TEMPORAL USER ATTRIBUTE-BASED APPROACH TO DETECT COMMUNITIES IN ONLINE SOCIAL NETWORKS

AMIN MAHMOUDI

THESIS SUBMITTED IN FULFILMENT FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

FACULTY OF INFORMATION SCIENCE AND TECHNOLOGY
UNIVERSITI KEBANGSAAN MALAYSIA
BANGI

2019

MASA PENDEKATAN INTERAKSI BERASASKAN CIRI PENGGUNA BAGI MENGESAN KOMUNITI RANGKAIAN SOSIAL ATAS TALIAN

AMIN MAHMOUDI

TESIS YANG DIKEMUKAKAN UNTUK MEMPEROLEH IJAZAH
DOKTOR FALSAFAH

FAKULTI TEKNOLOGI DAN SAINS MAKLUMAT
UNIVERSITI KEBANGSAAN MALAYSIA
BANGI

2019

DECLARATION

I hereby declare that the work in this thesis is my own except for quotations and summaries which have been duly acknowledged.

3 January 2019

Amin Mahmoudi
(P85882)

ACKNOWLEGDEMENTS

First of all, I praise God for showing me the right way, without His help I would not have been able to achieve my aims. He has helped me to deal with all the difficulties along my way and in each stage of this work has paved the way for me to reach its conclusion.

I would also like to express my deepest gratitude to Professor Dr. Azuraliza Abu Bakar for her guidance, advice and support throughout my research project. I would also like to thank Dr. Mohd Ridzwan Yaakub, who kindly agreed to be my co-supervisor, for his invaluable guidance and encouragement. My appreciation also goes to Dr. Mohd Ridzwan Yaakub for providing funding through research grant FRGS/1/2017/ICT02/UKM/02/4. And, last but not least, my heartfelt thanks to my family who supported and encouraged me throughout my endeavours.

Finally, I would also like to thank the research and administration staff of the Sentiment Analysis Lab (Faculty of Science & Technology) at UKM for their technical assistance.

ABSTRACT

Online social networks (OSNs) allow the detection of specific communities of users. This has led to the development of community detection (CD) algorithms. However, these algorithms are unable to define the time intervals needed to detect communities in time-varying OSNs; they are edge- and modularity-based so they only consider the number of connections between users, not the user attributes; and their computational complexity is high. OSN are dynamic and the key players are humans whose geo-location, density of interactions and user weight change over time and influence the formation of user communities. Therefore, this study aims to propose a new method to compute the time interval; a method to compute the user weight and a probability function of friendship based on geo-location; and a new CD algorithm based on the user attributes and time interval. Statistical functions are used to create a simulated model which is tested on six different datasets to identify the time interval. The user weight is computed by a simple exponential smoothing method which is tested on the Universiti Kebangsaan Malaysia (UKM) dataset and the result is compared with three existing methods using the pairwise F measure. Three large-scale datasets are analysed to determine the relation between geo-location and OSN ties. The proposed CD algorithm which is named Recently Largest Interaction (RLI) deploy gravitational search method and is tested on the Travian and UKM datasets and compared with the Dynamic Structural Clustering Algorithm for Network (DSCAN), edge-betweenness, label propagation, Walktrap, Infomap, leading eigenvector, and fast greedy algorithms using normalized mutual information, and adjusted rand index measures. The results show that the proposed approach is able to identify time interval accurately and subsequently the user weight computation method outperforms the three existing methods. In addition, estimating the probability of friendship based on the users' attributes outperforms the scenario in which only geo-distance is considered, and above all, RLI algorithm can detect communities more accurately than existing algorithms and improve time and space complexity.

ABSTRAK

Rangkaian Sosial Atas Talian (OSNs) membenarkan pemgesanan komuniti pengguna yang khusus. Algoritma Pengesanan Komuniti (CD) telah dibangunkan bagi memenuhi keperluan OSN. Walau bagaimana pun, algorima-algoritma ini tidak mampu untuk mentakrifkan selang masa yang diperlukan untuk mengesan komuniti dalam perubahan masa OSNs. iaitu yang bersandarkan pinggir (edge) dan bersandarkan kemodulan, oleh itu ia akan menimbangkan jumlah hubungan di antara pengguna dan bukannya sifat-sifat pengguna. dan kekompleksan pengkomputeran adalah tinggi. Rangkaian Sosial Atas Talian adalah dinamik dan pemain utamanya ialah manusia yang mana lokasi grografinya, kepadatan interaksi dan keakraban pengguna berubah sepanjang masa dan mempengaruhi formasi komuniti pengguna. Oleh itu, tujuan kajian ini ialah mencadangkan satu kaedah baru untuk menghasilkan selang masa secara berkomputer; satu kaedah mengukur keakraban pengguna secara berkomputer dan satu kebarangkalian perhubungan persahabatan bersandarkan lokasi geografi; dan satu algoritma CD bersandaran kepada sifat-sifat pengguna dan selang masa. Fungsi statistik digunakan untuk mencipta satu model simulasi yang telah diuji menggunakan 6 set data berbeza bagi mengenal pasti selang masa. Keakraban pengguna dihitung menggunakan kaedah pelicinan eksponen mudah, di mana ia diuji menggunakan set data UKM dan keputusannya dibandingkan dengan 3 kaedah sedia ada menggunakan ukuran cara berpasangan F. Tiga set data berskala besar telah dianalisa untuk menentukan hubungan antara ikatan OSN dan lokasi geografi. Algoritma CD yang dicadangkan Di mana ia dinamakan sebagai Interaksi Terkini yang Paling Besar (RLI) yang memasang atur kaedah pencarian gravity dan diuji pada set data Travian dan UKM, dan kemudian dibandingkan dengan algoritma-algoritma algoritma Pengelompokan Struktur Dinamik bagi Rangkaian (DSCAN), pinggir betweennwss, label rambatan, Walktrap, peta bermaklumat, vector eigen yang mendahului dan ketamakan yang pantas menggunakan maklumat matang yang dinormalkan dan ukuran indeks rand yang diubah. Keputusan hasilan menunjukkan pendekatan mampu untuk mengenal pasti selang masa dengan tepat dan kemudian kaedah pengiraan berat pengguna adalah lebih baik berbanding tiga kaedah yang sedia ada. Ditambah, anggaran kebarangkalian Persahabatan berasaskan ciri-ciri pengguna telah mengatasi keadaan di mana hanya jarak geografi saya yang diambil kira, dan di atas semua, algoritma RLI dapat mengesan komuniti lebih tepat daripada algoritma sedia ada dan meningkatkan kerumitan masa dan ruang.

CONTENTS

LIST OF TABLES

LIST OF ILLUSTRATIONS

LIST OF SYMBOLS

n	Number of periods in an OSN
t	Refers to a specific time interval
u(t)	Number of users in time interval t
c(t)	Number of connections in time interval t
f (u, c)	OSN growth function based on number of users and number of connections
r_ (per each time)	Growth rate per each time interval
r_total	Growth rate in all time intervals
p(t)	Population of users or connections at time interval t
⌈n/2⌉	Denoted ceiling (n/2)
SE	Denoted standard error
$C_d(v_i)$	degree centrality of node i
$C_e(v_i)$	Eigenvector centrality of node i
$C_{Katz}(v_i)$	Katz centrality of node i
$C_p(v_i)$	PageRank of node i
$C_b(v_i)$	Betweeness centrality of node i
$C_c(v_i)$	Closeness centrality of node i
Q	Modularity
Md	Median
μ_x	Mean
σ_x	Standard of deviation.

LIST OF ABBREVIATIONS

Community Detection	CD
Social network	SN
Online Social Network	OSN
Common neighbors	CN
Event detection	ED
Non polynomial	NP
Girvan & Newman algorithm	GN
Minimum Spanning Tree	MST
Pairwise F measure	PWF
Cumulative Distribution Function	CDF
Joint Degree Distribution	JDD
Weakly connected component	WCC
Largest connected component	LCC
Skewness	SK
Standard deviation	SD
Location based social network	LBSN
Area Under the Curve	AUC
Normalized Mutual Information	NMI
Adjusted Rand Index	ARI

CHAPTER I

INTRODUCTION

1.1 OVERVIEW

Each and every day, a large volume of transactions is stored in online social networks (OSNs) such as Facebook, Instagram and Twitter. Generally, OSNs are online services which facilitate communications in a social network. Understanding the characteristics of such networks is very important because they represent a rich source of information, the analysis of which can contribute to cyber security, event detection, marketing and urban planning. Furthermore, the huge amount of transactions present in OSNs provides a good opportunity to extract knowledge regarding the relation between people who form themselves into communities. Thus the detection of communities in networks is one of the most important problems currently being considered by numerous researchers in the field of computer science (Atay et al. 2017; Bu et al. 2013; Cai et al. 2016; Dev et al. 2014; Dhumal & Kamde 2015; Li et al. 2017; Lu et al. 2015; Moradian Zadeh & Kobti 2015; Sharma & Annappa 2016; Tabarzad & Hamzeh 2017). Researchers working on community detection (CD) attempt to develop methods to find a group of nodes which have a higher connection with each other than with other nodes in the rest of the network (Fortunato, 2010; Newman & Girvan, 2003; Sudhakaran & Renjith, 2016; Wang et al. 2015). Community detection was first introduced as a problem by Girvan and Newman (2003), who proposed the GirvanNewman (GN) algorithm as a solution designed for use in static networks. They tested the GN algorithm on a physics collaboration network and achieved some success. After that, many researchers entered the CD

domain and began to propose other CD models and approaches (Blondel et al. 2008;Clauset et al. 2004; De Meo et al. 2012; Gregory 2007; Liu et al. 2011).

Essentially, a network consists of two main types of entity – nodes and edges – and a network can be categorized into two different types – static or dynamic. In a static network, such as a collaboration network, the nature of the nodes and edges does not change over time; however, in a dynamic network, such as an OSN, the nodes and edges do change with time. Hence, when developing a CD method, the nature of the network needs be taken into account. (Note that, in this research, the term 'nature' refers to some common characteristics of nodes and edges in a network such as geo-location, weight, number of interactions and life span.)

Several algorithms have been proposed for the CD problem, but they have some notable limitations. First, they mostly are edge based, which means that they are affected by the number of edges in the network, such as betweenness-based algorithms, where betweeness is computed as how many times an edge needs to be used in reaching other nodes with respect to the shortest path between nodes or vertices. The existing algorithms also use the modularity metric which assigns the same weight to each connection and only considers the number of connections between all pairs of nodes. However, in OSNs, the weight of connections changes over time according to the relevant user attributes (lifespan, geo-distance, density of interaction and user weight). In other words, some relationships during a given time frame are passive when friends do not interact with each other (Wilson et al. 2012). Second, according to Statista (2016), there will be three billion OSN users in 2020, and based on Dunbar's Number (McCue 2013) each user having 150 friends, this will mean that 450 billion edges will exist across all OSNs. The output of existing CD algorithms is in the form of a graph or a dendrogram (a hierarchical tree structure that consists of a number of levels and shows the community in each level) (Newman, 2004). However, a graph structure representation would have high space complexity when applied to such a huge network (big data) as a graph consists of all the edges, while, on the other hand, the dendrogram has two main limitations when dealing with an enormous volume of data, namely, a complicated implementation and it is not an

auto generate structure, while it uses only in modularity based CD algorithms and formed based on modularity value. However, according to Wilson et al. (2012), OSN users do not have interactions with 50% of their friends and the most active users receive comments from only 5% of their friends, ergo it should be possible to prune the network based on user-interaction attributes. It has also been noted that the CD problem is categorized as an NP-hard problem (Mothe et al. 2017).

Therefore, in light of the foregoing, this study aims to overcome the above-mentioned drawbacks of existing CD methods by proposing a new algorithm for detecting communities that takes into account the relevant user attributes in order to reduce time and space complexity, and above all, to improve detection accuracy to identify meaningful communities that are relevant to the real world. To solve the above-noted CD problems, this study needs to identify the variables that are important for CD research. The correct identification and definition of the characteristics of OSNs is fundamental to the analysis of huge networks and related problems such as CD. Attributes such as user weight, time, density of interaction and geo-location play crucial roles in OSNs. Thus, before proposing a new CD algorithm, it is essential to try to answer the following major questions: such as in which time interval must the proposed CD algorithm be applied? How can the important users be identified and what roles do they play in the forming of communities? What roles do density of interaction and geo-location play in OSNs and in the CD problem? Moreover, the accurate definition of the main user attributes is a limitation in previous works on OSNs (Aston & Hu 2014; Blondel et al. 2008; De Meo et al. 2012) that will be addressed herein.

The remainder of this chapter is organized as follows: section 1.2 describes the research background. Next, section 1.3 outlines the problem statement. Then, sections 1.4 and 1.5 present the research questions and objectives, respectively. After that, section 1.6 highlights the significance of this research. This is followed by section 1.7 which covers the scope of this research, and finally, section 1.8 outlines the organization of this thesis.

1.2 RESEARCH BACKGROUND

As can be inferred from the title of this thesis, this research consists of three main parts that cover (1) the use of a time-based approach, (2) OSN user attributes and (3) CD algorithms. Thus the research background is also divided in three parts. The first part focuses on the time dimension and OSN behaviour, where OSN behaviour refers to users' activities in OSNs. The second part is concerned with the main attributes that are common to every OSN as this study tries to employ these attributes to develop a new algorithm for detecting communities in OSNs. The third and final part is related to the CD problem.

