Analysis of Email Ego Networks

An exploratory study of ego networks in an email network

Angel X. Chang
Computer Science Department
Stanford University
Stanford, CA, USA
angelx@stanford.edu

ABSTRACT

For this project, we look at network properties of ego networks in a email network from a research lab. Due to the nature of the email network, most of the ego networks have a star-like structure. After the removal of the ego, there is one large connected component that roughly corresponds to email addresses within the lab, and isolated nodes corresponding to the external addresses. We look at the effect of tie strength on the connectedness of the ego network. We also try to group nodes in the ego network using Girvan-Newman and by tie strength. We found that using tie strength gives more clearly defined clusters.

1. INTRODUCTION

While there has been many studies on the global properties of a network, there has been less systematic study on the local structure of the neighboring network around a focal node. In this project, we study ego networks where the ego network is defined as the first order neighbourhood of the ego. It is the network consisting of the ego and the nodes directly connected to the ego, and the ties between them.

2. MOTIVATION AND BACKGROUND

Ego networks have been studied in social networks to understand the role an individual plays and how they interact with others.[1] [6] Typical analysis involves calculating various graph properties such as density, degree centrality, closeness centrality, betweenness centrality, and the number of cliques and components, and clustering alters into groups. The expectation is that the clusters would correspond to natural groups such as family or other groups based on school, religion, and hobby. However, these studies are are usually limited in scope. Often, only a small number of ego networks is examined. Ego network information is obtained via questionaire by asking individuals about people they know and in what context.[7]

By studying ego networks, we hope to gain a similar understanding of the differences between individuals based on network structure. For instance, in a organization we would expect ego networks of different people to differ depending their position. In a university, we would expect the ego network of a professor and student to be very different. Thus, by examining ego network of an individual, we can try to characterize the structure of the network and predict what role the individual plays.

Another question of interest is whether we can understand the global properties of the network by looking at the network properties of a sample of ego networks. Often, we may only have data on a sample of the population and we would like to understand properties of the whole network from that sample. In addition, certain measures may be too difficult to calculate directly for the entire network, but can be approximated from a sample of ego network. In other cases, some measures are not necessary meaningful for a large network or may simply be different depending on the ego. For instance, how to cluster nodes into groups may differ depending on the ego since each individual may want to group they people they know differently, because people play different roles in different peoples' lives.

3. EGO NETWORK ANALYSIS

3.1 Email Data

The email network is formed from email data from a research institution, IJS, over a span of 803 days. Because of a time gap in the data (there are virtually no messages from 525 to 797 days), we focus our analysis on the initial 525 days. The email data comprises of roughly 2.4 million messages exchanged between 287,755 email addresses. For our analysis, we discard any messages that have multiple recipients.

We focus on the ego networks of the members of the research lab for which we have the name and the department information. There are 1266 such individuals across 43 departments, with a total of 1407 email addresses (since several individual have multiple email addresses). We have email messages corresponding to 1218 out of the 1407 email addresses. For each of these, we extract the ego network and perform analysis on the ego network. We subdivide the remaining email addresses into the following three groups (we use fake departments to represent these groups):

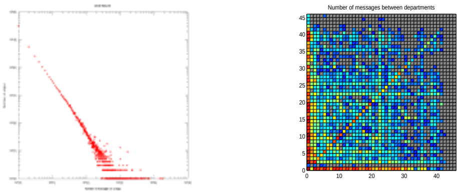
1. External (191,151 addresses, department 0) - External email address that does not appear to relate to IJS

(does not contain ijs.si in the email address).

2. External IJS (74,894 addresses, department 1) - External email address containing the string ijs.si.

3. IJS Other (20,492 addresses, department 2) - Email address belonging to the ijs.si domain that does not correspond to a known individual.

Using each email address as nodes, we create a edge between two nodes if a message was sent from one address to another. To start, we perform a initial analysis of the entire network to get a feel for the global structure of the network. For instance, we take a look at the network degree distribution, which appears to follow a power law. We observe that the distribution of messages across edges also appear to follow a power law (see Figure 1(a)). In Figure 1(b), we show the distribution of messages across departments. The bright diagonal indicates that interally, most of the messages are between nodes of the same department.



