Using Social Network Analysis to Study Business Partnerships

دراسة الشراكات التجارية بين رواد الأعمال باستخدام آليات تحليل الشبكات الإجتماعية

by

ROLA RASHAD ARIF ABDUL KARIM

Dissertation submitted in fulfilment

of the requirements for the degree of

MSc INFORMATICS
(KNOWLEDGE AND DATA MANAGEMENT)
at

The British University in Dubai

November 2018

DECLARATION

I warrant that the content of this research is the direct result of my own work and that any use made in it of published or unpublished copyright material falls within the limits permitted by international copyright conventions.

I understand that a copy of my research will be deposited in the University Library for permanent retention.

I hereby agree that the material mentioned above for which I am author and copyright holder may be copied and distributed by The British University in Dubai for the purposes of research, private study or education and that The British University in Dubai may recover from purchasers the costs incurred in such copying and distribution, where appropriate.

I understand that The British University in Dubai may make a digital copy available in the institutional repository.

I understand that I may apply to the University to retain the right to withhold or to restrict access to my thesis for a period which shall not normally exceed four calendar years from the congregation at which the degree is conferred, the length of the period to be specified in the application, together with the precise reasons for making that application.

_____

Signature of the student

COPYRIGHT AND INFORMATION TO USERS

Abstract

Social network analysis is gaining increased interest due to the expansion of social media and networking websites. Network analysis also significantly contributes to the biomedical sector in analysing protein networks and interactions. This paper proposes a new domain for network analysis of analysing business partnerships as an interesting economic aspect. Datasets of business partners featuring trade licenses created in 2015, 2016, and 2017 are transformed into graph datasets with nodes representing business partners, links representing a relationship between two partners, and a link's weight representing the number of trade licenses shared between the two connected partners. The resulting weighted undirected network is analysed using community detection algorithms. Characteristics of the top seven communities discovered from the 2015 data are discussed for which common social network motifs are captured. The behaviour of the seven discovered clusters are also analysed over the subsequent two years for deeper insights into business partnerships behaviours.

الخلاصة

تشهد تقنيات تحليل الشبكات الإجتماعية إهتماماً واسعاً خلال السنوات الأخيرة نظراً لانتشار وسائل و مواقع شبكات التواصل الاجتماعي. كما أن استخدام تقنيات تحليل الشبكات قد ساهم بشكل ملحوظ في مجال الطب الحيوي لاسيما تحليل شبكات البروتينات وطرق تفاعلها مع بعضها البعض. يقدم هذا البحث استغلالاً فريداً لآليات تحليل شبكات التواصل الإجتماعي في أحد أهم المحاور في المجال الإقتصادي وهو الشراكات التجارية بين رواد الأعمال. وذلك من خلال تحويل قواعد البيانات الخاصة بالرخص التجارية التي تم إصدارها في السنوات 2015 و 2016 و2017 إلى صيغة ملائمة لإنشاء شبكة تواصل. تم إنشاء الشبكة الإجتماعية من خلال تمثيل رواد الأعمال كنقاط إتصال ترتبط ببعضها البعض وفقاً للشراكات التجارية المُقامة بين كل شخصين. كما تم تحديد وزن لكل رابط بين رواد الأعمال يُمثل عدد الرخص التجارية المُفعلة بين رواد الأعمال. بعد أن تمّ إنشاء شبكة الشراكات التجارية ، تم استخدام خوارزميات تحليل الشبكات الإجتماعية لاستخلاص العناقيد والمجموعات الأساسية لرواد الأعمال. حيث تم اكتشاف سبعة عناقيد رئيسية لبيانات العام 2015. وقد شملت الدراسة تحليل لأهم خصائص هذه العناقيد والتي أخذت في تكوينها أنماطاً شائعة بين المجموعات المكتشفة ضمن شبكات التواصل الاجتماعي. كما تم أيضاً ضمن هذه الدراسة تحليل سلوك العناقيد التي تم اكتشافها خلال العامين اللاحقين بهدف الاطلاع بشكلٍ أعمق على السلوك التجاري لرواد الأعمال.

Dedication

This dissertation is dedicated to my father, mother, sisters, and brothers. Thank you for the unlimited support throughout this journey. Thank you for your prayers and for encouraging me to pursue my Master's degree.

Additional thanks to all my friends and colleagues who were so supportive and inspirational.

I would also like to thank my dissertation supervisor for the great support and valuable feedback throughout this process and for providing an outstanding mentoring experience.

Acknowledgements

In the name of Allah, the Most Gracious and the Most Merciful.

First and foremost, I acknowledge my deepest gratitude to Allah who granted all that was needed to complete this program and dissertation. This achievement could never have been possible without his blessings and guidance.

I would like to acknowledge Prof. Sherief Abdallah, my supervisor, and Prof. Khaled Shaalan, my academic advisor, for guiding and supporting me over the past two years. Thank you for your patience and motivation. I would also like to thank Dr Cornelius Ncube for his support and encouragement.

I want to express my appreciation to the staff of the Dubai Statistics Center and Department of Economic Development for their assistance and contribution in providing the required dataset for this dissertation. Special thanks to Mr Abdullah Al Hossani for his generous support.

Finally, I want to thank the members of the doctoral training centre, students' administration, library, and IT services at the British University in Dubai for their support and co-operation during the past two years.

Table of Contents

List of Figures

List of Tables

Chapter One: Introduction

Understanding peoples' behaviour is crucial to decision makers in many different sectors including politics, marketing, business, economics, social, education, and health. Analysing behaviour can significantly impact the process of updating and imposing new policies, regulations, and rules. This research sheds light on the significant economic aspect of business partnerships among entrepreneurs.

Kasseeah (2016), Naudé (2014), and Wennekers and Thurik (1999) agreed that entrepreneurship significantly influences economic development and growth. Audretsch, Keilbach and Lehmann (2010) claimed that studying firms' datasets of different German regions indicated a positive link between entrepreneurship and economic performance. However, Zaki and Rashid (2016) found a considerable negative relationship between new startups and economic growth in some emerging countries. Thus, analysis of business partnerships is essential at the government level to understand economic progress that affects decision making and strategic planning. At the individual level, analysing business partnerships has great value for business owners and those planning to establish new business partnerships. According to Ward (2018), up to 70% of business partnerships eventually fail. In 2013, the rate of failed business partnerships was close to 80% and claimed to be more complicated when involving more than two partners (Neville, 2013). Therefore, understanding the key elements of successful business partnerships is of interest within the field of business and economics.

This study provides a novel technique for exploring and analysing hidden patterns and characteristics of business partner communities. The approach is based on graph mining and network analysis techniques in which business partnerships data is represented as a graph reflecting the relations between partners. The business partners dataset was obtained from the Department of Economic Development (DED) in Dubai city. Having been recognized as the most cosmopolitan city in the world in 2015 with 83% foreign residents (Kapadia, 2016) and hosting residents from more than 200 nationalities according to Gulf News (23 January 2018), Dubai city with such significant diversity of population nationalities and backgrounds could considerably impact social, economic, and cultural aspects. Several studies revealed that ethnic diversity positively impacts economic growth (Alesina 2016; Bellini et al. 2013;

Kemeny 2017). The author considers all these aspects for developing the objectives of this study and establishing the research questions described in Sections 1.3 and 1.4.

1.1 Graph mining

A graph is a set of nodes connected by edges (or links) used to model relationships between the nodes (Mihalcea & Radev 2011). According to Chakrabarti and Faloutsos (2006), any many-to-many relationship of a relational database can be represented as a graph. The nodes can be any set of objects connected through a specified relationship, such as friendship, kinship, business relationship or protein interaction. Using the graph to represent and analyse such relationships is effective for large datasets (Quirin et al. 2010) as a graph can capture more information in the links between the nodes. Labels can be assigned to nodes and edges within the graph denoting attributes. For example, in social networks, nodes can be assigned with nationality labels and edges with the type of relationship connecting nodes, such as friendship or kinship. Moreover, the strength of the relationship between nodes is captured by assigning a weight to each edge. These details contribute to more accurate and informative data mining tasks implemented on the graph (Chakrabarti & Faloutsos 2006).

A graph may be directed or undirected, and it can also have self-loops (Inokuchi, Washio & Motoda 2003). Directed graphs reflect directional information in the relationship or link between two nodes (Nettleton 2013). However, undirected graphs provide no information about the direction of flow between nodes (Nettleton 2013). Self-loop in a graph is an edge with same node as its end nodes (Deo 2017). Fischer and Meinl (2004) defined graph mining as the search in the structure of all possible subgraphs discovered within a graph database. Frequent subgraph mining is a typical technique in graph mining (Takigawa & Mamitsuka 2013). Two steps are required to build a structure of frequent subgraphs. First, candidate generation uses small subgraphs to create a new, larger graph. Second, support computation specifies the frequency of newly created subgraphs (Fischer & Meinl 2004). Graph mining is vital for many applications, including anomaly detection, simulation studies, realisation of samples, and graph compression (Chakrabarti & Faloutsos 2006), and involves many tasks, measures, and statistics that provide deeper insight into the data. Details on these aspects of graph mining are discussed in the next chapter.

1.2 Social network analysis

The expansion of social media and the rapid development of social networks through the Internet contributed to increasing interests in social network analysis (Scott 2017). Moreover, sociologists have contributed to the graph mining field utilising Social Network Analysis (SNA) to analyse the structure of social groups and organisations (Chakrabarti & Faloutsos 2006). SNA is defined as the act of conducting sociological analysis augmented with methodologies and techniques to discover hidden patterns within social relationships connecting individuals and groups (De Nooy, Mrvar & Batagelj 2018; Scott 2017). Social relationships among individuals and groups are represented as a network of nodes and edges. While the terms network and graph may be used interchangeably (Mihalcea & Radev 2011), some researchers have identified characteristics to distinguish the two. According to Chakrabarti and Faloutsos (2006), the difference between a network and graph is in the size of each where graphs tend include hundreds of thousands of nodes and millions of links. Chakrabarti and Faloutsos (2006) claimed social networks tend to have fewer nodes, as in a few hundred. Another difference is the research problem attached to the graph mining and social network analysis. Chakrabarti and Faloutsos (2006) argued that SNA tends to utilise the social role of its nodes, unlike graph mining.

On the other hand, Mihalcea and Radev (2011) differentiated a network from a graph in that a network indicates a naturally generated relationship among the nodes while a graph captures an automatically generated relationship among the nodes. Furthermore, the authors claimed networks tend to have a more complex structure in comparison with some types of graphs (Mihalcea & Radev 2011). Yet, Johnson (2013) proposed there exists no precise rule to differentiate between network and graph, and suggested the term network can be used in cases where the links between nodes represent the transferring of an object between connected nodes, such as sending messages, while graph can be used whenever edges between nodes simply represent a link between nodes, such as the case of an interest graph .

De Nooy, Mrvar and Batagelj (2018) described four techniques for network analysis of "definition of a network, network manipulation, determination of structural features, and visual inspection." The definition of a network involves building a graph of vertices and edges

by preprocessing the dataset to include lists of vertices and edges representing the links between these vertices. After building the network, further manipulation is applied, such as reducing the network size for proper analysis by selecting a representative subnetwork. De Nooy, Mrvar and Batagelj (2018) stated that visualisation works better for small to medium-sized networks of hundreds of vertices instead of large networks with thousands of vertices. The authors also suggested focusing on analysing one relation when the graph dataset includes multiple relationships between the vertices. The determination of structural features for the entire network, subnetworks, or vertices can be achieved using a calculation of structural indices, such as vertex centrality, which is claimed to be more accurate than visual aspects. The fourth aspect of network analysis is visualisation, which facilitates recognition of network concepts. However, the authors stressed that optimal network layouts must be implemented to highlight the structural features of interest. Overall, they concluded the result of SNA depends on the type of network being analysed.

Chakrabarti and Faloutsos (2006), on the other hand, pointed out that the classic structure of a social network is a clique in which a group of vertices are connected within one subgraph. The authors also highlighted that a core-periphery structure is common for a social network, which is a subgraph with cohesive core vertices linked to other sparse peripheral vertices within the same subgraph. In terms of graph statistics and measurements, the authors argued that vertices' centrality is an essential aspect in SNA and can be captured using a set of measures, such as node degree and betweenness centrality values.

In view of all that has been mentioned so far, the researcher suggests that business partnership; as a natural relationship among partners; can be visualized as a network and analysed using SNA techniques. This research objectives and questions it aims to answer are expressed in the following section.

1.3 Research goals and objectives

This research contributes to the economic sector by exploring and analysing business partners' networks utilising the power of graph mining techniques. The goal of this novel study is to provide decision makers with a deeper insight into the relationships and behaviours that govern business partners within a community.

1.4 Research questions

The questions this research answers are the following:

1: How can we visualise business partnership data as a social network?

2: What patterns can we discover within business partnership data using graph mining techniques?

3: Can common motifs be identified within a business partners network?

4: How diverse are the formed partnerships in terms of nationality and background?

1.5 Uniqueness of this research

Upon exploring and reviewing other contributions to the field of graph mining and SNA, this study offers unique research in applying SNA techniques to business partnerships data. The novel datasets utilised for this study are collected from the Department of Economic Development in Dubai city from 2015, 2016, and 2017.

Chapter Two: Literature Review

This section explores related work and implementations by other researchers to gain deeper insight into the field of SNA using various graph mining techniques and approaches. Also, contributions to the field of business partnership and ownership analysis are examined.

Keywords: Graph mining, Social network analysis, Graph-based clustering, Community detection, Anomaly detection in graphs, Business entrepreneurship analysis.

2.1 Entrepreneurship analysis using data mining techniques

The following reviews different implementations of business partnership and entrepreneurship analysis using data mining techniques. This investigation of applied approaches and techniques will be compared to the proposed approach of using graph mining techniques for analysing business partnerships.

The role of entrepreneurship in economic growth and development has gained interest in recent research as economists and policymakers argue that the level of entrepreneurship contributes to public success (Glaeser, Kerr & Kerr 2015). Similarly, in consideration to a significant impact of entrepreneurship, Hartmann et al. (2016) and Hochsztain, Tasistro, and Messina (2015) conducted two entrepreneurship analysis studies aiming to support decision making for successful start-ups. The focus of Hartmann et al. (2016) analysed the business models of data-driven start-up firms to identify commonalities between the models pursued by the firms. Hochsztain, Tasistro, and Messina (2015), on the other hand, focused on applying data mining classification techniques to anticipate a successful business project and discover key factors to entrepreneurship success or failure. Distinct methodologies were applied by each study where Hartmann et al. (2016) exploited unsupervised data mining techniques and Hochsztain et al. implemented a supervised technique.

The approach applied by Hartmann et al. (2016) involved the utilization of the *k*-medoids algorithm in implementing the data mining technique of clustering. Two steps were applied to identify attributes of interest. First, they identified six dimensions common among business models, including key resources possessed by the firm, data-related main activities, such as data preprocessing and transformation, value proposition, defined as the product or service the customer's value, targeted customers, revenue model, and cost structure. The second step

was to identify the features related to each dimension. For example, the main features related to a data-driven firm's resources were the data and data sources, and the main activities dimension's features included data generation, acquisition, processing, aggregation, analytics, visualisation, and dissemination. The second step resulted in 35 features covering the six business model dimensions.

The output of these two process steps is a data-driven business model (DDBM) framework (Hartmann et al. 2016). The authors prepared a sample of 100 randomly selected data-driven firms for the analysis where data on these firms' business models were collected from different resources, which were then coded against the developed DDBM framework resulting in binary feature vectors (Hartmann et al. 2016). With the prepared dataset, the analysis was applied through four stages of selection of clustering variables, selection of a clustering algorithm and similarity metrics, specifying the number of clusters, and results in validation and analysis. The authors selected nine related variables out of the 35 features captured by the DDBM framework for the clustering task. Seven clusters were next specified, and the *k*-medoids and the Euclidean distance measure algorithms were used to execute the clustering task. For verification, the clustering analysis was repeated using a different algorithm, the silhouette coefficient (Han, Pei & Kamber 2011), to verify the quality of the cluster, and case studies were applied by the authors to review clusters significance. The result of the clustering task was seven clusters of data-driven start-up firms, and Hartmann et al. (2016) neglected one of these clusters due to insufficient similarity among its firms. Each of the six clusters was analysed in terms of the seven cluster variables where four distinct business models were discovered for the data source variable and three patterns were identified in term of key activities.

