**Department of Electrical and Computer Engineering**

**North South University**

---


# Senior Design Project

## Meaning in Structures:

## What lower manifold graph network embeddings tell us

**Name: Mahmud Elahi Akhter** **ID # 1721498042**

**Name: Zahin Ahmed** **ID # 1711020042**

**Faculty Advisor:**

**Dr. Mohammad Ashrafuzzaman Khan**

**Assistant Professor**

**Department of ECE**

## Fall 2020

# LETTER OF TRANSMITTAL

January, 2021

To

Dr. Mohammad Rezaul Bari

Associate Professor and Chairman,

Department of Electrical and Computer Engineering,

North South University, Dhaka.

Subject: Submission of Capstone Project on “Meaning in Structures: What lower manifold graph network embeddings tell us”.

Dear Sir,

With due respect, we would like to submit our Capstone Project Report on “Meaning in Structures: What lower manifold graph network embeddings tell us” as a part of our BSc program. In this project we analyzed internet networks and tried to discover meaningful patterns in them. We learned quite a bit about non-linear dimensionality reduction, high dimensional clustering and came to appreciate NP-complete problems and efficient algorithmic solutions to them. We also learned about manifold learning and graph embeddings. Our work can be extended to generate better network graph embeddings for the purpose of knowledge discovery.

We will be highly obliged if you are kind enough to receive this report and provide your valuable judgment. It would be our immense pleasure if you find this report useful and informative to have an apparent perspective on the issue.

Sincerely Yours,

.....

Mahmud Elahi Akhter

Department of ECE

North South University, Bangladesh

.....

Zahin Ahmed

Department of ECE

North South University, Bangladesh

# APPROVAL

The capstone project entitled “Meaning in Structures: What lower manifold graph network embeddings tell us” by Mahmud Elahi Akhter (ID 1721498042) and, Zahin Ahmed (ID 1711020042) is approved in partial fulfillment of the requirement of the Degree of Bachelor of Science in Computer Science and Engineering on January,2021 and has been accepted as satisfactory.

**Supervisor:**

_____

**Dr. Mohammad Ashrafuzzaman Khan**

**Assistant Professor**

Department of Electrical and Computer Engineering

North South University

Dhaka, Bangladesh.

**Department Chair:**

_____

**Dr. Mohammad Rezaul Bari**

**Associate Professor and Chairman**

Department of Electrical and Computer Engineering

North South University

Dhaka, Bangladesh.

# DECLARATION

This is our truthful declaration that the “Capstone Project Report” we have prepared is not a copy of any “Capstone Project Report” previously made by any other team. We also express our honest confirmation in support of the fact that the said “Capstone Project Report” has neither been used before to fulfill any other course related purpose nor it will be submitted to any other team or authority in future.

.....

Mahmud Elahi Akhter

Department of ECE

North South University, Bangladesh

.....

Zahin Ahmed

Department of ECE

North South University, Bangladesh

# ACKNOWLEDGEMENT

Firstly, we would like to thank our supervisor Dr. Mohammad Ashrafuzzaman Khan for providing us with this research idea and for supporting us with necessary computational resources.

We would like to thank the wonderful people behind CRAN repository as conducting this research without their resources and vignettes would have been impossible.

We would also like to thank the ECE department of North South University for providing us with the opportunity to conduct this research.

We would also like to thank Ebna Mannan, Imtiaz and Mir Imtiaz Mostafiz for their support and guidance during methodology development and result analysis.

Finally, we would like to thank our families and everybody else who supported us for the completion of this project.

# ABSTRACT

As the internet becomes more accessible all around the world, the different networks around the internet such as p2p, email etc. are also growing in size. Therefore, analyzing large networks for the purpose of knowledge discovery, in order to learn different governing patterns that dictate these large networks has become quite necessary. With that goal in mind, in this study we studied nine different sized networks. These varied from small to large networks such as more active human driven networks like email communications, social networks to more passive networks such as p2p. We employed unsupervised learning techniques to visualize the underlying structures of these networks by representing them as 2 dimensional manifolds. We also used high dimensional clustering to find pattern in these representations. We have also tried to explain these pattern through an ablation study. With these results in mind we would like to hypothesize that network structures are more affected by the behavior of the nodes/users of the networks instead of having a predefined shape. We found that same type networks (p2p network) have similar structures within certain error margin, while networks of different types (social network vs p2p network) will be different in terms of network structure. Based on our findings, we also propose a hierarchical categorization of networks in a broader sense, such as communication networks, have hierarchies within their structures, where we can observe the structures changing in a certain pattern or trend. We have named this as a galaxy model for communication for its self-repeating pattern in large networks. As per galaxy model assumptions, this kind of behavior based unsupervised knowledge discovery methods can help us find further meaningful patterns in large random human networks which than can be used to identify and generalize different networks such as migration networks, criminal networks or corruption networks.

Keywords: Unsupervised Learning, Data Visualization, Knowledge Discovery

# Table of Contents

# List of Tables



# List of Figures

# CHAPTER 1: INTRODUCTION

As the number of internet users grows, so does the amount of data that is being generated. These data are propagated through many different networks. Networks are the central structure for the internet. Many types of networks range from more formal email networks to more informal instant messaging networks to more anonymous file-sharing networks. However, if we were to look beyond these categories and types, we would see that the internet exists for connectivity and communication. Communication is the central idea behind most of these networks. This naturally raises the question of whether networks shape different communication types or do communications shape the networks. We can also move a step forward and ask how these various networks' structures vary depending on the different communications they carry out. Finding communication patterns or hierarchies from abstract network structures can give us a lot of insight into how different styles/types of communication shape the networks. This can also help see how one communication type grows from another. To better understand these mechanisms, we studied different networks and tried to answer how the human behavior behind communication and the volume of the data can shape the network structure.

To do so, we used network centralities to work with the networks. In order to understand network, we need to understand it's nodes and edges better, and centralities have been used over the years to answer the characteristics of different nodes. Centralities such as Degree, Closeness, Betweenness, Crossclique, Pagerank, etc., have all been proposed and developed [1] over the years to answer "Which node is the most influential?" in various different applications [2]. "Significance" or "importance" of nodes varies from one context to another. For example, in some cases, it is essential to identify which node(s) propagate more information locally or globally in a graph [3], whereas in other cases, detecting the central node(s) might be of more value [4]. Centralities can, therefore, be seen as features/characteristics of a graph. We can also discriminate between networks to some extent centrality values [5].

In this study, the centrality measures helped us visualize and analyze the structure of real-world networks. As stated earlier, the central idea behind the internet is communication. Therefore, we hypothesize that the structure of a network is not random but is instead dictated and or largely affected by the behavior of the nodes or, in this case, users of the network. We found that, as different centrality measures express different characteristics of the nodes of a network, the information given by the combination of centrality measures offer similar visual representations

for similar networks (such as email networks of two institutions). Apart from that, we also saw that different networks have different correlation between their centralities. From this we came to the conclusion that there is a hierarchical nature to communication and this nature is dictated by communication type. We also found that we can compare this hierarchy to a galaxy if we have enough volume of data where different solar systems can represent different network structures. The following sections present our study. Chapters 2 and 3 discuss the use of centralities. Chapter 4 outlines the datasets used in the study, and chapter 5 denotes the ablation studies that was carried out. Chapter 6 goes into the results and discussion of our findings.

# CHAPTER 2: CENTRALITIES

Our study largely revolves around using centralities for their usefulness in network characteristics representation. They can be used to explain the node's importance both globally or in their local communities/cliques. Centrality measures essentially describe a node's connectivity within a network, based on factors such as number of connections, geodesic distances with other nodes, placement in between different cliques etc. [6,7,8]. The centralities used in this study were chosen based on whether they expressed the global or local presence of the node, resource intensiveness, and computability. The centralities that were selected are: Degree, Eigenvector, Pagerank, Authority, Hubscore, Betweenness, Closeness-latora, Eccentricity, Density of Maximum Neighborhood Component (DMNC), Lobby index, Leverage, Local Bridging.

**Degree:** Total number of edges going into and out of a node. Denoted by:

$$C_D(j) = \sum_{j=1}^{n} A_{ij}$$

where $A_{ij}$ is a dense adjacency matrix of the graph.

**Eigenvector:** Ranking of nodes based on how many well-connected nodes they are connected to. Denoted by:

$$x_v = \frac{1}{\lambda} \sum_{t \in M(v)} x_t = \frac{1}{\lambda} \sum_{t \in G} a_{v,t} x_t$$

where $M(v)$ is set of neighbors of $v$ and $\lambda$ is a constant.

**Pagerank**: Ranking of nodes based on the frequency of their appearance in a random traversal of the network. Denoted by:

$$C_{PR}(v) = (1-d) + d\left(\frac{C_{PR}(t_1)}{C(t_1)} + \cdots + \frac{C_{PR}(t_n)}{C(t_n)}\right)$$

where $t_i, i = 1, \ldots, n$ are the nodes that have a directed edge to node $d$, $C(v)$ is the number of edges going out from $v$, and $d$ is the damping factor ($d \in [0,1]$).

**Authority**: Ranking of nodes based on the principal eigenvector of

$$v = A^t \cdot A$$

where A is the adjacency matrix of the graph.

**Hubscore**: Ranking of nodes based on the principal eigenvector of

$$u = A \cdot A^t$$

where A is the adjacency matrix of the graph.

**Betweenness**: Ratio to define how many shortest paths pass through a node amongst total shortest paths between all the node pairs of the network. Denoted by:

$$C_B(v) = \sum_{s \neq t \neq v \in V} \frac{\sigma_{st}(v)}{\sigma_{st}}$$

$\sigma_{st}$ is the number of shortest paths between node s and t and $\sigma_{st}$ (v) is the number of shortest paths passing on a node v out $\sigma_{st}$

**Closeness-latora [9]**: Closeness is not well defined for networks with disconnected components. Therefore, closeness-latora was used which is denoted by:

$$E(\boldsymbol{G}) = \frac{\Sigma_{i \neq j \in \boldsymbol{G}} \in_{ij}}{N(N-1)} = \frac{1}{N(N-1)} \sum_{i \neq j \in \boldsymbol{G}} \frac{1}{d_{ij}}$$

**Eccentricity [10]**: Maximum distance between a node v and all other nodes. Denoted by:

$$C_E(v) = \frac{1}{\max\{dist(u,v) : u \in V\}}$$

**DMNC [11]**: Density of the Maximum Neighborhood Component of a node. Denoted by:

$$\frac{|E(MNC(v))|}{|V(MNC(v))|^{\varepsilon}}$$

Where $1 \leq \varepsilon \leq 2$

**Lobby index [12]**: Defines the largest integer $k$ such that the node has at least $k$ neighbors which have degree of at least $k$.

**Leverage [13]**: Relationship between the degree of a node and the degree of all of its neighbors, averaged over the total number of neighbors. Denoted by:

$$l_i = \frac{1}{k_i} \sum_{N_i} \frac{k_i - k_j}{k_i + k_j}$$

**Local Bridging [14]**: A variant of global bridging centrality which can be calculated locally requiring only limited local neighborhood graph information. Denoted as

$$LBC = C_{ibet} \cdot \beta_C \; and \; \beta_C = \frac{\frac{1}{d(v)}}{\Sigma_{i \in N(v)} \frac{1}{d(i)}}$$

# CHAPTER 3: LITERATURE REVIEW

In recent years, network analysis techniques have evolved [15] in proportion to the rapid growth of real world networks. Much research has been done on networks such as social networks [16,17], communication networks [18], citation networks etc. As time progresses, these networks become bigger and more complex, consequently holding vast amounts of interesting information which can be used for various purposes. Network analysis techniques include, but are not limited to, using random graph models to capture or derive the properties of real world networks [19] subgraph isomorphism [20]**,** graph simulation etc. A significant portion of the research conducted on networks has been on centralities [3] to deduce, "which" are the more important nodes in the graph i.e. central nodes or nodes that propagate most information, and how traffic flows through them. Over the years, network centralities have been used for many different purposes that range from traffic network to brain networks i.e., analyzing social networks, such as tweet classification [21], detecting political discussion practices [22] and many other such applications [23-25]; has involved the generation of multiple centralities. Centralities have also been used in analyzing road traffic networks, such as finding road network patterns [26], tourism management [27] etc. Another important application of centrality measurements is in analyzing biological networks, as can be seen in [13,28-30].