1.2.1 Measurement of OSN Behaviour

Time is a main dimension of an OSN and needs to be considered accurately by every method that is developed for analysing a dynamic network. This is because the nature of an OSN changes according to the number of users and connections over time. Thus, OSN behaviour also changes with time. Hence, the time interval needs to be computed in order to adequately identify network behaviour changes. (Note that in this research that the term 'network behaviour changes' is used to refer to the amount of activities which change in an OSN over time). However, to date, not much research has been conducted on this specific topic. Nevertheless, among the related research that does exist, Sulo et al. (2010) proposed a good method to discretize data in dynamic network based on variance and the compression ratio, which can induce its idea in order to use in the proposed algorithm. However, the variance showed how the data was distributed around the mean value; it did not reveal the amount of distribution. These works are valuable in order to support this research hypothesis. Kazienko et al. (2011) presented a three-dimensional model for social network analysis in which the role of time was emphasized. In their work, a 'time window' was defined as a period of time with well-matched size. However, they did not present a method for identifying said time window. Rather, they suggested a new model for analysing a network based on three dimensions: layer hierarchy, time and group, but they did not propose an experimental model for specific domains. In other words, their model was conceptual. Also, Nicosia et al. (2013) analysed the role of time in networks and found that the relationship between two nodes is not persistent over time. Later, Sekara et al. (2016)

conducted an interesting analysis of dynamic social networks in which they showed that the mobility of individuals in a social network can lead to the prediction of social behaviour over time. On the other hand, some researchers have tried to introduce the OSN' measure (Allcott et al. 2007; Baagyere et al. 2016; Ghali et al. 2012; Himelboim 2017; Mislove et al. 2007; Newman 2003; Santoro et al. 2011; Zafarani et al. 2014). However, the above-cited works have mostly analysed the effectiveness of structural measures such as centrality and have not put forward measures for monitoring user behaviour. Some researchers have also used OSN measures in their respective methods (Ahmed et al. 2010; Jiang et al. 2013).

Recently, in light of the exponential growth rate in the use of OSNs, some researchers working in OSN domain adopted a different approach by assuming that the users were the main players in OSNs and consequently began to analyse user behaviour in OSNs. For example, Benevenuto et al. (2009) investigated how users communicate with each other in an OSN, especially in Orkut, a social network run by Google. They also studied user behaviour in other OSNs and observed that it cannot be represented as a normal distribution with comparable mean and variance. Later, Wilson et al. (2012), in their significant work, analysed the role of time and users' activities in OSNs and tested their approach on the Facebook dataset. Their study showed that users tend to communicate with a small group of friends and often do not have any interactions with 50% of their friends on Facebook. This means that, over time, even though many users join Facebook, the growth rate of the interactions between them is not that high. This implies that finding the recent time interval is fundamental for identifying current user communities.

1.2.2 User Attributes

In recent years, many research studies have been conducted to identify the important attributes of dynamic networks. These attributes include geo-location and influential nodes (The term 'Nodes' refer to 'users' in all parts of this thesis). In an OSN, each user has a specific weight, which refers to the influence that the user has in the context of the OSN, and the weight of each user is different. A user's weight is a key indicator of the user's influence on the OSN, where the weight of the user is greater, the more influence that user has on the OSN as compared to other users. Freeman (1978) was

the first to use the centrality measure to estimate node weight in social networks. Work in this area culminated in the study by Brin and Page (1998) who introduced the PageRank algorithm for ranking websites. Over a decade later, Shafiq et al. (2013) presented a method to identify four groups of users in OSNs, namely, followers, introvert leaders, extrovert leaders and neutrals. They named their method, the longitudinal user centred influence (LUCI) method, and their research is discussed in detail in the literature. Another noteworthy work is that of Trusov et al. (2010) who proposed a method to identify the influential users in an ego-centred network from the marketing standpoint. However, all of the above works considered the weight of nodes in a static network and did not present a solution that was suitable for OSNs as not only are the weights of the nodes in an OSN different in each time interval, the latest time intervals are more important than other time intervals. Hence a new measure needs to consider this issue to assign the correct weight to each node in an OSN.

Another important variable in an OSN is the user geo-location characteristic, but it has not yet been analysed in depth and no consensus on its role has been reached. Geo-location and its effect on OSN ties has been considered by some researchers (Cho et al. 2011; Cranshaw et al. 2010; Huang & Liu 2015; Kaltenbrunner et al. 2012; Lambiotte et al. 2008; Lengyel et al. 2015; Lengyel et al. 2013; Liben-Nowelly et al. 2005; Scellato et al. 2010), where some of the above have shown that the effect of geo-location on OSN ties is negligible, whereas others have shown that the relationship is strong. However, none of the existing works has produced a solid formula for the relationship between social network ties and distance that considers that the user attribute also has an effect on OSN ties, and this is the main limitation of these works.

1.2.3 Community Detection Problems

As discussed in section 1.1 of this introduction, CD algorithms have been proposed for two types of network: static and dynamic. In this section, CD in these two types of network is considered.

a. Static network

The basic algorithm for CD is the GN algorithm (Newman & Girvan, 2003). This algorithm works on static networks, and the basis of this algorithm and similar algorithms such as those proposed by Gregory (2007), Moradian Zadeh and Kobti (2015), Takaffoli et al. (2011) and Thang N. Dinh and My T. Thai (2011), is modularity (Newman & Girvan 2003), which is basically a measure for evaluating CD algorithms that rely on betweenness. Many of the algorithms for static networks try to detect communities based on the maximum interconnection and lowest intra-connection between communities (Aktunc et al. 2015; De Meo et al. 2012; Liu et al.2014; Moradian Zadeh & Kobti 2015; Newman 2004; Pizzuti 2008; Radicchi et al. 2004; Reihanian et al. 2015; Salter 2015; Sudhakaran & Renjith 2016; Thang N. Dinh & My T. Thai 2011; Xie et al. 2013). However, the drawbacks of these algorithms are that they only consider the edges between nodes and give the same weight to all of them. These algorithms do not consider dynamic properties such as lifespan (time), weight, density, and location, which are crucial features for detecting communities in social networks. Moreover, the problem of maximum modularity has been proved to be NP complete (De Meo et al. 2012). Furthermore, the complexity of the algorithms developed by the above-mentioned studies is high, where the time complexity of the basic algorithms is O (n^3) (Clauset et al. 2004; Hecking et al. 2013; Liu et al. 2014; Newman 2004), and in the best case the time complexity O (n^2). Hence, in other words, none of them have been able to improve on a space complexity of O (n^2).

b. Dynamic network

Social networks in the real world are dynamic, so some algorithms have been developed to try to detect communities in dynamic social networks and real-world settings (Aston & Hu 2014; Ferreira & Zhao 2015; Gauvin et al. 2014; Hecking et al. 2013; P. Nguyen et al. 2014; Takaffoli et al. 2011; Tantipathananandh & Berger-Wolf 2007). Tantipathananandh and Berger-Wolf (2007) presented an approximation algorithm for modelling a dynamic community structure as a graph-colouring problem. Their notion was related to changes in communities over time. They consider each time step as group of individuals which have interactions to each other, where they do not define the threshold of interactions. They also considered this

problem as a class of optimization problem, but their algorithm is modularity based. Gauvin et al. (2014) presented a time-varying adjacency matrix of a temporal network in order to detect communities in dynamic social networks. They stated that their proposed method was “intrinsically temporal and allow[ed them] to simultaneously identify communities and to track their activity over time” (2014: 2). However, they used some constant time intervals such as 5, 15, 30, and 60 minutes, also they did not address time complexity or space complexity. Nevertheless, their work is valuable because, by considering the time component in their work, they changed the way in which the CD problem is viewed. Earlier, Hecking et al. (2013) tried to develop new approaches to identify the optimal time slice size because they intended to conduct a temporal analysis of the community in dynamically evolving networks, however, they only show that the changing of communities with regard to different time slice, while their work do not present a method to define time interval. On the other hand, Stattner and Collard (2015) used the frequent sub-graph discovery method to detect communities which is used to search for frequent patterns in social networks with the aim of identifying the sub-graphs that occurred frequently in a single very large network. However, similar to other modularity-based methods, they concentrated on the quantity of the connections between nodes instead of the quality (time based or interaction-based) of the nodes as their model was designed to search for the sub-graph with the highest connection between nodes. Basically, these dynamic algorithms consider the community in some snapshots and they apply the modularity measure to each snapshot in order to detect communities.

Some researchers have attempted to find ways to detect communities in complex networks. For instance, You et al. (2016) presented a new algorithm for detecting communities based on partition density in complex networks, which they named the IsoFdp algorithm. Their algorithm does not need to have the number of community as prior input (as an advantage), also it can be used in large-scale real-world datasets. Hu and Liu (2015) also introduced a novel algorithm for detecting communities in complex networks. They named their algorithm Infomap-SA and claimed that it had higher modularity and lower computational complexity than the basic Infomap algorithm (Rosvall & Bergstrom 2008). The time complexity of their

algorithm is better than that of the GN algorithm, but it is still $O(n^2)$ and the output is also a graph. In addition, their algorithm is modularity based.

More recently, Dev et al. (2014) presented a state-of-the-art method to detect communities in OSNs that was based on user interaction. The key idea for their study came from the following two observations: (i) the degree of interaction between each pair of users varies widely and reflects the strength of the tie between them and (ii) for each pair of users, the interactions with mutual friends (known as group behaviour) play an important role in determining belongingness to the same community. Earlier, Pietilänen and Diot (2012) analysed the role of temporality in OSNs and proposed a methodology to break the temporal contact graph into clusters of nodes that meet more frequently and for longer periods of time during an experiment. They named these clusters 'temporal communities'.

On the other hand, some researchers have used meta-heuristic algorithms to try to solve the CD problem (Atay et al., 2017; Cai et al. 2016; Li et al. 2017; Lu et al. 2015; Pizzuti 2008; Sharma & Annappa 2016; Tabarzad & Hamzeh 2017). However, they are modularity based, which means that the objective function works based on the modularity metric, also they are not time based and are suitable only for static networks. While Yang et al. (2015) and Pei et al. (2014) used the universal gravitation formula, they too used the modularity metric for evaluation.

1.3 PROBLEM STATEMENT

This section describes how this research aims to address three main problems in relation to the shortcomings in the solutions that have so far been proposed for CD in OSNs. First of all, to the best of the author's knowledge, none of the existing studies have proposed a robust method to define time intervals for CD problem with respect to OSN behaviour. This means that the current methods confine all kinds of OSNs to a constant time interval without taking into account the network behaviour changes (amount of activities) that take place over a period of for example 1 month, 2 months or 1 year. This issue is very important in the context of the CD problem as the real behaviour of OSNs can be captured from the most recent time intervals. This is because, during some periods of time, some connections are passive and users do not

have any interaction with each other. If these passive users and their lack of connections are considered by CD methods, this leads to less accuracy in the detection of community members. However, how can time be discretized? In other words, the main question that arises is which time interval needs to be selected to analyse the network more accurately? The time interval metric must be precise as a large time frame would lead to the loss of some information while too small a time frame would result in an increase in noise (Sulo et al. 2010). The lack of a metric to compute time intervals that covers the maximum number of changes in the network and network behaviour is the main shortcoming of existing studies, where the maximum number of changes and network behaviour refer to user activities in different time intervals. However, the computation of time intervals would lead to an evolution in OSN analysis.

The second main issue in the area of CD, is that the user attributes interfere with the formation of communities because OSNs are a human-centric domain, but existing CD algorithms only use the structural graph measure in order to detect communities in OSNs, thus identifying the user attributes accurately defined as the second main problem. Every OSN consists of three key variables: user weight, geo-distance and density of interaction. Also, due to the time-varying nature of OSNs as dynamic networks each of these variables can change over time. So that, a small change in these variables leads to new forms of OSN. In other words, these features play a fundamental role in OSN behaviour, thus it is crucial to accurately identify their characteristics based on time before employing them in OSN analyses, such as CD. In addition, according to Wilson et al. (2012) for the majority of the users, around 70% of their interactions with only 20% of their friends and according to Wrzus et al. (2013) people in social networks tend to make new connections with new members. This means that, over time, many users are passive and do not have any interaction with each other at all. Therefore, it is clear that the real picture of OSN behaviour is depicted in recent time intervals, or in other words, the most recent time interval has higher priority than the other time intervals. However, the existing studies do not assign priority to time intervals when identifying the importance of users. Therefore, this study aims to propose a new time-based metric to compute the above-mentioned user attributes in order to utilize them in the CD algorithm.

