UNIVERSITAT DE BARCELONA

Data Driven Approach to Enhancing Efficiency and Value in Healthcare

Richard E. Guerrero Ludueña

Part IV

Mining social networks

Chapter 6

Organisational email knowledge extraction with Social Network Analysis (SNA): Concepts and empirical analysis

Chapter 2 introduced the Social Network Analysis (SNA) and its relation to other fields such as Data Mining. Social Network Analysis conceptualises individuals or groups as *points* and their relations to each other as *lines* [192]. It is concerned with the patterns formed by these points and lines, exploring these patterns, mathematically or visually, in order to assess their effect on the individuals and organisations that are the members of the 'networks' formed by the intersecting lines that connect them. Treating a social structure as a network is the cornerstone of SNA [209, 42].

This chapter describes the application of SNA for organisational email mining. Thanks to its efficiency, low cost and compatibility of diversified types of information, email is one of the most popular forms of communication between co-workers and organisations [12, 49].

This chapter is structured as follows: Section 6.1 introduces the problem and research aims; Section 6.2 presents the research methods and design. The intervention strategies and results are presented in Section 6.3; Section 6.4 introduced a new framework; Section 6.5 will focus upon conclusions and study limitations; and Section 6.6 summarises the chapter.

6.1 Introduction

The study was conducted at the Wessex Academic Health Science Network (Wessex AHSN), one of the 15 AHSNs across England, established by the NHS in 2013 to spread innovation, improve health for patients, and generate regional economic growth across the healthcare sector. AHSN are the only bodies that connect NHS and academic organisations, local authorities, the third sector and industry [152].

This project started with a request from a senior manager, who wanted to identify the network of stakeholders and organisations related with the Wessex AHSN, and also identify key contacts within the network. Social Network Analysis was selected as the most suitable methodology to answer this questions.

The study was developed using Social Network Analysis (SNA) (see Section 2.4.2). Email was selected as a data source for the straightforward approach required to collect the information, but more importantly, because it was identified as the most rich data source of the given options.

In the next section, we will discuss the application of SNA to study email contact and email network property analysis, we will also present email visualisation using SNA techniques. Applying a SNA framework to merged email datasets differentiates our research and is the lens through which we address the following questions:

1. How can we use email mining and SNA to identify the properties and structure of the communication network in one specific Wessex AHSN programme?

2. What are the agent specific patterns of different individuals within the organisation?

3. What are the properties and structure of the communication networks formed merging data from different individuals?

6.2 Methods

Email activities represent human social, organisational relationships. Emails networks are social networks where the nodes are emails contacts and the edges are email interaction [49]. One node can be a hub for several email addresses belonging to the same individual [42]. Edges are created according to particular criteria, for example, one edge represents an email exchange between two contacts, or can be present only if the contacts exchange emails more than a defined number of times [12].

Email networks record rich information about communication among people and organisations: understanding the properties of an email network could help us recognise how people communicate with each other, to understand how an organisation interacts with stakeholders, and also to identify key individuals within the network. Nevertheless, very few organisations use email mining as a management tool.

In [205], the author presented a survey analysis of the main tasks, techniques and tools related with Email mining. The author analysed five main tasks related to email mining: spam detection; email categorisation; contact analysis; email network property analysis; and email visualisation.

There are two types of email networks: a personal, or egocentric, email network and a complete, or whole email network. A personal email network is a network built from one individual's email account, centred by the node of the email account owner. A whole email network is a network built from an email data corpus of an organisation and can be viewed as a combination of many egocentric networks.

Graphs were used to model the email social networks; the email networks were modelled as a social network graph $G = (V, E)$, were V is the set of email addresses as nodes and E is the email interactions as edges.

This social structure based approach will extract a social network formed by emailing activities. The analysis here presented, was developed using email mining and SNA, and following the methodology described at the section 2.4.2.3:

Problem and network to focus on. The analysis focuses on the identification of the network of stakeholders and organisations related with the Wessex AHSN, as well as the identification of key contacts within those networks.

Data sources. Different options for data sources where evaluated, e.g. a Customer relationship Management (CRM) system; major social networks as Facebook, Twitter, and Linkedin; Network surveys, asking people to manually characterise their relationship with other people; and finally, examining an existing stakeholders database. Email traffic was selected as a data source mainly because of the low effort level required to collect the information, but also because it was identified as the richest data source between of the options discussed above.

Network members. Four individuals within the organisation were selected for this study.In all cases, participation was voluntary. Network members were defined as all the emails contacts for those individuals.

Edges, nodes and type of network. Nodes are defined as email contacts and the edges are email interactions, including received and sent emails directly or as a Cc. Each edge represents a directed relation from one sender to one recipient as captured

From: (Name)	From: (Address)	To: (Name)	To: (Address)
Guerrero-Luduena R.E.	/o=University of Southampton/ou=Exchange Administrative Group (FYDIBOHF23SPDLT)/cn=Recipients/cn=rel1a13581	'rel1a13@soton.ac.uk'	rel1a13@soton.ac.uk
Guerrero-Luduena R.E.	/o=University of Southampton/ou=Exchange Administrative Group (FYDIBOHF23SPDLT)/cn=Recipients/cn=rel1a13581	'rel1a13@soton.ac.uk'	rel1a13@soton.ac.uk
Guerrero-Luduena R.E.	/O=UNIVERSITY OF SOUTHAMPTON/OU=EXCHANGE ADMINISTRATIVE GROUP (FYDIBOHF23SPDLT)/CN=RECIPIENTS/CN=REL1A13581	rel1a13@soton.ac.uk	rel1a13@soton.ac.uk
Guerrero-Luduena R.E.	/O=UNIVERSITY OF SOUTHAMPTON/OU=EXCHANGE ADMINISTRATIVE GROUP (FYDIBOHF23SPDLT)/CN=RECIPIENTS/CN=REL1A13581	rel1a13@soton.ac.uk	rel1a13@soton.ac.uk
Guerrero-Luduena R.E.	/O=UNIVERSITY OF SOUTHAMPTON/OU=EXCHANGE ADMINISTRATIVE GROUP (FYDIBOHF23SPDLT)/CN=RECIPIENTS/CN=REL1A13581	Guerrero-Luduena R.E.	/O=UNIVERSITY OF SOUTHAMPTON/OU=EXCHANGE ADMINISTRATIVE GROUP (FYDIBOHF23SPDLT)/CN=RECIPIENTS/CN=Rel1a13581
Guerrero-Luduena R.E.	/O=UNIVERSITY OF SOUTHAMPTON/OU=EXCHANGE ADMINISTRATIVE GROUP (FYDIBOHF23SPDLT)/CN=RECIPIENTS/CN=REL1A13581	Guerrero-Luduena R.E.	/O=UNIVERSITY OF SOUTHAMPTON/OU=EXCHANGE ADMINISTRATIVE GROUP (FYDIBOHF23SPDLT)/CN=RECIPIENTS/CN=Rel1a13581
Guerrero-Luduena R.E.	/O=UNIVERSITY OF SOUTHAMPTON/OU=EXCHANGE ADMINISTRATIVE GROUP (FYDIBOHF23SPDLT)/CN=RECIPIENTS/CN=REL1A13581	Guerrero-Luduena R.E.	/O=UNIVERSITY OF SOUTHAMPTON/OU=EXCHANGE ADMINISTRATIVE GROUP (FYDIBOHF23SPDLT)/CN=RECIPIENTS/CN=Rel1a13581
Guerrero-Luduena R.E.	/o=University of Southampton/ou=Exchange Administrative Group (FYDIBOHF23SPDLT)/cn=Recipients/cn=rel1a13581	Guerrero-Luduena R.E.	/o=University of Southampton/ou=Exchange Administrative Group (FYDIBOHF23SPDLT)/cn=Recipients/cn=rel1a13581
Guerrero-Luduena R.E.	/o=University of Southampton/ou=Exchange Administrative Group (FYDIBOHF23SPDLT)/cn=Recipients/cn=rel1a13581	Guerrero-Luduena R.E.	/o=University of Southampton/ou=Exchange Administrative Group (FYDIBOHF23SPDLT)/cn=Recipients/cn=rel1a13581
Guerrero-Luduena R.E.	/o=University of Southampton/ou=Exchange Administrative Group (FYDIBOHF23SPDLT)/cn=Recipients/cn=rel1a13581	Guerrero-Luduena R.E.	/o=University of Southampton/ou=Exchange Administrative Group (FYDIBOHF23SPDLT)/cn=Recipients/cn=rel1a13581
Guerrero-Luduena R.E.	/o=University of Southampton/ou=Exchange Administrative Group (FYDIBOHF23SPDLT)/cn=Recipients/cn=rel1a13581	Guerrero-Luduena R.E.	/o=University of Southampton/ou=Exchange Administrative Group (FYDIBOHF23SPDLT)/cn=Recipients/cn=rel1a13581

