DE MONTFORT UNIVERSITY

Encouraging Inactive Users towards Effective Recommendation

by

Eseosa Oshodin

A thesis submitted in partial fulfillment for the degree of
Doctor of Philosophy

in the
Faculty of Technology
School of Computer Science and Informatics
Centre for Computational Intelligence
De Montfort University, Leicester

March 2019

Acknowledgements

Firstly, I would like to thank the Almighty God for giving me strength in carrying-out and completing this research. There were several times I almost gave up on completing the research but God gave me hope to continue till the end.

Secondly, I would like to acknowledge and appreciate my first supervisor, Professor Francisco Chiclana for his support and guidance throughout the research work. His involvement also encouraged me to have more confidence towards achieving reasonable results from the research. I would also extend my thanks to my previous second supervisor, Dr Samad Ahmadi who before he left the University, worked effectively with my first supervisor to provide good advise on my research work. Another supervisor, Dr Archie Khuman who stepped in for Dr Samad Ahmadi as my second supervisor worked tirelessly to review my thesis and provide me with feedback.

I would also like to thank all my PhD colleagues and staff of the Centre for Computational Intelligence (CCI) for providing a friendly, supportive study environment and good facilities to support the completion of my research work.

Lastly, special thanks to my family members for their love, support and care. I would like to appreciate my parent for all the financial support they provided towards my Education. I will also like to appreciate my wife for her love, care and patience exhibited throughout my Master's degree and PhD study.

Abstract

Disagreement amongst users in a social network might occur when some of them have different opinion or preferences towards certain items (e.g. topics). Some of the users in the social network might have dynamic preferences due to certain situations. With these differences in opinion amongst the users, some of the users might decide to become either less-active or inactive in providing their opinions on items for recommendation processes to be possible or effective. The current state of the users will lead to a cold-start problem where the recommender system will be unable to find an accurate preference information of the users for a recommendation of new items to be provided to them. It will also be difficult to identify these inactive or less-active users within a group for the recommendation of items to be done effectively.

Attempts have been made by several researchers to reduce the cold-start problem using singular value decomposition (SVD) algorithm, but the disagreement problem amongst users will still occur due to the dynamic preferences of the users towards items. It was hypothesized in this thesis that an influence based preference modelling could resolve the disagreement problem. It is possible to encourage less-active or inactive users to become active only if they have been identified with a group of their trustworthy neighbours. A suitable clustering technique that does not require pre-specified parameters (e.g. the number of clusters or the number of cluster members) was needed to accurately identify trustworthy users with groups (i.e. clusters) and also identify exemplars (i.e. Cluster representatives) from each group. Several existing clustering techniques such as Highly connected subgraphs (HCS), Markov clustering and Affinity Propagation (AP) clustering were explored in this thesis to check if they have the capabilities to achieve these required outputs. The suitable clustering technique amongst these techniques that is able to identify exemplars in each cluster could be validated using pattern information of past social activities, estimated trust values or familiarity values. The proposed method for estimating these values was based on psychological theories such as the theory of interpersonal behaviour (TIB) and rational choice theory as it was necessary to predict the trustworthiness behaviour of social users. It will also be revealed that users with high trust values (i.e. Trustworthy users) are not necessarily exemplars of various clusters, but they are more likely to encourage less active users in accepting recommended items preferred by the exemplar of their respective cluster.

List of Publications

In the duration of this PhD research, the following works have been published during the research while some other publications are still being prepared based on the findings discussed in the thesis.

Conference Papers

1. Oshodin, E. and Chiclana, F. Effect of Influential users on recommendation. *In SAI Intelligent Systems Conference(IntelliSys), IEEE, 2015, pp.731-732*.
2. Oshodin, E., Chiclana, F., and Ahmadi, S. Social Trust in a Familiar Community. *Research and Development in Intelligent Systems XXXII. Springer, 2015, pp. 119-132*.

Contents

List of Figures

List of Tables

Dedication

Dedicated to my parents, wife and kids. With your care and support, I was able to achieve my dream. Love you all.

Chapter 1

Introduction and Background

1.1 Brief Preview

As we live in the age where isolation is no longer an option, it is vital to understand the behaviour of users in a community [1]. A community can be considered to be a system environment where interaction or engagement between users of the system occurs. Most systems usually have a network structure where entities (such as events, items, people or animals) represent a set of nodes and their relationship represent the edges (i.e. ties or links) in the network [77, 211]. Engagement or interaction between users in a social network is expected to improve knowledge growth within the network [81].

It is impossible to achieve and maintain success within a network without links or relationships, as there is the need for information flow and updates to assist each user in their daily decision making. For instance, residents within a street may not be aware initially of a planned power outage for certain construction work scheduled to take place during a period of time in their street. But a resident member of that street who could be a staff or representative of the construction company is expected to pass on information on the scheduled work to other members of the street for them to plan ahead in having an alternative power supply. Also, if the company decides to suspend the power outage, this same staff of the company can also provide updates to residents of the street. Even though the representative of the construction company is not close to some neighbours, he will trust other neighbours that are close to these neighbours to deliver any information to them. This type of social network from the example is known

[1] A group of entities created from a network based on their similarity.

as an ***advisory/knowledge network*** [95] where individual members benefit and receive information within the network to make their decisions.

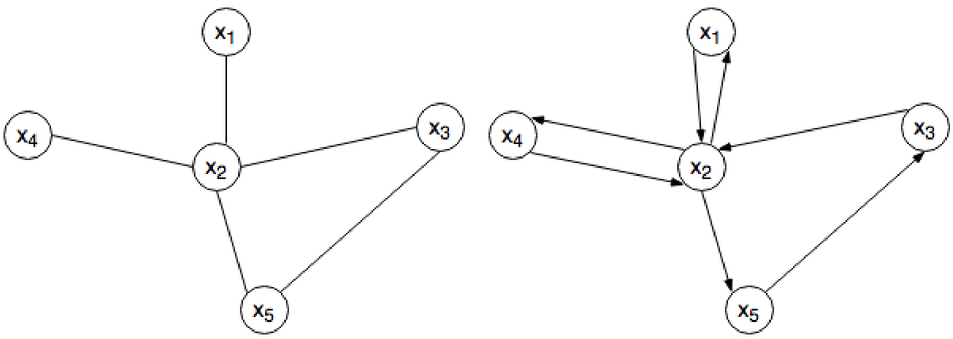
Structure of a network created by the interaction between users can be a useful means for retrieving information to support decision making. According to Caldarelli [38], understanding the creation mechanism of a network or graph can be a useful means in supporting predictions towards decisions to be made. Graph theory [211] have been proven to provide a description of connected structure between entities, as it has been applied to different fields in the real world. Examples can be seen under network structures in the economic field [88], where the market trends can be observed to improve the goods or services provided to consumers; network structures in the finance field [20, 124], where payments, trades and securities can be observed to resolve financial risk; network structures in the ecological field [110, 138], where biological interaction between organisms are observed to understand their adaptation, survival or resources depletion in their environment; network structures in transportation field [17, 111, 200], where traffic situations on different routes for a journey can be observed to determine the best route to a destination; network structures in communication field[10], where information flow is observed to understand all communication traffic or patterns in order to reveal the best channel for transmitting and receiving messages via phone exchanges or knowledge via online media (such as email and other social media). Analyses that can be carried out to examine the pattern of relationships in all these network structure examples are referred to as ***network analysis*** [211].

1.2 Network Analysis and Structure

With ties between entities in a network, information or knowledge is easily exchanged amongst themselves, cooperation will be improved and uncertainty will also be reduced [55]. This relation between entities from network structures supports decision making when important information such as preferences or possible solutions based on patterns or outcomes from past events are estimated. Aral & Van Alstyne [10] examined the correlation between network structure and benefits of information, where they pointed out that there will be better performance only if members of a community have the ability to access and discern non-redundant information. It is therefore important to measure the properties of a network to determine the information required to support decision making.

Network analysis measures the structural properties of a network such as patterns, duration and outcomes of past events or activities in order to determine either the formation of a network or dissolution of a network. Measuring concepts such as *sociometry*, introduced by Moreno[140] was amongst the initial method of network analysis which measures the interpersonal relationship between entities within a community. Moreno [140] was curious in visualizing the structure of a group and also understanding the human behaviour within a group. This was achieved with the invention of *sociogram*[140] which displayed the relations as lines that connect humans who are represented as points (nodes) in the display. The sociogram was not only described as a display tool but it was also referred to as an *"exploration"* tool which reveals social patterns.

Network analysis could be further understood with mathematical representation to model the relation between entities. One commonly used representation is the graph theory [211] which is the study of social graphs $G(V,E)$ where entities belong to vertices set V and relations or ties belong to edges set E. Social graphs might either be undirected (i.e. a graph where there is no distinction between edges $\{x_i,x_j\} \in E$ and $\{x_j,x_i\} \in E$ associating two entities $x_i \in V$ and $x_j \in V$) as shown in figure 1.1(a) or directed (i.e. a graph where there is a distinction between edges $\{x_i,x_j\}$ and $\{x_j,x_i\}$ associating two entities $x_i \in V$ and $x_j \in V$) as shown in figure 1.1(b).



((A)) Undirected Graph ((B)) Directed Graph

FIGURE 1.1: Types of Social Graph

For an undirected graph G with n entities, the set E must contain a maximum of $\frac{n(n-1)}{2}$ edges for the graph to be considered as a *complete graph* [211, 214] and if the graph is a directed graph, the set E must have $n(n-1)$ directed edges for the graph to be considered as a *complete digraph* [41, 211]. Both types of graph are relevant in understanding interaction pattern of entities as edges E play a key role in the observation of vertices' connection to others.

Exploration of social network structure with support from graph theory reveals the importance of some network concepts such as:

- *Homophily*[132], which is the likelihood for an individual to relate with similar members in a community. That is, an entity will have more preference for members in their group.
- *Clustering coefficient* [154, 212], which is the measure for the likelihood of nodes to be connected or clustered together. This can be used in determining if nodes can form a clique (complete subgraph) with mutual opinions or attributes.
- *Centrality* [65, 147, 211], which is the 'status' measure of a node being the most important member or central member reachable by all other members of a network.
- *Reciprocity* [34, 173] is the degree to which two nodes mutually exchange something (e.g messages or ratings) within a community. It is usually observed in a directed graph where the degree of exchange from pair of nodes might not necessarily be the same.
- *Structural holes* [36], which is the structural gap (differences) between two members of a network, where an intermediary might be able to retrieve unfamiliar ideas from complementary opinions or idea of members in a network.
- *Density* [69, 79], which is the proportion of relations(edges) compared with the expected number of relations in a network. This reveals whether a network is close to being a completely-connected graph or not.
- *Structural Cohesion* [139, 163], which is a measure for the minimum amount of nodes to be removed from a group to cause a disconnection in that group.
- *Structural Equivalence*[211], which is the degree to which two or more nodes share ties to the same other nodes. This structural equivalent nodes are also considered to be similar to each other based on their behaviour, i.e. Similar structured nodes are expected to have similar behaviour.

Some of these concepts and their relation with a system that might support users decision making will be discussed below to understand their relevance in analysing network structures. Most systems require network analysis for exploring navigation and position of entities within a community while others are focused on the causes of events or the behaviour (e.g competitive nature, activeness and in-activeness of entities) and similarities of entities.

1.2.1 Network Analysis via Homophily

Communities can be formed based on a concept where only similar members connect with one another. This concept is known as Homophily [132], which was referred to be a coordinating concept in networks. Earlier studies by Wellman [213] on community formation presented both friendship and play communities of kids which were formed based on demographic characteristics (e.g. Age, gender, religion and educational level). This revealed the correlation between affiliation and similarity of individuals or entities. Initial modern studies [30, 122] further revealed how similarity in attributes affects tie strength [2] of individuals. It was clearly seen that group members with similar attributes based on gender and education tend to have stronger interaction as the group is formed based on the similarity in experience and knowledge features between the members. Other groups with race/ethnic and religion attributes had their members engaging with themselves based on their beliefs as their attitude will always be similar.

Other studies [37] on community formation by homophily were based on psychological attributes (such as attitudes). This approach to measuring homophily was presented with the idea that peers will always influence each other's behaviour. In other words, peers will always aspire to be like each other in the community. Freeman [66] described the work done by Almack [6] as one which involved the comparison of peers' intelligence based on their school performance (grade) that affects their choice of sending membership invitation to other peers. The actions of similar members in a group are well coordinated in accordance with their mutual understanding in the community. We can also refer to this type of homophily as the behavioural pattern based homophily as it is a measure of the similarity in member's activities within a community.

Byrne [37] stated that similarity of members' actions in a community influences their attraction to each other. He proved this concept by analysing attitude-ratings of existing individuals towards an unknown group of individuals where the ratings were based on their intelligence, knowledge of existing events and integrity. If an individual has good morals and honesty in making the right decisions to various events, they could be considered to be a knowledgeable and an intelligent member. Other individuals

[2]Tie strength is a measure of solidity of a relationship between two members in a network.

will be attracted to this individual based on the condition that they have these same characteristics.

Apart from the qualities of an individual that affects attraction, other factors such as the 'importance' of issues and closeness were considered by other researchers. Byrne [37] revealed how certain group members considered the degree of subjects' importance from other members in evaluating their attitude similarities. As the importance of an issue, event or person is considered by an individual, the attitude of the individual is expected to be consistent and this, in turn, will enhance the attraction of others to the individual. Marsh [123] described 'importance' factor as a subjective means of determining the number of benefits to be expected from events or situations, as two similar events or situations could be rated differently based on 'importance' at different times. An important issue, event or person to agents could improve their closeness in a network. This 'importance' factor is comprehended with the concept of centrality where an entity's position in a network has to be measured to determine its' closeness to others in the network.

1.2.2 Network Analysis via Centrality

Centrality measure reveals the position or location of an entity in a network which in turn indicates the importance degree of the entity to others in the network. According to previous research [211], important entities are said to be located in "strategic" positions of a network in order to be more reachable to other members of the network.

Moreno [140] revealed in his work that the attraction between entities in a network will always cause them to remain close to each other. He demonstrated how less-active members of a network will only be close to few members that they like but this does not necessarily mean that the less-active member should be completely ignored by the community. This could be viewed in situations where a new member is initally close to an 'important' member of a community. In this type of situation, a new member might be accepted by other members into the community due to relationship with the 'important' member. An 'important' entity was described by Moreno [140] to have several properties which includes:

1. ability to give equal opportunity to all members to provide their opinion

2. ability to protect the weak from the strong

3. having no bias opinion towards alternatives

According to Borgatti [27], the importance of entities can be measured with centrality based on information transfer within a network. As a particular entity is close to other entities in a network, their proximity will determine how well information from the entity will be accessible by others, thereby making this entity important in the network. Freeman's work [65] is believed to have revealed the most reliable measure of centrality which consists of degree, closeness and betweenness concepts [211]. Previous researchers [65, 211] referred to degree centrality as an index for the activeness of a node who has the most ties in the network; closeness centrality as an index for the efficiency of a node to reach out to every other node in a network; betweenness centrality as an index for measuring the ability of a target node to control or influence other nodes who are separated or linked together by this target node. There is a correlation between the definition of closeness centrality in the previous researches [65, 211] with that given by Borgatti[26] who referred to it as an evidence for the expected time of an item (e.g. information) from a member to reach another member via the shortest path in the network. Both definitions refer to the flow time based on the position and proximity of a node with other nodes as a measure of centrality as this reveals the capability of a node to distribute or transmit information.

Research work in this thesis will reveal how all the three measures of centrality previously mentioned could be applied to the proposed model in identifying the most 'important'/influential node in a network (See figure 1.2) as activeness, efficiency and controllability [3] characteristics of a node will determine the fitness of the node. The identified influential node will be considered to be a motivator for less-active nodes to become active in the community or group.

1.3 Knowledge Transfer within Communities

Information transfer in a homophily community is limited to only members of that community with similar characteristics. According to Mcpherson [132], the closeness of two members can be translated into the number of links (edges) in which information flows between them (i.e "path length" [211]), which also correlates with homophily

[3]Controllability is the attribute of a person who creates an indirect link between two or more unlinked individuals in a network whereby resources or information can then be shared amongst themselves.

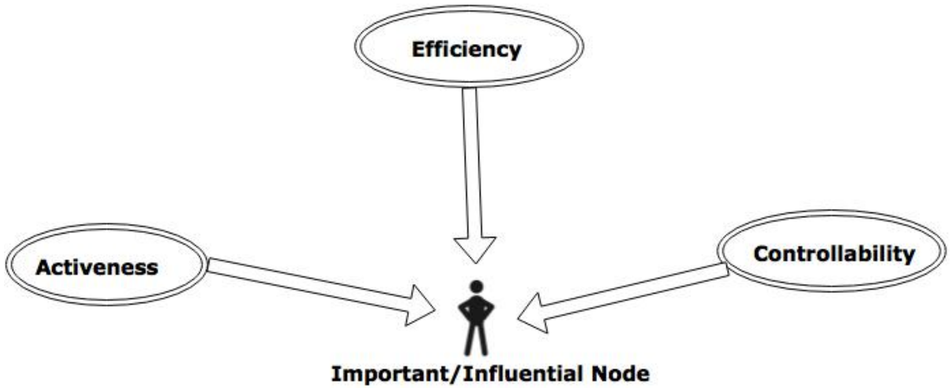


FIGURE 1.2: Attributes for identifying an Influential Node

[122, 132] that exist between the two members. It also implies that members (i.e. nodes) are reachable, i.e. there is a path [4] between them in a network [211].

The strength on edges for a path can be referred to various terms used as supporting information for analysis in different fields. In the communication field, it is the number of information or messages that have been transmitted or discussed between nodes (e.g Aral & Alstyne [10]); economist and financial officers refer to it as the rate of transactions between customers and banks or two banks/businesses (e.g. Battiston et.al.[20], Masi et.al.[124]); psychologist refer to it as rate of actions in response to a situations, effect or another person, which can be seen in a network between causes and events resulting from the causes[5] (e.g. Hevey et.al. [93], Almack [6]); transportation analyst refer to it as the amount of traffic or distance between start and destination point of a vehicle's journey (e.g. Sun et.al.[200], Levinson[111]); and ecologist refer to it as the consumption, access or feed rate on species by other species (e.g. Lever et.al. [110], Montoya et.al. [138]). The strength of the relationships [30] in these networks can be used as information to predict the tendency for future association or engagement.

Performance can be predicted based on access to novel information from diverse communities' ideas as individuals require new knowledge or innovative idea to support their decision making [10, 85]. Information from communities/groups has diverse and important contents (e.g. ideas, opinions or perspectives) which will create an innovative knowledge for members when combined together. This can be seen where familiar or similar individuals of a community have access to resources from unfamiliar or dissimilar individuals of diverse communities. This may also imply that there is the tendency

[4] Wassermann [211, p. 107] described path as the trail that can be used in keeping track of the communication channels between nodes in a network graph.

[5] The relationship in psychological network includes the relation whereby stress or depression could lead to illness, addictive personality could lead to overeating and overeating could lead to obesity.

of a particular node from a community to be connected with unfamiliar nodes of diverse communities based on their access to shared resources of information.

Reagan & Zuckerman [172] revealed the possibility of harmonizing both knowledge and benefits from network structures, whereby either diverse information from different groups or within a group can be beneficial. The novel information from various communities shared with a particular community will be effective in decision making by members of this particular community that experience new problems. An example of such situation will be when a customer of a Bank learns about a new idea for better services from another bank and he/she would prefer his/her own bank to adopt this new idea for better services. Even though the customer now has the preference for the new services from other banks, he or she will still remain with his or her bank probably because of his/her colleagues from work are still patronizing the same bank. The bank will then have to adopt or merge diverse idea from the different bank to meet its customer's need.

Earlier research work by Pfeffer [168] had explored other research work to point out that sharing of information and communication within long-existing communities tend to diminish due to familiarity of idea or behaviour within these communities. A related question of interest that needs answering will be: *Do members of a community who accept innovative idea or information from diverse communities also trust members of these communities?*. Subsequent sections will provide an insight of the research conducted that relates to this interesting question.

1.4 Understanding Behaviours in Social Communities

Earlier sections of this thesis have introduced the homophily features in a community where similar or familiar nodes will always be associated with each other but there is still the possibility for a target node with preferences of items to be connected to unfamiliar nodes, who might have changed preferences to some of these items, to become neighbours to other existing contacts of the target node in the network. There are cases where users (e.g. customers) add value to an item by making use of it for another user to value it. The effect in this kind of case can be referred as ***network effect or network externalities*** [190] which depends on the number of members involved in adding value to the item. Similarly, active nodes' interaction with a new or inactive node will have an effect on others to interact with the same new or inactive node.

Naive members of a community will find decision-making easier when other members share their experience and information about a new or non-active member; as the naive members need to be convinced before they can decide to interact.

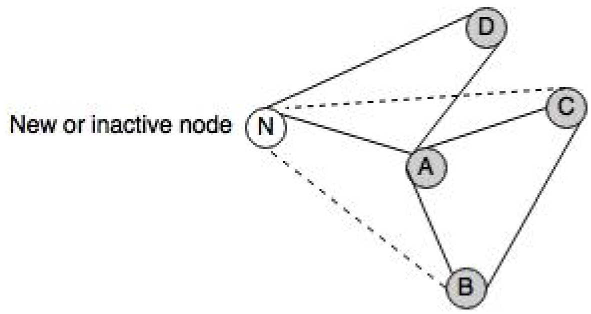


FIGURE 1.3: Network Effect where broken lines represent possible ties.

Even though there can be unfamiliar actions by an individual, other individuals will still need to know the reason behind such behavioural change of the individual. There could be a change in behaviour and this new behaviour must be related to the usual behaviour with a reasonable explanation to clarify the change. Steven Hayes [89] gave an example of how soldiers who are meant to fight a battle could abandon the battle for various reasons. An unusual behaviour of an entity could occur towards a familiar event where victory outcome is expected at the end. There is the need to determine the reasons for the change in behaviour as this will also deduce if the actions were genuine or not.

Previous research [120] suggested the use of ***casual history explanation*** as means of linking all different behaviours to determine a "common denominator" from the reasons to each behaviour. Looking at the battlefield example, the soldiers might have abandoned the battle probably because there might have been no ammunition available to use in the battlefield, no means to ship in ammunition or their government had no more funds to purchase the ammunition. All these reasons can be generalized to be the unavailability of ammunition which is causing the actions such as "the abandonment of battle" to happen. But in a social context, a general factor that affects the behaviour of entities (soldiers) towards another entity (battle event) will be "trust". We could then reason and say that the soldiers abandoned the battle because they do not trust the amount of ammunition that have been provided for the battle, do not trust the shipment of more ammunition or do not trust the government in purchasing more ammunitions.

It is important for each member of a community to understand the behaviour of other members in order to build their trust and cooperation with them. Looking at the battlefield example once again, we could consider the trust amongst soldiers as a factor for the abandonment behaviour. What if the few soldiers do not trust other soldiers based on their lack of fitness, health or morale that affect their performances on the battlefield? The type of network structure from this case will be the relationship between the soldiers which is different from the network structure in the previous case where the network is between soldiers and the battle event. In this thesis, two types of network related to this cases of the example (i.e network between different classes of nodes and network between a single class of nodes) will be explored to determine the trustworthiness of an entity.

1.4.1 Predicting Trust behaviours in Social Communities

For an intelligent system, such as a recommender system, there is the need to know the trustworthiness of either the system or individuals interacting with the system. Either the system or the individual might decide to act in a way to manipulate each other to obtain benefits. Alan Turing[206] had earlier predicted that a day will come when humans will be deceived during communication without being aware of who they are actually conversing with, as it could be either another human or a computer. We can then suggest the use of behavioural pattern of the entity as a means to providing explanation tools in resolving the trusting issue.

As intelligent systems are based on a communication network where entities need to interact in order to share resources or messages, trust is required to predict the future interaction between the entities. An entity might tend to trust other entities if the resources or information provided by these entities are important, credible and presented when required [169]. But it is also possible for an entity to change their trusting behaviour when they are influenced by a familiar entity in their community; this effect was referred as ***social norm*** in [70, 128].

As earlier discussed, people may tend to identify themselves with groups based on homophily in attitude where their behavioural pattern are similar. Social norms act as the standard set of behavioural rules that guides all members in a particular group. Group members may decide to comply with or accept an item suggested by other similar members in a group. They decide to act this way probably because they believe that members of the group are trustworthy and not deceptive since they are guided by the social norm. One common theory for understanding this behaviour is known as the conformity theory [130], as the change in behaviour or beliefs must be in accordance with a group's acceptable behaviour.

1.4.2 Conforming to a Social Community

Conformity which was described to be a type of social influence [130] was initially studied by Jenness [99] with an experiment carried out with participants who initially provided their individual estimate of the number of beans in a bottle before they were all grouped to provide a group estimate. The group was then separated in order for each participants to carry out a comparison between their initial estimate and the group estimate for them to either adjust to the group's opinion or stick to their initial opinion. The final result from the experiment showed that nearly all the participants changed their opinion from their initial estimate to the group estimate. Another similar but popular experiment is the Asch's line experiment[6] [127] where the accuracy level of a naive participant's opinion is traced to the pressure of a group who had planned beforehand on their opinion. The experiment concluded that most of the naive participants conformed to group's opinion because of their fear of being judged by others or their thought of behaving abnormally. Kelman [107] referred to this type of conformity as ***Compliance*** as members decide to conform just to be accepted by the group and avoid being punished by the group.

Other types of conformity mentioned by Kelman[107] that seems to be more appreciated are ***Identification*** and ***Internalization*** where individuals change their behaviour based on their desire to be established with a group and the idea behind the induced behaviour of group members, respectively. A person whose behaviour is based on 'Identification' conformity only decides to have a relationship with an influential person to boost up his/her own status in the community. In the case of 'Internalization' conformity, a person accepts an influential person [7] who usually acts in an honest and selfless way to support in solving important problems. Deutsch & Gerrard [54] referred to the reason behind the 'Internalization' behaviour as ***Informational conformity*** where the

[6] Asch's line experiment involved 50 naive participants with seven other participants who were associates to Solomon Asch, the researcher. The naive participants who were unaware of the other participants being associates to Solomon Asch was asked to provide his or her opinion along with other participants' opinion on the length of line from a card when compared to another card with several lines of different lengths.

[7] A person who has the ability to induce his/her behaviour on another person.

person who accepts an induced behaviour is a naive or ill-informed person that decides to compare his/her own behaviour with other members of a community.

1.5 Behavioural patterns in Decision Making

Psychologist's studies on network structure show that understanding the causes of events will enable us to predict or plan for better future outcomes [89, 93]. They pointed out the need for network analysis to understand patterns and their causes. Hevey et.al.[93] pointed out the benefits from network analysis as it reveals:

- the history of events in relation to possible future outcomes.
- the causes and their degree of occurrence in past events or outcome.
- the pattern of behaviour that can further show the stability of the causes of events.

The change in interaction pattern of individuals in a network can be observed and utilized in predicting the future engagement of individuals with other members. Steven Hayes had previously pointed out in his research [89] that understanding how things work or relying on behavioural patterns of people will provide easy solutions to prediction problems. People trust other people based on their usual behaviour even though there are changes in behaviour after certain times or during certain situations. Hayes [89] stated:

"*If you do anything different in the presence of events that normally lead to patterns, you are helping to create more psychological flexibility.*"

This means that if there is a change in behaviour of an entity based on certain situations, there should always be a relationship between the previous behaviour and the new behaviour. This relationship reveals the justification for the change in the behaviour of the entity, as the past events or situations are analysed. Kashdan & Rottenberg [105] defined psychology flexibility to be a measure of change in perception and the adaptation to varying events.

Behavioural patterns of an entity which can reveal the implicit preferences [21, 75] of the entity are required for evaluating the recommendation for the entity. But an entity in a social group could have dynamic preferences to either items or other entities which might lead to consensus problem. For instance, a target user of a system believed to belong to a group based on the fact that the target user had once interacted with one or two other users of the group in the past, could disagree with the group's general opinion or become inactive on specific items accepted by the group. Here the target user inactivity could be due to their lack of interest on some of the items presented. The in-activeness of this user might then affect the identification of the user with a group and the accuracy in recommendation of new items to the user.

There could be other possible causes for users being inactive in a social network. These might include:

- the user's fear of being judged by others
- the user's fear of compromised privacy.

Most naive people would not want to experience failure or to be accessed by others when they are revealing their opinion in a social network; they will rather remain inactive. This inactiveness might also be due to situations where some people prefer their opinion or information kept private. All this will then cause consensus problem to occur where inactive members that exist in a social group disagree with the group's opinion, thereby being in a state of cold-start problem [125, 186, 189].

1.6 Research Aim and Objectives

The research aim is mainly to explore if inactive members of a social group could be encouraged by influential members to become active in order for accurate prediction of their preferences to be estimated for an effective recommendation. The previous section of the thesis revealed that consensus problem might occur when inactive members of a group decide to disagree with other members on the group's opinion on recommended items. This disagreement is mostly observed from past activities where inactive members decide to be less active towards items which their neighbours are interested in. It is possible to resolve the consensus problem by considering the influence concept as inactive members could be influenced by their trustworthy neighbours to change their behaviour.

From the main aim of the research, several objectives were drawn out. They include:

1. To formulate a trust metric to measure the trustworthiness of each entity or node of a social network.

2. To explore suitable archived social network data to define the important social features that could be used as parameters in the formulated trust metric. There is the need to obtain trust data which predicts the trust relationship amongst entities.

3. To test and implement suitable clustering algorithms on the trust data that reveals the relationship between entities. Each generated cluster along with its influential member could reveal how other members are attracted to the clusters.

4. To validate an influence-driven recommendation framework which integrates social network analysis, trust concept and clustering. This will be carried out to check if an influential member of a cluster will always be a motivator to inactive members of the cluster.

The adoption of these objectives in the study offers insight on how the proposed recommendation algorithm which is based on influence could predict if less active users will have interest on items that their trustworthy neighbours mostly prefer.

1.7 Research Methodology and Datasets

The method used in this research involved applying natural science (i.e. knowledge that describes and explain how things in the world behave) on design science [94] (i.e. knowledge that reveals a new phenomenon that will support the needs of people). The phases of the research to be discussed later in the section will clearly describe this research approach.

For the past few years, the initial research focus has been on network structure and the factors that affect relationships in a network which could provide a better understanding of behavioural patterns in the network. These include both theoretical and empirical analysis carried out on archive datasets which will be discussed later in the section. The theoretical analysis in the research work involved the study of natural science where social interactions between entities in different fields were required to reveal the behaviour of the entities and this phase provided the awareness of problems that exist in the research area. The empirical analysis of the research work involved the artificial (design) science where the knowledge from the theoretical study was used to support the design for an innovative approach in the research area.

1.7.1 Research Methodology

In order to address the research aim mentioned in the previous section (Section 1.5), the research tasks were structured as follows:

Phase 1. Theoretical Investigation: In-depth studies and critical reviews were carried out to provide evidence for the existence of the problem in preference prediction as current evaluations have been inaccurate and ineffective due to inactiveness (cold-start problem)[125] which are not considered in the computation of preferences. Some preference computation will be inaccurate in a system's 'learning process' if they fail to consider justification for the actions in past situations as we cannot always rely on usual behaviours for future predictions. Previous research [224] referred to this type of computation problem as centralized computation.

Discovery from the literature review revealed that in-activeness within a network which leads to the cold-start problem [125, 219] also affects preference prediction. This in-activeness problem occurs when users are new and they have no knowledge of items or services offered by the system. Based on natural phenomena, users can only be influenced by their trustworthy neighbours to change their behaviour or opinions. The factors that affect trusting behaviour in a network was then required to determine if activeness in the network can be built or improved by influence.

In cases where there are no explicit feedbacks for learning and predicting preference, the implicit information are relied upon but this also faces the inaccuracy problem as actions might have been carried out based on so many reasons [1]. For instance, it will be very difficult to compute the preference for a person who might have purchased items for other people(e.g. friends). Palmisano [157] attempted to resolve the problem by partitioning users according to their context of purchase which could be used to predict each user's preference. The concept of partitioning (clustering) was also decided to be applied as part of the concepts to explore its effect on changing user's behaviour within a social network.

Phase 2. Conceptual Model: From the literature review carried in phase 1 of the research work, various concepts were reviewed and the important ones were chosen to be utilized in designing a new model for encouraging inactive members of a social network. It was decide in the research to identify influential members who are also trustworthy members to the less active or inactive members in a social network data. From the observation of the social data, there was the need to rely on implicit feedbacks [106, 109, 149] (modelled from the theory of interpersonal behaviour [13, 165, 203]) as part of the dimensions for preference prediction since explicit feedbacks are either misunderstood or not provided. Previous research has revealed that explicit feedbacks were described to be inaccurate since different persons might have different understanding of a particular rating. For instance, two different persons might have different opinions on what exact value to represent a high value.

As trust is an element in a social network [144] that might enable a specific behaviour, the best dimension for measuring it is implicit feedbacks such as viewing count [160, 161], interaction count [162] and purchase count [45]. There is the need to accurately measure trust since it is defined and applied differently in various areas of life. The formulated trust metric was expected to measure the trustworthiness of each node in a social network to reveal their degree of activeness, efficiency and controllability in the network. That is, a trustworthy entity is expected to be more active, reachable and able to control information than other members of the network.

The computation of trust proposed in this thesis was based on the theory of interpersonal behaviour (TIB) [13, 165, 203] which considers habits and 'situational condition' in predicting future behaviour (See figure 1.4). Habits or patterns of entities can only be understood from entities' repetitive behaviour which could be used to estimate situational condition [204]. The situational condition, also known as *facilitating condition* was described by Triandis [203] as a factor that aids agreement amongst entities in a social community. Triandis further pointed out that 'intention' might not necessarily predict the behaviour of an entity as a situational condition (such as environmental condition or logical condition) can directly cause a behaviour to occur. For example, there might be an intent of a person initially to carry out a task (e.g. accessing a document or interacting with unknown users) but if an environmental condition (e.g. insufficient security) does not allow such task to be performed by the person then the action will be halted.

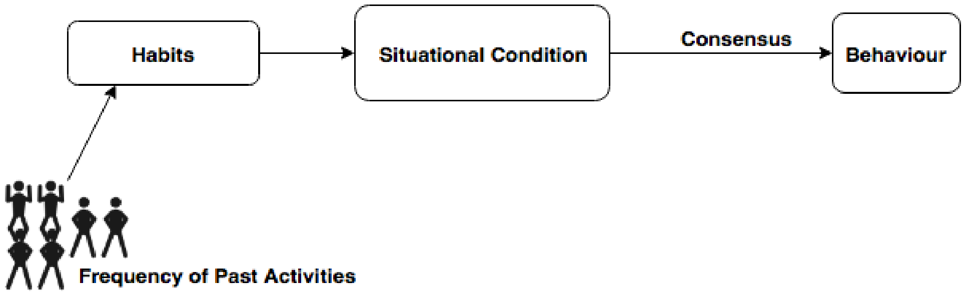


FIGURE 1.4: Theory of Interpersonal Behaviour (TIB)

The use of association rule [2, 187] was initially required in the research to analyze how frequent a node from a specific class will interact with another node of the class based on communication on an item (a different class of node). The decision to use interaction count for the research work was based on the type of dataset (described in the subsection below) which consists of static social interaction (direct or indirect interaction) between entities. The use of rational choice theory [47] was also considered in the aggregation of all entity's previous behaviour as each entity acted previously based on its own preferences amongst all available alternatives and constraints at those times.

From the trustworthiness of each node, the influence concept was to be explored in determining the influential node. Here, it was decided to simulate how users could be identified with a group based on influence or attraction by a trustworthy user. Various clustering techniques were used in identifying different groups and their influential node based on either connectivity or similarity. A comparative analysis of the various clustering techniques was carried out to determine the best groups and their influential nodes based on the degree of activeness. ***R*** software [170] with relevant packages were used in the clustering analysis carried out in the research work.

The concept of learning and understanding social factors that could support in predicting the relationship between an influential node and other members in its group were considered based on the theory of social cognitive which was introduced by Bandura [14, 15] as a concept which involves social learning of entities' social behaviour (i.e. feedback from past experience) that could influence a target entity. The theory considers the evaluation of social factors affecting entities of a social group which might have an impact on a behaviour of a target entity in the same group.

Other concepts that were reviewed but not considered for modelling behaviour in this research work of the thesis includes:

1. Belief, desire and intention theory (BDI)
2. Theory of Planned Behaviour(TPB).

BDI theory was introduced by Bratman[31] as a theory for comprehending an 'intention' towards a behaviour. Here, 'desire' is considered to be a motivation for enabling someone to act [191]. It will be more reasonable to consider 'situational condition' as an important feature in modelling a behaviour instead of relying on the 'desire' to act or behave which is guided by the 'belief' in satisfaction of the desire. Situational condition reveals the need for an entity or person to act based a current situation where belief is not required.

Belief is usually misunderstood for 'trust' but there are some differences between both concepts. The dictionary referred belief to be the acceptance of something existing or being true while 'trust' was defined to be the belief in the truth or strength of something/someone. Belief is an expectation towards only one outcome while trust might lead to either a negative or a positive outcome. In some cases, there could be belief in something/someone without the existence of trust but a persistent belief will lead to trusting behaviours. For example, a new driver will initially be made to believe with no trust that an airbag from his vehicle will always deploy at the point of any collision but the driver could develop trust after several experience with the vehicle in such situation. From these differences between belief and trust, we could therefore see 'trust' as a more stronger factor that could be used in predicting behaviours.

TPB theory which is another theory for understanding behaviour considers volitional control [4] where entities have the free-will to behave or not behave in a certain situation. This theory lacks other features such as 'habit' and 'situational condition' that have strong predictive effect in determining future behaviour.

Phase 3. <u>Evaluation</u>: This last phase was to create a simulation of a recommender system that will reveal identified groups and their influential nodes to target nodes (inactive nodes) who might be encouraged to accept the recommendation of items provided by the system. This phase simulates influential nodes motivating inactive nodes in their group to accept items which were previously preferred by them

(Influential node). Here, an existing recommendation algorithm to test this framework was considered to validate the possibility of the influential nodes acting as recommenders.

1.7.2 Dataset

An archived dataset previously collected as a Facebook-like social network and used by Opsahl [151] was initially chosen to be used in this research as it consists of social context. This dataset (See example in table B.1) known as a weighted one-mode network, represented social ties between nodes (anonymous persons) of a single set that exchanged messages amongst themselves. The data consist of index number for identifying each node and weighted values to reflect the total number of messages sent or received by the node. Social factors to be used in the proposed trust metric were evaluated from the dataset and the activeness of a node was measured based on direct interaction with another node. This dataset was not considered for further research as it does not reflect any information on the messages exchanged between anonymous persons. Further description of a one-mode network from this type of dataset will be analysed and compared with another type of network later in the thesis.

Another type of dataset from the resource in Opsahl's research work [151] was considered for further analysis to observe a different type of social network where a non-direct interaction between nodes exist. The dataset is a Facebook-like forum network of anonymous persons' activities towards topics. The dataset which was known by previous researchers [28, 151] as a weighted two-mode dataset represent the social relationship between a set of nodes (i.e. anonymous persons) and another set of nodes (i.e. discussed topics). The network structure from the dataset consists of the anonymous persons having ties with 'topics' based on their posted messages (implicit ratings) towards the 'topics' and both sets of nodes are identified with numbers. The weight on ties or edges between the different sets of nodes are represented in the data as the total number of post to a 'topic' from each anonymous person. This type of dataset was chosen as a suitable dataset that could reveal each node's initial preferences to items (e.g. topics) based on their past activities.

