SNAP Final Paper - Team 3

Jordan Bass, Elijah Hansen, Marcus Mok, Pun Tichachol, and Derek Yu

June 8, 2023

1 Introduction

The president of a major European research institution wants to throw social events to improve the relationships among its members. However, previous events have suffered from low turnout, leading to the need to find out who the influential members are that can be reached out to for help in boosting attendance. The president is up to date with Harvard Business Review and knows that informal social networks do not spontaneously promote better production and innovation. Hence, our firm, SNAP, has a goal to present the tools needed to consciously manage these social networks for the betterment of both member productivity and mental health. Most importantly, our firm will show which employees in this social network are brokers, or essential for both intradepartmental and interdepartmental communication.

The main form of communication in this research institution is via email. The raw dataset for the email communication was available via Stanford University's SNAP website by Jure Leskovec. Our SNAP team was able to obtain the network generated using email data from the institution. The emails exclusively represent communication between institution members (the core), and do not contain incoming or outgoing messages outside of the institution. A directed edge (u, v, t) means that employee u sent an email to employee v at time t. An edge is created for each recipient of an email. For the institution as a whole, there are 332,334 temporal edges (representing 24,929 static connections) spanning 803 days and 986 nodes. Within this network, there are 4 sub-networks that hold all communication for each department in the institution. Department 1 has 61,046 temporal edges (3,031 static connections) spanning 803 days with 309 nodes. Department 2 has 46,772 temporal edges (1,772 static connections) spanning 803 days with 162 nodes. Department 3 has 12,216 temporal edges (1,506 static connections) spanning 802 days with 89 nodes. Department 4 has 48,141 temporal edges (1,375 static connections) spanning 803 days with 142 nodes.

There are some benefits and limitations that come with the Eu Core dataset that our team would like to discuss. To start, a strength of our dataset is that it contains a high volume of data spanning over two years of communication at the institution. This allows for

thorough analysis of communication. However, our team also found that this data can be very sparse when truncated by time or number of nodes. This meant that complex analysis via Relational Event Model (REM) for temporal tie prediction resulted in numerous errors for matrix sparsity. Our team believes this sparsity originates from the fact that a member might only send a few emails a day, so when their data is truncated, it leads to lots of sparse data. Another limitation in the chosen dataset is that there are not many roles or labels to match nodes on because the data includes only the sender, receiver, and the timestamp.

2 In-Depth Questions

Our team believes that if we can answer the following questions, we will be able to thoroughly satisfy the needs of our client and their institution. If the President wants to promote social gatherings for their members, they must know who to contact in order to spread information. We know that if we can identify a broker within the network, the President can use this broker to consciously manage the network of members. Hence, we must answer question (a): **Who are the actors in the network with the highest centrality?** Our team plans to utilize local and global properties of networks to answer this question. We will primarily use measures of centrality such as in-degree, out-degree, and betweenness to identify brokers in the network that we believe will better facilitate communication to members that would otherwise not participate in social gatherings.

In addition to finding which members are essential for brokering communication, our team believes that finding which members are isolated from their colleagues is essential for helping our client promote a more supportive and more productive environment for all of their members. Thus, we will answer question (b): **Are there any actors that rarely or never email colleagues within or outside their department?** After identifying isolated individuals within the network, we hope to uncover any patterns among them to minimize isolation in future occurrences.

After obtaining the information on which members are essential for spreading the information of the social event, we can use REM to determine how these members should act to get as many responses as possible. For example, if we find that the likelihood of individual i sending a message to individual j is greater if j has sent a message to i recently, the President can tell the members with high connectivity to send messages as that action is likely to lead to more responses. Because the data has timestamps, we will use REM to answer question (c): **What are the factors that contribute to the formation or dissolution of ties between individuals within the network?**

3 Analysis & Findings

3.1 Data Description

For this project, we obtain data from the email-Eu-core temporal network. This dataset is a text file that contains rows of three numbers in the format of "u v t" that represents a directed edge for the action of person u sending an email to person v at time t (in seconds). Note that even though we have four sub-networks corresponding to the communication between members of four different departments at the institution, we cannot label the data in the core network since node IDs in the sub-networks do not correspond to the same node ID in the core network. However, we are still able to plot these sub-networks independently to study their components. Our network study consists of two parts: local properties analysis and REM.

