Int. J. Advance. Soft Comput. Appl., Vol. 6, No.2, July 2014
ISSN 2074-8523; Copyright © SCRG Publication, 2014

Flight MH370 Community Structure

Mohammed Z. Al-Taie[1], Siti Mariyam Shamsuddin[2], Nor Bahiah Ahmad[3]

[1,2] *UTM Big Data Centre*
Universiti Teknologi Malaysia (UTM), Skudai, 81310 Johor, Malaysia
e-mail: mza004@live.aul.edu.lb, mariyam@utm.my

[3] *Soft Computing Research Group*, Faculty of Computing
Universiti Teknologi Malaysia (UTM), Skudai, 81310 Johor, Malaysia
e-mail: bahiah@utm.my

Abstract

Studying community structure has proven efficient in understanding social forms of interactions among people, quantitatively investigating some of the well-known social theories and providing useful recommendations to users in communities based on common interests. Another important feature for community structure is that it allows for classification of vertices according to their structural positions in clusters such that some vertices may have an important function of control and stability within the community while others may play an important role of leading relationships and exchanges between different communities. Studying the community structure of Flight MH370 will help us finding patterns that emerge from that structure which can lead to demystify some of the many ambiguous aspects of that flight. The aim of this study is to analyze the mesoscopic and macroscopic features of that community using social network analysis. Pajek, which is a program for social network analysis, is used to generate a series of social networks that represent the different network communities.

Keywords — *Community Detection, Community Structure, Flight MH370, Graph Theory, Pajek, Social Network Analysis, SNA*

1 Introduction

Community structure (also called clustering) is the organization of vertices in clusters, where many edges join vertices of the same cluster and few edges join vertices of different clusters [1]. Finding communities in networks has recently become an active research area for scientists and researchers.

In [2], the author discussed a number of methods for community detection. For example, from the computer science perspective, the problem is addressed by dividing the nodes of a network into a number of groups while minimizing the number of edges between different groups. Many algorithms stand here such as the *Spectral Bisection Algorithm* (which is based on the properties of the spectrum of the Laplacian matrix) and the *Kernighan-Lin Algorithm* (which is one of the earliest approaches proposed and is still frequently used but mostly in combination with other methods. It is based on the idea of partitioning electronic circuits onto boards). Other popular methods for graph partitioning include the *geometric algorithm*, *level-structure partitioning* and *multilevel algorithms* [1].

Social sciences have adopted a different approach for finding communities which is the use of *Hierarchical Clustering* based on developing a measure of similarity between pairs of vertices. Hierarchical clustering techniques can be classified in two categories (a) *Agglomerative algorithms*, where clusters are iteratively merged (if their similarity is sufficiently high), and (b) *Divisive algorithms*, where clusters are iteratively split by removing edges connecting vertices with low similarity [1].

More recent approaches for community detection include the *Modularity Maximization Algorithm* and the *Resister Network Algorithm* [2]. Modularity, which was first introduced by Newman and Girvan in 2003, is the most used and best known quality function and the one which attempted to achieve a first principle understanding of the clustering problem [1]. This technique consists of two steps: the first one: '*Divisive'* which incorporates iteratively removing edges from the network and thus breaking it into communities, and the second step: '*Recalculation'* where betweenness scores are re-evaluated after the removal of every edge. The second step (missing in the previous algorithms) has the primary importance in the algorithm and thus gives power to the current one [3].

Through harnessing one of the community detection methods mentioned earlier, particularly Modularity, in this study, to investigate the general mesoscopic and macroscopic characteristics of Flight MH370 in order to analyze the statistical and topological characteristics that describe the organizational patterns of the flight community. That Malaysia Airlines Flight (MH370), on March 8 2014, left Kuala Lumpur International Airport at 12.41am on a scheduled flight and disappeared from radar screen about 40 minutes later while over the South China Sea. The aircraft, Boeing 777-200ER, was carrying 239 people (12 crew members and 227 passengers) was decided to land in Beijing Capital International Airport at 6.30am

the same day. However, this has never happened. The passengers were from far-flung parts of the world: China (153), Malaysia (38), Australia (6), Indonesia (7), India (5), France (4), United States 3 (including two toddlers), New Zealand (2), Canada (2), Ukraine (2), Russia (1), Taiwan (1), Italy (1), Austria (1) and Holland (1). They were grandparents, parents, married couples, construction workers, artists, tourists and even children. Different groups of people were onboard: for example; a group of prominent China artists, a group of 20 top management employees of a US company; Freescale Semiconductor, and so on. Hence, the main contribution of this work is that it is the first academic study that addresses, from a scientific point of view, the community structure of Flight MH370.