Figure 6.1: Email dataset structure before data munging.

in the email data. Recipients were retrieved from the *To*, *From*, and *Cc* fields in the emails, without further distinction between recipients *To* and *Cc*. Edges reflect contacts, meaning that if one contact sent one email to four other contacts, four edges would be formed. The edges are weighted by the cumulative frequency of emails exchanged between any pair of individuals.

The analysis is based on four email egocentric networks, analysed individually and as a new network (formed after combining the four egocentric networks into one network).

Data collection. Email communication was extracted from four individuals within the organisation, with emails hosted in two different email servers. As noted, email was selected as a data source for the straightforward approach required to collect the information, but more importantly, because it was identified as the most rich data source of the given options.

In this project, network data collection was performed by exporting a year of emails for the four individuals part of the analysis. Emails were exported to comma separated values files. Despite several available options to extract information from emails automatically, a manual data extraction was selected because of email privacy policies. Each individual performed the data extraction from their emails with the support from the researcher performing the analysis. Email messages from *Inbox* and *Sent Items* folders were exported, selecting the fields: *From: (Name)*, *From: (Address)*, *To: (Name)*, *To: (Address)*, *Cc: (Name)*, and *Cc: (Address)*.

Data munging. A iterative data preparation (data munging) process was used to transform the email datasets into a usable form for visualisation and analysis (e.g., subsetting and filtering data, aggregating data, merging data, reshaping data, and data rename). Email addresses that do not refer to actual individuals, e.g., all staff, lists, etc. were removed from the dataset. This includes a manual process to identify email address used to communicate with lists of individuals in different organisations. Figure 6.1 includes an example of one of the datasets analysed.

One of the main challenges of the project was related with data pre-processing and merging, because although the four individuals are part of the same organisation, they

are formally employees by different organisations, using two different email servers and email clients (i.e., different data structures, and different names for the same contact). Different email addresses belonging to the same person were merged to create one node (e.g., R.E.GuerreroLuduena@soton.ac.uk, rel1a13@soton.ac.uk).

Network visualisation and analysis. Network visualisation and analysis was developed using the algorithm ForceAtlas 2 [100] from the software Gephi [139]. ForceAtlas 2 is a good choice when we want to put nodes into communities.

The email network analysis was divided into three main parts describes as follows:

- *N1 Individual 1 network (research question 1).* The network of one of the Wessex AHSN's main programmes was analysed using the programme lead emails as a source of information (Individual 1). An egocentric email network (N1) was built using emails sent and received. Cc emails were also included, and edges between the other nodes in the networks (existing when they exchange emails), including when our individual in the study is part of the recipients included in copy. Multiple recipients and Cc contacts are modelled as different nodes and edges in the network. The project manager included in the analysis is the main contact with co-workers, other internal programme leads, and with external organisations and stakeholders involved in the programme. Therefore, a SNA of her email network is a good representation of the entire programme network. Both node-specific analysis and aggregate network metric were performed.

- *N2, N3, N4 email networks (research question 2).* Three other Wessex AHSN staff email networks (N2, N3, N4) were analysed to compare similarities and differences in the networks developed using contacts from individuals working in one specific programme and individuals working across programmes. N1 is based on the emails contacts of a data analyst (Individual 2) working across different programmes; N3 is based on the emails of an operational researcher Individual 3) involved in one specific programme; and N4 is based on emails of a project support officer (Individual 4). Node-specific analysis and aggregate network metrics were performed for each network.

- *N5 network (research question 3).* After the four individual egocentric email network analysis, a new email network (N5) was developed combining the four individual-centred networks into one, and an aggregate network analysis was performed to analyse the potential impact of email mining and SNA as a management tool for the organisation.

Table 6.1: Number of nodes and edges of the five networks

	N1	N2	N3	N4	N5
Nodes	3,239	1,853	401	2,409	7,225
Edges	91,877	132,920	3,494	13,399	236,247
Network density	0.018	0.077	0.044	0.005	0.009

Network statistics. Nodes, edges and network metrics described in Section 2.4.2.3 were estimated. One of the key problems in network science and SNA is the identification of the most important (or central) nodes [67]. The four best-know centrality measures are Degree, Betweenness, Closeness and Eigenvector centralities [67], each of which views centrality from a different perspective, focusing on certain traits that make nodes central (or important) to the network [116]. Estimation of those centrality measures, as well as nodes and edges statistics will be described in the following section.

6.3 Results

We have used a number of different datasets representing a variety of email networks (N1 - N5). Network and edge metrics described in Section 2.4.2.4 have been used to provide an overview of, and compare, email networks. Table 6.1 report the number of nodes and edges for the five networks.