3.2 Local Properties

To study the local properties, we create a weighted static network in which the weight is calculated by the amount of times u sends an email to v. This allows us to calculate centralities such as in-degree, out-degree, and betweenness. These are the definitions we used:

- In-degree: This is the number of incoming edges to a node in a directed graph. Nodes with high in-degree are seen as influential or popular.
- Out-degree: This counts the number of outgoing edges from a node in a directed graph. Nodes with high out-degree are seen as influential or active.
- Betweenness: This measures the number of times a node acts as a bridge between other nodes in a network. Nodes with high betweenness are seen as important for communication and information flow.
- In-closeness: This calculates how easily a node can reach other nodes in the network via incoming edges. Nodes with high in-closeness are seen as central to the network and can access information from other nodes quickly.
- Out-closeness: This calculates how easily a node can reach other nodes in the network via outgoing edges. Nodes with high out-closeness are seen as central to the network and can disseminate information to other nodes quickly.
- Hub score: This measures a node's ability to connect to other nodes with high out-degree. Nodes with high hub scores are connected to many other nodes with high out-degree and can serve as hubs of information dissemination.

	In-Degree		Out-Degree		Betweenness	
Rank	**Node**	**Score**	**Node**	**Score**	**Node**	**Score**
1	91	211	91	333	91	72626.497
2	160	178	773	226	952	37695.392
3	215	168	121	221	3	27174.022
4	121	156	215	203	121	24704.122
5	952	153	952	201	160	24682.977

Table 1: The top 5 nodes for centrality measures in-degree, out-degree, and betweenness.

	In-Closeness		Out-Closeness		Hub Score	
Rank	**Node**	**Score**	**Node**	**Score**	**Node**	**Score**
1	91	4.62e-4	358	1	775	1
2	701	4.62e-4	537	1	569	2.98e-2
3	952	4.48e-4	773	3.82e-4	451	1.69e-2
4	891	4.45e-4	891	3.80e-4	41	1.05e-2
5	121	4.45e-4	92	3.77e-4	587	1.02e-2

Table 2: The top 5 nodes for centrality measures in-closeness, out-closeness, and hub score.

From Tables 1 and 2, we observe that node 91 is ranked first in in-degree, out-degree, betweenness, and in-closeness. Node 91 holds a position of influence and prominence within the email network. It appears to be a central hub that receives a high volume of emails, actively engages in communication with various individuals or departments, serves as a critical intermediary for information flow, and maintains close proximity to other nodes in terms of incoming communication. Other nodes that should be considered are nodes 358 and 537 because of their high out-closeness. Nodes with high out-closeness are influential in terms of information dissemination. They play a crucial role in ensuring that messages, announcements, or important updates reach a wide audience in the network quickly and effectively. These nodes can be considered as key information hubs or communicators. Lastly, node 775 has the highest hub score, so it is connected to other nodes with high authority or importance. In an email network, nodes with a high hub score are those that frequently send emails to nodes with high in-degree or receive emails from nodes with high out-degree.

Figures 1 and 3 suggest that there are only a few nodes with a high in-degree and out-degree. The log-log plot of in-degree (Figure 2) has an alpha value of 1.077389. The log-log plot of out-degree (Figure 4) has an alpha value of 3.02758. Both plots show a relatively linear negative relationship. The alpha value represents the slope of the line in the log-log plot. A larger alpha value indicates a steeper slope, meaning that the distribution has more high-degree nodes compared to low-degree nodes. Conversely, a smaller alpha value indicates a shallower slope, indicating that the distribution has more low-degree nodes compared to high-degree nodes. In both networks, the log-log plots show a relatively linear negative relationship, indicating a power-law distribution. The alpha values of both networks' in-

degree and out-degree distributions are greater than 1, indicating that the distributions are right-skewed, with a few nodes having a very high degree and most nodes having a lower degree.