Lastly, the existing CD algorithms are mostly edge and modularity based, which means that they assign the same weight to each connection and consider only the amount of connections between users. However, in OSNs, relevant user interaction attributes (geo-distance, density of interaction and user weight) can change the weight of the connections. This means that some connections are included in an analysis even though they are passive and no interactions are taking place between them. Hence, the existing CD methods that are modularity based fail to recognize meaningful communities because they do not take into account the effect of user attributes. In fact, modularity is not a suitable metric for CD algorithms that are applied to OSNs. Furthermore, it has been reported that modularity has a resolution limit so it cannot find small well-defined communities in large-scale networks. It is therefore necessary to define the requisite user attributes in order to develop a new kind of algorithm to replace the modularity-based algorithm. On the other hand, as mentioned earlier, the output of existing CD algorithms is in the form of a graph so this would lead to time and space complexity in the graph structure. Moreover, the solution for maximum modularity is NP complete (De Meo et al. 2012). However, a huge OSN can be pruned based on the user attributes, which would lead to reducing time complexity. Hence, this study attempts to represent the CD output in the form of a tree, specifically a minimum spanning tree (MST) (Pettie & Ramachandran 2002) as it is envisaged that this will lead reduced space complexity. It has already been shown that the complexity of the graph traversal problem in a network can be improved by using the MST in linear time (Megiddo et al. 1988) where the detection of a community is a type of graph traversal. The time complexity by MST is $O\,(nlogn)$, and the space complexity is $O(n)$, where n is the number of nodes (Neumann & Wegener 2006; Wu & Chao 2004), In contrast, the time complexity of most existing CD algorithms is $O(n^3)$ (Alzahrani & Horadam 2016; Dhumal & Kamde 2015; Hecking et al. 2013), essentially, the MST structure improves space complexity by cutting out the passive connections.

Figure 1.1 depicts the research problem of this study. It shows that this study will leverage user attributes to deal with the limitations of the modularity based CD algorithm and that it will deploy a proposed algorithm whose output is presented in the form of a MST in order to improve time and space complexity.

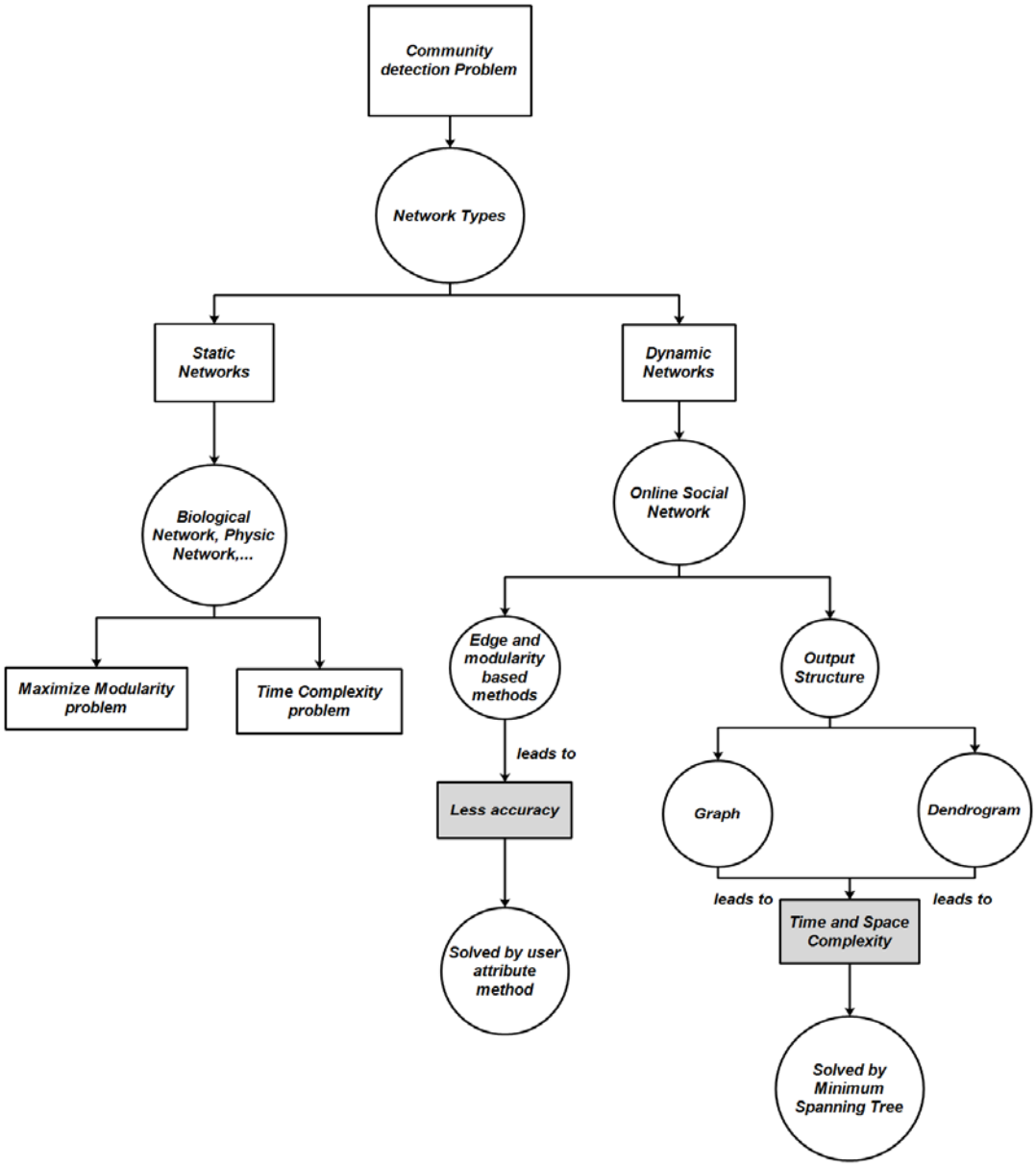


Figure 1.1 Community detection problem statement

1.3.1 Theory and Hypothesis Development

The main theory underpinning this study is derived from the fact that OSNs consist of some discrete time intervals that do not have the same weight and priority, which consequently affects the results of analyses conducted on OSNs. In addition, this study recognizes that the real picture of an OSN is revealed in recent time intervals (Wilson et al. 2012). By applying this theory, it should be possible to rewrite the results of many studies that could not adequately address the issue of time in their particular

contexts. Girvan and Newman were the first to use the modularity metric for evaluating the output of a CD algorithm. Then, due to the developments in OSNs and the growing interest in this domain among scientists, some researchers considered OSNs as dynamic social networks. It was also found that the density of interaction between OSN' members is different in regards to different time period (Wilson et al. 2012). The CD problem has also been studied from the sociological perspective. For instance, Wrzus et al. (2013) found that the social network that is relevant to people changes as they age and that people tend to establish connections with new persons over time. Accordingly, the modularity metric, which works on the basis of the amount of connections only, is not a suitable metric for OSNs. Moreover, due to the exponential rate of growth in the amount of data in OSNs, it has now become an area of interest to those working in the field of big data (Sharma & Oliveira 2017), which means that it has become necessary to create new algorithms that are online and scalable in terms of memory and computational resources. Moreover, new methods need to be able to prune these huge OSNs. Therefore this research study proposes a new user attribute-based approach to overcome the above-mentioned problems in the field of CD.

1.4 RESEARCH QUESTION

Based on problem statement of this research four main questions are presented as follow:

1. How can the time intervals in OSNs be computed in order to cover most of the behavioural changes and evolution in OSNs?
2. What is the effect of user attributes on CD in OSNs?
3. Is an algorithm based on user attributes more accurate than edge- and modularity-based algorithms?
4. How can the time and space complexity of existing CD algorithms for CD in OSNs be improved?

1.5 RESEARCH OBJECTIVES

In order to answer the research questions, the three main objectives of this research were to:

1. To Propose a new method to compute time intervals based on network behaviour changes in order to cover the most OSN behaviour changes;
2. To propose types of user attributes that can enhance the performance of CD in OSNs; and
3. To Develop a CD algorithm based on user attributes and time interval in order to achieve better accuracy, time and space complexity compared to the existing methods.

1.6 SIGNIFICANCE OF RESEARCH

This research is timely and significant for the domains CD, event detection, recommender systems and marketing because due to the exponential growth of users as well as the dynamic nature of OSNs the main concern of many research studies in those domains is finding a way to identify the important users in OSNs and other attributes such as the accurate definition of time intervals that cover the maximum changes that take place in such networks. In addition, it is predicted that the number of OSN users will reach three billion in 2020 and it has also been estimated that people spend around 2 hours per day on OSNs (statista 2016). Alongside this rapid rise in the usage of OSNs and the advantages of such networks, there is also some negative activity and abuse taking place via such networks, including terrorism, child abuse and organized crime. Thus, it has become more necessary than in the past to attempt to manage this messy network in which half of the people around the world are involved. The existing CD algorithms try to arrange similar nodes into groups. However, they in this huge database (OSNs) are complicated as mentioned in the background and also they are modularity based.

Today, OSNs are considered to offer a good opportunity to find out about many of the events taking place around or involving OSN members, which is of help to people generally and also to cybersecurity agencies. Hence a significant aspect of

this study is its contribution to enhancing the ability to detect events based on the volume of interactions and changes in communities, rather than on the text content. Such events include momentous global events before they happen, such as earthquakes, which other studies have tried to find ways to detect and report on, rather like global news agencies. However, this study also is significant because it can be used to detect personal life events such as the immigration of OSN members, which could be important to the members of the group and to national security departments. Some researchers have used a text-based approach to detect the above types of event in which a keyword or the text of a message is considered and events are detected based on the analysis of the text (Sayyadi et al. 2009). However, such methods cannot accurately identify the many different types of event that take place, especially personal life events, while the text in tweets or messages is sometimes so ambiguous that text-based methods alone cannot identify an event properly. Other researchers have deployed communication volume in order to detect an abnormal number of messages which acts an indicator of an event (Chierichetti, et al. 2014; Krumm & Horvitz 2015). This prior research suggests that it is possible to detect events based on changes in the user community, where changing communities occurs as the result of a specific event. For example, when a person moves to another country they may join new communities on OSNs. In fact, a community confines the space of event-detection problems into smaller groups that are more relevant to local events. Tan et al. (2014) developed a multilevel method that detects global and local events based on user communities by using a fast unfolded CD algorithm that was proposed by Blondel et al. (2008). Unfortunately, their CD algorithm only works on a static network and is modularity based, which means it is based on the number of connections while other user attributes such as user weight, geo-location and lifespan are ignored, which then leads to the detection of the wrong communities. In order to detect events in a timely and effective manner, a dynamic algorithm based on user behaviour is needed. This is because the main players in personal life events are human beings. Therefore, such an algorithm should detect communities based on user behaviour.

In addition, the current study is of significance because the proposed CD algorithm can be applied to the link prediction problem in OSNs. Link prediction has

become a hot topic and many research studies have been conducted on this issue in recent years (Dhote et al. 2013; Gao et al. 2015; Liu et al. 2013; Peng et al. 2015; Valverde-Rebaza & de Andrade Lopes 2013). There are two main types of link prediction method. The first is a structural method that is based on assessing the similarity between nodes and consists of two subgroups: local and global structural information methods. There are several local similarity methods: the common neighbours (CN) algorithm, Adamic–Adar (AA) index, Jaccard coefficient (JC), resource allocation (RA) and preferential attachment (PA) (Dhote et al. 2013; Peng et al. 2015; Srinivas & Mitra 2016). The CN algorithm is widely used as a similarity method. The CN algorithm has demonstrated some success in predicting links between nodes when there are many common neighbours, but it fails to perform as well when the nodes have a low number of common neighbours. The second main type of link prediction method is based on social theory and considers user attributes (Peng et al. 2015).

With regards to the emergence and disappearance of people in OSNs, unfortunately, existing methods have failed to accurately predict new links in OSNs. This is because the previous studies mostly considered the network to be a static environment and frozen in time. However, this assumption is not true to the reality of OSNs that are time-variant. Not only did the previous studies ignore time, they did not consider user attributes, even though it has been shown that user attributes such as geo-location and user weight interfere in social network ties (Bliss et al. 2014; He et al. 2015; Papadimitriou et al. 2012). Moreover, the premise on which the methods proposed in previous studies were based was that the largest number of common neighbours over the whole lifespan of the user would act as the indicator for new links. However, in an environment of changing communities, the criteria for making connections (links) is not a function of the number of common neighbours because in a new community there may be no common neighbours between the newcomer and existing members of the new community. Hence some friendships (links) in OSNs can exist between users who have the lowest number of common neighbours. The existing methods are mostly structure-based and use the CN approach to predict links. However, a user who changes their community does not have as many neighbours in common with the members of their new community as they do with the members of

their older community. Therefore, the current research contributes to predict links in OSNs based on the user community changing.

This study works on the CD problem in the OSN domain with a particular focus on the area of cyber security. The proposed CD algorithm contributes to a vast verity domains as depicts in Figure 1.2. The aim here is to contribute to the important work of security agencies in preventing terrorist activities in societies around the world and other malicious activities in and via OSNs, as highlighted by Figure. 1.2.