The remainder of this paper is organized as follows. Section 2 describes a number of previous studies that addressed community structure from a social network analysis (SNA) perspective. In section 3, and since SNA is based on graph theory, we provide a brief introduction to the basic notions of graph theory. In section 4, we provide an introduction to SNA including its historical development, the fields that it is applied to, some of the important measures and some of the widely used modeling tools. Section 5 describes the data used in the study and the analysis of Flight MH370 community structure. Finally, section 6 gives the study conclusions and the main challenges faced during our work.

2 Related Work

The area of community structure has its roots in the problem of graph partitioning whose initial contributions go back to the 1970s. Several studies have been conducted to investigate the community structure of real social networks. In some cases, these studies are stimulated by practical applications, while in others it is more related to cultural analysis. However, all of them are based on the algorithmic background of community discovery.

Since this issue has been studied in a variety of environments, we are interested only in studies with the utilization of SNA:

In order to study the community structure of one celebrated online social network, the authors, in [4], conducted an in-depth analysis of Facebook friendship networks at five American universities during a single-time snapshot in 2005. They investigated each network's community structure -consisting of clusters of nodes- and employed graphical and quantitative tools to measure the correlations between the network communities and the demographic labels included in the data.

Aiming to study and analyze the properties of a number of large and different networks, the authors, in [5], presented an empirical analysis of the statistical properties of large communities of different types such as communication, technological, biological, and social networks. Their results show that the mesoscopic organization of these networks (such as density of communities,

community size distribution, and the average shortest path length) is remarkably similar which is reflected in several characteristics of community structure.

To compare the performance and results of applying different community detection techniques, the authors, in [6], analyzed the performance of three community detection methods (Infomap, Louvain and clique percolation methods) by using them to identify communities in a large social network representing mobile phone call records. They focused on the analysis of several mesoscopic features such as density distribution, neighborhood overlapping and the community size. Their results show that all the three methods can detect communities in some respects but they still come short in others. Also, and according to the same study, there is a hierarchical relationship between communities detected by different methods.

Not like other studies which consider nodes when discovering communities, the authors, in [7], went in a different path of identifying communities which is by considering links rather than nodes. According to the study, it was found that link communities incorporate overlap and show hierarchical organization. Their study spans many networks such as biological and social networks.

Taking another well-recognized online social network into consideration, the authors, in [8], discussed the notion of *Strength of Weak Ties* in the context of Twitter. Twitter is a social platform that maintains a directed graph (in contrast to Facebook which incorporates an undirected graph) and is represented with multiple types of edges. Social ties, in Twitter, represent hierarchical connections (forming follower and followed users). The researchers analyzed the clusters of the network formed by the basic type of connections and, according to the study, the network bears valuable information on the localization of more personal interactions between users and it has some users that act as brokers of information between groups.

After introducing a number of studies that sought to analyze the characteristics of communities found in different environments, in section 3, we will give an introduction to the basic concepts of graph theory, since it is strongly connected to SNA- the analysis tool that is used here to study the characteristics of Flight MH370 community.

3 Graph Theory

Leonhard Euler, in 1736, was the first to talk about graph theory in his writings. His work was on the well-known problem of the seven bridges of town Konigsberg. The problem was how to find a way around that town so that a man would, in the daily movements, go over the seven bridges only once. This problem had led to a new branch of mathematics called *Graph Theory* [15].