(a) Per Edge (b) Between Departments

Figure 1: Number of messages a) Number of messages per edge b) Log of the number of messages between departments

We give a brief summary of some other global properties of the email network in Table 1.

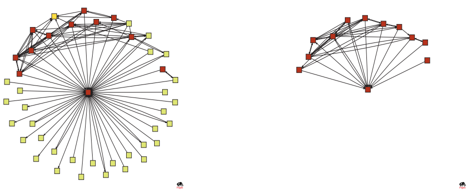
Table 1: Global Properties of Email Network

	All	Internal
Nodes	287755	1218
Edges	397915	20691
Diameter	5.61891	3.93913
Clustering Coefficient	0.2992	0.3804

While the entire email network is large with 287,755 nodes, we will focus most of our analysis on a subset of the network, which is restricted to the 1218 addresses of people in the research lab.

3.2 What does the Ego network look like?

Using the entire email network, with all the 287,755 email addresses as nodes, we find that the ego network is a star-like structure with one large connected component (Figure 2(a)). Upon closer examination, it becomes obvious that many of the spokes in the network corresponds to external addresses (shown in yellow), while the large connected component consists mainly of internal addresses (shown in red). This happens to be an artifact of the nature of the email data. Because we only have emails for the institution, we do not have the data for communications between the external nodes. Thus, we do not know if the external addresses are connected or not. In the rest of our analysis, we will restrict ourself to the 1218 email addresses of known individuals in the research lab. The resulting ego network now consist mainly of one cluster (Figure 2(b)).



(a) All (b) Internal

Figure 2: Sample Ego Network. a) Star-like with one large connected component if all email addresses included. b) Less star-like, one large connected component if only internal email addresses used.

3.3 Effect of Tie Strength

We study the effect of tie strength on an individual's ego network. In particular, we look at how the ego network changes as we throw out weak ties. Using the number of messages as an indication of tie strength, we only include edges between nodes if more than a certain number of messages are exchanged. Taking the directionality of the message into consideration, we construct edges in our ego networks in two different ways:

1. Unidirectional. We include edge from $a \rightarrow b$, if there are x or more messages from a to b. In most of our analysis, we treat the graph as undirected.

2. Bidirectional. We include edge from $a \leftrightarrow b$, if there are x or more message from a to b *and* there are x or more messages from b to a.

We then plot how various network properties change as we increase the threshold on the number of messages, x. For each graph, we plot the uni-directional maximum (blue), average (red), median (green), and the bi-directional maximum (yellow), average(magenta), median (cyan).

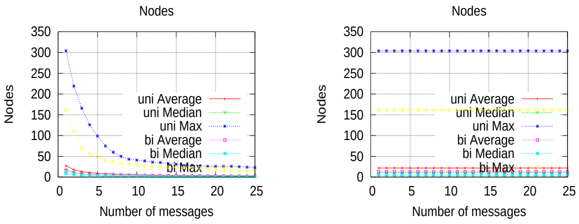
There are also two different ways of including nodes as we construct our ego network based on tie strength depending on whether we choose to keep all alters in the ego network or not:

1. Alters Removed. We remove alters once they are no longer connected to the ego. As an example, if x is 3 and we will not keep any alters in the network that exchanged only 1 or 2 messages with the ego.

2. Alters Kept. We keep alters in the network. In this case, if x is 3, we will still keep the alters that exchanged only 1 or 2 messages with the ego.

3.3.1 Ego network size

Obviously, if we remove alters from the ego network as we increase the tie strength, the size of the ego network will shrink as the number of messages increases. If we keep the alters in the network, then the size of the ego network will be constant. This is shown in Figure 3.