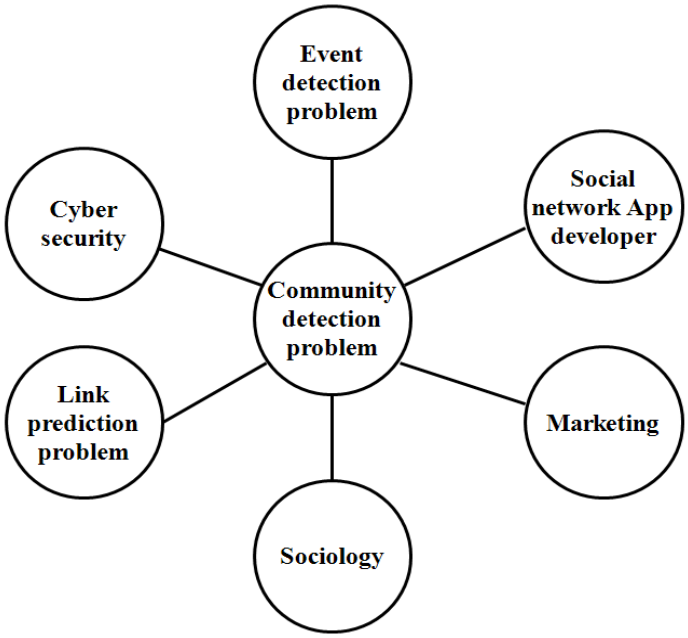


Figure 1.2 Research significance

1.7 RESEARCH SCOPE

This research discusses user attributes in OSNs in the context of CD in order to overcome the problems associated with CD that are described above in the problem statement. This study does not discuss optimization algorithms. However, a future research avenue could investigate the suitability of the proposed algorithm for optimization problems. This is because the CD problem has the characteristics of an optimization problem. Whilst, the proposed CD algorithm is also nature inspired. Also, even though this research highlights the drawbacks of meta-heuristic algorithms in relation to solving the CD problem in OSNs this is from the viewpoint of accuracy

rather than from the optimization problem perspective. Hence the main focus of this research is on the characteristics of user attributes.

In addition, it should be noted that this research considers CD algorithm for both types of network (static and dynamic) because previous research studies have usually used a static algorithm in a dynamic domain by dividing an OSN into some snapshots. It should also be noted that in order to conduct this research a temporal social network is needed. In other words, this research works on a temporal network with time label connections. Hence the data used in this research is collected from a vast variety of resources such as Statista, a provider of market and consumer data company, the BrightKite, Gowalla and Foursquare location-based social networks (LBSNs) and the Stanford University and Irvin California University datasets, as well as the Travian dataset, which is a popular browser-based real-time strategy game with more than five million players, and a synthetic dataset created specifically for this research named the UKM dataset. Therefore, the scope of this research is defined by the above-mentioned datasets, and these datasets have diversity in terms of both time and scale.

In order to find the time interval, the entire dataset needs to be analysed, hence it is not possible to use part of the data from each dataset. In other words, sampling cannot be used in this study (exception for an experiment in section 4.3.2, where a random user is selected). However, it should be noted that for a truly comprehensive CD study, a dataset that contains a time-labelled friendship graph and location information in each time interval for every user is needed. The dataset should be time labelled and the ground truth should be available. However, the ground truth is only available for the Travian and UKM datasets. Nevertheless, this research employed the other datasets as and when relevant to each objective and proposed method.

The last point to note in terms of research scope is that the time needed to obtain a result on betweenness centrality, which is a major feature of existing CD methods, is time consuming. In addition, it takes around 1 week to apply a label propagation algorithm to the Travian dataset.

1.8 THESIS OUTLINE

This thesis is organized into seven chapters, including this introductory chapter, as follows:

- Chapter II contains a review of the literature in the OSN domain. This chapter covers OSN behaviour studies, user attributes and CD algorithms. The first part of the chapter reviews the significant works that consider the characteristics of OSNs. The second part describes the main attributes that have an effect on community formation and finally, the chapter concludes with a review of the state-of-the-art CD algorithms.

- Chapter III presents the methodology adopted for this study. This chapter describes the phases of the research in detail. It also presents an evaluation of the proposed methods for each of the three objectives of this study. In addition, it provides the details of the experimental datasets.

- Chapter IV presents the proposed method for defining time intervals, which is the first objective of this research. The chapter also discusses the method proposed for monitoring OSN behaviour changes. In addition, it provides the result of the experiments conducted on these two proposed methods when applied to six different datasets.

- Chapter V contains an explanation of the proposed method for identifying user attributes, which is the second objective of this research. The chapter presents a new method for computing the user weight. It also provides the results produced by the proposed method which are compared to those of the LUCI, node degree and node centrality methods. In addition, the chapter introduces a new formula for computing the probability of friendship in OSNs. The results produced by this formula are presented and compared with those produced by state-of-the-art methods in which the probability of OSN ties formation is computed based on only distance.

- Chapter VI introduces the proposed method for detecting communities in OSNs, which considers the role of distance, user weight and density of interaction. The chapter also contains a results section in which the output of the proposed method is compared with that of six other algorithms that are considered to be the main and standard algorithms in this domain.

- Finally, Chapter VII highlights the implications of this study and presents the conclusion of this research.

CHAPTER II

LITERATURE REVIEW

2.1 INTRODUCTION

This chapter contains a review of the literature that was conducted for this research. The aim of this chapter is to identify the gaps and limitations in the existing works. This literature review is organized into three main parts. The first part in section 2.2 reviews the structure, topology and behaviour properties of OSNs. Next, the second part in section 2.3 discusses the relevant literature on user attributes in OSNs. Then, the third part in section 2.4 presents the prominent algorithms in the CD domain. Finally, the chapter ends with a summary of the drawbacks and limitations that were identified by the review.

2.2 ONLINE SOCIAL NETWORKS

This section contains a review of the prominent studies. This part of the review was conducted in order to analyse the topology, measures and behaviour of OSNs. First, the definitions of main measures for OSNs which have been used and defined by many researchers are presented. Then the studies which describe OSN behaviour are reviewed. One of the main aims of this first part of the literature review is to determine whether the existing studies have considered time-related measures, such as the milestone, time interval and exciting period, which are all introduced in the current study.

The definitions of interest to this study are as follows:

- **Definition 2.1** OSN milestones are the time points in OSNs in which the users' activities change markedly.

- **Definition 2.2** A time interval is a series of time periods that contains different amounts of user activities.
- **Definition 2.3** OSN behaviour changes refers the specific states of OSNs in which the amount of user activities changes.
- **Definition 2.4** An exciting period is a time period in OSNs in which the amount of activity is very high (abnormal).

2.2.1 OSN Measures

This section presents the main measures that are common to all OSNs. These measures are assortativity, the clustering coefficient, network closure, propinquity, centrality, density, the clique, cohesion and path length (radius and diameter). These measures fall into three main categories, namely, connections, distributions and segmentation, and are discussed under their respective category (Allcott et al. 2007; Baagyere et al. 2016; Cordeiro et al. 2018; Ghali et al. 2012; Himelboim 2017; Jiang et al. 2013; Mislove et al. 2007; Santoro et al. 2011; Wilson et al. 2012; Zafarani et al. 2014).

a. Connections

The connections category of social network measures encompasses the connection characteristics that exist between users in OSNs.

i. Assortativity

Wilson et al. (2012), Mislove et al. (2007) and Newman (2003) defined the assortativity coefficient (r) of a graph measure as the probability of the nodes in a graph linking to other nodes of similar degree. To determine this measure, the Pearson correlation coefficient is computed for the degrees of the node pairs for all the edges in a graph and this calculation returns values in a range between -1 and 1. An assortativity value of less than zero shows that nodes connect to others with dissimilar degrees. On the other hand, an assortativity value that is greater than zero indicates that nodes are likely to connect with those with similar degrees.

ii. Network Closure

Allcott et al. (2007) stated that the concept of network closure is used to indicate the level of connectivity between friends of friends. So, if friends in a network have common friends (neighbourhood), then this network is considered to exhibit a high level of network closure. In contrast, if the friends in a network have different friends then low network closure is present. The difference between the two types of closure is illustrated in Figure. 2.1.

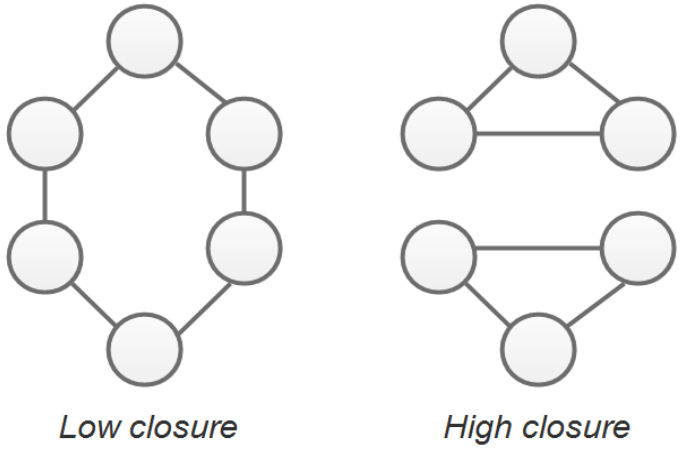


Figure 2.1 Types of closure in a network

iii. Propinquity

The term propinquity is used in social psychology in which it is understood that people who live near to each other are more likely to be friends. This measure is important because it shows the effect of geo-distance on the formation of communities in OSNs (Reagans 2011). Thus it one of the measures that is of great interest to this study because it underpins one of the hypotheses of this research.

b. Distributions

The distributions category contains measures that describe how users are distributed in a network. This is very important especially for identifying the influential nodes in a network.

i. Centrality

Zafarani et al. (2014), Nicosia et al. (2013) and Himelboim (2017) described centrality in social networks and showed that several important types of centrality, such as betweenness centrality, closeness centrality, eigenvector centrality, alpha centrality and degree centrality. These types of centrality are described in detail in section 2.3 of this chapter as they are used to identify influential users and user weights.

ii. Density

The density measure is calculated as the number of existing connections in a network divided by the number of possible connections in a network (Himelboim, 2017; Santoro et al, 2011; Wilson et al. 2012). Figure 2.2 provides an example of some connections in a network and Eqs. (2.1) and (2.2) show how its density is calculated.

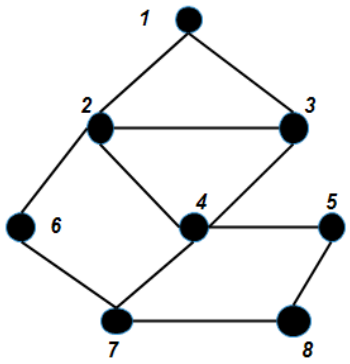


Figure 2.2 Example of connections in a network

$$density = d = \frac{number\ of\ existing\ connections}{number\ of\ possible\ connections} \tag{2.1}$$

$$number\ of\ possible\ connections = \frac{n*(n-1)}{2} \tag{2.2}$$

Thus, for the network in Figure 2.2 the density is computed as:

$number\ of\ possible\ connections = \frac{8*7}{2} = 28$ and $d = \frac{11}{28}$

iii. Path length

To establish the path length, the radius and diameter are calculated using the eccentricity of each node in a social graph. Eccentricity is defined as the maximum distance between a node and any other nodes in the graph. The radius is the minimum number of all eccentricities, while the diameter is the maximum. The average path length is simply the average of all-pairs-shortest-paths on the social graph (Santoro et al. 2011; Wilson et al. 2012).

c. Segmentation

The segmentation category of measures addresses the issue of clustering in OSNs, and the measures in this category, especially cliques, are widely used in CD research.

i. Clustering Coefficient

The cluster coefficient (CC) is a fraction of the possible interconnections in a network. The value of CC is between 0 and 1. The CC of a whole network is the average of CC of each user (Santoro et al. 2011; Wilson et al. 2012).

So, the CC of user 4 in Figure. 2.2 is computed as

$$CC(v) = \frac{2* N_v}{K_v*(K_v-1)} = \frac{2*(1)}{4*(3)} = \frac{1}{6} \quad (2.3)$$

where K_v is a degree of *v* and N_v is the number of connections between the neighbours of *v*.

ii. Clique

A clique in a network is formed when all the nodes have connections with each other. This measure is usually used in CD algorithms as it can show how dense the connections are in a community. Figure 2.3 provides an example of a clique.

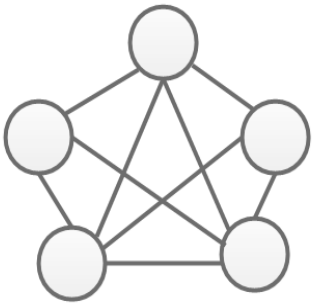


Figure 02.3 Connections between a clique in a network

All of the above measures relate to the network structure. However, the current study suggests that the other measure that is very important in OSN analysis is network behaviour, which can be ascertained through an analysis of network time intervals and network milestones. The network time interval is an OSN measure that shows how the activity of a user is distributed over the lifespan of the OSN, yet this measure has not been analysed in existing studies.

Mislove et al. (2007), in their significant work, were the first researchers to attempt to analyse the meaning of the term OSN and to try to introduce this network measure. They analysed four popular OSNs, namely, Orkut, LiveJournal, Flickr and YouTube to discover the distribution of links in OSNs. Their findings revealed that OSNs have small world; power-law distribution and scale-free properties. The small world measure represents a small diameter and high clustering while power-law distribution represents the probability that a node with degree k is proportional to $k^{-\gamma}$, where γ is the power-law coefficient. The scale-free property indicates that high-degree nodes tend to connect to other high-degree nodes. They also found that a social network has a small diameter and short path length. However, it should be noted that, although Mislove et al. (2007) conducted a novel study to reveal the properties of OSNs, they did not present a time-based measure, which implies that they analysed a static network. While they did analyse the above-mentioned parameters for the whole lifespan of an OSN, the characteristics of these parameters can differ in each time interval as OSNs change over time.