From the Table 6.1 we can see the differences between the four egocentric networks (N1 - N4). N1 is the network with the largest number of node (3,239), and N2 is the network with the largest number of edges (132,920). N3 is the shortest network, with 401 nodes and 3,494 edges.

Network density refers to the portion of the potential connections in a network that are actual connections. A potential connection is an edge that could potentially exist between two nodes, regardless of whether or not it actually does. As our analysis is based on an undirected (or symmetric) networks, density is calculated relative to the number of unique pairs (or potential connections).

Given a network $G = (V, E)$ where $V = 1, \ldots, n$ and $E = e_1, \ldots, e_m$, the number of edges of G is at most $\frac{n(n-1)}{2}$. In SNA this value represents the potential connections k_n, and then a network density is calculated as $\frac{|V|}{k_n}$. It is clear from Table 6.1 that N2 is the network with the highest density (0.077). Nevertheless, from the definition presented in Section 2.4.2.4, all the networks are sparse, with a link density ≈ 0.

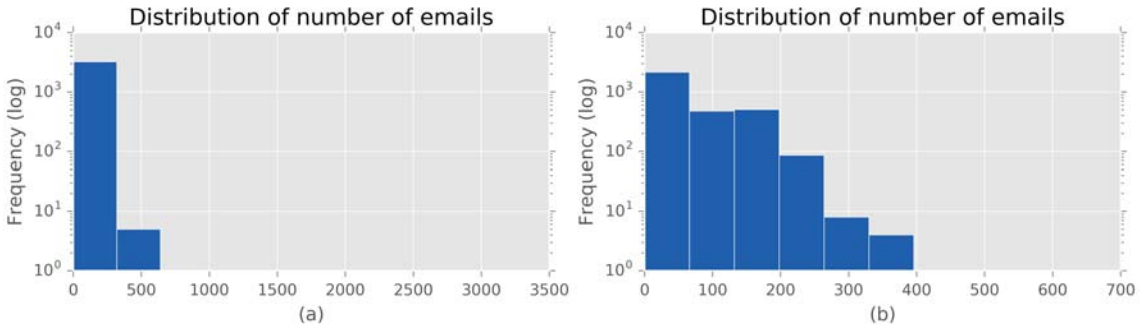

Figure 6.2: Distribution of number of emails sent and received for each contact. (a) All nodes included. (b) In order to facilitate the graph evaluation, two nodes were excluded due to the relative high frequency of email communication. The first is the network owner (frequency = 8,965), and the second is the clinical lead of the project (frequency = 3,352 emails)

6.3.1 Analysis of N1

As noted, N1 is an egocentric email network built using emails sent and received for a programme manager, who is the main contact with co-workers involves in different projects, and with external organisations and stakeholders.

Figure 6.2 shows a histogram with the distribution of emails sent and received for each node of the network. The figure also represents the distribution of emails, after deleting the two contacts with highest frequency. As expected in an egocentric network analysis, the network owner is the contact with the higher frequency of communication within the network. Unexpectedly, a second actor appeared as a contact with a high influence within the network. A interview with the network owner showed that this contact is the programme clinical lead.

Centrality is a measure used to determine the relative importance of a node within the network. Closeness, Betweenness, Eigenvector and Degree are all measures of centrality. Those centrality measures were calculated for each of the 3,239 nodes of the network N1. Figure 6.3 shows the distribution of centralities of email network N1. Betweenness centrality computes the node importance within the network by taking into account the node's neighbour connectivity. It describes the number of indirectly connected nodes, through node's direct links. Closeness centrality measures the proximity of a node from the rest of the network, by the inverse of the sum of the shortest distances between each node and every other node in the network. Degree measures the number of edges to other nodes in the network. Eigenvector is

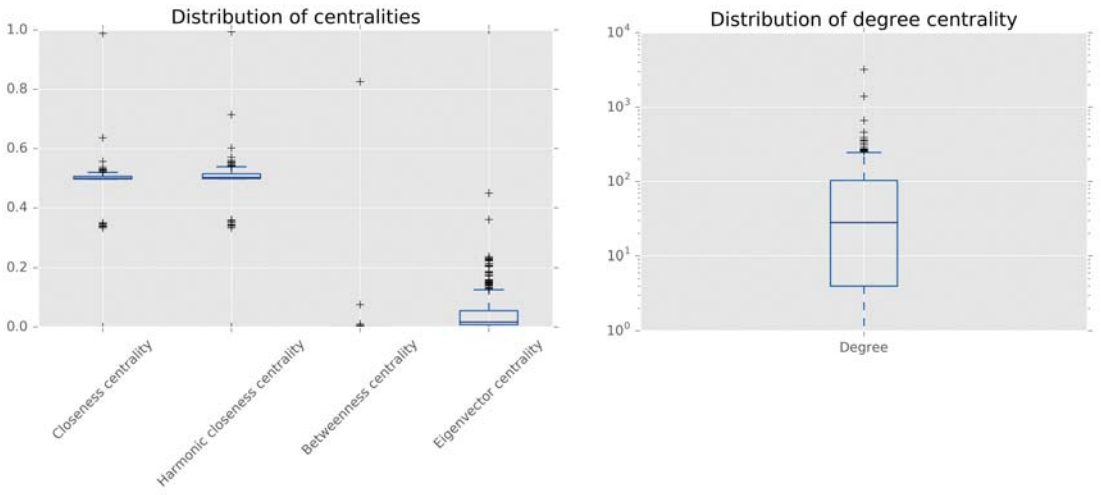

Figure 6.3: Box plots with distribution of centralities of email network N1.

a extension of degree centrality, based on the hypothesis that a node is important if it is linked to by other important nodes.

Despite it is not presented in this Thesis, because data protection, those centrality measures were used to identify the relative importance of different contacts within the network.

Degree centrality was used to identify very connected and potentially popular individuals within the network. The assumption is that those individuals are likely to hold the most information and are the individuals who can quickly connect with the wider network.

Betweenness centrality was used for finding the individuals who influence the information flow through the network, i.e., individuals acting as 'bridges' between nodes and clusters in a network.

Individuals who are the best placed to influence the entire network most quickly were identified using Closeness centrality.

6.3.2 Visualisation of email network N1

Following the statistical analysis of N1, the network was visualised using colour codes for centralities and to identify cluster within the network.