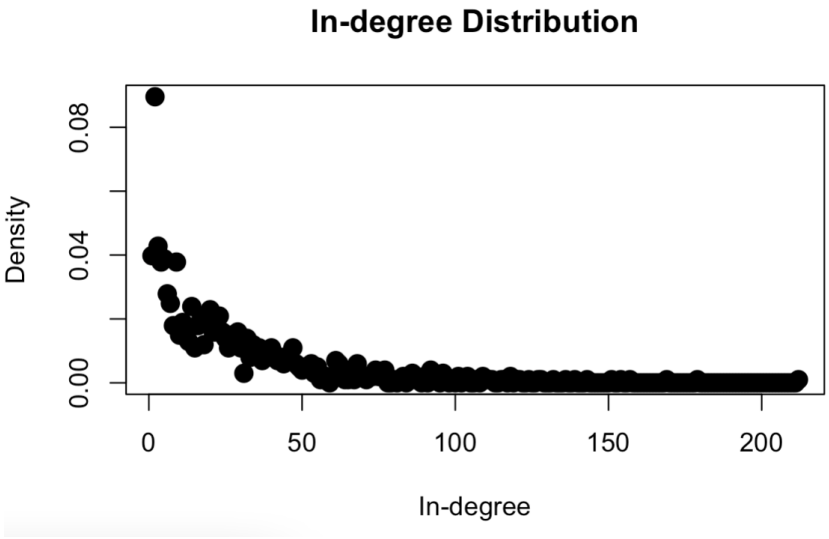

Figure 1: The in-degree distribution for the network.

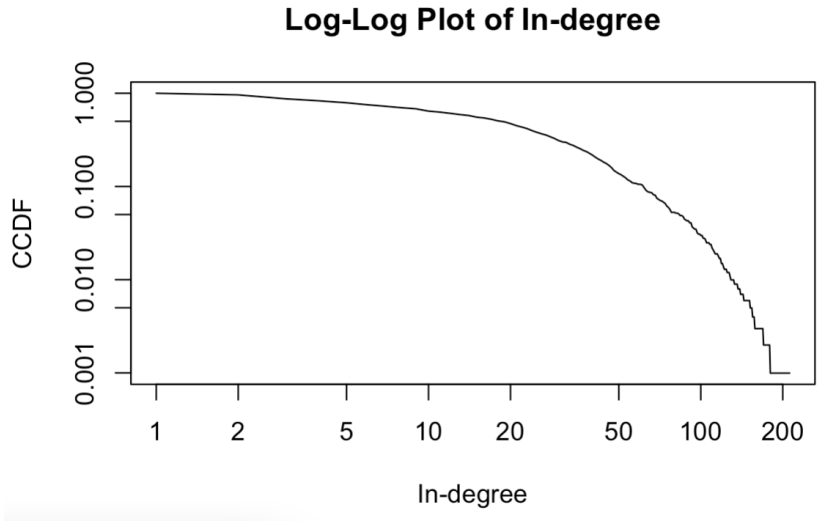

Figure 2: The log-log plot of the in-degree distribution.

Our team decided to plot sub-department graphs to study the components and isolates of the network. Plotting the entire network yields no important information because there are too many nodes crowded together. Showing individual departments allowed us to see different network structures forming.