Santoro et al. (2011) claimed that most existing measures which have been proposed for social network analysis fail to capture the dynamics of the social network. So they tried to define related indicators based on temporal and atemporal viewpoints. They used a sequence of short time-varying graphs to deal with the static measurement problem. On the one hand, they introduced the time-varying graph (TVG) model, while on the other hand they used an evaluation model for OSN atemporal measures such as density, CC, modularity, degree power law and conductance. In addition, based on their classification approach some measures fell into the temporal category, such as distance, centrality, betweenness, closeness and diameter and eccentricity. Although their model was time based and considered a sequence of time intervals, they did not present a method to find the time interval, meaning that their model considered some time windows which time windows are defined by default given value. So the lack of a method to compute the time interval to monitor the network evaluation is the main limitation of their work.

Himelboim (2017) divided the social media measure into three levels: node, link and network. At the node level the author selected centrality as the main measure and this measure consisted of degree, betweenness, closeness, eigenvector centrality and reciprocity. For the link level the author defined and included link weight and link reciprocity. At the network level, the author included density, reciprocity and centralization. In addition, the author described the group, cluster and community, power law and preferential attachment attributes. However, the lack of a time-based analysis to address network behavioural changes in the relevant time interval is a shortcoming in this author's work.

Cordeiro et al. (2018) presented a social network measure that included degree, betweenness, closeness, eigenvector centrality and Laplacian centrality. In addition, they introduced some other measures at the network level, such as edge bursts, fluctuability, volatility, reachability latency, temporal efficiency, diameter, radius, average geodesic distance, average degree, reciprocity, density and global CC. They tried to represent the above-mentioned measures with regard to time as they recognized the need to consider the social network as dynamic. They computed the time window as T/n, where n was the number of snapshots. In the case of a dynamic

network, their method used the changes in the number of edges and nodes to define criteria to compute the time interval. However, they did not define a threshold for changing the number of edges and nodes. In another part of their work, they discussed three other methods that could be used to define the time interval to graph data windowing strategies, namely, landmark windows, non-overlapping sliding and overlapping sliding, while they defined a constant time period as the length of each window as they stated "dynamic networks are discretized in time by converting temporal information into a sequence of n static snapshots" (Cordeiro et al. 2018; 8). The main problem of their work and similar works is that they monitor the user behaviour changes not OSN behaviour changes.

2.2.2 OSN Behaviour

This section contains a review of the studies that have discussed user behaviour in OSNs. Specifically, this section looks at how users interact with each other over the lifespan of the OSN.

Benevenuto et al. (2009) analysed the role of user behaviour and its properties in OSNs. They undertook their study to better understand user behaviour in order to facilitate better website designs and address traffic issues on the internet. The authors used clickstream data that was collected over a 12-day period from 37024 user activities on four of the most well-known OSNs, namely, Orkut, Hi5, LinkedIn and MySpace. The data was collected via a social network aggregator through which users were able to access multi-OSNs by single authentication. In this way, the authors were able to analyse the four networks by measuring the properties of users such as the total time spent on these OSNs per user, frequency of access to OSN sites, session length and dominant activities. They presented four findings:

- Browsing is the most popular activity for users in OSNs, accounting for 92% of activity;
- Users of Orkut not only interact with friends who are 'one hop' 22% of them had interactions with friends who are two or more hops away;
- The majority of users (63%) access social networks only once during the 12-day period; and

- 51% of users spend no more than 10 minutes on a social network over a 12-day period.

While their initial aim was to analyse user behaviours such as time spent on social networks and so on, their study revealed that the latent activities of users were more significant in OSN analysis than previously realized. Hence, their work is valuable because they changed the view of how to analyse OSNs by considering that user activity is a more important measure than the number of connections. However, they did not present a new practical approach for analysing OSNs based on user behaviour.

Sulo et al. (2010) tried to identify the best temporal resolution for revealing critical changes in a network data stream. They intended to define a temporal resolution that would be able to achieve a balance between reduction of noise and loss of information. To do so, they used variance and the compression ratio measure. Then they defined a goodness measure based on the threshold value in order to evaluate each time window. Their algorithm considered the window size that had a close or equal amount of compression ratio and variance. They tested their proposed method on the Enron, Reality Mining, Barabasi, Haggle Infocomm, Grevy's Zebra and Plains Zebra datasets and the results showed that "the changes of different types and scales need to be analysed at different resolutions" (Sulo et al. 2010: 134). They also noted that interesting network behaviour happened in different time intervals. Although their study proposed a novel framework to define a better temporal resolution for analysing network behaviour, the computational cost of their algorithm was high because it needed to find the variance and compression ratio for all possible time windows. In addition, while the variance showed how the data was distributed around the mean, it did not show the amount of the distribution. Furthermore, their study' results are based on static time interval in order to show loss of information and also their method did not prioritize the time intervals, which, is crucial because OSN behaviour in recent time intervals provides the real picture of user activities and thus communities.

Wilson et al. (2012) argued that gaining an understanding of the role of user interactions was vital for studying social applications such as OSNs, rather than

relying on the social graph approach. Therefore, in their study, they intended to determine whether social links could be used as valid indicators of real user interactions, and if not, they sought to ascertain how these factors could be quantified to form a more accurate model for evaluating socially enhanced applications. In their study, they used interactions among users of a regional Facebook dataset consisting of the 10 largest cities in the world. They considered only explicit interactions and crawled the Facebook dataset with a multithreaded crawler in order to find them. They observed how interactions took place across time and analysed these interactions over a fine-grained time scale. For this purpose, they used a number of measures such as the cumulative distribution function (CDF), CC, joint degree distribution (JDD), assortativity coefficient, graph distance, radius, diameter and average path length. Due to their approach incorporating a wide range of measurements, their study is now widely considered to be a seminal study that contains some valuable findings, which are summarized below:

- Users in OSNs use a public display to show their status. This means that, at a glance, all users seem to look like each other.
- User interactions deviate from a social link pattern according to time, the method of interaction and type of user.
- A user tends to communicate with a small group of theirs friend.
- Interaction graphs reveal different OSN properties when compared to the social graph such as larger graph diameters, lower CC and higher assortativity.
- Although Facebook grows over time, user interactions do not vary with time. This means that over time, even though more users join Facebook, the growth rate in interactions between users does not increase in tandem.
- The majority of user accounts on Facebook can be categorized as a single and weakly connected component (WCC).
- Connections between high-degree nodes are numerous based on the assortativity coefficient (the measure of probability that nodes will have links with other nodes that have a similar degree).

- Low-degree nodes tend to connect with similar nodes and vice versa based on the degree correlation function, which is used to represent the average degree of all nodes that are connected to nodes with a given degree.
- The removal of the super node from a network causes the breakdown of the network.
- For the majority of users, around 70% of their interactions are made with only 20% of their friends.
- Surprisingly, the most active users receive comments from only 5% of their friends.
- Half of all the interactions on the network belong to 10% of the well-connected nodes based on their degree.
- There is a difference in the activities of new and old users.

Thus the Wilson et al. (2012) study is considered to be a significant analysis of OSNs. They conducted a comprehensive study on OSNs in order to show the importance of focusing on user interactions rather than on the social graph. They evaluated some important measures for OSNs such as social degree, CC, degree correlation, roles of supernodes, interactions among friends, the role of the lifetime of users and more. Finally, they introduced the idea of using an interaction graph instead of a social graph for analysing OSNs because they argued that not all social links are equally useful in analysing OSNs. They showed that the social link is not a valid indicator for analysing user interactions as the most active users receive comments from only 5% of their friends. However, their study did not present a method or algorithm for OSN problems such as CD.

Jin et al. (2013) conducted a comprehensive review of state-of-the-art research on user behaviour in OSNs from several perspectives. Their findings threefold: i) with regards to their review of works specifically on Facebook crawling, these studies showed that users tend to interact with half of their friends and often do not have any interaction with half of their friends; ii) latent interactions among users on OSNs generally are significantly more prevalent than visible interactions; and iii) “Latent interactions are passive actions of OSN users (e.g., profile browsing) that cannot be observed by traditional measurement techniques” (Jin et al. 2013: 145). In conclusion,

although their study did not present a method for detecting communities, it did point out that interaction times among users in OSNs could play a role in detecting communities, whereas existing algorithms assigned the same weight to all connections between users (edges).

Wrzus et al. (2013) studied social network changes and life events from a psychological standpoint. They found that "(a) the global social network increased up until young adulthood and then decreased steadily, (b) both the personal network and the friendship network decreased throughout adulthood, (c) the family network was stable in size from adolescence to old age, and (d) other networks with co-workers or neighbours were important only in specific age ranges" (Wrzus et al. 2013:1). Their findings clearly indicate that researchers cannot view an OSN as a static network as many parameters influence the forming of communities over time. Thus their findings are of particular relevance to the current study.

Baagyere et al. (2016) considered the main important features in OSNs, namely, degree distribution, path length distribution and CC. They investigated the above-stated properties in four different datasets, namely, cit-HepTh, the Erdös and Réyni network model, the power-law/scale-free network model and the Watts–Strogatz small world network. They described the major metrics of OSNs as:

- Aggregate network metrics, which is a global metric that can be used to address an entire network system feature such as network density that is computed for each network by Equation. 2.4:

$$D = \frac{2|E|}{|V|(|V|-1)} \tag{2.4}$$

where E is the number of edges and V is the number vertexes

- Node-specific network metrics, which can be used to analyse the position and influence of users in an OSN. To calculate this metric, the authors suggested using five important measures: degree centrality, betweenness centrality, closeness centrality, eigenvector centrality and PageRank centrality.

Nevertheless, their study and other existing studies in this domain did not present a temporal analysis of OSNs, which is the main limitation of these earlier studies.

In the absence of related works that have sought to formulate the OSN growth rate, similar work in this domain is reviewed, as well as existing studies that have tried to identify the characteristics of OSNs.

One of the notable studies in this area is that of Berlingerio et al. (2013), who attempted to identify the turning points in network evolution. Their work was based on a dissimilarity measure that was computed by using the JC between two temporal snapshots of the network. Earlier, Rajaie et al. (2010) tried to estimate the growth of the user population of the Twitter and MySpace OSNs. Their study estimated the account creation time by examining the relationship between the user ID and the time that had elapsed since the user's last login. They introduced the concept of 'tourist nodes', which are short-lived users that are scattered across the whole ID space. Using these users and the last login time made it possible to compute the account creation time of all the users. The authors divided the user population growth into three distinct phases: initial, expansion and maturity. The results showed that the growth of the two studied OSNs was different in each of these three phases. The most important finding of their study was that the popularity of an OSN is strongly correlated with that of other OSNs. In a similar vein, Gonzalez et al. (2013) presented a model to evaluate the user growth of Google+ and compare it with the growth of Twitter and Facebook. They computed the growth of the number of largest connected component (LCC) users in different time frames to determine the number of users who departed or arrived between two consecutive time intervals. Also, Ribeiro (2014) proposed a model to show the growth mechanism of a membership website by focusing on the daily number of users (DAU). Although the above-mentioned works are significant in terms of their analysis of the growth rate of the network, as yet existing studies have not proposed a robust approach for detecting the milestones of OSNs.

In those works that do exist in relation to milestones, Musial et al. (2013) investigated the growth and dynamics of evolution in complex networks. In their

analysis of users' activities, they showed that users tend to be active only in a one-time window that lasts 90 days. The authors also analysed network characteristics such as the CC, shortest path and degree distribution during a time interval of 30 days. However, they did not mention how or why they selected this time interval. Budka et al. (2012) stressed that there is a need to define an accurate time window in order to analyse social networks accurately. To this end, they tried to define the time window in order to enhance the prediction accuracy of node degree, shortest path, triad census distribution, clustering coefficient distribution, Katz score distribution, and betweenness centrality distribution in Enron dataset. In their work, they used a divergence measure (Jensen-Shannon divergence) and considered the time window problem as a constrained optimization problem and evaluated their method based on link prediction problem. However, they use a time windows with step 10 days in their method, and in each time window, they observed a prediction horizon. On the other hand, the Jensen-Shannon divergence works based on mutual information between two variables, their method consider two distribution as two variables in stated divergence. Consideration of prediction horizon and structural measures in its computation is main limitation of their study, where they mentioned that they consider the distribution of the above-mentioned measure.

Another study of note is that of Liu et al. (2013), who investigated the behaviour of active users by constructing a user activity graph that illustrated the activities of users. They used three OSN apps: Hugged, iSmile and iHeart. They categorized the users into three different states: active, alive and quit. An important achievement of their study is that it proposed a model that was able to predict the number of active users based on the number of active, alive and new joining nodes in a week *t*. They showed that users mostly perform their activities in the initial phase of the timespan; around 50% of users are active only in the first 20 weeks. The authors used the static time interval of 1 week to monitor the network. However, the use of static time granularity is the main limitation of their work. Defining a time interval so as to monitor OSN behaviour is an essential task. Guo et al. (2014) presented a mathematical method to analyse the behaviour of Twitter users, which consisted of two models: a Gaussian mixture to show the daily pattern of users' posting behaviour and a logarithm to show the relation between the number of retweets and the posting

times of a specific message. Sekara et al. (2016) in their work showed that knowing the mobility of individuals in a social network can assist in predicting social behaviour during a specific time period. Although their study demonstrated that user activities are different in each specific time period, they did not present a method to formulate this behaviour.