Figure 6.4 shows a visualisation of the email network N1 and the nodes of influence within the whole network. A colour code was used to represent node's betweenness centrality (from blue to red, with red as the nodes with highest betweenness central-

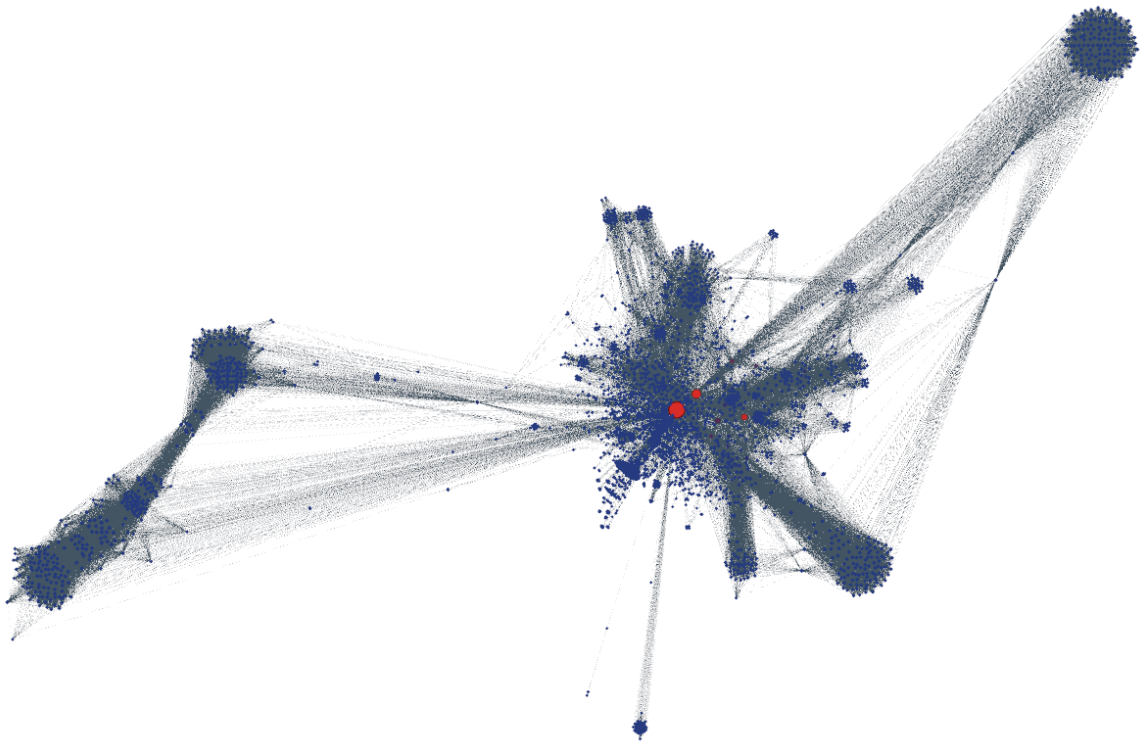

Figure 6.4: Visualisation of email network N1. Colour represents betweenness centrality (from blue to red), and node size represents degree centrality. Dark grey represents high edge weight (number of emails between two contacts).

ity). Node size represent degree centrality. Edge's weight was also represented using colours, where dark grey represents high Edge weight (number of emails between two contacts).

A combination of analysis of centrality measures and a visual inspection helped to identify key individuals within the network. A colour code was used to represent the clusters identified in the network. Figure 6.5 shows the results of the communities detection using the modularity feature. Colour coding of the visualisation is used to present communities. Degree centrality is represented with node size.

6.3.3 Correlation between centrality measures

Table 6.2 reports the correlations between centrality measures of the network N1. Correlation between measures varied. The highest correlation was between Closeness centrality and Harmonic Closeness centrality (r=0.98), perhaps because harmonic centrality reverses the sum and reciprocal operations in the definition of closeness

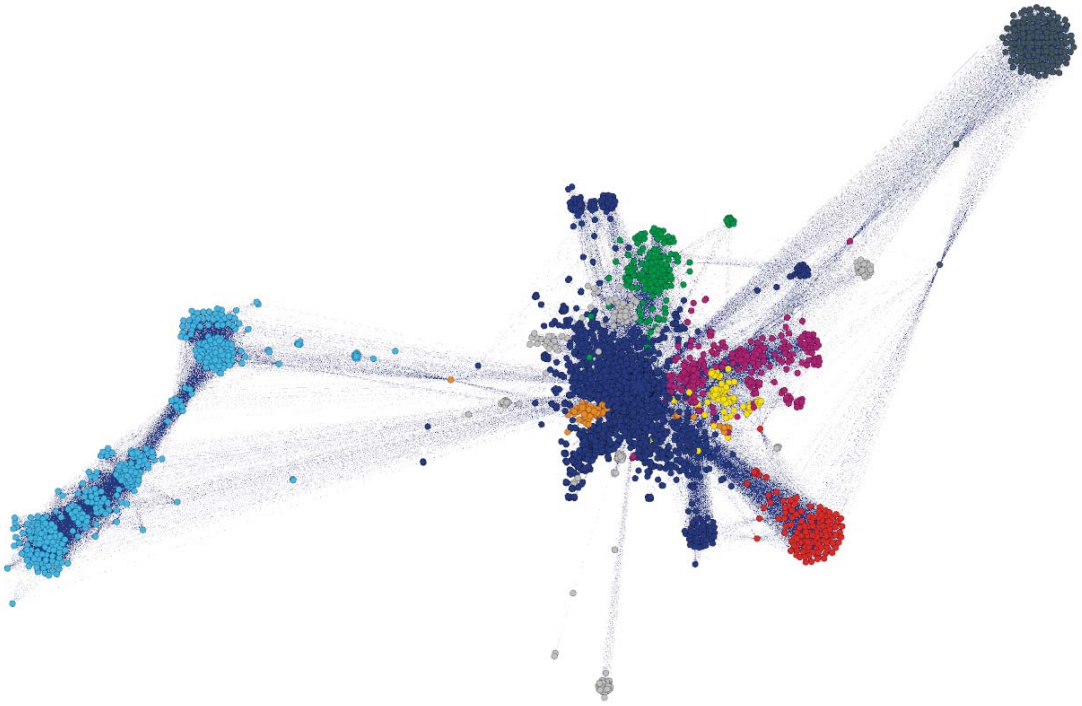

Figure 6.5: Visualisation of email network N1. Communities are presented using a colour code, node's size shows degree centrality

centrality [48]. The next highest correlation was between Eigenvector centrality and degree (r=0.86) followed by Betweenness centrality and Degree (r=0.65).

This results are consistent with the analysis reported in [215], where the author studied the correlation between centrality measures for 62 sociometric networks in different settings (e.g., professional and personal network of physicians, network of members of voluntary organisations, and information exchange in an IT department).

Figure 6.6 shows a scatter plot matrix with five centrality measures for the Individual 1 network. The diagonal includes a kernel density estimate (KDE) plot. Table 6.2 shows the standard correlation coefficient between centralities.

In [215], the author suggested that if measures are not highly correlated, they indicate distinctive measures likely to be associated with different outcomes.

6.3.4 Analysis of email networks N2, N3 and N4

Three additional email networks (N2, N3, and N4) were analysed. Table 6.3 summarises statistics for edges, nodes and networks.