The importance of graphs, in our study, is that SNA establishes its basic concepts on mathematical graph theory. This is why we are going to provide an introduction to the basic concepts in graph theory [15] and [19]:

A graph is simply a set of points and lines that connect some pairs of points. Points are called 'vertices', and lines are called 'edges'. A graph G is a set X of vertices together with a set E of edges and is written: $G = (X, E)$. However, we can also describe a graph by listing all its edges. For example, for graph G, the edge list, denoted by $J(G)$, is as follows:

$$J(G) = \{\{x1,x2\},\{x2,x3\},\{x3,x4\},\{x4,x5\},\{x1,x5\},\{x2,x5\},\{x2,x4\}\}.$$

The degree of vertex x is the number of all vertices adjacent to x, denoted by $d(x)$. The maximum degree over all vertices is called the maximum degree of G, denoted by $\Delta(G)$. The set of edges incident to a vertex x is denoted by $E(x)$. A cut-vertex (also called cut-point) is a vertex whose removal increases the number of components.

A loop is an edge that connects a vertex to itself. If a vertex has no neighbors, then that vertex is said to be 'isolate'. If there are many edges connecting the same pair of vertices, then these edges are called 'parallel' or 'multiple'. A graph in which all vertices can be numbered in such a way that there is precisely one edge that connects every two consecutive vertices and there are no other edges, is called a 'path', and the number of edges in that path is called the 'length'.

Vertices (or nodes), along with edges, are the elementary components of a graph. An undirected graph (where edges have no directions) consists of a set of vertices and a set of edges (unordered pairs of vertices), while a directed graph (where edges have directions) consists of a set of vertices and a set of arcs (ordered pairs of vertices). Undirected graphs without loops or multiple edges are called 'simple graphs'.

In a graph, an ordered pair of vertices is called an 'arc'. A graph in which all edges are ordered pairs is called a 'directed graph', or 'digraph'. A digraph $N = (X,A)$ is called a 'network', if X is a set of vertices (also called nodes), A is a set of arcs, and to each arc $a \in A$ a non-negative real number $c(a)$ is assigned which is called the capacity of arc a.

For any vertex $y \in X$, any arc of type (x,y) is called 'incoming', and every arc of type (y, z) is called 'outgoing'. A digraph is called 'weakly connected' if its underlying graph is connected and it is called 'strongly connected' if from each vertex there is a directed walk to all other vertices.

A simple adjacency between vertices occurs when there is exactly one edge between them. A graph is called 'connected' if any two vertices are connected by some path; otherwise it is known as 'disconnected'.

Let $G = (X,E)$ be a graph, $x,y \in X$: The distance from x to y, denoted by $d(x,y)$, is the length of the shortest (x,y)-path. If there is no such path in G, then $d(x,y) = \infty$. In this case, G is disconnected and x and y are in different components.

A graph, in which every pair of vertices has an edge, is called 'complete' because we can't add any new edge to the graph and still maintain a simple graph. The diameter of G denoted by $diam(G)$ is max $max_{x,y \in X}\ d(x,y)$, which means that it is the distance between the farthest vertices.

In a graph G, a walk is an alternating sequence of vertices and edges where every edge connects preceding and succeeding vertices in the sequence. It starts at a vertex, ends at a vertex and has the following form: $x_0,e_1,x_1,e_2, \ldots, e_k,x_k$.

A graph $G = (X,D)$ is called 'weighted' if each edge $D \in D$ is assigned a positive real number $w(D)$ called the weight of edge (D). In many practical applications, the weight represents a cost, distance, time, probability, capacity, resistance, etc.

A 'cycle' is a connected graph in which every vertex has degree equals 2 and denoted by C_n (where n is the number of vertices). If we have a graph $G = (X,E)$ and a vertex $x \in X$: The deletion of x from G means excluding x from set X and removing from E all edges of G that contain x. This introductory material to the basic concepts of graph theory was crucial due to its close connection to SNA.

4 Social Network Analysis

In this section, we are going to provide an introduction to SNA including its definition, a historical background, current uses, important network measures and finally a number of SNA modeling tools.

4.1 What is Social Network Analysis?

Social network analysis is a research approach that analyzes the structures of social networks, and the relationships among its members [9]. It aims at studying the structures and processes of networks and hence it focuses on the relationships between people and the syntheses formed through interactions, instead of the specific attributes of individuals (such as age, gender, occupation etc.). SNA can quantify these relationships through the use of graphical representations, where nodes represent individual units and ties represent relationships between individuals [10].