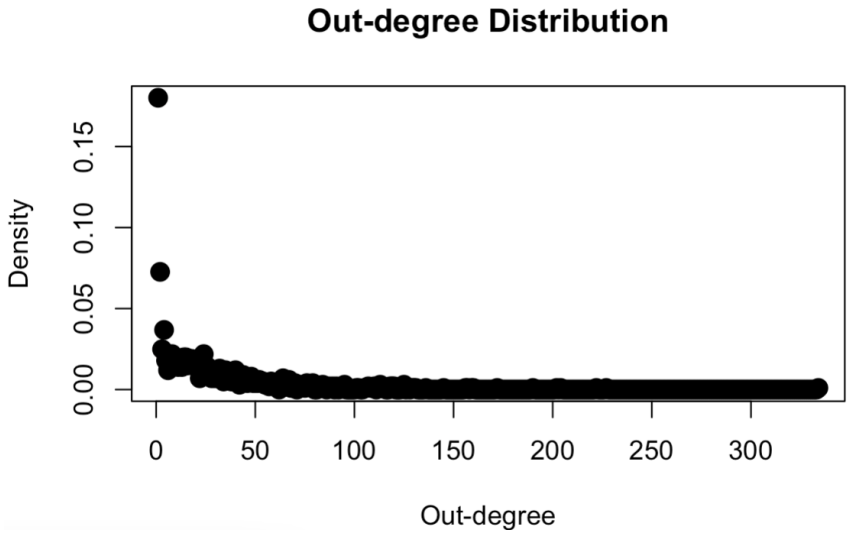

Figure 3: The out-degree distribution for the network.

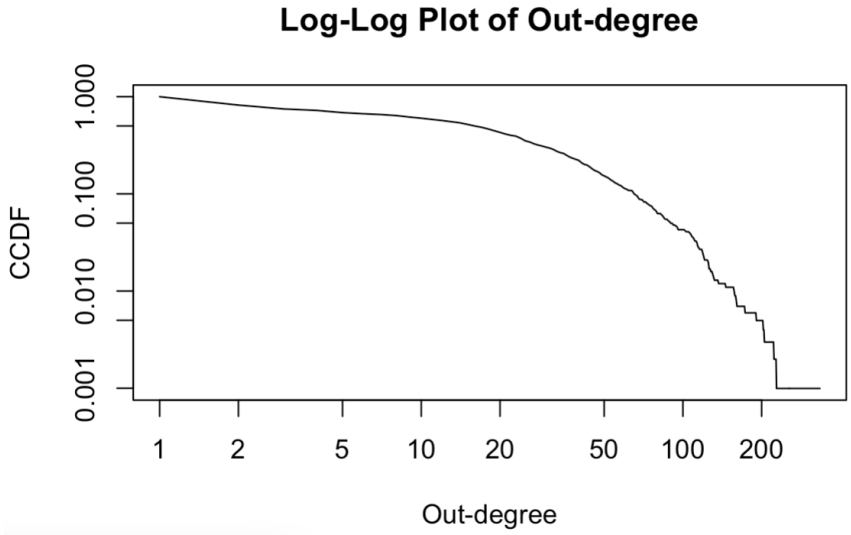

Figure 4: The log-log plot of the out-degree distribution.

We can visualize each of the individual departments by observing Figures 5 through 8. By plotting each of them separately, we can identify some key characteristics that are reinforced by the tables. First, we notice that department 3 appears to be much more connected than the other departments. There exists a large group of nodes that all seem to be closely connected to each other, so much so that it is difficult to identify any clear patterns because they are all so closely clumped together. In contrast, departments 1, 2, and 4 are grouped more in distinct components that are separated from one another. Within the network for department 1, we can see that there are a total of 9 components. Four components in the