As stated earlier, in their work, Wang and Krim showed that discrimination of graph networks is possible through centrality measures [5]. They showed that degree centrality and clustering coefficients were enough to discriminate networks and adding Betweenness Centralities, Eigenvector Centralities or Closeness Centralities degraded the results. However, they only used two small sample datasets for their study. In his study Dwyer showed that visual analysis to explore and compare the centralities within a given network was possible [8]. In the study the centralities were drawn on a 2D plane that was mapped to 3D plane. Three different methods were employed to this purpose and these were 3D parallel Coordinates based Comparison, Orbit based comparison and hierarchy based comparison.

Sarrac´en and Rosso used Betweenness centrality to automatically summarize text [31]. In order to do so, they represented the text as an indirected weighted graph where each sentence was represented as a bag of words. They also used similarity/dissimilarity criterion to represent the semantic relation between nodes. Afterwards, a ranking algorithm was used which was based on Betweenness centrality. Wu proposed a novel graph clustering algorithm which used Betweenness

centrality recursively to create groups of clusters called LEADER which guided the algorithm to cluster the whole network [32].

Huang et al proposed a visual analytics method to explore urban traffic mobility patterns [33]. They used Pagerank and Betweenness centralities to calculate the more central/ important streets. Pagerank detected hub streets and Betweenness detected street/region that acted as back-bone in urban networks. Crucitti used Closeness, Betweenness, Straightness and information centralities to capture street patterns of different cities of the world [1]. They proposed that a hierarchical clustering based on distribution of centrality measures are capable of distinguishing different cities to some extent used degree, betweennesss and closeness centrality to calculate different road patterns [26]. The aim of their study was to discriminate different road patterns using centralities. The centralities were calculated using a topological network representation of the road networks.

# CHAPTER 4: DATASETS

All the datasets that were used in this study were taken from SNAP [34]. The datasets are categorized as Social Networks, Citation Networks, Collaboration Networks, Communication networks and Internet P2P networks. The description of each network is given below.

## 4.1 Social Network

Only one social network was used for this study. It was an ego network. The Ego networks are essentially networks centered around one particular node. Such networks are formed by taking one node and finding all the vertices that are directly connected to it while also finding the connections between those vertices. The ego network used in this study is the "Ego-Facebook" dataset, which contains 4039 nodes and 88233 edges. It combines ten different ego networks. The Facebook ego network dataset was published in 2012 [35], and it was collected through a survey using the Facebook app.

## 4.2 Communication Networks

The communication networks used in this study were "Email-EuAll" and "Email-enron" networks. "Email-EuAll" has 265214 nodes and 420045 edges and was collected from a European institute from October 2003 to May 2005. The nodes represent email addresses, and a directed edge signifies that the source node sent at least one email to the target node. "Email-enron" has 36692 nodes and 367662 edges and was collected and made public from the Enron Corporation when it was being investigated. This network is undirected and contains an edge between nodes if any email was exchanged between them.

## 4.3 Collaboration Networks

The Arxiv collaboration networks were collected from the e-print arXiv website. The edges represent collaborations between authors, so if one author collaborated with another, an undirected edge exists between them. The data consists of papers published between January 1993 to April 2003. The "Ca-HepTH" dataset has 9877 nodes and 25998 edges, and represents collaborations in papers of the High Energy Physics. - Theory network. The "Ca-HepPH" dataset has 12008 nodes and 118521 edges and represents collaborations in papers of the High Energy Physics - Phenomenology network.

## 4.4 Citation Networks

The Arxiv citation networks were also collected from the e-print arXiv database. These datasets represent which papers cite each other within this database, and if one paper cites another, a directed edge is drawn from the former to the latter. Information regarding cited papers that do not exist within the database is not present in the networks. The data was collected from papers published between January 1993 and April 2003. The "cit-HepPH" dataset contains 34546 nodes and 421578 edges and consists of papers from the High Energy Physics- Phenomenology network. The "cit-HepTH" dataset contains 27770 nodes and 352807 edges and consists of papers from the High Energy Physics- Theory network.

## 4.5 Internet P2P Networks

The peer-to-peer network used in this study was constructed from the Gnutella file-sharing network by taking nine snapshots across a few days in August 2002. The ones used in this study are: "P2p-Gnutella04" (10876 nodes and 39994 edges), which was collected on 4th August, "P2p-Gnutella08" (6301 nodes and 20777 edges) was collected on 8th August, "P2p-Gnutella24" (26518 nodes and 65369 edges) was collected on 24th August, "P2p-Gnutella25" (22687 nodes and 54705 edges) was collected on 25th August, "p2p-Gnutella30" (36682 nodes and 88328 edges) was collected on 30th August and finally "p2p-Gnutella31" (62586 nodes and 147892 edges) was taken on 31st August.

## 4.6 Dataset Statistics

Below we present the different statistics of each dataset of each category that were used in this study.

| Dataset name | No. Nodes | No. Edges | Average Degree | Graph Density | Graph Transitivity |
|---|---|---|---|---|---|
| **Ego-Facebook** | 4039 | 88233 | 43.69051745 | 0.005408581 | 0.519189302 |

*Table 4-1: Social Networks*

| Dataset name | No. Nodes | No. Edges | Average Degree | Graph Density | Graph Transitivity |
|---|---|---|---|---|---|
| **Email-Enron** | 36,692 | 183,831 | 20.04044478 | 0.00027309 | 0.085310796 |
| **Email-EUAll** | 265,214 | 420,045 | 3.167592963 | 5.97E-06 | 0.004106431 |

*Table 4-2: Communication Networks*

| Dataset name | No Nodes | No Edges | Average Degree | Graph Density | Graph Transitivity |
|---|---|---|---|---|---|
| **CA-HepPh** | 12,008 | 118,521 | 39.47534977 | 0.00164371 | 0.659477009 |
| **CA-HepTH** | 9,877 | 25,998 | 42.20754315 | 0.001124215 | 0.318001581 |

*Table 4-3: Collaboration Networks*

| Dataset name | No Nodes | No Edges | Average Degree | Graph Density | Graph Transitivity |
|---|---|---|---|---|---|
| **Cit-HepTh** | 27,770 | 352,807 | 25.40921858 | 0.000457494 | 0.119569073 |
| **Cit-HepPh** | 34,546 | 421,577 | 24.4067041 | 0.000353249 | 0.145656674 |

*Table 4-4: Citation Networks*

| Dataset name | No Nodes | No Edges | Average Degree | Graph Density | Graph Transitivity |
|---|---|---|---|---|---|
| **p2p-Gnutella04** | 10,876 | 39,994 | 7.354542111 | 0.000338109 | 0.005402029 |
| **p2p-Gnutella08** | 6,301 | 20,777 | 6.595556 | 0.00052354 | 0.02066042 |
| **p2p-Gnutella24** | 26,518 | 65,369 | 4.930271 | 9.30E-05 | 0.004101532 |
| **p2p-Gnutella25** | 22,687 | 54,705 | 4.822497 | 0.000106288 | 0.004535649 |
| **p2p-Gnutella30** | 36,682 | 88,328 | 4.815876997 | 6.56E-05 | 0.005163701 |
| **p2p-Gnutella31** | 62,586 | 147,892 | 4.726040968 | 3.78E-05 | 0.003872019 |

*Table 4-5: P2P Networks*

# CHAPTER 5: ABLATION STUDY

In order to understand how the centralities affected the representation of our model, an ablation study was carried out. In order to do so, as per the centrality definitions, we categorized them into three categories. These categories are *Node based centrality*, *Community/Neighborhood based centrality,* and *Node and Community based centrality*. Degree, Eigenvector, PageRank, Authority, Hubscore, Betweenness and Closeness, these centralities were categorized as Node based centrality. The Community based centralities are Density of Maximum Neighborhood Component (DMNC), Lobby index and Local Bridge Centrality. The last category of both Node and Community centrality contains Leverage Centrality and Eccentricity Centrality. Afterward, for ablation study nine ablation models were designed based on assumptions that again stemmed from the centrality definitions. These model assumptions are described below. The categorization of centralities is also given below.

| Node Centrality | Community Centrality | Both Node and Community |
|---|---|---|
| Degree | DMNC | Eccentricity |
| Eigenvector | Lobby index | Leverage |
| Pagerank | Local bridging | |
| Authority | | |
| Hubscore | | |
| Betweenness | | |
| Closeness | | |

*Table 5-1: Centrality Categories*

Model 1 is our baseline model. In model 1, all the 12 centralities were kept. The goal of the ablation study was to see the effect of these centralities on the baseline model.

Model 2 is the first ablation model. The primary assumption of model 2 was that degree is one of the most important centrality. Therefore, we wanted to see how much it affected the whole representations on its own. Thus, in this model, we removed degree centrality but kept the rest of the 11 centralities.

In model 3, we removed the community-based centralities. Our assumption was that these centralities work on groups or communities; therefore, these centralities should influence the creation of meaningful local structures in the representation. In order to verify that, we removed both the categories of *community based centrality,* and *node and community based centrality*.

In model 4, all the centralities that belonged to the category *Node based centrality* were removed. The assumption was that these centralities hold a global structure, not local ones.

Therefore, removing these should give us disjointed meaningful local clusters that shouldn't be able to propagate information globally very well due to their disjointedness.

In model 5, we removed all the ranking algorithms such as eigenvector, pagerank, hubscore, and authority centrality. These centralities calculate the influence of a node in the network. Therefore, this model studies the representation that results in by removing these centralities.

In model 6, betweenness, eccentricity, and closeness centrality were removed. These centralities tell us about the information propagation throughout the nodes in the network. Therefore, this ablation model studies the impact of the removal of these measures have on the baseline model.

Model 7 is a mix between the categories *Node based centrality* and *Node and Community based centrality*. From the category *Node based centrality* we have degree, pagerank, authority, hubscorem betweenness and eccentricity. The assumption was that we would see a representation that would show a subset of the baseline representation but give us better interpretability of which centralities were relatively more important for that representation.

Model 8 also similar to model 7. Here, we combined all three categories. The assumption was that this would also give a subset of the representation and show more of the impact of these centralities and the ones that were removed.

In model 9 we again combined all three categories. In model 9, *Community/Neighborhood based centrality* and *Node and Community based centrality* categories have more presence compared to *Node based centrality*. The assumption is almost similar to model 4; however, the goal was to more interpretability with the mix of a few *Node based centrality*.