4.2 Historical Development of SNA

SNA, which describes how patterns of interpersonal relations are associated with various outcomes (such as emotions, cognition and behavior) draws from traditions of psychology, sociology, and other disciplines [11].

Much of contemporary SNA today builds on the bases established during two periods: the 1950s and 1970s which can be described as the golden age of social psychology. The advances in computing and communications led to better applying SNA. Growth is also attributed to peoples' engagement in online social activities and services [11].

The most apparent development that contributed to the building of network analysis has been the growth of interest among physicists to apply network ideas to social phenomena. One remarkable key area, highlighted in their work, was network dynamics, an area gained little attention, if at all, from sociologists working in network analysis [12]. Dynamic social network analysis investigates the behavior of social network overtime to detect recurrent patterns, formation and deformations of structures and discover anomalies and formalities of network communities. This branch is gaining more and more interest from researchers in the field [13].

It was in 1930s when network thinking emerged as a distinctive approach to social structures. Moreno was the first to introduce the idea of depicting social structures as network diagrams – called socoigrams – consisting of points and lines. In 1934, he put his famous publication *'Who Shall Survive?'* which was a turning point for the development of the field. He imagined society as a kind of physics that has its own social atoms and laws of social gravitation [14]. Sociometry, the name he gave to his approach, became a major field in education and social psychology [12].

In the 1940s and 1950s, work on social networks advanced at different perspectives, particularly in matrix algebra and graph theory, which made the discovery of emergent groups in network data possible. This development was followed by a wide spread of computer software products that benefited from the improvements achieved in computation [14].

By 1980, SNA has become an established field within the social sciences, with annual conferences being held every year and a professional organization called INSNA (International Network for Social Network Analysis) that follows up with all issues, developments and events in the field [14]. Also, a number of centers for network searching and training have opened worldwide and university courses are offered to students and practitioners [15].

4.3 SNA as an Emergent Analysis Tool

In the early days of SNA, research was targeted towards small groups and small social networks. Later on, social network analysis has benefited from the quantity and ubiquity of information in social media, web pages, sensors, mobile devices and others to apply structural analyses [13]. Today, it can be applied to large-scale social networks of up to millions of nodes that form online social networks [9].

One of the principal areas of application for SNA is the investigation of corporate power and interlocking directorships. A second major area is community detection which is about discovering groups of entities that show high connectivity within a group and low connectivity between different groups [13]. SNA has been also used in political and policy networks, social movements, criminality and terrorism, religious networks [12], consulting management, public health and crime/war fighting [14].

In SNA, four major areas of future development can be forecasted: (a) the exploration into the cultural context of social network models (b) the exploration of new methods for visualization (c) the use of statistical significance test, and finally (d) the development of models for longitudinal change [12].

4.4 SNA Metrics

SNA metrics fall into two types: some provide information about the individuals themselves and how they interact, and some provide information about the global structure of the social network [15]. Common concepts of social network analysis are as follows [9] and [16]:

- *Actors*: network members that could be individuals, organizations, events etc. Actors are usually connected in a network. However, if an actor is not connected to any actor in the network, then it is called 'isolate'. If two actors share more than one tie, we call this multiplicity [16].
- *Tie*: a link that connects two or more nodes in a graph.
- *Component*: it is a segment of a network where actors are connected to each other either directly or indirectly. A separate component is called 'isolate'.
- *Density*: the total number of relational ties divided by the possible total number of ties. It is one of the basic measures of network analysis and commonly used in social terminology.
- *Centrality measures*: there are a number of measures that are usually used to identify the important actors who are widely involved in relations with others in a network: (a) degree centrality: the sum of all actors who are directly connected to an actor, (b) closeness centrality: how far an actor is to all other actors in the network (c) betweenness centrality: the number of times that an actor connects between pairs of actors, who otherwise cannot reach each other.
- *Reachability*: it measures how reachable an actor is to the rest of actors in a network.
- *Cohesion*: the extent to which two actors are directly connected, by cohesive bonds, to each other.

4.5 SNA Modeling Tools

The advancements in computing power provided a number of new models and methods, never found before, to analyze networks [17].