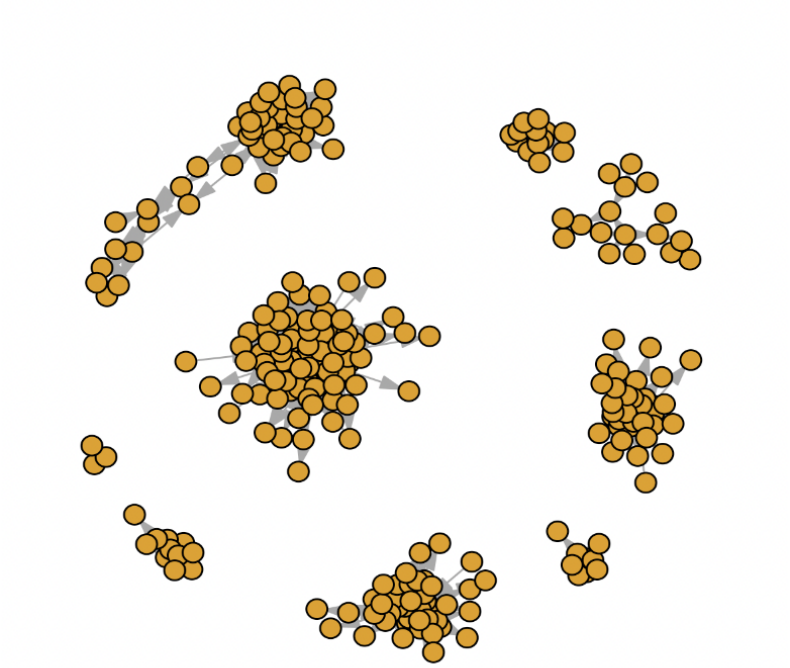

Figure 5: The network graph for members in department 1.

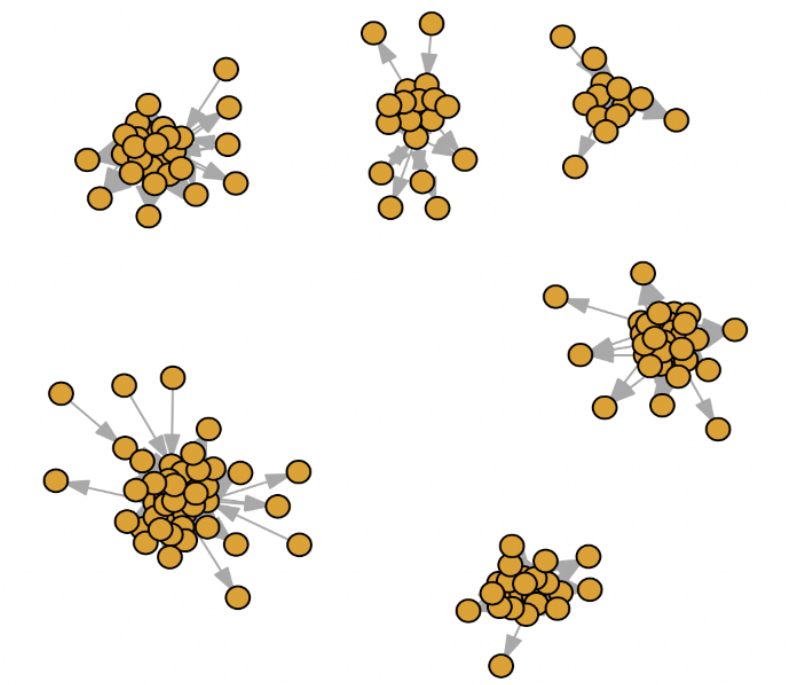

Figure 6: The network graph for members in department 2.

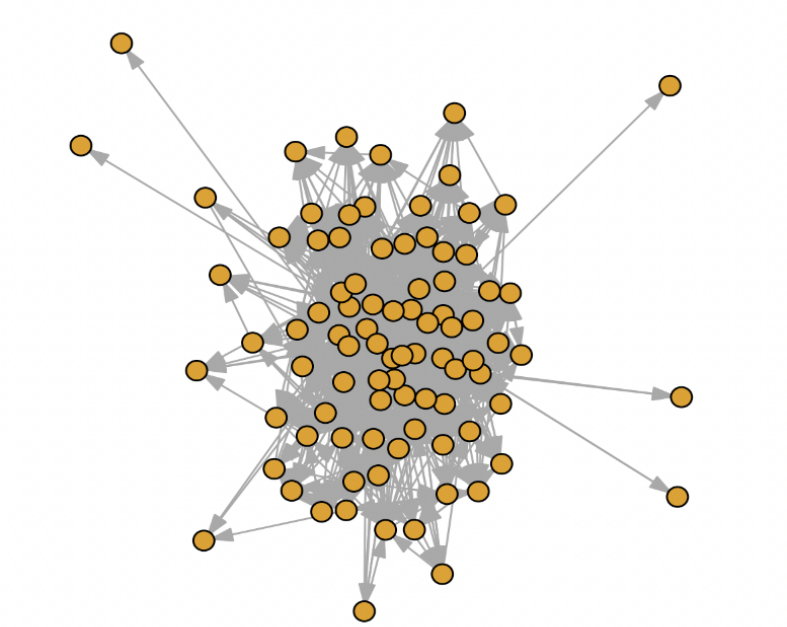

Figure 7: The network graph for members in department 3.