Below a matrix of the ablation study models and their selected centralities are given. The different categories have been annotated using different colors.

| | Degree | Eigenvector | Pagerank | Authority | Hubscore | Betweenness | Closeness | Eccentricity | DMNC | Lobby index | Leverage | Local Bridging |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Model 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| Model 2 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| Model 3 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 |
| Model 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 |
| Model 5 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| Model 6 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 1 | 1 | 1 | 1 |
| Model 7 | 1 | 0 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 |
| Model 8 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 0 | 1 |
| Model 9 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 1 |

*Table 5-2: Ablation model matrix*

# CHAPTER 6: RESULTS ANALYSIS & DISCUSSION

## 6.1 Results Analysis

As stated earlier, the chosen centralities were used to visualize the networks. For the visualization purpose we used t-stochastic neighbor embedding (tsne). Due to the close proximity of centrality values, tsne failed capture any patterns in the datasets. For the purpose of pattern recognition, we used a clustering algorithm on the original non-linear high dimensional dataset and fitted those clusters on the lower manifold tsne representation. As the datasets were non-linear and high dimensional, we used both spectral clustering and spherical kmeans. However, we found that spherical kmeans was much faster than spectral clustering. Therefore, spherical kmeans was used throughout the whole study. Even though the datasets were non-linear, we used principal component analysis (pca) to check for variable contribution. The pca was used mainly to interpret the ablation study models.

In this section we will present our results. As stated earlier that the datasets were categorized in social network, communication network, p2p network, citation network, and collaboration network. From this point onward we will denote these categories as class. Therefore, the results were divided between within class analysis i.e. analysis between “eu-All” and “enron” and between class analysis i.e. p2p networks and communication networks. The within class analysis uses boxplots and correlation matrix to find the similarity between networks. On the other hand, we visually analyze the lower dimensional representation of the networks for the between class analysis.

### 6.1.1 Within dataset

For the purpose of within dataset analysis, we have used correlation matrix and boxplot. Due to the varying nature data volume of different networks, we emphasized more on correlation matrix in order to better understand the underlying characteristics. The correlation matrix better expresses the numerical relation among the centralities compared to the boxplot as the numerical distributions can vary within datasets for many different reasons. However, boxplots were kept as extra validation for our analysis.

From figure 6-1 we can see the boxplot and correlation matrix comparison between Eu-All and Enron. We can see from Eu-all and Enron that most node based centralities have positive correlation with each to varying degree. However, enron has more correlation among its

centralities. Enron is an example of a dishonest network. There are anomalies in it that will be discussed in the visualization section. However, we can also the correlation of leverage with dmnc and, correlation of local bridge with lobby and leverage is somewhat similar for both networks. The increase of correlation between node based centralities for Enron can be attributed to its nodes and their internal global cliques. As we know, enron high officials were part of a massive accounting scandal, which can be reflective of these global network cliques. Also notable is the negative correlation of closeness and leverage of Eu-All. This can be explained as, more central nodes that propagate more information tend to be less influential in an email network. The negative correlation between eccentricity and leverage of the Eu-All can be

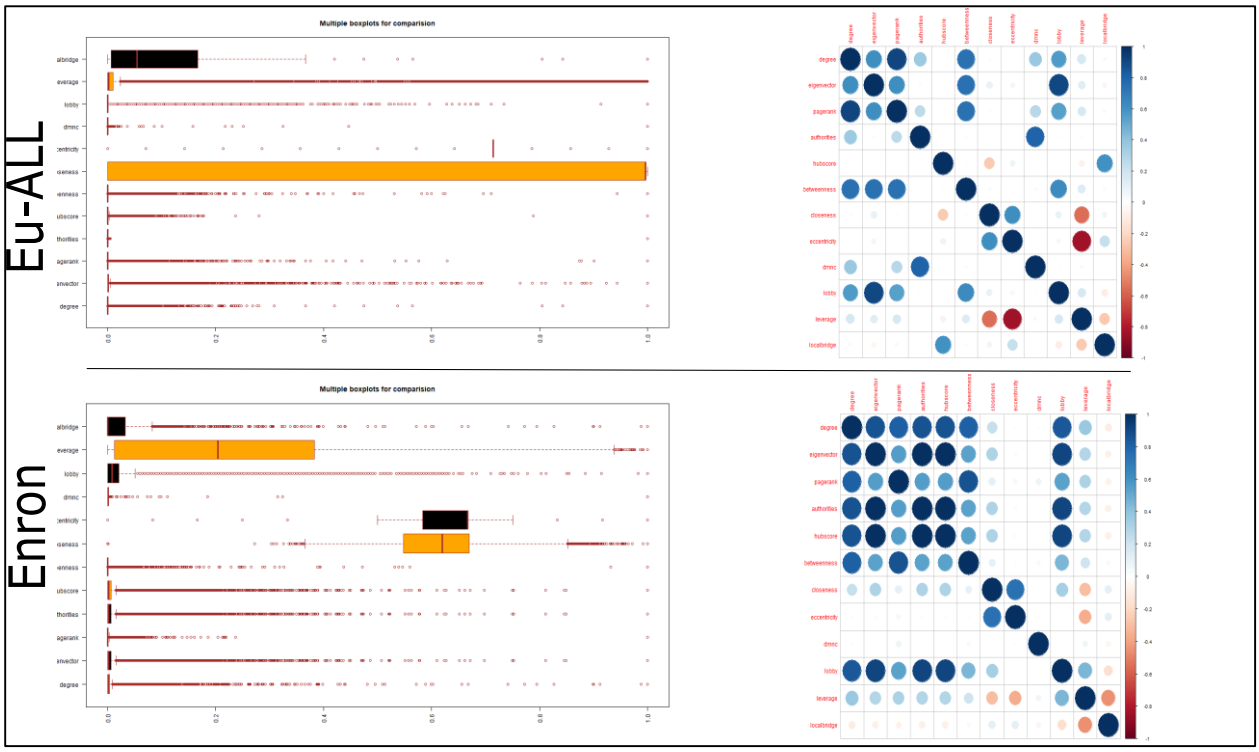


*Figure 6-1 Communication Networks*

attributed to having no influential nodes in a neighborhood instead having influential neighborhoods. For Enron, local bridging centrality is mostly slightly negatively correlated with most node based centrality except closeness.

The boxplots do have some form of similarity and the boxplot for Enron can be thought of as a subset of the Eu-All boxplot for most centralities.

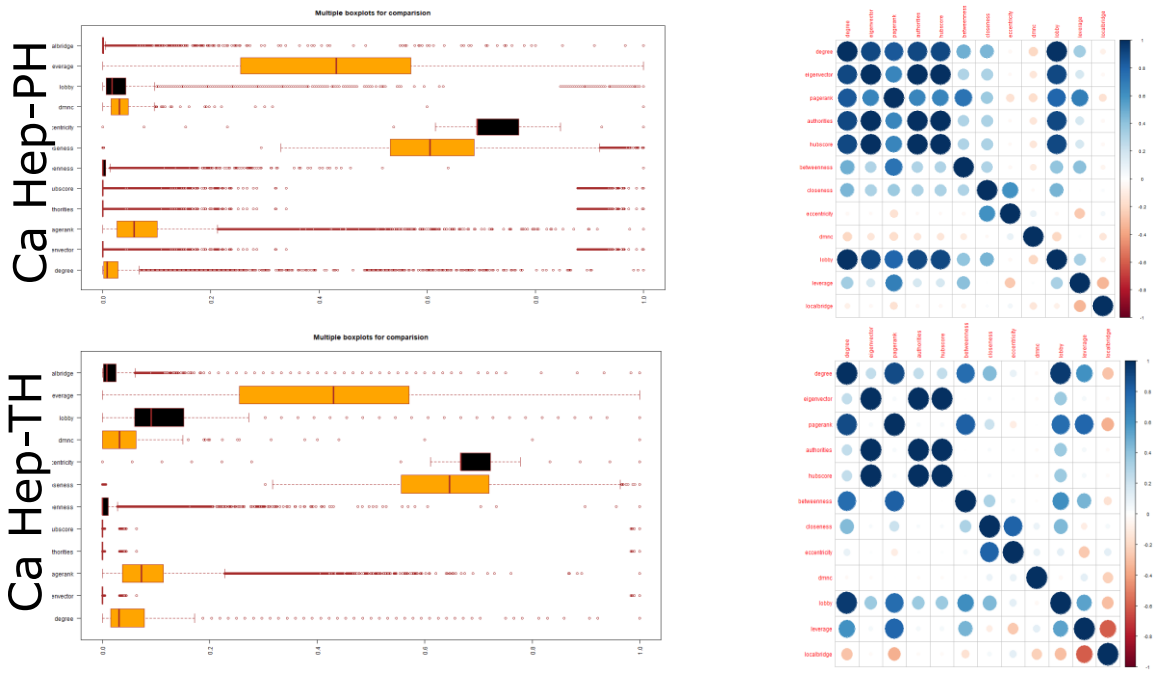


*Figure 6-2 Collaboration Networks*

From figure 6-2, we can see that the boxplot between Ca Hep-Ph and Ca Hep-Th are similar. Both are collaboration networks. So, this kind of distribution are not surprising. However, the dissimilarity between the correlation matrices are quite interesting. Ca Hep-Th has less correlation among its centralities. This can be due to theoretical physics having less collaboration compared to phenomenology. However, this requires additional statistics to verify.

From figure 6-3, we can see that both Cit Hep-Ph and Cit hep-Th networks have similar boxplots and correlation matrices. Citation is not necessarily a direct form of communication. It is a very indirect form, therefore there is not much change when it comes to citation networks.

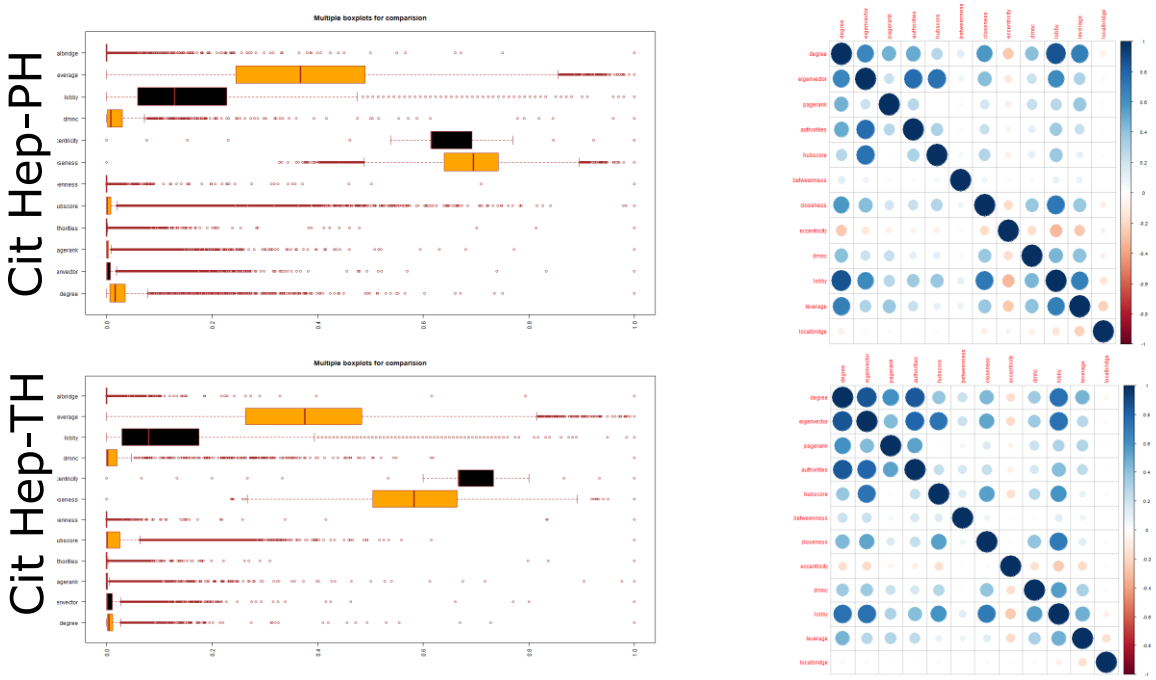


*Figure 6-3 Citation Networks*

Figure 6-4 contains the facebook ego network which represents social networks in our study. In most other communication network, eccentricity and closeness did not have any negative correlation. However, we can see that Facebook ego network has negative correlation between eccentricity and closeness.