A random sample of data (See Appendix B.2) was extracted from the two-mode dataset [152] to carry out all empirical test in the research work of the thesis. The sample data consisted of 20 user nodes and 211 item nodes. Proper care was taken (with no bias) to ensure that the sample data collected covered all kinds of users which include both active and inactive users in the network. These type of users were suitable for the simulation process of the proposed framework which is expected to encourage inactive entities in the network. The network from this weighted two-mode dataset will be further discussed and compared with the type of network from the initial dataset (i.e. one-mode network) in Chapter 3 of the thesis.

Another data (See Appendix B.4) extracted from a similar weighted two-mode dataset, 'Hollywood film-music' dataset presented by Vladimir Batagelj & Andrej Mrvar [18] was also considered in the research for further analysis to validate the proposed framework. The 'Hollywood film-music' dataset consists of a collaboration between a producer of a film and a composer of music where the network from this dataset consider the composer as a common item that links two or more producers who had at once employed the composer to create at least a music (soundtrack) for their film (i.e. the number of music composition by a composer for a film producer is the weight on the tie or edge between them). A data (See Appendix B.4) of 30 producers (indexed between 1 and 62) with 35 composers (indexed between 63 and 102) were randomly retrieved from the 'Hollywood film-music' dataset to carry-out further empirical test for the validation of the proposed framework in the thesis research.

1.8 Measure of Success

The following criteria were used as a measure to determine the success of the research work:

- The ability to reveal how the probability between pairs of nodes from a set having ties with other nodes from another set could be used in predicting if the pairs will remain active in the network.
- The ability of the proposed trust metric to accurately measure the trustworthiness of each node in a network.
- The ability of the trust data to reflect and predict the activeness of each node in the network.
- The ability to accurately identify groups (i.e. several neighbour of trusted members) and their influential nodes from the trust data using a suitable clustering technique.

- The ability of an *influence based recommendation framework* to provide accurate recommendation and encourage an inactive node to become active; the identified influential nodes are expected to motivate the inactive nodes to be active in supporting the recommendation process.

1.9 Outline of the thesis

The remaining chapters of the thesis presents all outcomes from each phase of the research work. The outline for the remaining chapters are follows:

- **Chapter 2** presents the reviews on existing recommendation approaches and their impact on the society today. The trust concept was also reviewed to present its impact on recommendation processes. Supporting reviews on pattern analysis were carried out to reveal social features that could be considered when evaluating the trustworthiness of an entity in a social network. Also, the influence concept was discussed to point out the characteristics of influential members in a social group. Various clustering techniques were also reviewed to provide a clear understanding of their impact on social network (trust network).

- **Chapter 3** presents the proposed trust metric which requires social features that are defined for the computation of trust. The literature review carried out in chapter 2 which revealed that there is an association between trust and similarity helped in the formulation of trust. This chapter also discusses how social activities from a data could be used to estimate the activeness of an entity (i.e. Node) which can be used to determine if the entity is trustworthy or not.

- **Chapter 4** presents the empirical tests carried out using clustering techniques on the trust data previously evaluated during the investigation discussed in chapter 3. Various clustering techniques reviewed during the phase 1 of the research work were compared during the tests to discover the most suitable technique for identifying the trust groups (clusters) and their potential influential members.

- **Chapter 5** presents a framework of enhancing recommendation by integrating trust and cluster output to encourage inactive members in accepting recommended items. Further investigation using singular value decomposition algorithm is expected to confirm if potential influential members could motivate inactive members of their cluster.

- **Chapter 6** summarises the research work that have been discussed in the thesis and suggests possible future works that could provide extended contributions beyond those from this research work.

Chapter 2

Related Works

2.1 Introduction

This study includes an investigation to determine if an influence based mechanism could improve the activities of inactive nodes for accurate recommendation to be provided for the inactive nodes and other nodes in a social network. The content of this chapter includes a review carried out on related areas on recommendation processes and the most common recommendation approaches previously presented and discussed by other researchers. The review of trust concept and its impact on recommendation processes will also be discussed to justify the decision for applying this concept in the research work. This review will also enable us to be aware of the social features that are being ignored in the computation of trust. In order to apply a suitable clustering technique on trust data, there was the need for a review on literature related to clustering techniques. Most importantly a review of influence concept was required to discover the relationship between trust and influence which affects change in the behaviour of entities within a social network.

2.2 Exploitation of Recommender System

A recommender system is used to predict the preferences of a user who might find it difficult to search and decide on the right items from a pool of item source [178]. This is done by the system's observation on the behavioural pattern of the user from previous experience to be able to predict other items that the user has no knowledge or experience. A lot of service providers use a recommender system to improve user's satisfaction with their services by providing new and relevant items [178]. Recommender systems have been used for a range of services which includes advisory service on supportive people[25], purchasing assistance on relevant items [114], advisory services on suitable travels [177], viewing support [137, 175] and advisory services on suitable finance [61, 62].

The provision of the recommendation services seems promising to deal with the information overload in various systems as the preferences of each user or customer are used to improve their selection of items. Previous research [115] described the outcome from the recommendation as a means to measure the interest of users. But users might find it difficult to retrieve the relevant items from a huge pool of dataset with items or from a short query which might return a lot of results. However, cold-start problem [125, 186] which occurs when users don't provide sufficient explicit ratings or opinions on items still persist in a recommender system.

In order to provide a recommendation for users, the users' preferences to items need to be modelled before the system can be able to recommend other items related or similar to the previous items that were preferred by the users. Previous researchers [133, 196] considered diversity [1] as an important factor in recommendation as people actually want to see diverse and slightly similar items on their recommendation list in order to make a proper selections or decisions. In other words, increasing the diversity of items to be recommended implies decreasing the similarity of items to a certain degree without compromising it. Previous research [196] used a bounded greedy algorithm which required both similarities between each case with a query and diversity of a case relative to other cases for the recommendation strategy. Users of various systems do not necessarily want the exact item according to the query or from previous experience [189] as they prefer new and diverse items in the recommendation list for them to easily make selections. The diverse items can also assist new users who are naive and require assistance in their decision making. But the interest of the users can be affected by diversity based on the nature of the items. For instance, it will be more difficult to evaluate the preference of users on news items than movie items since the news item changes frequently.

A system's dataset or memory of previous cases with the same problem but different outcomes obtained from the application of various actions makes learning process in

[1] Smyth & Mcclave [196] described diversity in recommendation as a phenomenon where there is relativity between cases which are dissimilar to each other.

recommendation difficult. For example, a case of a news item where a user who recently has read and saved a particular type of news but on other occasions, he or she has either read and shared this same type of news to someone else or ignored the news. All these news-item cases have their individual satisfaction level based on different situations where the user might have read and saved the news because either he or she needed well-detailed information about available jobs, he/she knew a friend who is searching for a job or he/she does not need the information as it is not sufficient enough for any job application. Frequency measure was considered in [76] where repetition of cases might exist with a different outcome. If the frequency of successes with an action p is higher in a particular memory than the frequency of failures with other actions q in different memories, then the action p will be more preferable than the other actions q.

Another recommendation issue can be seen in the “learning stage” of recommendation process when the user’s query is unavailable or there is insufficient information about the item (cold-start problem) [186, 219]. Previous researchers [3, 33, 194, 195] stated that the possibility of having a recommendation list of items for a user without the provision of an explicit query to specify the needs of a user. With the use of the hybrid technique [33, 35] by combining both collaborative and case-based techniques, this will support the learning process to provide recommendation list. We could refer to collaborative approach as a process of using implicit queries since the preferences of users who have similar behavioural patterns to the target users are used in the recommendation. The needs of the target user are inferred from the similar users who are active for the system to observe them. But we should also ask the question if these similar users can actually influence or motivate the target user to be more active in the system; are the preference of the target user always similar to the similar users?

2.2.1 Recommendation with Content Based Filtering

The main idea in the content based filtering [115, 136, 164] is to recommend new items that share attributes or features with group items that have been rated with high preference by a target user. The content-based filtering algorithm will search for items with similar features f_t that a target user has rated in the past and identifies other items with similar features f_s to f_t, where $f_s, f_t \in F$. The feature of the rated items (e.g keyword of a web-page or document) from the target user’s profile will be compared with features of new items using a similarity measure.

According to previous researches [115, 164], similarity measure will initially require weights that represent the degree of feature's or term's relevance in an item (e.g webpage or document as a vector). This weight $w_{(f,i)}$ is known as Term frequency-inverse document frequency (TF-IDF) which is based on how frequent features or terms $freq_{f,i}$ occur in an item i (i.e. TF) that are also rare in other items (i.e. IDF) could be considered relevant to the main point of the item.

$$w_{(f,i)} = TF \times IDF = freq_{f,i} \times \log\left((\frac{N}{n^i_f}) \right) \tag{2.1}$$

The weight vectors was described in [115] to be normalized using cosine normalization to disallow complex items (e.g lengthy documents) from being retrieved. This normalization ensures that the weight range between 0 and 1.

$$w_{(f,i)} = \frac{freq_{f,i} \log(\frac{N}{n^i_f})}{\sqrt{\sum (freq_{f,i})^2 \log(\frac{N}{n^i_f})^2}} \tag{2.2}$$

Where: N represents the number of items in the collection while n^i_f represent the number of items that have the feature or term f.

The weights of each feature to an item are evaluated using equation (2.2) and the feature with the highest weight is considered the most relevant feature of the item. However, in order to carry out the evaluation two types of preference information are important to build the user profile: Information via user's history (e.g explicit or implicit feedbacks) on items and information via main feature (keyword) of the item that describes the item.

Example 2.1. *For example, given a user's explicit ratings on certain items with their features:*

Items & Features				
Football Club	English	Spanish	French	Rating
Arsenal FC	2	14	5	Like
Man United	12	6	3	?
Monaco	1	0	20	Dislike
FC Barcelona	0	17	6	Like
PSG	0	2	18	?

TABLE 2.1: Ratings from a user on several items

$w_{(Spanish,Arsenal)} = 0.95$ $w_{(Spanish,Barcelona)} = 1.00$ $w_{(Spanish,ManUnited)} = 0.21$

$w_{(English,Arsenal)} = 0.31$ $w_{(English,Barcelona)} = 0.00$ $w_{(English,ManUnited)} = 0.97$

$w_{(French,Arsenal)} = 0.00$ $w_{(French,Barcelona)} = 0.00$ $w_{(French,ManUnited)} = 0.00$

$w_{(Spanish,PSG)} = 1.00$ $w_{(Spanish,Monaco)} = 0.00$

$w_{(English,PSG)} = 0.00$ $w_{(English,Monaco)} = 0.99$

$w_{(French,PSG)} = 0.00$ $w_{(French,Monaco)} = 0.00$

By observing table 2.1, it can be inferred that the user has a preference for football clubs with the majority of their players as Spanish players. Do we think the user will consider the items, Man United and PSG as relevant to follow?

Previous research [179] revealed that an item is said to be relevant to a user if he/she rates the item with explicit positive feedback (e.g. 'like', '+' or binary value **1**). But in the example2.1, the item's features along the user's explicit feedback are used as means in predicting the relevance of items. Initially, the weight of each feature in an item w_{fi} needs to be determined to represent the item as a vector in n-dimensional space (i.e. item $I = \{w_{f_{1i}}, w_{f_{2i}} \dots w_{f_{Ni}}\}$). The weight of feature 'Spanish' was evaluated to be higher than any other feature (English and French) on the items, 'Arsenal' and 'Barcelona'. The feature 'French' is insignificant or informative as it occurs in every item.

To compare and match the item representation (i.e. as vectors) with a potential item for a recommendation, a suitable similarity metric must be used. The most common and appropriate similarity measure used by previous researchers [46, 115, 181] for vector space model is the cosine similarity where the similarity between items are evaluated based on their relevant feature.

$$sim(I,J) = \cos(\overrightarrow{I}, \overrightarrow{J}) = \frac{\overrightarrow{I} \cdot \overrightarrow{J}}{\| \overrightarrow{I} \| \times \| \overrightarrow{J} \|} = \frac{\sum_{f=1}^{N} w_{fi} \cdot w_{fj}}{\sqrt{\sum_{f=1}^{N} (w_{fi})^2 \cdot \sum_{f=1}^{N} (w_{fj})^2}} \tag{2.3}$$

Using equation 2.3, items 'Arsenal' and 'Barcelona' are confirmed to be more similar to each other as the similarity value is estimated to be 0.95. Another item considered to be similar to ‘Arsenal’ is ‘PSG’ as their weights of various features are closely related where the weight of ‘French’ in any of the items is irrelevant (i.e. Either weight of ‘French’ in any item is zero). Therefore, we can consider item ‘PSG’ as an item to be recommended to the user based on its similarity of features in both ‘Arsenal’ and ‘Barcelona’ that have been rated previously as items the user likes. Item ‘Man United’ is considered to be irrelevant to the user as it is not similar to both ‘Arsenal’ and ‘Barcelona’ but more closely related to item ‘Monaco’ that the user dislikes.

Content-based filtering technique is transparent as justification for the recommended list to an active user is clearly based on similar features that exist in items previously rated by the user. The features of items in this technique have been described in [115] as a trust indicator for the user to accept the recommendation. However, the recommendation might be inaccurate or impossible to retrieve due to cold-start problem [125, 186, 189, 219] where there is insufficient or no information (user’s ratings on items) to model a user’s preference. The user might decide to be inactive due to the fact that he/she is new or based on privacy issues where the user might have lack of trust to share their preference information.

Another problem that the content based filtering approach could experience is *Over-specialization* where a recommended item with similar features with items that an active user had previously rated as ‘like’ might not necessarily be a new or novel item to the user. Also, it could be that items with exact features to previously preferred items are not relevant to the active user. For instance, item ‘Barcelona’ recommended to the active user might not be relevant to the user as there could be other reasons why previously preferred item was chosen by the user. Therefore, we can state that the use of features in describing items is insufficient in distinguishing an item from another item that could be of interest to an active user.

With all these existing problems in Content-based filtering, researchers have decided to consider another recommendation technique, *Collaborative filtering* [11, 90, 185] as an alternative to resolving the problem where an active user has been inactive in rating items or acting (For example, discussing, viewing or purchasing) towards items. The profile of other users similar to the active user will be used to recommend certain new items to the active user.

2.2.2 Recommendation with Collaborative filtering

Collaborative filtering requires ratings or feedbacks from active users $U = \{u_1, u_2, u_3 \ldots u_m\}$ on items $I = \{i_1, i_2, i_3 \ldots i_n\}$ which they have previously encountered or utilised. This is then used as a means for predicting new items I_a to active users who have never had any experience with these items. The ratings $r_{ab} \in R$ for items are usually represented on a user-item matrix $(m \times n)$ where each row represent a user u_a and columns represent items i_b rated by each user.

In order to predict suitable new items (Known as top-N recommendation list [51]) for active users u_a, there is the need to also predict ratings $\hat{r_a}$ for all unknown items I_a where N suitable items ν (Where $\nu \subset I_a$) will be retrieved based on their high predicted ratings [182] where $\forall_{x \in \nu} \forall_{y \in I_a} : \hat{r_{ax}} \geq \hat{r_{ay}}$.

Breese et.al. [32] classified the collaborative filtering CF algorithm into two classes, **Memory based CF** (Neighbourhood-based) and **Model based CF**.

2.2.2.1 Memory Based Collaborative filtering

The memory-based CF requires the whole set of rating data for the prediction of the user's preference. An Examples of the Memory-based CF algorithm is the user based collaborative filtering where the aggregation of a set of N_u users' rating $r_{u',i}$ on an item i estimates the rating on the item $\hat{r_{u,i}}$ for an active user u who is similar to these set of users $u' \in U$ also known as nearest neighbour [52]. The aggregation methods required for the estimation of the rating for the active user includes:

1. Mean Ratings of similar users

$$\hat{r_{u,i}} = \frac{1}{N_u} \sum_{u' \in U} r_{u',i} \tag{2.4}$$

2. Ratings weighted by similarity

$$\hat{r_{u,i}} = \frac{1}{|\sum_{u' \in U} sim(u,u')|} \sum_{u' \in U} sim(u,u') r_{u',i} \tag{2.5}$$

3. Deviation from mean ratings (Rating Normalization)

$$\hat{r_{u,i}} = \bar{r_u} + \frac{1}{|\sum_{u' \in U} sim(u,u')|} \sum_{u' \in U} sim(u,u')(r_{u',i} - \bar{r_{u'}}) \tag{2.6}$$

Where: $\bar{r_u}$ is the mean rating for all rated items by user u.

The N_u users u' are considered to be neighbours to user u if they have high similarity $sim_{u,u'}$ with u. **Equation 2.4** requires only the ratings of users u' on a particular item that is new to user u but it does not consider the fact that neighbours of user u could have a different degree of similarities with user u [52]. For example, considering the user-item rating matrix below (Example 2.2) where the missing rating from an active user u_4 on item i_3 is required. If the nearest neighbours of user u_3 are users u_1, u_2 and u_4, it will be more reasonable to consider user u_1 as more similar than the others to user u_3 due to their close ratings for certain items. Various similarity measures will be discussed and applied to this example later in the section.

Example 2.2.

$$\begin{array}{c c}
 & \begin{array}{ccccccc} i_1 & i_2 & i_3 & i_4 & i_5 & i_6 & i_7 \end{array} \\
\begin{array}{c} u_1 \\ u_2 \\ u_3 \\ u_4 \\ u_5 \end{array} &
\begin{pmatrix}
5 & 1 & ? & 4 & 5 & 1 & 1 \\
3 & ? & 4 & 5 & 1 & ? & ? \\
4 & ? & 1 & 2 & 5 & 4 & 5 \\
? & 2 & ? & 5 & 4 & 2 & 2 \\
? & ? & ? & 2 & ? & 1 & ?
\end{pmatrix}
\end{array}$$

This degree of similarities are then incorporated as weights in **equation 2.5** where they are normalized to prevent sum of weights from going out of range (i.e $\sum_{u' \in U} sim(u,u') > 1$). But from this equation, the ratings still require normalization [32] as several ratings could lead to conflicting appraisal to a certain item with the same level of satisfaction or acknowledgement. **Equation 2.6** considers this normalization by transforming $r_{u',i}$ to a mean centred [2] where the average rating of user u' on all items is subtracted from its rating on the target item i. The normalization tends to improve the prediction of preference in situations when ratings are not widely distributed [91].

Another example of the memory-based CF algorithm is the item-based CF which uses a similar concept like the user-based CF but it relies on ratings $r_{u,j}$ by the active user

[2] Desrosiers & Karypis [52] described the process of checking if ratings are either positive or negative by comparing them to their mean rating

towards items j that could be similar to potential recommended item i. For example, an active user u_4 in example 2.2 will accept a new item i_1 based on similar items i_4, i_5 which the active user has encountered and liked. Here, similar rating pattern towards a pair of items will reveal how similar the items are to each other. The rating of an item can be predicted in similar ways like equation 2.5 and equation 2.6 but the difference in the estimation is that the ratings are weighted by the similarity between items $sim(i,j)$.

$$\hat{r_{u,i}} = \frac{1}{|\sum_{j\in I_u} sim(i,j)|}\sum_{j\in I_u} sim(i,j) r_{u,j} \tag{2.7}$$

$$\hat{r_{u,i}} = \bar{r}_i + \frac{1}{|\sum_{j\in I_u} sim(i,j)|}\sum_{j\in I_u} sim(i,j)(r_{u,j} - \bar{r}_j) \tag{2.8}$$

Where: $\bar{r}_i$ is the mean rating towards item i from all users and I_u is the set of items rated by user u that are similar to item i.

The similarity weights in the prediction were described by Desrosiers & Karypis [52] as a means to reveal that neighbours of an active user, to be used in the prediction, are trustworthy. The most popular similarity measures used in retrieval of information are:

- **Cosine vector similarity** where ratings from users or ratings towards items are considered as rating vectors [51, 92].

  For similarity between users u and u' that have rated set of $I_{uu'}$,

$$CVsim(u,u') = \cos(\vec{u}, \vec{u'}) = \frac{\vec{u}\cdot\vec{u'}}{\|\vec{u}\|\times\|\vec{u'}\|} = \frac{\sum_{i\in I_{uu'}} r_{u,i}\cdot r_{u',i}}{\sqrt{\sum_{i\in I_u} r_{u,i}^2 \sum_{i\in I_{u'}} r_{u',j}^2}} \tag{2.9}$$

  For similarity between items i and j that have been rated by set of users U_{ij},

$$CVsim(i,j) = \cos(\vec{i}, \vec{j}) = \frac{\vec{i}\cdot\vec{j}}{\|\vec{i}\|\times\|\vec{j}\|} = \frac{\sum_{u\in U_{ij}} r_{u,i}\cdot r_{u,j}}{\sqrt{\sum_{u\in U_{ij}} r_{u,i}^2 \sum_{u\in U_{ij}} r_{u,j}^2}} \tag{2.10}$$

  Where: $I_{u,u'}$ is the set of items rated by both user u and user u'. Also, I_u is the set of items rated by only user u.

Previous researchers [52, 183] pointed out that the differences in ratings of items with the item based cosine similarity measure were not considered for accurate estimation. This could be resolved by subtracting the mean rating of individual users $\bar{r_u}$ from their respective ratings. This approach was referred as **Adjusted Cosine Similarity**:

$$ACsim(i,j) = \frac{\sum_{u \in U_{ij}} (r_{u,i} - \bar{r_u}) \cdot (r_{u,j} - \bar{r_u})}{\sqrt{\sum_{u \in U_{ij}} (r_{u,i} - \bar{r_u})^2 \sum_{u \in U_{ij}} (r_{u,j} - \bar{r_u})^2}} \tag{2.11}$$

- **Pearson Correlation coefficient** where deviation between ratings from their mean ratings are considered [92].

  For Similarity between users u and u' that have rated set of items $I_{uu'}$

$$PCsim(u,u') = \frac{\sum_{i \in I_{uu'}} (r_{u,i} - \bar{r_u})(r_{u',i} - \bar{r_{u'}})}{\sqrt{\sum_{i \in I_{uu'}} (r_{u,i} - \bar{r_u})^2 \sum_{i \in I_{uu'}} (r_{u',i} - \bar{r_{u'}})^2}} \tag{2.12}$$

  For similarity between items i and j that have been rated by set of users U_{ij},

$$PCsim(i,j) = \frac{\sum_{u \in U_{ij}} (r_{u,i} - \bar{r_i})(r_{u,j} - \bar{r_j})}{\sqrt{\sum_{u \in U_{ij}} (r_{u,i} - \bar{r_i})^2 \sum_{u \in U_{ij}} (r_{u,j} - \bar{r_j})^2}} \tag{2.13}$$

The problem experienced with the memory/neighbourhood based collaborative filtering is the difficulty in predicting accurately the preference of a user towards an item which he/she have not rated before. This problem known as data sparsity [182] has led other researchers to consider the model-based CF as a better algorithm for a recommendation.

2.2.2.2 Model Based Collaborative filtering

Model-based CF which are considered to be the uncommon CF approach requires the user's set of rating data for determining or learning a model that will be used for predicting the user's preference. In this approach, the user-item relationship is a model from latent characteristics of users (preference) and items (category). Types of this approach can be classified as a probability based type where predicting the probability of ratings being specific values correlates to the process of estimating the user's preference for items. An example is a Bayesian Network and clustering(Classification)[32, 135].

In the Bayesian classification approach, a common set of users who have same preferences towards certain items will enable a model to be learnt for a recommendation. The preferences of the items revealed from their ratings $R = (r_1, r_2, ...r_n)$ are conditionally independent as the main idea with the approach is to reveal the distinction between these items in a hidden class variable c_z for the system to recommend accurate items. The model depends on naive-Bayes formulation [32] where the probabilities of membership in classes $Pr(c_z)$ and the conditional probability of ratings given its classes $Pr(r_i|c_z)$ must both be estimated from the training set of ratings. Thus, the probability of an item belonging to a class c_z is given by:

$$Pr(c_z|r_1, r_2, ...r_n) = Pr(c_z) \prod_{i=1}^{n} Pr(r_i|c_z) \tag{2.14}$$

In order to determine the possible class of an exemplar, the probability of each class must be initially estimated for the exemplar to be assigned to the class with the maximum probability. For m Classes $c_1, c_2, ...c_m$, rating r_i can be predicted to be a member c_z iff:

- $Pr(c_z|r_i) > Pr(c_y|r_i)$ for $1 \leq y < m$
- $Pr(r_i|c_z)Pr(c_z)$ is maximized

An example of a Bayesian classification can be observed in Appendix A.1 where the preference for item i_5 was initially unknown. From the evaluation, the target item i_5 will belong to the class of 'likes' for the group of users with their given rating features on other items.

According to previous research [126], the main problem with the Bayesian clustering/-classification method is that it relies on certain assumptions such as :

1. the parameter set r_i seen in Appendix A.1 are mutually independent.
2. all feasible hypothesis (e.g target item will be liked and disliked) are considered for observing the data.

In the proposed framework, the probabilistic method will still be applied as the part of the social network analysis to predict the trust between nodes or determine the probability of future engagement between the nodes in the network.

Another type of model-based algorithm is the latent semantic indexing (LSI)[49, 64] where singular value decomposition (SVD)[73, 182] is applied to the algorithm for improving the performance of the recommendation process in terms of scalability. This type of technique has been confirmed from previous research [182] to be more effective as it reduces the dimensionality of the predicted data from a recommender system. It was considered as an acceptable technique for text classification in retrieving hidden information from documents [49, 64]. The concept of the LSI/SVD is focused on reducing a high dimensional dataset containing the relationship between users and items to a low dimensional space where substructures are generated to clearly reveal how items can be categorized based on factors inferred from user's feedbacks (i.e. either explicit or implicit ratings).

The SVD algorithm that can be applied to a recommendation is a common matrix factorization method where an $m \times n$ matrix P is decomposed into three different matrices, U, Σ and V^T.

$$P = U_{m\times m} \cdot \Sigma_{m\times n} \cdot V^T_{n\times n} \tag{2.15}$$

Matrices U (also known as left singular vectors) and V (right singular vectors) are considered to be orthogonal matrices [3] that both have their columns as the eigenvectors of PP^T and P^TP respectively. According to Sarwar et.al. [182], U matrix represent the latent features of users in accordance to their frequent interaction with items while V matrix represents the latent features of items in accordance to the users that have interacted with or utilized them. Σ (also referred as scaling matrix [73]) is a diagonal matrix having positive entries known as singular values that are the square root of eigenvalues from either U or V and at least one of these values must be different (i.e. anisotropic scaling).

The application of SVD algorithm (See algorithm 1) on a matrix with data representing the rating to items v by users u can be observed in appendix A.2. This algorithm was based on the SVD computation process which was demonstrated by Hampton [83]

[3] A matrix M is said to be orthogonal if $M \cdot M^T = M^T \cdot M = I$, where Matrix M^T is the transpose of matrix M and I is an identity matrix which has main diagonal entries of one with every other entry of zeros.

Algorithm 1 Singular-Value Decomposition Algorithm

1: **procedure** GIVEN: $m \times n$ Matrix P
2: Find the Matrix product PP^T.
3: Determine the m eigenvalues λ using $\left|PP^T - \lambda I\right| = 0$
4: Take the square root of each eigenvalues to obtain the singular values to be elements in matrix Σ.
5: Substituting the eigenvalues λ into $PP^T - \lambda I$ and resolve to matrices.
6: Apply the Gauss-Jordan elimination method [8] to obtain the reduced row echelon form of the matrices.
7: Find the unit-length vector in the kernel of these matrices to obtain m vectors $(\vec{u_1}, ..\vec{u_m})$ of matrix U.
8: Find the Matrix product $P^T P$.
9: Repeat similar process from step 3 to step 7 for n eigenvalues Σ from $P^T P$ to obtain matrix V

where he applied both the Gauss-Jordan elimination method and the method in determining the kernel of a matrix [8]. The unit-length vectors obtained from the combination of the methods then reveals the vectors that make-up either matrix U or matrix V. Hampton [83] was able to prove that this method was more stable than the Gram-Schmidt algorithm [8] that have been used in previous research.

In this research thesis, the prediction of preference to certain items i for a user u_k will be described using a prediction metric described in previous research works [98, 182] which considered all the generated matrices from a decomposed matrix with relationship information of the users with items.

$$\hat{r}_{u_k i} = \bar{r}_{u_k} + U_k \times \Sigma \times V_i^T \tag{2.16}$$

Where: $\bar{r}_{u_k}$ is the average rating of user u_k, U_k is the row vector for user u_k from the U matrix and V_k^T is the column vector for u_k from V^T matrix.

The model-based algorithms are more efficient than the memory/neighbourhood based algorithm as less time is required to observe part of the dataset (model) instead of the whole dataset. However, it is possible that since the whole data are not being used, the prediction might be inaccurate.

2.3 Feedbacks for Recommendation

As it has been revealed in the previous section that recommendation algorithms (both content-based recommendation and collaborative filtering recommendation) can rely on explicit feedbacks (ratings) from users in order to predict the preferences of a target user, we also need to know the importance of implicit feedbacks on recommendation algorithm. With the content based recommendation, the target user must be active towards some items for the evaluation of the user's preference to be carried out, while for the collaborative filtering algorithm, some number of similar users to the target user needs to be active for the target user's preference to be predicted. But if users are inactive, how can the preferences be predicted?

As mentioned in chapter 1, users might decide not to be active in providing explicit feedbacks for several reasons. These includes:

1. The user's non-interest for certain items even though they might have interacted with them in the past.
2. The user's fear for compromised privacy.
3. The user's fear of being judged by other users.

It might be possible to resolve the cold-start problem by relying on information based on influence concept where a person can be influenced to change their behaviour of revealing preferences towards items only by those they consider to be trustworthy. The research in this thesis focuses on this concept where it is possible to encourage an inactive member of a social group to become active in the presence of trustworthy members. It is therefore important to understand trust and its' impact in a social group. Previous researchers [201] had also revealed that trust in a system is inspired only by an effective recommender system.

2.4 An insight on Trust

In every part of everyone's life, trusting decisions are made where risks are involved in them. We sometimes decide to communicate with someone without actually having an idea of who is on the other end of the communication. There is the risk of being deceived by others who may cause harm such as loss of benefits or unsatisfactory outcomes. Apart from deceiving cases which lead to harm, there are cases where someone who might either be ignorant or uncertain about the truth can later be convinced or persuaded by others to accept the truth. There are also cases where a system or an individual exploits other individuals' ignorance, making them succumb to offers. A deceiving act by one with the intention to persuade another who is ignorant or uncertain should shield light and provide understanding on the truth from experience.

Gambetta[71] had previously pointed out that ignorance or uncertain situation requires trust as support for decision-making and this trust being referred to as belief has more priority over the benefits from cooperation or engagement. There is a slight difference between belief and trust, as belief is the view that leads to the acceptance of something's existence or truthfulness without any reasoning or evidence to support it while trust can be seen as a kind of belief either based on a person's direct experience or the social-reasoning concept [205] where other people's experience can be used in confirming the truthfulness of something. This definition then shows that reasoning or evidence plays an important role in decision-making as there must be a relationship between the past and future decisions. There are several other definitions of trust which will be discussed below. Some of them are related or mean the same thing while some are slightly different

2.4.1 Defining Trust

Various definition of trust in diverse fields all point to trust being an important means of evidence for making a decision. Marsh[123] had previously described trust as a useful tool for a person's decision-making with evidence from the experience of others' behaviour. Gambetta[71] defined trust as a subjective probability of an entity predicting another entity's action or behaviour which could determine the possibility of their engagement or cooperation. Both researchers in [71, 123] described trust to be of free-will and not forceful in making someone wrongfully accept unwanted items. Harwood[87] described a decision problem to trust as "utilitarian" concept where the consequences of the decision to be made will be considered before any judgement, that is, utility based on conditions or situations plays an important role as support in trusting decisions.

The definition given by Barber [16] and Rotter [180] are closely related as they both point to reliance based on the expectancy of individual's performance, competence or adequate behaviour. Rempel [174] also referred to trust as a predictive rating assigned to events reoccurring in the future. Trust can be seen as a feature in social environment[16, 116, 123] where the expectation of the future must be determined. Marsh [123] listed three kinds of expectation from the research work by Barber[16] which include:

- expectation of competence in performing roles, where there is the belief that a person will accomplish a task to a satisfactory level.
- expectation of persistence, where there is the belief that a person should always behave in a certain manner or pattern.
- expectation of selfless interest, where one places the interest of others before his/her own.

People will fail to trust if disappointed with all these expectations not being met by others within a community. The strength of trust relies on these expectations which maintain a community based on cooperation. A social community with trust can be seen as means of reducing the complexity that leads to disagreement when diverse opinions, views or goals exist amongst people. A group cannot be formed when there are conflicts based on diverse opinions of various individuals. Previous research [116] revealed that the best way of building a social and cooperative community that will reduce complexity is by ensuring the existence of trust. An exiting community might shrink in size or be destroyed when the opinions or goals of individuals gradually deviate from each other at different times, which then shows the need of trust in generalisation [16], as this enables someone to view an uncertain situation in a general view of others. Each situation of events within a community needs to be considered in order to determine trust.

Also in the view of generalized expectancy, trust can be considered as an adaptive tool in a social community where new members will need to determine if they fit into the group. Most new potential members may decide to join the group based on their high interest being similar to existing members of the group while some existing members may decide to break from the group when they disagree on the opinions of others towards certain subjects or items. Members may decide to remain with the group if they have continuous familiarity with other members or an exemplar of the group [155]. Luhmann [116] previously stated that trust exists in familiar communities where the change in behaviour affects the state and structure of the community. This behavioural change might be due to unfamiliar situations, such as new items being introduced or presented to members of a community.

Research works in the communication field [74] described trust as a communication tool which cannot be transferred through the same path of a particular communication. This means that trust given to an information source is not based on a single communication path with an individual but it should be based on multiple communication paths. The communication path between individuals cannot be easily controlled without trust concept. In our complex world today, a person needs more than one person to confirm the reliance or competence of items. Gerck [74] referred to two main kinds of individuals that support inducing trust to another: trusted witnesses who provide testimonies of their experience with certain actions and trusted inducers who see the current situation of an individual is facing and then provides a solution.

With all these views on trust from various researchers in diverse fields [19, 48, 71, 74, 116], the computation of trust is being affected. As trust is usually misrepresented or misinterpreted, it is still very difficult to measure it. Marsh [123] previously asked what it means for someone to trust another person. Other researchers [40, 199] referred to this trust as generalized or social trust as it is the belief of a person on how others will behave towards him or her in a social community. Does it mean that the trust of a person towards another is more than 60% or any threshold value? The generalised trust is not clearly expressed as we cannot really judge or make an accurate decision by relying only on this information. There is the need to consider the context in a trusting relationship for a better understanding of the information before making any decision. For example, the trust of a person **A** towards a person **B** in delivering a lecture will not be the same with the trust based on person **B**'s ability to play for person **A**'s football team. Here, competence seems to be the first and most important feature to be considered in measuring trust.

The generalized trust model represented in figure 2.1 is slightly similar to that of Harwood [87] but the persistent and selfless interest property was considered as relevant in trust computation.

Trust can be used as part of a predicting measure as it is possible to apply the trust information in determining the preferences of entities based on the behaviour of the entity. The persistent behaviour of the entity will reveal accurate preferences of the entity, provided the behaviour do not vary at any time.

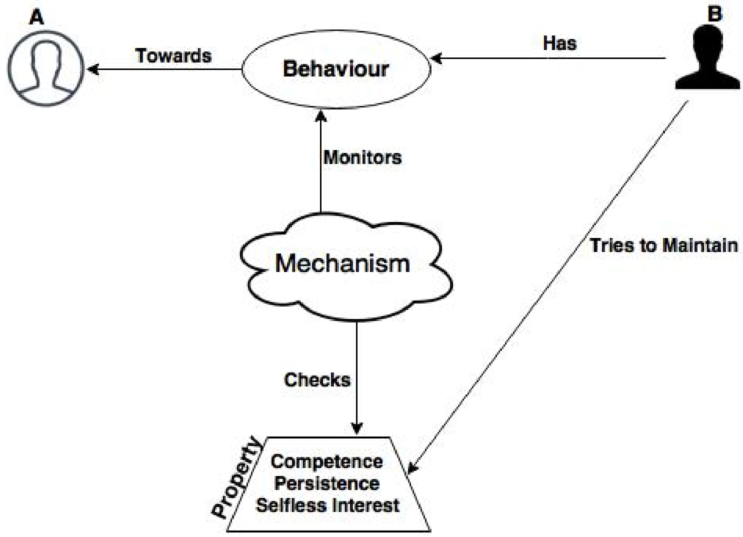


FIGURE 2.1: Generalized Trust Model

2.4.2 Trust Based on Conformity

As there is misrepresentation of trust, there is the need to understand the various types of trust in order to know the best way to represent trust in a social community. Types of trust can be drawn out from the behaviour of individuals. Previous researchers [211] had earlier viewed the behaviour of one person towards another as a source for trust types. Behaviours from both parties (interactors) are useful as there is need to know why they both acted or made decisions in a certain way. Deutsch [53] had earlier presented an example with a story of *The Lady or the Tiger*[4] where a princess' suitor is discovered by a king. The king is annoyed with the status of the suitor then throws him into a dungeon with two exit doors which he must choose one to exit the dungeon to avoid being punished by the king. One of the doors has a hungry tiger behind it while the other has a beautiful lady who the princess considers to be a rival that also has affection for her suitor. The princess being aware of what will be behind both doors then points to a door for the suitor to choose. Readers of this story will ponder on what door the princess actually chose for the suitor. Did she suggest the door with the beautiful lady to ensure the survival of the suitor from being killed by the tiger or the door with the tiger believing that she might lose her suitor completely to her rival and so she will rather lose the suitor to the tiger? On the other hand, the suitor will always trust the princess' suggestion without knowing her intention of making those suggestions for him.

[4] Adapted from a publication in The Century magazine written by Frank Stockton in 1882.

Trust drawn out from the behaviour of individuals can enable one to predict the future actions of the individuals or others who are relying on the individuals for their own decisions. The intention or reason behind an individual's action, for example, the intention of the princess from the story, could have been a knowledge source for the suitor to decide accurately on whether to trust the princess or not. If the suitor in the story had knowledge of the princess' intention (e.g desire to lose him to the tiger rather than the lady) he might want to save his life by not following the princess's suggestion and then he will rather choose the other door with the lady. But from the story, it seems that the suitor has a general trust where the suitor has never experienced a situation like this before and so he will always believe that the princess has his best interest to always keep him safe from any harm. This type of trust in the story can be referred to as **compliance based trust** as the suitor makes his choice based on his belief that the princess' suggestion will be rewarding to him.

Compliance based trust derived from the conformity type previously described by Kelman [107] as a situation where a member of a community conforms to opinions of another member or the whole community due to his or her fear of failure in having an accurate opinion to a task. An example of this type of conformity is the Asch's experiment [127] where a person gives an incorrect answer to a line judgement task due to social pressure from a group. The person in the experiment is unaware that the other participants actually have "scripted" behaviours to observe and induce their opinion to the person. This person who conforms to the group's opinion without knowing their intention has the belief that the group is trustworthy to always provide the right answer to a task. The person also conforms to the group's opinion to avoid being mocked or ridiculed by others. In the story of the lady or the tiger, the suitor decided to make a choice to show his braveness to the princess and avoid being ridiculed as a fearful person by the king or others present. Also in the Asch's experiment, the main participant decides to follow the group of other participants to avoid being judged and described as a naive person by them. This type of trust can be described as forceful, as both the suitor and the main participant still reserve their own opinion or beliefs towards their individual tasks, even though they follow the advice or opinions from the princess and other participants respectively.

The trust sometimes seen as forceful in certain situations can have either a positive or negative consequence. The positive consequence can be seen when a decision has to be made by entities to avoid worse or negative conditions than the new condition based on the decision to be made. Deutsch [53] referred to the trusting behaviour in deciding as despair where there is no hope in which the entities find themselves. An example can be seen in the previous story where the suitor will be forced to make a decision and trust the princess to avoid being punished by the king. Another example can be seen in the Asch's experiment where the main participant will be forced to provide an answer by trusting the group's opinion to a given task in order to avoid being mocked or labelled as a naive person.