Table 6.2: Correlations between centrality measures of N1.

	Degree	Closeness centrality	Harmonic closeness centrality	Betweenness centrality
Closeness centrality	0.494279			
Harmonic closeness centrality	0.624362	0.977837		
Betweenness centrality	0.659469	0.411961	0.390434	
Eigenvector centrality	0.863210	0.407801	0.567065	0.321111

Table 6.3: Statistics of networks, nodes and edges for N2, N3, and N4.

	N2	N3	N4
Network overview			
Avg. degree	143.465	17.426	11.124
Avg. weighted degree	298.172	62.688	41.911
Modularity	0.68	0.431	0.425
Node overview			
Avg. clustering coefficient	0.921	0.914	0.884
Edge overview			
Avg. path length	1.937	2.055	2.453

Degree represents the number of edges to other nodes in the network. Not surprisingly, the network with highest value of average degree (and also average weighted degree) is N2 (143), the network with the highest density (see Table 6.1).

Modularity is a measure of the structure of the networks [192]. Networks with high modularity have dense connections between the nodes within clusters but sparse connections between nodes in different clusters. N2 is the network with highest modularity (0.68), following by N3 (0.431).

Clustering coefficient is the probability that two nodes that are connected to another, are also connected together. The lowest clustering coefficient was for the network N4 (0.884). A higher clustering coefficient means a dense network.

Path length of a pair of nodes is the distance between them in the network. Average path length is the average of these distances between all pairs of nodes. N2 is the network with lowest average path length (1.937)

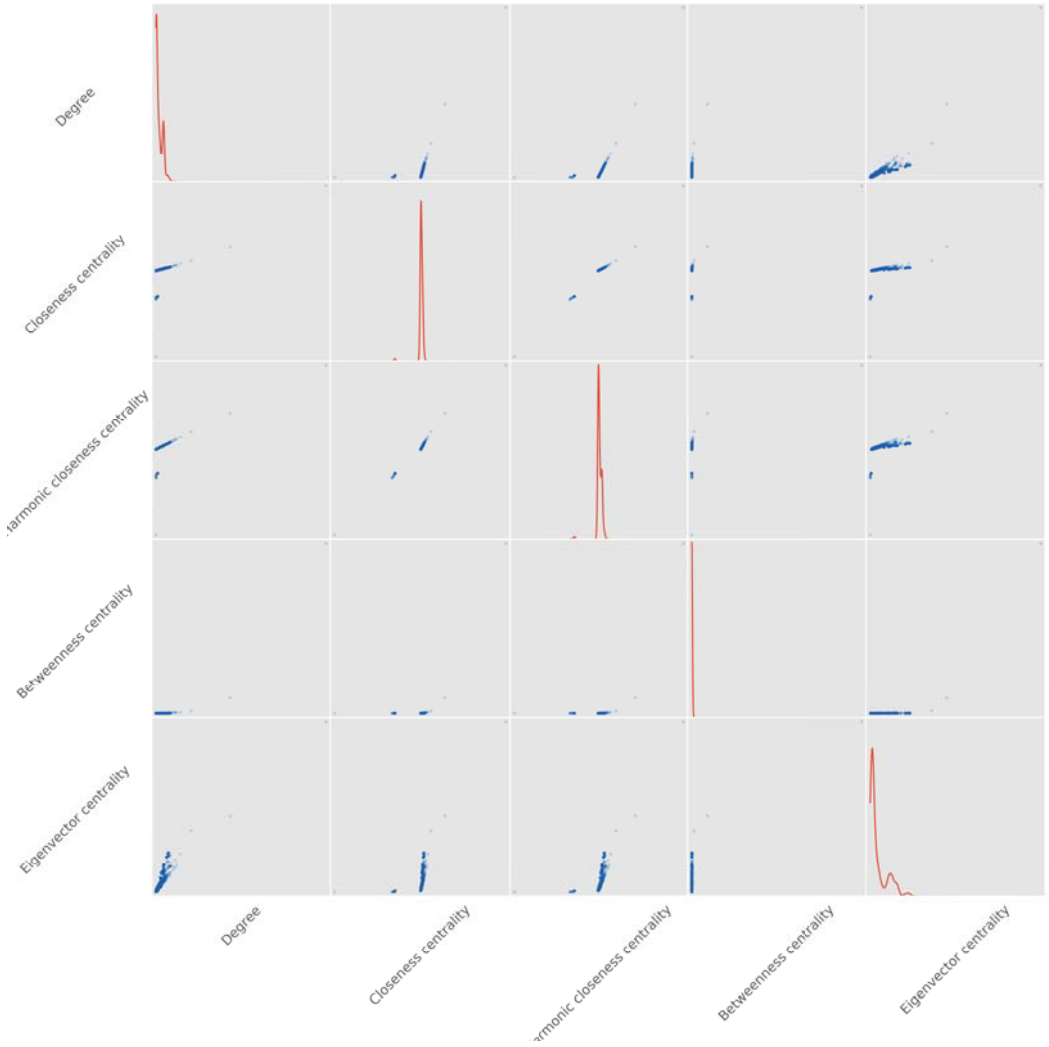

Figure 6.6: Visualisation of centralities of each node of email network for network N1.

6.3.5 Visualisation of email networks N2, N3 and N4

Figure 6.7 shows the structural differences of the four egocentric networks included in the analysis, in all the cases, the algorithm ForceAtlas 2 was used to identify communities. Figure 6.8 shows the same networks after the application of colours to represent the different clusters and node's size to represents the node's degree centrality.