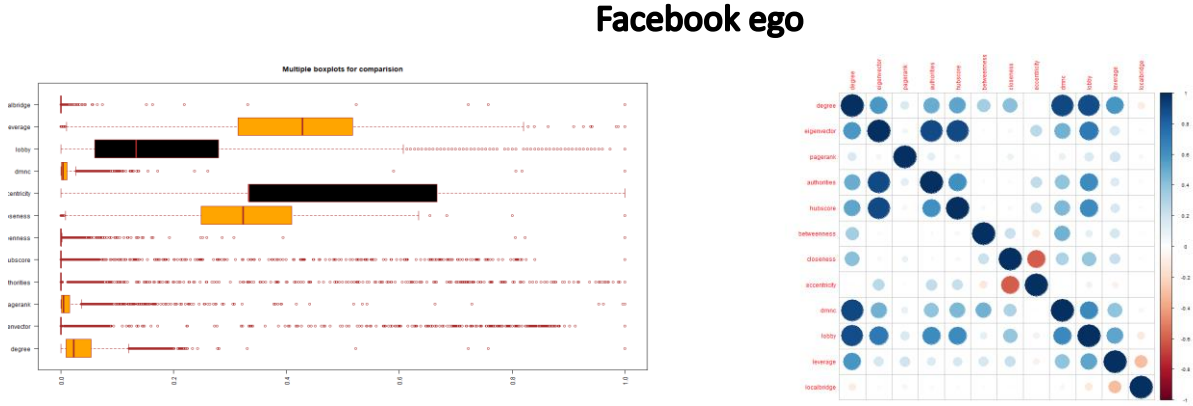


*Figure 6-4 Social network*

From figure 6-4, we get the correlation matrices of P2P networks. P2P networks are very much dynamic networks and vary community to community. Therefore, it is rather chaotic. For this reason, the boxplot for P2P was omitted here. From the correlation matrices we can see that all of them have the same characteristics.

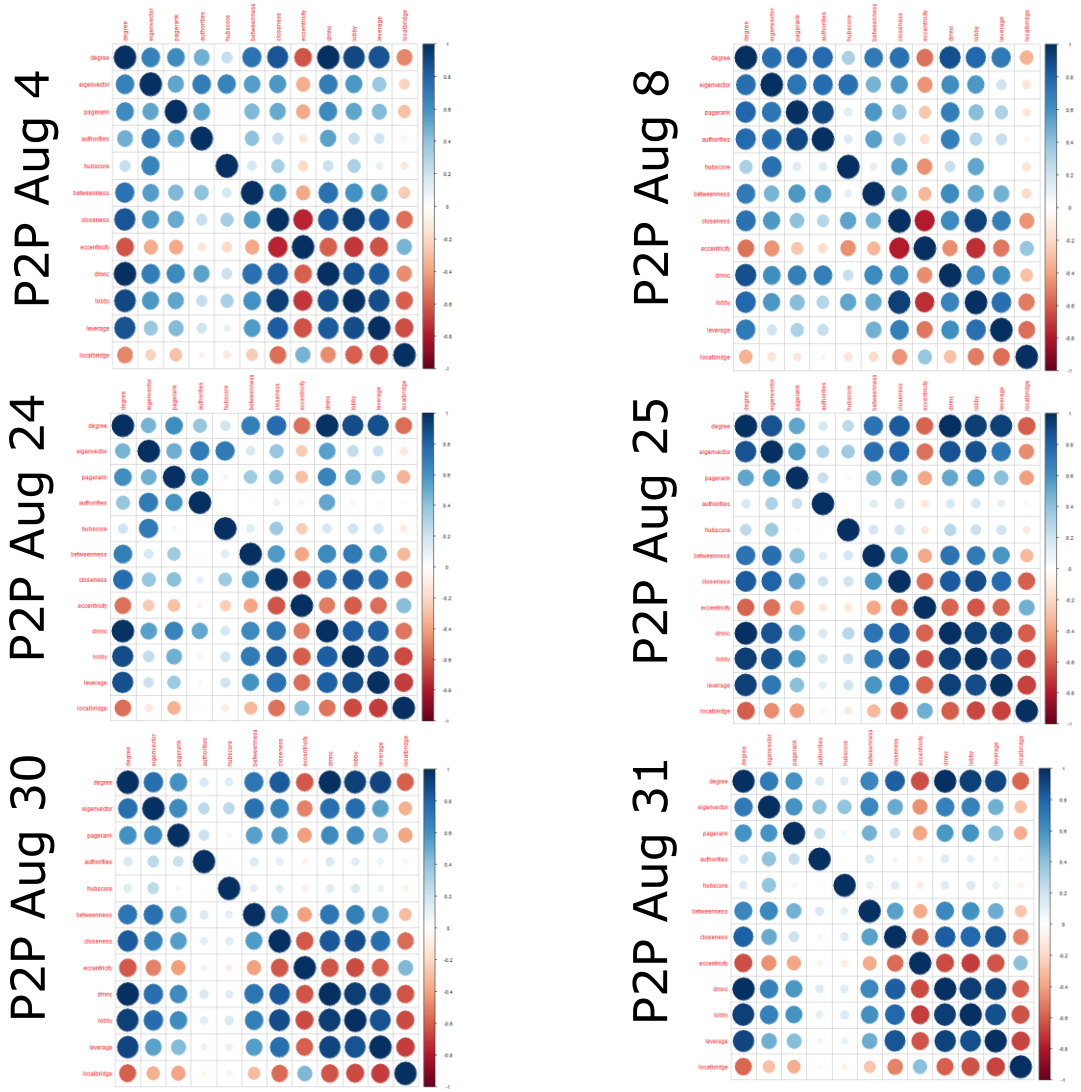


*Figure 6-5 P2P Networks*

In this section, we mostly analyzed the results of networks that share the same class. Here we saw that same class networks do share similar characteristics.

### 6.1.2 Between Datasets

In this section, we will be analyzing the network characteristics between network classes. From figure 6-6 we can see three different class of networks. What's interesting is the similarity between Ca Hep-Ph and enron. Collaboration requires much more frequent communication compared to a corporate email network like enron. However, we can see that the collaborative behavior of collaboration network is present in the enron correlation matrix to quite some extent. This can be due enron's high officials tie to corruption. If we are to compare the research institute network of Eu-all to collaboration, we can see what the difference should be between research institute emails and collaborative communications.

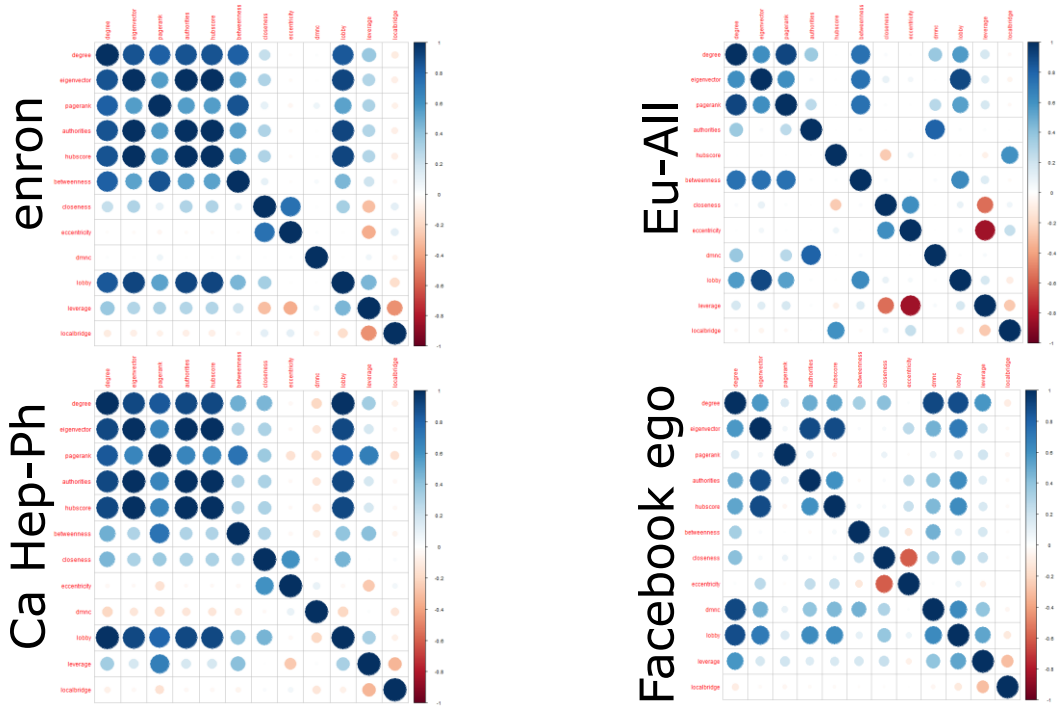


*Figure 6-6 Comparison between datasets*

From the comparison between facebook ego and Eu-all, we can see that they also share some underlying characteristics. This can be attributed the social messaging aspect of facebook that resembles email networks.

### 6.1.3 Visual analysis

In this section, we analyze the lower manifold representation of the networks and patterns that were found in them. First we will go over the representations and patterns and later we will try to interpret these with the help of our ablation model.

#### *6.1.3.1 Analysis of TSNE*

In this section we will analyze the tsne representations of the datasets along with clusters that were found. From figure 6-7 we can see the difference between Eu-all and enron. We would like to