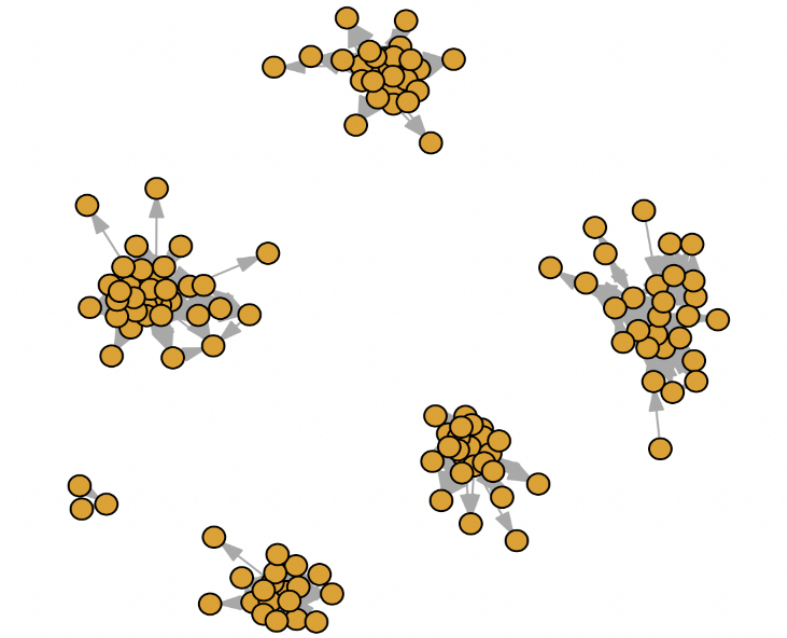

Figure 8: The network graph for members in department 4.

network are relatively larger than the rest. As for department 2, the sizes of the components are more equal in size compared to department 1. The components for department 4 are

also very similar in size, except for one which only contains 3 nodes. Another observation is that, although department 3 as a whole appears to be much more closely connected, there seem to be many nodes that are a bit isolated. As shown from the graph, there are a few actors who are only connected to a singular actor within the network. If you compare this to department 1 (and to an extent the others), there does not appear to be any nodes that are very isolated from the others.

3.3 Relational Event Model (REM)

In the second part of our project, we use REM to identify factors that contribute to the formation or dissolution of ties between individuals within the network. This is appropriate because each directed edge in the network has a timestamp. REM is useful in testing hypotheses and determining which kinds of messaging behavior should be promoted to create new ties to help spread the information about the social events. Here are the following hypothesis that will be tested:

Hypothesis 1: The likelihood of individual i sending a message to individual j is greater if i has sent a message to j recently.

Hypothesis 2: The likelihood of individual i sending a message to individual j is greater if j has sent a message to i recently.

Hypothesis 3: The likelihood of an individual i contacting an individual with a high degree of connectivity is greater compared to contacting those with lower degree of connectivity.

Hypothesis 4: The likelihood of individual i sending a message to individual j is greater if the message immediately preceding that event is j sending a message to i.

Hypothesis 5: When an individual i sends a message to another person j, that person has a greater likelihood of communicating with another individual k, thereby creating an information-sharing chain.