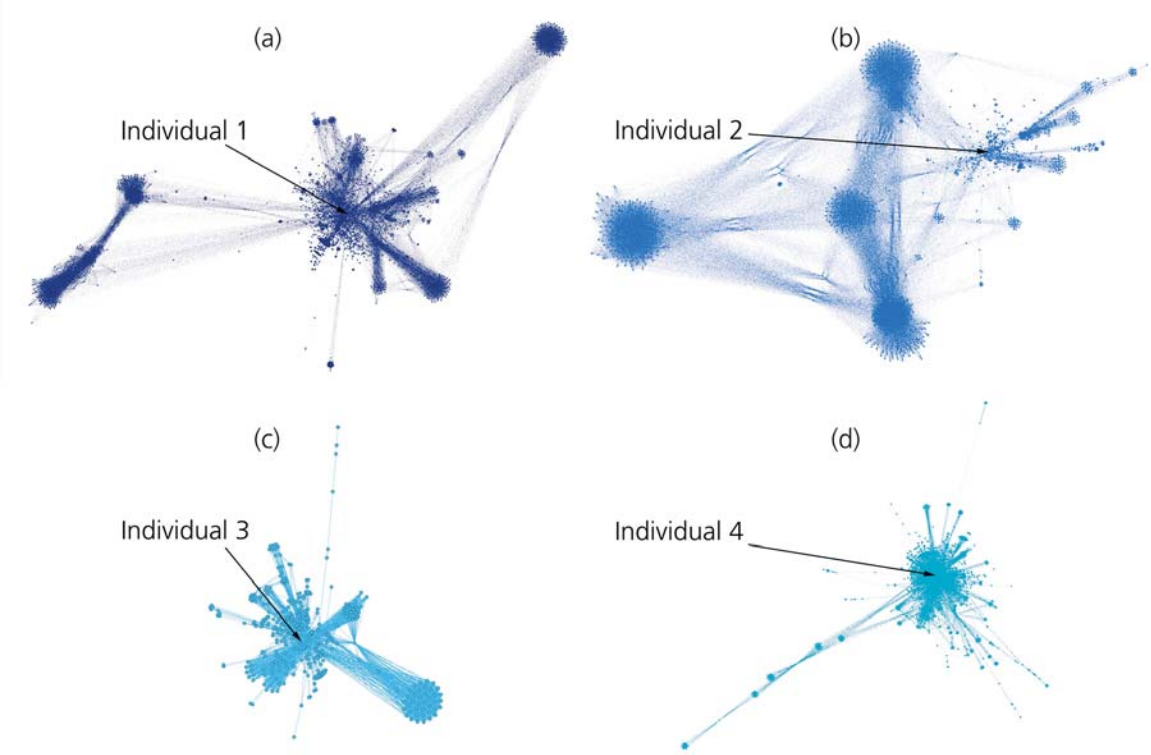

Figure 6.7: Visualisation of the four egocentric networks (N1 - N4).

6.3.6 Analysis and visualisation of email network N5

As described above, network N5 is the network developed combining the four individual egocentric email networks into one. The same analysis was performed for this network. Figure 6.9 shows the visualisation of N5, with the position of each individual within the network. Colour represents clusters of contacts. The new network is formed for 7,225 nodes and 236,247 edges. Table 6.4 summarises the statistics estimated at the network and nodes level, and compares the results with the individual-centred networks (N1-N4).

Figure 6.10 shows an example of one of the analysis performed to identify individuals acting as 'bridges' within the network. The network was filtered by betweenness centrality, removing nodes with betweenness centrality under 0.006, a value selected empirically.

Table 6.4 summarises the statistics calculated for the five email networks (N1 - N5). Table includes statistics for networks, nodes and edges.

As defined in Section 2.4.2.4, the average path length give us a way of measuring the performance of a network. From the Table 6.4 we can observe that N2 is the

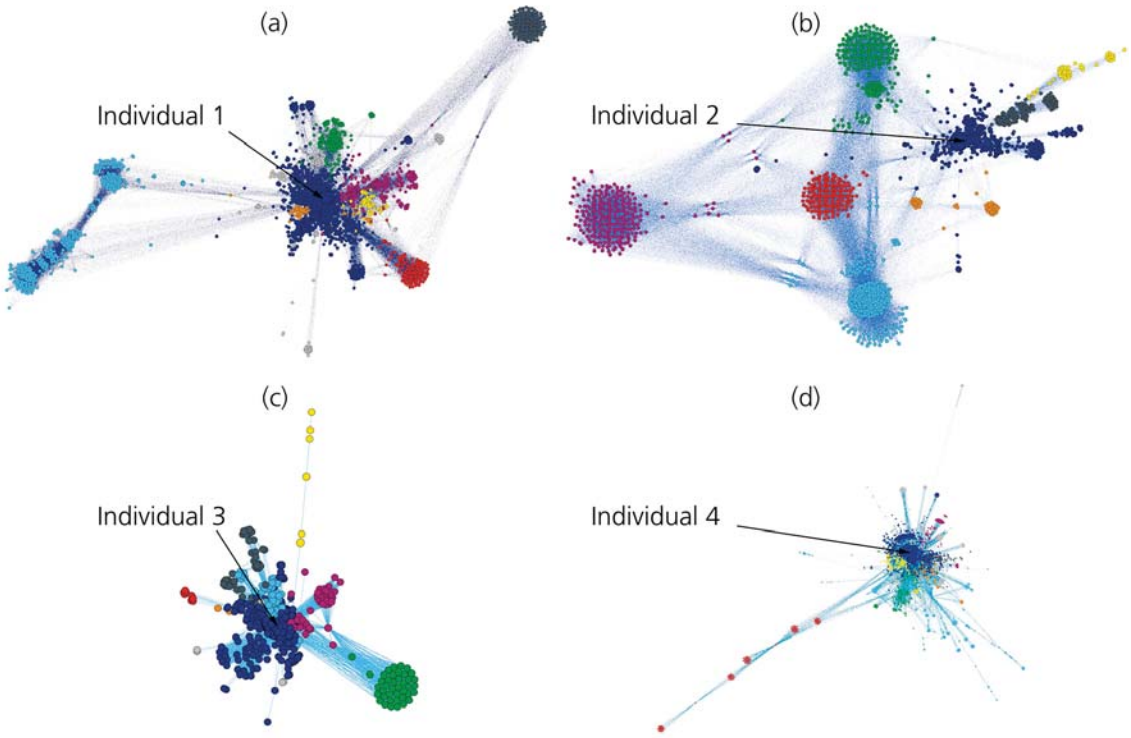

Figure 6.8: Visualisation of the four egocentric networks included in the analysis. Colour represents clusters and node's size represents node's degree centrality.

network with the highest performance (the bigger is average path length the smaller is its performance). This metric is also related with the concept of network's vulnerability (i.e., lack os resistance of the network to the deletion of nodes and edges). The bigger is average path length the bigger network's structural vulnerability.

The average clustering coefficient is high for all networks ($0 < C(G) \leq 1$). N2 is the network with higher concentration of neighbourhood of nodes, i.e., the network where two nodes with a common neighbour are most likely to connect each other (they form a triangle).