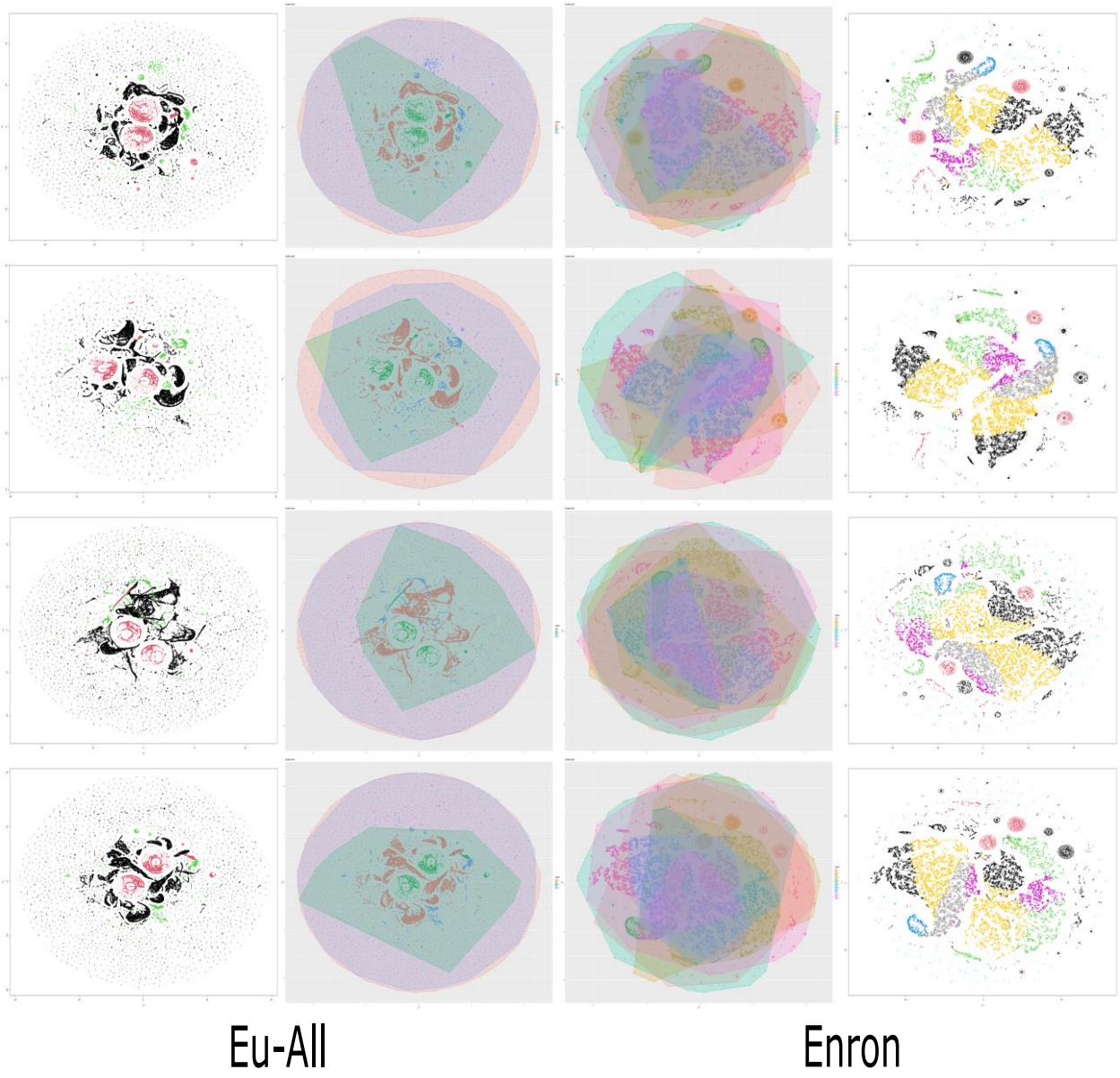


*Figure 6-7 Communication TSNE*

mention here that in order to ascertain the correct network structure, we applied four different tsne models with different parameters. We found that the representations present here are accurate and different tsne only rotated the structures with very little deformation as evident from the figure. We would also like to point out that whereas eu-all has only three clusters in the network. Enron has nine clusters in it. This can be again attributed its dishonest nature. The convex hull of the representations shows the spread of the cluster points.

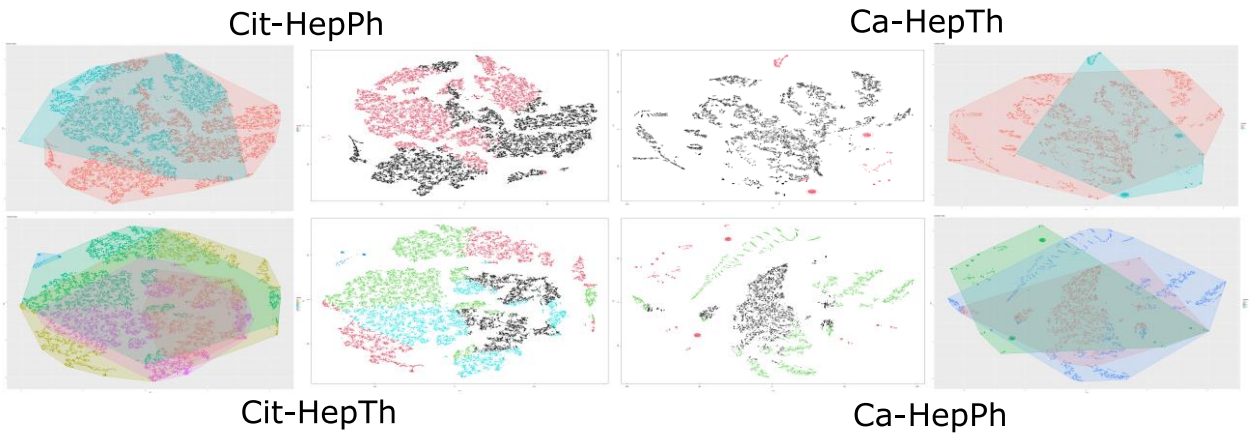


*Figure 6-8 Citation vs Collaboration TSNE*

From figure 6-8 we can see the differences between citation and collaboration networks. One interesting aspect of Cit hepTh is that it is theoretical work therefore it requires more varied citation for support, rebuttal and other means of written communication. Whereas, Cit HepPh is more experimental work, therefore has less citation as it has less diverse interest groups. However, more people collaborate on experimental domains and this can be seen in the Ca HepPh collaboration representation. It has three clusters that are quite spread out compared to Ca HepTh which only two clusters with one being the dominant one.

From figure 6-9, we can see that facebook ego has three clusters where one is rather disjoint from the others. The facebook ego network did not contain enough data. Therefore, bigger social networks are necessary to come to more concreate decisions when it comes to finding patterns and characteristics in these networks.

Facebook ego network

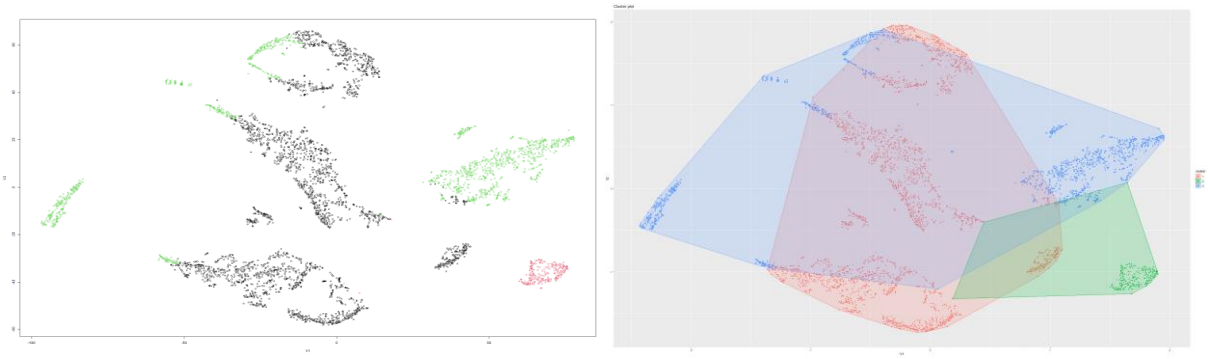

*Figure 6-9 Social network*

From figure 6-10, we see the how the P2P networks grew within a month. We can only see three clusters in P2P aug8. The rest of them all have only two clusters. P2P networks are rather chaotic, therefore, finding pattern in them might not be the best possible solution.

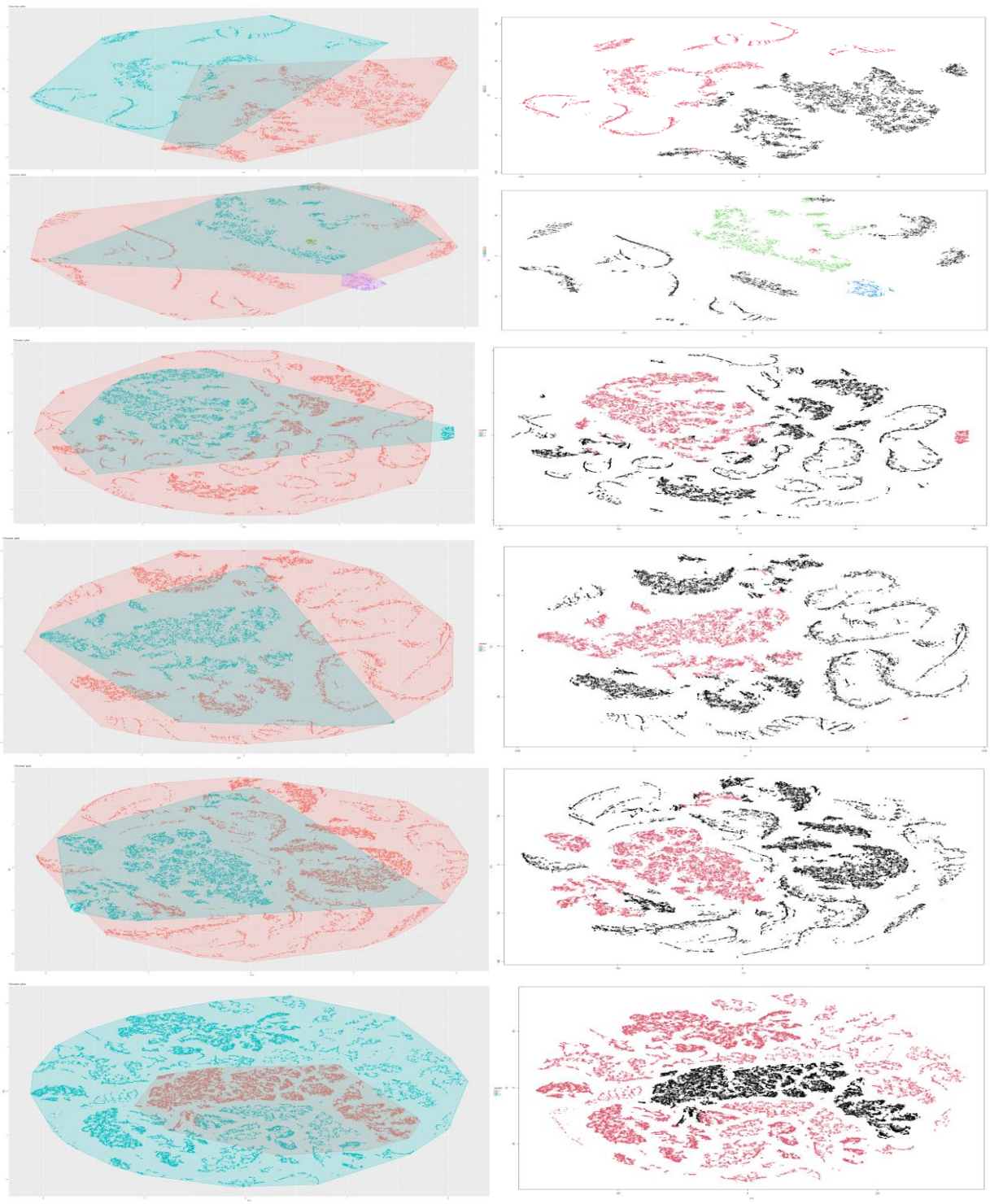

*Figure 6-10 P2P Growth*

#### *6.1.3.2 Ablation Study results*

In this section we present our ablation study. We carried out studies on nine ablation models including the baseline. We found that node based centralities are better for finding structures and patterns compared to community based centralities.