As described, the Hartmann et al. (2016) study was centred around a business domain of data-driven companies, and they did not consider aspects related to business owners. For the approach adopted by Hochsztain, Tasistro and Messina (2015), classification was the data mining technique chosen for analysing business start-ups data. The three classification algorithms were applied are decision tree algorithm, Apriori algorithm, and Logistic regression algorithm (Hochsztain, Tasistro & Messina 2015). The analysis process began with the acquisition of entrepreneurship datasets where records of 63 entrepreneurs who participated in the CCEEmprende program for entrepreneurial development support between 2007 and 2010 were collected (Hochsztain, Tasistro & Messina 2015). The dataset includes a

set of features related to the entrepreneurs including age, gender, education level, employment status, reasons for establishing entrepreneurships, type of support (family, economic, moral), reason for participating in the CCEEmprende program, project funding, and success indicators specified as the creation of the enterprise and generated income (Hochsztain, Tasistro & Messina 2015).

For the implementation of the classification task, Hochsztain, Tasistro and Messina (2015) first applied the decision tree algorithm using SPSS to identify the path associated with enterprise success or failure. Next, Tanagra data mining software was used to generate the association rules. Two rules were related to enterprise success, called 'having funds' and 'independent labour situation,' and two for enterprise failure, called 'having no fund' and 'dependent labour situation' (Hochsztain, Tasistro & Messina 2015). The authors used support and confidence measures to observe association rules quality verifications. Finally, logistic regression including the Wald test[1] was performed highlighting the two significant variables of 'having fund' and 'entrepreneur pre-existing employment situation' (Hochsztain, Tasistro & Messina 2015). The authors concluded the two key factors related to entrepreneurship success are the entrepreneur's financing and previous employment status.

In summary, both reviewed studies analysed entrepreneurship from different perspectives using alternate tools and methodologies. Hartmann et al. (2016) focused on clustering data-driven start-ups relying on firms' attributes. While clustering is also the approach in this paper, we also utilise the data related to entrepreneurs and the partnerships regardless of the firms' economic activities. Hochsztain, Tasistro and Messina (2015), on the other hand, focused on entrepreneurs' attributes with an objective to predict entrepreneurship success or failure. Due to present limitations and unavailability of related variables, a prediction of enterprise status is out of scope for this research.

2.2 Graph mining approaches and techniques

In this section, several studies conducted using graph mining techniques are reviewed to gain a deeper insight into this analysis approach as graph mining is adopted by this research.

Nettleton (2013) analysed online social network data as a graph of nodes and links. The focus of the study was the implementation of link prediction and the identification of common sub-

[1] https://www.statisticshowto.datasciencecentral.com/wald-test/

graphs. The author claimed graphs of online social networks are distinguished from other graphs by the small world phenomenon, which is described as the need of only a few links to connect two nodes into a larger online social network graph (Nettleton 2013). The graph data processing techniques discussed by in this work included streaming, sampling, and searching. As the aim of data processing is to enhance memory usage and processing time of large graph data, the extraction of community structures was discussed using different algorithms and metrics. Nettleton highlighted the utilisation of two algorithms. The first was one proposed by Newman and Girvan (2004) in which most the metrics of central edges and least central edges are used to identify groups within social networks. The second algorithm was proposed by Blondel et al. (2008), which is an optimisation of the Newman and Girvan (2004) algorithm. The optimisation was achieved by saving the computation cost where nodes of the same community are aggregated to form a new network in each iteration.

Expanding on these techniques, the following reviews cover graph mining implementations for clustering and classification purposes. Wang, Lio, and Chen (2017) discussed the implementation of contextual graph mining to investigate and analyse hidden features affecting patterns of students in cross-college course enrollments as well as the significance of the extracted graph-based features for analysing behaviours. The authors claimed analysing course enrollment behaviour is a crucial aspect in the educational data mining field. However, few studies examined graph-based features for identifying the potential impact on students' enrollment behaviour (Wang, Lio & Chen 2017). The authors considered students and courses outside their college as inputs with the output set to be if the student enrols in the course. A random forest algorithm was used to build the course enrolment model with node2vec to explore the relationships between the graph nodes and neighbourhood nodes. Feature importance measurements were calculated for each feature of interest to evaluate the impact of each on the prediction model accuracy (Wang, Lio & Chen 2017). According to the authors' results, the graph-based features concerning the distance between student and course improved analysis accuracy in comparison with other features. They also identified 'student course preference' as the main feature affecting students' cross-college course enrollments, which claimed to be consistent with previous studies.

Moreira et al. (2017) exploited graph mining to identify key role players in the development of the refractory field and its supporting technologies. The method analysed publications from major journals covering the field of refractories over 21 years and represented features of

interest as a graph to leverage the tools from network analysis. The required network was implemented with a Python script to map the connections between countries for each publication based on their contributions. With this map, several important attributes were extracted from the graph including the accumulated impact factor, degree of collaboration, betweenness centrality, and modularity (Moreira et al. 2017). By applying their graph mining results, characteristics and factors affecting trends in the refractory field were highlighted. For example, the authors claimed they discovered the countries with the highest contribution to the refractory field over the 21-year period of the study. To identify the technological impact of each country, a cumulative impact factor was utilised instead of the number of publications released by each country. An additional result noted by authors was the degree of collaboration, which emphasised the significance of cooperation among research centres around the world. Another interesting result was the identification of the most commonly used raw materials over the time frame based on publication keywords (Moreira et al. 2017).

2.3 Communities detection

Community detection is a common task in graph mining and is the focus of this research. According to Fortunato (2010) "[c]ommunities are parts of the graph with a few ties with the rest of the system. To some extent, they can be considered as separate entities with their own autonomy." For a better understanding of the communities detection task, the following review includes papers discussing various aspects of community detection in networks.

Ríos and Videla–Cavieres (2014) discussed the implementation of data mining clustering techniques on a large retail dataset using an innovative approach. The authors represented the transactional dataset as a graph of nodes and edges to discover hidden communities of products. They claimed their proposed approach enhanced the results and computational costs of clustering such large datasets. An essential task was the construction of the graph in a way that accurately represents the products and their relationships. Two approaches to generate the graph of products were discussed including a Bipartite transaction products network, where products are linked to the associated transaction, and the co-purchased product network, where nodes are the products and edges link products purchased in the same transaction (Ríos & Videla–Cavieres, 2014). The authors adopted the latter approach, and after constructing the network, they next filtered the network to remove bogus non-frequent relationships. A threshold was calculated based on a proportion average of the highest three edge weight.

After filtering the network using 5% of the computed threshold, two overlapping community discovery algorithms were applied, COPRA and GANXiS. According to Ríos and Videla–Cavieres, both algorithms grouped products of the same label to one community on each iteration. A retail analyst then analyzed the result of this community discovery task for verification. The authors also analysed the discovered product communities to identify the best period for the communities' stability. To achieve this task, Ríos and Videla–Cavieres used the Jaccard Index to measure the similarity of discovered communities over different time windows, such as day, week, month, quarter, semester, and year. They claimed the best time window to maintain a communities' stability was the month, which is aligned with retail's time window in practice (Ríos & Videla–Cavieres, 2014).

Another application of graph-based clustering was discussed by Umamaheswari and Geetha (2014) with the implementation of an event mining task on Universal Networking Language (UNL) semantic subgraphs of sentence constituents. The input to the clustering algorithm was primarily the information extracted from 112,000 news documents after applying semantic interpretation using UNL. A UNL graph was then constructed with key concepts represented as nodes and edges corresponding to the relations between concepts (Umamaheswari & Geetha, 2014). The authors claimed their applied approach is distinguished by how the feature sets were selected for clustering and for the similarity measures applied between event-context graphs.

Four methods for clustering were applied by Umamaheswari and Geetha, including a properties-based approach, attribute similarity score between two event-context graphs, continuity-based context similarity, and time-, person-, and place-specific similarity. In the properties-based approach, the authors focused on the semantic constraints and degree of connectivity between concepts to build an event-context graph (Umamaheswari & Geetha, 2014). In the second method, the authors enhanced the similarity measurement between context graphs by utilising the similarity of event attribute values, such as time, location, and people involved. The third method adjusted the attribute-based similarity method by introducing a weight for connected events. The fourth method was instead concerned with clustering the events with respect to time, location, and people using an agglomerative algorithm to enhance the results indexing implemented by event-based search engines (Umamaheswari & Geetha, 2014). The results of clustering UNL subgraphs corresponding to the trained news documents were claimed to be considerably improved through all four

clustering methods. The authors attributed these improved results to increasing the number of considered features and the clusters' specificity (Umamaheswari & Geetha, 2014). Their proposed methodology of clustering was evaluated in terms of both events clustering and events mining effectiveness with evaluation measures of intra-cluster similarity, inter-cluster similarity, and a silhouette constant for measuring how properly the data were clustered. In comparison with other state-of-the-art approaches, Umamaheswari and Geetha suggested their proposed clustering methodology using an agglomerative algorithm achieves 30% improvement in precision while the recall level was reduced by 20%. For an improvement to their methodology, the authors suggested incorporating an increased number of events and relations as well as giving special considerations to temporal and naming convention aspects.

He et al. (2018) examined the detection of hidden communities in complex networks. According to the authors, for certain domains, identifying hidden communities can be of great value, such as identifying criminal groups in a social network. Criminal groups would appear as hidden or weak communities compared to families or location communities. The authors defined weak communities as those hidden community structures with a majority of their nodes also being a part of stronger communities within a network (He et al. 2018). This adopted approach is novel, and the authors proposed a meta-approach named Hidden Community Detection (HICODE). HICODE was implemented using various community detection algorithms as in its base method while running through iterations augmented with three communities of weakening methods, including Remove Edge, Reduce Edge, and Reduce Weight (He et al. 2018). Applying the weakening method helps identify hidden communities in a network. The Remove Edge method is concerned with removing intra-community edges while Reduce Edge randomly removes some inter-community edges. Reduce Weight, on the other hand, reduces inter-community edges' weights in weighted networks. The authors applied their proposed methodology on two real-world datasets. One was extracted from Facebook for seven different university networks and the second was taken from SNAP[2] of three networks with ground truth communities (He et al. 2018).

Described as a challenging task, He et al. claimed that identifying an adequate number of community layers resembling the ground truth communities improved the modularity of the detected layers. Jaccard-based metrics were then applied to verify the relativeness of the

[2] http://snap.stanford.edu

selected layers to the ground truth communities. Based on hiddenness values, ground truth communities were partitioned into subnetworks and, according to the authors, the hiddenness value determines the portion of nodes connected to stronger communities (He et al. 2018). For this work, the HICODE results outperformed those achieved by overlapping and disjoint detection methods with Reduce Weight achieving the highest detection accuracy followed by Reduce Edge and Remove Edge. Multiple evaluations were measured by applying Normalized Mutual Information (NMI), F1, and modularity. The authors attributed the HICODE improved results to the refinement stage of their methodology, which determined the appropriate number of hidden community layers.

2.4 Anomaly detection in graph mining

In social networks, anomaly detection is the identification of deviated users' behaviours by analysing patterns hidden in the network (Bindu & Thilagam, 2016). In their paper, Bindu and Thilagam reviewed and discussed various techniques and approaches for graph-based anomaly detection on social network datasets. According to the authors, anomaly detection in graph data features a set of distinguishing characteristics, including the nature of the networks' inputs that define if the network is static or dynamic and attributed or unattributed. Another characteristic of graph-based anomaly detection is the type of anomalies in terms of if it is related to an edge, node, subgraph or event (Bindu & Thilagam, 2016). The authors discussed the characteristics first needed to determine the appropriate anomaly detection technique for a network. For static unattributed networks, they suggested clustering or community-based, network structure-based, and signal processing-based techniques. They recommended clustering or community-based and network structure-based techniques for static attributed networks. With a dynamic unattributed network, the authors argued the appropriate three anomaly detection techniques include a matrix or tensor decomposition-based, community-based, and probability-based techniques. For dynamic attributed networks, they suggested tensor decomposition-based, probability-based, and signal processing-based techniques (Bindu & Thilagam, 2016). The authors described how anomaly detection in a network is differentiated from traditional anomaly detection in that it relies on the node interactions within the network. The two scenarios are similar in that they are highly domain-specific tasks, so it is difficult to decide the proper technique or algorithm to apply. They also highlighted challenges to anomaly detection that include computational cost, streaming

networks, maintaining a history of updates, and performance evaluation (Bindu & Thilagam, 2016).

Zhang, Kiranyaz, and Gabbouj (2017) discussed the detection of outlier edges in graph mining achieved through understanding and analysing the attributes of the edges corresponding to the graph structure around each edge. The methodology adopted relies on identifying the authenticity of the edges as a mechanism to identify unexpected outlier edges within a graph. Authenticity was calculated based on the difference between the number of actual existing edges between two neighbouring subgraphs and their expected number of edges. Accordingly, edges with a low authenticity score are assumed to be outliers. Then, random graph generation models were applied to determine the number of expected edges between two undirected and unweighted subgraphs. So, this outlier detection algorithm proposed by Zhang, Kiranyaz, and Gabbouj (2017) is simply based on calculating the authenticity of edges with the assumption that outlier edges are those with low authenticity rates compare to a given threshold.

To experiment, the authors utilised a real-world graph with previously identified outlier edges, in which the graph was randomly injected by false edges connecting two unlinked nodes (Zhang, Kiranyaz & Gabbouj 2017). The dataset included more than 58,000 nodes and 214,000 edges with 1,000 random edges injected using random algorithms. According to the authors, the proposed outlier detection algorithm effectively identified the outlier edges while providing good results compared with state-of-the-art algorithms by achieving a high score of 11% for the global clustering coefficient.

Zhou et al. (2017) utilised an anomaly detection technique to discover rare categories in time evolving datasets represented as graphs. They proposed a novel approach for rare class detection in graphs using the incremental algorithms, SIRD and BIRD, with which the anomaly detection model is updated instead of being recreated over data changes in real-time applications (Zhou et al. 2017). The authors claimed this proposed approach contributes to enhanced computational cost and performance.

The two incremental algorithms proposed by Zhou et al. (2017) were intended to capture edge changes over time. SIRD was applied for single edge update in a time scenario and BIRD was applied for edge batch update scenarios. Another algorithm was introduced by the authors, called BIRD-LI, to handle the case in which obtaining exact priors of rare classes is difficult

and relies on an estimated upper bound of the former state instead of the actual one to calculate the size of all rare classes at a given time (Zhou et al. 2017).

In addition to handling rare class detection in time-evolving graphs, Zhou et al. (2017) discussed the optimisation of queries required for labelling these classes. Query locating and distribution were discussed to improve the efficiency of finding the optimal time step to discover rare classes and also to find the appropriate number of queries required over different time steps. Using Matlab 2014a, Zhou et al. evaluated the performance of the SIRD, BIRD, and BIRD-LI algorithms using six time-evolving graph datasets and three real datasets to measure the query allocation and query distribution. The results of their proposed novel approach outperformed those achieved by state-of-the-art algorithms, and the authors considered this as the first attempt to use dynamic settings in rare class detection.

Chapter Three: Dataset Acquisition

A unique dataset is required for this study, and since the focus of this research is to explore and analyse business partnership networks in Dubai city, the Department of Economic Development (DED) was contacted to provide data. After explaining the objective of this research and the attributes and records of interest, the dataset was provided from recently created trade licenses over three years (2015, 2016, and 2017). According to the DED, the criteria applied to query the dataset involved selecting all records with an active license status and an active partner status within each license over the past three years. The dataset excludes records of inactive licenses and partners. Details on the dataset attributes and preprocessing tasks are described in the following sections.