Using the library nusoniclab/rem, each of these hypotheses corresponds to the following parameter statistics. For H1, we use the RRecSnd parameter which represents reciprocity. This describes a receiver j sending a message to sender i after receiving a message from sender i. A positive effect indicates reciprocity being observed at a higher-than-random frequency. A negative effect indicates reciprocity being observed at a frequency lower than could be expected by chance. For H2, we use the RSndSnd parameter which represents repetition. This describes a sender sending multiple messages in a sequence. A positive effect indicates repetition from senders being observed at a higher-than-random frequency. A negative effect

indicates repetition from senders being observed at a frequency lower than could be expected by chance. For H3, we use the NTDegRec that represents a total degree measure of receiver effects. The rate of the receiver sending or receiving messages normalized by the distribution of messages. A positive effect indicates higher total degree of the receiver than expected by chance. A negative effect indicates a lower total degree of the receiver than expected by chance. For H4, we use the parameter PSAB-BA. This represents a pattern of message-sending activity ("participation shift"). The first person A sends a message to person B, and then person B sends a message to person A (reciprocity immediately following the initial message). A positive effect indicates the presence of the pattern at a rate higher than can be expected by chance. A negative effect indicates the presence of the pattern at a rate lower than can be expected by chance. Lastly for H5, we use the parameter PSAB-BY. This represents a pattern of message-sending activity ("participation shift"). The first person A sends a message to person B, and then person B sends a message to any other person (creation of a message chain based on the initial message). A positive effect indicates the presence of the pattern at a rate higher than can be expected by chance. A negative effect indicates the presence of the pattern at a rate lower than can be expected by chance.

Reading from the email-Eu-core-temporal, we see that our data had 332,334 rows. Therefore we decided to sample only rows from 50 randomly selected nodes. We have tried to run the REM models with 50-100 nodes, but the execution time was over an hour and resulted in exceeding the memory limit.

Parameter	**MLE**	**p-value**	**hazard-ratio**
RRecSnd	3.254e-1	<2.2e-16	1.384584
RSndSnd	3.623e-2	<2.2e-16	1.036894

Table 3: REM Model 1 (BIC = 23512)

Parameter	**MLE**	**p-value**	**hazard-ratio**
RRecSnd	3.412e-1	<2.2e-16	1.406635
RSndSnd	2.071e-2	<2.2e-16	1.020926
NTDegRec	2.513e-1	<2.2e-16	1.285696

Table 4: REM Model 2 (BIC = 23498)

From the results, we can see that model 2 has the lowest BIC, suggesting that it has the best fit and therefore we will be analyzing the numbers from Table 4. The coefficients for the variables RRecSnd (reciprocity), RSndSnd (repetition), and NTDegRec are statistically significant, as indicated by their p-values being less than 0.05. All coefficients are also positive. However, for PSAB-BA and PSAB-BY, we got p-values of above 0.05 in model 3, meaning that we cannot reject the null hypothesis, and will not be interpreting the results of these parameters and hypotheses 4 and 5.

Parameter	MLE	p-value	hazard-ratio
RRecSnd	4.233e-1	<2.2e-16	1.526992
RSndSnd	5.648e-2	<2.2e-16	1.058105
NTDegRec	2.301e-1	<2.2e-16	1.258726
PSAB-BA	-1.981e-1	0.09234	-
PSAB-BY	-1.642e-1	0.06541	-

Table 5: REM Model 3 (BIC = 23507)

For RRecSnd (reciprocity), the coefficient is $3.412e^{-1}$. Taking the exponential of this coefficient, we find that the relative likelihood of an individual sending a message to their prior communication partner is approximately 1.406635. This suggests that an individual is slightly more likely to send a message to their prior communication partner compared to a random chance.

For RSndSnd (repetition), the coefficient is $2.071e^{-2}$. Taking the exponential of this coefficient, we find that the relative likelihood of someone responding to another (repeating the messaging sequence) is approximately 1.258726. This indicates that an individual is slightly more likely to respond to another and repeat the messaging sequence compared to a random chance.

For NTDegRec, the coefficient is $2.513e^{-1}$. Taking the exponential of this coefficient, we find that the relative likelihood of an individual contacting another individual with a high degree of connectivity is greater is approximately 1.285696. This indicates that certain individuals are more connected or have a higher level of activity in terms of message sending and receiving compared to what would occur randomly.