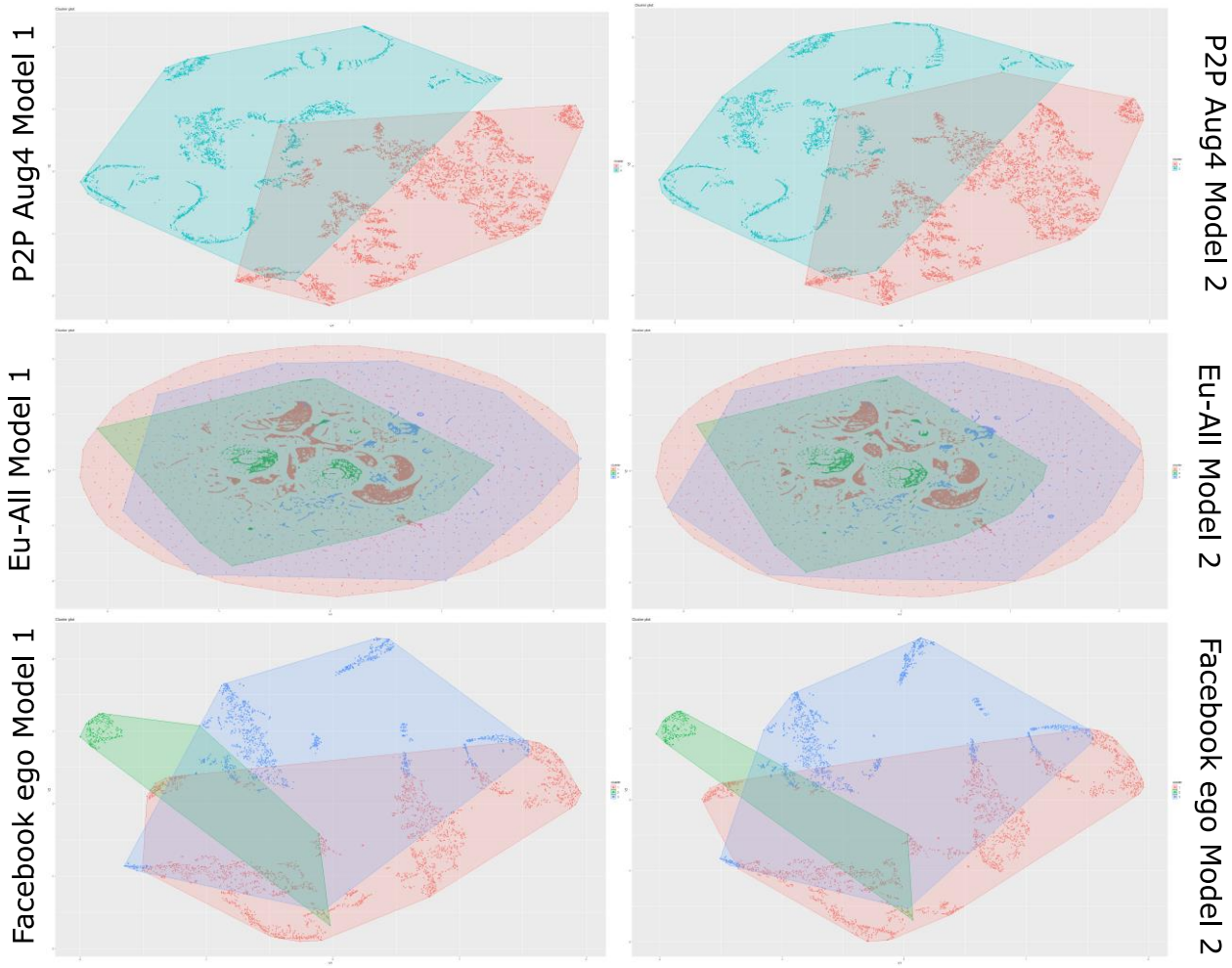


*Figure 6-11 Removal of Degree*

One of our assumptions was that degree in an influential centrality and its removal should affect our representation. However, from figure 6-11, its evident that removal of degree doesn't change the representation much apart from slight spread of the cluster points.

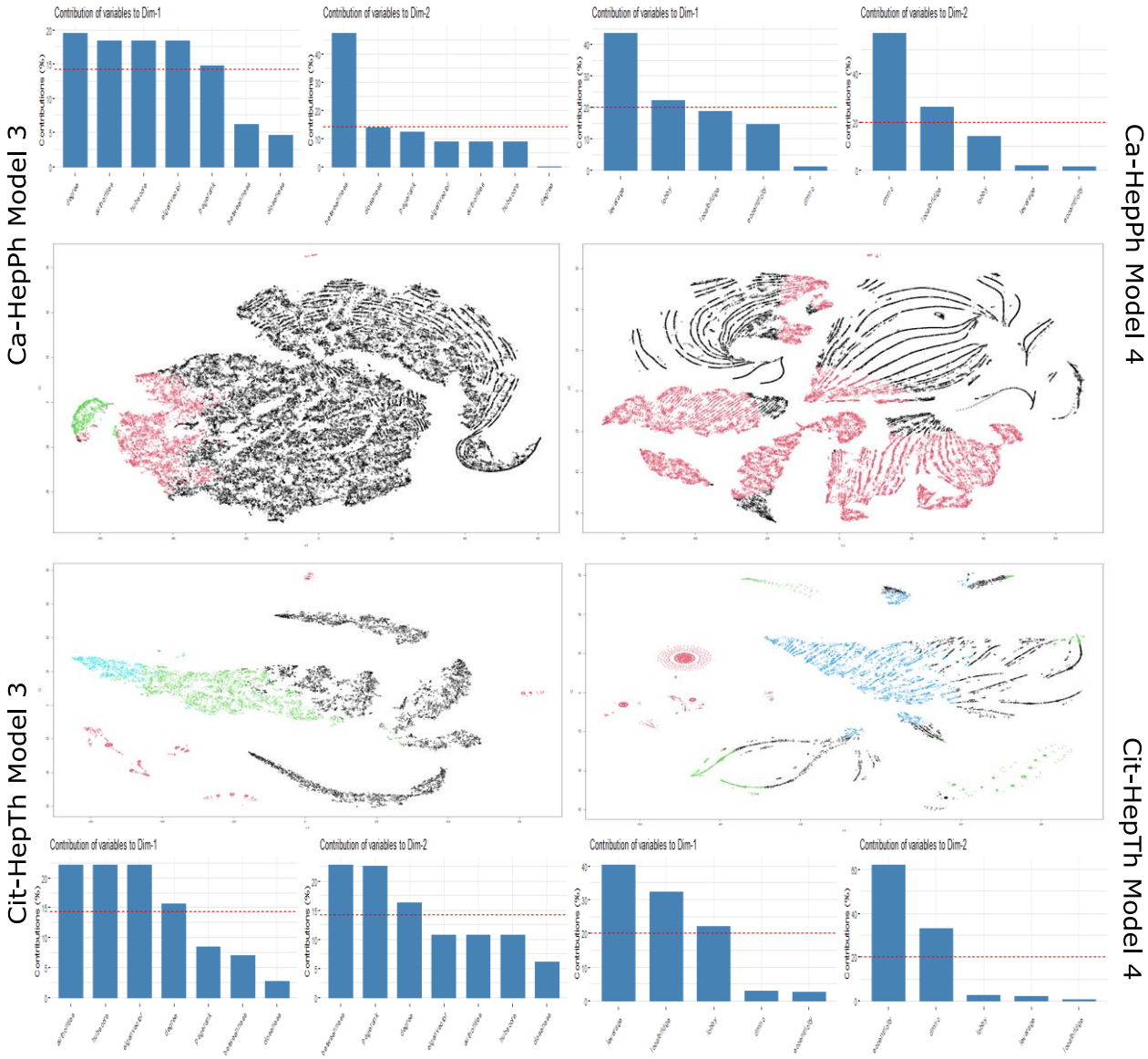


*Figure 6-12 Eccentricity effect*

From figure 6-12, we can see the effect of node based centralities and community based centralities. Node based centralities give more meaningful global structures. We assumed community based centralities would give meaningful local structure, however it is not always the case. They mostly break down to string like structures with occasional meaningful structures as seen in Cit Hep-Th model 4. We also found that eccentricity plays a major role at breaking down the more solid structures into string like structures. We can see this phenomenon in other networks also as evident by figure 6-13.

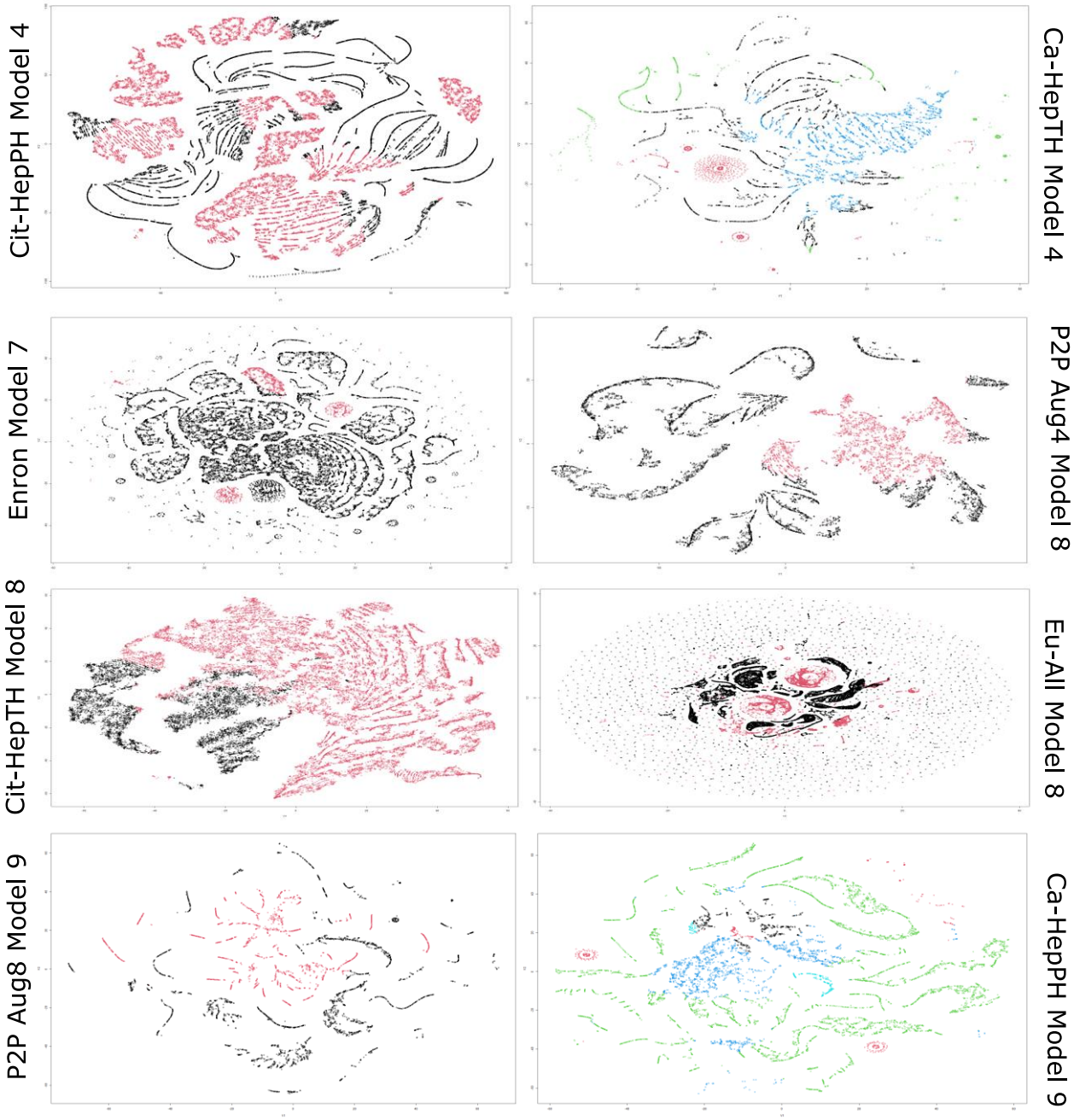


*Figure 6-13 Node centralities vs Community centralities*

## 6.2 Discussion

In this section, we would like to propose a hierarchical model of communication based on our findings.