3.1 Dataset description and characteristics

The key attribute to the business partners dataset for implementing this study effectively is the unique anonymous identifier assigned to each business partner labelled the "Person Serial No". The purpose of this attribute is to identify business partners across all the trade licenses in which they are registered. The "Person Serial No" value is also unique for each business partner across different license categories, economic activities, and years. Therefore, tracing the business activities of partners is possible using this as an anonymous identifier. Another important attribute is the "License Number" as it creates the link between partners.

Furthermore, demographic attributes are essential to explore patterns and extract valuable information thoroughly. However, a partner's nationality, sex, and date of birth are the only demographic attributes available. Table 1 presents the dataset attributes and their details.

The dataset includes 89,547 records for 2015, 89,260 for 2016, and 82,710 for 2017 all with the same attributes. Each record represents an economic activity for each license where one trade license can be attached to multiple economic activities.

Attribute Name	Data Type	Description
Issue Year	numeric	The year in which the license was issued.
Issue Month	numeric	The month in which the license was issued.
License Category	text	The name of the authority that issued the license.
License No	numeric	A unique number for each license.
Person Serial No	numeric	A unique identifier for each business partner.
Partner Birth Date	date	The date of birth for each business partner.
Sex	text	The business partners sex (Female/Male).
Partner Nationality	text	The name of the country.
Person Category	text	The category of the business partner (Person/Body Corporate).
Partner Legal Type	text	The partnership type (Manager/Owner/Partner).
Partner Share	numeric	The ownership percentage as a decimal for the amount each partner has in the corresponding business.
Legal Type	text	The legal type of the license (e.g., Limited Liability Company).
Actv Master Grp	text	The primary economic activity attached to each license (e.g., Commercial/ Professional).
Activity Category	text	A detailed description of the economic activity attached to the license.

Table 1. *The attributes of the business partners dataset utilised for analysis in this study.*

3.2 Data preprocessing and preparation

Since the source dataset is semi-structured data, it requires preprocessing and transformation to be ready for graph mining tasks. The details for the data preparation are described in the following:

3.2.1 Data preprocessing

- **Conversion into a structured format.** While the data was supplied in a tabular format, many cells needed to be merged. For example, the license column had to be merged across all business partners as well as with the year, month, and license

category columns. A re-organisation of the data table spreadsheet converted the dataset into a structured format where each row defines the business partner details across all columns. The transformed tabular data was imported into an MS SQL Server database to facilitate additional preprocessing tasks.

- **Filtration.** Due to limitations in resources and scope, a filter on the "Person Category" column to select only business partner records with "Person" was applied to reduce the dataset size to enhance the research focus. This ensures the analysis considers only human business partnership behaviour separate from corporate behaviour. Also, since this study aims to analyse partnership behaviour among business owners, licenses with single business owners were also excluded.
- **Missing Values.** Missing values were identified in the business partner date of birth and sex fields. Since the count of records with missing values was small, these records were simply eliminated.
- **Feature Extraction.** Additional fields were added to the dataset to enhance the analysis results, including:
    - **Age** of the business partner in relation to the license year of issuance.
    - **Longitude** of each country representing the business partner's nationality.
    - **Latitude** of each country representing the business partner's nationality.
- **Data Mapping.** The data for longitude and latitude fields were obtained from the Harvard WorldMap dataset[3]. A mapping between the partner's nationality field and corresponding country name from the Harvard WorldMap dataset provided accurate longitude and latitude values for each country listed in the DED datasets.

After applying these preprocessing tasks, the remaining records total 48,111 for 2015, 46,708 for 2016, and 43,751 for 2017. Table 2 shows a sample of records after the data preprocessing.

[3] https://worldmap.harvard.edu/data/geonode:country_centroids_az8

Issue Date (Year)	Age	License No	Person Serial No	Sex	Nationality	Longitude	Latitude
2017	39	574792	150254	Male	United Arab Emirates	54.30017	23.90528
2017	45	574792	207435	Male	United Kingdom	-2.865632	54.12387
2017	49	619801	645267	Male	Turkey	35.16895	39.0616
2017	58	763041	609888	Male	Syria	38.50788	35.02547
2017	35	763041	573147	Female	United Arab Emirates	54.30017	23.90528

Table 2. *Sample records of the business partners dataset after applying the data preprocessing filter tasks*.

3.2.2 Data preparation

The next step following data preprocessing is to prepare the data for the graph mining task. This process involves transforming the filtered dataset into an adequate format. To implement graph mining effectively, the business partners datasets must be transformed into a layout suitable to construct a graph. After transforming purchasing data into a customer-product matrix (Sarwar et al. 2000a, 2000b, 2001). Kim, Kim, and Chen (2012) proposed two approaches for constructing products network. The first approach builds a customer product network where the products and customers are nodes linked by sale transactions. The second approach builds a co-purchased product network where products are connected through weighted edges that reflect the number of customers who purchased the two products (Kim, Kim & Chen 2012).

Both approaches were considered for the business partners dataset by first defining a list of nodes and edges. Based on available fields and features, the two connections for building a business partners network included a License-Partner network and a Partners network. In the license-partner network model, trade licenses and business partners are the nodes with edges representing the trade license registration of each connected partner. The edge weights are assigned as the ownership percentage of each business partner registered with that license. Figure 1 illustrates this first approach. For the partners network model, business partners are the nodes, and the edges resemble the partnership between two business partners regarding a trade license. An edge weight is equal to the number of trade licenses the two business partners are registered in together. Figure 2 illustrates this second approach for translating the

business partners dataset as a graph. The graph type for both approaches is an undirected weighted graph as the relationship among business partners does not involve directional information or indications. Also, no self-loop exists in the graph because only licenses with two or more partners were selected to study the business partnerships behaviour via the preprocessing filters. Due to the enhanced representation of the relationship and connectivity between business partners, which is the focus of this study, the second approach of Figure 2 is selected for this analysis.

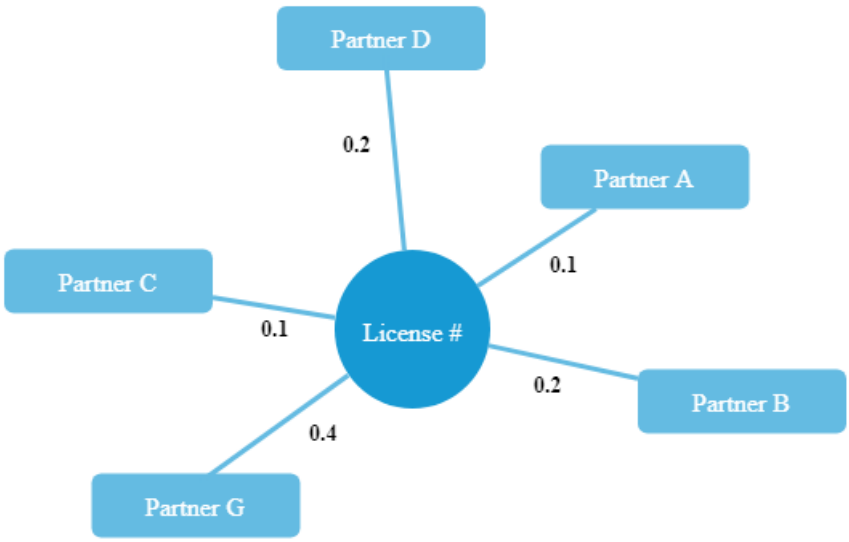

Figure 1. *License-partner network structure proposed for constructing business partnerships as a graph.*

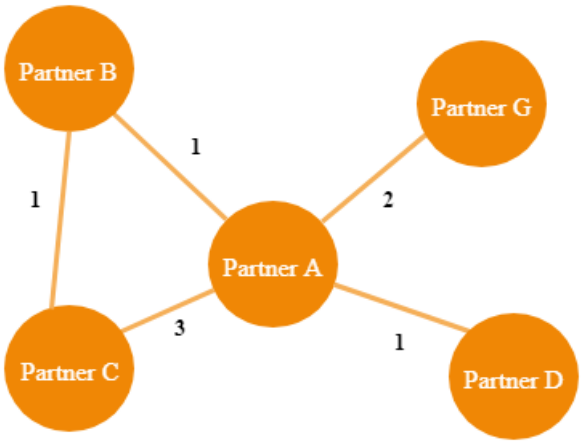

Figure 2. *Partners network structure proposed for constructing business partnerships as a graph.*

Having partners network structure selected, business partnerships dataset has been transformed into two lists: list of nodes and list of edges. List of nodes provides details of each business partner and list of edges provides records on source and target partners in addition to the weight attached to each edge. The generated lists support construction of business partnerships network using partners network structure shown in Figure 2. Next chapter discusses the implementation of network analysis on the generated network.

Chapter Four: Methodology and Results

Several graph mining techniques and methodologies presented by other researchers were investigated in Chapter 2 to provide insight into the approach for this research. Under this chapter, the researcher demonstrates the methodology applied to utilize SNA techniques to analyze business partnerships dataset as well as a discussion of the results.

4.1 Tools and techniques

The graph to be built will present business partners as nodes with edges that represent the partnership agreement. A weight is assigned to each edge equivalent to the number of licenses linking two partners. The graph is undirected with no self-loops. The tool selected for implementing the graph mining is Gephi version 0.9.2. Using MS SQL Server, a query is executed to generate a list of the edges and nodes, which are imported into Gephi to visualise the data as a graph and prepare it for network analysis.

4.2 Methodology

The methodology adopted here begins with exploring the dataset as a graph to identify opportunities for further filtering or tuning. The next step computes graph statistics and metrics to discover hidden clusters and sub-graphs. Based on these calculated metrics, force-based algorithms and a readability optimisation algorithm provided by Gephi are applied to detect existing communities. The resulting communities are from a clustering task that is then further investigated. Nodes and edges are re-formatted to enhance the network analysis process, and the details of this methodology are presented in the following.

4.2.1 Graph exploration

The first step in the network analysis is exploring the imported nodes and edges by investigating the initial state of the dataset. This process will assess how much tuning is required to bring the data into the most suitable state for applying graph mining algorithms and extracting frequent patterns. Table 3 shows the number of nodes and edges available for each year. Upon exploring the graph, a set of unconnected sub-graphs with low degree values are shown in Figure 3. Thus, the next filtration process is to eliminate

the nodes and edges of low interest. The following section provides details on how these nodes and edges are highlighted and excluded.

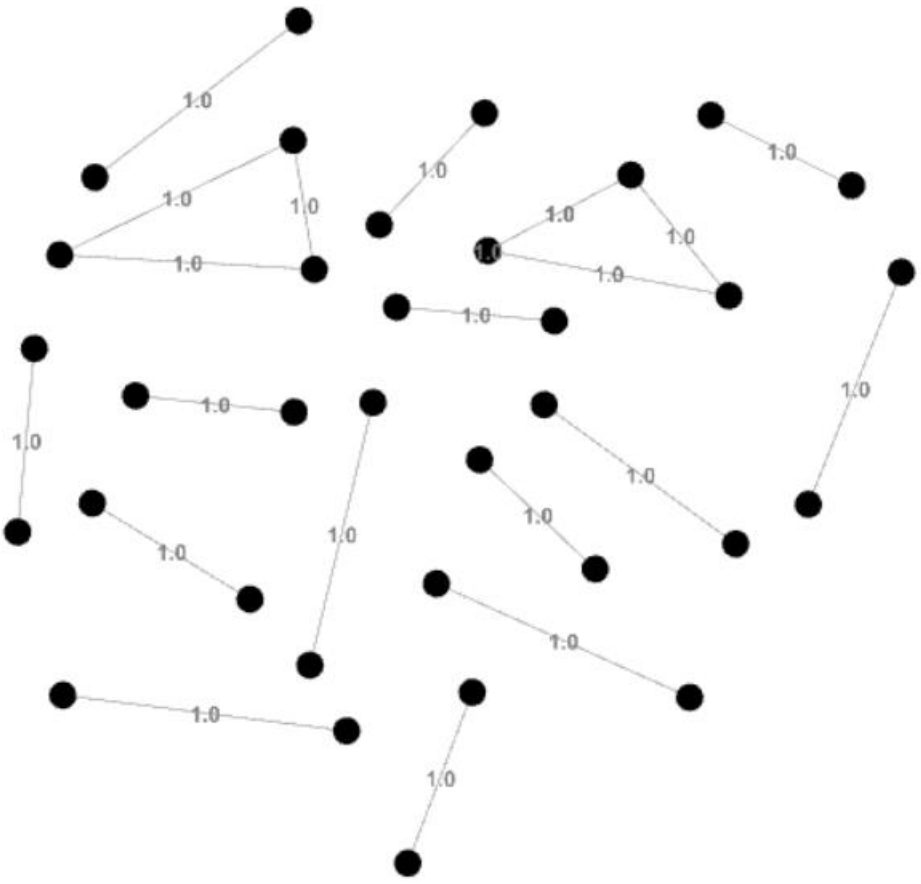

Figure 3. *The set of isolated low degree sub-graphs captured after initially constructing the business partnerships network.*

4.2.2 Network metrics and measures

Statistical metrics are calculated to offer insight into the business partnerships network characteristics. These metrics are exploited in graph filtration and community detection tasks as details in the following:

- **Average Degree.** With this measure, the degree distribution is generated for each year, which provides the degree value for each node in the graph that reflects the number of business partnerships for each node (i.e., partner). Figure 4 shows the degree distribution graph for 2015, 2016, and 2017 where the average degree of each year is listed in Table 3. Based on the average degree value, a filter is applied. However, since the edge weight is also a key measure to identify frequent partnership patterns, they are considered while filtering the graph nodes. Details on this filtration process are discussed in the following section.

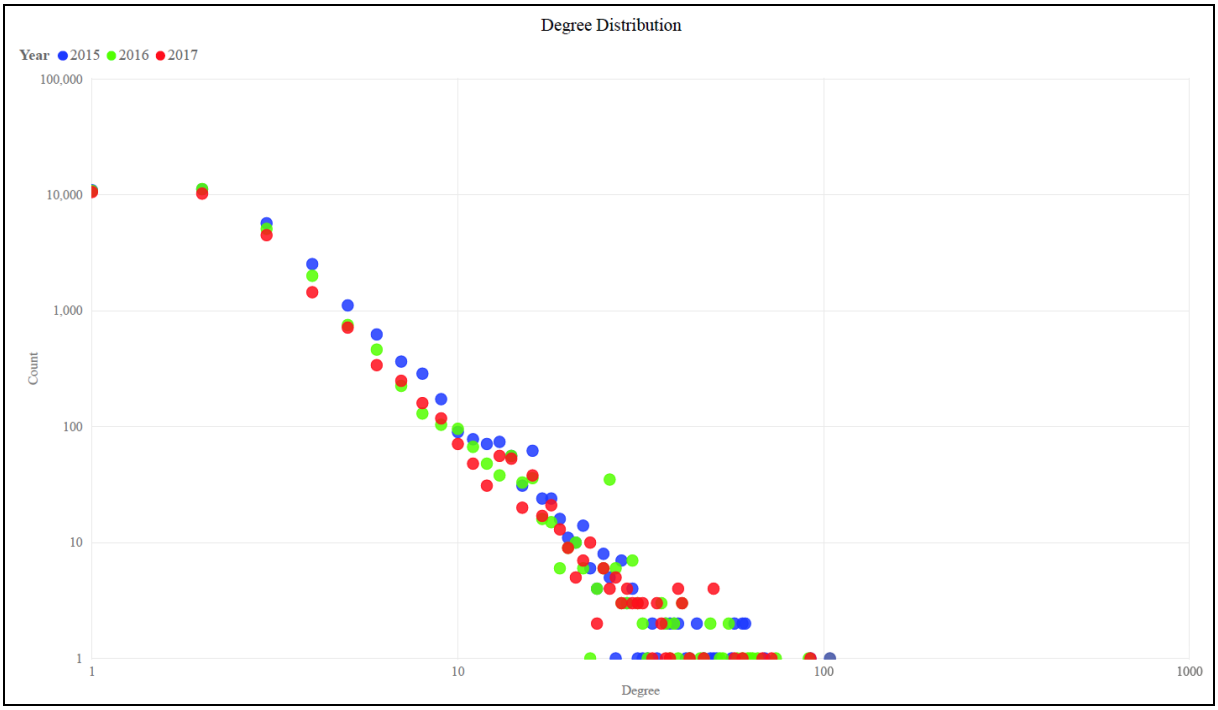

Figure 4. *The degree distribution graph of the business partnerships data for 2015, 2016, and 2017. The degree value represents the distinct number of partnerships maintained by each business partner, which appears to be consistent over the three years.*

- **Average Weighted Degree.** Since the graph is an undirected weighted graph, calculating the average weight of the available edges is important. This is especially this case since the weight represents the number of licenses two business partners own together. Thus, the value of this metric is important to identify partner business activity.
- **Average Edge Weight.** Using R, an average edge weight is calculated for the entire graph. So, this measurement value is utilised in the node filtration process as it shows the average edge weight assigned to graph edges.
- **Average Clustering Coefficient.** Watts and Strogatz (1998) defined the clustering coefficient as a measure of how strongly a node's neighbourhood is associated. The value of the clustering coefficient for a node is higher if more nodes within its neighbourhood are connected. This measure is calculated and graphed to explore the tendency of graph nodes to cluster together. Figure 5 is a line chart of the clustering coefficient distribution for each year, and the average clustering coefficient values for each year are included in Table 3. The average clustering

coefficient is nearly equal to 0.9 over the three years of data, which reflect the nodes' high tendency to group and form clusters.