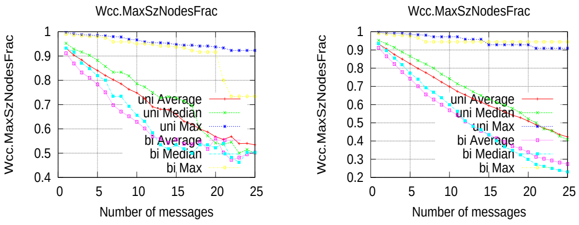


(a) Alters Removed (b) Alters Kept

Figure 3: Size of ego network

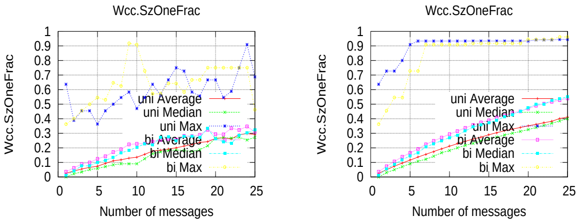
3.3.2 *Distribution of connected components*

Next, we look at the distribution of the sizes of the connected components after the removal of the ego. In Figure 4, we plot the fraction of the number of nodes in the largest connected component, which decreases as we increase the tie strength. Figure 5, shows a corresponding increase in the fraction of the number of isolated nodes. In both cases, we only include ego networks of at least size 10.



(a) Alters Removed (b) Alters Kept

Figure 4: Size of largest connected component (as fraction of ego network size) after removal of the ego

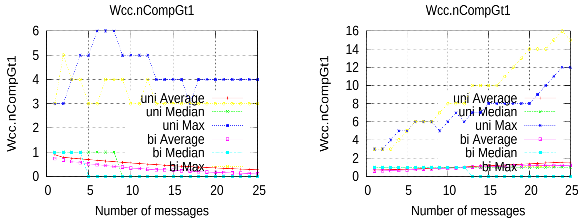


(a) Alters Removed (b) Alters Kept

Figure 5: Number of isolated alters (as fraction of ego network size) after removal of the ego

The increase in the number of isolated alters and decrease in the size of the largest connected component as we increase the tie strength, indicates the ego network is becoming more star-like, with just one connected cluster. To verify that, we examine the number of connected components of size great than one (i.e. components that are not isolated nodes).

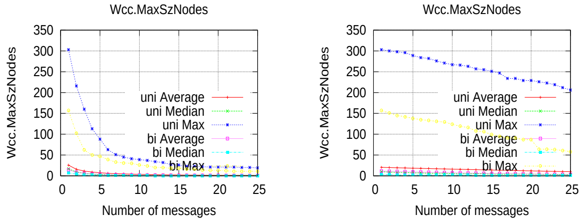
Figure 6 show that the number of non-isolated connected components is very low if we discard alters as we increase the threshold on the number of messages. It is somewhat higher if we keep the alters, but the average is still around 1 or 2, confirming that there is just one large cluster in the ego network. This is likely to be a reflection of the fact that the email network is mainly work related, so we do not get nice clustering of alters into family, work, and other groups. In addition, since we are looking at the all messages across a span of two years, eventually everyone in the research lab that ever communicated will become aggregated into one connected cluster.



(a) Alters Removed (b) Alters Kept

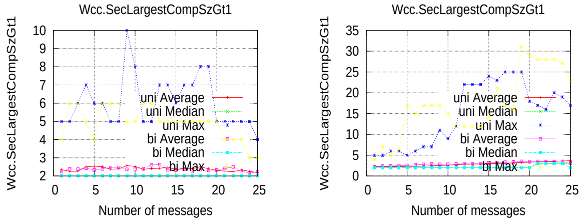
Figure 6: Number of non-isolated connected components after removal of the ego

If we look at the size of the largest (see Figure 7) and the secondlargest connected component (see Figure 8), we see that sizable connected components emerge as we increase the tie strength, but still keep the alters in the ego network. This suggest a natural way to group the alters by using their tie strength to each other. We will explore this later we do hierarchical clustering of the alters based on the number of messages exchanged.



