 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 scal-free networks, Phys. Rev. E, Vol 65, 035108(R)

[3] 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

[4] Grippa F., Zilli A., Laubacher R., Gloor P., E-mail may not reflect the social network, International Sunbelt Social Network Conference 2006, 2006

[5] Costenbader, E., Valente, T. W.; The stability of centrality measures when networks are sampled, Social network 25 (2003) 283-307

[6] 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*

[7] Gloor P.; Swarm Creativity: Competitive Advantage through Collaborative Innovation Networks; Oxford University Press, 2006.

[8] 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

[9] Fruchterman, T.M.J & Reingold, E.M. (1991), Graph drawing by force directed placement. Software: Practice and Experience, 21(11), 1991.

[10] Gloor, P. Zhao, Y. Visualizing Time in Social Networks with TeCFlow, (Web document http://www.ickn.org/JoSS_subm/TeCFlow4JoSS.htm) submitted, 2005.