Year	Nodes Count	Edges Count	Average Degree (AD)	Average Weighted Degree (AWD)	Average Edge Weight (AEW)	Average Clustering Coefficient
2015	33,650	43,333	2.576	2.921	1.134	0.905
2016	31,434	37,988	2.417	2.988	1.236	0.899
2017	28,886	33,937	2.35	2.681	1.141	0.894

Table 3. *Summary statistics of the graphed datasets prepared for 2015, 2016, and 2017. The nodes represent business partners, and edges represent a connection of at least one trade license shared between two business partners.*

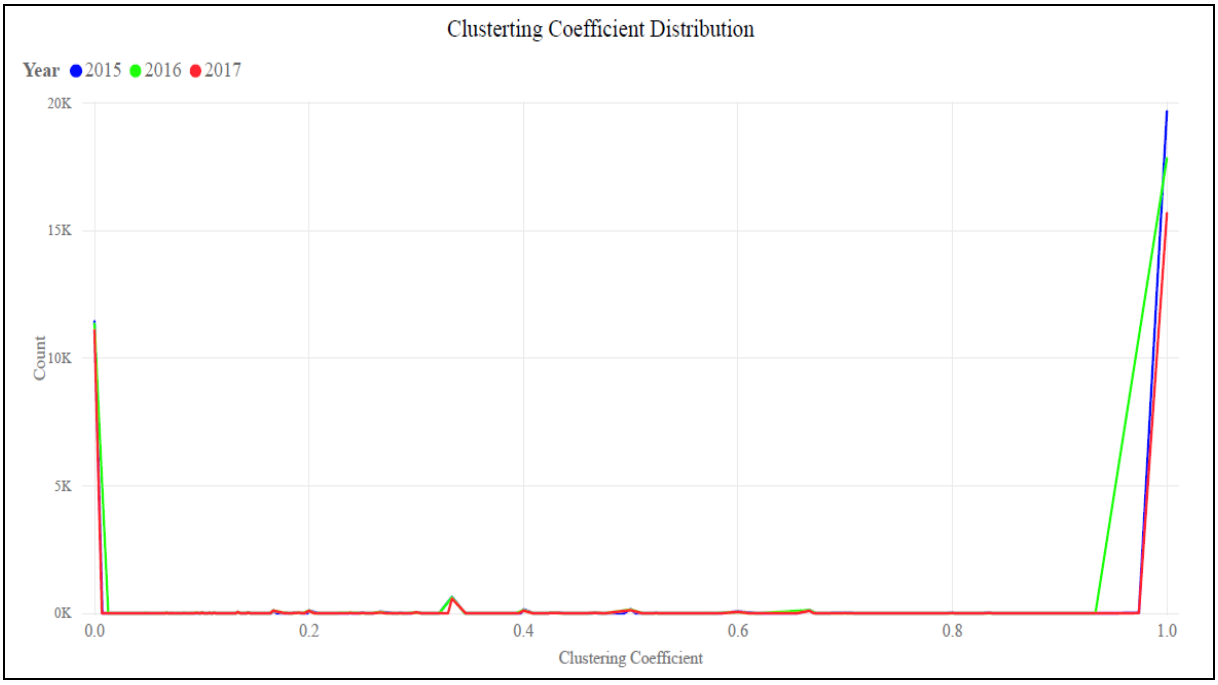

Figure 5. *A line chart of the clustering coefficient distribution for the business partners datasets for 2015, 2016, and 2017. The clustering coefficient value is proportional to the node's neighbourhood connectivity and is nearly the same for the three years, except for 2016 in which fewer nodes appear to have a clustering coefficient greater than 0.93 and less than 1. Also, most nodes have a clustering coefficient value of 1 with 2015 including the greatest number of nodes with a clustering coefficient equal to 1.*

- **Betweenness Centrality.** This measure plays an important role in identifying nodes with high density of connections. In Gephi, this measure describes how

often a node appears on the shortest paths between nodes in the network[4]. Moreira et al. (2017) leveraged this value to identify key players (e.g., nodes) in a network. Nodes with the highest betweenness centrality are those defined as key players in the linked nodes connectivity.

- **Modularity.** A community detection algorithm, this tool defines how well a network decomposes into modular communities[5] by algorithmically detecting communities based on the density of links within each community compared to the density of links between communities (Blondel et al. 2008). The modularity algorithm implemented by Gephi is the one proposed by Blondel et al. (2008) and is based on two phases applied for each algorithm iteration. The first phase assigns a different community to each node such that the initial partition produces some communities equal to the number of nodes in the graph. After community allocation, local modularity for each node is evaluated, and nodes are reallocated to a different community or remain within the same community. This phase is repeated until no further improvement is achieved. The second phase of the Blondel et al. algorithm constructs a new graph with nodes that are aggregated into the communities identified during the first phase.

4.2.3 Nodes and edges filtration

For analysing partnership patterns, a filtration for maintaining nodes and edges of interest is essential to identify the frequent patterns. This requires specifying a threshold to filter the nodes and edges against. One approach from Videla–Cavieres and R´ıos (2014) proposed the value of the threshold should equal the average value of the top three heavy (in terms of weight) edges. Thus, edges with a weight less than the threshold value are removed. However, this research suggests the top three heavy edges threshold (*tthet*) is not adequate for filtering a business partnerships network because *tthet* might eliminate nodes with multiple low weighted edges in favor of nodes with few or even a single heavily weighted edge. Both node types are considered interesting for analysing a

[4] https://github.com/gephi/gephi/wiki/Betweenness-Centrality

[5] https://github.com/gephi/gephi/wiki/Modularity

business partners network because the node degrees and edge weights represent business partnership activity.

A second approach for filtering nodes and edges is to remove edges with a weight lower than the average weight (Kim, Kim & Chen 2012). However, this approach might also cause a loss of nodes with a high degree but low weighted edges. Therefore, a third approach is proposed here to determine the appropriate filtration criteria. The nodes to be removed are identified by calculating the graph's average degree (AD) and graph average edge weight (AEW). Those nodes of degree (N_d) that are less than AD and with an edge weight (Ew) less than AEW are removed. Thus, only nodes of a degree higher than AD or an edge weight higher than AEW are selected for analysis. The union of these two conditions to select graph data is illustrated in Formula 1, which ensures that nodes with few high weighted edges and nodes of many low weighted edges are maintained.

Partnership Network Filtration Formula: $(N_d > AD) \cup (E_w > AEW)$ **(1)**

After applying this filter, around 30% of the nodes and 5% of edges from each year are selected for analysis.

4.2.4 Community detection

With the business partnerships network tuned with the nodes and edges of interest identified and selected, the next task is to implement the graph mining task of community detection, which is the process of identifying substructures corresponding to significant functions (Fortunato & Barthe´lemy, 2007). The two measures to discover communities within the business partnerships network are betweenness centrality and modularity and are estimated in Table 4. After running the modularity algorithm provided by Gephi, a list of detected communities together with corresponding the percentage of nodes is captured. Additional details on this process are included in the next section.

Year	Nodes Count	Edges Count	Modularity	Network Diameter[6]
2015	11,433	1,644	0.935	3
2016	9,323	1,920	0.892	3
2017	7,991	1,456	0.919	3

[6] https://github.com/gephi/gephi/wiki/Diameter

Table 4. *The filtered graph dataset for each year together with the corresponding modularity results*.

4.3 Results

The number of communities detected by the modularity algorithm was high with 10,821 identified for the year of 2015. Thus, another level of filtration is applied at this stage to retain the top seven largest communities for the analysis phase. These top seven are selected because the percentage of nodes within each is between 0.25% and 0.12% while each of the remaining 10,814 clusters contains only between 0.09% and 0.01% nodes. This approach to consider only the largest communities discovered was also implemented by Ríos and Videla–Cavieres (2014).

To visually reflect the results of modularity and betweenness centrality on the graph, formatting is applied to the nodes and edges using the value of these measures to enhance readability and facilitate the analysis. The nodes are partitioned based on the community class they are related to, and nodes size is set to be proportional to its betweenness centrality value. Moreover, the edge thickness is set to represent the corresponding edge weight. Network spatialization is another important step to enhance the analysis and is performed by applying two layout algorithms, Force Atlas 2[7] and Fruchterman Reingold[8], as shown in Figure 6. Table 5 provides details on the communities detected within the business partnerships network for 2015. A detailed analysis of these results is discussed in the following chapter.

[7] https://github.com/gephi/gephi/wiki/Force-Atlas-2

[8] https://github.com/gephi/gephi/wiki/Fruchterman-Reingold

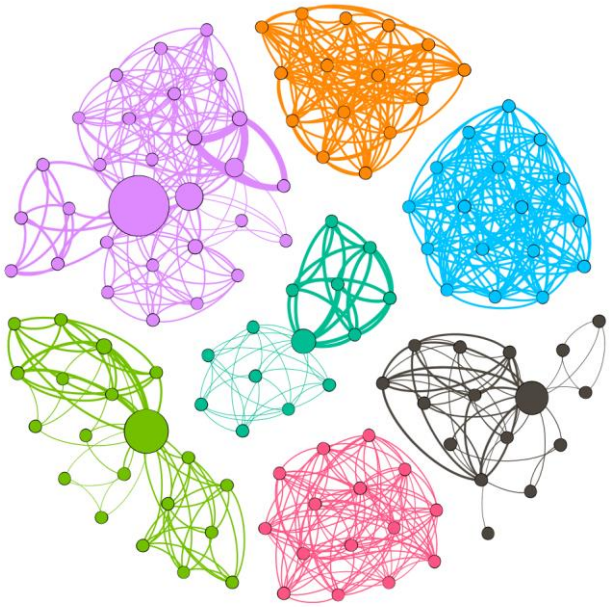

Figure 6. *A visualisation of the seven largest communities detected in the business partnerships network for 2015. Each cluster is presented with a unique colour, which is consistent throughout this work for easier reference. The node sizes are proportional to the betweenness centrality value that reflects a node's role within its cluster. The bigger the node size, the more active it is in building business partnerships among the neighbourhood partners. The edge's thickness, on the other hand, is proportional to edge weight. The thicker the edge is the greater the number of trade licenses established between the connected partners.*

Year	Community #	Number of Nodes (Percentage)	Cluster Colour
2015	1	0.25%	**Purple**
	2	0.17%	**Light Green**
	3	0.15%	**Blue**
	4	0.14%	**Black**
	5	0.13%	**Orange**
	6	0.13%	**Pink**
	7	0.12%	**Green**

Table 5. *The top seven largest communities detected within the business partners network for 2015. The percentage of the number of nodes that form each discovered community is listed with the communities sorted in descending order such that community #1 is the largest in terms of the number of nodes forming it. Each community is marked with a unique colour as shown in the cluster colour column for referencing purposes.*

Chapter Five: Analysis

This chapter discusses the communities detected within the business partners network to answer the research questions. This review involves analysing the discovered patterns and investigating the attributes of each.

5.1 Research findings and results analysis

The analysis of the business partnerships network is implemented in two stages. First, the largest seven communities discovered for 2015 are investigated. Second, the behaviour of the largest seven cluster nodes in 2015 is traced over the subsequent 2016 and 2017 datasets.

5.1.1 Analyzing business partners communities detected in 2015

Beginning in 2015, as shown in Figure 6, cluster 1 is significantly larger than all other identified clusters and includes 29 of the total graph nodes (0.29%). Also in cluster 1, two prominent nodes with high betweenness centrality (BC) values are seen within the cluster. The size of the nodes is proportional to its betweenness centrality value, which implies a key role played by these two nodes in the connectivity of the cluster. Table 7 presents the main characteristics of cluster 1 and, for more insight, Figure 7 presents the attributes of the cluster 1 nodes. Business partner nationality and age attributes are labelled on each node. The business partners with Egyptian and French nationalities have the highest betweenness centrality values and are both in their late 30s. The weighted degree (WD) value together with the degree (D) value of these nodes implies a high engagement of business activities.

Another notable characteristic of the largest community identified for 2015 is that the business partners identified are from various foreign nationalities and no Emirati partner is identified within this cluster. As edge thickness represents the weight, cluster 1 reflects the case of several companies established among business partners with diverse nationalities and few partners connected across these companies. Also, as seen in cluster 1, four nodes are connected by markedly thick edges involving business partners from Pakistan, Canada, Ireland, and the United Kingdom (UK). The D and WD of these four partner nodes are also high with significantly lower betweenness centrality. This condition implies a scenario of multiple companies established among these business partners. Table 6 presents the main characteristics of the six key nodes from cluster 1.

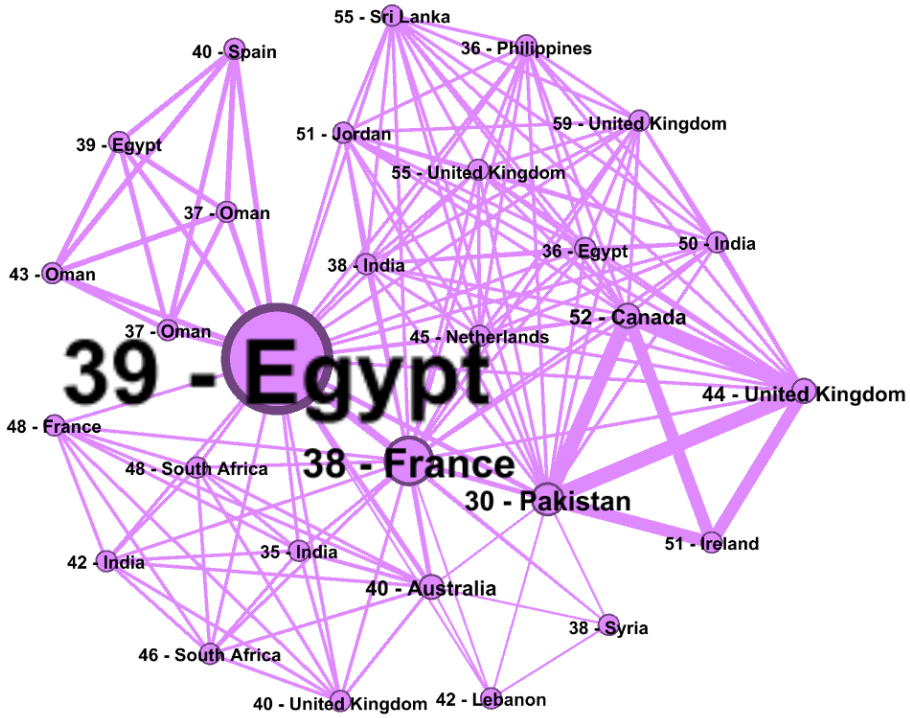

Figure 7. *Cluster 1 discovered in the business partnerships network of 2015. Each node is labelled with the business partner nationality and age. The nodes size corresponds to the BC value and edges thickness reflects the number of trade licenses connecting two partners. The network shows multiple subgroups of business partners where few partners of high BC value are connected across these subgroups.*

Node Id	Nationality	Sex	Age	D	WD	BC
311499	Egypt	Male	39	27	163	167.57
479607	France	Male	38	22	120	52.57
569732	Pakistan	Male	30	17	143	21.90
607768	Canada	Male	52	14	128	7.04
607763	United Kingdom	Male	44	14	128	7.04
586589	Ireland	Male	51	3	63	0

Table 6. *The characteristics of the six key nodes identified within cluster 1 of the business partnerships network for 2015.*

The next largest cluster is cluster 2 with characteristics presented in Table 7. As seen in Figure 8, one prominent node is connecting the remaining nodes within the cluster. This is a similar case to cluster 1 in which a business partner is involved in different trade activities across different companies of different partners. A variation on the edge weights is easily identified implying a variation of trade activity size among the partners. Also, as seen in Figure 8, the cluster 2 motif is similar to that cluster 1. In terms of demographical attributes, most cluster 2

partners are of Indian nationality with the dominant partner from Oman. The average partner age is nearly 44.