SNA methods can be conducted by a number of specific pieces of software. The most widespread for a long period of time has been UCINET which was developed at the University of California, Irvine and includes multiple analytical tools to efficiently explore and measure social networks. Pajek, another software package, developed at the University of Ljubljana, has been used to handle large data sets and large-scale networks with the assistance of methods of graphical representation [12]. Other tools include StOCNET that provides statistical methods focusing on probabilistic models, NetDraw and Gephi that allow for good network visualization, and last but not least; E-Net and KeyPlayer that specialize in ego-network analysis and key-player operations [18].

The software that we used in our study for analysis was Pajek. This preference came due to a number of reasons that will be summarized in sub-section 5.3. In the next section, we will dive into the Flight MH370 community structure.

5 Flight Mh370 Community Structure

5.1 Data Source

Our dataset has been collected during March and April 2014 from online sources (such as YAHOO! News Malaysia, CNN.com, the Economic Times, The New Indian Express, India Today and the Daily Express). Information gathering was applied by surfing hundreds of WebPages that addressed Flight MH370 from when it was announced missing by the Malaysian Airlines on the 8th of March 2014 until late of April when the international efforts aiming to find the wreckage of the missing plane declined.

Online sources have addressed this topic from different perspectives such as (a) telling the story of the missing aircraft, how it disappeared from radar and the possible spots of its current location, (b) providing short personal profiles for some of the passengers onboard, (c) showing how the relatives of the passengers have been dealing with this issue and (d) picturing the efforts made by the international society to locate the missing plane.

These sources obtained their information from interviews with the Flight MH370 passengers families, focusing on the way their relatives lived and the words they uttered just before the plane took off. Investigating the places where the passengers had lived, worked or the people who used to mix with was the second major information source for these websites, in addition to the information revealed by official channels, which were very concise due to different political and security considerations.

Since the data, which the current study is based on, is based on online sources, such as journals, blogs and official websites, some information in one place may not be the same as in another place. However, we tried our best to rely on information with better frequency. Therefore, it’s noteworthy to say that the

dataset used in our study is not final and is subjected to further updates especially when new information about the passengers becomes available or new facts are discovered.

5.2 Data Description

The data are in .xls format and can be downloaded from the first author's website [(1)]. The data represent the known relationships of people before joining the flight. For example, in the *Artists Group,* which incorporated painters and calligraphers from China, each one is connected to the rest of the group members (since they were in an exhibition in Kuala Lumpur and certainly knew each other). However, for the *Singapore-Career Group*, some travelers are confirmed to know each other but it is not confirmed that all the individuals who worked in Singapore were acquainted to each other. Thus, all of these nodes are connected to node 'Singapore' but only three of them are directly connected (since they already knew each other). So, we added the vertex 'Singapore' to connect the Chinese passengers who were reported to work in Singapore. In addition, all the passengers are connected to a main node called 'Flight' which expresses that all of them were present in the same environment (even if they were alone travelers) which means that they were all connected to one network: the aircraft network.

For convenience, we preferred, throughout our study, to use labels when referring to the flight passengers. For example,

$P_1, P_2, P_3, ..., P_n$ correspond to passenger numbers in our dataset.

We had to reformulate Flight MH370 passenger names and put them in more recognizable formats. For example, we removed unnecessary word capitalizations and suffixed the second name of each passenger with his/her age to avoid name confusions (for instance, 'Zhang Yan' is the name of two different passengers; P_{225} and P_{226}). We preferred, for simplicity purpose, to use only the first and second name of each passenger followed by age. So, for example, P_9 in our dataset is Bibinazli Mohd Hassim, 62 years old from Malaysia. Her corresponding name label is Bibinazli Mohd (62).

However, some information in the list of passengers revealed by the Malaysian Airlines still needs to be reconsidered. For example: regarding passenger P_{106}, all sources mentioned that her age was 32 while the manifest says that her age was 52 year old. Also, according to different sources, passenger P_{72}, 2 years old, was the youngest child in the trip and he had the American nationality. However his name is not on the list! (We can solve this issue by considering that he was attached to his mother, passenger P_{225}, 36 years old, in her passport).