The negative consequence of a forceful trust can be seen when a decision is made due to no or insufficient information to support the decision-making which might lead to unfavourable outcomes. There is the need for individuals to know the reason for another individual's or group's induced behaviour on them before making any decision on a subject matter or an item. Individuals who have no information will be vulnerable to others that will persuade them to conform to their own opinion for selfish reasons. For example, an e-commerce company might force customers to purchase new items, making them believe that these items are similar to their previous purchases and therefore they will be beneficial to them. It could be that the company's main intention is to make a profit from sales of the items and not to recommend the suitable items for the customers.

Conformity theory [130] states that individuals who lack knowledge will conform to a group believed to be more knowledgeable than themselves. The conformity of individuals is based more on the consistency and the idea of behind the induced behaviour from the group that requires the individuals to change their views permanently to the group's general views. We can then refer to this type of trust as **informative trust** as knowledge can be gained from a mass group believed to have similar opinions. This type of trust was seen in the Sherif's experiment [130] [5] where an individual who seems to be uncertain about the answer to a solution will always conform to a group they find themselves in.

Most people decide to trust others because of their closeness or proximity to themselves. They do not consider the intentions of the other party believing that they will never be deceived and disappointed by the other party. Conformity theory [130] referred to this type of conformity as *identification* due to the fact that individuals conform to a group just to be identified with a group. The individuals who conform to a group do not necessarily have to change their views completely to fit with the group as they are expected

[5] Sherif carried out the experiment by grouping two or more persons that have a similar opinion (estimates) from an observation on the movement of light on a screen even though it may seem stationed on a position.

to always carry out a self-evaluation on themselves to determine their fitness in the group. Previous research [63] had revealed people identifying themselves with groups based on their self-evaluation. Comparison test carried out during the self-evaluation process by individuals actually checks if there are either similarities or differences in their roles, beliefs or attitudes with that of others in a group. Also, the self-evaluation reveals the consequences of changing or preserving behaviours. An experiment carried out by Zimbardo [129] showed that the participants who had roles of either a prisoner or prison guard fitted quickly in their various roles within the prison community. The prisoners carried out self-evaluation on themselves to observe the prison guards' behaviour towards them and to compare their roles with that of the guards. This stimulated them to comply with the prison guards on the prison rules and therefore building their trust with them. The type of trust here can also be described as forceful as the prisoners were dependent and obedient to the guards after they were harassed and humiliated by them. Even though the prisoners were forcefully conformed to the rules of the prison community, they still had free-will to decide either to obey or disobey the prison rules.

Trust should not be considered as a forceful belief but it should be seen more as a belief based on free-will. The proposed concept for guiding users and recommending accurate items for them will adopt the informative trust concept which is based on internalisation conformity [130] where the intentions of 'guardians' (trustees) or recommenders are evaluated before acceptance of the guidance or recommendation by the users.

2.5 Computing Trust within a Social Network

The computation of trust will still remain difficult to accurately achieve due to the diverse representation or view of trust that exist in various fields. Mui et.al [142] previously revealed that reputation could be a means of measuring trust as the reputation is considered as a function of feedback rating [176]. Houser & Wooders [97] described reputation as the probability to measure the competency of an entity to fulfil an expected task or action. They provided an example of an auction case scenario where a buyer may have the probability of making payments if he/she wins or a seller having the probability of presenting the auctioned item to the buyer when payment has been received. This previous research did not reveal how trustworthiness of an entity could be computed from their reputation.

Several researchers had previously introduced the computation of trust in a social network. Golbeck [78] was amongst the first researchers that introduced the computation of trust between people in a social network where their personal traits and preferences based on their personal social profile are determined. Trust was described in [78] as a 'label' on a tie between two persons who are considered similar. Sinha & Swearingen [193, 201] had earlier pointed out a recommender system will be trusted if there is similarity in user's preference. Vedula et.al.[207] proposed the computation of trust based on the structural properties of a social network and the content from an interaction of members in the network. Karmar et.al.[103] focused on the computation of trust based on the authenticity of files sent or received by peers.

The proposed trust metric discussed later in the thesis will reveal how trust can be evaluated without the content from interaction as activeness and similarities in preference could be considered to be more relevant measures for trust. The activeness and behavioural pattern (habits) of a node in a network will determine if the node actions are genuine. With the theory of interpersonal behaviour, it is possible to model the trustworthy behaviour of an entity. This will be described later in chapter 3, where the activeness of a node will be revealed to be a condition that could affect prediction of the node's behaviour.

According to Triandis [203], a behaviour can be considered as a function of habits, intention and situational condition. But the research of this thesis will not consider intention as a behaviour could be understood from both situational conditions and habits. Situational conditions such as trust [113] can predetermine the intention of entities. For example, a person x will accept the opinion of another person y if person x has trust for person y, believing that person y's intent is to genuinely support him or her. The habits observed from past activities could also reveal intentions as the situational conditions that affect intentions are estimated based on the habits from frequent activities. Gardner [72] further revealed that intention might not necessarily be associated with a behaviour as an impulse from a situational condition (e.g. Trust) could be cause for the behaviour.

2.6 Analysing Patterns in Social relationship

As most users might not realise the importance of being active in a system, it is then important to know the credibility of the system and other users in the system as it might enable them to see clearly how the system can be beneficial to them. According to Ricci et.al.[178], if there is no information (e.g. Activities towards items) on a target user, the recommender system will recommend the same items that will be recommended normally to an average user.

In order to measure the trustworthiness of any entity in the social network, there is the need to identify their pattern of relationship with other entities. Based on real-world situations, two persons will decide to trust each other only if they are familiar with themselves and they also consider themselves to be similar. Both persons need to rely on each other's opinion in order to make their individual decisions. Similarity which determines the strength of a relationship between entities has been considered in several types of research as a means to yield attraction[132], to identify groups (cluster analysis) [118] and to predict new items for users in recommender system [90, 164]. Previous researchers [84, 146, 211] had revealed that patterns could be examined through the measure of similarities between nodes based on their relations with others.

In order to evaluate the similarity between two entities, it is important to measure their distance (also known as dissimilarity) as it reveals their disagreement, incompatibility or distinction based on certain features. Each feature represents the coordinate of an entity in the feature space. In the proposed research, the path distance considered from a social relation between nodes were measured based on their interaction pattern observed from structural equivalence where they might have common connecting ties with other nodes. Details of this analysis will be discussed in chapter 3.

Hamming distance measure which can be applied to nodes' sequences or patterns of equal length is usually considered for the measure of similarities between nodes in terms of their frequent actions towards common nodes. Newman[146] pointed out that there is a correlation between the Hamming distance and Euclidean distance since they both consider the distinction between two nodes. When using the Hamming distance, the patterns of two nodes are compared to carry out checks for the minimum amount of errors that could be corrected to change one's pattern of activity to the other [82]. According to Richard Hamming's geometrical model for error detecting and error correcting [82], the distance $d(i,j)$ in a space of 2^n points for n-dimensional cube (i.e a cube is a 3-dimensional object in geometry) is based on the least number of edges for a travel path between point i and point j in the 3-dimensional space.

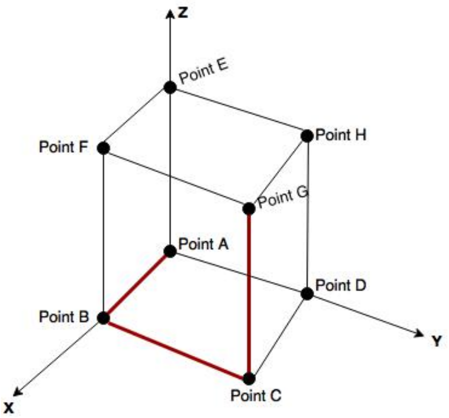


FIGURE 2.2: 3-dimensional space of the Cartesian coordinate system

From Figure2.2, the Hamming distance between **point** A and **point** G is **3** as the minimum number of paths for point A to reach point G is 3. This relates to the closeness centrality as the distance between A and G irrespective of the path taken by A to reach G will still be 3; which also means that for A to be like G (or vice versa) it needs 3 corrections to be made.

Hamming distance could also be observed and measured when comparing two vectors **i = [8, 9, 0, 2, 1]** and **j = [2, 9, 1, 4, 7]** that assigned rank of importance to **5** items.

TABLE 2.2: Vector ranking on 5 items

$Node$	$item1$	$item2$	$item3$	$item4$	$item5$
i	8	9	0	2	1
j	2	9	1	4	7

The Hamming distance between i and j is determined based on the number of replacement required to change a node's rating pattern to the other node's pattern. Therefore, the distance between i and j is **4** as there are **4** distinction between i and j or there are **4** replacements expected to be carried out on either vector i or j for both vectors to be equivalent or completely similar.

Previous research [222] considered 'importance' rating value for each point in determining the distinction between points $H_{(i,j)}$, where the hamming distance between two points i and j is normalized:

$$d(i,j) = \frac{H_{(i,j)}}{N} \tag{2.17}$$

Considering the previous vector example with equation 2.17,

$d(i,j) = \frac{4}{5} = 0.80$

This Hamming distance measured could also be correlated to the probability of these two points being dissimilar as it still satisfies the four properties in distance measure which includes:

- $d(i,j) = 0$ if and only if i = j, as every value or symbol possessed by i is also possessed by j (Identity of indiscernibles).
- $d(i,j) \geq 0$ if $x \neq y$ where there is at least a non-zero value possessed by one of the nodes.
- $d(i,j) = d(j,i)$ (Symmetry).
- $d(i,k) \leq d(i,j) + d(j,k)$ (Triangle Inequality)

The similarity between nodes from their normalized hamming distance is given by:

$$S_{i,j} = 1 - d(i,j) \tag{2.18}$$

Where: *Similarity of* 1 *indicates that both i and j are definitely equal while similarity of* 0 *indicates that i and j are definitely different in pattern.*

From the above example,

$S_{i,j} = 1 - 0.80 = 0.2$

This means both i and j are not exactly similar in their pattern of rating the item. It seems that the measure favours nodes who have similar high frequency of no weight degree (i.e. weight = 0.00) on ties with other neighbours (i.e. Items in this case). That is, it will be so unclear in cases where two nodes who don't share any ties with any neighbour and the Hamming distance measure will result to both node being similar in pattern based on their similar weight degree (i.e. weight = 0.00). Therefore, Hamming distance measure cannot be considered suitable for measurement of similarity in the research work of this thesis as this measure is more suitable for ordinal variables where the number of distinction between variables are considered.

Another strategy that could be used in analysing dissimilar pattern between nodes is the Cosine similarity [146, 221] where structural equivalence is based on the pattern of social ties between the nodes. From a 3-dimensional plane shown in Figure2.3, the smallest angle separating point C and point G is considered in measuring how similar both points are alike based on their features (i.e. quantitative variables).

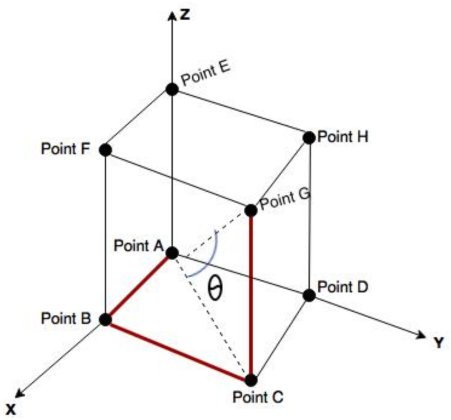


FIGURE 2.3: 3-dimensional space with an angle separating two points

Previous research [84, 211] considered an adjacency matrix in obtaining insight on how the position of a particular node to others could be used in determining if the node is similar to these other nodes based on structural equivalence in the network. Based on other researches [146, 171, 221], cosine similarity between two nodes is considered as a preferred measure where the dot products of the features between the two nodes are divided by their magnitudes; this measure considers the common neighbours that other pairs of nodes share in the network.

$$S_{I,J} = cos\theta = \frac{I \cdot J}{||I|| \cdot ||J||} \tag{2.19}$$

$$= \frac{\sum_{n=1}^{m} I_n J_n}{\sqrt{\sum_{n=1}^{m} I_n^2 \cdot \sum_{n=1}^{m} J_n^2}}$$

***Where**: I_n and J_n are features of vector I and J.*

The similarity values from this measure lie between 0 and 1 where a value of 0 indicates that the two nodes do not have a common neighbour while a value will indicate that both nodes are similar as they definitely have the same neighbours. Newman [146] pointed that if one or both of the nodes have a degree of zero, their similarity will be taken as zero in accordance with the convention in a social network.

Using the cosine similarity measure on the previous vector example,

$$S_{i,j} = \frac{8*2+9*9+0*1+2*4+1*7}{\sqrt{8^2+9^2+0^2+2^2+1^2}\cdot\sqrt{2^2+9^2+1^2+4^2+7^2}} = \frac{112}{150.50} = 0.74$$

This similarity value between i and j clearly shows that they are similar to a certain degree as they both have some common neighbours (i.e. they have a connection to most of the same items) in the network. This also implies that i and j are reachable to a certain degree to rely on or trust each other.

As earlier discussed in section 2.2.2.1, Cosine similarity measure is referred not to be a suitable measure for measuring the similarity between users in determining their neighbourhood and the importance of neighbours. It was suggested that Pearson correlation measure seems to be more suitable as it considers the use of deviation in users' rating to determine how similar the users are to themselves.

Wasserman & Faust [211] had earlier pointed out similarity measures do not always provide the same results from the same relations. It was stated that Euclidean distance measure is not a suitable measure for structural equivalence when considering similarity in patterns as it only measures the identity of ties where each degree of ties is considered in the evaluation. The results from the measure will not reveal the pattern information between nodes but only reveals the potential ties between the nodes based on their tie weights[60].

2.7 Activeness and Centrality in Social Network

Before considering the proposed approach in improving recommendation, there is the need for a social network of users or entities to be analysed as the social network is considered to be an example of a platform using the recommender system. One important measure from all measures in social network analysis discussed in Chapter 1 of this thesis is the centrality measure [211]. A metric space, referred by Mendelson [134] as a space where pairs of points (nodes) from a set, are measured based on their closeness centrality [65, 211] which determines how efficient a node will be reachable to the others in its set. The closeness centrality concept is useful in the proposed approach in motivating inactive nodes in a network as it will define the relationship between similarity based on path or reachability and trust based on neighbourhood.

From a social network, nodes that interact with each other might be connected by a central node. This central node could be of the same or different class (e.g. human class and item class) with other nodes that are connected to it. For instance, two nodes of the same class (e.g. humans) could be connected together via their link to another class (e.g. topic, subject or item) of nodes. Borgatti et.al [28] and Opsahl [151] referred to this type of structure as two-mode network. Newman[146] described a recommender network as a two-mode network (i.e. Bipartite network [211]) where there are two types of nodes, one indicating the items that other type of node directly interacts with (shown as connecting edges in the network graph 2.4). Newman described the network to be a useful tool in collaborative filtering algorithm for a recommendation as it is easier to analyse from the network the items linked to neighbours of a particular node that could also be preferred or relevant to the node.

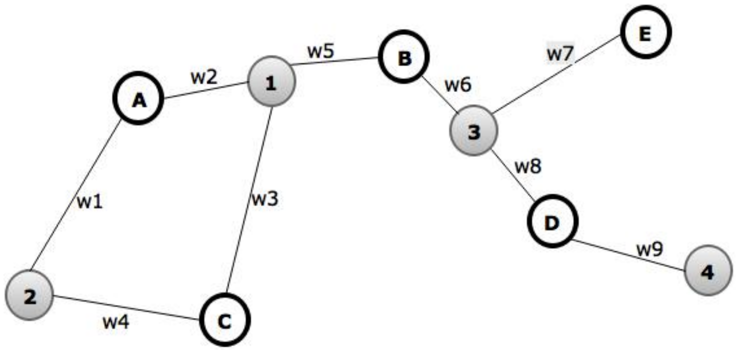


FIGURE 2.4: Two-mode network

The activeness of nodes from the same set (i.e node A, node B, node C, node D and node E) can be determined by the number of actions towards another set of nodes (i.e node 1, node 2, node 3, node 4). Granovetter [80] and Newman [146] pointed out that a function of duration or exchange of services could be the strength of an edge(i.e. a tie weight) between the nodes. The overall weight between two nodes of the same class can be determined when we consider their individual weights towards other nodes of a different class that links them together. An example could be observed in a relationship between two players who might have played for common teams in the past and their weights towards each team will be the number of matches they have played for the respective team. The overall weights between the two players will then depend on the number of matches each player had played for the respective teams.

Opsahl [151] suggested that a two-mode network still needs to be transformed (known as projection) to a one-mode network in order to carry out further analysis that will reveal the relationship between nodes of the same set (A, B, C, D, E). Padron et.al [156] also suggested that the transformation is required to predict potential competition and successful interaction. Two nodes of the same set are linked together if they have a connection with common nodes from another set. Opsahl [151] defined the transformed two-mode network by two methods which include: Sum and Newman's method [147].

Applying the sum method of transformation on a two-mode network, the sum of weights from all ties a node (e.g node A from Figure 2.4) shares with another node (e.g node C from Figure 2.4) of the same set towards nodes of a different set will be the weight on the directed tie towards the node in the transformed two-mode network (i.e One-node network).

$$w_{ij} = \sum_t w_{i,t} \tag{2.20}$$

***Where**: $w_{i,t}$ is the weight on ties that node i had towards a context node t which is also linked with node j.*

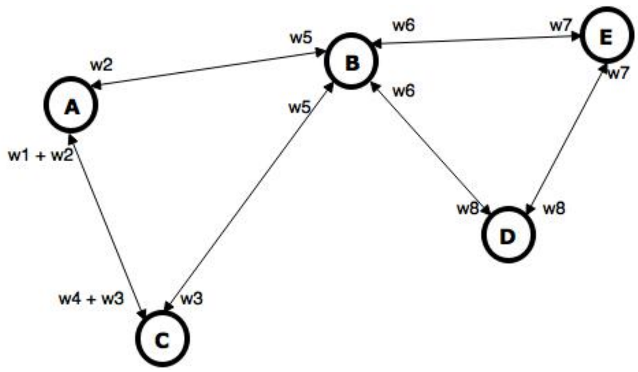


FIGURE 2.5: One-mode Network using Sum Method

The weight to the target node derived from the sum method correlates to the degree of a node that Wasserman & Faust [211] described to be the number of other nodes adjacent to it . The ties weight of a target node from a set can be seen as the number of nodes from another set which is interpreted here as the sum of interaction between the two nodes of the different set. According to Opsahl [151], this weight is then directed towards the target node by another node of the same set that share ties with nodes of another set

With the Newman's method of transformation on a two-mode network, the weight on the ties in a transformed two-mode network based on the fact that the number of ties to a particular node will affect the strength of their ties [147]. For example, if there are a lot of users purchasing a particular item, their ties will become weaker as they may decide to purchase different or unique items used by few users.

$$w_{ij} = \sum_t \frac{w_{i,t}}{N_t - 1} \tag{2.21}$$

Where: *N_t is the number of nodes that are connected to the context node t.*

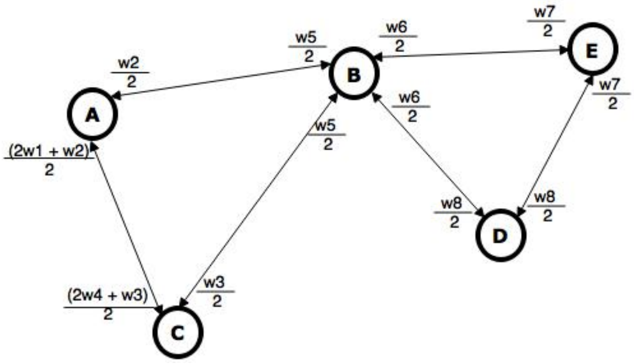


FIGURE 2.6: One-mode Network using Newman's Method

Degree centrality is considered to be applied to a projected two-mode network as where a certain node could be centrally based on its number of connection to other nodes based on their links to context nodes. Wasserman & Faust [211] considered this central node as the most active node amongst other nodes. From figure 2.5 or figure 2.6, node B is considered to be central as this node has more ties in the network. Also, the two-mode network could be analysed with the degree centrality as links between a node of a certain set and the context nodes are analysed. From 2.4, node A, B and C are considered to be the most active nodes as they had more ties with the context nodes.

Nepal [145] considered the count of interaction towards a context as a means to determine how popular a target member is within the social network. It also reveals how well other members of the social network will trust the target member based on the activeness of the node. Vedula [207] described the activeness of a node as the rate of the node's influence towards another node; and it was also revealed that a popular member of a network will not get into a competition with another popular member. The popular members will then have to separate into their individual group to avoid conflict. It is therefore important to have an understanding of clusters where influential members based trust can be identified.

The next section will reveal the reason for not considering cliques in the proposed framework where nodes of a social network need to be accurately identified to their groups with an influential member.

2.8 Cliques from One-mode network and Two-mode network

Based on the concept of the real social world, entities tend to be clustered together into completely connected groups which can be referred to as cliques. Friends of an active person will become friends with each other and create a one-mode network structure. On the other hand, neighbours or contacts of an active node in a two-mode network cannot be completely connected together as the same measure used on a one-mode network for determining the structural ties between nodes cannot be applied directly on a two-mode network [28, 151].

As earlier discussed, a projected two-mode network (i.e. One-mode network) will reveal the connection between two or more nodes of a certain set based on their relationship with common nodes (i.e. items such as topics/subjects) of a different set from the original two-mode network. Newman [146] stated that the projected two-mode network will form cliques based on the relationship from the two-mode network but the main disadvantage with the projected two-mode network is that the network will not actually reveal the information of common nodes shared by the nodes of the particular set.

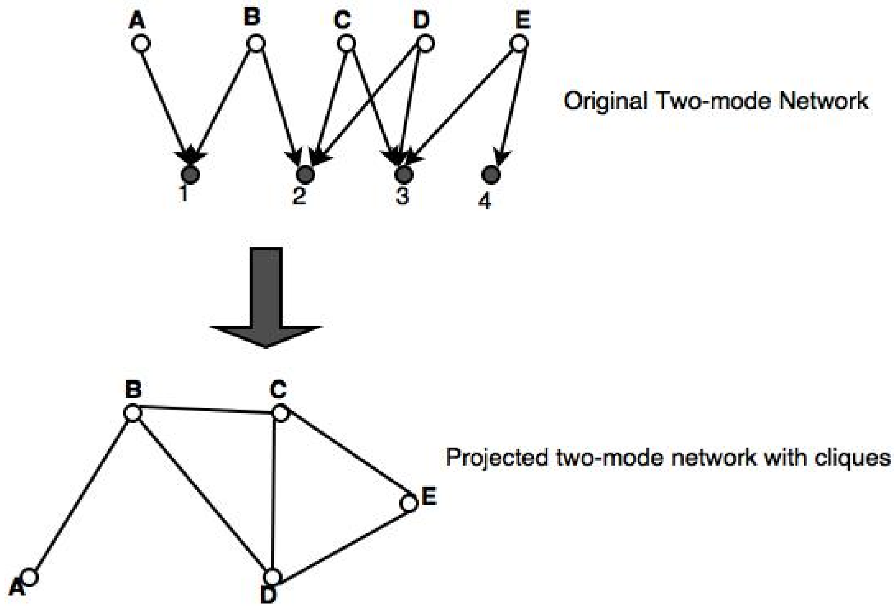


FIGURE 2.7: Cliques from a Projected Two-mode Network

With the given two-mode network in 2.7, nodes (i.e. nodes A,B,C,D,E) of a certain set have relationships with nodes of another set(nodes 1,2,3,4) with no clique present but in the projected two-mode network cliques are formed. The number of cliques formed can be limited by the presence of some inactive nodes that have no tie with any of the nodes from the other set. The cliques here are B,C,D and C,D,E as they are completely connected subgroups where both node C and D are members. Certain nodes belonging to several cliques will then lead to consensus problem where nodes in those cliques might have conflicting opinions towards the other set of nodes.

2.9 Cluster Analysis on Social Graphs/Networks

As earlier revealed, there are cases where a node or set of nodes could belong to several different cliques [211] which might lead to the issue with lack of consensus where there might be different opinions on subjects/items in the different cliques. Members of a clique are expected to have similar interest or opinions on items as this will enable them to feel accepted and identified by the group. This clique formation based on similar interest or opinion was described in [131] to ensure isolation amongst cliques which will build confidence amongst members of individual cliques who might require support.

Previous research [158] has suggested the use of clustering as a possible means to resolve conflict or disagreement within a social group. The clustering is usually applied to identify groups (clusters) after ties between pairs and exemplars have been detected. The formation of clusters is based on similarity measures (or distance) which determine the members that belong to each cluster and ensures that members of a particular cluster are closely similar than to members of other clusters. The similarity that is measured within a clustering process will improve the efficiency of determining the distances between entities (i.e. users or items) which are normally useful in recommendation techniques such as Collaborative filtering and content-based filtering. Various model of clusters are based on:

- Distance connectivity where the distance (similarity) between pairs of data points will cause the points to either merge or split into clusters. An example is Hierarchical clustering [121].
- Centrality where a distance between a central member (exemplar) that represent a cluster and potential members determines the members of the cluster. An example is the K-means clustering [104, 118].

- Density where the clusters are determined based on how closely-packed or dense an area of data points appears to look. An example is the Density-based spatial clustering of applications with noise (DBSCAN) [59].

- Probability distribution where the probability of membership to clusters can be determined. That is, individual data point will belong to each cluster with certain probabilities. An example is the Expectation Maximization (EM) clustering [29, 50].

- Graph connectivity where a clique formed from a subset of nodes is considered as a cluster which is based on the degree of social ties for each node with other nodes in a social graph. An example is the Highly connected subgraphs(HCS) clustering[86, 131].

Popular clustering technique includes the K-means clustering that requires a pre-specified number of clusters and the DBSCAN that requires a pre-specified minimum number of members in a cluster to be generated. This pre-specification in both clustering techniques could be considered as an inaccurate procedure involved in generating the clusters as the number of clusters or members in a cluster is constrained by any random pre-specification. Even though clustering technique such as Hierarchical clustering does not require pre-specification for the number of clusters required, the technique still needs a cut-off point on the dendrogram [6] to generate the clusters [121]. This procedure is also considered inaccurate as the clusters can be constrained to be viewed from any cut-off point.

Expectation-maximization clustering [119] considered as probabilistic approach, assigns each data point to various clusters but data points might have certain degree of membership to various clusters (fuzzy clustering) [23], that is, a data point might not belong to a single cluster if its degree of membership in a cluster is lesser than other data points. For example, a player can be described as a member with 50% fitness in a football team but we cannot state that the player has 50% membership in the team.

In this research, the graph connectivity model will be one of the models that will be explored in chapter 4 of this thesis as the research focuses on the network structure of social ties. Other clustering techniques to be considered are the Markov clustering and Affinity propagation as both have been referred in previous researches [56, 67,

[6]Hierarchical clustering is usually viewed as a dendrogram where clusters are arranged in a tree-like form based on similarity between pairs of data-points.

184] as clustering model on graphs. Social factors are expected to be derived from the estimation of the clusters with these clustering models.

2.9.1 Highly connected subgraph Clustering

Hartuv & Shamir[86] had previously revealed that elements being the vertices of a social graph will be considered similar to be members of the same cluster if their social ties (based on similarity) are more than half of the total vertices. The clustering algorithm was referred as **Highly connected subgraphs**(HCS) which does not require a pre-specified number of clusters but relies on the similarity graph to generate the clusters from similarity data between vertices compared with a threshold value. A single vertex will not be considered as a cluster but as a singleton set which has the vertex as the only element.

The HCS algorithm requires:

- A graph **G** of n vertices
- the minimum number of edges $\lambda(G)$ that will disconnect the graph when removed (also known as edge connectivity that measures structural cohesion of a graph) [211]
- minimum cut C which is the cut with the minimum set of edges that will disconnect the graph G when removed [86, 211].

Highly connected subgraphs are formed as clusters if and only if $\lambda(G) > \frac{n}{2}$ and every diameter[7] between two vertices must be at most two for the graph to be highly connected, that is their longest distance between themselves must either be one or two.

The HCS algorithm is initially applied to a graph G to check if it is highly connected and if yes, the graph G will be returned, otherwise, C will be applied to separate the graph into two subgraphs g' and g''.

[7] A diameter between two vertices in a social graph is the maximum length of the shortest path from one vertex to the other.

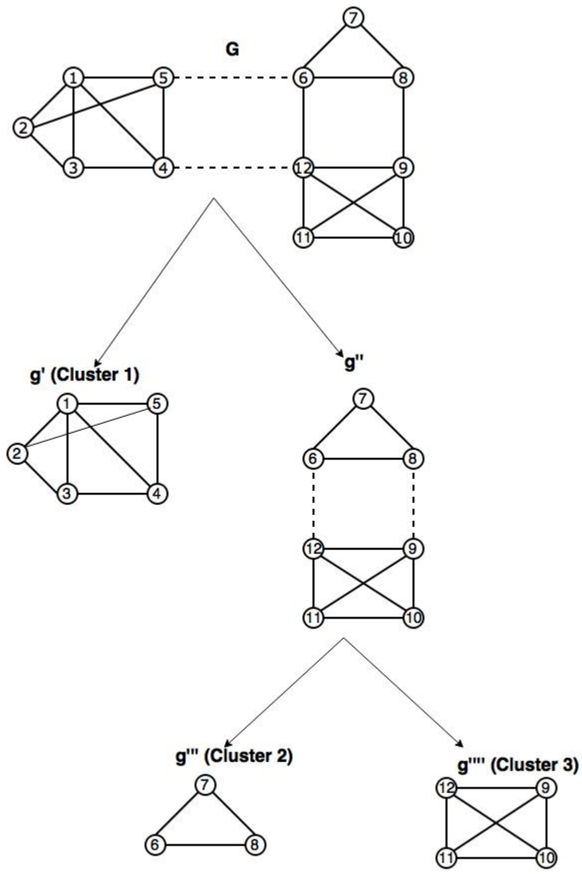


FIGURE 2.8: Partioning a Similarity graph with HCS clustering algorithm

From the example shown in figure 2.8, the HCS algorithm on graph **G** will return subgraph g' as one of the clusters while subgraph g'' will require further splitting since it is not highly connected; subgraph g'' has some vertices with maximum distance of more than two to certain vertices and edge connectivity $\lambda(G) = 2$ which is less than $\frac{n}{2}$ = $\frac{7}{2}$. Subgraphs g''' and g'''' will be generated from subgraph g'' to become cluster 2 and cluster 3 respectively. Note that the broken lines in the above diagram represent the minimum cut C that split either the graph or the subgraphs.

Algorithm 2 HCS Clustering Algorithm

```
1: procedure GIVEN: Graph G
2:     Evaluate minimum cut C of the graph.
3:     if λ(G) > n/2 then
4:         Return G as cluster
5:     else
6:         Split G with C to yield subgraphs
7:         Repeat step 3 on subgraphs.
```

The clusters generated from HCS algorithm has the property of homogeneity and separation. The property of homogeneity exists with the clustering algorithm as two vertices in a highly connected graph (a cluster) must have at least a common vertex neighbour making all vertices similar; in other words, the distance between two vertices must be at most two. The property of separation is based on the splitting of the graph into sub-graphs [8] by the removal of edges at each iteration which will be linear unlike the final clusters that have a quadratic number of edges.

A drawback of this algorithm could be the time complexity in finding minimum cut for each subgraph where the size of both vertices and edges affects the process [42, 198].

2.9.2 Markov Clustering

Markov clustering (MCL) is an unsupervised method for applying clustering on a graph or network to form groups of nodes based on their similarity of interactions in the graph or network. This clustering technique was introduced by Stijn van Dongen in his PhD thesis [56] where mathematical concepts were used to prove the effectiveness of the technique. The main focus was to tackle the scalability issue in clustering where the size of the network has to be considered as a factor in determining the number of clusters to be formed. The whole idea was derived from a random walk representation of a network G where the similarities between two nodes S_{ij} can be interpreted as the probability p_{ij} that a node i will end-up meeting the other node j after a random walk within the network.

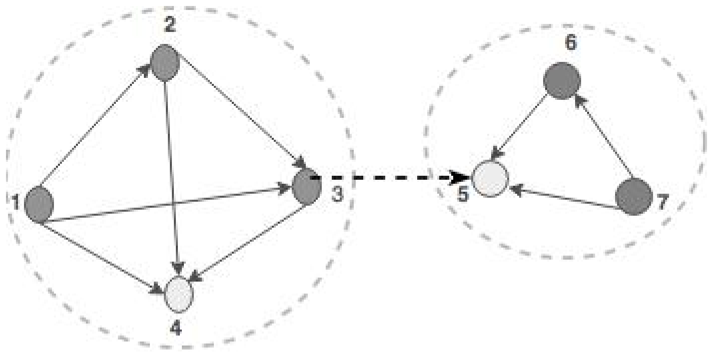


FIGURE 2.9: Random walk within clusters in a network G

There is a high probability for the walk path of a node to be through certain nodes to eventually reach a target node. These connected nodes based on the random walk will rather remain with their cluster than walk across to another cluster with different nodes. From Figure 2.9, the probability of node 1 to travel to either node 2, 3 or 4 will be

[8]Where vertices from different sub-graphs are considered to be dissimilar

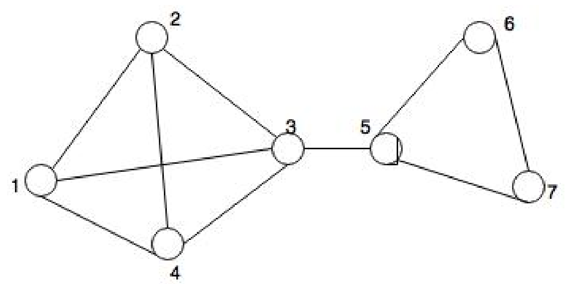
equivalent to 0.33 [9] while the probability of the node 1 to travel across to either node 5, 6 or 7 in another cluster will be 0.00. With the probability values from n nodes, the probability matrix is formed as $n \times n$ matrix.

Algorithm 3 Markov Clustering

1: **procedure** GIVEN: a network of entities **G(V,E)**, expansion parameter **e** and inflation parameter **r**.
2: Create a probability or correlation matrix M_{ixj} where $i, j \in V$.
3: Add self loops to matrix (i.e $p_{ii} = 1$)
4: Normalize the matrix
5: Expand the matrix with parameter **e**.
6: Apply inflation parameter r to matrix.
7: **while** matrix state = unsteady **do**
8: Repeat step 5 and 6
9: Clusters and their attractors are obtained.

From the MCL algorithm, a graph $G(V,E)$ is required to reveal the connection (edges E) between entities (vertices or nodes V) based on their interaction. A weighted graph is expected to be a non-directed graph in order for a symmetric (or correlation) matrix to be generated as input for the algorithm. But this is not always the case as a transition matrix or probability matrix can be generated as a non-symmetric matrix from a non-weighted graph where probability values are considered as the non-negative elements in the matrix.

Using the network structure in Figure 2.9 but representing it as a non-directed graph, the probability matrix can be generated as shown below.



The probability of travel path for each node to other nodes is represented as an element within the matrix. Note that the probability of either node 3 or node 5 to travel across clusters exist with an edge connecting both nodes. The probability of each travel path for node 3 is $\frac{1}{4}$ as it is connected to four different nodes (nodes 1,2, 4 and 5) in the graph.

[9]Probability of a node travelling to another node in the same cluster where there are n members is $\frac{1}{n-1}$

$$\begin{bmatrix} 0.00 & 0.33 & 0.25 & 0.33 & 0.00 & 0.00 & 0.00 \\ 0.33 & 0.00 & 0.25 & 0.33 & 0.33 & 0.00 & 0.00 \\ 0.33 & 0.33 & 0.00 & 0.33 & 0.00 & 0.00 & 0.00 \\ 0.33 & 0.33 & 0.25 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.25 & 0.00 & 0.00 & 0.50 & 0.50 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.33 & 0.00 & 0.50 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.33 & 0.50 & 0.00 \end{bmatrix}$$

FIGURE 2.10: Probability matrix

The probability for each node is an implicit way of measuring the similarity between the node with its connected nodes. Van Dongen [56] described the similarity space as a pair between node or vertex V and a symmetric function s (similarity measure such as Euclidean distance) that maps $V \times V$ to $\mathbb{R}_{\geq 0}$. But as the initial graph is a non-weighted graph, the number of edges are considered as weighted elements linking nodes to a target node in a column of the matrix. A probability matrix can then be obtained from the probability values which are the division of each element in a column by the sum of the elements in the column. This concept for determining the probability matrix was applied and demonstrated in Chapter 3 of this thesis.

Addition of self-loops [10] to each node in the matrix eliminates parity dependence of flow spread on the length of the random walk.

$$\begin{bmatrix} 1.00 & 0.33 & 0.25 & 0.33 & 0.00 & 0.00 & 0.00 \\ 0.33 & 1.00 & 0.25 & 0.33 & 0.33 & 0.00 & 0.00 \\ 0.33 & 0.33 & 1.00 & 0.33 & 0.00 & 0.00 & 0.00 \\ 0.33 & 0.33 & 0.25 & 1.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.25 & 0.00 & 1.00 & 0.50 & 0.50 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.33 & 1.00 & 0.50 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.33 & 0.50 & 1.00 \end{bmatrix}$$

FIGURE 2.11: Applying self loop to matrix

There is also the need to normalize the matrix to ensure that it has the property of probability matrix whereby the sum of each column element will be equal to 1. To normalize the matrix, each element in a column is divided by the sum of elements in that column.

[10] Self-loop is an edge or link that connects a node to itself in a graph. In a matrix, it is indicated with a value assigned at the main diagonal area which corresponds to the connection between vertex with itself.

$$\begin{bmatrix} 0.503 & 0.166 & 0.125 & 0.166 & 0.000 & 0.000 & 0.000 \\ 0.166 & 0.503 & 0.125 & 0.166 & 0.166 & 0.000 & 0.000 \\ 0.166 & 0.166 & 0.500 & 0.166 & 0.000 & 0.000 & 0.000 \\ 0.166 & 0.166 & 0.125 & 0.503 & 0.000 & 0.000 & 0.000 \\ 0.000 & 0.000 & 0.125 & 0.000 & 0.503 & 0.250 & 0.250 \\ 0.000 & 0.000 & 0.000 & 0.000 & 0.166 & 0.500 & 0.250 \\ 0.000 & 0.000 & 0.000 & 0.000 & 0.166 & 0.250 & 0.500 \end{bmatrix}$$

FIGURE 2.12: Normalizing the matrix

With the application of expansion operator [11] where a normal product of the matrix $M \times M$ is taken i.e M^2, flow will be distributed across different regions of the graph as nodes within a cluster are expected to have stronger ties than nodes outside the cluster.

$$\begin{bmatrix} 0.33 & 0.21 & 0.17 & 0.21 & 0.03 & 0.00 & 0.00 \\ 0.21 & 0.33 & 0.19 & 0.21 & 0.17 & 0.04 & 0.04 \\ 0.22 & 0.22 & 0.31 & 0.22 & 0.03 & 0.00 & 0.00 \\ 0.21 & 0.21 & 0.17 & 0.33 & 0.03 & 0.00 & 0.00 \\ 0.02 & 0.02 & 0.13 & 0.02 & 0.34 & 0.31 & 0.31 \\ 0.00 & 0.00 & 0.02 & 0.00 & 0.21 & 0.35 & 0.29 \\ 0.00 & 0.00 & 0.02 & 0.00 & 0.21 & 0.29 & 0.35 \end{bmatrix}$$

FIGURE 2.13: Applying Expansion to matrix

Applying inflation to the expanded matrix will strengthen existing strong ties and weakens the existing weak ties. The inflation operation involves taking the power [12] for individual element between vertices p and q in each column of the matrix M with parameter r (i.e $\mathbb{R}_{>1}$) and then normalizing the each column to retrieve a matrix $\Gamma_r M$. This resulting matrix was defined in [56] as :

$$(\Gamma_r M)_{pq} = \frac{(M_{pq})^r}{\sum_{i=1}^{k}(M_{iq})^r} \tag{2.22}$$

Where: *Γ is the inflation operator and $\sum_{i=1}^{k}(M_{iq})^r$ is the sum of elements (values) in a column of the matrix which vertex q receives.*

[11] Taking the power of e (expansion parameter) on the transition matrix M i.e. M^e

[12] This power operation is known as Hadamard power [96].

$$\begin{bmatrix} 0.11 & 0.04 & 0.03 & 0.04 & 0.00 & 0.00 & 0.00 \\ 0.04 & 0.11 & 0.04 & 0.04 & 0.03 & 0.00 & 0.00 \\ 0.05 & 0.05 & 0.10 & 0.05 & 0.00 & 0.00 & 0.00 \\ 0.04 & 0.04 & 0.03 & 0.11 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.02 & 0.00 & 0.12 & 0.10 & 0.10 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.04 & 0.12 & 0.08 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.04 & 0.08 & 0.12 \end{bmatrix}$$

$$\begin{bmatrix} 0.458 & 0.167 & 0.136 & 0.167 & 0.000 & 0.000 & 0.000 \\ 0.167 & 0.458 & 0.182 & 0.167 & 0.130 & 0.000 & 0.000 \\ 0.208 & 0.208 & 0.455 & 0.208 & 0.000 & 0.000 & 0.000 \\ 0.167 & 0.167 & 0.136 & 0.458 & 0.000 & 0.000 & 0.000 \\ 0.000 & 0.000 & 0.091 & 0.000 & 0.522 & 0.333 & 0.333 \\ 0.000 & 0.000 & 0.000 & 0.000 & 0.174 & 0.400 & 0.267 \\ 0.000 & 0.000 & 0.000 & 0.000 & 0.174 & 0.267 & 0.400 \end{bmatrix}$$

FIGURE 2.14: Applying Inflation to expanded matrix - **(Above)**Taking power r for elements in each column $(M_{pq})^r$ and **(Below)**After Normalizing the matrix.