### 6.2.1 Different Communications are related

As we have shown in the results section, there are underlying characteristics that are present in all

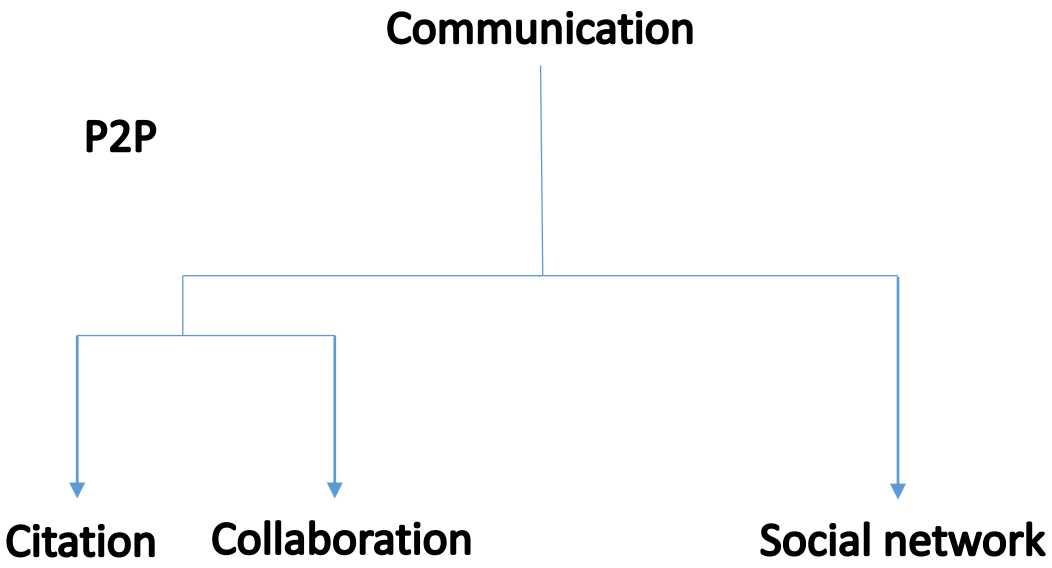


*Figure 6-14 Communication Hierarchy*

the networks except P2P. With that in mind, figure 6-14 shows how there is a hierarchy between different forms of communication. The tree diagram is not scaled to length. However, social networks should be closer to communication like email compared to collaboration and citation. Citation is very indirect written form of communication. Therefore, the hierarchy should read like formal communication like email, informal communication like social networks have more similarity compared to citation network. Collaboration is a mix of formal and informal communication. In figure 6-15, we present the hierarchy found in the datasets we used. We found that enron and Ca-HepPh shared similarities and citation and collaboration had similar underlying characteristics.

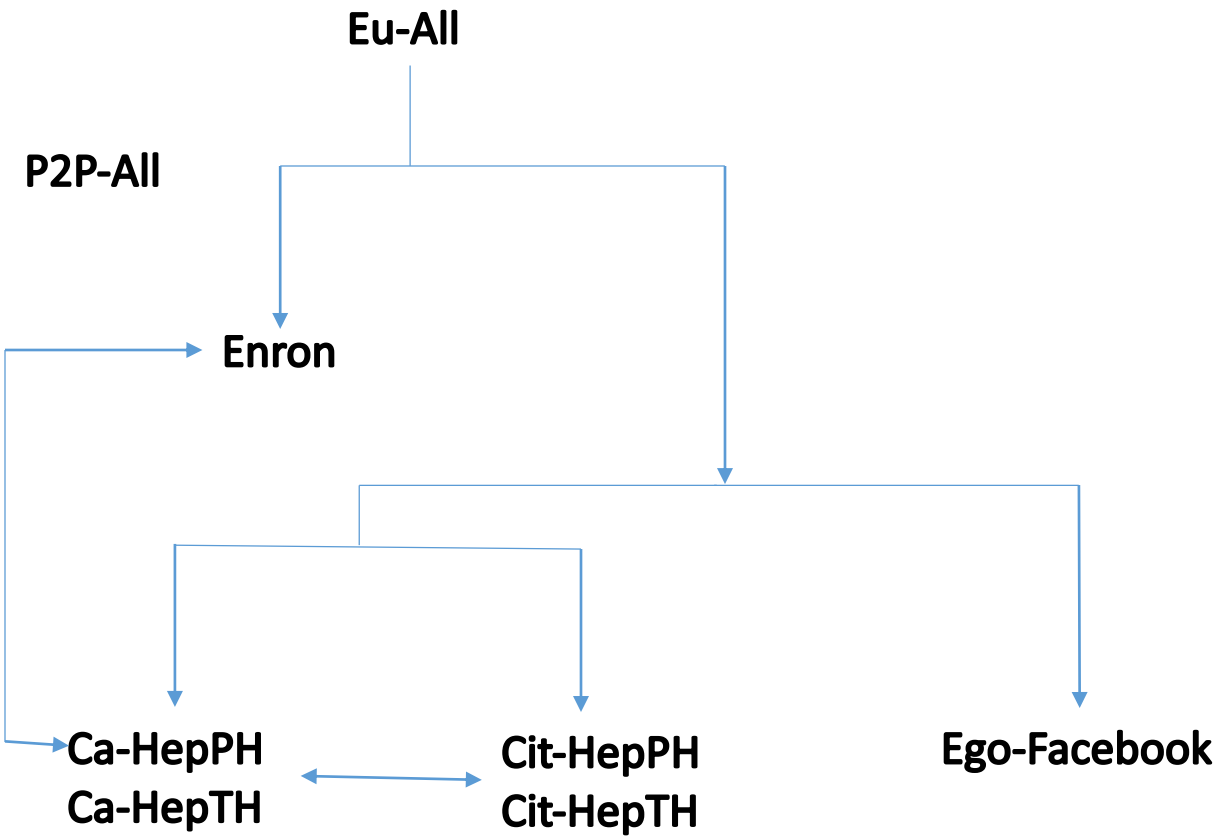


*Figure 6-15 Datasets Hierarchy*

### 6.2.2 Is Communication a Galaxy?

Lastly, we would like to ponder the question, if communication hierarchy is a galaxy. Figure 6-16 represents the Eu-All network. If we break it down different sections of the Eu-All network, we can see that they also show patterns in them that are similar to what we have seen in other networks i.e. figure 6-17. Figure 6-17 can be compared to figure 6-16 D as they show similarity structure wise. In our communication hierarchy facebook was close to eu-all as per correlation matrix. Therefore, if the self-repeating nature of communication is equivalent to a galaxy, then that can have meaningful implications. For instance, we can analyze national communication data and find meaningful structures that can help us identify dishonest networks like Enron.

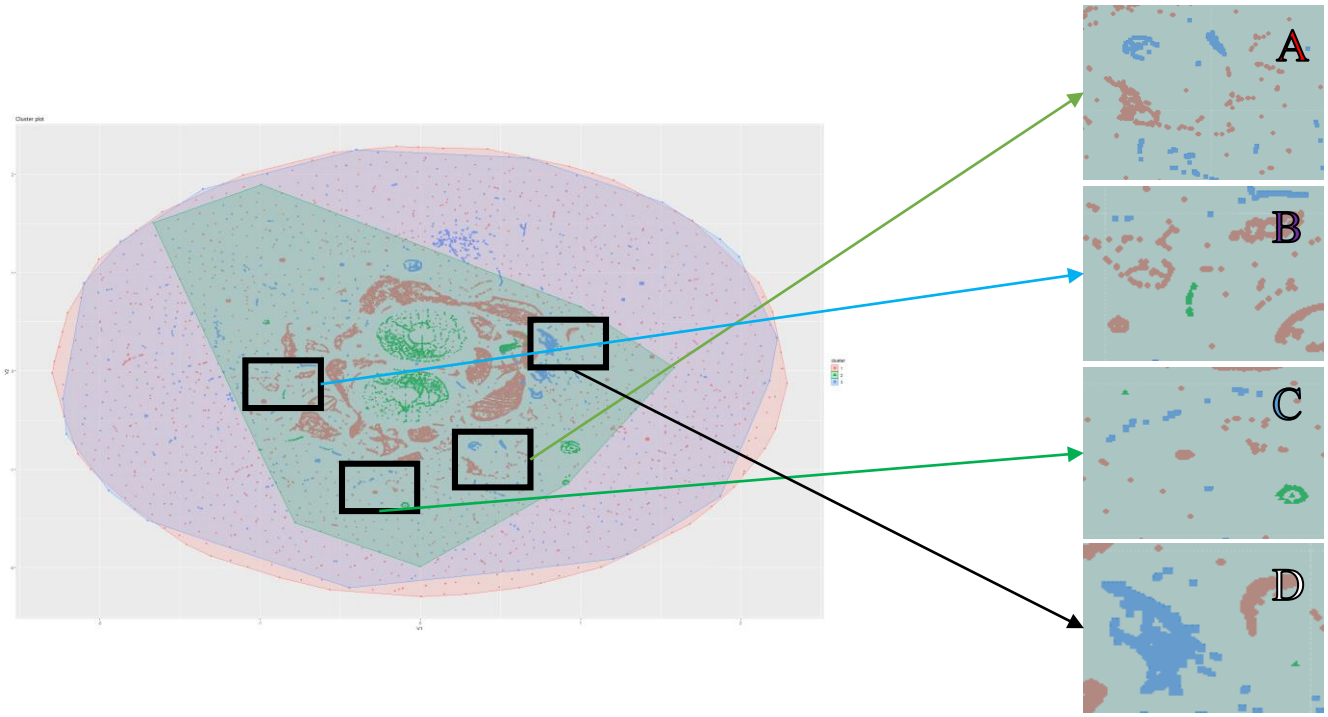


*Figure 6-16 Galaxy Model of Communication*

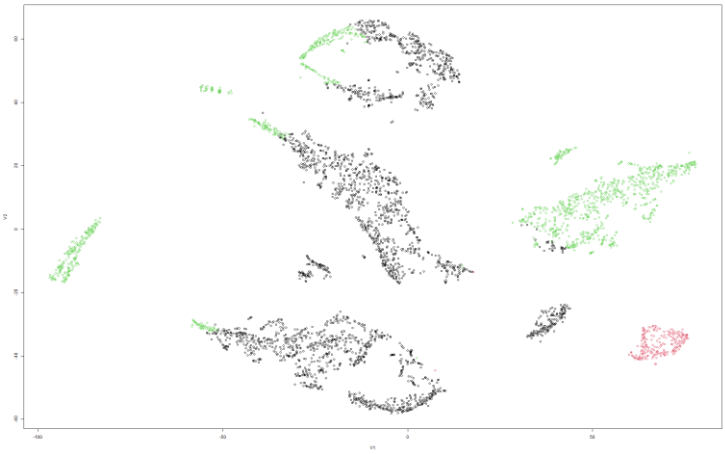

*Figure 6-17 Facebook ego network*

Therefore, more analysis is required with larger network datasets to better understand this phenomenon. However, the biggest drawback of analyzing large network datasets is they become computationally intractable.

# CHAPTER 7: CONCLUSION AND FUTURE WORK

The primary goal of this study was to analyze underlying characteristics of different networks and also visualizing these networks to discover meaningful pattern within them. For this work we employed the help for non-linear dimensionality reduction algorithms and visualized these networks in a two dimensional manifold. Although this let us peek at the surface of what these networks have to offer as knowledge, we believe more study is necessary. We saw that even though we could come to some conclusion from our results, we still needed some network specific knowledge to make those findings more meaningful. Therefore, more quantitative analysis of different networks is necessary to be able to make knowledge discovery through visualization. For future, we would like to employ graph neural networks to try to find knowledge in higher dimension and see if we can come to an intersection of high dimensional and low dimensional knowledge discovery.

In conclusion, in this work we studied different internet networks to find how human behavior drives their structure. We also found that some networks are less meaningful like P2P, whereas email networks like enron and eu-all are more meaningful when it comes to knowledge discovery. We also proposed that different forms of communications have a hierarchy among them. From this, we also proposed a galaxy model that contains this hierarchy and self-replicating nature of networks.