(a) Alters Removed (b) Alters Kept

Figure 7: Size of the largest component after removal of the ego



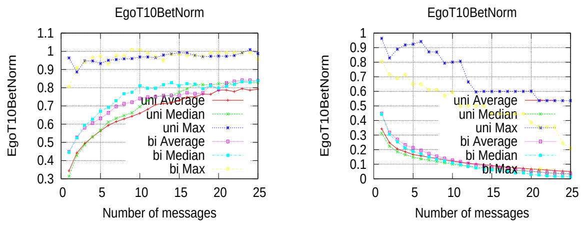
(a) Alters Removed (b) Alters Kept

Figure 8: Size of the second largest component after removal of the ego

3.3.3 *Ego Betweenness*

We also study the betweenness of the ego. In some sense, the betweenness of the ego is a good measure of the ego's importance and how star-like is the ego network. In Figure

9, we plot the normalized betweenness of the ego. Not surprisingly, the ego betweenness increases if we remove alters as we increase the tie strength (since the ego network is becoming more star-like). In this case, the average appears to be asymptoting to 0.8. If we do not remove alters that are isolated from the ego, then the ego betweenness decreases.

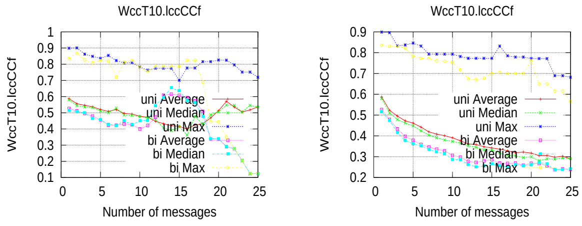


(a) Alters Removed (b) Alters Kept

Figure 9: Normalized betweenness of the ego

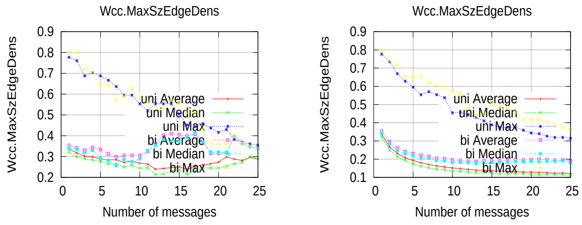
3.3.4 *Graph density*

Next, we examine the graph density of the ego networks as we increase the tie strength. Figure 10 and Figure 11 shows the clustering coefficient and edge density of the ego networks as we increase the threshold on the number of messages.



(a) Alters Removed (b) Alters Kept

Figure 10: Cluster coefficient for the largest connected component (after removal of ego)



(a) Alters Removed (b) Alters Kept

Figure 11: Edge density of the largest connected component (after removal of ego)

If we remove alters isolated from the ego, both graphs indicate an slow decrease in the bidirectional graph density as we initially increase the number of messages. But starting at around 10 messages, the graph density increases, until it abruptly drops off at around 20 messages. The bumps indicates that as we start to shrink the ego network based on tie strength, we are perhaps cutting away alters that are not strongly interconnected. Then at around 10-20 messages, we encounter a tighly connected cluster of alters. But as we increase the threshold further, we cut into the tighly connected cluster and it also falls apart, leaving us with just disconnected alters. This effect is not noticeable if we keep all alters in the ego network. Instead, we see that the average edge density asymptoting to around 0.2 (for the bidirectional case). This may indicate the global edge density for the given threshold on the number of messages. Further analysis is required to understand both phenomenons.

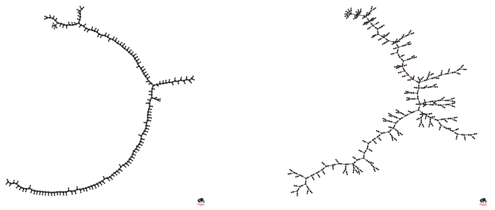
3.4 Grouping of Nodes within the Ego Network

For a given ego network, it is natural to try to cluster the alters into different groups. We attempted hierarchical clustering of nodes within the ego network using two methods:

1. Girvan-Newman[4]. We perform hierarchical clustering by removing edges with the highest betweenness and seeing how the network falls apart.

2. Tie strength. We perform hierarchical clustering by removing edges in increasing order of strength of tie (i.e. number of messages sent between two nodes).