In recent years, many other researchers have tried to introduce new metrics for dynamic networks (Ullah & Lee 2016; Zignani et al. 2014). For instance, Nicosia et al. (2013) analysed time changes in a network and found that the relationship between two nodes is not persistent. Also, they observed that assigning the same weight to all edges can lead to misleading conclusions. To solve these problems, their study proposed a TVG in order to include the time aspect to connections in the network as an additional dimension. They introduced some important characteristics of the temporal network such as reachability, connectedness, component, distance, efficiency, temporal clustering and modularity. Although their work is considered to be a comprehensive study because it defined the major characteristics of dynamic network and emphasized the role of time in temporal networks such as OSNs, they did not define a metric to compute the time frame nor did they discuss the importance of each node in the network based on taking into account the element of time.

Earlier, Ahmed et al. (2010) proposed a new time-based method in order to sample the social network activity graph for reducing the size of this huge network (social network). In this way preserve the characteristics of the network is the main concern of them.They called their method streaming time node sampling (STNS). They introduced three types of sampling: node sampling, edge sampling and topology-based sampling. Their method combines the strenght of edge based sampling so as to preserve path length and node based so as to preserve degree distribution. their algorithm chooses edges such as $e = (v_i, v_j, t)$ where $t \geq st_{vi}$ and $t \geq st_{vj}$ where st_v is a sample time for each node v. “In other words, we only sample activity edges that involve a node after it has been added to the sample” (Ahmed et al. 2010, 5). In their method, for a user to be considered active, the user has post or receive one message. Their method was developed to monitor the activity of users in a network in a constant time period of k weeks, where k was equal to 6, 12 and 18 weeks. They

tested their method on a Facebook and a Twitter dataset and measured three main measures, namely, degree, clustering and average path length,in a daily and monthly time period. The results of their experiment showed that the pattern of the defined time periods (monthly) differs from that for a cumulative time period (i.e., the whole lifespan of the OSN).

The main shortcoming of the works reviewed in this section is that they did not define a time interval based on OSN behaviour changes. Rather, they defined a series of snapshots based on changes in the number of nodes and edges (without defining a robust criterion) or a static time window lasting, for example, for 1 day, 1 month or 1 year. Therefore, in light of the above, the current research study suggests the addition of two new measures to the OSN behaviour model: the exciting time period, which is defined in section 4.3, and the milestone, which is defined in section 4.2.

2.3 USER ATTRIBUTES IN OSNS

User attributes are of fundamental interest to this research as the proposed CD algorithm is based on user attributes. For the purpose of this study, the literature on user attributes is divided into two categories, user weight and geo-location.

2.3.1 User weight

Definition 2.5 User weight refers to the influence of a user in an OSN, which attracts other users to him/her such as degree centrality (Freeman, 1978).

In recent years, many researchers have sought to identify the influential users and to formulate user weight (Erlandsson et al. 2016; Hangal et al.2010; Heidemann et al.2010; Ilyas & Radha 2011; Jianqiang et al.2017; Shafiq et al. 2013; Zafarani et al. 2015). This is because, in an OSN, each user has a specific weight, which refers to the influence of the user on the OSN, and the weight of each user is different. A user's weight is a key indicator of the user's influence on the OSN; where the greater the weight of the user, the more influence that user has on the OSN as compared to other users. An accurate understanding of the role of users is fundamental to the solving of many OSN domain problems, such as CD, event detection and marketing.

In 1978, Linton C Freeman introduced the concepts of node centrality, closeness centrality and betweenness centrality (Freeman 1978). He used these centrality concepts as measures to identify the influential nodes in social networks. However, in his work the social network was frozen in time and was represented in graph form. Two decades later, Brin and Page (1998) introduced a method to rank the pages in a website, which attracted a lot of attention in later years. Indeed, their method was fundamental in the development of the Google Inc. search engine for ranking websites and web pages. Their idea has also been used to find influential nodes in social networks.

More recently, Zafarani et al. (2014) reviewed seven methods that used the different types of centrality to identify the important nodes in OSNs. These types of centrality were degree centrality, eigenvector centrality, Katz centrality, PageRank, betweenness centrality and group centrality, each of which is defined below.

- Degree centrality: This metric ranks the nodes based on their connections (see Equation. 2.5) (Freeman, 1978):

  $C_d(v_i) = d_i$ (2.5)

  where, d_i is the degree of v_i.
- Eigenvector centrality: This measure considers the importance of neighbours (see Equation. 2.6):

  $C_e(v_i) = \frac{1}{\lambda} \sum_{i=1}^{n} A_{j.i}\, C_e(v_j)$ (2.6)

  where λ is the eigenvalue and A is the adjacency matrix of a graph.
- Katz centrality: Eigenvector centrality has a limitation in that it does not have the capability to recognize important nodes in a directed graph. Katz centrality was developed in order to solve this problem (see Equation. 2.7):

  $C_{Katz}(v_i) = \alpha \sum_{i=1}^{n} A_{j.i}\, C_{Katz}(v_j) + \beta$ (2.7)

  where α is controlled value and β is the bias term that avoids the zero centrality value.

- PageRank: Katz centrality and eigenvector centrality have a limitation when a node in a directed graph with high centrality passes all its centrality along all its outgoing links (Brin & Page , 1998). The PageRank measure was proposed to address this issue (see Equation. 2.8).

$$C_p(v_i) = \alpha \sum_{i=1}^{n} A_{j.i} \frac{C_p(v_j)}{d_j^{out}} + \beta \quad (2.8)$$

- Betweenness centrality: This measure refers to nodes that establish connections between nodes (see Equation. 2.9) (Freeman, 1978):

$$C_b(v_i) = \sum_{s \neq t \neq v_i} \frac{\sigma_{st}(v_i)}{\sigma_{st}} \quad (2.9)$$

where σ_{st} is the number of shortest paths from node *s* to *t* and $\sigma_{st}(v_i)$ is the number of shortest paths from s to t via v_i.

- Closeness centrality: In this measure, it is assumed that the more central nodes can reach nodes more quickly than other nodes (see Equation. 2.10) (Freeman, 1978):

$$C_c(v_i) = \frac{1}{\bar{I}_{v_i}} \quad (2.10)$$

where $\bar{I}_{v_i} = \frac{1}{n-1} \sum_{v_j \neq v_i} I_{i.j}$ (2.11)

- Group centrality: This measure is defined “as the number of nodes from outside the group that are connected to group members” (Zafarani et al. 2014: 85).

All of the above measures consider the weight of nodes in a static network and therefore do not offer a solution for OSNs because not only are the weights of nodes in an OSN different in each time interval, according to Wilson et al. (2012) the later time intervals are more important than the other time intervals. Hence a new measure needs to be developed that takes this fact into account when assigning a weight to a node.

In recent years due to the emergence of OSNs, the issue of finding influential users has attracted much more attention. Hangal et al. (2010) carried out a study in order to find the best path from user A to user B, as the authors assumed that the shortest path in an OSN is not the best measure to identify the best search solution as

the weights of friends are not equal in OSNs. So they tried to find influential users in networks by defining the “influence from A to B, influence (A, B) as the proportion of B’s investment on A. let invests (B, A) be the investment B makes on A” (Hangal et al. 2010: 3). This investment pointed to the number of interactions being of importance. In other words, a user is influential if he/she has influence on many other users, as they explain in the following: “A high influence from A to B corresponds with a high probability that B will forward A’s message to the desired target, whether that be the end goal or another intermediary along the path” (Hangal et al. 2010: 3). They tested their methods on the DBLP and Twitter datasets and their results showed that the shortest path for searching in OSNs is not effective and it is not the best path; rather, the best path considers influential users. In conclusion, although their study was conducted on OSNs, the authors did not address the role of time and differences of interaction in each time interval.

Trusov et al. (2010) proposed a new method to identify the influential users in an ego-centred network from the marketing perspective. Their method used a multi-layer mechanism to count the number of logins to a website. If the website usage by the friends connected to a specific member increased in line with the usage by that member, then their method identified that person as influential. Conversely, if a member’s usage went up or down and the usage among the people connected to him/her did not rise and fall in line with that usage, their method classified that person as uninfluential. However, they did not propose a framework for defining the time interval, which means that they did not consider which time interval in the network should be evaluated by their method. Also, their method was only developed for the marketing domain and the page viewing activities of users, so it is cannot be generalized to other OSNs.

Erlandsson et al. (2016) proposed a method based on association rules to identify the influential users in social media. Essentially, their method worked by finding a common interest among users. For example, if users A, B and C share a common interest, and users A and B make a comment on this topic, then there is a chance that user C will also comment on that topic. However, they did not consider

the time interval during which social media content should be observed. Hence, their method is frozen in time and it monitors the network only in one time frame.

The above-mentioned studies did not consider the time in their proposed methods and instead presented static methods. However, Shafiq et al. (2013) with regard to the significance of strong nodes and their roles in OSNs, tried to identify four types of nodes (user groups), which they called introvert leaders, extrovert leaders, followers and neutrals. In their proposed model, introvert leaders are followed by many individuals even though these leaders interact very little with their followers. On the other hand, extrovert leaders have a high number of interactions with their numerous followers. Followers usually interact with their friends based on mutual relations, while neutrals interact with their friends regardless of mutual interactions. Hence, Shafiq et al. (2013) used user interaction information instead of content-mining, which is useful in cases where text content is not available. They called their method the longitudinal user centred influence (LUCI) method and computed two coefficients: the ego and the network coefficient. They stated that: "The ego coefficient tries to quantify the correlation between the past and future outward interaction of users. The network coefficient tries to quantify the correlation between users' past inward interaction and future outward interaction" (Shafiq et al. 2013: 2). They tested their method on the Everythings2 and Facebook datasets. The key findings of their study were as follows: i) followers are part of closely connected communities and have the highest CC compared to the other user groups; ii) extrovert leaders have more friends than introvert leaders; and iii) extrovert and introvert leaders have shorter path lengths than other user groups. The main limitation of their work is that they only considered inward and outward interactions in constant time period where authors stated "In our experiments, the duration of time periods is set to be one month for Facebook data and six months for Everything2 data" (Shafiq et al. 2013; 621). Also, their method did not assign a weight to the edges (interactions) over time, which is a drawback because the most recent interactions are more important than those made in the past in terms of establishing the connections among users.

Zhao et al. (2017) developed a method to identify influential nodes by using the topological connections among neighbours and the number of neighbour nodes. They used neighbour nodes to apply the bridge concept to network analysis. Then they

presented a decreasing function to compute the local CC which they called the coefficient of local centrality. Their study assumed that the network was unweighted, but this assumption is not relevant to the real world. Nevertheless, their idea of using the number of CCs to determine whether a user is at the centre of network in which neighbour connect with each other is interesting. However, it only works for a static environment and does not take into account the dynamic nature of social networks.

Wu et al. (2018) proposed a new method based on a topic-behaviour influence tree in order to identify influential users in social networks. Their method was based on the correlation between messages and behaviours from two aspects: messages→topic and topic→social behaviour. They defined six types of relationship in their experimental network: similarity of messages, hashtag title similarity, retweet, reply, mention and follower/followed. They considered user *u* as influential if user *u* was interested in topic *z* and influenced other users' opinions about topic *z* and evaluated this relationship by taking into account the minimum propagation time path of *u* to each affected user. One of the main significant findings of their study was that influential users have different influences in different communities. This finding underpins one of the hypotheses of the current study, as this study attempts to compute the user influence based on the user position in each community.

In addition, the current study also considers that this influence can change over time. In other words, recent time intervals have higher priority over other time intervals. However, previous studies, such as those by Hangal et al. (2010), Shafiq et al. (2013), Zafarani et al. (2014), Erlandsson et al. (2016), Nan et al. (2016), Munger and Zhao (2015) and Jianqiang et al. (2017) did not give priority to the recent time intervals in their attempts to identify the relative importance of users. Thus defining a metric to compute a time interval that covers the maximum changes in the network is a major shortcoming of such earlier studies. In order to conclude and summarize this section, Table 2.1 shows the characteristics of the existing user weight computation methods.

Table 2.1 Existing methods for user weight computation

Authors	Method	Time-based	Assign priority to time interval
Freeman, (1978)	Degree centrality	-	-
Freeman, (1978)	Betweeness centrality	-	-
Sergey Brin and Larry Page (1998)	Page rank	-	-
Trusov et al. (2010)	multi-layer mechanism to count the number of logins	-	-
Shafiq et al. (2013)	interaction information	√	-
Erlandsson et al. (2016)	association rules	-	-

2.3.2 Geo-location

A variety of measures can effect on community forming among nodes (in this case, human OSN users), such as job, interests and activities. However, some algorithms use geo-location because it has been proved that users who are near to each other are more likely to form relationships with each other than with users who are further away from them (Allamanis et al. 2012; Cho et al. 2011; Cranshaw et al. 2010; Huang & Liu 2015; Kaltenbrunner et al. 2012; Lengyel et al. 2013; Liben-Nowelly et al. 2005; Scellato et al. 2011).