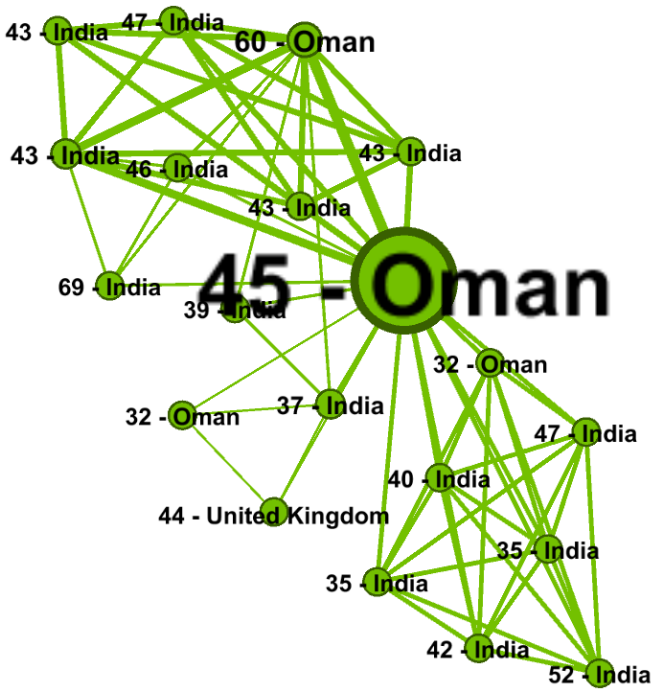

Figure 8. *The network structure of cluster 2 for 2015. The graph shows one dominant business partner is involved in business activities across different subgroups of business partners.*

The business partnerships network presented in cluster 3 features a different structure compared with the networks of clusters 1 and 2. Figure 9 shows the partnership network of cluster 3 with no dominant nodes and edges that are intensively showing a connecting cluster of nodes representing a clique. This scenario is also seen through the measures of D, DW, and BC for cluster 3 as listed in Table 7. BC is zero for all nodes, which explains why no prominent node shows within cluster 3, the average D and average WD are greater than those calculated for cluster 1 and cluster 2 and, hence, the edges are the most remarkable element.

The motif of cluster 3 can be interpreted as a case of common business activities shared among the same group of business partners. This is in contrast to cluster 1 and cluster 2 where the network motif indicates the existence of multiple business activities among different subgroups of business partners where particular business partners are connected across these subgroups. Cluster 3 also features partners of diverse nationalities.

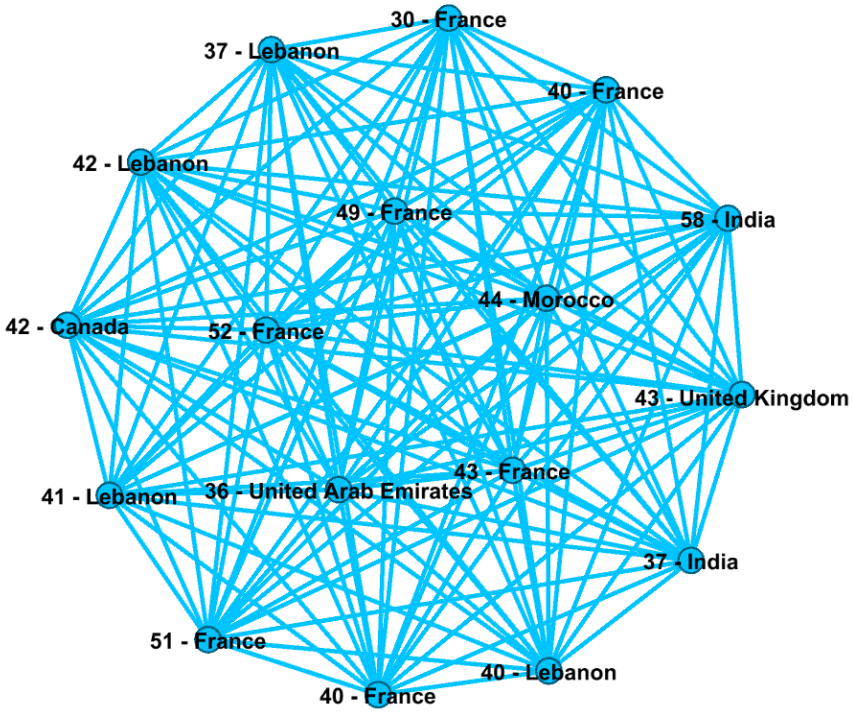

Figure 9. *The business partners network of cluster 3 for 2015. The cluster nearly shows a partnership paradigm where the same business partners are connected through multiple business activities. This can be deduced from consistent node sizes, which represents the BC measure, and via the edge thickness and connectivity*.

Cluster 4 has a similar motif to cluster 2 as seen visually in Figure 10 and through calculated measures listed in Table 7. However, for cluster 4, the D, WD, and BC measures are smaller than those for cluster 2. The paradigm of business activity among partners within cluster 4 is similar to that of cluster 2 except that cluster 4 partners are involved in fewer business activities.

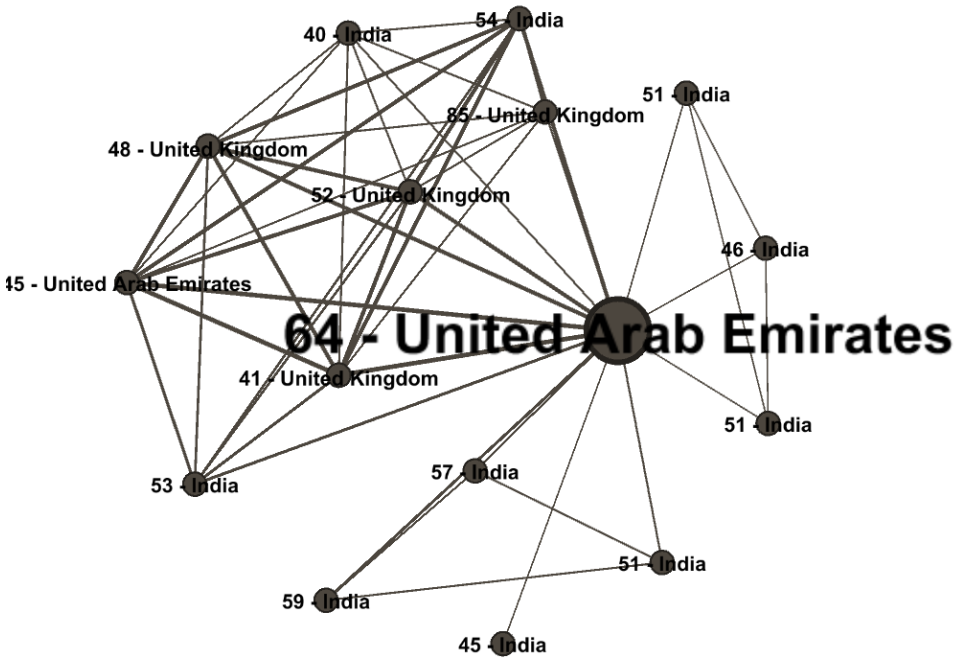

Figure 10. *The business partners network of cluster 4 for 2015. The node sizes are proportional to the BC value, and the edge thickness represents the number of trade licenses between partners.*

Cluster 5 and cluster 6 resemble cluster 3 but with less average D, WD, and BC values as well as different demographical attributes. The motif of both clusters simulate the existence of one, or multiple common business activities shared among the same group of business partners. In cluster 5, all business partners are of Indian nationality except one that is British. For cluster 6, most business partners are Emiratis. Another key difference is the average WD that is visually identifiable in both clusters. Cluster 5 features thicker edges than those of cluster 6, which indicates an increased size of business partnerships among the cluster 5 members. Figure 11 and Figure 12 show the cluster 5 and cluster 6 networks with cluster details included in Table 7.

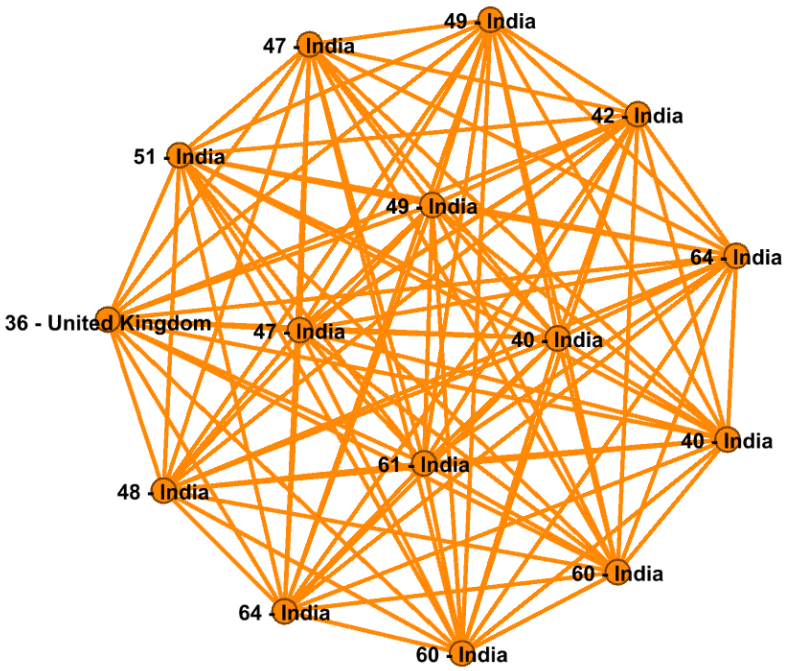

Figure 11. *The business partners network of cluster 5 for 2015. The node sizes are proportional to BC, and the edge thickness represents the number of trade licenses connecting two partners.*

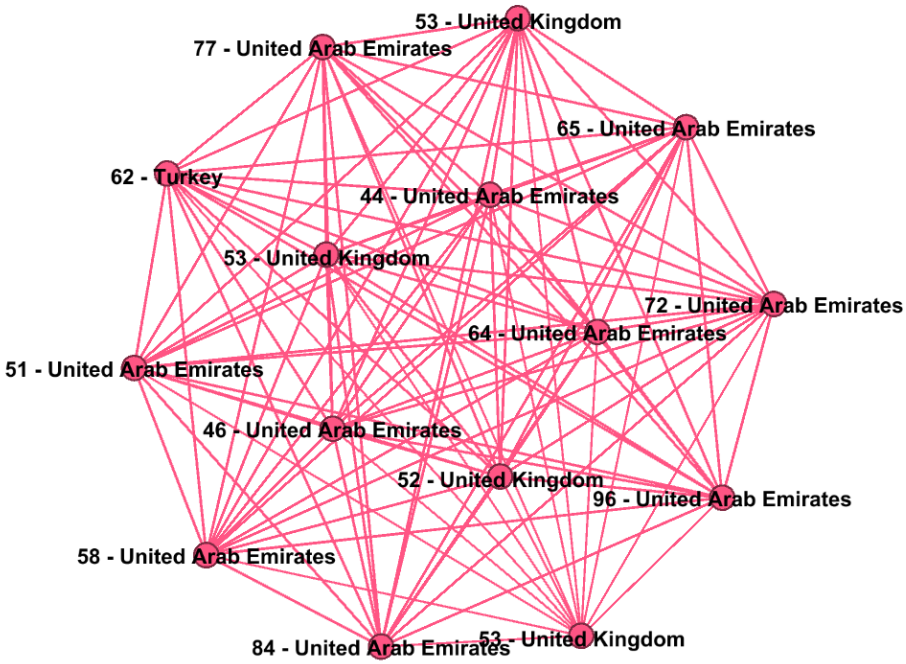

Figure 12. *The business partners network of cluster 6 for 2015. The node sizes are proportion to BC, and the edge thickness represents the number of trade licenses connecting two partners.*

The last cluster 7 seen in Figure 13 is similar to the motif of cluster 1 and cluster 2. However, a distinct difference exists as a prominent node of an Indian businessman is clearly shown linking two subgraphs (e.g., two groups of business partners). One of these two subgraphs is showing an excessive partnership among its members, and the other shows a lesser partnership among its members. This scenario is deduced from the thickness of the edges in both subgraphs, which represents the number of business licenses between two connected partners. The characteristics of cluster 7 are presented in Table 7.

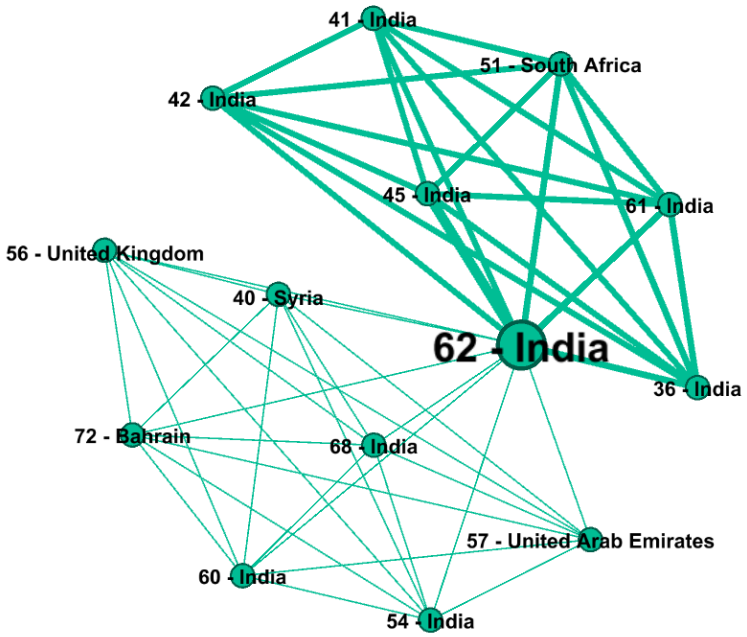

Figure 13. *The business partners network of cluster 7 for 2015. The node sizes are proportional to BC, and the edge thickness represents the number of trade licenses connecting two partners. The network shows a prominent business partner connected with two subgroups of partners. Observing the edge thickness, one subgroup of partners is involved in increased business activities compared to the other subgroup members.*

As observed above, most business partners in the 2015 data are foreigners of diverse nationalities. In terms of partner gender, most are male. Figure 14 is a projection of business partner nationalities on a world map to visualise the diversity within business partners in each cluster. Its corresponding colour identifies each cluster following Table 5.