In Figure 12, we show dendrograms of sample clusterings using the 2 methods for a given ego. We observe that using betweenness gives a highly skewed tree, where we are essentially just cutting away whisker nodes. Using tie strength gives more clear clusters (notice the higer level of branching in the dendrogram). However, it is still uncertain whether these clusters correspond to natural groups such as the departments of the research lab. More analysis is required.



(a) Betweenness (b) Tie Strength

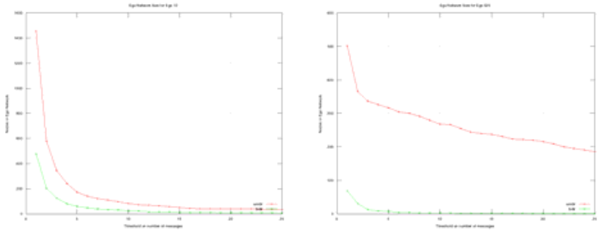
Figure 12: Dendrograms showing hierarchical clustering of ego networks

3.5 Variations between Egos

There is a large variation amongst ego networks for certain network properties. For instance, the size of the ego network (effectively the degree of the node) has high variability, as the distribution follows a power law. Also there are difference in the number of incoming edges and outgoing edges between different egos. There are some nodes that have just in edges, and no out edges, and others that have just out edges and no in edges.

These variations can be indicative of the nature of the ego. For instance, we can compare the ego network size of Ego 625 and Ego 10 for different tie strength (see Figure 13).

Ego 10, which is more typical of other egos in the email network, shows an exponential decay in network size as the threshold on the number of messages increases. Ego 625, on the other hand, has a much flatter decrease in network size. In addition, Ego 625 has very few alters with which it has bi-directional communication. These differences in ego network is suggestive that we should indeed be able to cluster egos based on their network structure.



(a) Ego 10

(b) Ego 625

Figure 13: Comparison of network size of two ego networks (ego network is created from entire email network)

4. CONCLUSIONS AND FUTURE WORK

In this project, we have performed preliminary analysis to study different properties of ego networks. We focused on examining the effect of tie strength on the connectivity of the ego network. Initial analysis indicate that most of the ego networks in the email network has a simple structure of being a star-like network with one large cluster. We showed that using tie strength, we can break apart the ego network into clusters. It still remains to be seen whether these clusters correspond to natural groups such as departments. Also, we saw that there can be large variation in network properties for different egos. It would be interesting to see if we can cluster the egos, either using different features of the ego network or directly through the structure of the ego network.

Another future direction would be to explore the evolution of ego networks over time. We can perform dynamic analysis of the ego networks over time[5], to understand the changing role of the ego as it acquires new contacts[2], and as it disengages from previous contacts. We can also trace the pattern of flow of information through a ego network[3]. Both dynamic aspects can also be used to help us cluster egos.

Finally, we can extend our analysis to other types of networks (e.g. phone networks, IM, Facebook, LinkedIn, etc). In this project, we used data from an email network, reflecting the interactions of individuals within a research lab. But it is limited in that it contained only work related connections, and had limited information about connections bteween individuals that are outside the institution. By looking at other networks, we can get richer information about people's interactions outside of work and have more interesting groupings of alters.

5. REFERENCES

[1] R. S. Burt. Models of network structure. *Annual Review of Sociology*, 6:79–141, August 1980.

[2] M. Everett and S. P. Borgatti. Ego network betweenness. *Social Networks*, 27(1):31–38, January 2005.

[3] J. K. G. Kossinets and D. J. Watts. The structure of information pathways in a social communication network. In *Proc. 14th ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining*, pages 435–443, 2008.

[4] M. Girvan and M. E. J. Newman. Community structure in social and biological networks. *Proc. Natl. Acad. Sci. USA*, 99:7821–7826, 2002.

[5] G. Kossinets and D. J. Watts. Empirical Analysis of an Evolving Social Network. *Science*, 311(5757):88–90, January 2006.

[6] N. Lin. Building a network theory of social capital. *Connections*, 22(1):28–51, 1999.

[7] C. McCarty. Structure in personal networks. *Journal of Social Structure*, 3(1), 2002.