Degree centrality represents the number of edges connected to a node. Betweenness centrality measures how often a node appears on shortest paths between nodes in the network [26]. Closeness centrality considers the important nodes to be those that are relatively close to all other nodes in the network, and is calculated as the average distance from a given starting node to all other nodes in the network. Finally, eigenvector centrality is a measure of a node's importance that considers the importance of the node's neighbours, where importance is estimated as a weighted

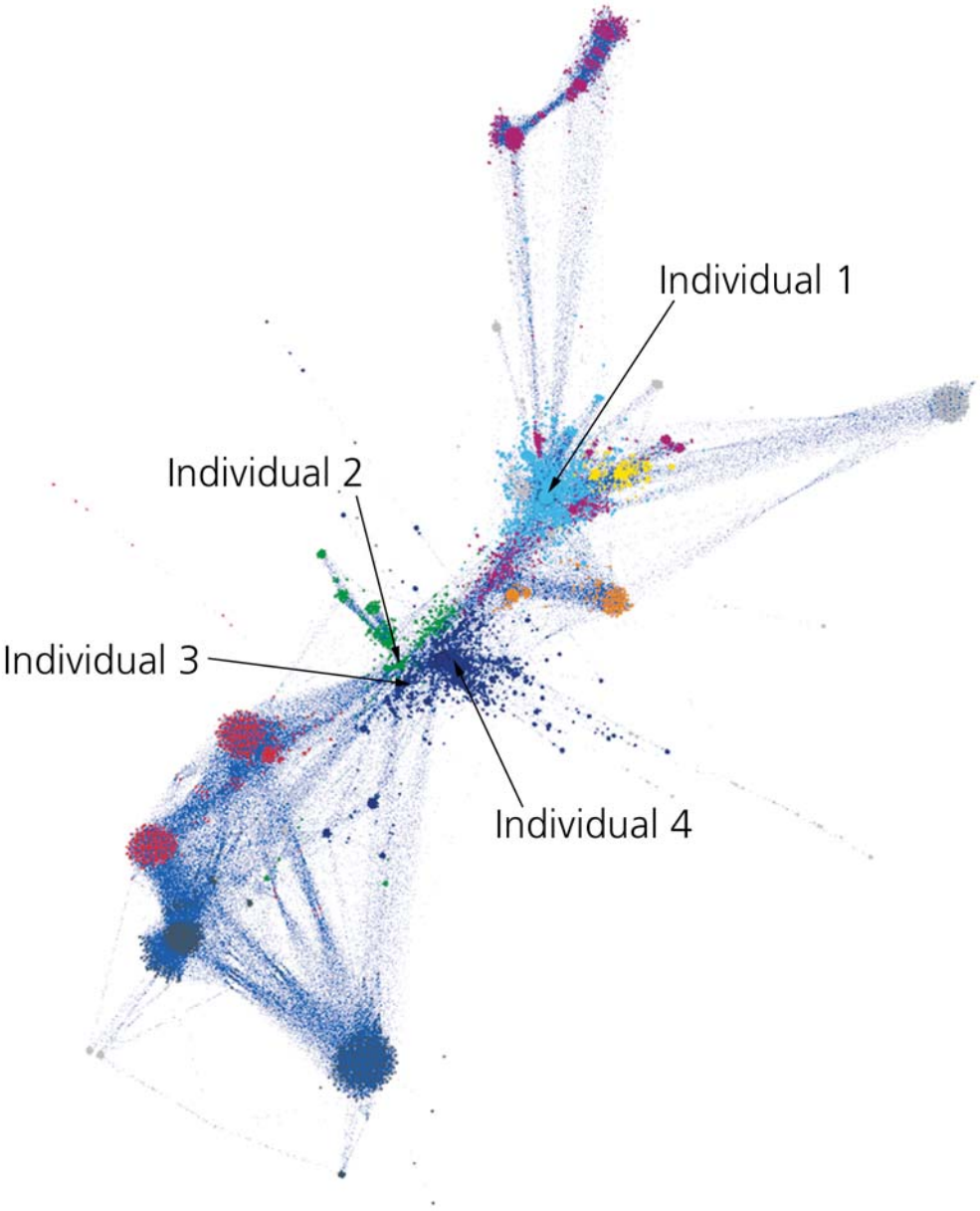

Figure 6.9: Visualisation of the network N5, formed for the four ego-networks included in the analysis. Colour represents clusters and node size represent degree centrality.

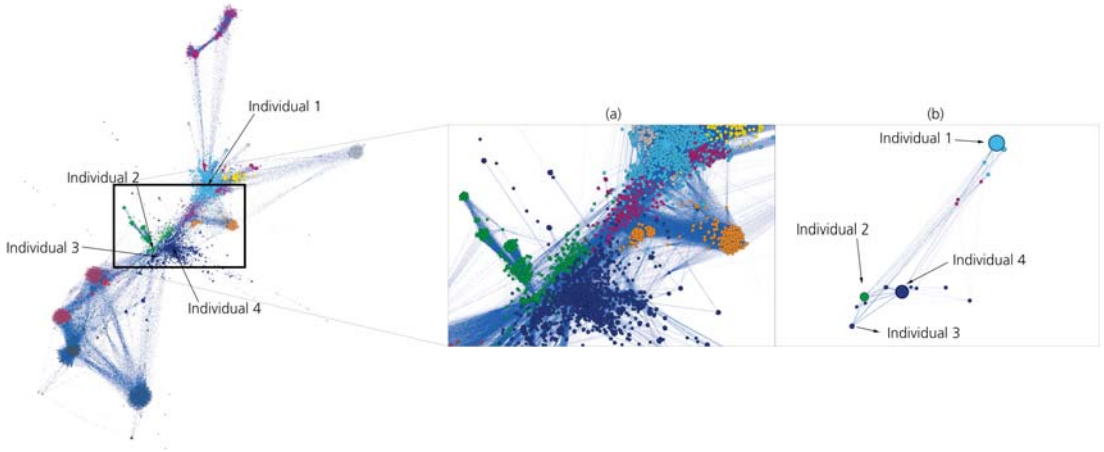

Figure 6.10: Visualisation of the network N5: (a) Zoom visualisation of part of the email network. (b) Filtering for node's betweenness centrality (> 0.006).

Table 6.4: Statistics of networks, nodes and edges for datasets used in experimentation. Final column represents statistics for the network formed for the four egocentric networks (Individuals 1 to 4).

	N1	N2	N3	N4	N5
Nodes	3,239	1,853	401	2,409	7,225
Edges	91,877	132,920	3,494	13,399	236,247
Network overview					
Avg. degree	56.732	143.465	17.426	11.124	65.397
Avg. weighted degree	186.422	298.172	62.688	41.911	177.5
Network diameter	4	5	7	7	9
Network density	0.018	0.077	0.044	0.005	0.009
Modularity	0.668	0.68	0.431	0.425	0.724
Node overview					
Avg. clustering coefficient	0.881	0.921	0.914	0.884	0.888
Edge overview					
Avg. path length	2.003	1.937	2.055	2.453	2.676