4 Implications & Recommendations

4.1 Local Properties

From our findings, we identified a few influential nodes. Notably, node 91 holds a position of influence and prominence within the email network. This suggests that it plays a crucial role as a central hub for information flow, communication, and coordination within the network. **Node 91 should be involved in disseminating important information, coordinating communication efforts, and fostering collaboration across different departments or groups.**

Nodes 358 and 537 have high out-closeness, indicating its influence in disseminating information. It ensures that messages and important updates reach a wide audience quickly and effectively. **Nodes 358 and 537 can be entrusted with ensuring that information reaches a wide audience in a timely manner, thereby increasing engagement and participation among the network members.**

Node 775 has the highest hub score, indicating its connection to other nodes with high authority or importance. It acts as a bridge or intermediary between influential individuals or groups within the network. **Node 775 can act as a facilitator in bridging communication gaps and fostering productive interactions within the network.**

Plotting the sub-department graph shows a concerning problem that there are unconnected components even in each department. This means some people would not receive the communication at all. **Our recommendation is for the institution to encourage a culture of collaboration and active participation within the institution.** This includes implementing strategies that promote open communication, knowledge sharing, and engagement among members of each department. This can include organizing regular team meetings, knowledge-sharing sessions, or collaborative projects to facilitate information flow and involvement.

4.2 Relational Event Model (REM)

From the findings of REM, we see that Hypotheses 1, 2, and 3 are supported. H1 suggests that individuals are slightly more likely to send a message to their prior communication partner compared to a random chance. This suggests that fostering reciprocal communication can be beneficial for building stronger relationships and improving overall communication within the institution. **This can be achieved through initiatives such as collaboration projects, mentorship programs, or regular networking events that facilitate ongoing connections and interactions.**

H2 suggests that individuals are slightly more likely to respond to another person and repeat the messaging sequence. This finding highlights the importance of follow-up communication and maintaining ongoing conversations. The institution should emphasize the value of timely and consistent responses to foster effective communication and collaboration among its members. **The institution should establish guidelines or protocols that encourage members to respond promptly, acknowledge messages, and maintain a consistent flow of communication to prevent information gaps or delays.**

H3 suggests that individuals are more likely to contact others with a high degree of connectivity. This finding suggests that certain individuals within the network have a higher level of activity and connectivity in terms of message sending and receiving. The institution should identify and leverage these well-connected individuals as potential brokers or influencers to facilitate communication and engagement among other members. **The institution can involve them in planning and organizing events to leverage their influence and improve member participation.**

Our team also suggests that the institution provide training or workshops on effective communication strategies, including active listening, clear and concise messaging, and building rapport. These skills can enhance overall communication within

the institution and contribute to a more supportive and productive environment.

Due to p-values being above 0.05, we cannot give any definitive answers on hypotheses 4 and 5. Further research must be conducted in order to determine whether or not to accept or reject these hypotheses.

By implementing these recommendations, the research institution can enhance communication, strengthen relationships among its members, and create a more supportive and productive environment for all.

4.3 Summary

Overall, the findings from local properties analysis align with the supported REM analysis. Reciprocity, repetition, and repetition is a common behavior in this institution's network and we can use this knowledge by utilizing important nodes like node 91, node 358, and node 775 to engage in communication with various individuals or departments. Of course, communication could be improved further by following the recommendations to create more well connected members.

5 Reflection

We were able to answer 3 of the questions we originally wanted to answer. One of our original questions was: **Which department(s) most often cooperate/communicate beyond their department?** We were unable to answer this question due to the node IDs in the sub-network datasets being different from the node IDs in the entire network dataset. Because of this, we couldn't identify which edges connected nodes from different departments.

Using the local properties analysis, we were able to answer question (a): **Who are the actors in the network with the highest centrality in the network?** This helped us reach the conclusion that only a few nodes are well connected and thus are very essential to establishing communications.

By plotting the network graph, we found that there were colleagues that do not interact with other departments to answer question (b): **Are there any actors that never or almost rarely email colleagues within/outside their department?** We also found it interesting that even in the sub-department graph there were many components especially in department 1. Thus, we listed recommendations to promote inter-group activities. Lastly, we were able to use REM to answer question (c): **What are the factors that contribute to the formation or dissolution of ties between individuals within the network?** We found that reciprocity, repetition, and repetition is a common behavior in this institution's network.