The expansion and inflation operation will be alternated continuously until the resulting matrix will not change when these operations have been applied. The steady state of the matrix (known as Equilibrium state matrix - ESM) will take the form of a doubly idempotent form [13] with the resulting matrix having homogeneous values in all column from a single row.

$$\begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

FIGURE 2.15: Steady state matrix after Markov clustering (ESM)

[13]Doubly idempotent state is where two processes will always have their same outcome when there are further iterations with the processes.

After 14 iterations, two clusters are formed with node 1,2, 3, 4 in one cluster and nodes 5, 6, 7 in the other cluster. The attractors from both clusters can be referred to as the influential nodes. They are usually identified as the node from the row of the matrix that has all homogeneous values (probability values) to its members in the cluster. The first cluster has node 3 as an attractor that attracts nodes 1, 2 and 4 while node 5 attracts nodes 6 and 7 in the second cluster.

Markov clustering has been considered in previous researches [5, 56] as a clustering technique that does not allow overlapping in clusters. But there are some cases (Such as isomorphic clusters [14]) where a node could be attracted by more than one attractors into their individual cluster. This might mean that the node has conflicting preferences as the node exists in different clusters. This will cause difficulty for the node to be influenced towards specific items by any of the attractors.

2.9.3 Affinity Propagation(AP) Clustering

Affinity propagation clustering algorithm (AP) introduced by Frey & Dueck [67] as a clustering technique which involves a concept of passing real-valued messages between data points of a dataset until clusters and their exemplars are discovered. All data points are initially considered as potential exemplars for clusters to be found. Unlike the k-means clustering technique [104] that requires the number of clusters specified, the affinity propagation technique requires the similarity between pairs of all data points as an input for the algorithm.

Another reason for not considering k-means technique as a suitable technique for the research is that the technique only finds it's centroid by computing the mean point of a cluster and this centroid might not necessarily be considered to be an ideal representative of a cluster as they are refined and derived from pre-specified k points (i.e. clusters) which were initially chosen randomly from the dataset.

Several research areas (such as Bioinformatics, image processing) have considered the AP clustering to be more effective in finding clusters and their exemplars from given data. In bioinformatics researches [141], AP clustering was used to understand and manipulate biological processes (such as protein-protein interactions and metabolism) when their network is prone to produce errors (e.g. false positive and false negative

[14]Isomorphism is the case when the vertices of a cluster or graph can be mapped to vertices of another cluster or graph.

test results) also referred to as 'noise' [141]. This clustering technique has been used in [141] to assist on untangling of the 'noise' network from the 'genuine' network with the accurate information. Image processing researches [68] made used of the AP clustering to reveal how an exemplar of an image is related to the original-input image. Also, the information/data mining research area [192] have used the clustering in finding the relevant information of items based on their similarity to existing knowledge.

The AP clustering algorithm considers a non-empty set of data points x_1 through x_n (Where $n \geqslant 2$) and a similarity function s using negative squared Euclidean distance (equation 2.23) that computes the similarity between pairs of data points $s(i, j)$; where $s(i, j)$ might not necessarily be equivalent to $s(j,i)$. The reason for this type of similarity measure to be used was due to the type of dataset where the data points are all real-valued data. Clustering with this type of similarity measure is said to be more efficient when compared with the normal Euclidean distance. Frey & Dueck [67] suggested this measure to minimize the squared error and optimization problem in the distances when computing the similarities. In chapter 4 of this thesis, further analysis will be carried out with several other similarity measures in supporting the AP algorithm.

Computed similarities between points $s(i, j)$ within the AP algorithm are represented in an $n \times n$ matrix where it indicates how suitable data point j could act as an exemplar to data point i. The closeness in the similarity values amongst data points determines how if they become members of the same cluster.

$$s(i, j) = -||x_i - x_j||^2, i \neq j \tag{2.23}$$

To determine and control the number of clusters, another feature known as the input preference $s(i,i)$ is required and it represents the suitability of data point i to act as an exemplar. A lower input preference value that is closer to the minimum similarity value will generate few clusters while a high input preference value that is closer to the maximum similarity value will generate more clusters. Usually, the best value to be used for AP algorithm is either the median value or the minimum value from the similarity matrix.

The real-valued messages to be exchanged between pairs of data points are referred to as the 'availability' $a(i,k)$ and 'responsibility' $r(i,k)$ [67] where both messages consider the competition between potential exemplars k as a representative for target data point i. $a(i,k)$ represents the message sent by potential exemplar k to target data point i which reveals how appropriate for target i to select k as its exemplar when considering other points $i^{'}$ that have chosen k as an exemplar.

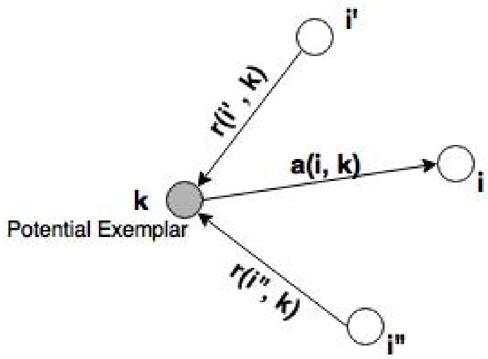


FIGURE 2.16: Availability sent by potential exemplar k to target point i.

$r(i,k)$ represents the message sent by data point i to a potential exemplar k which reveals how ideal it would be for k to serve as its exemplar even though other potential exemplars (k' and k" as shown in 2.17) have indicated their availability to compete with k.

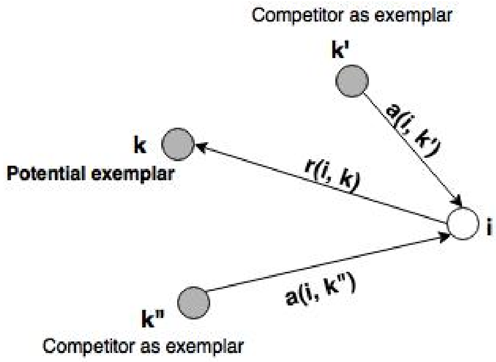


FIGURE 2.17: Responsibility sent by data point i to potential exemplar k.

Frey & Dueck described how normally constraints could exist where each data point might have constrained choices in selecting an exemplar based on their similarity but having all data points to interact with each other could resolve the problem. From the diagram shown above, a data point will send message to a potential exemplar who has been chosen by other data points as their representative, revealing the list of other potential exemplars that have signified interest in representing the data point. A potential exemplar will send messages to every point informing them of the degree of agreement between their constrained choices and the list of other points that have chosen the exemplar as a representative.

In the algorithm, the exchange of the messages ('Responsibility' and 'Availability') are iterated until it determines the best exemplars from the data set before other data points cluster around their suitable exemplar based on their similarities. Each iteration of the algorithm involve:

- All 'responsibilities' being updated (with given 'availabilities') and distributed to potential exemplars. Updates are done using:

$$r(i,k) \leftarrow s(i,k) - \max_{k' \neq k} \{a(i,k') + s(i,k')\} \tag{2.24}$$

- All 'availabilities' being updated (with given 'responsibilities') and distributed to all data points. Updates are carried out with:

$$a(i,k) \leftarrow \min\left\{0, r(k,k) + \sum_{i' \notin \{i,k\}} \max\{0, r(i',k)\}\right\} \tag{2.25}$$

  and

$$a(k,k) \leftarrow \sum_{i' \neq k} \max(0, r(i',k)) \tag{2.26}$$

  Where: a(k,k) is known as 'self availability' which reveals that *k* had received 'responsibilities' from other data points.

- Combination (i.e. sum) of both 'responsibilities' and 'availabilities' to observe the decisions on exemplar $a(i,k) + r(i,k)$

The iteration will be terminated when the decisions remain the same for certain number of iteration. The exemplars k_i are identified as those whose combination (sum) of their 'responsibility' and 'availability' is positive (i.e. Where $k = i, (r(i,i) + a(i,i)) > 0$) or when k maximizes the combination.

$k_i = \text{argmax}[a(i,k) + r(i,k)]$

Algorithm 4 Affinity Propagation

procedure GIVEN: a Similarity data (matrix) $s(i,j)_{i,j \in 1,\ldots n}$ and input preference $s(i,i)$.

  Initialize 'Availabilities' to zero i.e. $a(i,k) = 0$.

  while Potential exemplar k = unsteady **do**

    update 'Responbility' and 'Availability' .

    Combine messages, 'responsibilities' $r(i,k)$ and 'availabilities' $a(i,k)$ of potential exemplar k.

    if k maximizes combination (Addition) of messages **then** k is an exemplar for point i i.e. $k = k_i$ or point i is an exemplar when $k = i$.

  Clusters c_i and their exemplars k_i are obtained.

Previous research [202] revealed how to evaluate and demonstrate the algorithm easily by considering the 'responsibility' values and 'availability' values in separate matrices that could be updated repeatedly to obtain the exemplars. As the algorithm simulates the interaction between 'actors', a consensus is expected to be reached when they are selecting their best representatives.

Example 2.3. *Given a data on actors that have acted several number of times with 4 items*

TABLE 2.3: Users' action towards Four Items

$Actors$	$item_1$	$item_2$	$item_3$	$item_4$
$user_1$	4	2	1	5
$user_2$	3	5	2	5
$user_3$	1	3	3	2
$user_4$	2	1	3	4
$user_5$	4	3	5	1

Using equation 2.23, similarity matrix to compare pair of actors :

$$\begin{array}{c} \\ user_1 \\ user_2 \\ user_3 \\ user_4 \\ user_5 \end{array} \begin{array}{c} \begin{array}{ccccc} user_1 & user_2 & user_3 & user_4 & user_5 \end{array} \\ \begin{pmatrix} 0 & -11 & -23 & -10 & -33 \\ -11 & 0 & -18 & -19 & -30 \\ -23 & -18 & 0 & -9 & -14 \\ -10 & -19 & -9 & 0 & -21 \\ -33 & -30 & -14 & -21 & 0 \end{pmatrix} \end{array}$$

The negative sum of the square of differences between the pairs of actions towards four different items shown in 2.3. For instance, the similarity between $user_1$ and $user_3$ (shown above) was evaluated as:

$$s_{(user_1,user_3)} = -[(4-1)^2 + (2-3)^2 + (1-3)^2 + (5-2)^2] = -23$$

With the input preference set to the lowest similarity value(-33), this value is inserted as the diagonal element of the matrix.

$$\begin{array}{c c} & \begin{array}{ccccc} user_1 & user_2 & user_3 & user_4 & user_5 \end{array} \\ \begin{array}{c} user_1 \\ user_2 \\ user_3 \\ user_4 \\ user_5 \end{array} & \begin{pmatrix} -33 & -11 & -23 & -10 & -33 \\ -11 & -33 & -18 & -19 & -30 \\ -23 & -18 & -33 & -9 & -14 \\ -10 & -19 & -9 & -33 & -21 \\ -33 & -30 & -14 & -21 & -33 \end{pmatrix} \end{array}$$

Using equation 2.24 and initializing the $a(i,k^{'})$ to zero, the responsibility matrix will be:

$$\begin{array}{c c} & \begin{array}{ccccc} user_1 & user_2 & user_3 & user_4 & user_5 \end{array} \\ \begin{array}{c} user_1 \\ user_2 \\ user_3 \\ user_4 \\ user_5 \end{array} & \begin{pmatrix} -23 & -1 & -13 & 1 & -23 \\ 7 & -22 & -7 & -8 & -19 \\ -14 & -9 & -24 & 5 & -5 \\ -1 & -10 & 1 & -24 & -12 \\ -19 & -16 & 7 & -7 & -19 \end{pmatrix} \end{array}$$

Equation 2.26 and 2.25 are then used to provide the diagonal and non-diagonal elements of the availability matrix.

$$\begin{array}{c c} & \begin{array}{ccccc} user_1 & user_2 & user_3 & user_4 & user_5 \end{array} \\ \begin{array}{c} user_1 \\ user_2 \\ user_3 \\ user_4 \\ user_5 \end{array} & \begin{pmatrix} 7 & -22 & -16 & -19 & -19 \\ -23 & 0 & -16 & -18 & -19 \\ -16 & -22 & 8 & -23 & -19 \\ -16 & -22 & -17 & 5 & -19 \\ -16 & -22 & -23 & -19 & 0 \end{pmatrix} \end{array}$$

To check if any of the potential exemplars maximizes the function $a(i,k)+r(i,k)$, the elements of the availability matrix is added to the elements of the responsibility matrix.

$$\begin{array}{c} \\ user_1 \\ user_2 \\ user_3 \\ user_4 \\ user_5 \end{array} \begin{array}{c} \begin{array}{ccccc} user_1 & user_2 & user_3 & user_4 & user_5 \end{array} \\ \begin{pmatrix} -16 & -23 & -29 & -18 & -42 \\ -16 & -22 & -23 & -26 & -38 \\ -30 & -31 & -16 & -18 & -24 \\ -17 & -32 & -16 & -19 & -31 \\ -35 & -38 & -16 & -26 & -19 \end{pmatrix} \end{array}$$

From this matrix, it is observed that two clusters are formed from two identified exemplars, $user_1$ and $user_3$ that have certain maximum value being the same in their column indicate the members of the individual cluster. The first cluster with $user_1$ as exemplar has a member $user_2$ while the second cluster with $user3$ as exemplar has members $user_4$ and $user_5$. These selected exemplars will be confirmed based on the maximum values in each column being greater than zero. Several iterations will be done to further confirm that the outputs reveal the exemplars are unchanged.

The evidence computed from the updates of both ‘Responsibility’ and ‘Availability’ are given to potential exemplars and members of a cluster respectively. This evidence is used to justify the selection of each exemplar and the decision by each member to join the cluster. Previous researchers have considered the AP clustering to be more effective because of the computation of the evidence.

The next section will discuss the relationship between clusters and the trustworthiness amongst their members as it could be possible to measure the strength of a cluster based on the degree of trust amongst members.

2.10 Cluster Analysis and Trust Network

Previous researchers [7] had pointed out that clustering techniques are considered to be an effective means to improve recommender system. Clusters can be considered as structures which might be equivalent to trusted networks as they promote interaction amongst respective members. The distance between cluster members derived basically from the similarities of their attributes will reveal within the structure how items such as information can be shared (flow) and accepted amongst the members.

Most people might consider privacy to be a factor to affect the sharing of information within a cluster. Previous research [150, 197] had revealed that people tend to share their information with others and ignore the importance of privacy. The reasons for their action may be that they may want to either receive privileges or build their reputation. Also, they would not want to be asked questions directly on certain information and so they decide to reveal the information to the public [150]. In this case, these people don't consider their trust towards others before their decision to share with a group. It is possible that the people's expected privileges from the group might not be in their favour or satisfy them as other members of the group might mislead them for their own benefit.

According to Segarra & Ribeiro [188], there is the need to apply clustering to model the trust network between entities to understand and know how an entity will trust another. The trust between entities should be determined based on the similarities in the entities' opinion as an entity will trust another entity's opinion if they are similar to their own opinion. It was revealed in [188] that the tie weight between entities in a bidirectional network can be the dissimilarities between the entities $A_X(x,x')$ which may satisfy the asymmetric property i.e. $A_X(x,x') \neq A_X(x,x') \; \exists \; x,x' \in X$. The clusters are determined based on the resolution value $\delta \geq 0$, where the boundary when $\delta = 0$ generates a singleton cluster (a cluster with a single node as a member) for each node and the boundary when the $\delta > 0$ generates clusters with different nodes. It was also revealed that for two nodes to share opinions their dissimilarity value (i.e. It indicates their distrust for each other) must be less than the resolution value δ. The resolution value δ in [188] was considered to be the tree cut value in a dendrogram of the hierarchical clustering on a trust network which also represents the degree of familiarity with items (e.g. Topic or issue) that links two nodes to interact.

2.11 Social Influence and Trust

From a real-life scenario, a person will decide to have the same behaviour or have the same opinion towards an item like his/her neighbours only if they are trustworthy. There could be some untrustworthy neighbours who might want to mislead the target person for certain reasons such as promoting or criticise an item [178] for their own benefit. So with the existence of trustworthy neighbours to a target node, the degree of influence will increase between the target node and a neighbour only if their path length or distance is small [52].

Previous research [117, 207] focused on the problem of identifying influential users in a social network based on how well the users can represent members of their trust network, that is, the pairwise relationship between nodes in the network indicated as an influence measure of one node towards another. Vedula et.al. [207] revealed that a user i will trust another user j when influenced by j to engage in a certain context. In other words, user j who is considered trustworthy is influential towards user i. Luini et.al[117] revealed that predicted trustworthiness of a member in a trust network is one important factor required in identifying an influential member. Apart from identifying the influential member of a network, it is also important to understand the impact of influence on both network structure and decision making.

2.11.1 Influence effect on social communities

As we have already seen from the previous section that trust has an effect on influence, it is required to determine if influence can change a social group with trustworthy members. Luini et.al.[117] had previously investigated how the social influence of a member in a network could affect the trusting behaviour of others due to the member's dynamic preference.

An individual will identify himself or herself with a group based on his/her information similar to a common group members' information. This can be referred to as social identity. Understanding the social history or experience of members will provide a clearer view of the members' identity to their groups. The social history which is represented in an initial network structure is observed to predict the type of behavioural change the members could undergo. Previous research by Kelman [107] had classified various behavioural changes into different types of conformity [15].

- Compliance conformity, where an individual will conform to a group's opinion or preference in order to receive reward or benefits. An example can be observed in Asch's line experiment[127] which was described in an earlier section of this Chapter.

[15] Mcleod [130] referred to conformity as a social influence that changes the behaviour or belief of an individual to match the belief or behaviour of a group

- Internalization conformity, where an individual will conform to a group's preference due to observed consistent activities or opinion by the group members. The individual will have the belief that the group members are more knowledgeable than other groups which might have exhibited inconsistent behaviour in the past. An example was also mentioned in a previous section of this chapter as Sherif's Autokinetic experiment [130].

- Identification conformity, where an individual will conform to a group's opinion in order to be recognized with a group. An example is the Zimbardo's prison study [129] which was also described in a previous section of this Chapter.

The changes in the behaviour of an individual in both Compliance and Identification are not permanent as the individual (which can be referred as an 'unbeliever') decides to conform with no genuine belief in the group's opinion or preference. To maintain control of consistent behaviour in the group, an influential member will be required to encourage the inactive who might have been misled by the individuals with a conflicting opinion. It is also possible for the influential member to convince the 'unbeliever' with a clearer understanding on the reason to completely believe in the group's opinion.

Centrality is one important concept in a social network that could be considered for the identification of either the inactive members who might require advice to make decisions on alternatives (e.g. topics to discuss and items to purchase) or the influential members of the network who might be able to provide the advice to the inactive members. With or without the knowledge on the level of reciprocity, the inactive members will imitate the actions of influential members only in situations where the actions affect the inactive members' belief [117].

2.11.2 Decision Making from Influence

Decision making is one popular topic under the social influence research that analyzes how and why people make a particular decision. Decision-making problem occurs when some individuals who might belong to their individual group are deviating from their group's opinion or they are unable to make decisions towards alternatives (e.g. products, topics or any other item). The previous sections had presented how individuals decide to join a group for various reasons based on the conformity theory.

A good example that demonstrates a group decision-making pattern is a social network such as Facebook and Twitter. The interaction between users within the network reveals their preferences towards items (e.g. topics or products). Users might decide to continue interacting with neighbours on new items that could be similar to items preferred from their past interactions. So it is possible that past experience will be an important factor for modelling influence. Another factor which plays an important role is the similarities between users' preferences.

Previous researchers [102, 166, 215] had considered applying social network analysis (SNA) concept for the development of a decision-making model based on interaction patterns and network structures. Kamis et.al. [102] revealed that a group decision-making problem can be resolved by using a consensus procedure where the level of group agreement is measured. The level of agreement is measured based on a feedback mechanism where the identification of an adviser (Influential member of a group) and the proximity measure [16] amongst group members' preferences are both required. The easiest way to accomplish this is by relying on neighbourhood information(clusters) obtained from clustering [102, 112, 167]. But identifying the influential member from each neighbourhood remains a difficult task. Various clustering techniques have been discussed in previous sections of the thesis but the most suitable clustering for identifying the influential member would be revealed later in Chapter 4.

The influential member is meant to advise/guide any other member (inactive member) that might be uncertain on decisions to be taken. For the inactive members to make an accurate decision, they will require a consolidated information on preferences from their individual group. Previous researchers [43, 44, 102] relied on using either ordered weighted averaging(OWA)[217] or induced ordered weighted averaging (IOWA)[9, 218] to carry-out the consolidation operation where information from various sources in a group are aggregated. The application of OWA or IOWA is outside the scope of the research as a new method for consolidating the preference information will be introduced in Chapter 3 of this thesis.

[16] A suitable proximity measure is the similarity measure which determines the closeness between two entities.

2.12 Summary

This chapter has presented the literature review which was divided into three parts. The review includes background on common recommendation approaches, clustering techniques on a social network, the effect of trust concept on a social network which reveals the relationship between trust and influence concept towards a change in the behaviour of a specific entity within the social network. As the social network platform is an example that utilizes the recommender system, there is the possibility of implementing the trust and influence concept.

The active research in recommender system actually revealed that there is still the need to resolve the cold-start problem in the recommender system where there are insufficient or no information from active users. The most common recommendation algorithm which are the content-based approaches and collaborative approaches were discussed with given examples to reveal their operations within the recommender system. The cold-start problem that could exist with the two approaches might due to either the irrelevance of items presented to the users, the fear of being judged by others or the fear of jeopardizing privacy.

The concepts that could be useful in the proposed approach in resolving the cold-start problem, include the trust concept which several researchers [16, 71, 116, 123, 144] have previously referred to be an element of social network. Luhmann [116] revealed that the only way to handle a complex neighbourhood will be through the understanding the trust between neighbours. From review comparing various similarity measures, it was revealed that the most suitable measure that can be used for the proposed framework is the cosine similarity measure as it relates reachability between two nodes and their similarities which will also imply their trust for each other.

It is also believed that trust can be drawn from the similarity in behaviour of neighbours within a group where each neighbour will believe that other neighbours are knowledgeable in making right decisions. This idea then relates trust with influence based on conformity theory where a neighbour will conform or behave like other members of their group since they are believed to be knowledgeable due to their consistent behaviour. Based on the review of various clustering algorithms, it was clear that there is also the need to understand how members identify themselves with groups(clusters) based on their activeness in a network. The review was also required to discover the possibility of accurately identifying influential members in the groups (clusters).

Chapter 3

Trust in Social Communities

3.1 Introduction

As earlier discussed in the previous chapter, the problem which recommender system still faces is the inaccuracy in predicting user's preference due to lack of activeness and the lack of conscious reasoning[1] applied to the learning process in the recommender system. Most preference prediction are based on explicit feedbacks (for example, ratings or declaration) but it will be very difficult to predict any preference when this type of feedbacks are not available or not clearly expressed.

To overcome this drawback, there is the need to rely on implicit feedbacks (for example, frequency of actions that occurred within a social environment) from behavioural pattern in predicting preferences where **trust** is expected to provide a better understanding of behaviours. A thorough understanding of trust initially reviewed will bring to light its relevance in decision-making within a system and also draw out important features that can enable an accurate formulation of trust. The correlation between trust and similarity which were earlier revealed from reviews in the previous chapter helped in justifying the formulation of the proposed trust metric.

[1]Conscious reasoning is the ability to understand something (e.g. behaviour) with rational justification.

3.2 Trust within Intelligent systems

As we are living in the era of intelligent system, information overload seems to continually affect accurate decision making. Individuals require relevant information in several situations: when they find themselves in unexpected situations, when they are naive about certain things or when they are new to an environment. In order to manage the diversity and dynamic attribute in online data generated either by explicit reviews on items or implicit behaviours towards items, the intelligent systems is required for retrieving relevant information that a particular user will use in making an accurate decision. Intelligent system is believed to have a significant effect on satisfaction of users or customers [178]. Users will only be satisfied when the intelligent system provides the relevant services or items to them. Another factor that could increase satisfaction is the usability or accessibility degree of the system interface as users will always want to enjoy using the system.

An intelligent system such as a recommender system can also be referred to as a recommender agent [210] for companies' online system. The recommender system is expected to understand and adapt to the needs of users. There is the need to enhance this feature of the recommender system by considering an influential agent that will understand the behaviour of both the target users and their similar neighbours in the system's community before convincing them to adjust their behaviour to match themselves.

As an intelligent system requires user's information for its learning process, there is the need to consider trust between users in situations when there are no or insufficient information available for the learning process before decision making. The intelligence in service-provider system can be improved when advisory features such as the recommendation are deduced based on trust relations between users of the system.

3.3 A New Approach for Computing Trust

Trust can be considered as an implicit rating in which various individuals could express towards others differently based on implicit behaviour. For example, in a social network, an individual may decide to view or repost most of the numerous items initially posted by another individual. In other words, we could consider trust as a condition from an entity's activeness that can be used to predict the trustworthy behaviour of the entity. Philosophers have described trust to be the attitude of people towards a particular person whom they consider to be trustworthy.

As earlier discovered from the literature reviews in Chapter 2, there is the need to consider social factors that affect the computation of trust. A person who is committed to being active or doing his/her job in a social environment will be trusted by other members of the environment. As the strength of ties between entities is determined by the frequency of their interaction, it is possible to compute trust between the entities based on their active condition.

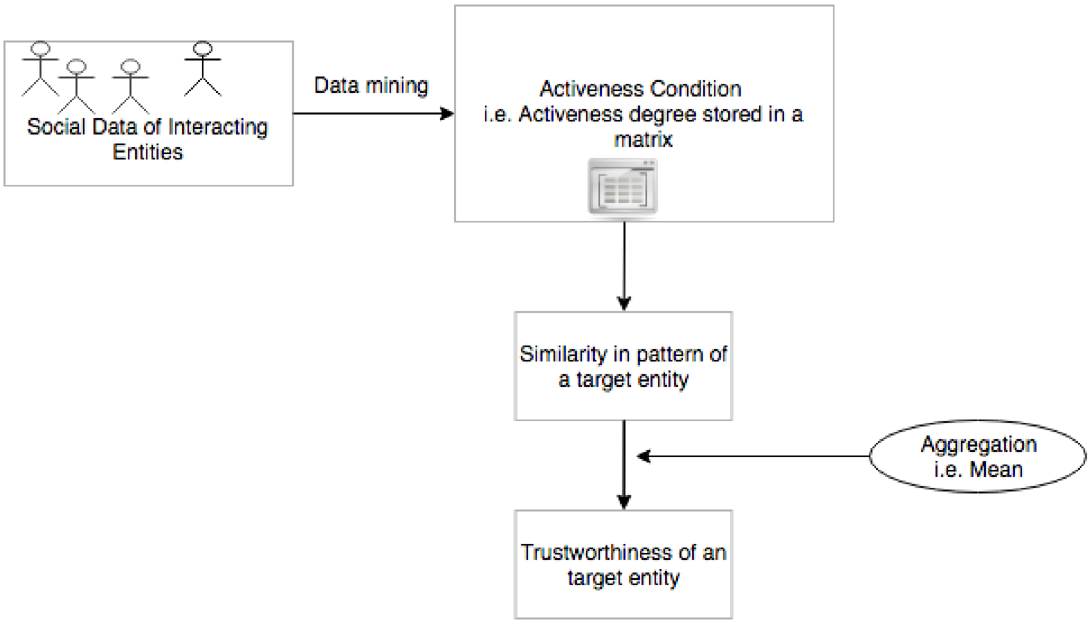


FIGURE 3.1: A Framework of the proposed trust computation.

The above framework of the proposed trust computation (See figure 3.1) is based on the theory of interpersonal behaviour [13, 203] where the expected trusting behaviour of a target entity will depend on the target entity's active situation (i.e. activeness degree in relation with others) and its pattern or habit with others (where similarity in predicted interaction could be estimated).

In the proposed trust computation, there are several steps (Described in subsections below) to be taken in order to determine the trustworthiness of a node. This can be seen in the pseudocode below (See Algorithm 5).

Algorithm 5 Familiarity based Trust Algorithm

1: **procedure** GIVEN: a two-mode network of n entities.
2: Analyse the network to retrieve activity information of entities (i.e. using either Sum method or Newman method).
3: **for** Each entity i **do**
4: Evaluate trust between a entity i and each entity j using a probabilistic measure to form elements in activeness matrix $M_{n \times n}$
5: From the activeness matrix, evaluate similarity between the entity i and each entity j as their familiarity degree.
6: Take the average of all similarities as Global familiarity degree Fam_i for entity i.
7: **return** Fam_i as entity i's measure of trustworthiness.

The algorithm 5 generates the trustworthiness for all entities in a given two-mode network. We expect to see if the activeness of a node will determine how trustworthy the node will be for future engagements with other nodes. The following subsections will further describe how each step in the algorithm was carried to obtain the outputs.

3.3.1 Analysing the Social Activities

Most researchers as earlier discussed in the Literature review might view a social data as a one-mode network without knowing or considering how it was derived. However, a clearer understanding of the derivation of the one-mode network could assist in the prediction of trust between entities. It will be best to view the two-mode network where a tie between different class of nodes (i.e a user and an item) can reveal how the tie between the same class of nodes (i.e two or more users) are formed.

Definition 3.1. A two-mode network is a graph $G^t = (V_u, V_s, E^t)$ which is made up of a set V_u of vertices for users, a set V_s of vertices for items and a set of edges E^t that might form association between $v_x \in V_u$ with $v_y \in V_s$. Two or more nodes from the set V_u might have ties with common nodes in set V_s.

An example of a dataset representing the information for a two-mode network can be seen in table 3.1 where four different nodes of the user class $V1$ interact with nodes of the item class $V2$.

TABLE 3.1: Dataset for a two-mode Network with social activities from Four users

V1	V2	V3
5	72	1
5	82	1
5	448	1
6	64	2
6	82	1
6	115	2
6	229	4
6	236	10
6	355	1
6	391	3
6	482	1
6	524	2
8	136	1
8	341	1
100	13	40
100	82	166
100	115	3
100	391	16
100	399	19

Thus, a two-mode network of the given dataset above (Table 3.1) can be represented in figure 3.2. The strength on a tie $V3$ represented in the dataset are the frequency of interaction by a user with an item. From the dataset, it can be observed that user node 100 is more active than other user nodes as it has more frequent interaction with items than others.

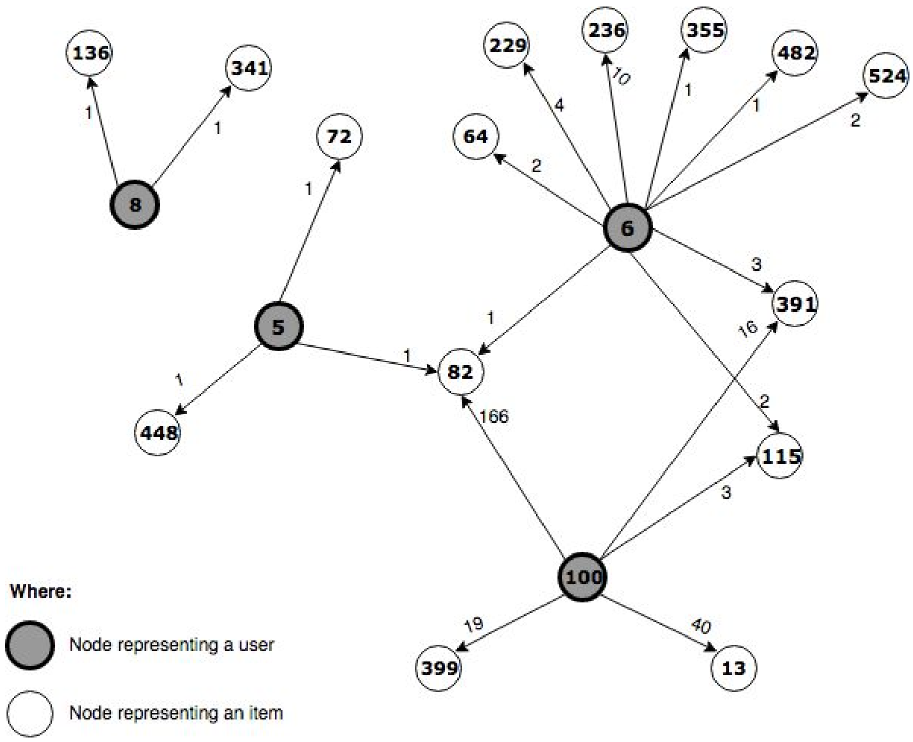


FIGURE 3.2: A graph based on the two-mode network with four users.

The dataset described in section 1.7.2 of Chapter 1 to be used during the empirical test is similar to the dataset given in table 3.1. This was derived randomly from the original dataset previous used by Opsahl [151] which consist of three variables $V1$, $V2$ and $V3$ representing user(actor) nodes, item nodes (i.e. topics or subjects) and strength of tie (i.e. Number of comments to an item or out-degree of user node to the item) respectively. As this network represent interaction between nodes of different class (i.e user and topic represented as $V1$ and $V2$ shown in table 3.1), there was the need to carry-out a measure of reciprocity on the network to comprehend the mutual exchange of messages between the nodes of the same class (i.e. user or actor class) based on their interaction on another class of node (i.e. topic, item or subject class). This can only be done after the two-mode network has been converted to a one-mode network.

Definition 3.2. A one-mode network is a graph $G = (V, E)$ which is made up of a set of vertices $v \in V_u$ and a set of edges E that form an association between $v_n \in V$ and $v_m \in V$ based on their common edges from a two-mode network G^t.

TABLE 3.2: One-mode Network Dataset: A projected Two-mode Network using Sum method

i	j	w
5	6	1
5	100	1
6	5	1
6	100	6
100	5	166
100	6	185

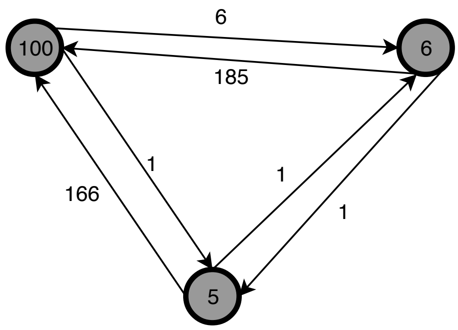


FIGURE 3.3: One-mode network graph of the projected two-mode network using Sum method

From the above definition, it may imply that two or more nodes from set V having ties with common nodes from set V_s of a two-mode network might trust themselves to a certain degree. A two-mode dataset (table 3.1) is more preferred to a one-mode dataset (See an example in table 3.2) since it has more information to justify how the ties between certain nodes were derived in a one-mode network. This one-mode dataset only represents a network between users (See figure 3.3) exchanging messages amongst themselves without any information about context or items (e.g. topics discussed or subjects studied) where i represent a user node that interacts with another user node j who assigns a weight w to i as the total number of out-degree weights from i towards several items in set V_s which j also interacted with.

Sample of a one-mode dataset was used in previous research work [155] to actual reveal social factors such as Familiarity based on a central value of similarities from a probability distribution and 'experience' based on engagement outcomes [145] evaluated from probability expectation of future outcome [101]. From further Literature review, it was concluded that 'experience' correlates with 'familiarity' as they are both equivalent to predicting trust between pair of nodes [145]. This prediction can only be done via datasets with activity information as observed in a two-mode network.

As earlier reviewed, previous researchers found it necessary to convert a two-mode network to a one-mode network if social network measures must be applied to understand the two-mode network. The methods of conversion proposed by Opsahl [151] was used in this research. One method, the sum method discussed in chapter 2 (section 2.7) was compared with the other method, Newman method, to determine which has a better result but it was discovered that there will be no significant difference when using any of the method. The weights obtained from the Newman method will be slightly different from that of sum method based on the fact that more nodes who acted on a common item will have lesser bond or strength in interaction than when fewer nodes acted on the common item [147, 151]. An example of a converted two-mode network (considering the two-mode network in figure 3.1) using both methods can be viewed in table3.2 and table3.3 respectively where four users assign weights to each other based on their frequent interaction (frequency of comments) to common items.

TABLE 3.3: Dataset for a One-mode Network: A projected Two-mode Network using Newman method

i	j	w
5	6	0.5
5	100	0.5
6	5	0.5
6	100	5.5
100	5	83
100	6	102

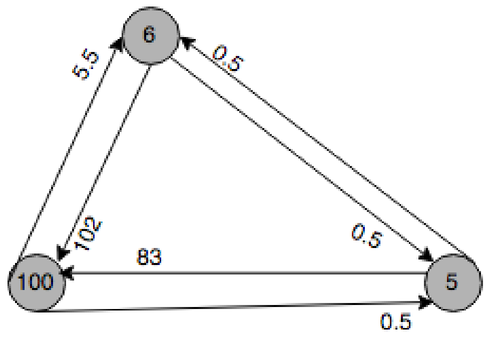


FIGURE 3.4: One-mode network graph of the projected two-mode network using Newman method

One main factor to be measured or derived from the social network is the familiarity between nodes as it is based on the real world scenario where two persons might trust each other only if they have similar patterns from their past interactions. This will be described later in subsection 3.3.3 where the prediction of a node's trustworthiness could reveal if the node is the most active node amongst other nodes.

3.3.2 Predicting Activeness

As earlier discussed in Chapter 2, most researchers [145, 151] consider the frequency of interaction between a pair of nodes to be the weight of ties between the nodes. Weights that indicate strong ties have been pointed out in previous researches to be a means in predicting the existence of ties between a pair of nodes. According to Barrat et.al [17], nodes that have a collaboration on items as observed in scientists collaboration network will have strong ties. It was also pointed out in [220] that individual tie weights cannot completely reveal the complexity of a network which consists of other nodes and alternative paths that could be used to support the estimation of nodes' strength.