# Appendix

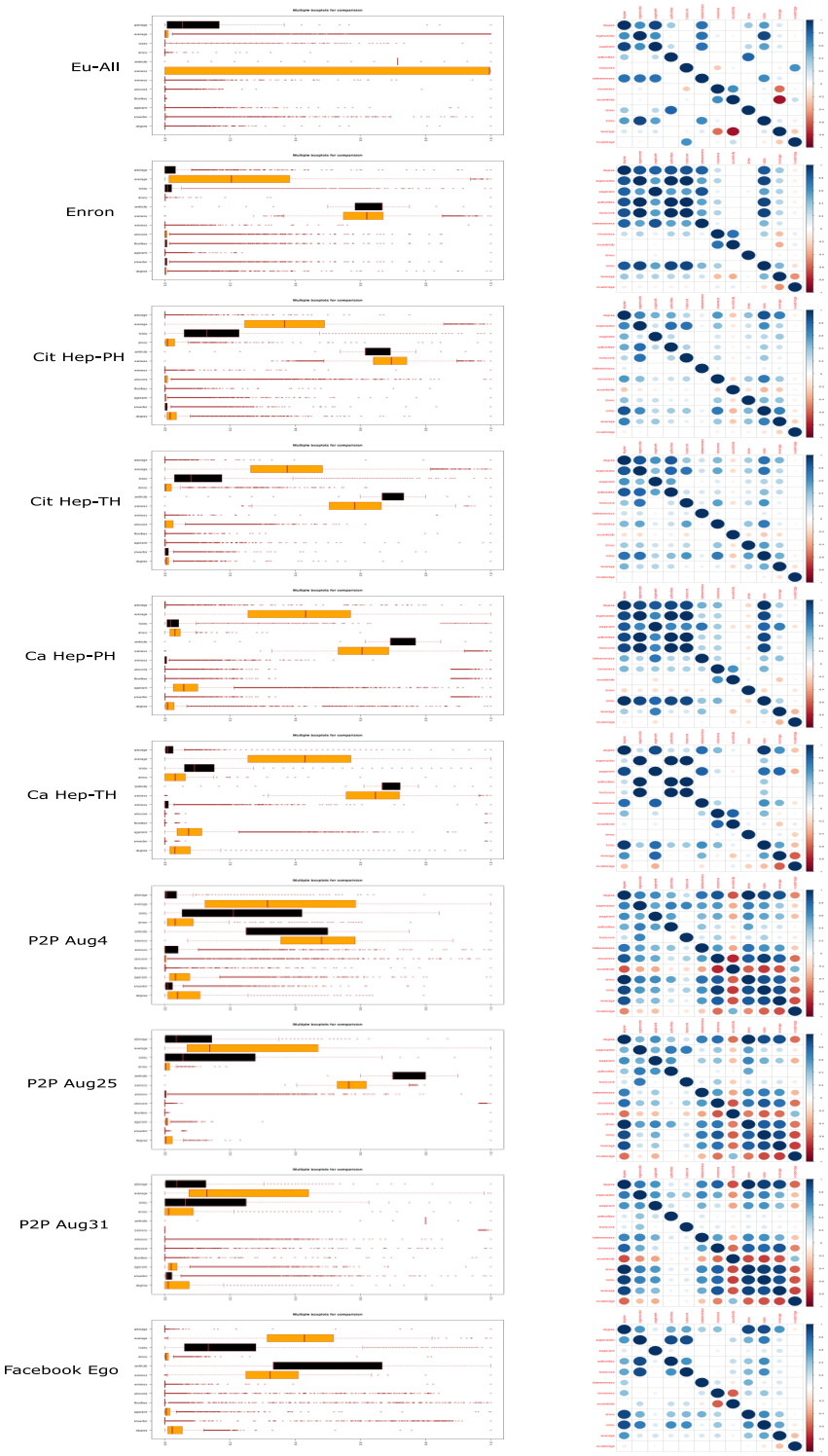


*Figure 8-1 Boxplot and Correlation matrix of datasets*

# REFERENCES


[1] P. Crucitti, V. Latora, and S. Porta, “Centrality in networks of urban streets,” *Chaos*, vol. 16, no. 1, 2006, doi: 10.1063/1.2150162.

[2] A. Landherr, B. Friedl, and J. Heidemann, “A Critical Review of Centrality Measures in Social Networks,” *Bus. Inf. Syst. Eng.*, vol. 2, no. 6, pp. 371–385, 2010, doi: 10.1007/s12599-010-0127-3.

[3] M. Newman, *Networks: An Introduction*. 2010.

[4] C. C. Chiu, P. Balkunid, and F. Weinberg, “When managers become leaders : The role of manager network centralities , social power , and followers ’ perception of leadership,” *Leadersh. Q.*, 2016, doi: 10.1016/j.leaqua.2016.05.004.

[5] T. Wang and H. Krim, "STATISTICAL CLASSIFICATION OF SOCIAL NETWORKS", in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2012.

[6] S. P. Borgatti, “Centrality and network flow ☆,” vol. 27, no. November 2003, pp. 55–71, 2005, doi: 10.1016/j.socnet.2004.11.008.

[7] U. Brandes and C. Pich, “Centrality estimation in large networks,” *Int. J. Bifurc. Chaos*, vol. 17, no. 7, pp. 2303–2318, 2007, doi: 10.1142/S0218127407018403.

[8] T. Dwyer, D. Kosch, and K. Xu, “Visual Analysis of Network Centralities,” no. Wuchty 2002, 2003.s

[9] V. Latora and M. Marchiori, “Efficient behavior of small-world networks,” *Phys. Rev. Lett.*, vol. 87, no. 19, pp. 198701-1-198701–4, 2001, doi: 10.1103/PhysRevLett.87.198701.

[10] P. Hage and F. Harary, “Eccentricity and centrality in networks,” *Soc. Networks*, vol. 17, no. 1, pp. 57–63, 1995, doi: 10.1016/0378-8733(94)00248-9.

[11] Chen, S.H.; Chin, C.H.; Wu, H.H.; Ho, C.W.; Ko, M.T.; Lin, C.Y. Cyto-Hubba: A Cytoscape plug-in for hub object analysis in network biology. In Proceedings of the 20th International Conference on Genome Informatics, Yokohama, Japan, 14–16 December 2009.

[12] A. Korn, A. Schubert, and A. Telcs, “Lobby index in networks,” *Phys. A Stat. Mech. its Appl.*, vol. 388, no. 11, pp. 2221–2226, 2009, doi: 10.1016/j.physa.2009.02.013.

[13] K. E. Joyce, P. J. Laurienti, J. H. Burdette, and S. Hayasaka, “A new measure of centrality for brain networks,” PLoS One, vol. 5, no. 8, 2010, doi: 10.1371/journal.pone.0012200.

[14] J. P. MacKer, “An improved local bridging centrality model for distributed network analytics,” *Proc. - IEEE Mil. Commun. Conf. MILCOM*, pp. 600–605, 2016, doi: 10.1109/MILCOM.2016.7795393.

[15] H. Chen, H. Yin, T. Chen, Q. Viet, H. Nguyen, and W. P. Xue, “Exploiting Centrality Information with Graph Convolutions for Network Representation Learning,” *2019 IEEE 35th Int. Conf. Data Eng.*, pp. 590–601, 2019, doi: 10.1109/ICDE.2019.00059.

[16] P. Bródka, K. Skibicki, P. Kazienko, and K. Musiał, “A degree centrality in multi-layered social

[17] A. Culotta and J. Cutler, “Predicting Twitter User Demographics using Distant Supervision from Website Traffic Data,” vol. 55, pp. 389–408, 2016.

[18] A. Mccallum, “Topic and Role Discovery in Social Networks with Experiments on Enron and Academic Email,” vol. 30, pp. 249–272, 2007.

[19] S. A. Williamson and M. Tec, “Random clique covers for graphs with local density and global sparsity,” *35th Conf. Uncertain. Artif. Intell. UAI 2019*, 2019.

[20] L. P. Cordella, P. Foggia, C. Sansone, and M. Vento, “A (sub)graph isomorphism algorithm for matching large graphs,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 10, pp. 1367–1372, 2004, doi: 10.1109/TPAMI.2004.75.

[21] J. Hussain and M. A. Islam, “Evaluation of graph centrality measures for tweet classification,” *2016 Int. Conf. Comput. Electron. Electr. Eng. ICE Cube 2016 - Proc.*, pp. 126–137, 2016, doi: 10.1109/ICECUBE.2016.7495209.

[22] P. R. Miller, P. S. Bobkowski, D. Maliniak, and R. B. Rapoport, “Talking Politics on Facebook : Network Centrality and Political Discussion Practices in Social Media,” 2015, doi: 10.1177/1065912915580135.s

[23] A. M. Cohn *et al.*, “Discussions of Alcohol Use in an Online Social Network for Smoking Cessation: Analysis of Topics, Sentiment, and Social Network Centrality,” *Alcohol. Clin. Exp. Res.*, vol. 43, no. 1, pp. 108–114, 2019, doi: 10.1111/acer.13906.

[24] G. Rossman, N. Esparza, and P. Bonacich, “I’d like to thank the academy, team spillovers, and network centrality,” *Am. Sociol. Rev.*, vol. 75, no. 1, pp. 31–51, 2010, doi: 10.1177/0003122409359164.

[25] B. Yang and J. Liu, “Discovering global network communities based on local centralities,” *ACM Trans. Web*, vol. 2, no. 1, 2008, doi: 10.1145/1326561.1326570.

[26] Y. Zhang, X. Wang, P. Zeng, and X. Chen, “Centrality characteristics of road network patterns of traffic analysis zones,” *Transp. Res. Rec.*, no. 2256, pp. 16–24, 2011, doi: 10.3141/2256-03.

[27] S. H. Lee, J. Y. Choi, S. H. Yoo, and Y. G. Oh, “Evaluating spatial centrality for integrated tourism management in rural areas using GIS and network analysis,” *Tour. Manag.*, vol. 34, pp. 14–24, 2013, doi: 10.1016/j.tourman.2012.03.005.

[28] D. C. Bell, J. S. Atkinson, and J. W. Carlson, “Centrality measures for disease transmission networks,” *Soc. Networks*, vol. 21, no. 1, pp. 1–21, 1999, doi: 10.1016/S0378-8733(98)00010-0.

[29] S. Narayanan, “The Betweenness Centrality Of Biological Networks A Study of Betweenness Centrality,” 2005.

[30] K. Park and D. Kim, “Localized network centrality and essentiality in the yeast – protein interaction network,” pp. 5143–5154, 2009, doi: 10.1002/pmic.200900357.

[31] G. L. De La Peña Sarracén and P. Rosso, “Automatic text summarization based on betweenness centrality,” *ACM Int. Conf. Proceeding Ser.*, vol. Part F1377, 2018, doi: 10.1145/3230599.3230611.

[32] Q. Wu, X. Qi, E. Fuller, and C. Q. Zhang, “‘follow the leader’: A centrality guided clustering and its application to social network analysis,” *Sci. World J.*, vol. 2013, 2013, doi: 10.1155/2013/368568.

[33] X. Huang, Y. Zhao, J. Yang, C. Zhang, and X. Ye, “TrajGraph : A Graph-Based Visual Analytics Approach to Studying Urban Network Centralities Using Taxi Trajectory Data,” vol. 22, no. 1, pp. 160–169, 2016.

[34] "Stanford Large Network Dataset Collection", Snap.stanford.edu, 2021. [Online]. Available: http://snap.stanford.edu/data. [Accessed: 19- Jan- 2021].

[35] J. Leskovec, J. Kleinberg and C. Faloutsos. Graph Evolution: Densification and Shrinking Diameters. ACM Transactions on Knowledge Discovery from Data (ACM TKDD), 1(1), 2007.