The majority of studies on the OSN phenomenon were published between 2008 and 2012, when a rapid growth in the membership of these sites occurred. In those years, the number of Facebook and Twitter members reached one billion and 200 million, respectively, and WeChat attracted around 200 million users in just one year (Statista, 2017). With the advent of these OSNs, people were able to find each other independent of their actual geo-location. So, many researchers decided to investigate the effect of distance on OSNs. The peak of these endeavours is limited to the above-stated period which saw the publication of several important works including, in date order, Gonzalez et al. (2008), Lambiotte et al. (2008), Cranshaw et al. (2010), Cho et al. (2011), Scellato et al. (2011) and Kaltenbrunner et al. (2012). After that, a few studies were undertaken to analyse the importance of geo-location in OSNs in other disciplines, such as psychology (Chorley et al. 2015), urban studies

(Herrera-Yagüe et al. 2015), CD (Huang & Liu 2015) and link prediction (Xu-Rui et al. 2015). However, the absence of a solid analysis of the effects of user mobility and other user attributes on OSNs is noticeable.

Currently, the roles that geo-location and distance play in OSNs are of great interest to researchers as these behaviours affect a range of domains. For instance, some researchers have attempted to discover the effect of a LBSN on user behaviour, including Chorley et al. (2015), who conducted a study to explain individuals' personalities by using a LBSN. On the other hand, Xu-Rui et al. (2015) introduced an algorithm to predict friendship formation in OSNs. Others have attempted to determine the effect of user behaviours such as mobility and geo-location on OSNs (Cho et al. 2011; Kaltenbrunner et al. 2012;). Similarly, the aim of the current study is to discover how user geo-location and user movement affects OSNs ties.

Many research studies have been conducted in order to explain the effect of distance on OSNs' ties, including Cho et al. (2011), Huang and Liu (2015), Lengyel et al. (2015), Scellato et al. (2011) and Xu-Rui et al. (2015). Some studies have concluded that the effect of distance on ties in OSNs is negligible (Cho, Myers, & Leskovec 2011), whereas others have stated that it is significant (Huang & Liu 2015; Liben-Nowelly et al. 2005). The latter two studies also emphasize the role of user location in OSNs. Previous studies have mostly tried to show the effect of distance on OSNs, but other attributes such as number of friends, number of interactions and user lifespan have not been explored in any detail. However, as noted some years ago by Scellato et al. (2011), the exact relation between OSN ties and distance is unclear, and to date it still needs to be fully elaborated. Nevertheless, there seems to have been no significant works that have attempted to formulate the relation between distance and OSN ties. Therefore, a review was conducted of the related works that have considered the effect of distance on users' connections in OSNs.

Nevertheless, there was some success in demonstrating the probability function of friendship and distance, as demonstrated by the following works:

Liben-Nowelly et al. (2005) experimented on the LiveJournal network and observed that when the distance between users increased, the probability of making a

connection decreased. They showed that this relation can be described as $P_{geo} \cong d^{-1} + \varepsilon$. Lambiotte et al. (2008) considered distance when introducing the gravity model for mobile communication networks. They showed that distance has an inverse proportional relationship with pairs of individuals' connections. However, they did not describe how this relation worked. They showed that this relation can be described as $P_{geo} \cong d^{-2}$. Allamanis et al. (2012) showed that the probability of making a new edge is a function of distance, where $P_{geo} \cong d^{-\propto}$ and α =.6, although they stated that the exact function for this relation was still under debate.

More recently, Lengyel et al. (2015) studied the effect of distance on ties in a Hungarian OSN. They analysed this problem at two levels: individual-level links and town-level links. They found that the effect is less at the individual level, but that the weights of town-to-town ties are strongly correlated to geo-location. A significant work in this domain is that by Kaltenbrunner et al. (2012), who investigated the effect of geographic distance on online social interactions, specifically the effect of distance on friendship and interaction. Their results showed that the relationship between user connections and geo-location is very strong, but the geographic location of the users does not affect the amount of interactions. Also, they discovered that OSN users interact with a small subset of friends, a finding also reported by Wilson et al. (2012). Earlier, Cranshaw et al. (2010) showed that there is a relation between the entropy of the location that the user visits and the number of that user's connections in the network. They introduced a location entropy measure that showed the diversity of unique visitors to a location. The authors also attempted to deploy other measures to describe each edge between users, such as intensity and duration, location diversity, specificity and structural properties. They found that users who visit highly diverse locations tend to have more friends in social networks. Prior to that, Gonzalez et al. (2008) showed that, despite the diversity of locations visited, human beings follow a regular, simple pattern to return to a few highly frequented locations.

However, Cho et al. (2011), in their significant work on the relation between friendship and mobility, showed that the relation between friendship and distance can be negligible. In their study, they deployed the Brightkite and Gowalla LBSNs and found that the impact of friendship on mobility twice as strong as the effect of

mobility on making friendships. Also, there are limits in using friendship alone to predict mobility. Overall, they concluded that the relationship can explain about 30% of the movement in LBSNs, while periodic movement can explain about 50% to 70%. Also, Scellato et al. (2011) stated that there is a weak positive relation between the number of friends and the average distance between them all. They made two null models (geo and social) to investigate the geographic and social properties of OSNs. They found that the socio-spatial structure cannot be explained by only one of these models. The authors also showed that the probability of a connection existing between users increases for short distances.

However, there is no available work in this area that has also considered in addition to distance the effects of other user attributes, such as user weight and density of interaction on OSN ties. In fact the absence of a solid relationship between social network ties and distance is a limitation of the above-mentioned works, which the current study seeks to address. To conclude and sumamrize this section, Table 2.2 shows the existing methods that have been proposed for the computation of the probabilty of OSN ties.

Table 2.2 Existing methods for the computation of the probabilty of OSN ties

Author	Proposed formula	Distance based	User attribute based
Liben-Nowelly et al. (2005)	$d^{-1} + \varepsilon$	√	-
Lambiotte et al. (2008)	d^{-2}	√	-
Allamanis et al. (2012)	$d^{-\propto}$ and $\alpha = .6$	√	-

2.4 COMMUNITY DETECTION ALGORITHMS

Before proposing a new output structure which is one of the main objectives of this research, this section first provides an overview of the three output structures that can be produced by CD algorithms because it is important to understand the problems that may arise from each type of output structure, such as search speed and high memory

consumption. Second part of this section provides a critical review of existing CD algorithms and methods.

2.4.1 Output Structure of Community Detection Algorithms

The basic form of an OSN consists of nodes and edges, where the nodes are represented as $(n_1, n_2, \dots, n_k)$ and the edges are represented as $(e_1, e_2, \dots, e_l)$, where k is the number of nodes and l is the number of edges. The application of a CD algorithm to an OSN leads to one of three output structures: the graph, dendrogram or tree.

a. Graph structure output

The graph structure is the main form of CD algorithm output. According to Gallier (2010) and Haggard et al. (2006), the graph is a data structure and consists of a set of vertexes and a set of edges. In the graph structure, the edges between the nodes can be directed or undirected. Each graph *G* can represent with a matrix number of rows and columns represents the node ID, where each cell takes the value of 0 or 1, and when an edge is between two nodes, then the relevant cell is 1, otherwise 0. The algorithms that are generally used to search graphs are the breadth-first search (BFS) and the depth-first search (DFS) algorithms. However, some other techniques can be used to optimize the search problem in a graph, such as the greedy best first and A^* algorithms. However, these techniques are beyond the scope of this study as this research does not intend to present an optimization algorithm for graph traversal.

However, this structure has two main limitations. First, there is memory complexity as the graph contains all of the edges which leads to a high search time. It is important to recognize this problem in CD research because in some domains such as cybersecurity it needs to traverse a specific community.

Every community can be shown in a graph (see Figure. 2.4):

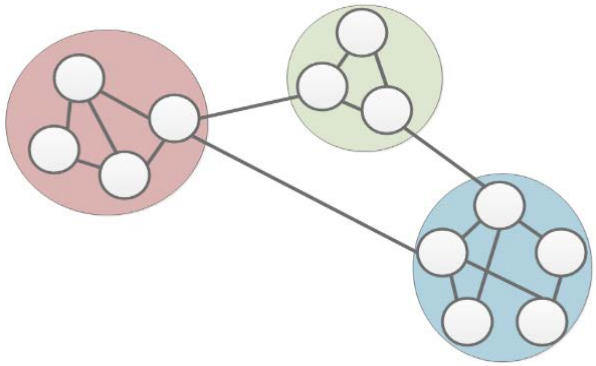


Figure 2.4 Example of graph structure

b. Dendrogram structure output

At first glance, a dendrogram seems like a tree structure. However, a dendrogram is hierarchical structure that shows how communities merge or split (Newman & Girvan 2003), which makes the dendrogram more complex than the tree structure. Moreover, the dendrogram is conceptual and should not be considered a data structure as such. The communities in each level of the dendrogram are revealed by cutting the edges based on a modularity value (see Figure. 2.5):

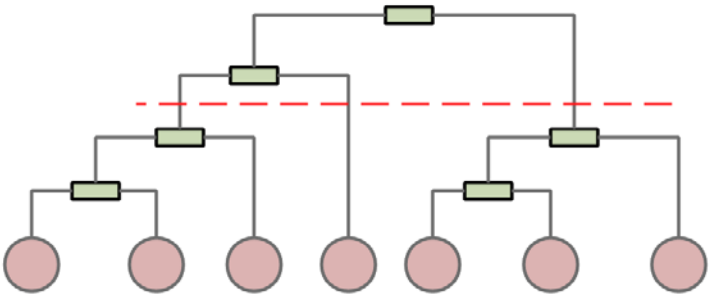


Figure 2.5 Example of a dendrogram structure

c. Minimum spanning structure output

The minimum spanning tree (MST) can be categorized as a hierarchical data structure. The MST is sub-graph of G that is connected. Like other types of tree structure, the MST is acyclic and includes all vertexes. The difference between a MST and other types of tree structure is that the MST has the lowest cost. The methods that are generally used to convert a graph into a MST are Kruskal's algorithm and Prim's

algorithm which are both greedy MST algorithms. The MST output for CD algorithms contains a set of MSTs that form a forest, which means that there are no connections between any of the communities (see Figure. 2.6):

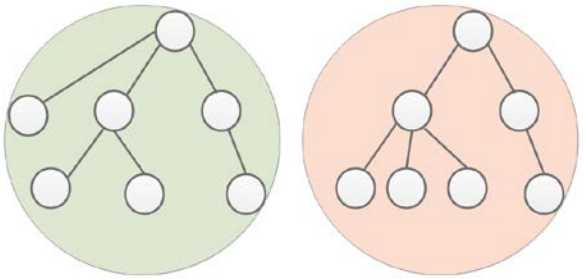


Figure 2.6 Minimum spanning tree structure

A directed acyclic graph (DAG) shows the directed acyclic flow in a graph or tree.

Table 2.3 shows the differences in the properties of the graph and tree structures. It can be seen from the table that the tree is less complex than the graph. This is because cycles and loops do not exist in the tree structure.

Table 02.3 The differences between the graph structure and tree structure

properties	**TREES**	**GRAPHS**
Loop	There are no circuits or loops in trees	There are loops and circuits in graphs
Root concept	There is only one root and a child can have only one parent	There is no root; the concept is not applicable in graphs
Directed Acyclic Graph (DAG)	Trees fall into the DAG category	Graph falls into the DAG category
No. of edges	Trees always have n-1 edges	There is no relation between edges and nodes in graphs
PathS	There is only one path available between two specific nodes in trees	There is more than one path that can be found between two specific nodes in graphs

to be continued…

…Continuation

TraversAL	Trees can be traversed post-order, in-order and pre-order, where a tree traversal is a type of graph traversal	The breadth-first search (BFS) and depth-first search (DFS) algorithms can be utilized to traverse graphs
Parent-Child relationship	This relationship exists in trees	There is no such relationship in graphs
Connection Rules	Trees encompasses some rules in order to make connection between nodes	There is no any rules in order to make connection between each pairs of nodes
Complexity	As there are no cycles or loops in trees, trees are less complex than graphs	As cycles and loops are present in graphs, graphs are more complex than trees
Different Types	There are many types of tree, such as the search tree, heaps binary tree, binary tree and AVL tree.	There are only two types of graph: directed and undirected
Applications	Trees can be used in searches such as tree traversal and binary search and also in sorting	Graphs can be used in algorithms, graph colouring, job scheduling and the colouring of maps
Model	Trees fall into the hierarchical model category	Graphs fall into the network model category

2.4.2 Community Detection Algorithms

This section reviews the existing CD algorithms which can used to detect communities in OSNs. Before to review the CD algorithms three important notations are described which is necessary for defining each algorithm. Generally, there are three notations that are used to interpret problem complexity, such as NP (non-polynomial) hard, NP complete and Big O, where the type of NP represents a set of decision problems that are solved by a non-deterministic algorithm in polynomial time. The two types of NP determination are defined as follows (Duarte et al. 2018; Talbi 2009; Hamalainen, 2006):