A node's strength which was described to be the sum of weights from all ties shared (in the case of a collaborative network) [17] correlates to the node's level of activeness which depends on the node's frequency of interaction.

As probability can be used to predict if an event will happen or not, we can compute the possibility of a node to remain active based on past activities in a social network. This measure can be referred to as the ***activeness trust value*** since it predicts a target node's social activeness with other nodes on items(e.g. Subjects or topics). This can be determined by considering each past interaction of a target node i with every other node j of the same user set V_u based on their actions towards certain items (node of another set V_s).

Definition 3.3. Based on probability expectation of future outcome [101], the probability of a target node's (i.e node i) activeness along with another node j towards all subjects (or items) c they both share is:

$$P^c_{i,j \in V_u} = \begin{cases} \dfrac{N^c_{ij}}{N^c_{ij} + N^c_{ji}}, & N_{ij} \neq 0 \\ 0, & N_{ij} = 0 \end{cases} \tag{3.1}$$

$i \neq j$
$P_{i,j} \in [0,1] \quad \forall i, j$
$P_{i,j} + P_{j,i} = 1$

Where: *N^c_{ij} is the sum of node i's actions towards all subject c which node j has also acted upon. N^c_{ji} is the sum of node j's actions towards all subject c which node i has also acted upon.*

Based on Nepal's [145] proposed *predicted engagement trust*, probability expectation value [101] was used to model the likelihood of a node i engaging with another node j. Probability expectation value which was defined with two parameters α and β that relate with the possible two outcomes that could occur from future engagements. With the probability of activeness using equation 3.1, there can only be an outcome from any of the following possible outcomes:

- Nodes will act more on shared items c. i.e. $P^c_{i,j} \geq 0.5$
- Nodes will act less on shared items c. i.e. $P_{i,j} < 0.5$

- None of the nodes will act on any item c . i.e. $P^c_{i,j} = P^c_{j,i} = 0$

This predicted activeness $P^c_{i,j}$ which can be referred to as ***activeness trust*** is based on the frequent interaction towards all shared subject/item c between target node i and node j. Probabilities of activeness between pair of nodes $P_{i,j}$ are distinct or asymmetric as the computation is based on node j's perception of node i's activeness. Any $P_{i,j}$ that is greater than or equal to a threshold of 0.5 indicates node i being active with node j.

The probability of activeness between pair of nodes is then represented in a matrix which can be referred to as ***activeness trust matrix***. An example of a generated activeness trust matrix from a one-mode network (Such as table 3.2) can be seen in figure 3.5, where equation 3.1 was used in determining the elements in the matrix. A definite 'activeness' trust (probability value = 1) can be observed on the diagonal of the activeness matrix where each node i consider itself active or act alone towards certain items.

$$
\begin{array}{c c}
 & \begin{array}{cccc} [5] & [6] & [8] & [100] \end{array} \\
\begin{array}{c} \mathbf{[5]} \\ \mathbf{[6]} \\ \mathbf{[8]} \\ \mathbf{[100]} \end{array} &
\begin{pmatrix}
1.00 & 0.50 & 0.00 & 0.99 \\
0.50 & 1.00 & 0.00 & 0.97 \\
0.00 & 0.00 & 1.00 & 0.00 \\
0.01 & 0.03 & 0.00 & 1.00
\end{pmatrix}
\end{array}
$$

FIGURE 3.5: Activeness matrix for Example dataset 3.1 using its one-mode dataset (See table 3.2)

In a case of a more active node [2], the node will have a high trust degree from the other nodes. As node i is more active than node j, node j will desire to interact more with node i but node i may want to have less future interaction with node j (i.e. lesser trust degree will be assigned to node j) due to inactiveness of node j. For example, node 100 as shown in figure 3.5 was considered as a more active node by node 5 and node 6 while both node 5 and node 6 were considered to be less active nodes by node 100. Node 5 assigned a higher trust value than node 6 to node 100 as node 100 has acted more frequently on items that node 5 also encountered.

From figure 3.5, non-active node 8 which was identified based on its non-sharing behaviour, will also be included in the matrix to indicate the non-reachability of the node by other nodes. As a node i can not be reachable by other nodes j, the node should be consider to be untrustworthy i.e. $P_{i,j} = P_{j,i} = 0.00$ where $N_{i,j} = N_{j,i} = 0$ implying there is no interaction between the pair of nodes.

[2]An active node usually has more frequent interaction towards several alternatives that other nodes have also interacted in the past.

There could be some other factors that could affect a node's trustworthiness not being predicted accurately. The activeness trust considered to be an attitude, could either represent apportioned values for deceptive nodes who might be pretending to be knowledgeable or apportioned values for genuine nodes who acted honestly towards items. There is still the need to measure the familiarity to predict a node as a trustworthy node based on the factor to be discussed in the next section. The possibility of nodes being in a particular cluster could also be based on their familiarity measure. A clustering model using familiarity as a measure to check the accuracy of the generated clusters will be discussed in the next chapter to reveal the performance of a possible clustering algorithm that could be applied to a social network. The next section will describe how familiarity value between pairs will be evaluated using their activeness trust information in the 'activeness' matrix.

3.3.3 Familiarity based Trust

Based on the previous discussion of social interaction (section 2.6), it was discovered that people who consider themselves similar in behaviour will have the tendency to interact in the future based on how familiar they were with each other in the past. It was pointed out in [84] that similiarity between persons can be easily observed from the pattern of their interaction. In cases where a target node has few past interaction, it is possible to rely on the information on past interaction of similar nodes to the target node [145, 155].

Definition 3.4 (Familiarity degree)**.** Familiarity degree between two nodes in a network $Fam_{i,j}$ is a value of similarity based on the nodes' habits from their interaction with other nodes.

From a generated activeness matrix, familiarity $Fam_{i,j}$ between a target node i and any other node j is computed as the cosine similarity (using equation 2.19) between the nodes' activeness rating pattern (i.e. activeness trust weights in the activeness matrix) to others. The choice of using cosine similarity measure was based on the review carried out in section 2.6 which revealed that the output from this measure indicates similarities in pattern of interaction.

Example 3.1. *Considering the probability(Activeness) matrix in figure3.5, the familiarity degree $Fam_{5,6}$ between node 5 and node 6 is computed as follows:*

$$
\begin{array}{cc}
 & \begin{array}{cccc} [5] & [6] & [8] & [100] \end{array} \\
\begin{array}{c} \textbf{\textit{[5]}} \\ \textbf{\textit{[6]}} \end{array} & \begin{pmatrix} 1.00 & 0.50 & 0.00 & 0.99 \\ 0.50 & 1.00 & 0.00 & 0.97 \end{pmatrix}
\end{array}
$$

$$Fam_{5,6} = Fam_{6,5} = \frac{1 \times 0.5 + 0.5 \times 1 + 0.99 \times 0.97}{\sqrt{(1^2 + 0.5^2 + 0.99^2)} \cdot \sqrt{(0.5^2 + 1^2 + 0.97^2)}} = 0.89$$

Similarly, other familiarity degree from the matrix in figure 3.5 includes:

$$Fam_{5,100} = Fam_{100,5} = 0.68$$

$$Fam_{6,100} = Fam_{100,6} = 0.68$$

$$Fam_{8,5} = Fam_{5,8} = Fam_{8,6} = Fam_{6,8} = Fam_{8,100} = Fam_{100,8} = 0.00$$

From the output of the above example, it can be observed that node 5 and node 6 have high familiarity degree because both nodes had the same amount of activeness towards an item in the past. As node 100 had shared some items either between node 5 or 6 in the past, there is a certain degree of familiarity evaluated based on the amount of activeness towards the shared items.

If the amount of activeness from one user node i towards the set of shared items is greater than that from another user node j, then their familiarity degree $Fam_{i,j}$ will be low but if the amount of activeness from both nodes are the same then their familiarity degree $Fam_{i,j}$ will be high. The closeness in the rating pattern (i.e. activeness trust rating) from the activeness matrix also determines how familiar pair of nodes will be in the future. For instance, the rating pattern from both node 5 and node 100 observed from the activeness matrices in figure 3.5 were quite different in cases where there was a huge difference in their trust values apportioned to node 5 and node 6. This leads to the familiarity degree $Fam_{5,100}$ being lower than the familiarity degree $Fam_{5,6}$ between node 5 and 6 who have more similar pattern ratings towards other nodes.

Definition 3.5 (Global Familiarity degree)**.** To measure how trustworthy a node i will be towards a social network of n members, the mean Familiarity value Fam_i is taken over all familiarity degree between the node i and each node j. This global familiarity degree represents a consensus level of trustworthiness for a target node reached by all nodes in the network.

$$Fam_i = \frac{\sum_{j=1}^{n-1} Fam_{i,j}}{n-1} \tag{3.2}$$

This correlates to the "popularity" trust of a member in a community described by Nepal [145] which is based on positive and negative feedbacks towards contexts. It was further revealed in previous research work [155] that this positive or negative feedbacks can be determined based on the differences in their frequency of interaction which implicitly measures how nodes appreciates the services (message distribution) from a particular node when they are satisfied. This satisfactory behaviour can also be observed in previous example (figure 3.5) where an active node will receive a high familiarity degree by the whole network only if it has received appreciation (equal or similar frequent interaction) from several other nodes, in other words, some or all of the computed trust values apportioned by both a particular active node and other nodes to each node in the activeness matrix were similar in pattern. An inactive or less active node who had previously had no or fewer activities in the past similar to others will receive a lower global familiarity degree (0.00 for inactive nodes and $Fam_i < 0.5$ for less active nodes) as there will be no similar pattern between the node and other nodes in the activeness matrix.

Example 3.2. *Considering all familiarity values between a target active node* 5 *and other nodes,* $Fam_{5,6}$, $Fam_{5,8}$ *and* $Fam_{5,100}$ *:*

the global familiarity degree for the active node 5 *is*

$$Fam_5 = \frac{Fam_{5,6} + Fam_{5,8} + Fam_{5,100}}{4-1} = \frac{0.89 + 0.00 + 0.68}{3} = 0.52$$

Similarly to the process in example 3.2, the global familiarity degree Fam_6 and Fam_{100} are 0.52 and 0.45 respectively. Fam_5 being equivalent to Fam_6 is justified with the familiarity degree $Fam_{5,6}$ or $Fam_{6,5}$ which was initially evaluated using the cosine similarity measure to reveal how similar node 5 and node 6.

An empirical test on a complete two-mode dataset sample (i.e. Social forum dataset described in section B.2) with 20 user nodes and 211 item nodes still show that both node 5 and node 6 will be familiar with each other based on their interaction to similar items. This further reveals that the trustworthiness of a node cannot only be evaluated from the activeness (predicted trust) of the node in an activeness matrix but it should be evaluated based on the similar pattern with several other nodes in apportioned trust observed from the activeness matrix. Unlike example 3.2, the empirical test on the complete data sample showed that node 100 will be more trustworthy than any other node (including node 5 and node 6) based on the similar pattern observed from the activeness matrix.

It was observed that there was no significant difference between the outputs from the activeness matrix using the sum method and Newman since they basically generated similar values of global familiarity computed for the nodes in the social network. Further observation using activeness matrix (See figure B.11) generated from another two-mode dataset (i.e. 'Hollywood film-music' dataset described in Section B.4) revealed a global familiarity output (See table B.3) similar to that (See table B.2) of previous two-mode dataset (i.e. Social forum described in section B.2) as it predicts the trustworthiness of each social actor (i.e. Producer) based on their activeness with other social actors.

3.4 Summary

This chapter has described the importance of trust in situations where there is no or insufficient information to support decision making. Trust was considered to be an attitude demonstrated by an entity towards another entity that has been active. A proposed algorithm for computing trust was presented with several steps which enable the algorithm to conform with the theory of interpersonal behaviour where both a situational condition (i.e. attitude) and habits (i.e. patterns) are factors that have an impact on behaviours. The algorithm was revealed to initially generate activeness trust values that are considered to be a predicted attitude towards contexts. As there is the possibility of entities being active or non-active in a social network, activeness trust values was estimated based on probability expectation value of possible outcomes [101]. The generated activeness trust values, stored in an activeness matrix were demonstrated with an example that described how a two-mode network are observed to be the source of information to measure the activeness of an entity.

Further evaluation was required in the proposed algorithm since a node could have deceptively acted just to mislead other nodes of its trustworthiness. Familiarity-based trust was then considered to be predicted as the characteristics or behaviour of trustworthiness and this can be computed based on similarity in patterns from an activeness matrix where all activeness trust relationship between nodes is stored. Cosine similarity measure, considered from previous literature review (See section 2.6) to be a suitable measure of patterns, was used in estimating familiarity-based trust. The familiarity-based trust can be considered to be a more realistic method for computing trust as it emulates the real-life scenario where a person will trust another person only if they have similar behaviours. A person or entity with high familiarity value does indicate that the person or entity is an active member of a social network with other people or entities.

Outliers amongst data-points of familiarity-based trust for entities should not be considered as irrelevant information since some of them represent the points for inactive or less active nodes that could be useful in further analysis of the data for recommendation processes. Experiments described in the next chapter of the thesis will further reveal the usefulness of both the activeness matrix and familiarity-based trust values where the behaviour of most active members could be clearly observed in their individual group generated using a suitable clustering technique.

Chapter 4

Generating Clusters of trustworthy members from an Activeness Matrix

4.1 Introduction

This chapter will reveal how clustering can be applied to an 'activeness' matrix generated from a social network data. There is the need to further explore the social network to determine how members belong to certain groups. Several clustering techniques could be applied to the 'activeness' matrix based on their ability to utilize social context in generating clusters.

In the real world, a person who seeks advice on a subject will accept the opinion of a group when there is a general agreement amongst the group members. The identification of an influential member via a suitable clustering technique is also considered to further support the theory that the most trustworthy member from a group will influence the co-members of its group. Current clustering techniques tend to either refer to identified 'attractors' [56] or 'exemplars' [67] as the clusters' influential members that are identified based on their degree of ties (i.e. *activeness*) with other members in their individual cluster. But there is the need to consider the trustworthiness of a potential influential member in motivating others in their group.

4.2 Effective Clustering on a Social Network

With social actors having related variables being measured, the similarity between themselves will enable groups (clusters) to be determined efficiently. Clustering seems to be the best possible measure to analyse and determine if small groups still maintain their members or lose them [209]. The most difficult problem from clustering technique is the identification of the best way to minimize the production of noisy cluster output which consist of outliers that affect the accuracy of the output. Clusters generated from data with outliers will be indistinct, and eliminating the outliers from the data before clustering might worsen the problem [216, 223]. These outliers should not be completely dismissed as they could be useful data-points representing less active entities who might have been less active for different reasons (Described in section 2.3).

Some possible clustering techniques that could be applied to a network structure have previously been discussed (Section 2.9). They include Highly connected clustering (HCS), Markov clustering and Affinity propagation (AP) clustering. Unlike other clustering techniques such as K-means clustering and density-based spatial clustering of application with noise (DBSCAN), these clustering techniques do not require pre-specified parameters (e.g. number of clusters and number of members in a cluster) to generate clusters, that is, the clustering result is not predetermined with a given parameter.

Also, clustering techniques such as the EM clustering has the uncertainty property that a data-point might belong to several clusters with different degrees i.e. a data-point might have several degrees of membership to various clusters. This type of clustering technique is not required in the current work of this thesis as it is expected that a data-point must certainly belong to a particular cluster. In the current work of this thesis, it is expected to use an accurate-generated group (cluster) to determine if a member of the group will behave like other members of the same. This idea is based on the theory of social cognitive [15] where a member will only behave like others after observing their similar behaviour.

The three clustering techniques, HCS, Markov and AP clustering are focused on social context where social entities (i.e. Nodes) interact with each. With the application of the three different clustering technique, social context will be evaluated to retrieve groups based on nodes' acceptance of belief in belonging to the groups. Analysis of the social context using clustering reveals how ideas or information could be shared amongst entities within groups. Various groups (clusters) are generated by clustering based on similarity in attributes (i.e. opinions or behaviours) amongst entities that could persuade or guide each other on the group's opinion to ideas or items. Persuasive members of groups are referred to as *influential members* who could be representatives of the groups as they strengthen ties amongst co-members of their group.

4.3 Applying Clustering to Activeness matrix

As it is required to further know and understand how the nodes are linked together based on their activeness towards items, there is the need to apply a clustering algorithm which will minimise the noise in its output. A one-mode social data (Similar to that discussed in figure 3.3.1) of the Facebook-like forum dataset (See Appendix B.2) which was generated with ties between nodes of the same set reveals their random walks amongst themselves in a social graph (See Figure 4.1). All this can further be represented in the activeness matrix (See figure B.1 in Appendix B.2) which predicts if pairs of nodes will be connected in the future. As a node can randomly travel or flow to several connected nodes, the node is more likely to remain in the same cluster with these connected nodes than being a member of other clusters.

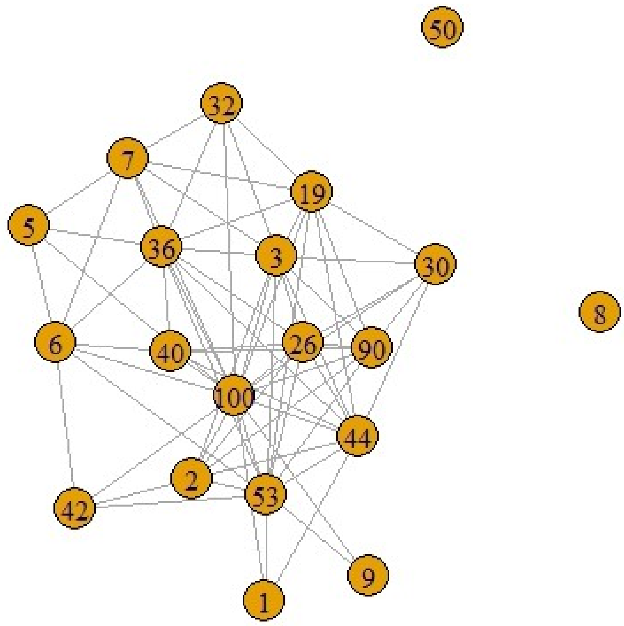


FIGURE 4.1: A social graph of sample data with 20 users from Facebook-like forum.

There is also the need to normalize column vectors of the activeness matrix (See Figure B.2 in Appendix B.2) to obtain unit vectors (i.e. length of each column from the matrix is 1). This will enable any of the applied clustering algorithm to easily analyse the relationship between the nodes from the social network data where social variables (amount of ties and activeness trust) are considered important in determining the clusters. Analysing the activeness matrix may reveal how some of the nodes could be identified as similar members based on either their similar active or inactive behaviour from previous social network data.

Different similarity measures, which play an important role in the clustering process, might only be suitable for certain cases. In the case of a social context where similarity is defined based on pattern, it was decided to test three different similarity measures with a suitable clustering technique to be explored later in this chapter. These include:

- Pearson Correlation measure
- Cosine similarity measure
- Negative squared Euclidean distance

Based on the review carried-out in chapter 2, the Hamming distance measure was decided to be disregarded as a suitable similarity measure to test along with clustering techniques since it seems to favour cases such as when certain nodes that have similar behaviour of having no ties with others (i.e. $P_{i,j} = 0.00$). With the measure, it will be evaluated that those nodes will be similar ($S_{i,j} = 1.00$) but it is inappropriate to consider them similar since their 'unreachable' behaviour to others does not necessarily infer that they have similar interaction pattern.

4.3.1 Applying Highly connected clustering (HCS)

The first clustering tested on an activeness matrix generated (either sum-based or Newman-based) from a sample data was the highly connected clustering (HCS) which has previously been discussed to be based on the fact that a cluster is formed only if the edge connectivity $\lambda(G)$ of a graph or subgraph is greater than half of the total vertices in the graph or subgraph. The three similarity measure earlier mentioned above were not considered suitable for the HCS clustering algorithm since the technique was originally developed by Hartuv & Shamir [86] with the definition of similarity to be based on the 'shortest distance' idea where nodes that can reach themselves within a short path will be considered similar to each other.

The activeness matrix had to be refined by converting it to an undirected graph even though it represents a directed graph relationship between nodes. The directional information (including weights on ties) is not required in the HCS algorithm since the most important information used in determining a cluster is the ties or connectivity between the nodes. This requirement also justifies the decision to exempt the self-loop information in the activeness matrix while applying HCS as a tie or edge that connects a node to itself will not obey the connectivity idea in HCS. As earlier mentioned, normalizing the matrix will further refine it to become a unit matrix for easy analysis. The 'refined' matrix can be referred to as transition matrix [148] where the probabilities of changes to various states from the activeness matrix are revealed. The weight elements in the transition matrix represent the degree of appreciation to past interactions between several nodes and a target node (See figure B.2 in Appendix B.2).

The HCS carried-out on a transition 'activeness' matrix (See Figure B.2) from an activeness matrix in figure B.1 resulted in five clusters being generated where four of the clusters, $cluster[1]$, $cluster[3]$, $cluster[4]$ and $cluster[5]$ were singletons (See figure B.3 in Appendix B.3.1). The single members in each of these singletons are either less active or non-active nodes in the social network. It will be reasonable to perceive that these singleton members will not be persuaded by each other or any member of another cluster.

Another cluster, $cluster[2]$ generated not as a singleton but based on the activeness (i.e. frequent interaction or amount of ties to others) from the sample social data where the nodes have ties with more active nodes in the network. Node 100 from the cluster is considered as the *most active node* (also known as the most trustworthy node observed in table B.2) that originally enabled all members of $cluster[2]$ (See figure B.3) to be reachable to each other in the social graph from the sample data (See figure 4.1) but these nodes were clustered together because they also had alternative paths to reach other active members of the cluster. But we cannot completely rely on this result as there is still the probability that some members of this cluster will decide to disengage from the group since their degree of activeness is probably not similar to others.

The presence of outliers in the activeness matrix kindled the application of HCS clustering to generate inaccurate clusters which seemed unclear to interpret. Even though there were less or no previous activities from these outliers, it will be a good idea to classify these members with active members of other clusters as they could seek advice from these active members. Other clustering algorithms previously discussed could be considered as alternatives for reducing this issue.

4.3.2 Applying Markov Clustering

Markov clustering algorithm requires a probability matrix and two parameters: expansion e and inflation r. With Markov clustering algorithm, the clusters formed are based on the social activeness amongst the nodes, as also observed from the HCS clustering algorithm, where nodes are considered to belong to the same cluster if they can randomly travel amongst themselves.

From the test carried out using Markov clustering technique on the transition 'activeness' matrix (See Figure B.2), five clusters (See figure B.5) were generated where one of the clusters formed is based on predicted non-activeness trust values from certain nodes observed in the original 'activeness' matrix (See Figure B.1), while the other clusters are formed based on predicted activeness values from potential influential members (i.e. 'attractors') also observed in the original activeness matrix .

TABLE 4.1: Generated clusters from the application of Markov clustering to the transition matrix in figureB.2

	Cluster	Attractor	Members
First Category	cluster[3]	None	3,6,8,26,44,50,100
Second Category	cluster[1]	53	1,9,53
	cluster[2]	2	2,42
	cluster[4]	36	5,7,19,32,36
	cluster[5]	30	30,40,90

The first category of cluster (See table4.1) generated based on activeness trust values (observed in figureB.1) revealed that members of this cluster were actually singletons that have low (i.e. $P_{i,j} < 0.5$) or no activeness trust values for other nodes. For example, $node100$ is predicted to have more less (i.e. low activeness trust value) or no interaction with certain nodes as compared with other nodes' predicted interaction. Nodes that have similar pattern, where a lot of low activeness trust value and few high activeness trust values ($P_{i,j} > 0.5$) for other nodes, will fall into this category of cluster. From Markov clustering, all members that fall into this first category are represented in the output matrix (i.e. Equilibrium State Matrix, ESM in figure B.5) as those nodes that have a self-loop but no ties with other nodes. Even though Node 100 was previously computed in Table B.2 to be the most trustworthy node (i.e. most active node), it is not considered as the most influential member amongst all members in this cluster since each member in the cluster will not accept or trust the opinions of other members.

The second category of clusters formed based on activeness of potential influential members is generated by considering the familiarity values computed (See Table B.2) in the previous chapter. These clusters are generated after completing the identification of nodes that fall in the first category of cluster previously discussed. The identification of the remaining nodes with other clusters that fall in the second category could be understood by following a number of steps.

1. Consider the highest trustworthy node (i.e. $Fam_i \geq 0.5$) amongst the remaining nodes from *'familiarity spectrum'* (E.g. table B.2) as a potential influential member i_p of a cluster. An influential member must also have at least one high predicted trust value for other nodes observed in the original activeness matrix.
2. Select other nodes j that have received the highest trust value (i.e.$P_{i,j}$) from the potential influential member i_p.
    - Node j will still be selected even if the trust value received from the influential member is shared by another node i_f (i.e. potential influential node or attractor already identified with an existing cluster).
    - Node j will still be selected if it received the highest trust value from an existing member or a node that has received the highest trust value from the influential member.
3. Then, consider the potential influential member i_p and the other nodes j identified from step 1 and step 2 respectively as members of a cluster.
4. Repeat step 1 to step 3 if some nodes have not been identified with a cluster or group.

Since Markov clustering algorithm still generates singletons which were represented as diagonal elements in the ESM and no attractor or potential influential member can be identified here, there is the need to consider other clustering algorithm that will accurately evaluate the similarity between entities and relating the similarities with either activeness matrix or transition matrix in generating clusters and influential members.

4.3.3 Applying Affinity Propagation (AP) clustering

Further analysis is required as previous clustering algorithms did not explore the use of various similarity measures to determine the clusters from a given data. As AP clustering algorithm requires input preference $s(i,i)$ in generating its clusters, the transition 'activeness' matrix (See figure B.2) could be considered as an input data where similarity matrix will be retrieved to be able to obtain the suitability of a data point (i.e. $s(i,i)$ which means suitability of node i) to be an exemplar or potential influential node. This similarity matrix is retrieved after a similarity measure has been applied to the transition 'activeness' matrix. As earlier discussed in chapter 2, $s(i,i)$ is best described as the median or minimum value amongst similarity values from the similarity matrix.

Considering the same transition 'activeness' matrix (See Figure B.2) used with previous clustering algorithms, the AP clustering can be applied along with various similarity measures to determine the cluster output. To avoid production of singleton clusters as observed when the required input activeness matrix consists of outliers, it will be best to include self-loop in the transition activeness matrix which will be made less-complex for cluster analysis. Apart from the similarity matrix required in determining how similar nodes are clustered together, the relation in both activeness matrix and transition matrix (i.e. normalized activeness matrix) is also to be used in explaining how nodes became members and how influential members were chosen.

1. Negative Euclidean Distance based AP clustering

From the use of negative Euclidean distance measure [67] along with the AP algorithm on transition matrix (with self-loop) (See figure B.2), six clusters were produced with an input preference value (i.e. -2.14) selected as the median value (See Figure B.7) amongst values from the similarity matrix. To further provide justification on how a cluster and its influential member were identified in this case, we can consider both $cluster1$ and $cluster4$ from the output (See figure B.7) for clear explanation.

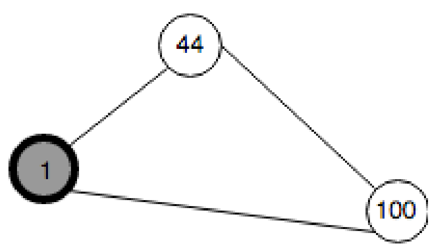


FIGURE 4.2: Graph of $cluster1$'s members from AP clustering using Negative Euclidean distance measure

Node 1, 44 and 100 are clustered together in $cluster1$ based on the fact that the less active $node1$ has high possibility of ties with $node44$ and $node100$ (i.e. $P_{1,44} = 0.75$ and $P_{1,100} = 0.56$ respectively) as observed in the activeness matrix (See figure B.1) where $node44$ also have a high predicted trust value as that of $node1$ for $node100$ (Most active node) and thereby forming a 'closed tie' [154] as shown in figure 4.2. From the transition 'activeness' matrix (See figure B.2), the exemplar (i.e. potential influential member) $node1$ was identified based on the fact that it received the highest weight (i.e. 0.31 received from the most active node 100) after comparing all weights shared amongst all members of the cluster. This can indicate that node 1 is more appreciated than node 44 and node 100.

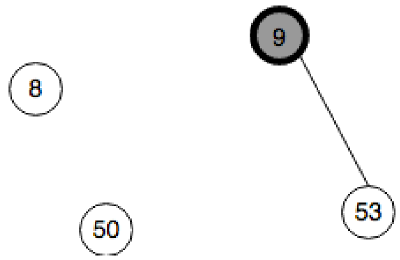
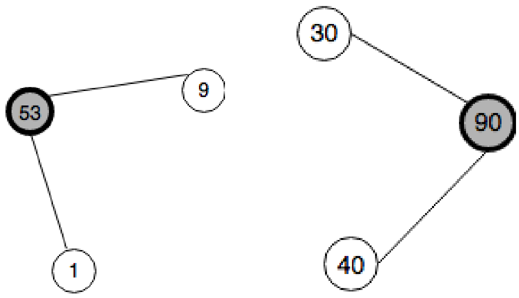


FIGURE 4.3: Graph of $cluster4$'s members from AP clustering using Negative Euclidean distance measure

$Cluster4$ consist of members who are either non-active or an active node considered to be closer to a less active node in the whole social network. This cluster reveals that a less active node (i.e. $node9$) has a more possible tie with an active node (i.e. $node53$) that assigned the highest activeness trust value ($P_{53,9} = 0.18$) to the less active node (See figure B.1). $Node9$ is identified as the exemplar based on the fact that it received the highest weight (i.e. 0.82 received from the most active node 53) in the transition 'activeness' matrix (figure B.2) after comparing all weights shared amongst all members of the $cluster4$.

The cluster output from this negative Euclidean distance based AP algorithm seems more suitable for simulating how non-active members could be attracted to a less active node of a cluster but there is still the need to compare this output with that based on other similarity measures for generating a similarity matrix since there might be conflict between the most active member and other members that seem closer to the exemplar of the cluster. The exemplar here is considered to be non-expert to advise its members as it lacks experience interacting with others on items (i.e. Less active behaviour with others observed in figure B.1). Also, in this case, there are still certain clusters (e.g. cluster 4) with less active members that are not structural equivalent (i.e. two or more members sharing common ties) with any other member in their individual cluster.

2. <u>Cosine similarity based AP clustering</u> Using Cosine similarity measure along with AP clustering algorithm generated more clusters (See figure B.10) than in the two previous cases as a higher input preference (i.e. median similarity value of 0.067) was obtained from the similarity matrix. After several iterations in this case, the problem of obtaining singletons with each non-active node being the member to a singleton is still observed in cluster 3 and cluster 5. Here, the members of some clusters have close ties to each other and most of them assign similar high activeness trust values to other members. For example, node 5 and node 6 of cluster 2 (See figure B.10) assign same trust values (i.e. 0.96 and 0.50 respectively as observed in figure B.1) to node 7 and node 36. Also, members of some clusters (i.e. members in cluster 6 and cluster 7) that might be less active nodes are not structural equivalent as their network structure are not closed where all members have ties with each other (See figure 4.4).



((A)) Network Structure of Cluster6

((B)) Network Structure of Cluster7

FIGURE 4.4: Network structure of some clusters from AP clustering using Cosine Similarity measure

The identification of exemplar for each cluster from the Cosine Similarity based AP algorithm is best comprehended in a similar way to that earlier explained for *negative Euclidean distance based AP algorithm* where an exemplar is the member of a cluster that has received the highest weight value amongst other members (See figure B.2).

3. Pearson correlation based AP clustering

Applying the AP algorithm along with Pearson correlation distance measure to the same transition matrix generated five clusters from input preference of -0.10 (See figure B.8). To justify why certain nodes are members of each cluster and a node chosen as their exemplar, $cluster1$, $cluster3$ and $cluster5$ were further explained below.

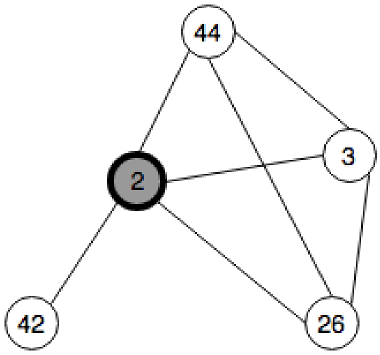


FIGURE 4.5: Graph of $cluster1$ from AP clustering using Correlation distance measure

From cluster 1, node 2,3,26,42 and 44 were clustered together based on the fact that node 3,26 and 44 assigned close or similar activeness trust values (i.e. $P_{3,2} = 0.46$, $P_{26,2} = 0.50$ and $P_{44,2} = 0.50$) to node 2 as observed from the activeness matrix. The transition 'activeness' matrix (See figure B.1) also reveals that these nodes are predicted to have ties with node 2 as they have close appreciation value (i.e. 0.17, 0.19 and 0.19 respectively) for node 2. Similar to the previous cluster output using the negative Euclidean distance, node 42 is still tied to node 2 based on the fact that node 42 receives more appreciation of its interaction (i.e. 0.60) by node 2 as observed from the transition 'activeness' matrix. The identified exemplar or potential influential member (i.e. Node 2) of this cluster is the node amongst members of the cluster with most ties with other members of the cluster and more appreciated (i.e. 0.15) by the less active Node 42 of the cluster as observed in the transition 'activeness' matrix (See figure B.2).

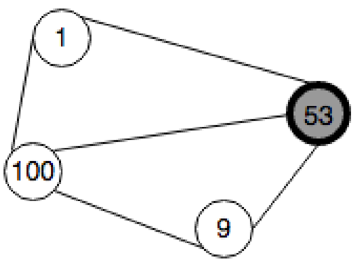


FIGURE 4.6: Graph of $cluster5$ from AP clustering using Correlation distance measure

The justification for members being in *cluster*5 (See figure 4.6) and a node being a potential influential member (i.e. exemplar) is similar to that earlier described for *cluster*1 (See figure 4.5). Even though two nodes (i.e. *node*53 and *node*100) were observed to have equal number of tie more than other members of *cluster*5 as shown in the network structure (See figure 4.6), the less active member (i.e. *node*9) from the cluster still appreciates *node*53 more (i.e. 0.10) from amongst the two nodes as observed from the transition 'activeness' matrix.

From the clusters generated using the *Correlation based AP algorithm*, it is clearly seen that structural equivalence amongst the members is obtained in most of the clusters. In *cluster*5, the node with the lowest activity (i.e. *node*9) in the whole social network (See figure 4.1) is more structural equivalent to other members in this cluster than its members in the cluster previously generated with the use of negative Euclidean distance measure.

*Cluster*3 (See figure B.8) shows that there might still be singletons from that cluster as there will always be a competition amongst node 8 and node 50, on who to be a representative of the cluster (See figure B.9). But it is still advantageous to rely on the initial cluster output from the correlation based AP clustering algorithm as it remains consistent after several trials of applying the algorithm on the transition matrix. This indicates that it is better to have these non-active nodes in the same cluster as it will be easier to recommend items preferred by an average node of another cluster than recommend their own preferred items to them. Node 8 was considered as an exemplar in this cluster (See figure B.8) by node 50 based on the fact that node 8 was more active with its own preferred items than node 50 in the original two-mode network data (See Appendix B.2).

Further analysis carried out with Correlation based AP clustering on an activeness matrix of another dataset (i.e. 'Hollywood Film-music' Dataset discussed in Appendix B.4) revealed similar cluster output (see figure B.15) with that in figure B.8. The output from this analysis revealed better clusters with one of them having all non-active nodes (i.e. nodes 21,24,27,29 and 61) clustered along with most active node (i.e. node 30) of the whole network. This might be due to some of these non-active nodes (i.e. nodes 24 and 27) previously having frequent contact with items that they do not share with other nodes.

From results using any of the AP clustering algorithms, there is still the need to confirm if these identified exemplars are actually influential members. The next chapter will reveal the main contribution of the thesis where a proposed framework will be presented with a further investigation using Singular value decomposition (SVD) to confirm this posteriori and determine which AP clustering algorithm is suitable for recommendation.

4.4 Summary

The exploration of how members belong to certain groups or clusters was carried with various clustering techniques that are compatible with social context. The AP clustering algorithm tends to be more effective as it was able to group members based on their degree of activeness and structural equivalence depending on the type similarity measure being applied with the algorithm.

As similarity measure is usually associated with clustering, three common similarity measures suitable in social context were tested for measuring the degree of ties (i.e. activeness) and similarity in the pattern of past activities (i.e. Interaction) of nodes or entities that could be structural equivalent. These similarity measures include the Euclidean distance, Cosine similarity and Pearson correlation measures. Even though the Euclidean distance measure is known to measure distances based on only the degree of ties [211], it was still necessary to consider it with AP clustering algorithm for the empirical test to check if clustering will be best without considering structural equivalence in generating clusters. It was proven that structural equivalence is observed mostly from correlation based AP algorithm as it seems to best justify the members joining the clusters generated. The application of correlation based AP clustering algorithm on an activeness matrix was also shown to be effective as outliers (i.e. non-active members) are not partitioned into singletons (i.e. individual cluster with a single member) but they are merged together in a single cluster. Further analysis with another data, where non-active members were actively in contact with items that they preferred revealed outputs with the non-active members being clustered with the most active node(MAN) of the whole data.

Chapter 5

A Framework for Motivating Inactive Members of a community: An Evaluation

5.1 Introduction

It was discovered earlier that there is the need to encourage inactive users in becoming active in a social community since these users find it difficult to decide on the way to act towards new items or cases presented to them. The main contribution of this thesis was to develop a framework that utilizes the proposed novel model in chapter 3 for determining the trustworthiness of entities (i.e. users) and the proposed model in chapter 4 for generating clusters of trustworthy users. This framework is expected to be an enhancement of a recommender system which requires preference information of inactive entities in carrying out an accurate recommendation of items for the entities. It has been shown in chapter 4 that these estimated trustworthy entities are clustered based on their activeness degree observed from activeness matrix generated as described in the example given in chapter 3 (Section 3.3.2). This chapter will reveal how the cluster output from a suitable clustering technique could support a recommendation. An empirical test using SVD recommendation algorithm on user-item data linked with the cluster output will be carried out to determine if members (especially the inactive or less active members) could be motivated by the most active member of a cluster.

5.2 Trust-Cluster based Recommendation using Singular Value Decomposition (SVD)

The novel framework of this thesis which is considered to be a solution in resolving the cold-start problem affecting inactive users of any system can be referred to as *Trust-cluster based Recommendation*. The framework can be considered as an enhancement for a recommendation process that requires every user's preferences to enable an accurate recommendation of items to be provided for each user. If there are few activities of some users available, it might be difficult to estimate the preferences of those users. But with the proposed framework (See figure 5.1), it is possible to encourage these less active users in accepting recommended items.