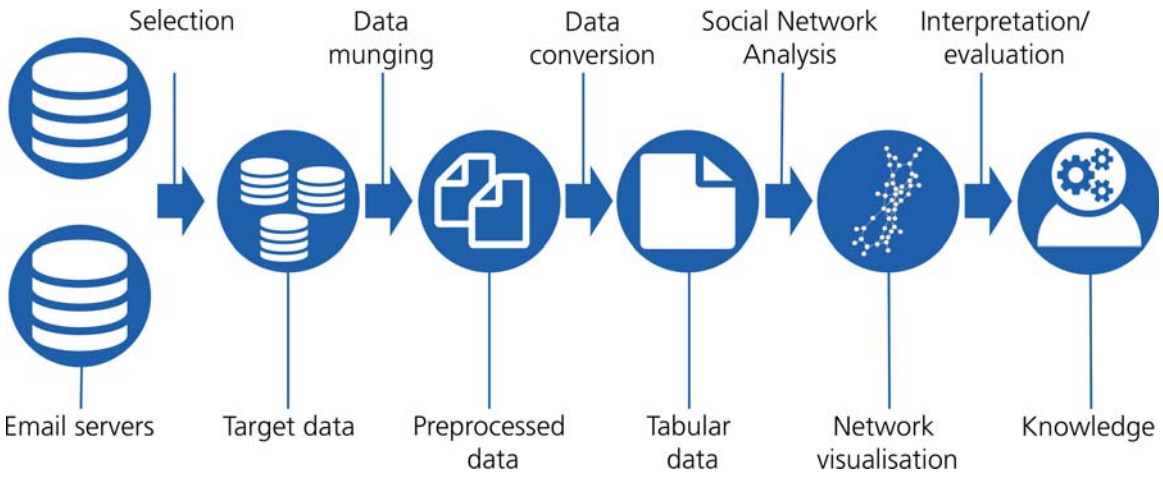

Figure 6.11: A proposed Email Social Network Analysis - Knowledge Discovery in Database process.

sum of direct connections and indirect connections or every length.

Betweenness centrality measures how disrupted the flow through a network would be if a person was removed, and helps to identify bridges spanners [131]. Degree centrality is a measure of popularity, and in our analysis, represents the number of messages a node sent and receives.

6.4 A proposed Email Social Network Analysis - Knowledge Discovery in Database process

Based on the review of SNA concepts in the literature, and the empirical analysis presented in the previous section, we developed a new framework, defined as Email Social Network Analysis - Knowledge Discovery in Database (ESNA-KDD), to integrate email mining theory and SNA. Figure 6.11 presents the different stages of ESNA-KDD.

The theoretical framework is based on the Knowledge Discovery in Database (KDD) process, a broad process of finding knowledge in data, popular in Data Mining [65]. The framework extends from understanding and extracting the data within email servers to the interpretation, evaluation and use of the results of the email SNA. In ESNA-KDD, SNA is defined as a stage within the ESNA-KDD process. The overall process of finding and interpreting mining and SNA from email data involves the repeated application of the following steps:

- Developing an understanding of the application domain, the relevant prior knowledge and the goals of the email mining and SNA. In our analysis, the application domain was related with the Wessex AHSN stakeholders and organisations part of the contact networks, as well as the programmes where the four individuals analysed are involved and the specific roles.

- Data selection and extraction from the email servers: selecting email datasets and focusing on a subset of variables or data samples, on which the analysis is to be performed. The outcome of this step is a subset of emails from the original email databases.

- Data munging in order to manually transform of email data to a usable format for more convenient use of the data with the help of semi-automated tools. Python was used for combining together the different datasets, and to perform data munging tasks (e.g., subsetting and filtering data, aggregating data, merging data, reshaping data, and data rename). Emails part of chain-emails or sent to a distribution list (e.g. all staff, etc.) were deleted in this step. Different email addresses belonging to the same individual were merged as part of the data munging process. The outcome of the task is a pre-processed dataset.

- Data structure conversion to prepare the data for the SNA. After this step, email information is stored in tabular format. In our study, this task was performed using Python, and the dataset was stored as a .csv file.

- Choosing the email SNA task, deciding whether the goal of the email SNA is the analysis of a personal (egocentric) network or a complete (whole) email network.

- Selecting SNA algorithms, SNA method, metrics and tools to perform the analysis.

- Email Social Network visualisation and analysis, searching for network structures, clusters, and key individuals within the network.

- Interpreting and evaluating results of the email SNA.

- Consolidating discovered knowledge extracted from the email SNA.

6.5 Conclusions

We identified several practical and theoretical implications in our study and particularly from the case study about mining email networks using Social Network Analysis (SNA). Specifically, practical contributions include the impact of SNA as a tool to identify 'important' members of a network and clusters or communities of stakeholders. Another contribution is the discussion about how rich information is hidden within email communications and to how to extract, analyse and visualise it. A new framework for email mining was proposed, combining theories from SNA and data mining.

Theoretical implications include the review of the core network concepts and future directions for an area of network research in knowledge extraction from email networks. SNA studies tend to focusses on a personal (egocentric) network, or on a complete (whole) network. We presented a methodology to develop and analyse a network formed by the combination of different egocentric networks. We also showed that impact of the centrality metric selection in the finding of 'important' members of the network, and demonstrated a low correlation between centrality metrics, meaning that they indicate distinctive measures likely to be associated with different outcomes.

The study was conducted at the Wessex Academic Health Science Network (Wessex AHSN), one of the 15 AHSNs across England, established by the NHS in 2013 to spread innovation, improve health for patients, and generate regional economic growth across the healthcare sector. AHSN are the only bodies that connect NHS and academic organisations, local authorities, the third sector and industry [152].

SNA was applied to identify the properties and structure of the communication network of one of the Wessex AHSN programmes, through the investigation of the project manager's email network, and we demonstrated that SNA can be applied to emails without violating important ethical issues (e.g., non required access to email content). Mining email networks was also used to identify specific individuals that could help to spread innovation and information to the whole network. Properties and structure of the communication network formed by combining different egocentric networks were also investigated.

Collecting email network data has several challenges in terms of reliability, validity, and ethical issues, especially for a whole-network research. Specifically, potential issues include omission of nodes/edges (due to the defined boundary of the network) or retrospective errors or simply because emails deletion from the servers or because different servers hosting emails using different data structure. Whole-email-network

studies require the participants to provide access to their email manager in order to extract specific variables and to download email datasets. If the privacy is not an issue, several options are available to automatise the data collection process.

Future research can consider the analysis of evolution email networks in time, using for example Dynamic Network Analysis. Other potential future research is a meta-network (i.e., multi-mode, multi-link, multi-level networks), through the integration of different data sources as, for example: online social networks; emails; and network surveys.

This chapter demonstrated that the combined application of SNA and Data Mining is a powerful tool to extract knowledge from data sources already available at the organisation, without extra costs as only open source tools were used for this analysis. This methodology can be used to identify key contacts within a stakeholders networks, but also can be used to identify potential risks of network disruptions or dependency on specific members.

6.6 Chapter Summary

Section 6.1 introduced the Social Network Analysis (SNA) and the research questions. Section 6.2 discussed the fundamentals of SNA, presented the methodology applied to this study, and described the five email networks analysed. Section 6.3 summarised the results of the networks analysis and visualisation. Section 6.4 introduced a new process to develop email mining, combining SNA and Data mining theories. Finally, Section 6.5 presented conclusions and future research.