[1] http://www.themesopotamian.com

5.3 *Pajek* Software

The software that we use in our analysis is *Pajek*, which is a program for analysis and visualization of large networks [20]. *Pajek* provides efficient methods for clustering social network data and has proven efficient in previous studies [19]. It is free software that can be downloaded from its official website and has a simple graphical user interface along with powerful visualization tools and several data analytic algorithms. It has a well-organized user's manual and free datasets that can be used for testing. It can deal with multi-networks and different types of networks at the same time. Also, *Pajek* incorporates powerful statistical analysis tools, such as R and SPSS [15]. The version that we used for community detection is 3.15 as of March 2014 [(2)]. Community detection in this version of the software is based on the *Multi-Level Coarsening and Multi-Level Refinement* algorithm [21] which uses modularity for graph clustering.

5.4 Analysis of Flight MH370 Community Structure

Before their last trip, passengers of the missing aircraft gathered in various groups based on social, friendship or work-related factors. For example, there were married couples, entire families, crew members who should provide the furthermost services to passengers and ensure safe arrival, company employees who were heading/departing work sites, people attending special events, etc.

The passengers were a mix of people hailed from 14 countries. However, it was learned later that two people on the manifest, an Italian and an Austrian, were not onboard. Most of the 227 passengers were Chinese, while the 12 missing crew members were Malaysian. The rest of passengers were from France, United States, the Netherlands, Indonesia, Australia, the, New Zealand, Ukraine, Canada, Russia and India. Some of those onboard were heading home while others just making a stopover. To some, it was only their first trip abroad. Due to shortage in the data available, no demographic information was used in our analysis. Let’s take a look at some of the overall flight network statistics:

[2] http://mrvar.fdv.uni-lj.si/pajek/

Table 1: Features of Flight MH370 Community Structure

Metric	Value
Graph Type	Undirected
Number of vertices (Dimension)	241
No. of edges	1563
No. of Loops	5
Network Density	0.05373530
Average Degree	12.97095436
Connected Components	1
Single-Vertex Connected Components	0
Maximum Vertices in a Connected Component	241
longest shortest path	Between nodes 11 and 194

It is an undirected network, which means that the links are symmetric. So, if a person 'A' knows/interacts with person 'B', then person 'B' also knows/interacts with person 'A'. (We can easily transform an undirected graph into directed by giving directions to all nodes).

The total number of edges is 1563 and of nodes (network dimension) is 241. As we mentioned before, we added two additional nodes to the original number of nodes (which is 230); one for connecting all passengers to a single network and the other to group the individuals who worked in Singapore in one network.

The network density is 0.05373530. The importance of density is that it provides us with some important network features such as the speed at which information diffuses among nodes (passengers), and the levels of social capital or social constraint of nodes [15].

The average degree is 12.97095436, which reflects the cohesive nature of the network. This measure is used to describe the structural cohesion of a network [15]. It is calculated by adding together all the average degrees and then divided by the total number of nodes in the network (for undirected networks, the number of edges should be calculated twice).

The number of components equals one meaning that we have zero disconnected components. This is because the aircraft network is considered one integral network with no isolates; any individual on the plane can interact with any other individual except in specific cases such as for pilots, children, disabled, etc.

Searching for the network diameter- the longest shortest path, reveals a path between passenger P_{11}, 54, Australia and passenger P_{193}, 27, China. This measure

is important for indexing the extensiveness of the network, i.e. how far the two furthest nodes are from each other [15].

5.5 Visual Representation of the Passenger Community

The visual analysis of the passenger community will help us study the mesoscopic and macroscopic features of this community. In Fig. 1, below, we can see that there are three larger components (in circles) within the MH370 network: the largest component (giant component) is the Artists component (29 vertices). The other two larger components are: the Freescale Semiconductor (14 nodes) and the Aircraft Crew (12 nodes).