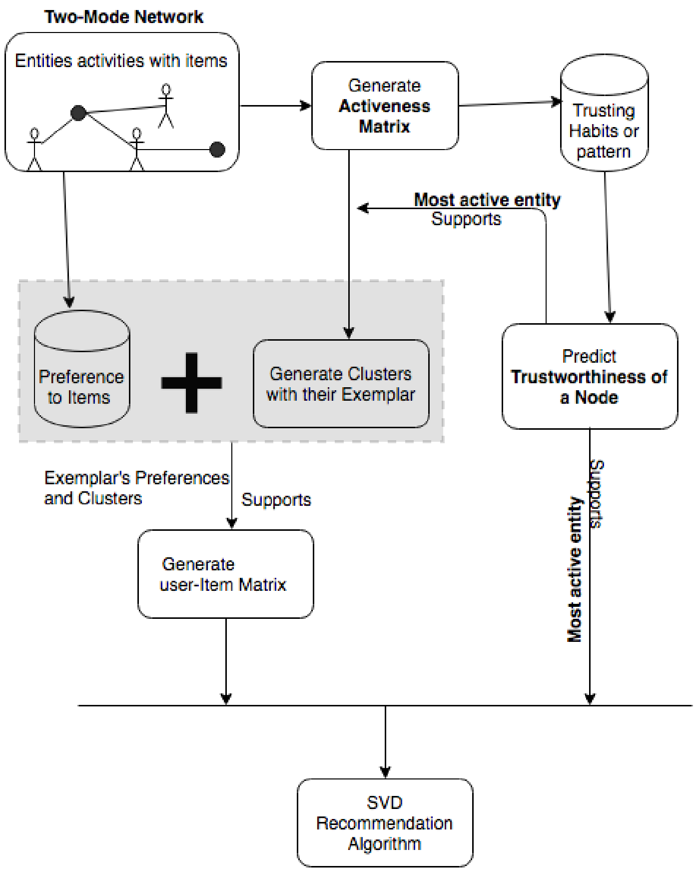


FIGURE 5.1: A framework of Trust-cluster based Recommendation

The two-mode network is shown in figure 5.1 to be an important requirement as there is the need to filter all preferred items by a potential exemplar amongst clustered-trustworthy users. These filtered preferences to items are required in the recommendation process for members of a cluster which the exemplar also belong to.

It was earlier revealed in the previous literature review that it is necessary to transform the two-mode network into a one-mode network for network analysis measures to reveal the association between two or more entities. This transformation was required in the framework for determining the activeness trust attitude between entities which is represented in the activeness matrix (See section 3.3.2). With habits or patterns of entities observed from a generated activeness matrix, the trustworthy behaviour of entities is predicted to support the idea that entities will decide to behave in a similar way as that of their trusted neighbours.

As it is possible that there might be some inactive members amongst the entities that require the recommendation of items, we can then rely on the preferences of an exemplar and a cluster of the exemplar to support the model for recommending items to trusted members of a cluster in the framework (See figure 5.1), where a user-item matrix is required. The generated clusters was previously explained in chapter 4 to be based on activeness information between entities observed in the activeness matrix. The exemplars of generated clusters were also described to be identified as potential influential members that have been rated higher than others by the most active entity [1].

The user-item matrix in the framework is a matrix of implicit ratings (e.g. frequency of visiting an item, frequency of conversing about an item or frequency of employing experts for a job) by trusted members of a cluster. SVD recommendation algorithm was previously revealed in chapter 2 to be an effective and efficient algorithm as it can be used in reducing a high dimensional user-item data to a low dimensional space. This algorithm is required in the framework for predicting if any member of a cluster will accept a recommended item preferred by the exemplar of the cluster.

5.3 Building a model towards recommending items for clustered-trustworthy users

Several components need to be in place before building the model towards recommending items preferred by a cluster's exemplar. One important component is the support from a most active node or entity that is also a member of a generated cluster. Information on the most active node is initially required in the framework (See 5.1) to support in

[1] Most active entity is identified amongst the members of a cluster with their trustworthy value (i.e. Familiarity based trust demonstrated in 3.3.3)

the identification of exemplars. Other information related to the most active node's behaviour towards items is also required in determining the quality of items recommended to members of the cluster to which both the exemplar and the most active node belong to. This would be possible based on the idea that a more active member of a group will add value to items by accepting these items or rating the items high.

As earlier revealed in chapter 4 that AP algorithm is a more suitable clustering technique to be applied on an activeness matrix, we expect to further validate which of the AP algorithm will be more suitable in the framework. Subsections below will describe in details how the required implicit rating matrix (i.e. user-item matrix) will be generated and how the recommendation of items for clustered members is carried out, demonstrated with examples.

5.3.1 Generating a user-item rating Matrix with cluster members

A user-item matrix will be required for SVD algorithm, where Y items are the column of the matrix presenting previous items (i.e. Items from original social data) that have previously been used as a context in past interaction by an exemplar or potential influential member of a cluster. The row of the matrix indicates X users that are members of the cluster in which the exemplar or influential member belongs to. The elements in the matrix are original elements (i.e *'frequent-interaction' value*) $r \in \mathbb{R}^{X \times Y}$ representing the frequency of interaction by either the exemplar node or other members of the cluster. If a member U_i of the cluster cl has frequent interaction with item P_j, then r_{ij} is the rating value in the matrix $M_{X \times Y}$, otherwise 0 is assigned to indicate that the value is unknown i.e. $r_{ij} = 0$.

$$
\begin{array}{c c}
 & \begin{array}{cccc} P_1 & P_2 & \dots & P_Y \end{array} \\
\begin{array}{c} U_1 \\ U_2 \\ \vdots \\ U_X \end{array} & \begin{pmatrix} r_{11} & r_{12} & \dots & r_{1Y} \\ r_{21} & r_{22} & \dots & r_{2Y} \\ \dots & \dots & \dots & \dots \\ r_{X1} & r_{X2} & \dots & r_{XY} \end{pmatrix}
\end{array}
$$

Items presented in this type of matrix are expected to be accepted by members of a cluster based on the fact that members of a group will consider accepting items preferred by an exemplar whom they trust. However, since the identification of cluster's exemplar was described in chapter 4 to be based on the appreciation level of the most active node of the cluster, we should not expect all items to be accepted by other members of the cluster.

5.3.2 Recommendation for Cluster members

Clusters generated by negative Euclidean distance based AP algorithm (See subsection 4.3.3) can further be analysed by applying SVD recommendation algorithm to the user-item rating matrix by cluster members representing the simulation for a recommendation of items by a cluster's exemplar to other members of the cluster. The analysis on clusters reveals that SVD algorithm which estimates the predicted rating or 'frequent-interaction' value for each item also indicate if inactive or less active members of clusters accept or reject an item. To demonstrate this, cluster 1 and cluster 4 previously generated in figure B.7 were considered since those are the clusters with inactive or less active nodes. Items preferred by each exemplar observed from original two-mode social network data are considered as the potential recommended items to be used in the user-item rating matrix.

Example 5.1. *Considering the exemplar and other members of cluster 1 generated from the use of negative Euclidean distance based AP algorithm (See figure B.7), a* $m \times n$ *user-item rating matrix with n preferred items (i.e. interacted items filtered out from original social data) by the exemplar is generated.*

$$
\begin{array}{c c}
 & \begin{array}{cccccc} [1] & [39] & [102] & [154] & [357] & [459] \end{array} \\
\begin{array}{c} \mathbf{[1]} \\ \mathbf{[44]} \\ \mathbf{[100]} \end{array} &
\begin{pmatrix} 1 & 9 & 2 & 1 & 5 & 1 \\ 0 & 10 & 23 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 8 & 0 \end{pmatrix}
\end{array}
$$

FIGURE 5.2: Matrix simulating recommendation of items to members of cluster 1

Some of those items in the matrix have ratings from some cluster members that have also interacted with these items in the past (Observed from the original social network data) while other members that have no experience with the items will have a rating of zero (0) indicating missing or unknown value for the items.

After decomposing the matrix by following the same steps demonstrated in example A.2, a scaling (non-uniform scaling) matrix with the same dimension as that of the user-item rating matrix is retrieved:

$$\Sigma = \begin{array}{c} \\ \textbf{[Dim1]} \\ \textbf{[Dim2]} \\ \textbf{[Dim3]} \end{array} \begin{array}{c} \begin{array}{cccccc} [Dim1] & [Dim2] & [Dim3] & [Dim4] & [Dim5] & [Dim6] \end{array} \\ \begin{pmatrix} 25.77 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 10.36 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 5.93 & 0.00 & 0.00 & 0.00 \end{pmatrix} \end{array}$$

Where: *DimG for distinct integer $G \in \{1,\ldots,k\}$ represent the scaling-dimension category of either the user or item features while the elements in the matrix are scale factors. k is the highest value amongst values (i.e. m and n) from the dimension of the user-item rating matrix.*

Users latent feature matrix U which is also obtained as part of the output from decomposition,

$$U = \begin{array}{c} \\ \textbf{[1]} \\ \textbf{[44]} \\ \textbf{[100]} \end{array} \begin{array}{c} \begin{array}{ccc} [Dim1] & [Dim2] & [Dim3] \end{array} \\ \begin{pmatrix} -0.2430 & 0.7503 & -0.6148 \\ -0.9685 & -0.2231 & 0.1106 \\ -0.0542 & 0.6223 & 0.7809 \end{pmatrix} \end{array}$$

If we decide to consider the latent feature matrix of user 100, taking the row vector representing this user:

$$U_{100} = \begin{array}{c} \\ \textbf{[100]} \end{array} \begin{array}{c} \begin{array}{ccc} [Dim1] & [Dim2] & [Dim3] \end{array} \\ \begin{pmatrix} -0.0542 & 0.6223 & 0.7809 \end{pmatrix} \end{array}$$

and Items latent feature matrix V which is also retrieved from the decomposition of the user-item matrix,

$$V = \begin{array}{c c} & \begin{array}{cccccc} Dim1 & Dim2 & Dim3 & Dim4 & Dim5 & Dim6 \end{array} \\ \begin{array}{c} \textbf{\textit{[1]}} \\ \textbf{\textit{[39]}} \\ \textbf{\textit{[102]}} \\ \textbf{\textit{[154]}} \\ \textbf{\textit{[357]}} \\ \textbf{\textit{[459]}} \end{array} & \begin{pmatrix} -0.0094 & 0.0724 & -0.1036 & -0.1097 & -0.9797 & -0.1097 \\ -0.4606 & 0.4364 & -0.7463 & -0.1029 & 0.1387 & -0.1029 \\ -0.8851 & -0.2902 & 0.3530 & 0.0448 & -0.0603 & 0.0448 \\ -0.0094 & 0.0724 & -0.1036 & 0.9875 & -0.0928 & -0.0125 \\ -0.0639 & 0.8424 & 0.5349 & -0.0056 & 0.0075 & -0.0056 \\ -0.0094 & 0.0724 & -0.1036 & -0.0125 & -0.0928 & 0.9875 \end{pmatrix} \end{array}$$

Considering latent feature matrix of item 154 *by transposing V and retrieving the vector representing the item:*

$$V_{154}^{T} = \begin{array}{c c} & [154] \\ \begin{array}{c} \textbf{\textit{[Dim1]}} \\ \textbf{\textit{[Dim2]}} \\ \textbf{\textit{[Dim3]}} \\ \textbf{\textit{[Dim4]}} \\ \textbf{\textit{[Dim5]}} \\ \textbf{\textit{[Dim6]}} \end{array} & \begin{pmatrix} -0.0094 \\ 0.0724 \\ -0.1036 \\ 0.9875 \\ -0.0928 \\ -0.0125 \end{pmatrix} \end{array}$$

Based on equation 2.16, the prediction of user 100 *preference for item* 154 *can be determined by considering the average rating of user* $100's$ *in the user-item matrix, the latent feature matrix for both user* 100 *and item* 154.

$$pred_{v154,u100} = 1.5 + U_{100} \times \Sigma \times V_{154}^{T}$$

$$= 1.5 + (-1.11 \times 10^{-16})$$

$$= 1.5$$

Considering the recommendation of a less-interacted item (For instance, item 154) amongst the items, the most active node (i.e. node 100) amongst members of cluster 1 seem not to be interested in this item as its predicted rating computed with equation 2.16 is lesser than that of the exemplar (i.e. node 1 which has predicted rating of 4.17

for the item 154) and the other member (i.e. node 44 which has a predicted rating of 5.50 for the item 154) of the cluster. Thus, the most active node in this cluster is not an influential member to other members of the cluster.

Example 5.2. *Considering exemplars and other members of cluster 4 generated from the use of negative Euclidean distance based AP algorithm (See figure B.7), a $m \times n$ matrix with n preferred items (i.e. interacted items filtered out from original social data) by the exemplar is generated.*

$$\begin{array}{c} \\ \mathbf{[8]} \\ \mathbf{[9]} \\ \mathbf{[50]} \\ \mathbf{[53]} \end{array} \begin{array}{c} \begin{array}{cc} [19] & [93] \end{array} \\ \begin{pmatrix} 0 & 0 \\ 1 & 1 \\ 0 & 0 \\ 2 & 7 \end{pmatrix} \end{array}$$

FIGURE 5.3: Matrix simulating recommendation of items to members of cluster 4

After decomposing the matrix by following the same steps demonstrated in example A.2, a scaling (non-uniform scaling) matrix with the same dimension as that of the user-item rating matrix is retrieved:

$$\Sigma = \begin{array}{c} \\ \textbf{\textit{[Dim1]}} \\ \textbf{\textit{[Dim2]}} \\ \textbf{\textit{[Dim3]}} \\ \textbf{\textit{[Dim4]}} \end{array} \begin{array}{c} \begin{array}{cc} [Dim1] & [Dim2] \end{array} \\ \begin{pmatrix} 7.39 & 0.00 \\ 0.00 & 0.68 \\ 0.00 & 0.00 \\ 0.00 & 0.00 \end{pmatrix} \end{array}$$

considering user data (row vector) for node 53 from generated latent feature matrix U of users:

$$U_{53} = \begin{array}{c} \\ \textbf{\textit{[53]}} \end{array} \begin{array}{c} \begin{array}{cccc} [Dim1] & [Dim2] & [Dim3] & [Dim4] \end{array} \\ \begin{pmatrix} -0.9856 & 0.1688 & 0 & (1.1102 \times 10^{-16}) \end{pmatrix} \end{array}$$

and item data (column vector) for item 19 from latent feature matrix V^T of items

$$\begin{array}{cc} & [19] \\ V_{19}^T = \begin{array}{c} \textbf{\textit{[1]}} \\ \textbf{\textit{[2]}} \end{array} & \begin{pmatrix} -0.2898 \\ -0.9571 \end{pmatrix} \end{array}$$

Then, the prediction of user $53's$ *preference for item* 19 *using equation* 2.16 *is:*

$$pred_{v19_u53} = 4.5 + U_{53} \times \Sigma \times V_{19}^T$$
$$= 4.5 + 2$$
$$= 6.5$$

Even though the most active node (i.e. node 53) amongst the members of cluster 9 is interested in the less-interacted item (i.e item 19), it was only able to motivate a less active member (i.e. node 9 which has a predicted rating of 2 for the item 19) and not the non-active members (i.e. node 8 and node 50). Which means there is no need for the non-active members to be members of this cluster since they have no predicted rating (i.e. a value of zero) for the recommended item (i.e. item 19).

Definition 5.1. Based on these examples, a member of a cluster will be influenced to accept item recommended by an exemplar if and only if the predicted rating from the *most active node (MAN)* of the cluster towards the recommended item is greater than or equal to that of the exemplar.

As clusters generated from the negative Euclidean distance based AP algorithm have been proven to be incompatible with trust-cluster based recommendation framework, there is the need to consider other clusters generated from both Correlation based AP algorithm and Cosine similarity based AP algorithm. The clusters from either of the algorithm might reveal a better result where a most active member of a cluster will motivate other members to either accept or reject a recommendation.

Considering members of cluster 2 and cluster 5 from correlation based AP algorithm (See figure B.8) along with SVD algorithm might best demonstrate how less active nodes of these clusters could be motivated by individual exemplar to accept or reject items recommended. A cluster (e.g. Cluster 3 in Figure B.8) having only non-active nodes as members indicates that they will probably receive the same recommended items. From the application of SVD algorithm on cluster 2 of correlation based AP algorithm, it was revealed that other members (i.e. node 5 and node 40) apart from the exemplar (i.e. node 6) were influenced by the most active member (i.e. node 36) of the cluster to reject some items (e.g. item 355) which the exemplar preferred in the recommendation list. Based on the predicted rating of MAN $pred_{MAN}$ for item 355 being less than that of the exemplar (See table 5.1), it indicates that MAN is not interested in the item.

TABLE 5.1: A Check for potential Influence by the Most active members in clusters from Correlation based AP algorithm

Cluster	Item recommended	Most active member (MAN) ID	$pred_{MAN}$	$pred_{Exemplar}$	Are other members influenced to accept item	Are other members influenced to reject item
Cluster1	45	3	0.26	3.11	No	Yes
Cluster2	355	36	0.11	2.89	No	Yes
Cluster4	121	19	3.83	3.83	Yes	No
Cluster5	482	100	14.65	11.60	Yes	No

Using cluster 5's members from correlation based AP algorithm in further analysis, MAN (i.e. 100) of this cluster has the potential of influencing other members to accept items which the exemplar preferred and recommended to the group since its predicted rating ($pred_{MAN}$) for a less interacted item (i.e. item 482) is higher than that of the exemplar ($pred_{Exemplar}$). Observation of similar behaviour with other clusters generated with the Pearson correlation based AP algorithm can be seen in table 5.1.

As observed in chapter 4 that a cluster 1 (shown in figure B.10) generated from Cosine similarity based AP algorithm have similar members as that of cluster 1 in figure B.8) generated from Correlation based AP algorithm, the prediction result for any of these clusters will be the same as shown in the first row of table 5.1. But considering other clusters generated from Cosine similarity based AP algorithm in the analysis using SVD shows that nodes who belong to the same cluster might be influenced by either the exemplar or MAN of their cluster; which indicates there might be conflicts amongst members. An example of this behaviour can be observed in cluster 2 where the exemplar (i.e. node 5) of this cluster has a predicted rating of 2 which is greater than that of MAN (i.e. node 36). Other members of this cluster, node 6 and node 7 are motivated by the cluster's MAN and the cluster's exemplar respectively.

Considering another example when applying SVD algorithm on cluster 6 which is also generated from Cosine similarity based AP algorithm reveals that its MAN being the exemplar of the cluster was not able to influence other members of the cluster due to the fact that the exemplar of the cluster is not structural equivalent with other members of the clusters since the exemplar does not share the same tie with any of these members in the cluster (i.e. Node 53 does not share the tie to node 1 with node 9 or Node 53 does not share the tie to node 9 with node 1).

TABLE 5.2: Predicted ratings for items recommended by Exemplar in clusters from Cosine similarity based AP Algorithm

Cluster	Item recommended	ID of Members	Predicted rating	Exemplar ID
Cluster1	45	2	3.11	2
		3(MAN)	0.26	
		26	0.07	
		42	0.11	
		44	0.07	
Cluster2	72	5	2.00	5
		6	0.33	
		7	7.33	
		36 (MAN)	0.33	
Cluster4	527	19	0.62	32
		32	25.88	
		100 (MAN)	8.13	
Cluster6	550	1	0.14	53
		9	0.06	
		53 (MAN)	11.60	
Cluster7	49	30	0.12	90
		40	0.04	
		90 (MAN)	4.6	

SVD algorithm prediction from all the analysis carried-out clearly showed that there will be a consensus in a cluster only if the rating pattern amongst members observed in the matrix simulating recommendation were similar. A better consensus was observed in the example (See Table 5.1) with clusters generated by Correlation based AP algorithm as each cluster have members being structural equivalent to each other.

Further observation of correlation based AP clusters from another data sample, 'hollywood film-music data' (See Figure B.15) revealed similar output where less-active members (i.e. film producers) of a cluster is predicted to either accept or reject items (music composers) based on the predicted preferences of the cluster's MAN. From all the observation, it was concluded that definition 5.1 implies that recommended items are only accepted by other members either when the MAN of the cluster is familiar with most of the items or when the MAN of the cluster is the exemplar of the cluster.

5.4 Summary

This chapter has revealed and evaluated the main contribution of the thesis where less-active entities of a social network could be encouraged in behaving like active entities. This framework was described to be important in situations where there is insufficient information (i.e. activities) of an entity in supporting the recommendation of an item for the entity. A framework for resolving the inactive problem was presented revealing how previously proposed trust computation in generating activeness matrix and predicting trustworthy behaviour (Chapter 3) are implemented in supporting the framework.

The framework also revealed that identifying entities with their group (cluster) and the exemplar of their group based on their trust attitude which can be observed from an activeness matrix (Demonstrated in Chapter 4) could support predicting their preferences. As exemplars identified from AP clustering algorithm might not necessarily be considered as influential members of clusters, a test with singular value decomposition (SVD) recommendation was carried out to verify if exemplars could be influential on its members. SVD algorithm was pointed out to be a suitable recommendation algorithm as it had been considered in previous researches to have the ability to reduce a high dimensional user-item data to a low dimensional space which makes recommendation easier and effective. A user-item matrix for simulating recommendation of items preferred by an exemplar of a cluster to its members was introduced with several examples demonstrating the use of clusters previously generated from different AP clustering algorithm based on the various similarity measures considered in chapter 4. The application of the SVD algorithm on a user-item matrix supports the prediction of preference ratings to recommended items which will indicate if members accept or reject the recommended items. It was proven with the clusters from correlation based AP algorithm that a most active node (MAN) considered as the most trustworthy member of each cluster will always motivate its members to accept or reject recommended items by the exemplar of their cluster. It is believed that every member in each cluster considers their behaviour to be useful in supporting others to make their decision. This obeys the theory of social cognitive [15] as the perception of certain members of a group motivate the group's less active or inactive members to have the same perception.

Chapter 6

Final Conclusion

6.1 Introduction

This chapter discusses all findings throughout the duration of the research. The main work of the research was originated from the idea that inactive members of a social group will always affect the accuracy of the recommendation of items as there will be conflicts on item preferences between them and their neighbours of a social group. Exploring social influence revealed that this conflict could be resolved to motivate the inactive members to change their behaviour to an active behaviour portrayed by other members of their social group. Previous research work related to relevant concepts which includes trust, social influence and clustering were all reviewed and combined to model how an inactive user could be motivated to act like other users similar to them in improving the prediction of their preferences to items.

Section 6.2 of the chapter will present a summary of the research work ; the main result and contribution to the research area will be presented in section 6.3 to reveal all the objectives that have been achieved; section 6.4 will discuss the problems experienced and issues not yet resolved from the research work that could be considered for future works.

6.2 Research Summary

As the prediction of preferences can be used in supporting the recommendation of new items to entities, there is the need to understand the social relationship (i.e. structural properties via patterns or habits) between entities via a network analysis concept to assist them in their decision making. Current social systems where users interact with themselves on items still experience inaccurate recommendation from the inaccurate evaluation of preferences for less active or inactive users (Section 1.5). The inaccuracy is due to the fact that the less active or in-active users might disagree on general perceptions or opinions of the group they belong to. So it cannot be completely assumed that a user identified with a group will always accept the same recommended items only estimated to be preferred by similar users as observed in current system such as the Collaborative recommender system (Section 2.2.2).

The proposed framework in the research to resolve the in-activeness problem was decided based on the real-life situation that people will decide to be active only when their trustworthy neighbours motivate or influence them. The framework was initiated with the theory of interpersonal behaviour [13, 165, 203] where the attitude and opinion of all entities are observed. This theory was important as observed patterns will be able to reveal communities or groups, the stability of communities or groups, vulnerability (exposure to deception) of members in communities and the important or influential node of each community. All these could be achieved only if the trusting behaviour amongst the entities is estimated and understood.

In the proposed framework, a learning mechanism was required to understand social actors' attitude toward items in the past and their potential ties with other social actors. Preferences from implicit feedbacks (e.g. actions towards items) could be observed from the interpersonal behaviour of entities where an entity's level of frequent engagement with items that are shared with other entities determines the activeness degree of the entity (section 3.3.2). Each entity's activeness degree (i.e. active condition) in relation with other entities were represented in a matrix (i.e. activeness matrix figure B.2) to reveal all predicted engagement between a pair of entities. But this degree on its own cannot be used to indicate the trustworthiness of an entity as other factors such as similarity in interaction pattern can also be applied along with the active condition from the activeness matrix in predicting the trustworthiness of the entity.

According to Triandis [13, 203], the performance of a behaviour (i.e. trusting behaviour in the case of this research) can only be affected by both pattern (i.e. frequent interaction with others in a social network) and situational conditions (i.e. degree of activeness or activeness trust in this case controls the future behaviour of entities). The activeness trust values of a target entity towards other entity could be compared with that of another entity to determine the similarity in their pattern. Comparing activeness trust values reveals how the target entity is appreciated by the other entity based on the similarity in their interaction which is best measured with the cosine similarity measure (i.e. Familiarity based trust in section 3.3.3). Therefore, taking the mean of all familiarity based trust for the target entity with each entity provides the trustworthiness of the target entity.

To model influence from trust, the theory of social cognitive [15] was considered based on the idea that the acceptance of an item depends on the knowledge (i.e. consequences) obtained from the observation of past experience with the item by a group of individuals. Every individual (including less or inactive active members) in a group have the belief that their behaviour in the group might be useful in supporting each other's behaviour; Bandura [15] referred to this as *self-efficacy*. In other words, a group of member's behaviour will influence each other based on their social or structural group determined by variation in predicted behaviour.

As predicted behaviour can be observed in an activeness matrix, several clustering techniques were used to check for the best set of social groups (clusters) generated to reveal that similar pattern in the predicted behaviour exists amongst members within a certain cluster. Empirical test carried out showed that Correlation based AP clustering algorithm was suitable to yield this output where structural equivalence amongst members can be better captured. It was revealed that identified potential influential members (i.e. either the central member, exemplar or attractor) from the clusters might not necessarily be influential members as they are not considered as the most trustworthy members to their neighbours.

Results from a further empirical test using Singular value decomposition (SVD) proved that clusters generated with Correlation based AP clustering had their most trustworthy members behaving as influential members to less active members and other members of their cluster. The results from the test showed that even though an exemplar's preferred items were recommended to members of the exemplar's cluster, members will only accept or like those items if the influential member (trustworthy or most active node) also accepts the recommended items. The prediction of acceptance to a recommended item was determined with the function (See equation 2.16) introduced by Jannich et.al [98] who used decomposed products (i.e. matrices with hidden information or features of vectors) from a matrix to formulate preference prediction to an item.

6.3 Contributions of the thesis

The main results obtained and contributions made from the research can be summarised as follows:

1. The proposed approach in computing trustworthiness of an entity is a novel approach proven to be useful as it was shown in section 5.2 to support the identification of an influential member from generated clusters or groups. It was revealed that the computation requires predicted information on how active entities will behave in the future amongst themselves (i.e. predicted behaviour indicated by activeness degree or trust). This information, represented in a matrix known as activeness matrix, can be observed by considering the similarities in predicted behaviour between a pair of entities to determine the trustworthiness of a specific entity.

2. As activeness matrix was shown to be a relevant data in determining how entities will trust themselves in the future based on their similarities in predicted behaviour with others (See 3.3.3), it was also considered to be a relevant source of data in supporting and explaining how members are clustered together (See section 4.3). The degree of how an entity appreciates another entity can only be viewed from a normalized activeness matrix B.2 and so it was more appropriate to cluster the normalized activeness.

3. Affinity propagation (AP) clustering algorithm [67] was proven empirically to be a suitable clustering algorithm as potential influential members of their individual clusters were identified as exemplars. However, a further empirical test revealed that the default similarity measure, negative Euclidean distance [57, 67] seemed not to be compatible with the algorithm in the case of social context. It was revealed that Pearson's correlation measure seems more suited for the algorithm as the members in the generated clusters were structural equivalent. The correlation-based AP clustering algorithm was also able to identify influential member for

each generated cluster. The influential member corresponds to a node considered to be the most trustworthy member amongst the members of an individual cluster.

4. The use of singular value decomposition (SVD) recommendation algorithm was able to further prove that members of clusters generated with the correlation-based AP clustering algorithm can be motivated by their most active member or most trustworthy member. In the research work, a novel recommendation framework where a matrix can be used to simulate the recommendation of items preferred by a cluster's exemplar was introduced and used as a required input data for the SVD algorithm. From an empirical test, it was revealed that less active or inactive members of clusters will always emulate the most active node of their individual cluster by accepting or rejecting items recommended to them.

A framework implementing the use of information from both activeness matrix and computed trustworthiness of entities (e.g. users or humans) could enhance current recommender systems that exist in media such as Netflix, Facebook and Amazon where less or inactive users are sometimes not presented with recommended items due to their insufficient activities on items or feedbacks on items that have been previously acted upon (i.e. purchase or converse). The current systems will be more effective if less active or inactive users could be motivated by their most trustworthy neighbours to either accept or reject recommended items.

6.4 Limitations and Future Work

As the learning mechanism within the trust computation from the proposed framework uses static social data in understanding the attitude of social actors, it might also be a good idea in future work to consider a longitudinal social data [1] [108, 153, 159] with the learning mechanism as time could also be used as one of the dimensions to determine the degree of activeness (i.e. activeness trust) for an entity. It could be used to enhance the proposed framework with adaptive features which could assist in keeping track of behavioural pattern for each social actor. This was not considered during this research as there was not sufficient time to collect and analyse this type of data; more attention was given to the analysis of static social data during this research. Analysing a longitudinal

[1] Longitudinal social data are data where samples can be repeatedly observed at different points in time. Time is considered as a factor in an analysis of the data to predict behaviours as the data will also include information on the time of interaction between entities.

data might support the mechanism to keep track of when entities started interacting on a specific item and how often the entities interact on the item for easy analysis of a network tie growth. It is important to know how long a tie between a pair of entities on an item will last to predict their trusting behaviour amongst themselves. An example of a longitudinal data or network that is still being studied by several researchers [12, 17, 208] is *network of airports* with scheduled flights where the tie between two airports will strengthen if there is an increase (per period e.g. daily or hourly) in volume of a flight route between the airports which may indicate that there is a predicted trust in the delivery of flight services between operators of both airports.

The novel Influence based framework for recommendation introduced in this thesis could be enhanced with the 'importance' of an item which can be modelled to support analysis in determining possible recommended items that non-active users might accept. Non-active members observed as singletons in the clustering example (See appendix B.9) from Facebook-like forum dataset revealed their behaviour of preferring or accepting previously engaged items instead of accepting new items recommended to them. Even though a better output with non-active members being clustered with and influenced by the MAN amongst social actors (i.e. a producer identified from TableB.3) was observed (See Appendix B.4.2) from another dataset (i.e. Hollywood film-music dataset in Appendix B.4), there is still the need to know how important the recommended items will be to the members. With the 'importance' of all items modelled, there is the possibility these non-active members might trust and accept certain new items based on their importance. This enhancement could be employed in resolving global conflict on certain issues (e.g. pollution, disarmament, overpopulation, peace and security) deliberated by a global community (such as United nation - UN and North Atlantic Treaty organization - NATO). The importance of items here could also be used in determining the trustworthiness of members in the global community where there could be some non-active or new members that might need support in their decision making on deliberated issues.

Unlike the novel consolidation method for predicting preferences in this thesis where aggregated (i.e. Mean) similarities of predicted ties between a target entity and other entities are estimated as the trustworthiness of the target entity, other researchers [43, 44, 102] have considered using OWA or IOWA operators in consolidating preferences of individuals to support the decision making of inactive or less active users. It might be a good idea to also implement these type of operators to identify the most influential member amongst all possible members since the trustworthiness of a member could be considered uncertain as the preference information of other individuals (i.e. in-active or less active users) are uncertain. With the OWA or IOWA operator, it might be possible to aggregate all preferences or opinions (activeness trust values) into a collective means for a decision to be made based on every individual preference or opinion [43, 102].

Hard clustering algorithms such as Highly connected subgroup, Markov clustering and Affinity propagation (AP) clustering, which all generate distinct clusters were considered within the proposed framework of this thesis for partitioning social actors based on their trust amongst themselves. But as trust attribute is ambiguous and difficult to define, fuzzy clustering [22, 58] might be a more suitable clustering algorithm for partitioning social actors; each social actor has the potential of belonging to more than one cluster. With fuzzy clustering, each data point of a social actor is expected to be assigned with a membership degree to each cluster they belong to. A clearer insight of social actors being identified based on their behavioural pattern could be observed with fuzzy clustering. From analysed patterns of a social actor, it is possible that the social actor trusts another social actor based on their several relations towards similar items in the past. These could be comprehended with fuzzy clustering which reveals how a social actor has a membership degree to various clusters based on several behavioural patterns with other social actors.

Bibliography

[1] ADOMAVICIUS, G., AND TUZHILIN, A. *Context-aware recommender systems*. Recommender systems handbook. Springer, 2011, pp. 217–253.

[2] AGRAWAL, R., IMIELIŃSKI, T., AND SWAMI, A. Mining association rules between sets of items in large databases. In *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data* (1993), SIGMOD '93, ACM, pp. 207–216.

[3] AGUZZOLI, S., AVESANI, P., AND MASSA, P. *Collaborative Case-Based Recommender Systems*, vol. 2416 of *Advances in Case-Based Reasoning: 6th European Conference, ECCBR 2002*. Springer Berlin Heidelberg, 2002, pp. 460–474.

[4] AJZEN, I. The theory of planned behavior. *Organizational behavior and human decision processes 50*, 2 (1991), 179–211.

[5] AKAMA, H., MIYAKE, M., AND JUNG, J. How to take advantage of the limitations with markov clustering?-the foundations of branching markov clustering (bmcl). In *IJCNLP* (2008), pp. 901–906.

[6] ALMACK, J. C. The influence of intelligence on the selection of associates. *School and Society* (1922), 529–530.

[7] AMATRIAIN, X., JAIMES, A., OLIVER, N., AND PUJOL, J. M. *Data mining methods for recommender systems*. Recommender Systems Handbook. Springer, 2011, pp. 39–71.

[8] ANTON, H., AND RORRES, C. *Elementary Linear Algebra: Applications Version*, 11th ed. Wiley Global Education, 2013.

[9] APPEL, O., CHICLANA, F., CARTER, J., AND FUJITA, H. A consensus approach to the sentiment analysis problem driven by support-based iowa majority. *International Journal of Intelligent Systems 32* (2017), 947–965.

[10] ARAL, S., AND ALSTYNE, M. V. Network structure & information advantage. In *Proceedings of the Academy of Management Conference, Philadelphia, PA* (2007), vol. 3, Citeseer.

[11] AVERY, C., AND ZECKHAUSER, R. Recommender systems for evaluating computer messages. In *Communication of The ACM*. ACM, 1997.

[12] AZZAM, M. Evolution of airports from a network perspective – an analytical concept. *Chinese Journal of Aeronautics 30*, 2 (2017), 513 – 522.

[13] BAMBERG, S., AND SCHMIDT, P. Incentives, morality, or habit? predicting students' car use for university routes with the models of ajzen, schwartz, and triandis. *Environment and Behavior 35*, 2 (2003), 264–285.

[14] BANDURA, A. *Social foundations of thought and action: a social cognitive theory*. Prentice-Hall, London, 1986.

[15] BANDURA, A. Social cognitive theory: An agentic perspective. *Annual Review of Psychology 52*, 1 (2001), 1.

[16] BARBER, B. *The Logic and Limits of Trust*. Rutgers University Press, New Jersey, 1983.

[17] BARRAT, A., BARTHELEMY, M., PASTOR-SATORRAS, R., AND VESPIGNANI, A. The architecture of complex weighted networks. *Proceedings of the National Academy of Sciences of the United States of America* (2004), 3747–3752.

[18] BATAGELJ, V., AND MRVAR, A. Pajek datasets. [Online], 2006. Available at http://vlado.fmf.uni-lj.si/pub/networks/data/.

[19] BATESON, P. The biological evolution of cooperation and trust. In *Trust: Making and breaking cooperative relations*, D. Gambetta, Ed. University of Oxford, Department of Sociology, University of Oxford, 1990, ch. 2, pp. 14–30.

[20] BATTISTON, S., GLATTFELDER, J. B., GARLASCHELLI, D., LILLO, F., AND CALDARELLI, G. *The structure of financial networks*. Network Science. Springer, 2010, ch. 7, pp. 131–163.

[21] BERNARD, H. R., KILLWORTH, P., KRONENFELD, D., AND SAILER, L. The problem of informant accuracy: The validity of retrospective data. *Annual Review of Anthropology 13* (10 1984), 495–517.

[22] BEZDEK, J. C. *Pattern Recognition with Fuzzy Objective Function Algorithms*. Plenum Press, 1981.

[23] BEZDEK, J. C., AND PAL, S. K. *Fuzzy models for pattern recognition: Methods that Search for Structures in Data*. IEEE, 1992.

[24] BODENHOFER, U., KOTHMEIER, A., AND HOCHREITER, S. Apcluster: an r package for affinity propagation clustering. *Bioinformatics (Oxford, England) 27*, 17 (Sep 1 2011), 2463–2464.

[25] BONHARD, P., AND SASSE, M. A. "i thought it was terrible and everyone else loved it"—a new perspective for effective recommender system design. In *People and Computers XIX—The Bigger Picture*. Springer, 2006, pp. 251–265.

[26] BORGATTI, S. P. Centrality and aids. *Connections* (1995), 112–114.

[27] BORGATTI, S. P. Centrality and network flow. *Social networks 27*, 1 (2005), 55–71.

[28] BORGATTI, S. P., AND EVERETT, M. G. Network analysis of 2-mode data. *Social Networks 19*, 3 (1997), 243–269.

[29] BRADLEY, P. S., FAYYAD, U., AND REINA, C. Scaling em (expectation-maximization) clustering to large databases. Tech. rep., Microsoft Research, Redmond, 1998.

[30] BRASHEARS, M. E. A longitudinal analysis of gendered association patterns: Homophily and social distance in the general social survey. *Journal of Social Structure 16* (2015), 1.

[31] BRATMAN, M. *Intention, plans and practical reason*. Havard University Press, 1987. Copyright, 1999.

[32] BREESE, J. S., HECKERMAN, D., AND KADIE, C. Empirical analysis of predictive algorithms for collaborative filtering. In *Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence* (1998), Morgan Kaufmann Publishers Inc., pp. 43–52.

[33] BRIDGE, D., GÖKER, MEHMET, H., MCGINTY, L., AND SMYTH, B. Case-based recommender systems. *The Knowledge Engineering Review 20*, 3 (09 2005), 315–320.

[34] BURGER, J. M., SANCHEZ, J., IMBERI, J. E., AND GRANDE, L. R. The norm of reciprocity as an internalized social norm: Returning favors even when no one finds out. *Social Influence 4*, 1 (2009), 11–17.

[35] BURKE, R. Hybrid recommender systems: Survey and experiments. *User Modeling and User-Adapted Interaction 12*, 4 (2002), 331.

[36] BURT, R. S. *Structural holes: The social structure of competition*. Harvard university press, 1995.

[37] BYRNE, D. Interpersonal attraction and attitude similarity. *The Journal of Abnormal and Social Psychology 62*, 3 (05 1961), 713–715.

[38] CALDARELLI, G. *Scale-free networks: Complex Webs in Nature and Technology*. Oxford University Press, 2007.

[39] CAREY, V., LONG, L., AND GENTLEMAN, R. *RBGL: An interface to the BOOST graph library*, 2017. R package version 1.54.0.

[40] CARL, N., AND BILLARI, F. C. Generalized trust and intelligence in the united states. *PloS one 9*, 3 (2014), e91786.

[41] CHARTRAND, G. *Introduction to graph theory*. Tata McGraw-Hill Education, 2006.

[42] CHEKURI, C. S., GOLDBERG, A. V., KARGER, D. R., LEVINE, M. S., AND STEIN, C. Experimental study of minimum cut algorithms. In *Proceedings of the Eighth Annual ACM-SIAM Symposium on Discrete Algorithms* (1997), SODA '97, Society for Industrial and Applied Mathematics, pp. 324–333.

[43] CHICLANA, F., HERRERA, F., AND HERRERA-VIEDMA, E. Integrating three representation models in fuzzy multipurpose decision making based on fuzzy preference relations. *Fuzzy Sets and Systems 97*, 1 (7/1 1998), 33–48.

[44] CHICLANA, F., HERRERA-VIEDMA, E., HERRERA, F., AND ALONSO, S. Some induced ordered weighted averaging operators and their use for solving group decision-making problems based on fuzzy preference relations. *European Journal of Operational Research 182*, 1 (10/1 2007), 383–399.

[45] CHOI, Y. K., AND KIM, S. K. An auxiliary recommendation system for repetitively purchasing items in e-commerce. In *2014 International Conference on Big Data and Smart Computing, BIGCOMP 2014* (2014), pp. 96–98.