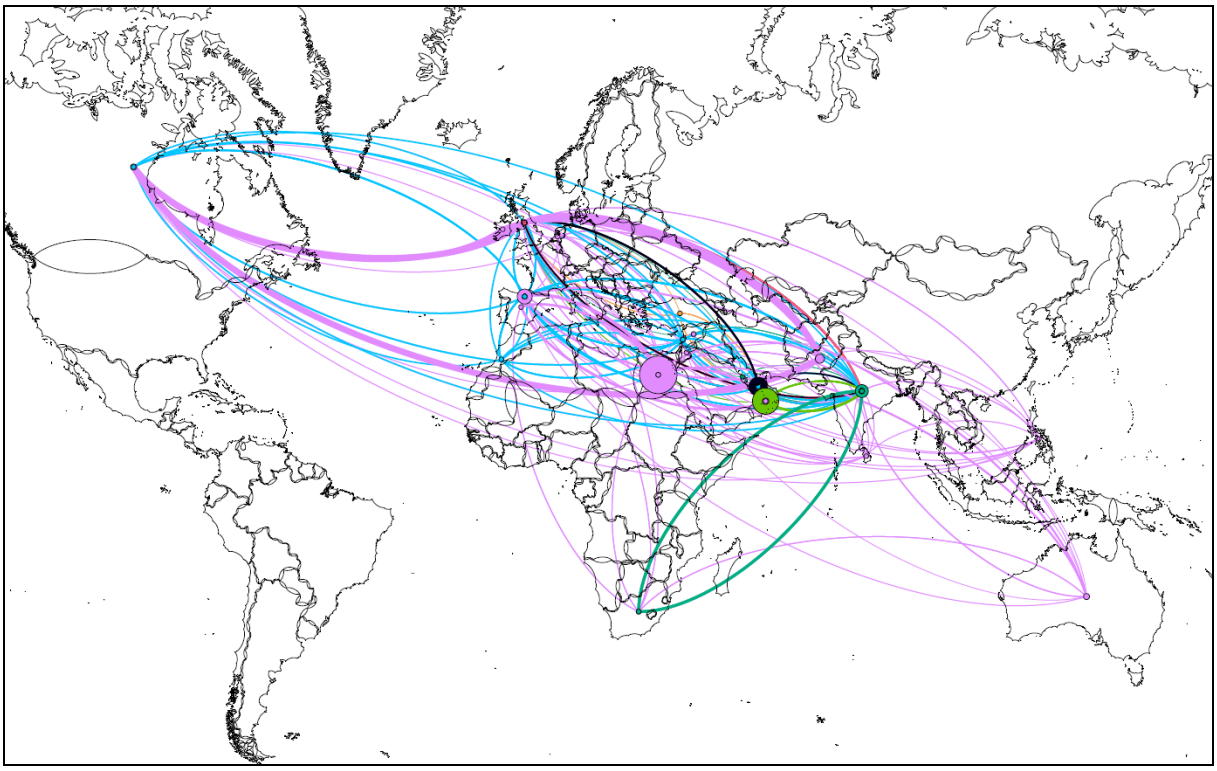

Figure 14. *A world map showing the top seven largest business partner clusters discovered within the business partnership dataset for 2015. Each is labelled with a unique colour where the business partner's nationality is projected to the corresponding country. The figure visualises the diversity of business partner backgrounds linked in business partnerships in Dubai city.*

In summary, the graph mining approach is successfully implemented to explore and analyse the business partnerships network. The largest communities are detected and analysed with two identified network layouts. The major network motifs for which the business partnerships were graphed are presented in Figure 15.

Cluster #	**Number of Nodes**	**Number of Female Partners**	**Number of Male Partners**	**Average Age**	**Nationality characteristic**	**Average D**	**Average WD**	**Average BC**
1	29	3	26	43.2	Multinational	10.6	64.3	9.1
2	20	0	20	43.7	India/Oman/UK	6.6	34.7	6.2
3	17	0	17	42.6	Multinational	16	112	0
4	16	1	15	52.6	India/UAE/UK	5.8	26	4.6
5	15	0	15	50.5	India	14	98.1	0
6	15	1	14	62	UAE	14	54.1	0
7	14	0	14	53.2	India	7	43	3

Table 7. *The main characteristics of the top seven clusters identified within the business partnerships network data of 2015.*

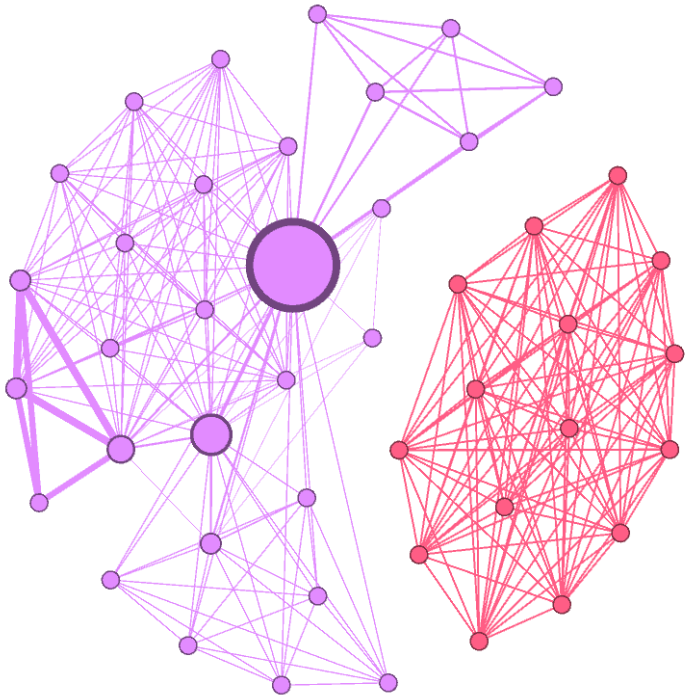

Figure 15. *The two network motifs of the business partnerships networks identified within the discovered clusters. The node size represents the betweenness centrality value. The purple subnetwork structure visualises a case of significantly active business partners connecting different subgroups of less active business partners. While the second red motif is showing as a group of business partners uniformly connected forming a clique, it is likely they are partners of a single or a set of firms.*

5.1.2 Business partner behaviours over time

The second stage of the business partnerships network analysis, the partner behaviour is captured over the subsequent two years. The business partners datasets for 2016 and 2017 were filtered by selecting only the 126 nodes that formed the seven largest clusters of 2015 and then executed the same graph mining tasks. Table 8 presents the seven clusters identified for 2016 and 2017. Of the 126 business partners in 2015, 106 were retained after filtering for 2016, and 113 business partners were retained in 2017. This result implies that most of the largest business partner communities tend to maintain their business activity over the following period.

Figure 16 and Figure 17 display the seven clusters detected within the business partnerships network during 2016 and 2017. Here, eight clusters are discovered in 2016 and 2017, but the eighth is represented by only one node and is excluded. Table 9 presents the characteristics of the 2016 and 2017 business partners dataset analysis after this stage.

Year	Community #	Number of Nodes (Percentage)	Cluster Colour
2016	1	0.19%	Purple
	3	0.16%	Blue
	5	0.14%	Orange
	6	0.14%	Pink
	7	0.13%	Green
	2	0.12%	Light Green
	4	0.10%	Black
2017	1	0.21%	Purple
	3	0.15%	Blue
	5	0.13%	Orange
	6	0.13%	Pink
	4	0.12%	Black
	7	0.12%	Green
	2	0.11%	Light Green

Table 8. *The business partner clusters for 2016 and 2017. Clusters are sorted in descending order within each year. The largest cluster is at the top, and the smallest is at the end. Cluster colour is unified over the three years for enhanced readability and analysis. Cluster 1 preserved its position over the three years whereas cluster 2 ranked at the bottom of the list in 2016 and 2017, which indicates that the business partners in cluster 1 maintained the highest business activity over the three years.*

Year	Nodes Count	Edges Count	Modularity	Network Diameter	Average D	Average WD
2016	105	521	0.751	3	9.9	120.6
2017	112	614	0.742	3	10.9	51.5

Table 9. *The characteristics of the filtered business partnerships datasets for 2016 and 2017.*

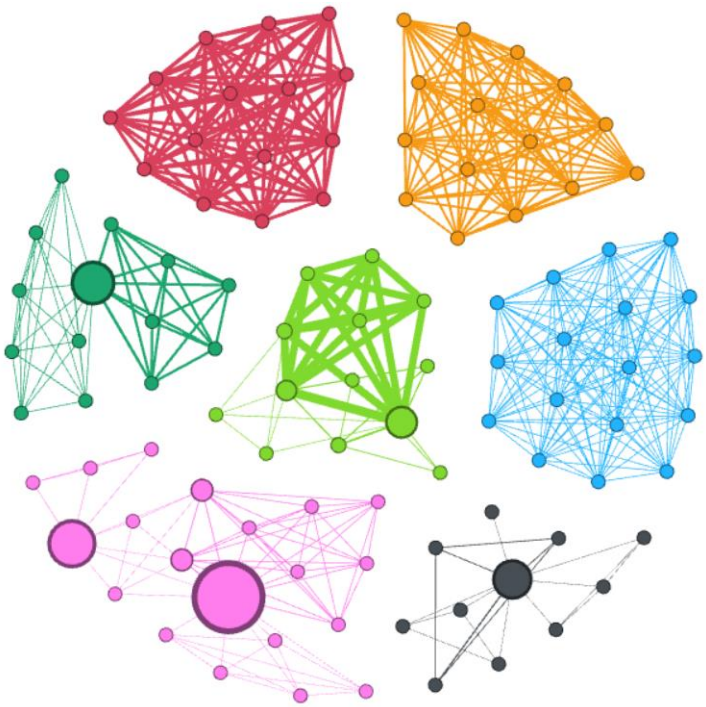

Figure 16. *The business partner clusters discovered in 2016 identified with a labelled colour similar to that applied for the corresponding cluster in 2015 for enhanced readability and comparison. As seen, each cluster maintained the same structure as the one captured in 2015.*

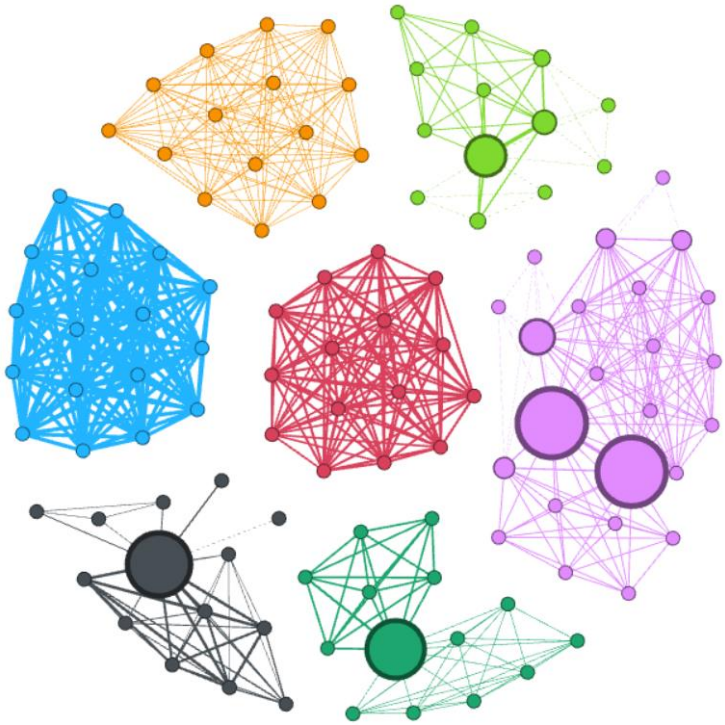

Figure 17. *The business partner clusters discovered in 2017 identified with labelled colour unified across 2015 and 2016 for enhanced readability and comparison. As seen, each cluster maintained the same structure as the one captured in 2015 and 2016.*

A visual comparison between the clusters discovered in 2015, 2016, and 2017 shows that the cluster network structures are similar over the three years indicating a balanced business partnership behaviour during this time. This observation is also consistent with the main communities captured and graphed using the two network motifs as shown in Figure 15. However, to gain a deeper insight into 2016 and 2017, the cluster characteristics, main graph measures, and metrics are calculated and presented in Table 10 and Table 11.

Cluster #	Number of Nodes	Number of Female Partners	Number of Male Partners	Average Age	Nationality characteristic	Average D	Average WD	Average BC
1	20	1	19	42.4	Multinational	6.9	18.7	7.7
3	17	0	17	43.6	Multinational	16	64	0
5	15	0	15	51.5	India	14	196.1	0
6	15	1	14	63	UAE	14	322	0
7	14	0	14	54.2	Multinational	7	67.2	3
2	13	0	13	46.4	India/Oman/UK	5.8	184.7	3.1
4	11	0	11	52.2	India/UAE/UK	3.4	8.3	3.3

Table 10. *The characteristics of the top seven clusters identified within business partners network for 2016. Clusters are ordered by the number of nodes forming each cluster. The largest clusters, cluster 1, is positioned at the top, and the smallest cluster, cluster 4, is at the end.*

Cluster #	Number of Nodes	Number of Female Partners	Number of Male Partners	Average Age	Nationality characteristic	Average D	Average WD	Average BC
1	24	3	21	46	Multinational	11.5	33.6	6
3	17	0	17	44.6	Multinational	16	144	0
5	15	0	15	52.5	India	14	28	0
6	15	1	14	64	UAE/UK	14	82.1	0
7	14	0	14	55.2	Multinational	7	23.1	3
4	14	1	13	55.2	India/UAE/UK	6	24.1	3.5
2	13	0	13	47.4	India/Oman/UK	5.8	16.1	3.1

Table 11. *The characteristics of the top seven clusters identified within the business partners network of 2017. Clusters are ordered by the number of nodes forming each cluster. The largest cluster, cluster number 1, is positioned at the top, and the smallest cluster, cluster 2, is at the end.*

Figure 18 presents the top seven clusters identified for each year again with a unified cluster colouring for enhanced readability and comparison. During the three years, no cluster demonstrates a major change in structure, which implies no dramatic change affected the business partnership behaviours in 2015, 2016, and 2017. However, some clusters show increased or decreased business activities. For example, the business partners of cluster 6 participated in an increased number of new business licenses in 2016 compared to 2015 and 2017. The partners of cluster 5 experiences fewer business activities in 2017 compared to 2015 and 2016, and such behaviour is observed through the changing in edge thickness connecting these partners within the cluster. Quantified measures of average degree and average weighed degree support this visual analysis of clusters 5 and 6 as listed in Table 7, Table 10, and Table 11, which suggests the power of visualisation provided by social network analysis techniques.

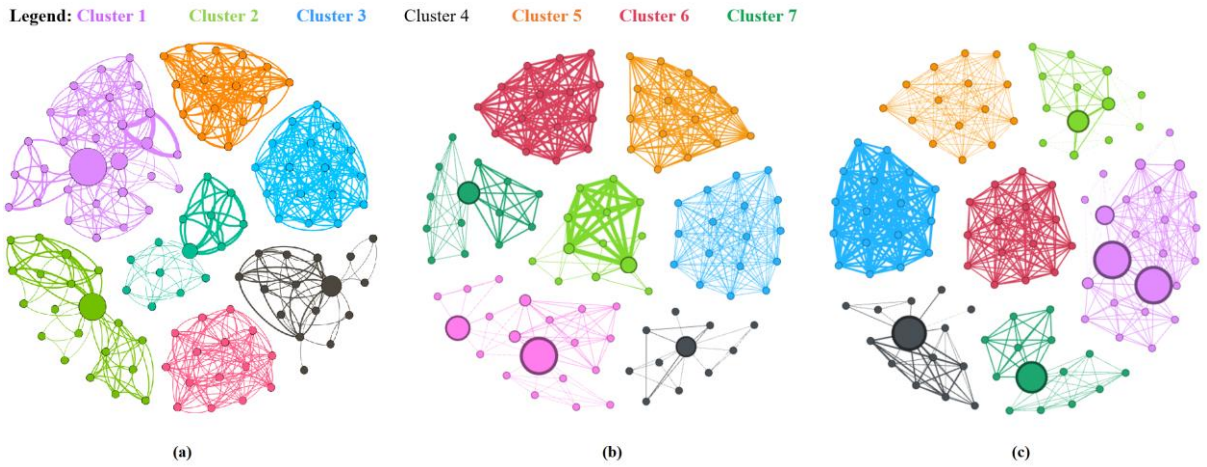

Figure 18. *The top seven communities detected in the business partnerships network over (a) 2015, (b) 2016, and (c) 2017. The clusters maintained their structure over the three years. However, a variation of the node sizes and edge thickness identify change affecting each cluster during the three years of analysis. Cluster 3, for example, shows increased partnerships among its members in 2017 comparing to 2015 and 2016.*

The ribbon charts presented in Figure 19 and Figure 20 were plotted using Power BI software to visualise the variance of the average degree and average weighted degree for each cluster over the three years. As seen in Figure 19, most clusters maintained the same average degree value which, as discussed above, implies no partnerships with new partners. However, clusters 1, 2, and 4 show less partnership activity in 2016 compared to 2015 and 2017. From Figure 20, 2016 marks a significant increase in business activities among the same business partners within the same cluster. This is deduced from the increased value of an average weighted degree and the stability of average degree value.