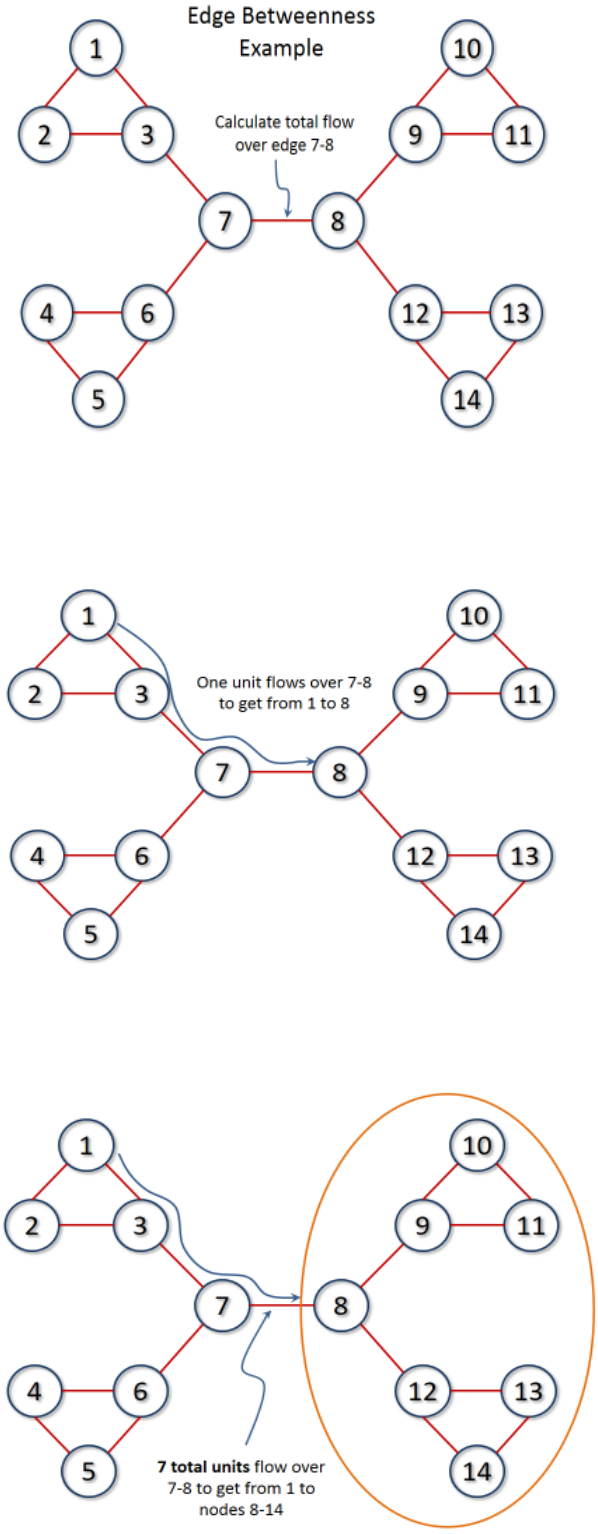
- **Definition 2.6 – NP hard**: A problem *P* is categorized as NP hard if it can be solved by a NP-complete problem such as *Z*, where *Z* is reducible to *P* in polynomial time.
- **Definition 2.7 – NP complete**: A problem is NP complete if it can be solved in polynomial time by a non-deterministic Turing machine.
- **Definition 2.8 – Big-O notation**: "An algorithm has a complexity $f\ (n) = O(g(n))$ if there exist positive constants n_0 and c such that $\forall n > n_0, f\ (n) \leq c \cdot g(n)$" (Talbi 2009:9).

As mentioned earlier, in recent years, due to the emergence of dynamic networks such as OSNs, static algorithms have become incapable of detecting communities that are relevant to the real world because basically the same weight is applied to the nodes and edges in a static network, but this is not an appropriate approach to adopt when seeking to identify communities in a dynamic network. Therefore, many researchers have attempted to overcome this drawback by developing a variety of dynamic algorithms (Aston & Hu 2014; Hecking et al. 2013; P. Nguyen et al. 2014). However, most have tried to develop approaches to CD in dynamic networks by dividing the network into a series of snapshots and then applying the modularity metric used by the GN algorithm to each snapshot.

As mentioned above, the basic algorithm for detecting communities is the GN algorithm, which was proposed by Newman and Girvan (2003). The GN algorithm is based on the maximum betweenness between the nodes in each community and the lowest interconnection between the nodes in different communities. To evaluate the results of the GN algorithm, the authors also introduced the modularity measure ($Q = \sum_i (e_{ii} - {a_i}^2)$), where $e_{ii}$is the percentage of edges in module *i* and a_i is the percentage of edges with at least one end in module *i*.

Edge betweenness is defined as how many times an edge is used to reach other nodes and it is used to calculate the shortest path between nodes or vertices.

The edge betweenness value is computed for all edges (see Figure. 2.7).



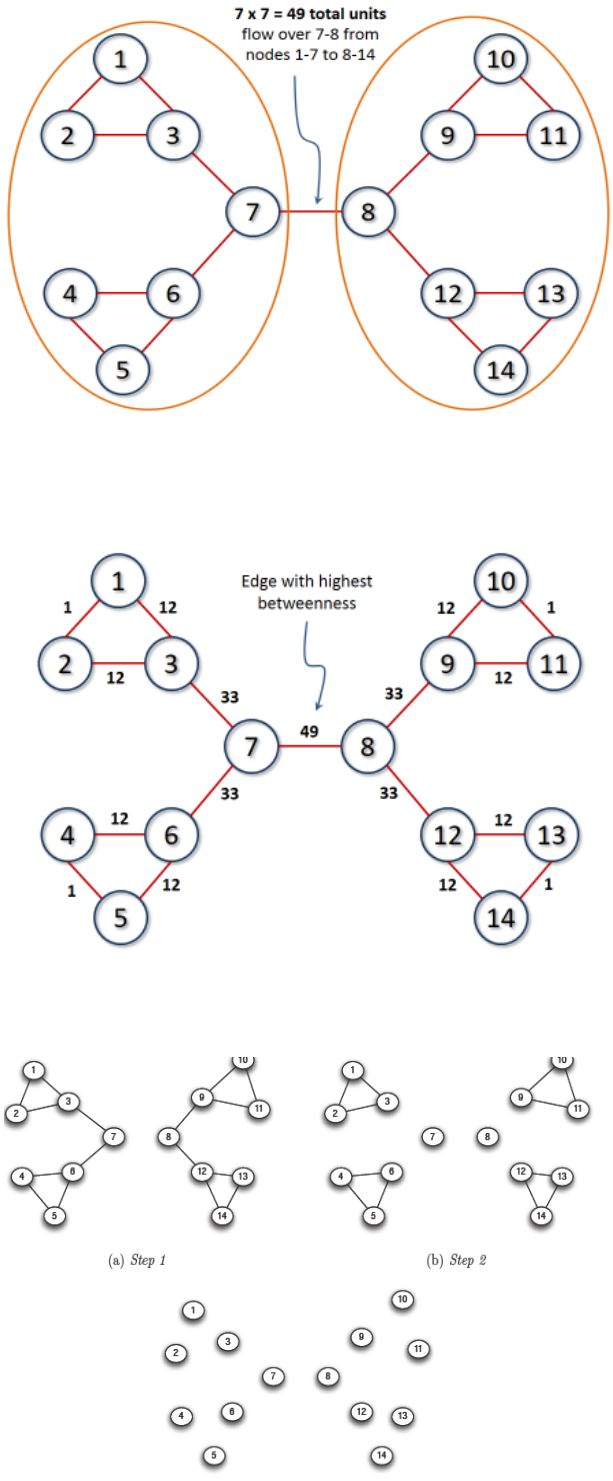


Figure 2.7 Examples of edge betweenness calculations for the GN algorithm (McCown, 2017)

Modularity is computed as the best division such that the greatest number of edges are within communities and the least are between communities (see Equation. 2.12) (Newman, 2004):

$$Q \text{ (modularity value)} = \sum_i (e_{ii} - a_i^2) \quad (2.12)$$

where:

e_{ii} = *percentage of edges in module i*

e_{ii} = $|\{(u, v): u \in Vi, v \in Vi, (u, v) \in E\}| / |E|$

a_i = *percentage of edges with at least one end in module i*

a_i = $|\{(u, v): u \in Vi, (u, v) \in E\}| / |E|$

Note that for high modularity there need to be more edges within the module than would occur by chance

The modularity measure was widely used by some researchers for several years after the publication of the GN algorithm in 2003. However, both modularity and the GN algorithm have some problems: i) the modularity function works on the quantity of connections, which means that it only computes the connections between the nodes in each community and only considers edges between nodes that have the same weight; ii) the algorithm is very complex (O (n^3)); iii) the space complexity is high at O(n^2); and iv) the output of the algorithm is in the form of a dendrogram, which considers all the edges and thus cannot reduce network complexity and has a high computational cost. To address these issues, Newman (2004) presented a new algorithm with a better modularity value than the GN algorithm and which was also faster. The time complexity and modularity of this new algorithm was evaluated by applying it to the karate club dataset, an American college football team, a jazz musician collaboration and a network of scientists in all branches of physics. Later, Liu et al. (2011) proposed a new measure called communicability (*C*) in order to identify communities, in which the value of *C* is the average of the summation of all the differences of inter-community density and intra-community density. The authors also expressed the view that the CD problem could be considered as a problem whose solution requires the finding of the best partition for a network such that its *C* value can be maximized. In their work, they assume that a large *C* value is the best measure

for portioning a given network. Thus, in their formulation, *n* is the number of nodes, and *m* is the number of modules (sub-graph) (see Equation. 2.13):

$$C=\frac{1}{m}\sum_{i=1}^{m}[\delta_{in}(G)-\delta_{out}(G)]=\frac{1}{m}\sum_{i=1}^{m}[\frac{G_{in}}{n_G(n_G-1)/2}-\frac{G_{out}}{n_G(n-n_G)/2}] \quad (2.13)$$

In Equation. (2.13) above, $\delta_{in}(G)$ is the ratio between the number of internal edges of the partition of G and the number of all possible internal edges. Similarly, the definition $\delta_{out}(G)$ is the ratio between the number of edges running from the vertices of *G* and the rest of the network and the maximum number of inter-community edges possible. Hence the value of the function is limited between -1 and 1 (Liu et al. 2011). However, it should be noted that Liu et al. (2011) considered only the edges between nodes; they did not consider the time aspect or the interaction volume in their work. Also, the output of their method was in the form of a graph.

Earlier, Clauset et al. (2004) developed an algorithm, which they called the Clauset-Newman-Moore (CNM) algorithm (and known as Fastgreedy), in order to optimize the computational cost, reduce the time complexity and detect meaningful communities. The authors claimed that the algorithms developed in previous works were slow because of their structure. They therefore presented a new method based on changing the modularity value in their algorithm. They used the ΔQ matrix instead of the adjacent matrix to save memory and time. They also used three data structures for their algorithm: (i) the sparse matrix for ΔQ, (ii) the max-heap and (iii) the ordinary vector array. They applied their proposed algorithm to the amazon.com purchasing network and the result was evaluated by calculating the modularity. They found that using the differences between the modularity value leads to the discovery of a local optimization solution. However, they did not discuss in which step the ΔQ should be computed.

In the same year Pons and Latapy (2005) presented an algorithm based on random walk, which they called Walktrap. The idea for their algorithm was motivated by the fact that random walk tends to become trapped when it is applied to a graph. So, for their proposed algorithm, they assumed that the nodes in the community are dense with respect to the number of edges. Therefore, in the random walk, the probability of reaching from node *i* to *j* is computed, for nodes in the same community

the probability value is high, they used two probability definition such as stationary distribution and reversibility. However, this assumption is not always true. In their work, the probability function is calculated based on the node degree, which implies that the output of their algorithm is a dendrogram, or in other words, a tree-type structure. At the time of publication, the Walktrap method was a breakthrough in the field of CD algorithms in terms of addressing the issue of time complexity. However, it did not consider user attributes, was not time based and, above all, it was based on the number of edges.

Backstrom et al. (2006) proposed a method to detect groups in a large-scale network. The key assumption that underpinned their work was that an individual will join a group if their friends have already joined that group. A notable aspect of their work is that they considered group changes over time in their analysis of a social network from the viewpoints of community membership, community growth and movement between communities. They made a decision tree based on the number of features and applied a decision tree technique over a given period time of 4 months. However, their consideration of a constant time period is a drawback of their work on social network communities as it is now known that user behaviour is not the same in each time interval. In same year Newman (2006) introduced Leading eigenvector algorithm. First, author computed the eigenvectors and eigenvalues of the modularity matrix in following his method computes leading eigenvector of the modularity matrix, this leads to improve modularity and iterate until the maximization not possible.

One year later, the label propagation algorithm (LPA) was proposed by Raghavan et al. (2007). Their algorithm did not consider the betweenness between the nodes; rather, it considered the number of neighbours. Therefore a node is only given a label if that label is the most common among its neighbours. The basic idea behind their method is that each node in the network has a label that can be propagated to other nodes. In fact, a node can change its label based on the number of neighbouring nodes that have the same label. Thus, this algorithm can be categorized as an edge-based algorithm. The authors evaluated their method by using a modularity metric. However, Rezai et al. (2015) reported that, in some cases, the LPA has a drawback in