[46] COHEN, W. W. Whirl: A word-based information representation language, April 2000 2000.

[47] COLEMAN, J., AND FARARO, T. *Rational Choice Theory*. Nueva York: Sage, 1992.

[48] DASGUPTA, P. Trust as a commodity. In *Trust: Making and breaking cooperative relations*, D. Gambetta, Ed. University of Oxford, Department of Sociology, University of Oxford, 1990, ch. 4, pp. 49–72.

[49] DEERWESTER, S., DUMAIS, S. T., FURNAS, G. W., LANDAUER, T. K., AND HARSHMAN, R. Indexing by latent semantic analysis. *Journal of the American Society for Information Science (1986-1998) 41*, 6 (1990), 391.

[50] DEMPSTER, A. P., LAIRD, N. M., AND RUBIN, D. B. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)* (1977), 1–38.

[51] DESHPANDE, M., AND KARYPIS, G. Item-based top-n recommendation algorithms. *ACM Trans. Inf. Syst. 22*, 1 (2004), 143–177.

[52] DESROSIERS, C., AND KARYPIS, G. *A comprehensive survey of neighborhood-based recommendation methods*. Recommender systems handbook. Springer, 2011, pp. 107–144.

[53] DEUTSCH, M. *The resolution of conflict: Constructive and destructive processes*. Yale University Press, 1973.

[54] DEUTSCH, M., AND GERARD, H. B. A study of normative and informational social influences upon individual judgment. *The Journal of Abnormal and Social Psychology 51*, 3 (11 1955), 629–636.

[55] DODGSON, M. Technological collaboration and innovation. *The handbook of industrial innovation* (1994), 285–292.

[56] DONGEN, S. M. V. *Graph Clustering by Flow Simulation*. PhD thesis, University of Utrecht, 2000.

[57] DUECK, D. *Affinity propagation: clustering data by passing messages*. PhD thesis, University of Toronto, 2009.

[58] DUNN, J. A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters. *Cybernatics and Systems 3* (1973).

[59] Ester, M., Kriegel, H.-P., Sander, J., and Xu, X. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining* (1996), AAAI Press, pp. 226–231.

[60] Everitt, B. S., Landau, S., Leese, M., and Stahl, D. *Cluster Analysis*, 5th ed. Wiley, 2011.

[61] Felfernig, A., Friedrich, G., Jannach, D., and Zanker, M. An integrated environment for the development of knowledge-based recommender applications. *International Journal of Electronic Commerce 11*, 2 (Winter2006 2006), 11–34.

[62] Felfernig, A., Isak, K., Szabo, K., and Zachar, P. The vita financial services sales support environment. In *PROCEEDINGS OF THE NATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE* (2007), vol. 22, Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, p. 1692.

[63] Festinger, L. A theory of social comparison processes. *Human relations 7*, 2 (1954), 117–140.

[64] Foltz, P. W. Using latent semantic indexing for information filtering. *SIGOIS Bull. 11* (1990), 40–47.

[65] Freeman, L. C. Centrality in social networks conceptual clarification. *Social networks 1*, 3 (1978), 215–239.

[66] Freeman, L. C. Some antecedents of social network analysis. *Connections 19*, 1 (1996), 39–42.

[67] Frey, B. J., and Dueck, D. Clustering by passing messages between data points. *Science 315* (2007), 972–976. Available from www.psi.toronto.edu/affinitypropagation.

[68] Frey, B. J., and Jojic, N. Learning the 'epitome' of an image. Tech. rep., University of Toronto, 2002.

[69] Friedkin, N. E. The development of structure in random networks: an analysis of the effects of increasing network density on five measures of structure. *Social Networks 3*, 1 (1981), 41–52.

[70] FRIEDKIN, N. E., AND JOHNSEN, E. C. Social influence networks and opinion change. In *Advances in Group Processes*, vol. 16. Emerald Group Publishing Limited, 1999.

[71] GAMBETTA, D. Can we trust trust? In *Trust: Making and breaking cooperative relations*, D. Gambetta, Ed. University of Oxford, Department of Sociology, University of Oxford, 1990, ch. 13, pp. 213–237.

[72] GARDNER, B., CORBRIDGE, S., AND MCGOWAN, L. Do habits always override intentions? pitting unhealthy snacking habits against snack-avoidance intentions. *BMC Psychology 3*, 1 (2015), 8.

[73] GENTLE, J. E. *Matrix transformations and factorizations*. Springer, 2007.

[74] GERCK, E. Towards real-world models of trust: Reliance on received information. [Online], (1998). Available from http://mcwg.org/mcg-mirror/trustdef.htm.

[75] GERGEN, K. J. Social psychology as history. *Journal of personality and social psychology 26*, 2 (05 1973), 309–320.

[76] GILBOA, I., SCHMEIDLER, D., AND WAKKER, P. P. Utility in case-based decision theory. *Journal of Economic Theory 105*, 2 (2002), 483–502.

[77] GIRVAN, M., AND NEWMAN, M. E. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences of the United States of America 99*, 12 (2002), 7821–7826.

[78] GOLBECK, J. A. *Computing and applying trust in web-based social networks*. PhD thesis, University of Maryland, 2005.

[79] GRANOVETTER, M. Network sampling: Some first steps. *American Journal of Sociology* (1976), 1287–1303.

[80] GRANOVETTER, M. S. The strength of weak ties. *American Journal of Sociology* (1973).

[81] GU, Q., SANTOS, E., AND SANTOS, E. E. Modeling opinion dynamics in a social network. In *Web Intelligence (WI) and Intelligent Agent Technologies (IAT), 2013 IEEE/WIC/ACM International Joint Conferences on* (2013), vol. 2, IEEE, pp. 9–16.

[82] HAMMING, R. W. Error detecting and error correcting codes. *Bell Labs Technical Journal 29*, 2 (1950), 147–160.

[83] HAMPTON, M. Svd computation example. Online, n.d. Available from http://www.d.umn.edu/~mhampton/m4326svd_example.pdf [Accessed: 14/05/2017].

[84] HANNEMAN, R., AND RIDDLE, M. Measures of similarity and structural equivalence. In *Introduction to Social Network Methods*. University of California, Riverside, 2005, ch. 13.

[85] HARDADON, A., AND SUTTON, R. I. Technology brokering and innovation in a product development firm. *Administrative Science Quarterly 42*, 4 (.12.1997 01), 716; 716. doi: pmid:.

[86] HARTUV, E., AND SHAMIR, R. A clustering algorithm based on graph connectivity. *Information Processing Letters 76*, 4 (2000), 175 – 181.

[87] HARWOOD, W. T. *The Logic of Trust*. PhD thesis, University of York, 2012.

[88] HAUSMANN, R., AND HIDALGO, C. A. The network structure of economic output. *Journal of Economic Growth 16*, 4 (2011), 309–342.

[89] HAYES, S. C. *Get out of your mind and into your life: The new acceptance and commitment therapy*. New Harbinger Publications, 2005.

[90] HERLOCKER, J. L., KONSTAN, J. A., AND RIEDL, J. Explaining collaborative filtering recommendations. In *Proceedings of the 2000 ACM conference on Computer supported cooperative work* (2000), ACM, pp. 241–250.

[91] HERLOCKER, J. L., KONSTAN, J. A., TERVEEN, L. G., AND RIEDL, J. T. Evaluating collaborative filtering recommender systems. *ACM Trans. Inf. Syst. 22*, 1 (2004), 5–53.

[92] HERLOCKER, J. L., KONSTAN, J. A., TERVEEN, L. G., AND RIEDL, J. T. Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems 22*, 1 (2004), 5–53.

[93] HEVEY, D., COLLINS, A., AND BROGAN, A. Methods: Network analysis. *The Physchologist 26*, 6 (June 2013), 430–431. Available from https://thepsychologist.bps.org.uk/.

[94] HEVNER, A., AND CHATTERJEE, S. Design science research in information systems. In *Design Research in Information Systems*. Springer, 2010, ch. 2, pp. 9–22.

[95] HILDRETH, P. M., AND KIMBLE, C. *Knowledge networks: Innovation through communities of practice*. Igi Global, 2004.

[96] HORN, R. A., AND JOHNSON, C. R. *Topics in Matrix Analysis*. Cambridge University Press, 1994.

[97] HOUSER, D., AND WOODERS, J. Reputation in auctions: Theory, and evidence from ebay. *Journal of economics & management strategy 15*, 2 (2006), 353–369.

[98] JANNACH, D., ZANKER, M., FELFERNIG, A., AND FRIEDRICH, G. *Recommendation Systems: An Introduction*. Cambridge University Press, 2010.

[99] JENNESS, A. The role of discussion in changing opinion regarding a matter of fact. *The Journal of Abnormal and Social Psychology 27*, 3 (10 1932), 279–296.

[100] JÄGER, M. L. *MCL: Markov Cluster Algorithm*, 2015. R package version 1.0.

[101] JOSANG, A., AND ISMAIL, R. The beta reputation system. In *Proceedings of the 15th Bled Electronic Commerce Conference* (2002), Citeseer, pp. 41–55.

[102] KAMIS, N. H., CHICLANA, F., AND LEVESLEY, J. Preference similarity network structural equivalence clustering based consensus group decision making model. *Applied Soft Computing* (2017). Available from http://www.sciencedirect.com/science/article/pii/S1568494617306853.

[103] KAMVAR, S. D., SCHLOSSER, M. T., AND GARCIA-MOLINA, H. The eigentrust algorithm for reputation management in p2p networks. In *Proceedings of the 12th International Conference on World Wide Web* (2003), ACM, pp. 640–651.

[104] KANUNGO, T., MOUNT, D. M., NETANYAHU, N. S., PIATKO, C. D., SILVERMAN, R., AND WU, A. Y. An efficient k-means clustering algorithm: analysis and implementation. *IEEE Transactions on Pattern Analysis and Machine Intelligence 24*, 7 (2002), 881–892.

[105] KASHDAN, T. B., AND ROTTENBERG, J. Psychological flexibility as a fundamental aspect of health. *Clinical psychology review 30*, 7 (11 2010), 865–878.

[106] KELLY, D., AND TEEVAN, J. Implicit feedback for inferring user preference: A bibliography. *SIGIR Forum 37*, 02 (2003), 18–28.

[107] KELMAN, H. C. Compliance, identification, and internalization three processes of attitude change. *The Journal of Conflict Resolution 2*, 1 (1958), 51–60.

[108] KOSSINETS, G., AND WATTS, D. J. Empirical analysis of an evolving social network. *Science 311*, 5757 (January 06 2006), 88–90.

[109] LEE, T. Q., PARK, Y., AND PARK, Y.-T. A time-based approach to effective recommender systems using implicit feedback. *Expert Systems with Applications 34*, 4 (2008), 3055–3062.

[110] LEVER, J. J., NES, E. H., SCHEFFER, M., AND BASCOMPTE, J. The sudden collapse of pollinator communities. *Ecology Letters 17*, 3 (03 2014), 350–359.

[111] LEVINSON, D. Network structure and city size. *PloS one 7*, 1 (2012), e29721.

[112] LI, L., SCAGLIONE, A., SWAMI, A., AND ZHAO, Q. Consensus, polarization and clustering of opinions in social networks. *IEEE Journal on Selected Areas in Communications 31*, 6 (2013), 1072–1083.

[113] LI, R., KIM, J., AND PARK, J. The effects of internet shoppers' trust on their purchasing intention in china. *JISTEM-Journal of Information Systems and Technology Management 4*, 3 (2007), 269–286.

[114] LINDEN, G., SMITH, B., AND YORK, J. Amazon.com recommendations: Item-to-item collaborative filtering. *IEEE Internet Computing 7*, 1 (2003), 76–80.

[115] LOPS, P., GEMMIS, M. D., AND SEMERARO, G. *Content-based recommender systems: State of the art and trends*. Recommender systems handbook. Springer, 2011, pp. 73–105.

[116] LUHMANN, N. Familiarity, confidence, trust: Problems and alternatives. In *Trust: Making and breaking cooperative relations*, D. Gambetta, Ed. University of Oxford, Department of Sociology, 1990, pp. 94–107.

[117] LUINI, L., NESE, A., AND SBRIGLIA, P. Social influence in trustors' neighbourhoods. *Journal of Behavioral and Experimental Economics 53* (2014), 97–110.

[118] MACKAY, D. J. *An Example Inference Task: Clustering*. Information theory, inference and learning algorithms. Cambridge University Press, 2003, ch. 20, pp. 284–292.

[119] MACKAY, D. J. *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, 2005, 2005.

[120] MALLE, B. F. *How the Mind Explains Behaviour: Folk explanation, meaning and Social Interaction*. The MIT Press, 2004.

[121] MANNING, C. D., RAGHAVAN, P., AND SCHÜTZE, H. *Hierarchical Clustering*, vol. 1 of *Introduction to information retrieval*. Cambridge university press, 2008, ch. 17.

[122] MARSDEN, P. V. Homogeneity in confiding relations. *Social networks 10*, 1 (1988), 57–76.

[123] MARSH, S. P. *Formalising Trust as a Computational Concept*. PhD thesis, University of Stirling, 1994.

[124] MASI, G. D., IORI, G., AND CALDARELLI, G. Fitness model for the italian interbank money market. *Physical Review E 74*, 6 (2006), 066112.

[125] MASSA, P., AND AVESANI, P. Trust-aware recommender systems. In *Proceedings of the 2007 ACM Conference on Recommender Systems* (2007), RecSys '07, ACM, pp. 17–24.

[126] MAXWELL CHICKERING, D., AND HECKERMAN, D. Efficient approximations for the marginal likelihood of bayesian networks with hidden variables. *Machine Learning 29*, 2 (Nov 1997), 181–212.

[127] MCLEOD, S. Asch experiment. [Online], 2008. Available from `http://www.simplypsychology.org/asch-conformity.html` [Accessed 05/04/2016].

[128] MCLEOD, S. Social roles. (Online), 2008. Available from `http://www.simplypsychology.org/social-roles.html`.

[129] MCLEOD, S. Stanford prison experiment. [Online], 2008. Available from `http://www.simplypsychology.org/zimbardo.html` [Accessed 05/10/2016].

[130] MCLEOD, S. What is conformity? [Online], 2008. Available from `http://www.simplypsychology.org/conformity.html` [Accessed 05/04/2016].

[131] McPherson, M., Smith-Lovin, L., and Brashears, M. E. Social isolation in america: Changes in core discussion networks over two decades. *American Sociological Review 71*, 3 (2006), 353–375.

[132] McPherson, M., Smith-Lovin, L., and Cook, J. M. Birds of a feather: Homophily in social networks. *Annual review of sociology* (2001), 415–444.

[133] McSherry, D. Recommendation engineering. In *Proceedings of the 15th European Conference on Artificial Intelligence* (2002), F. Van Harmelen, Ed., IOS Press, pp. 86–90.

[134] Mendelson, B. *Introduction to Topology*, third ed. Courier Corporation, 1990.

[135] Miyahara, K., and Pazzani, M. J. *Collaborative Filtering with the Simple Bayesian Classifier*. Springer Berlin Heidelberg, Berlin, Heidelberg, 2000, pp. 679–689.

[136] Mladenic, D. Text-learning and related intelligent agents: a survey. *IEEE Intelligent Systems and their Applications 14*, 4 (1999), 44–54.

[137] Mobasher, B., Dai, H., Luo, T., and Nakagawa, M. Discovery and evaluation of aggregate usage profiles for web personalization. *Data Mining and Knowledge Discovery 6*, 1 (2002), 61–82.

[138] Montoya, J. M., Pimm, S. L., and Solé, R. V. Ecological networks and their fragility. *Nature 442*, 7100 (07/20 2006), 259–264.

[139] Moody, J., and White, D. R. Structural cohesion and embeddedness: A hierarchical concept of social groups. *American Sociological Review* (2003), 103–127.

[140] Moreno, J. *Who shall survive?Foundations of sociometry, group psychotherapy and socio-drama*, 2nd ed. Beacon House Inc, 1953.

[141] Morris, Q., Frey, B. J., and Paige, C. J. Denoising and untangling graphs using degree priors. In *NIPS* (2003), pp. 385–392.

[142] Mui, L., Mohtashemi, M., and Halberstadt, A. A computational model of trust and reputation. In *System Sciences, 2002. HICSS. Proceedings of the 35th Annual Hawaii International Conference on* (2002), pp. 2431–2439. ID: 1.

[143] Murphy, K. P. *Machine Learning : A Probabilistic Perspective*. MIT Press, Cambridge, 2014.

[144] NEPAL, S., BISTA, S. K., AND PARIS, C. Behaviour based propagation of trust in social networks with restricted and anonymous participation. *Computational Intelligence 31*, 04 (2015), 642–668.

[145] NEPAL, S., SHERCHAN, W., AND PARIS, C. Strust: A trust model for social networks. In *Proceedings of the 10th International Conference on Trust, Security and Privacy in Computing and Communications* (2011), IEEE, pp. 841–846.

[146] NEWMAN, M. *Networks: An Introduction*. Oxford University Press, 2010.

[147] NEWMAN, M. E. Scientific collaboration networks. ii. shortest paths, weighted networks, and centrality. *Physical review E 64*, 1 (2001), 016132.

[148] NORRIS, J. R. *Continuous-time Markov chains I*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1997, p. 60–107.

[149] OARD, D. W., AND KIM, J. Implicit feedback for recommender systems. In *Proceedings of the AAAI workshop on recommender systems* (1998), pp. 81–83.

[150] OLSON, J. S., GRUDIN, J., AND HORVITZ, E. A study of preferences for sharing and privacy. In *CHI '05 Extended Abstracts on Human Factors in Computing Systems* (2005), CHI EA '05, ACM, pp. 1985–1988.

[151] OPSAHL, T. Triadic closure in two-mode networks: Redefining the global and local clustering coefficients. *Social Networks 35*, 2 (5 2013), 159–167.

[152] OPSAHL, T. Datasets. [Online], n.d. Available at http://toreopsahl.com/datasets/.

[153] OPSAHL, T., AND HOGAN, B. Growth mechanisms in continuously-observed networks: Communication in a facebook-like community. *CoRR abs/1010.2141* (2010). Available from http://arxiv.org/abs/1010.2141.

[154] OPSAHL, T., AND PANZARASA, P. Clustering in weighted networks. *Elsevier Journal on Social Networks 31* (2009), 155–163.

[155] OSHODIN, E., CHICLANA, F., AND AHMADI, S. *Social Trust in a Familiar Community*. Research and Development in Intelligent Systems XXXII. Springer, 2015, pp. 119–132.

[156] Padrón, B., Nogales, M., and Traveset, A. Alternative approaches of transforming bimodal into unimodal mutualistic networks. the usefulness of preserving weighted information. *Basic and Applied Ecology 12*, 8 (2011), 713 – 721.

[157] Palmisano, C., Tuzhilin, A., and Gorgoglione, M. Using context to improve predictive modeling of customers in personalization applications. *IEEE Transactions on Knowledge and Data Engineering 20*, 11 (2008), 1535–1549.

[158] Pan, X., Papailiopoulos, D., Recht, B., Ramchandran, K., and Jordan, M. I. Scaling up correlation clustering through parallelism and concurrency control. In *DISCML workshop at NIPS* (2014).

[159] Panzarasa, P., Opsahl, T., and Carley, K. M. Patterns and dynamics of users' behaviour and interaction: Network analysis of an online community. *Journal of the American Society for Information Science and Technology 60*, 5 (05 2009), 911–932. M3: Article.

[160] Parsons, J., and Ralph, P. Generating effective recommendations using viewing-time weighted preferences for attributes. *Journal of the Association for Information Systems 15*, 8 (08 2014), 484–513.

[161] Parsons, J., Ralph, P., and Gallagher, K. Using viewing time to infer user preference in recommender systems. In *AAAI Workshop - Technical Report* (2004), vol. 09, pp. 52–63.

[162] Pasi, G. Implicit feedback through user-system interactions for defining user models in personalized search. *Procedia Computer Science 39* (2014), 8–11.

[163] Pattillo, J., Youssef, N., and Butenko, S. *Clique relaxation models in social network analysis*. Handbook of Optimization in Complex Networks. Springer, 2012, pp. 143–162.

[164] Pazzani, M. J., and Billsus, D. *Content-based recommendation systems*. The adaptive web. Springer, 2007, pp. 325–341.

[165] Pee, L. G., Woon, I. M. Y., and Kankanhalli, A. Explaining non-work-related computing in the workplace: A comparison of alternative models. *Information Management 45*, 2 (2008), 120–130.

[166] PÉREZ, L. G., MATA, F., AND CHICLANA, F. Social network decision making with linguistic trustworthiness–based induced owa operators. *International Journal of Intelligent Systems 29*, 12 (2014), 1117–1137.

[167] PERONY, N., PFITZNER, R., SCHOLTES, I., TESSONE, C. J., AND SCHWEITZER, F. Enhancing consensus under opinion bias by means of hierarchical decision making. *Advances in Complex Systems 16*, 06 (2013), 1350020.

[168] PFEFFER, J. Organizational demography: Implications for management. *California management review 28*, 1 (1985), 67–81.

[169] PLANLOCAL. Getting people involved. In *PlanLoCaL energy Efficiency and community Green deal*. PlanLocaL, 2013, ch. 6. Available from http://www.planlocal.org.uk.

[170] R CORE TEAM. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2015. Available at https://www.R-project.org/.

[171] RAWASHDEH, A., RAWASHDEH, M., DÍAZ, I., AND RALESCU, A. Measures of semantic similarity of nodes in a social network. In *International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems* (2014), Springer, pp. 76–85.

[172] REAGANS, R., AND ZUCKERMAN, E. W. Networks, diversity, and productivity: The social capital of corporate r&d teams. *Organization Science 12*, 4 (2001), 502–517.

[173] REGAN, D. T. Effects of a favor and liking on compliance. *Journal of experimental social psychology 7*, 6 (1971), 627–639.

[174] REMPEL, J. K., HOLMES, J. G., AND ZANNA, M. P. Trust in close relationships. *Journal of Personality and Social Psychology 49*, 1 (1985), 95–112.

[175] RESNICK, P., IACOVOU, N., SUCHAK, M., BERGSTROM, P., AND RIEDL, J. Grouplens: an open architecture for collaborative filtering of netnews. In *Proceedings of the 1994 ACM conference on Computer supported cooperative work* (1994), ACM, pp. 175–186.

[176] RESNICK, P., AND ZECKHAUSER, R. *Trust among strangers in internet transactions: Empirical analysis of eBay' s reputation system*. Elsevier, 2002, pp. 127–157.

[177] RICCI, F. Travel recommender systems. *IEEE Intelligent Systems 17*, 6 (2002), 55–57.

[178] RICCI, F., ROKACH, L., AND SHAPIRA, B. *Introduction to Recommender Systems Handbook*. Recommender systems handbook. Springer, 2011, ch. 1, pp. 1–29.

[179] RICH, E. User modeling via stereotypes, 1979.

[180] ROTTER, J. B. Generalized expectancies for interpersonal trust. *American psychologist 26*, 5 (1971), 443–452.

[181] SALTON, G., AND BUCKLEY, C. Term weighting approaches in automatic text retrieval. Technical report, Cornell University, 1987. Available from https://ecommons.cornell.edu/handle/1813/6721.

[182] SARWAR, B., KARYPIS, G., KONSTAN, J., AND RIEDL, J. Application of dimensionality reduction in recommender system-a case study. In *WebKDD Workshop at the ACM SIGKKD* (2000).

[183] SARWAR, B., KARYPIS, G., KONSTAN, J., AND RIEDL, J. Item-based collaborative filtering recommendation algorithms. In *Proceedings of the 10th International Conference on World Wide Web* (2001), ACM, pp. 285–295.

[184] SCHAEFFER, S. E. Graph clustering. *Computer Science Review 1*, 1 (2007), 27 – 64.

[185] SCHAFER, J. B., FRANKOWSKI, D., HERLOCKER, J., AND SEN, S. *Collaborative filtering recommender systems*. The adaptive web. Springer, 2007, pp. 291–324.

[186] SCHEIN, A. I., POPESCUL, A., UNGAR, L. H., AND PENNOCK, D. M. Methods and metrics for cold-start recommendations. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (2002), ACM, pp. 253–260.

[187] SCHMITZ, C., HOTHO, A., JÄSCHKE, R., AND STUMME, G. Mining association rules in folksonomies. In *Data science and classification*. Springer, 2006, pp. 261–270.

[188] SEGARRA, S., AND RIBEIRO, A. Hierarchical clustering and consensus in trust networks. In *2013 5th IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)* (2013), pp. 85–88.

[189] SHANI, G., AND GUNAWARDANA, A. *Evaluating Recommendation Systems*. Recommender systems handbook. Springer, 2011, ch. 8, pp. 257–297.

[190] SHAPIRO, C., AND VARIAN, H. R. *Information Rules: A Strategic Guide to network economy*. Havard Business School Press, 1999.

[191] SHEN, A. X., CHEUNG, C. M., LEE, M. K., AND CHEN, H. How social influence affects we-intention to use instant messaging: The moderating effect of usage experience. *Information Systems Frontiers 13*, 2 (Apr. 2011), 157–169.

[192] SHI, C., KONG, X., HUANG, Y., YU, P. S., AND WU, B. Hetesim: A general framework for relevance measure in heterogeneous networks. *IEEE Transactions on Knowledge and Data Engineering 26*, 10 (2014), 2479–2492.

[193] SINHA, R. R., AND SWEARINGEN, K. Comparing recommendations made by online systems and friends. *DELOS workshop: personalisation and recommender systems in digital libraries Vol.1* (2001).

[194] SMYTH, B., AND COTTER, P. Surfing the digital wave, generating personalised tv listings using collaborative, case-based recommendation. In *Case-Based Reasoning Research and Development, Proceedings of the Third International Conference on Case-Based Reasoning ICCBR99* (1999), B. K. Althoff K.D., Bergmann R., Ed., Springer Verlag, pp. 561–571.

[195] SMYTH, B., AND COTTER, P. A personalised tv listings service for the digital tv age. *Knowledge-Based Systems 13* (2000), 53–59.

[196] SMYTH, B., AND MCCLAVE, P. Similarity vs. diversity. In *International Conference on Case-Based Reasoning* (2001), Springer, pp. 347–361.

[197] SPIEKERMANN, S., GROSSKLAGS, J., AND BERENDT, B. E-privacy in 2nd generation e-commerce: Privacy preferences versus actual behavior. In *Proceedings of the 3rd ACM Conference on Electronic Commerce* (2001), EC '01, ACM, pp. 38–47.

[198] STOER, M., AND WAGNER, F. A simple min-cut algorithm. *JACM 44*, 4 (1997), 585–591.

[199] Sturgis, P., Patulny, R., Allum, N., and Buscha, F. *Social Connectedness and Generalized Trust: A Longitudinal Perspective*. Handbook of Research Methods and Application in Social Capital. Edward Elgar, 2015, ch. 5, pp. 76–90.

[200] Sun, Y., Mburu, L., and Wang, S. Analysis of community properties and node properties to understand the structure of the bus transport network. *Physica A: Statistical Mechanics and its Applications 450* (5/15 2016), 523–530.

[201] Swearingen, K., and Sinha, R. Beyond algorithms: An hci perspective on recommender systems. In *ACM SIGIR Workshop* (2001), Citeseer, pp. 1–11.

[202] Thavikulwat, P. Affinity propagation: A clustering algorithm for computer-assisted business simulations and experiential exercises. *Developments in Business Simulation and Experiential Learning 35* (2014).

[203] Triandis, H. C. *Interpersonal Behaviour*. Brooks/Cole Pub.Co., 1977.

[204] Triplett, N. The dynamogenic factors in pacemaking and competition. *The American Journal of Psychology 9*, 4 (1898), 507–533. Available from http://www.jstor.org/stable/1412188.

[205] Turiel, E. *The development of social knowledge: Morality and convention*. Cambridge University Press, 1983.

[206] TURING, A. M. I.computing machinery and intelligence. *Mind LIX*, 236 (October 01 1950), 433–460. Available from http://mind.oxfordjournals.org/content/LIX/236/433.

[207] Vedula, N., Parthasarathy, S., and Shalin, V. Predicting trust relations among users in a social network: The role of influence, cohesion and valence. In *KDD Wisdom 2016 Workshops* (2016), ACM.

[208] Verma, T., Araújo, N. A. M., and Herrmann, H. J. Revealing the structure of the world airline network. In *Scientific reports* (2014).

[209] Voinea, C. F. *Dynamic Social Impact Theory and Model*. John Wiley & Sons, Ltd, 2016, pp. 107–137.

[210] Wang, W., and Benbasat, I. Recommendation agents for electronic commerce: Effects of explanation facilities on trusting beliefs. *Journal of Management Information Systems 23*, 4 (2007), 217–246.

[211] Wasserman, S., and Faust, K. *Social Network Analysis: Methods and Applications*. Cambridge University Press, 1994.

[212] Watts, D. J., and Strogatz, S. H. Collective dynamics of 'small-world' networks. *Nature 393*, 6684 (1998), 440–442.

[213] Wellman, B. The school child's choice of companions. *The Journal of educational research (Washington, D.C.) 14*, 2 (1926), 126–126.

[214] Wilson, R. *Combinatorics: Ancient Modern*. OUP Oxford, 2013. Accessed from http://search.ebscohost.com.proxy.library.dmu.ac.uk/login.aspx?direct=true&AuthType=ip,shib&db=e000xww&AN=598871&site=ehost-live.

[215] Wu, J., and Chiclana, F. A social network analysis trust-consensus based approach to group decision-making problems with interval-valued fuzzy reciprocal preference relations. *Knowledge-Based Systems* (2014).

[216] Wu, X., and Zhu, X. Mining with noise knowledge: Error-aware data mining. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans 38*, 4 (2008), 917–932.

[217] Yager, R. R. On ordered weighted averaging aggregation operators in multicriteria decisionmaking. *IEEE transactions on systems, man, and cybernetics 18*, 1 (1988), 183–190.

[218] Yager, R. R., and Filev, D. Operations for granular computing: mixing words and numbers. In *1998 IEEE International Conference on Fuzzy Systems Proceedings. IEEE World Congress on Computational Intelligence (Cat. No.98CH36228)* (1998), vol. 1, pp. 123–128.

[219] Yin, B., Yang, Y., and Liu, W. Exploring social activeness and dynamic interest in community-based recommender system. In *Proceedings of the 23rd International Conference on World Wide Web* (2014), WWW '14 Companion, ACM, pp. 771–776.

[220] Yook, S.-H., Jeong, H., and Barabási, A.-L. Modeling the internet's large-scale topology. *Proceedings of the National Academy of Sciences 99*, 21 (2002), 13382–13386.

[221] Zhang, J. *Understanding and Augmenting Expertise Networks*. PhD thesis, University of Michigan, 2008.

[222] ZHANG, J., GHORBANI, A. A., AND COHEN, R. A familiarity-based trust model for effective selection of sellers in multiagent e-commerce systems. *International Journal of Information Security 6* (2007), 333–344.

[223] ZHU, X., AND WU, X. Class noise vs. attribute noise: A quantitative study. *Artificial Intelligence Review 22*, 3 (2004), 177–210.

[224] ZIEGLER, C.-N. *Towards Decentralized Recommender Systems*. PhD thesis, University of Freiburg, 2005.

Appendix A

Recommendation Example

A.1 Model based Collaborative Recommendation: Bayesian Classification approach

Example A.1. *For example, given a table of similar relation as the matrix in previous example but the numerical ratings are transformed to linguistics representations or terms which can be referred as classes. Given this classes as 'likes' and 'dislikes',* $0 \leq ratings < 3$ *could be classified as 'likes' while* $3 \leq ratings \leq 5$ *could be classified as 'dislikes'. Also, the last row contains some predicted ratings for some items except item* i_5 *and* i_3*. Will* i_5 *be predicted to have 'likes' or 'dislikes' ?*

	i_1	i_2	i_3	i_4	i_5	i_6	i_7
u_1	likes	dislikes	?	likes	likes	dislikes	dislikes
u_2	likes	?	likes	likes	dislikes	?	?
u_3	likes	?	dislikes	dislikes	likes	likes	likes
u_4	?	dislikes	?	likes	likes	dislikes	dislikes
u_5	?	?	?	dislikes	?	dislikes	?
Class label/prediction	likes	dislikes	?	likes	?	dislikes	dislikes

Assuming there is no prediction for target item i_5 *to either belong to the class of 'likes' or 'dislikes' based on group of users' rating features on other items* $r_i = (r_{i_1} = likes, r_{i_2} = dislikes, r_{i_3} = ?, r_{i_4} = likes, r_{i_6} = dislikes, r_{i_7} = dislikes)$.

The Probability of item i_5 will be liked or disliked regardless of the ratings features on other items:

$Pr(c^{i_5} = likes) = 3/4 = 0.75$

$Pr(c^{i_5} = dislikes) = 1/4 = 0.25$

conditional probabilities : $Pr(r_{i_1} = likes|c^{i_5} = likes) = 2/3 = 0.67$

$Pr(r_{i_1} = likes|c^{i_5} = dislikes) = 1/1 = 1$

$Pr(r_{i_2} = dislikes|c^{i_5} = likes) = 2/3 = 0.67$

$Pr(r_{i_2} = dislikes|c^{i_5} = dislikes) = 0/1 = 0$

$Pr(r_{i_3} =?|c^{i_5} = likes) = 2/3 = 0.67$

$Pr(r_{i_3} =?|c^{i_5} = dislikes) = 0/1 = 0$

$Pr(r_{i_4} = likes|c^{i_5} = likes) = 2/3 = 0.67$

$Pr(r_{i_4} = likes|c^{i_5} = dislikes) = 1$

$Pr(r_{i_6} = dislikes|c^{i_5} = likes) = 2/3 = 0.67$

$Pr(r_{i_6} = dislikes|c^{i_5} = dislikes) = 0$

$Pr(r_{i_7} = dislikes|c^{i_5} = likes) = 2/3 = 0.67$

$Pr(r_{i_7} = dislikes|c^{i_5} = dislikes) = 0$

Therefore, based on maximum posterior hypothesis [143], the probability that the group of users have all these rating features and they also like i_5 :

$Pr(r_i|c^{i_5} = likes) = 0.67 * 0.67 * 0.67 * 0.67 * 0.67 * 0.67 = 0.09$

while the probability of the user group who have these rating features also dislike i_5:

$Pr(r_i|c^{i_5} = dislikes) = 1 * 0 * 0 * 1 * 0 * 0 = 0$

To find the class that maximizes $Pr(r_i|c^{i_5}) * Pr(c^{i_5})$:

$Pr(r_i|c^{i_5} = likes) * Pr(c^{i_5} = likes) = 0.09 * 0.75 = 0.07$

$Pr(r_i|c^{i_5} = dislikes) * Pr(c^{i_5} = dislikes) = 0$

$argmax\{0.07, 0\} = 0.07$

A.2 Singular value Decomposition Example

Example A.2. *An example can be demonstrated with the given matrix M below which represent the user-item relationship where users have initially given their ratings to items.*

$$M = \begin{bmatrix} 2 & 1 & 2 \\ 3 & 2 & -2 \end{bmatrix}$$

From equation 2.15, one of the decomposed product from the original matrix M is matrix U which can be determined by initially by taking the transpose of matrix M.

$$M^T = \begin{bmatrix} 2 & 3 \\ 1 & 2 \\ 2 & -2 \end{bmatrix}$$

As the matrix U is expected to be an orthogonal matrix, we are required to find both eigenvalues and eigenvectors of $M \cdot M^T$. So $M \cdot M^T$ will be:

$$M \cdot M^T = \begin{bmatrix} 2 & 1 & 2 \\ 3 & 2 & -1 \end{bmatrix} \cdot \begin{bmatrix} 2 & 3 \\ 1 & 2 \\ 2 & -1 \end{bmatrix} = \begin{bmatrix} 9 & 6 \\ 6 & 14 \end{bmatrix}$$

Then, to find the eigenvector $\overrightarrow{v}$ of MM^T, $MM^T \times \overrightarrow{v}$ must be equivalent to the scalar multiple of $\overrightarrow{v}$. That is,

$$MM^T \times \overrightarrow{v} = \lambda \times \overrightarrow{v} \tag{A.1}$$

With this equation (equation A.1), the eigenvector v will be scaled by eigenvalue λ; in other words, the eigenvalue measures the magnitude change of the eigenvector with MM^T. Based on polynomial characteristics, the equation is still equivalent to :

$$\left| MM^T - \lambda I \right| = 0 \qquad \text{Where } I \text{ is the identity matrix of } MM^T. \tag{A.2}$$

Equation A.2 is a characteristic equation based on a characteristic polynomial of the matrix MM^T. The elements of the eigenvector can be determined by resolving the eigenvalues using equation A.2.

$$\begin{bmatrix} 9 & 6 \\ 6 & 14 \end{bmatrix} - \lambda . \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = 0$$

This can further be resolved to reveal the eigenvalues λ of MM^T:

$$\lambda^2 - 23\lambda + 90 = 0$$

$$(\lambda - 18)(\lambda - 5) = 0$$

$$\lambda = 18, \lambda = 5$$

Taking the square root of each eigenvalues will then give the singular values that will be diagonal elements in Σ.

$$\Sigma = \begin{bmatrix} \sqrt{18} & 0 & 0 \\ 0 & \sqrt{5} & 0 \end{bmatrix} = \begin{bmatrix} 4.2426 & 0 & 0 \\ 0 & 2.2361 & 0 \end{bmatrix}$$

Substituting eigenvalues λ of MM^T in $MM^T - \lambda I$:

For $\lambda = 18$,

$$MM^T - 18I = \begin{bmatrix} -9 & 6 \\ 6 & -4 \end{bmatrix}$$

To find the reduced row echelon form of the resulting matrix, the Gauss-Jordan elimination method [8] is applied.

$$\begin{bmatrix} 1 & -2/3 \\ 0 & 0 \end{bmatrix}$$

Finding the unit-length vector in the kernel of this matrix

$$u_1 = \begin{bmatrix} -0.5547 \\ -0.8321 \end{bmatrix}$$

Repeating the same process from substituting λ (i.e. For $\lambda = 5$) to finding the unit-length vector in the kernel of matrix,

$$u_2 = \begin{bmatrix} -0.8321 \\ 0.5547 \end{bmatrix}$$

Therefore, Matrix U is:

$$U = \begin{bmatrix} -0.5547 & -0.8321 \\ -0.8321 & 0.5547 \end{bmatrix}$$

To determine the eigenvalues of $M^T M$, the same process applied when evaluating the eigenvalues of MM^T is applied.

$$\lambda = 0, \lambda = 18, \lambda = 5$$

All similar steps previously taken to obtain the vectors for matrix U is also repeated but with $\lambda = 0, 18$ and 5 to obtain matrix V and then it's transpose

$$V = \begin{bmatrix} -0.8498 & 0 & 0 \\ -0.5230 & 0.1240 & 0 \\ -0.0654 & -0.9923 & 0 \end{bmatrix}$$

$$V^T = \begin{bmatrix} -0.8498 & -0.5230 & -0.0654 \\ 0 & 0.1240 & -0.9923 \\ 0 & 0 & 0 \end{bmatrix}$$

Therefore, the SVD output from matrix M is:

$$M = U_{m\times m} \cdot \Sigma_{m\times n} \cdot V^T_{n\times n}$$

$$= \begin{bmatrix} -0.5547 & -0.8321 \\ -0.8321 & 0.5547 \end{bmatrix} \begin{bmatrix} 4.2426 & 0 & 0 \\ 0 & 2.2361 & 0 \end{bmatrix} \begin{bmatrix} -0.8498 & -0.5230 & -0.0654 \\ 0 & 0.1240 & -0.9923 \\ 0 & 0 & 0 \end{bmatrix}$$

Appendix B

Empirical Test Data

B.1 Sample of one-mode network Data

TABLE B.1: Five Actor Nodes network: A one-mode network sample similar to the data initially used for empirical test

$\mathbf{Ve_1}$	$\mathbf{Ve_2}$	$\mathbf{Ve_3}$
1	2	1
1	3	32
1	32	1
1	36	12
3	1	35
3	2	7
3	32	19
3	36	9
32	1	1
32	3	8
36	1	6
36	3	5
36	32	2

B.2 Sample from a Two-mode Dataset: Social Forum Dataset

A two-mode dataset of a social network can be considered for empirical test to estimate the degree of activeness for each entity in the social network. An example of such data is the randomly selected data sample of a Facebook-like forum which was retrieved from original data used by Tore Opsahl in his research [151]. This data sample reveals interaction of 20 users $V1$ that share ties with 211 items $V2$. The data sample have been partitioned here according to the index representing a user $v \in V1$. Each item has weight $w \in V3$ based on the frequency of interaction by a user.