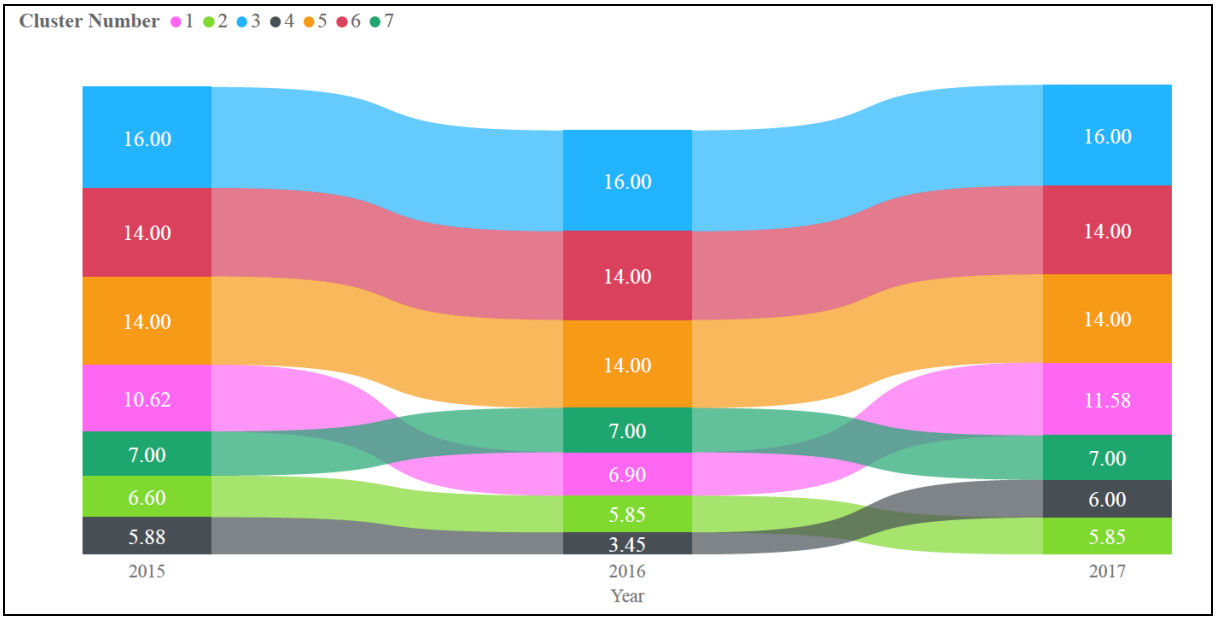

Figure 19. *The average degree value for each of the seven clusters identified in the business partnerships network during 2015, 2016, and 2017.*

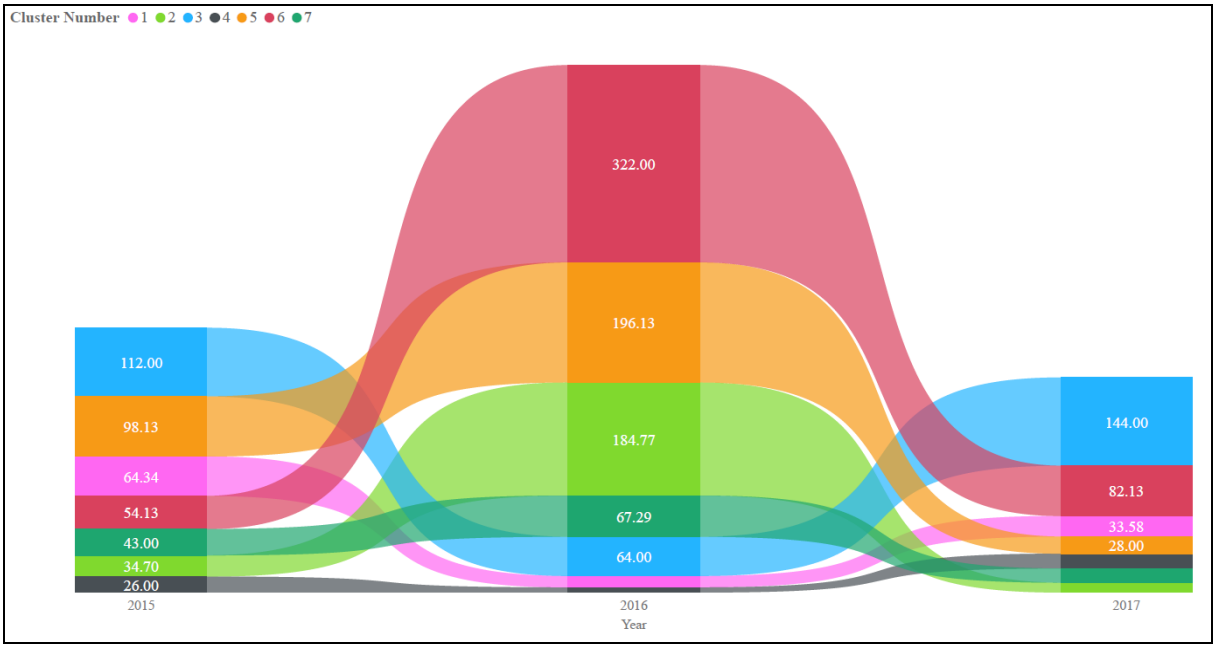

Figure 20. *The average weighted degree value for each of the seven clusters identified in the business partnerships network during 2015, 2016, and 2017.*

The betweenness centrality measure is also observed for the discovered clusters over the three years. As discussed above, betweenness centrality indicates the establishment of superior business partners within a cluster. As shown in Figure 21, cluster 1 maintains rank one over the three years. Overall, the figure shows reverting of the dominant role to the business partners in clusters 1, 2, 4, and 7 during 2016 and 2017.

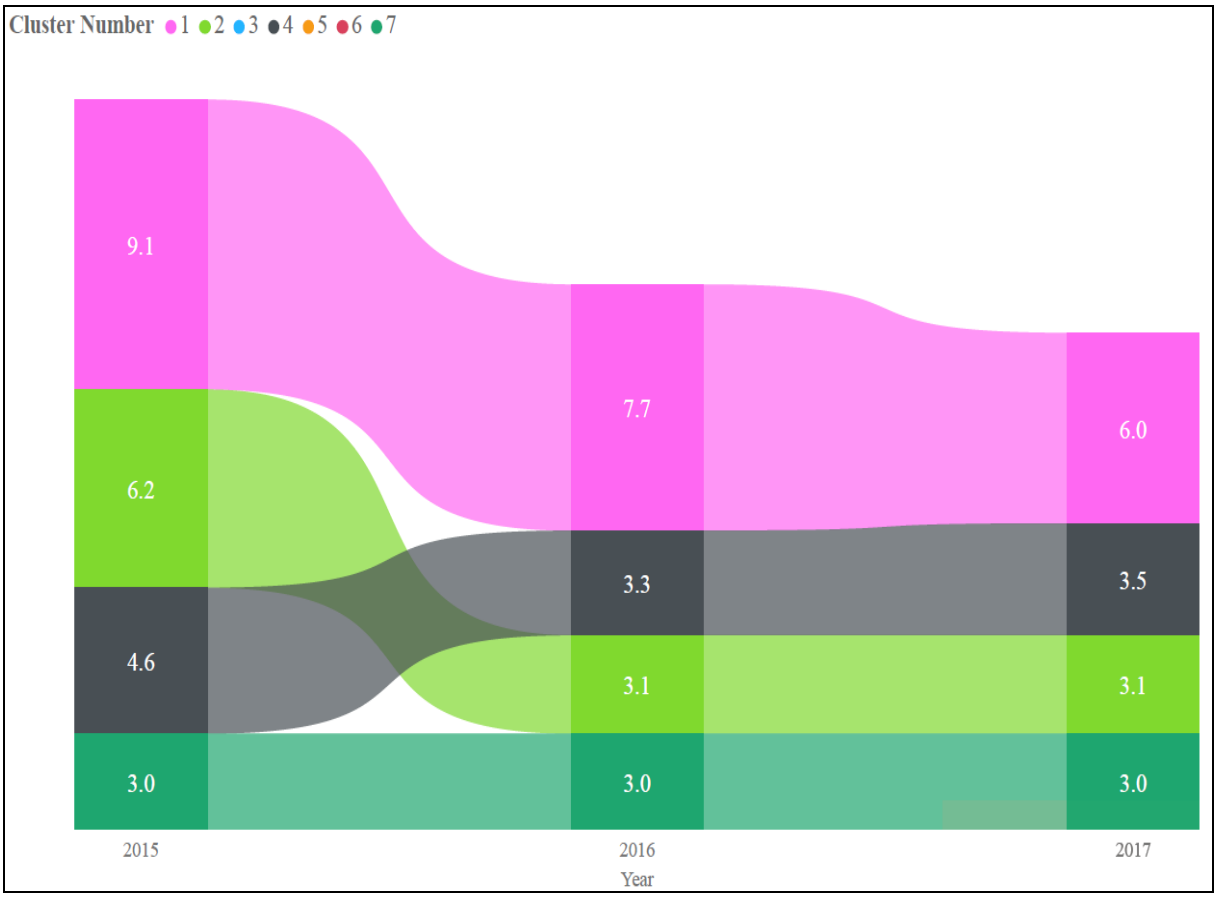

Figure 21. *The average betweenness centrality of the business partners clusters during 2015, 2016, and 2017. Clusters of zero betweenness centrality value are not shown in the figure.*

Finally, a world map is utilised again to visualise the nationalities of business partners linked together within each cluster between 2016 and 2017. Presented in Figure 22 and Figure 23, the colours represent clusters, node size is proportional to the betweenness centrality value, and edge thickness is proportional to the number of trade licenses connecting two partners.

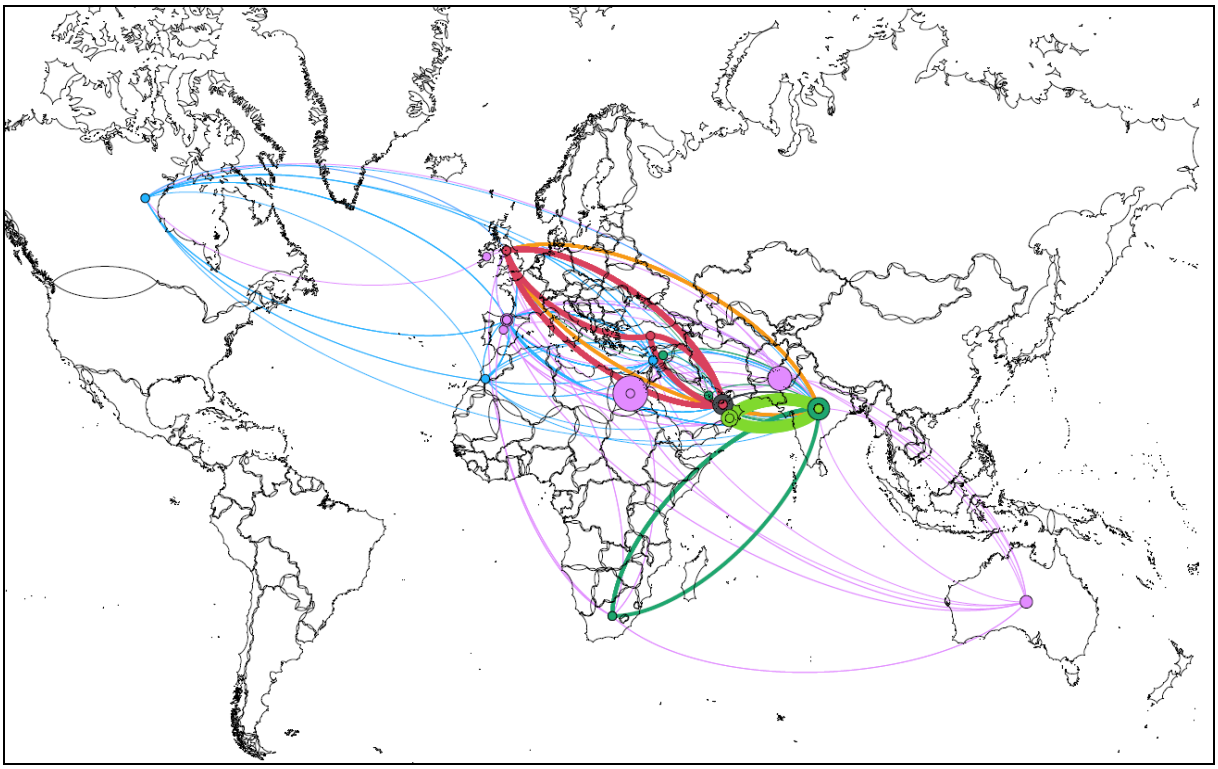

Figure 22. *A world map showing the business partnerships network for each cluster discovered in 2016. The node size is proportional to the betweenness centrality, and edge thickness is proportional to the number of business partnerships established between two partners.*

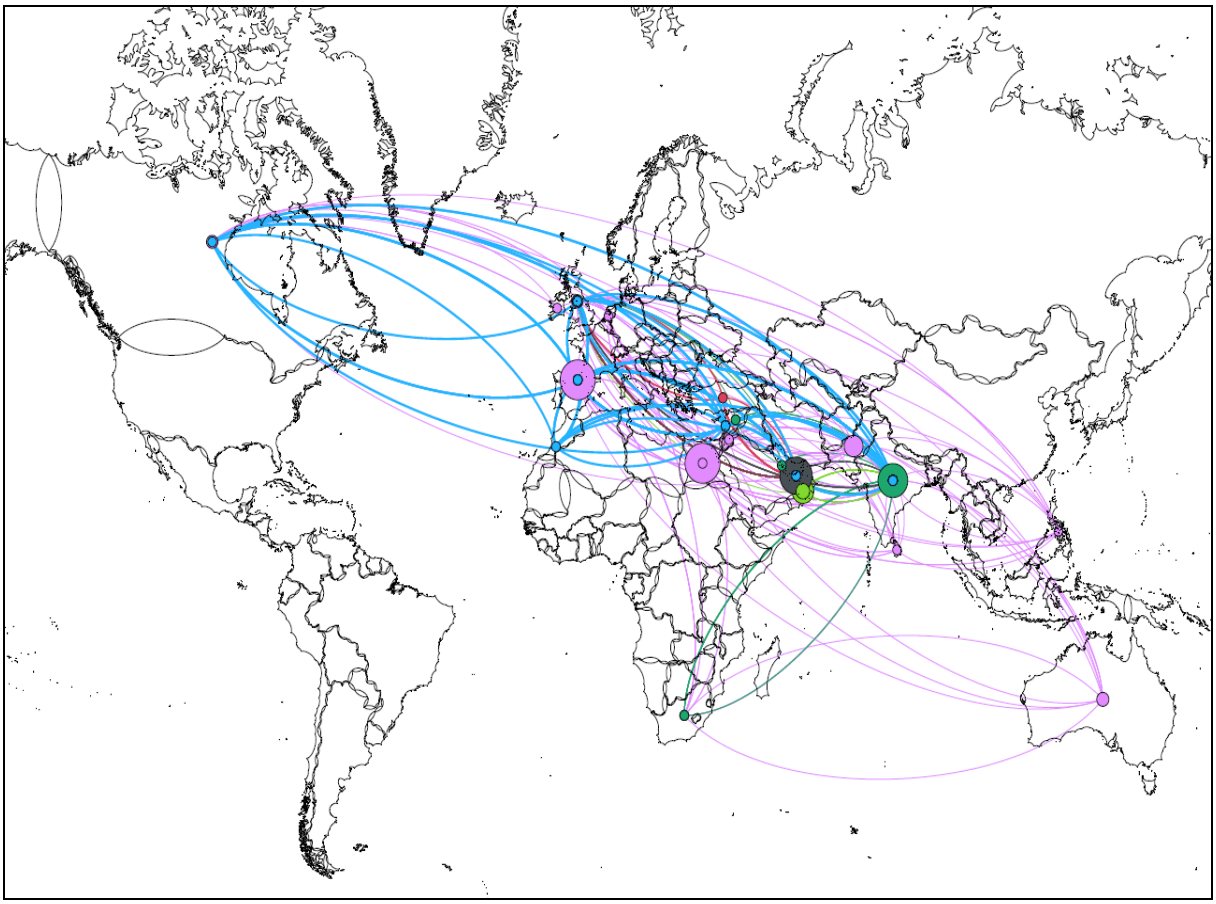

Figure 23. *A world map showing the business partnerships network for each cluster discovered in 2017. The node size is proportional to the betweenness centrality, and edge thickness is proportional to the number of business partnerships established between two partners.*

Chapter Six: Conclusion and Future Work

6.1 Conclusion

Social network analysis is gaining increased interest due to its use in many applications. The need to extract knowledge from big data has urged the data science community to utilise social network analysis as an alternative to traditional data mining techniques (Pushpa 2012). This study proposed a novel domain of business partnerships analysis to utilise social network analysis techniques. The proposed approach aims to reduce costs and effort incurred from using traditional clustering techniques, especially with large datasets.

Business partnerships were successfully analysed as a form of a social network in which the social (partnership) network was created by representing business partners as nodes connected with links representing trade licenses with participating companies. Furthermore, a weight was allocated to each link representing the number of trade licenses established between both

partners. The constructed network was an undirected weighted network with no self-loops. Partnerships with organisations or companies (non-human entities) were excluded from this work since the scope of this study is to analyse individual business behaviours.

The top seven largest clusters discovered within the business partnerships dataset from 2015 was analysed followed by their behaviour during the subsequent 2016 and 2017. The main identified clusters were characterised by diverse business partner nationalities and backgrounds with an average partner age of 43 years. Two distinct motifs were identified within the discovered clusters. The first network structure formed a clique, which can be analysed as a group of business partners linked together through the same business activities. The second motif was characterised by specific dominant business partners connected to different subgroups of business partners forming a core-periphery structure. These two motifs are claimed to be the most common network structures discriminating social networks according to Chakrabarti and Faloutsos (2006), which is another interesting finding of this work.

The clustering coefficient was relatively high over the three years, and the network diameter was small, which are common measures for naturally occurring graphs (Chakrabarti & Faloutsos 2006). Also, high modularity and clustering coefficient values determined over the three years indicated and verified the presence of community structures within the business partnerships network (Newman 2006). In general, the seven clusters identified in 2015 maintained the same structure into 2016 and 2017 indicating a balanced behaviour of business partners over the three years. However, some clusters marked an increased engagement of new business activities and others showed less engagement.

6.1.1 Answers to research questions

In this section, an answer to each of the research questions is provided.

1: How can we visualise business partnership data as a social network?

Answer. Since the aim of this study is to analyze the business partnerships network, utilization of a graph to represent and explore this relationship was effective as discussed in Section 3.2 where the graph was demonstrated a suitable layout to visualize the relationship among business partners showing key aspects such as common partnership patterns and the strength of the links among business partners.

2: What patterns can we discover within business partnership data using graph mining techniques?

Answer. The seven most common patterns, as described in Sections 4.3 and 5.1, were extracted and analysed from the 2015 data as a base year. The behaviour of these seven clusters was then examined over the subsequent 2016 and 2017.

3: Can common motifs be identified within a business partners network?

Answer: Two common network motifs were discovered within the business partnerships network over the three years of analysis as presented in Figure 15. Both identified motifs, clique and core-periphery structure, are considered the most common network structures that characterise social networks (Chakrabarti & Faloutsos 2006).

4: How diverse are the formed partnerships in terms of nationality and background?

Answer. The dominating partner nationalities of each cluster discovered over the three years of analysis are highlighted and presented in Table 7, Table 10, and Table 11. Moreover, a visualisation of the nationalities of the business partners linked in each cluster was provided on a world map for a compelling observation in Figure 14, Figure 22, and Figure 23. The largest cluster identified featured diversity in the business partners' backgrounds. However, other clusters included the same nationality among all business partners.

Finally, several important limitations must be considered. First, the study analysed business partnerships behaviour only over three years. However, capturing this behaviour over an extended period, such as five or ten years, would provide deeper insight into the business partner behaviours. Second, a limited set of business partner attributes were provided. Additional attributes of interest, such as level of education, employment status, professional level, and number of residency years in the United Arab Emirates, were not available. Third, verification of the graph clustering accuracy relied on quantified measures and extracted community motifs. However, due to scope limitation, an extended verification of the results through repeating the analysis using different tools, using alternate clustering techniques or by discussing the results with domain experts should be considered for future work.

6.2 Future work

The scope of this study will be extended to overcome the highlighted limitations and to enhance the potential for the proposed work. By expanding the data not to filter out the corporate partners would provide extra value to the economic field since business-to-business (B2B), and business-to-consumer (B2C) activities and services are significant relationships

(Dotzel & Shankar 2016), and this analysis approach could lend important insight. Moreover, the implementation of an evolution prediction to the business partnerships network would provide additional value to this study. Applying the technique of link prediction enables the prediction of missing links within a network and allows the prediction of future new or dissolution of links (Wang 2015). Implementation of a link prediction can be extended to a recommendation system suggest a business partner to another. Finally, the inclusion of enterprise success and failure indicators would allow the business partnerships network to be constructed and analysed using the License-Partner network topology. In this case, clustering can be applied to identify the key success and failure factors experienced by companies or partnerships. Such factors can be utilised for identifying classification and prediction tasks to be performed by successful or failed businesses.

References

Alesina, A., Harnoss, J. and Rapoport, H., (2016). Birthplace diversity and economic prosperity. *Journal of Economic Growth*, *21*(2), pp.101-138.

Audretsch, D.B. and Keilbach, M., (2010). Entrepreneurship and growth. In *Knowledge Intensive Entrepreneurship and Innovation Systems* (pp. 309-320). Routledge.

Bellini, E., Ottaviano, G.I., Pinelli, D. and Prarolo, G., (2013). Cultural diversity and economic performance: evidence from European regions. In *Geography, institutions and regional economic performance* (pp. 121-141). Springer, Berlin, Heidelberg.

Bindu, P.V. and Thilagam, P.S., (2016). Mining social networks for anomalies: Methods and challenges. *Journal of Network and Computer Applications*, *68*, pp.213-229.

Blondel, V.D., Guillaume, J.L., Lambiotte, R. and Lefebvre, E., (2008). Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, *2008*(10), p.P10008.

Chakrabarti, D. and Faloutsos, C., (2006). Graph mining: Laws, generators, and algorithms. *ACM computing surveys (CSUR)*, *38*(1), p.2.

De Nooy, W., Mrvar, A. and Batagelj, V., (2018). *Exploratory social network analysis with Pajek*. Cambridge University Press. 3rd edn. Viewed 13 November 2018. https://books.google.ae/books?hl=en&lr=&id=HgpaDwAAQBAJ&oi=fnd&pg=PR13&dq=Exploratory+Social+Network+Analysis+with+Pajek&ots=GmXnyH97rU&sig=E0hlkNfJhaEXExSwhyfu7vD573U&redir_esc=y#v=onepage&q=Exploratory%20Social%20Network%20Analysis%20with%20Pajek&f=false

Deo, N., (2017). *Graph theory with applications to engineering and computer science*. Courier Dover Publications.

Dotzel, T. and Shankar, V., (2016). The Effects of B2B Service Innovations on Firm Value and Firm Risk: How do they Differ from those of B2C Service Innovations?.

Fischer, I. and Meinl, T., (2004), October. Graph based molecular data mining-an overview. In *Systems, Man and Cybernetics, 2004 IEEE International Conference on* (Vol. 5, pp. 4578-4582). IEEE.

Fortunato, S. and Barthelemy, M., (2007). Resolution limit in community detection. *Proceedings of the National Academy of Sciences*, *104*(1), pp.36-41.

Fortunato, S., (2010). Community detection in graphs. *Physics reports*, *486*(3-5), pp.75-174.

Glaeser, E.L., Kerr, S.P. and Kerr, W.R., (2015). Entrepreneurship and urban growth: An empirical assessment with historical mines. *Review of Economics and Statistics*, *97*(2), pp.498-520.

Gulf News. (2018). UAE promotes tolerance and respect for others. 23 January. [Accessed 10 November 218]. Available at: https://gulfnews.com/opinion/editorials/uae-promotes-tolerance-and-respect-for-others-1.2161752

Han, J., Pei, J. and Kamber, M., (2011). *Data mining: concepts and techniques*. Elsevier.

Hartmann, P.M., Zaki, M., Feldmann, N. and Neely, A., (2016). Capturing value from big data–a taxonomy of data-driven business models used by start-up firms. *International Journal of Operations & Production Management*, *36*(10), pp.1382-1406.

He, K., Li, Y., Soundarajan, S. and Hopcroft, J.E., (2018). Hidden community detection in social networks. *Information Sciences*, *425*, pp.92-106.

Hochsztain, E., Tasistro, A. and Messina, M., (2015), September. Data Mining applications in entrepreneurship analysis. In *Data Mining with Industrial Applications (DMIA), 2015 International Workshop on* (pp. 25-29). IEEE.

Inokuchi, A., Washio, T. and Motoda, H., (2003). Complete mining of frequent patterns from graphs: Mining graph data. *Machine Learning*, *50*(3), pp.321-354.

Johnson, J., (2013). *Graphs and networks*. [blog entry] [Accessed 13 November 2018]. Available at https://shapeofdata.wordpress.com/2013/08/13/graphs-and-networks/

Kapadia, Y. (2016). Dubai most cosmopolitan city in world; 83% foreign residents. *Khaleej Times*. [online] 16 January. [Accessed 10 November 2018]. Available at: https://www.khaleejtimes.com/nation/dubai/dubai-most-cosmopolitan-city-in-the-world

Kasseeah, H., (2016). Investigating the impact of entrepreneurship on economic development: a regional analysis. *Journal of Small Business and Enterprise Development*, *23*(3), pp.896-916.

Kemeny, T., (2017). Immigrant diversity and economic performance in cities. *International Regional Science Review*, *40*(2), pp.164-208.

Kim, H.K., Kim, J.K. and Chen, Q.Y., (2012). A product network analysis for extending the market basket analysis. *Expert Systems with Applications*, *39*(8), pp.7403-7410.

Mihalcea, R. and Radev, D., (2011). *Graph-based natural language processing and information retrieval*. Cambridge university press.

Moreira, M., Pelissari, P.I.B.G.B., Parr, C., Wohrmeyer, C. and Pandolfelli, V.C., (2017). Data mining on technical trends and international collaborations in the refractory ceramic area. *Ceramics International*, *43*(9), pp.6876-6884.

Naudé, W., (2014). Entrepreneurship and economic development. *International Development. Ideas, Experiences and Prospects*.

Nettleton, D.F., (2013). Data mining of social networks represented as graphs. *Computer Science Review*, *7*, pp.1-34.

Neville, A., (2013). Why Partnership Is Harder Than Marriage. *Forbes* [online]. [Accessed 09 November 2018]. Available at: https://www.forbes.com/sites/amandaneville/2013/03/01/why-partnership-is-harder-than-marriage/#55abb6e87ec9

Newman, M.E. and Girvan, M., (2004). Finding and evaluating community structure in networks. *Physical review E*, *69*(2), p.026113.

Newman, M.E., (2006). Modularity and community structure in networks. Proceedings of the national academy of sciences, 103(23), pp.8577-8582.

Pushpa, S., 2012. An efficient method of building the telecom social network for churn prediction. *International Journal of Data Mining & Knowled Management Process*, *2*(3), pp.31-39.

Quirin, A., Cordón, O., Vargas-Quesada, B. and de Moya-Anegón, F., (2010). Graph-based data mining: A new tool for the analysis and comparison of scientific domains represented as scientograms. *Journal of Informetrics*, *4*(3), pp.291-312.

Ríos, S.A. and Videla–Cavieres, I.F., (2014). Generating groups of products using graph mining techniques. *Procedia Computer Science*, *35*, pp.730-738.

Sarwar, B., Karypis, G., Konstan, J. and Riedl, J., (2000a). *Application of dimensionality reduction in recommender system-a case study* (No. TR-00-043). Minnesota Univ Minneapolis Dept of Computer Science.

Sarwar, B., Karypis, G., Konstan, J. and Riedl, J., (2000b), October. Analysis of recommendation algorithms for e-commerce. In *Proceedings of the 2nd ACM conference on Electronic commerce* (pp. 158-167). ACM.

Sarwar, B., Karypis, G., Konstan, J. and Riedl, J., (2001), April. Item-based collaborative filtering recommendation algorithms. In *Proceedings of the 10th international conference on World Wide Web* (pp. 285-295). ACM.

Scott, J., (2017). *Social network analysis*. Sage. 4th edn. Viewed 12 November 2018. https://books.google.ae/books?hl=en&lr=&id=i5EmDgAAQBAJ&oi=fnd&pg=PP1&dq=social+NETWORK+anALYSIS&ots=DCS69QZif5&sig=LCriSw4Kup3yG1Y_7VH_IWgW_Yo&redir_esc=y#v=onepage&q=social%20NETWORK%20anALYSIS&f=false

Takigawa, I. and Mamitsuka, H., (2013). Graph mining: procedure, application to drug discovery and recent advances. *Drug discovery today*, *18*(1-2), pp.50-57.

Umamaheswari, E. and Geetha, T.V., (2014). Event Mining Through Clustering. *Journal of Intelligent Systems*, *23*(1), pp.59-73.

Videla-Cavieres, I.F. and Rios, S.A., (2014). Extending market basket analysis with graph mining techniques: A real case. *Expert Systems with Applications*, *41*(4), pp.1928-1936.

Wang, P., Xu, B., Wu, Y. and Zhou, X., (2015). Link prediction in social networks: the state-of-the-art. *Science China Information Sciences*, *58*(1), pp.1-38.

Wang, Y., Liu, X. and Chen, Y., (2017). Analyzing cross-college course enrollments via contextual graph mining. *PloS one*, *12*(11), p.e0188577.

Ward, S., (2018). Why Business Partnerships Fail. *The Balance Small Business* [online]. [Accessed 09 November 2018]. Available at: https://www.thebalancesmb.com/why-business-partnerships-fail-4107045

Watts, D.J. and Strogatz, S.H., (1998). Collective dynamics of 'small-world'networks. *nature*, *393*(6684), p.440.

Wennekers, S. and Thurik, R., (1999). Linking entrepreneurship and economic growth. *Small business economics*, *13*(1), pp.27-56.

Zaki, I.M. and Rashid, N.H., (2016). Entrepreneurship impact on economic growth in emerging countries. *The Business & Management Review*, *7*(2), p.31.

Zhang, H., Kiranyaz, S. and Gabbouj, M., (2017). Outlier edge detection using random graph generation models and applications. *Journal of Big Data*, 4(1), p.11.

Zhou, D., Karthikeyan, A., Wang, K., Cao, N. and He, J., (2017). Discovering rare categories from graph streams. *Data mining and knowledge discovery*, *31*(2), pp.400-423.