With the conversion of the above two-mode network to a one-mode network, activeness trust values are estimated and stored in an activeness matrix (See in figure B.1) which reveals the activeness trust value assigned by each node i to other nodes j as well as the node's self-loop (i.e. $P_{i,i} = 1.00$).

	[1]	[2]	[3]	[5]	[6]	[7]	[8]	[9]	[19]	[26]	[30]	[32]	[36]	[40]	[42]	[44]	[50]	[53]	[90]	[100]
[1]	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.75	0.00	0.29	0.00	0.56
[2]	0.00	1.00	0.54	0.00	0.00	0.00	0.00	0.00	0.00	0.50	0.00	0.00	0.00	0.00	0.60	0.50	0.00	0.73	0.60	0.89
[3]	0.00	0.46	1.00	0.00	0.00	0.21	0.00	0.00	0.14	0.60	0.71	0.26	0.80	0.00	0.00	0.58	0.00	0.56	0.32	0.76
[5]	0.00	0.00	0.00	1.00	0.50	0.96	0.00	0.00	0.00	0.00	0.00	0.00	0.50	0.00	0.00	0.00	0.00	0.00	0.00	0.99
[6]	0.00	0.00	0.00	0.50	1.00	0.96	0.00	0.00	0.00	0.00	0.00	0.00	0.50	0.33	0.44	0.00	0.00	0.50	0.00	0.97
[7]	0.00	0.00	0.79	0.04	0.04	1.00	0.00	0.00	0.75	0.00	0.00	0.98	0.27	0.00	0.00	0.00	0.00	0.00	0.00	0.91
[8]	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
[9]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.82	0.00	0.96
[19]	0.00	0.00	0.86	0.00	0.00	0.25	0.00	0.00	1.00	0.00	0.63	0.92	0.57	0.00	0.00	0.69	0.00	0.94	0.67	0.92
[26]	0.00	0.50	0.40	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.62	0.00	0.20	0.11	0.00	0.74	0.00	0.80	0.33	0.85
[30]	0.00	0.00	0.29	0.00	0.00	0.00	0.00	0.00	0.37	0.38	1.00	0.00	0.00	0.00	0.00	0.76	0.00	0.00	0.25	0.58
[32]	0.00	0.00	0.74	0.00	0.00	0.02	0.00	0.00	0.08	0.00	0.00	1.00	0.13	0.00	0.00	0.00	0.00	0.00	0.00	0.50
[36]	0.00	0.00	0.20	0.50	0.50	0.73	0.00	0.00	0.43	0.80	0.00	0.87	1.00	0.33	0.00	0.62	0.00	0.94	0.00	0.96
[40]	0.00	0.00	0.00	0.00	0.67	0.00	0.00	0.00	0.00	0.89	0.00	0.00	0.67	1.00	0.00	0.80	0.00	0.00	0.50	0.75
[42]	0.00	0.40	0.00	0.00	0.56	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.70	0.00	0.82
[44]	0.25	0.50	0.42	0.00	0.00	0.00	0.00	0.00	0.31	0.26	0.24	0.00	0.38	0.20	0.00	1.00	0.00	0.55	0.00	0.52
[50]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00
[53]	0.71	0.27	0.44	0.00	0.50	0.00	0.00	0.18	0.06	0.20	0.00	0.00	0.06	0.00	0.30	0.45	0.00	1.00	0.12	0.68
[90]	0.00	0.40	0.68	0.00	0.00	0.00	0.00	0.00	0.33	0.67	0.75	0.00	0.00	0.50	0.00	0.00	0.00	0.88	1.00	0.86
[100]	0.44	0.11	0.24	0.01	0.03	0.09	0.00	0.04	0.08	0.15	0.42	0.50	0.04	0.25	0.18	0.48	0.00	0.32	0.14	1.00

FIGURE B.1: Activeness matrix for Data sample with 20 users and 211 items

Applying cosine similarity measure on the activeness matrix will generate global familiarity value for each user (See table B.2). This values which reveals the level of activeness to the social network predicts if the users will be trustworthy to global social users. This values could be used to identify the most active node (**MAN**) amongst group of nodes as the most influential member in the group.

TABLE B.2: Global Familiarity for each user from activeness matrix of Social Data sample

Social $Node_i$	Fam_i
1	0.34
2	0.48
3	0.54
5	0.32
6	0.39
7	0.38
8	0.00
9	0.35
19	0.51
26	0.52
30	0.41
32	0.32
36	0.52
40	0.39
42	0.38
44	0.52
50	0.00
53	0.50
90	0.47
100	0.56

As the activeness rating from each node to other nodes are not widely distributed, there is need to normalize the activeness matrix to obtain accurate prediction of user's engaging preference. This matrix referred to as **Transition 'Activeness' matrix** could be used as the input matrix for a clustering technique to determine trust neighbourhood of users.

	[1]	[2]	[3]	[5]	[6]	[7]	[8]	[9]	[19]	[26]	[30]	[32]	[36]	[40]	[42]	[44]	[50]	[53]	[90]	[100]
[1]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.12	0.00	0.04	0.00	0.04
[2]	0.00	0.00	0.10	0.00	0.00	0.00	0.00	0.00	0.00	0.11	0.00	0.00	0.00	0.00	0.39	0.08	0.00	0.09	0.20	0.07
[3]	0.00	0.17	0.00	0.00	0.00	0.07	0.00	0.00	0.05	0.13	0.21	0.07	0.19	0.00	0.00	0.09	0.00	0.07	0.11	0.06
[5]	0.00	0.00	0.00	0.00	0.18	0.30	0.00	0.00	0.00	0.00	0.00	0.00	0.12	0.00	0.00	0.00	0.00	0.00	0.00	0.07
[6]	0.00	0.00	0.00	0.48	0.00	0.30	0.00	0.00	0.00	0.00	0.00	0.00	0.12	0.19	0.29	0.00	0.00	0.06	0.00	0.07
[7]	0.00	0.00	0.14	0.04	0.01	0.00	0.00	0.00	0.29	0.00	0.00	0.28	0.07	0.00	0.00	0.00	0.00	0.00	0.00	0.07
[8]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
[9]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.10	0.00	0.07
[19]	0.00	0.00	0.15	0.00	0.00	0.08	0.00	0.00	0.00	0.00	0.19	0.26	0.14	0.00	0.00	0.11	0.00	0.12	0.23	0.07
[26]	0.00	0.19	0.07	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.18	0.00	0.05	0.06	0.00	0.12	0.00	0.10	0.11	0.06
[30]	0.00	0.00	0.05	0.00	0.00	0.00	0.00	0.00	0.15	0.09	0.00	0.00	0.00	0.00	0.00	0.12	0.00	0.00	0.09	0.04
[32]	0.00	0.00	0.13	0.00	0.00	0.01	0.00	0.00	0.03	0.00	0.00	0.00	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.04
[36]	0.00	0.00	0.04	0.48	0.18	0.23	0.00	0.00	0.17	0.18	0.00	0.25	0.00	0.19	0.00	0.10	0.00	0.12	0.00	0.07
[40]	0.00	0.00	0.00	0.00	0.24	0.00	0.00	0.00	0.00	0.20	0.00	0.00	0.16	0.00	0.00	0.13	0.00	0.00	0.17	0.06
[42]	0.00	0.15	0.00	0.00	0.20	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.09	0.00	0.06
[44]	0.18	0.19	0.08	0.00	0.00	0.00	0.00	0.00	0.12	0.06	0.07	0.00	0.09	0.12	0.00	0.00	0.00	0.07	0.00	0.04
[50]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
[53]	0.51	0.10	0.08	0.00	0.18	0.00	0.00	0.82	0.02	0.04	0.00	0.00	0.01	0.00	0.20	0.07	0.00	0.00	0.04	0.05
[90]	0.00	0.15	0.12	0.00	0.00	0.00	0.00	0.00	0.13	0.15	0.22	0.00	0.00	0.29	0.00	0.00	0.00	0.11	0.00	0.06
[100]	0.31	0.04	0.04	0.01	0.01	0.03	0.00	0.18	0.03	0.03	0.12	0.14	0.01	0.15	0.12	0.08	0.00	0.04	0.05	0.00

FIGURE B.2: Normalized Activeness matrix with no self-loop of each node (Transition 'activeness' matrix)

B.3 Clustering Transition 'Activeness' Matrix

B.3.1 Cluster Analysis using HCS

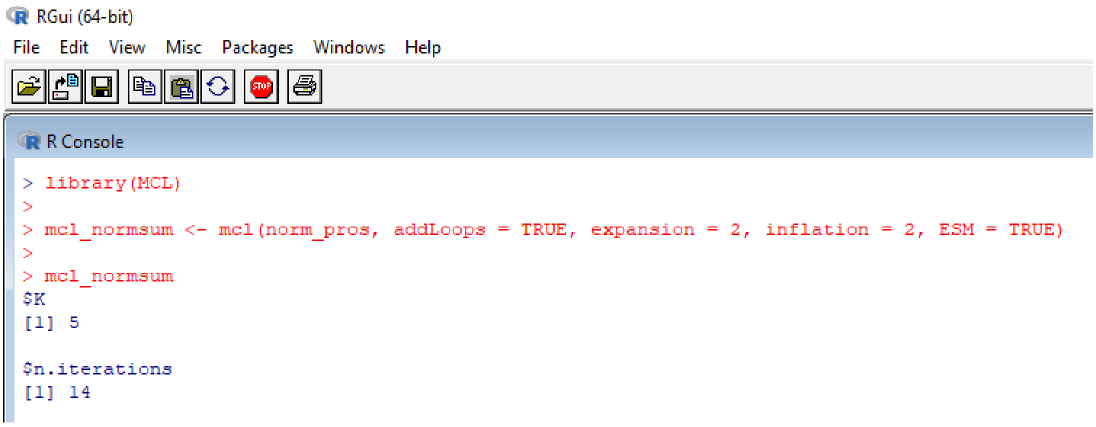
Using R, a package known as "RBGL" [39] which is required for applying 'highly connected subgraphs'(HCS) clustering was initially retrieved from the bioconductor source and installed.

Applying HCS clustering algorithm to the social graph in figure 4.1, five clusters including four singletons were generated.

FIGURE B.3: HCS clustering on social graph from figure 4.1.

B.3.2 Cluster Analysis using Markov Clustering

Using R, a package known as "MCL" [100] was installed before applying Markov clustering algorithm on transition matrix or probability matrix.



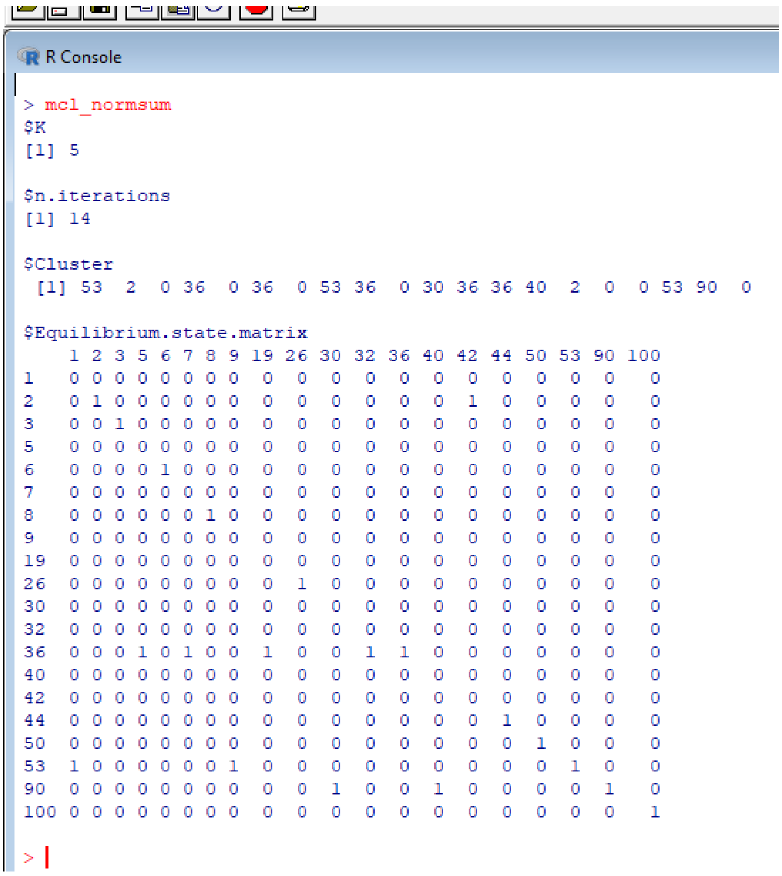
RGui (64-bit)
File Edit View Misc Packages Windows Help
R Console

```
> library(MCL)
>
> mcl_normsum <- mcl(norm_pros, addLoops = TRUE, expansion = 2, inflation = 2, ESM = TRUE)
>
> mcl_normsum
$K
[1] 5

$n.iterations
[1] 14
```

FIGURE B.4: Loading and application of the MCL function

Five clusters are generated along with an equilibrium matrix to reveal the attractors or influential members, the clusters and their members.



R Console

```
> mcl_normsum
$K
[1] 5

$n.iterations
[1] 14

$Cluster
 [1] 53  2  0 36  0 36  0 53 36  0 30 36 36 40  2  0  0 53 90  0

$Equilibrium.state.matrix
    1 2 3 5 6 7 8 9 19 26 30 32 36 40 42 44 50 53 90 100
1   0 0 0 0 0 0 0 0  0  0  0  0  0  0  0  0  0  0  0   0
2   0 1 0 0 0 0 0 0  0  0  0  0  0  0  1  0  0  0  0   0
3   0 0 1 0 0 0 0 0  0  0  0  0  0  0  0  0  0  0  0   0
5   0 0 0 0 0 0 0 0  0  0  0  0  0  0  0  0  0  0  0   0
6   0 0 0 0 1 0 0 0  0  0  0  0  0  0  0  0  0  0  0   0
7   0 0 0 0 0 0 0 0  0  0  0  0  0  0  0  0  0  0  0   0
8   0 0 0 0 0 0 1 0  0  0  0  0  0  0  0  0  0  0  0   0
9   0 0 0 0 0 0 0 0  0  0  0  0  0  0  0  0  0  0  0   0
19  0 0 0 0 0 0 0 0  0  0  0  0  0  0  0  0  0  0  0   0
26  0 0 0 0 0 0 0 0  0  1  0  0  0  0  0  0  0  0  0   0
30  0 0 0 0 0 0 0 0  0  0  0  0  0  0  0  0  0  0  0   0
32  0 0 0 0 0 0 0 0  0  0  0  0  0  0  0  0  0  0  0   0
36  0 0 0 1 0 1 0 0  1  0  0  1  1  0  0  0  0  0  0   0
40  0 0 0 0 0 0 0 0  0  0  0  0  0  0  0  0  0  0  0   0
42  0 0 0 0 0 0 0 0  0  0  0  0  0  0  0  0  0  0  0   0
44  0 0 0 0 0 0 0 0  0  0  0  0  0  0  0  1  0  0  0   0
50  0 0 0 0 0 0 0 0  0  0  0  0  0  0  0  0  1  0  0   0
53  1 0 0 0 0 0 0 1  0  0  0  0  0  0  0  0  0  1  0   0
90  0 0 0 0 0 0 0 0  0  0  1  0  0  1  0  0  0  0  1   0
100 0 0 0 0 0 0 0 0  0  0  0  0  0  0  0  0  0  0  0   1

>
```

FIGURE B.5: Outputs from the Markov clustering algorithm

B.3.3 Cluster Analysis using AP Clustering

Using R, a package known as “apcluster” [24, 67] was installed to enable the negative euclidean distance measure and the affinity propagation (AP) algorithm to be used on the data.

R Console

```
> norm_pros_loop
         1      2      3      5      6      7 8      9     19     26     30     32     36     40     42     44 50     53     90    100
1   1.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.1177  0 0.0361 0.0000 0.0415
2   0.0000 1.0000 0.0964 0.0000 0.0000 0.0000 0 0.0000 0.0000 0.1124 0.0000 0.0000 0.0000 0.0000 0.3947 0.0785  0 0.0909 0.2048 0.0660
3   0.0000 0.1742 1.0000 0.0000 0.0000 0.0652 0 0.0000 0.0549 0.1348 0.2107 0.0737 0.1942 0.0000 0.0000 0.0911  0 0.0697 0.1092 0.0564
5   0.0000 0.0000 0.0000 1.0000 0.1786 0.2981 0 0.0000 0.0000 0.0000 0.0000 0.0000 0.1214 0.0000 0.0000 0.0000  0 0.0000 0.0000 0.0734
6   0.0000 0.0000 0.0000 0.4762 1.0000 0.2981 0 0.0000 0.0000 0.0000 0.0000 0.0000 0.1214 0.1919 0.2895 0.0000  0 0.0623 0.0000 0.0720
7   0.0000 0.0000 0.1411 0.0381 0.0143 1.0000 0 0.0000 0.2941 0.0000 0.0000 0.2776 0.0655 0.0000 0.0000 0.0000  0 0.0000 0.0000 0.0675
8   0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 1 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000  0 0.0000 0.0000 0.0000
9   0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0 1.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000  0 0.1021 0.0000 0.0712
19  0.0000 0.0000 0.1536 0.0000 0.0000 0.0776 0 0.0000 1.0000 0.0000 0.1869 0.2606 0.1383 0.0000 0.0000 0.1083  0 0.1171 0.2287 0.0682
26  0.0000 0.1894 0.0714 0.0000 0.0000 0.0000 0 0.0000 0.0000 1.0000 0.1840 0.0000 0.0485 0.0640 0.0000 0.1162  0 0.0996 0.1126 0.0631
30  0.0000 0.0000 0.0518 0.0000 0.0000 0.0000 0 0.0000 0.1451 0.0854 1.0000 0.0000 0.0000 0.0000 0.0000 0.1193  0 0.0000 0.0853 0.0430
32  0.0000 0.0000 0.1321 0.0000 0.0000 0.0062 0 0.0000 0.0314 0.0000 0.0000 1.0000 0.0316 0.0000 0.0000 0.0000  0 0.0000 0.0000 0.0371
36  0.0000 0.0000 0.0357 0.4762 0.1786 0.2267 0 0.0000 0.1686 0.1798 0.0000 0.2465 1.0000 0.1919 0.0000 0.0973  0 0.1171 0.0000 0.0712
40  0.0000 0.0000 0.0000 0.0000 0.2393 0.0000 0 0.0000 0.0000 0.2000 0.0000 0.0000 0.1626 1.0000 0.0000 0.1256  0 0.0000 0.1706 0.0556
42  0.0000 0.1515 0.0000 0.0000 0.2000 0.0000 0 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 1.0000 0.0000  0 0.0872 0.0000 0.0608
44  0.1786 0.1894 0.0750 0.0000 0.0000 0.0000 0 0.0000 0.1216 0.0584 0.0712 0.0000 0.0922 0.1163 0.0000 1.0000  0 0.0685 0.0000 0.0386
50  0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000  1 0.0000 0.0000 0.0000
53  0.5071 0.1023 0.0786 0.0000 0.1786 0.0000 0 0.8182 0.0235 0.0449 0.0000 0.0000 0.0146 0.0000 0.1974 0.0706  0 1.0000 0.0410 0.0504
90  0.0000 0.1515 0.1214 0.0000 0.0000 0.0000 0 0.0000 0.1294 0.1506 0.2226 0.0000 0.0000 0.2907 0.0000 0.0000  0 0.1096 1.0000 0.0638
100 0.3143 0.0417 0.0429 0.0095 0.0107 0.0280 0 0.1818 0.0314 0.0337 0.1246 0.1416 0.0097 0.1453 0.1184 0.0754  0 0.0399 0.0478 1.0000
>
> sim_apned.normloop <- negDistMat(norm_pros_loop, r = 2)
```

FIGURE B.6: Negative Euclidean distance applied to a transition matrix with self-loops

From the similarity matrix generated using negative Euclidean measure and an input preference value (By default, the median value is selected from the matrix), six clusters were generated with their individual exemplars.

RGui (64-bit) - [R Console]

File Edit View Misc Packages Windows Help

```
> apned_normloop <- apcluster(sim_apned.normloop, details = TRUE)
>
> apned_normloop

APResult object

Number of samples     =  20
Number of iterations  =  175
Input preference      =  -2.143026
Sum of similarities   =  -21.1172
Sum of preferences    =  -12.85816
Net similarity        =  -33.97535
Number of clusters    =  6

Exemplars:
   1 2 5 9 19 90
Clusters:
   Cluster 1, exemplar 1:
      1 44 100
   Cluster 2, exemplar 2:
      2 42
   Cluster 3, exemplar 5:
      5 6 36
   Cluster 4, exemplar 9:
      8 9 50 53
   Cluster 5, exemplar 19:
      7 19 30 32
   Cluster 6, exemplar 90:
      3 26 40 90
>
```

FIGURE B.7: Outputs from AP clustering algorithm using negative Euclidean distance

From the similarity matrix using pearson correlation measure (Installed from the 'apcluster' package), five clusters where generated with non-active nodes being the only members of the cluster.

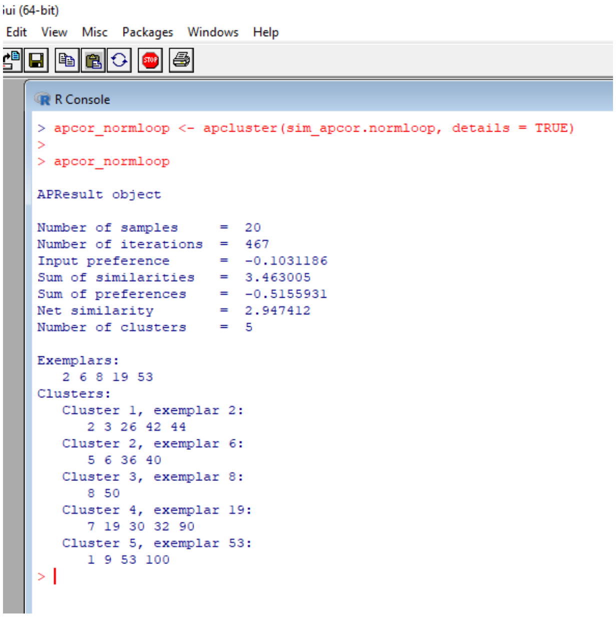


FIGURE B.8: Outputs from AP clustering algorithm using Pearson correlation measure

It is would be better if the non-active nodes (i.e. node 8 and node 50) were separated to form singletons as this will prevent them from competing for the exemplar role of their cluster (i.e. cluster 4).

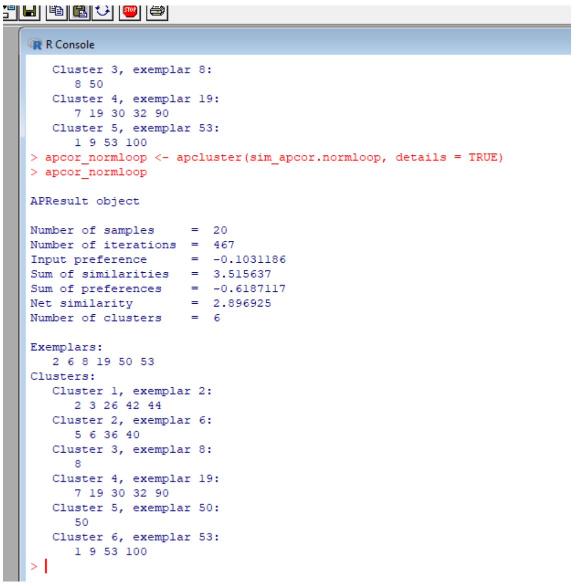
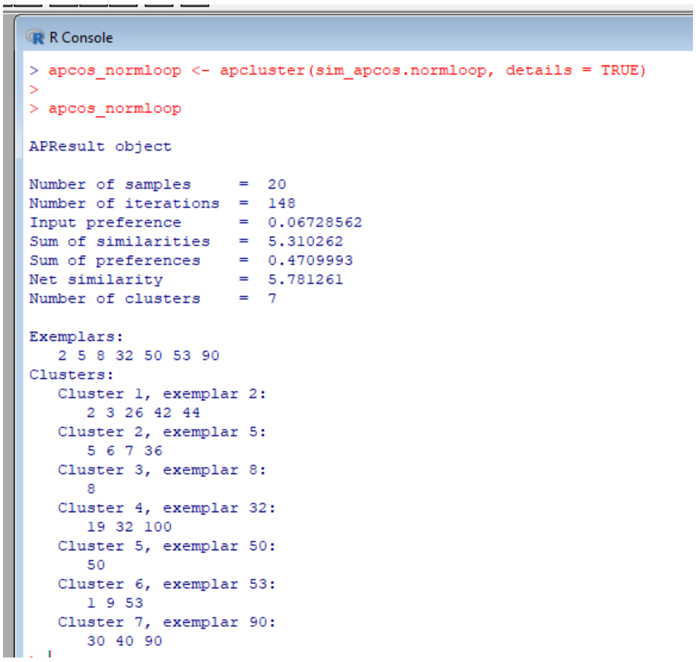


FIGURE B.9: Possiblity of node 8 and node 50 after several trials of the pearson correlation based AP algorithm

Considering the AP clustering algorithm using the cosine similarity measure(Installed from ‘lsa’ package), seven clusters were generated with non-active nodes being members of singletons.



```
R Console
> apcos_normloop <- apcluster(sim_apcos.normloop, details = TRUE)
>
> apcos_normloop

APResult object

Number of samples     =  20
Number of iterations  =  148
Input preference      =  0.06728562
Sum of similarities   =  5.310262
Sum of preferences    =  0.4709993
Net similarity        =  5.781261
Number of clusters    =  7

Exemplars:
   2 5 8 32 50 53 90
Clusters:
   Cluster 1, exemplar 2:
      2 3 26 42 44
   Cluster 2, exemplar 5:
      5 6 7 36
   Cluster 3, exemplar 8:
      8
   Cluster 4, exemplar 32:
      19 32 100
   Cluster 5, exemplar 50:
      50
   Cluster 6, exemplar 53:
      1 9 53
   Cluster 7, exemplar 90:
      30 40 90
```

FIGURE B.10: Outputs from AP clustering algorithm using Cosine similarity measure

After observing the three outputs above, the application of the AP clustering algorithm along with Pearson correlation measure appeared to be a more suitable method as it was able to cluster the non-active members together. The less-active members were also clustered along with a member that had the most interactions with items and also had high frequency towards those items.

B.4 Sample from a Two-mode Dataset: ‘Hollywood Film-music’ Dataset

The following activeness matrix generated is based on a data sample retrieved from ‘Hollywood film-music’ dataset presented by Vladimir Batagelj & Andrej Mrvar [18]. The data sample consist of collaboration between 30 film producers $V1$ (indexed between 1 and 62) and 35 music composers $V2$ (indexed between 63 and 102) , where a network consider the composers as entities that must have previously been employed by any of the film producers at least once (i.e. Frequency of music composition for the film producer $V3$). The transformed one-mode network from the two-mode network was used to estimate the activeness value revealed in the activeness matrix shown below.

	[1]	[2]	[4]	[7]	[8]	[9]	[13]	[14]	[16]	[17]	[21]	[22]	[23]	[24]	[27]	[29]	[30]	[31]	[33]	[35]	[37]	[39]	[43]	[45]	[47]	[49]	[50]	[54]	[61]	[62]
[1]	1.00	0.40	0.25	0.25	0.25	0.25	0.40	0.25	0.00	0.25	0.00	0.25	0.00	0.00	0.00	0.00	0.43	0.00	0.50	0.67	0.50	0.50	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
[2]	0.60	1.00	0.33	0.33	0.43	0.33	0.50	0.33	0.00	0.33	0.00	0.43	0.50	0.00	0.00	0.00	0.40	0.00	0.38	0.50	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.67	0.00	0.33
[4]	0.75	0.67	1.00	0.50	0.50	0.50	0.67	0.50	0.00	0.50	0.00	0.50	0.00	0.00	0.00	0.00	0.50	0.00	0.00	0.00	0.00	0.50	0.50	0.50	0.50	0.00	0.00	0.00	0.00	0.00
[7]	0.75	0.67	0.50	1.00	0.50	0.50	0.67	0.50	0.00	0.50	0.00	0.50	0.00	0.00	0.00	0.00	0.50	0.00	0.00	0.00	0.00	0.50	0.50	0.50	0.00	0.00	0.00	0.00	0.00	0.00
[8]	0.75	0.57	0.50	0.50	1.00	0.50	0.67	0.50	0.00	0.50	0.00	0.50	0.00	0.00	0.00	0.00	0.60	0.00	0.33	0.00	0.00	0.00	0.00	0.00	0.33	0.00	0.00	0.00	0.00	0.00
[9]	0.75	0.67	0.50	0.50	0.50	1.00	0.67	0.50	0.00	0.50	0.00	0.50	0.00	0.00	0.00	0.00	0.50	0.17	0.17	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.17	0.00	0.00	0.00
[13]	0.60	0.50	0.33	0.33	0.33	0.33	1.00	0.33	0.40	0.33	0.00	0.33	0.25	0.00	0.00	0.00	0.33	0.00	0.00	0.00	0.00	0.00	0.25	0.00	0.00	0.00	0.00	0.00	0.00	0.00
[14]	0.75	0.67	0.50	0.50	0.50	0.50	0.67	1.00	0.00	0.50	0.00	0.50	0.00	0.00	0.00	0.00	0.50	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
[16]	0.00	0.00	0.00	0.00	0.00	0.00	0.60	0.00	1.00	0.00	0.00	0.00	0.33	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.33	0.00	0.00	0.00	0.00	0.00	0.00	0.00
[17]	0.75	0.67	0.50	0.50	0.50	0.50	0.67	0.50	0.00	1.00	0.00	0.50	0.00	0.00	0.00	0.00	0.50	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
[21]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
[22]	0.75	0.57	0.50	0.50	0.50	0.50	0.67	0.50	0.00	0.50	0.00	1.00	0.50	0.00	0.00	0.00	0.50	0.00	0.50	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.60	0.00	0.00
[23]	0.00	0.50	0.00	0.00	0.00	0.00	0.75	0.00	0.67	0.00	0.00	0.50	1.00	0.00	0.00	0.00	0.00	0.00	0.50	0.00	0.00	0.00	0.50	0.00	0.00	0.00	0.00	0.00	0.00	0.00
[24]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
[27]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
[29]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
[30]	0.57	0.60	0.50	0.50	0.40	0.50	0.67	0.50	0.00	0.50	0.00	0.50	0.00	0.00	0.00	0.00	1.00	0.00	0.33	0.50	0.38	0.33	0.33	0.00	0.00	0.00	0.00	0.67	0.00	0.33
[31]	0.00	0.00	0.00	0.00	0.00	0.83	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.60	0.50	0.00	0.50	0.00	0.00	0.00	0.00	0.60	0.00	0.00	0.00
[33]	0.50	0.62	0.00	0.00	0.67	0.83	0.00	0.00	0.00	0.00	0.00	0.50	0.33	0.00	0.00	0.00	0.67	0.40	1.00	0.67	0.50	0.40	0.00	0.00	0.00	0.00	0.50	0.00	0.00	0.50
[35]	0.33	0.50	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.50	0.50	0.33	1.00	0.33	0.40	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.33
[37]	0.50	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.62	0.00	0.50	0.67	1.00	0.50	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.50
[39]	0.50	0.00	0.50	0.50	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.67	0.50	0.60	0.60	0.50	1.00	0.50	0.50	0.00	0.00	0.67	0.00	0.00	0.00
[43]	0.00	0.00	0.50	0.50	0.00	0.00	0.75	0.00	0.67	0.00	0.00	0.00	0.50	0.00	0.00	0.00	0.67	0.00	0.00	0.00	0.00	0.50	1.00	0.50	0.00	0.00	0.00	0.00	0.00	0.00
[45]	0.00	0.00	0.50	0.50	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.50	0.50	1.00	0.00	0.00	0.00	0.00	0.00	0.00
[47]	0.00	0.00	0.50	0.00	0.67	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.50	0.00	0.00	0.00	0.00
[49]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.50	1.00	0.00	0.00	0.00	0.00
[50]	0.00	0.00	0.00	0.00	0.00	0.83	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.40	0.50	0.00	0.00	0.33	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00
[54]	0.00	0.33	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.40	0.00	0.00	0.00	0.00	0.33	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00
[61]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00
[62]	0.00	0.67	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.67	0.00	0.50	0.67	0.50	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00

FIGURE B.11: Activeness matrix from a 'Hollywood film-music' data sample

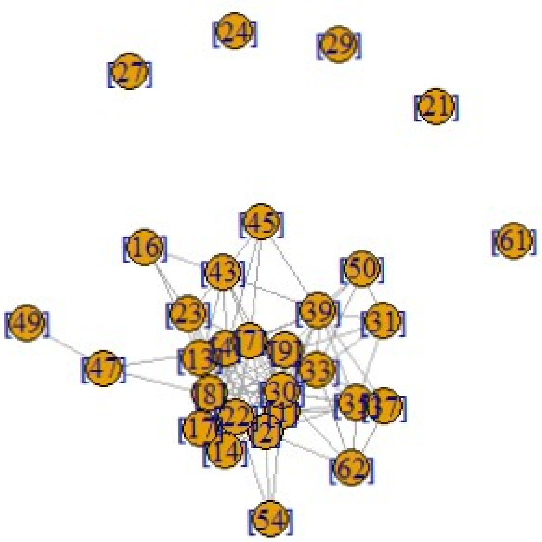


FIGURE B.12: A social graph of sample data with 30 producers from Hollywood film-music data

With the application of cosine similarity measure to compare the activeness relationship between node, the global familiarity value can be obtained to identify the most active node (**MAN**) amongst the nodes in the network.

TABLE B.3: Global Familiarity for each producer from activeness matrix of 'Hollywood film-music' data

Producer $Node_i$	Fam_i
1	0.42
2	0.43
4	0.43
7	0.43
8	0.42
9	0.43
13	0.41
14	0.40
16	0.13
17	0.40
21	0.00
22	0.42
23	0.24
24	0.00
27	0.00
29	0.00
30	0.46
31	0.20
33	0.38
35	0.29
37	0.25
39	0.32
43	0.28
45	0.16
47	0.12
49	0.03
50	0.16
54	0.16
61	0.00
62	0.24

B.4.1 Clustering Transition 'Activeness' Matrix of 'Hollywood film-music' Data

With the activeness matrix from the Hollywood film-music data sample, its transition 'activeness' matrix is retrieved for the generation of clusters along with their individual exemplar using Affinity propagation clustering algorithm. The different similarity measure used as possible measures in previous analysis were also considered for the cluster analysis on the matrix.

```
APResult object

Number of samples     =  30
Number of iterations  =  156
Input preference      =  -1.462451
Sum of similarities   =  -30.07697
Sum of preferences    =  -8.774705
Net similarity        =  -38.85167
Number of clusters    =  6

Exemplars:
   [16] [37] [45] [49] [50] [54]
Clusters:
   Cluster 1, exemplar [16]:
      [13] [16] [23] [43]
   Cluster 2, exemplar [37]:
      [1] [35] [37] [62]
   Cluster 3, exemplar [45]:
      [4] [7] [45]
   Cluster 4, exemplar [49]:
      [8] [47] [49]
   Cluster 5, exemplar [50]:
      [9] [31] [33] [39] [50]
   Cluster 6, exemplar [54]:
      [2] [14] [17] [21] [22] [24] [27] [29] [30] [54] [61]
>
```

FIGURE B.13: Clusters from activeness matrix of Hollywood film-music data using Negative Euclidean distance

```
APResult object

Number of samples     =  30
Number of iterations  =  138
Input preference      =  0.02870754
Sum of similarities   =  7.230835
Sum of preferences    =  0.315783
Net similarity        =  7.546618
Number of clusters    =  11

Exemplars:
   [7] [16] [21] [24] [27] [29] [37] [47] [50] [54] [61]
Clusters:
   Cluster 1, exemplar [7]:
      [7] [14] [17] [45]
   Cluster 2, exemplar [16]:
      [13] [16] [23] [43]
   Cluster 3, exemplar [21]:
      [21]
   Cluster 4, exemplar [24]:
      [24]
   Cluster 5, exemplar [27]:
      [27]
   Cluster 6, exemplar [29]:
      [29]
   Cluster 7, exemplar [37]:
      [1] [35] [37] [62]
   Cluster 8, exemplar [47]:
      [4] [8] [47] [49]
   Cluster 9, exemplar [50]:
      [9] [31] [33] [39] [50]
   Cluster 10, exemplar [54]:
      [2] [22] [30] [54]
   Cluster 11, exemplar [61]:
      [61]
>
```

FIGURE B.14: Clusters from activeness matrix of Hollywood film-music data using Cosine Similarity measure

```
APResult object

Number of samples     =  30
Number of iterations  =  143
Input preference      =  -0.06834713
Sum of similarities   =  5.458
Sum of preferences    =  -0.4100828
Net similarity        =  5.047917
Number of clusters    =  6

Exemplars:
   [7] [16] [37] [47] [50] [54]
Clusters:
   Cluster 1, exemplar [7]:
      [7] [14] [17] [45]
   Cluster 2, exemplar [16]:
      [13] [16] [23] [43]
   Cluster 3, exemplar [37]:
      [1] [35] [37] [62]
   Cluster 4, exemplar [47]:
      [4] [8] [47] [49]
   Cluster 5, exemplar [50]:
      [9] [31] [33] [39] [50]
   Cluster 6, exemplar [54]:
      [2] [21] [22] [24] [27] [29] [30] [54] [61]
> |
```

FIGURE B.15: Clusters from activeness matrix of Hollywood film-music data using Pearson Correlation measure

B.4.2 Applying Singular Value Decomposition (SVD) Algorithm to Cluster Members

Considering members of cluster output (See Figure B.15), their predicted preferences for recommended items are presented below (See Table B.4). The recommended items are cluster exemplar's preferred items (i.e. music composers V2 that the exemplar previously employed to compose a music for his/her produced film) which can be observed in the data sample from the 'Hollywood film-music' dataset.

TABLE B.4: Predicted ratings for items recommended by Exemplar in clusters from Correlation similarity-based AP Algorithm

Cluster	Item recommended	ID of Members	Predicted rating	Exemplar ID
Cluster1	79	7(MAN)	2.00	7
		14	0.50	
		17	0.50	
		45	1.50	
Cluster2	64	13 (MAN)	6.00	16
		16	4.00	
		23	2.00	
		43	2.00	
Cluster3	83	1 (MAN)	0.33	37
		35	0.67	
		37	2.00	
		62	1.33	
Cluster4	85	4 (MAN)	1.33	47
		8	1.33	
		47	2.33	
		49	0.33	
Cluster5	81	9 (MAN)	6.25	50
		31	1.50	
		33	1.75	
		39	0.25	
		50	2.25	
Cluster6	84	2	0.33	54
		21	0.00	
		22	0.67	
		24	0.00	
		27	0.00	
		29	0.00	
		30 (MAN)	0.33	
		54	2.33	
		61	0.00	