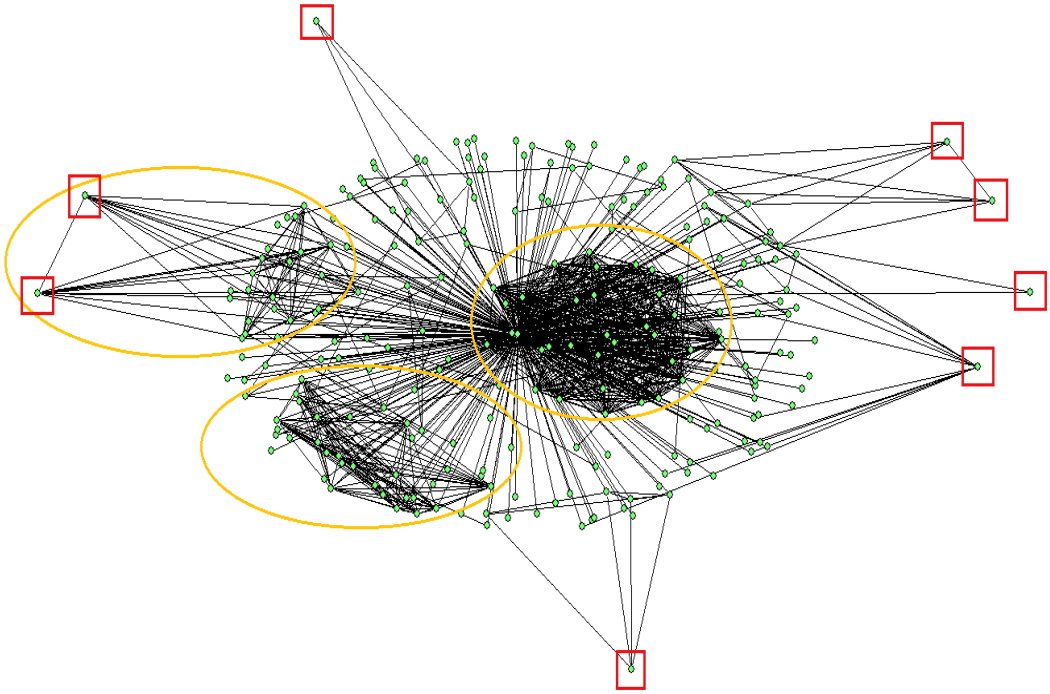

Fig. 1 Visual Representation of Flight MH370 Community Structure

In this air travel community, we notice the existence of a large number of communities whose size is small compared to a small number of communities whose size is big. This holds true also for other types of networks such as information, communication and biological [22][24].

Also, we can see eight peripheral nodes enclosed in boxes. These nodes are not directly connected to node 'Flight', which means that they are also not directly tied to the rest of nodes in the network unless through their own sub-network members. These nodes are: (a) the two aircraft pilots; P_{209} and P_{31}. Usually, pilots can only be reached through the rest of other crew members which means, in our

case, only through ten other nodes in the network, (b) five toddlers who could not interact normally with others unless through their family members and (c) an old visually-impaired woman, passenger P_{11}, 54 years old, who can only be interacted with through either her husband, passenger P_{130}, or one of their two friends; passenger P_{107} and passenger P_{131}. The details of the three larger groups are as follows:

1. *Artists*: A group of Chinese individuals were returning back after attending a cultural exhibition in Kuala Lumpur themed "*Chinese Dream: Red and Green Painting*". The artists, who came from Beijing, Shandong, Jiangsu, Sichuan and Xinjiang, comprised famous figures such as China Calligraphy Artists Association vice-chairman P_{111}, P_{94}, and P_{169}. The Association chairman said that thirteen of the artists on the delegation were members of the association. Some of the overall network statistics are shown in Table 2:

Table 2 Features of the Artists Group Community Structure

Metric	Value
Number of vertices (Dimension)	30
No. of edges	841
No. of Loops	0
Network Density	1.86888889
Average Degree	56.06666667
Connected Components	1

The group members were: P_2, P_{228}, P_6, P_{13}, P_{28}, P_{33}, P_{37}, P_{49}, P_{73}, P_{77}, P_{84}, P_{94}, P_{96}, P_{98}, P_{105}, P_{108}, P_{111}, P_{140}, P_{169}, P_{188}, P_{197}, P_{200}, P_{215}, P_{222}, P_{234}, P_{236}, P_{237}, P_{238}, and P_{239}. Network visualization is found in Fig. 2, below. We can see that all the nodes are connected to the main node 'Flight' along with their own connections.

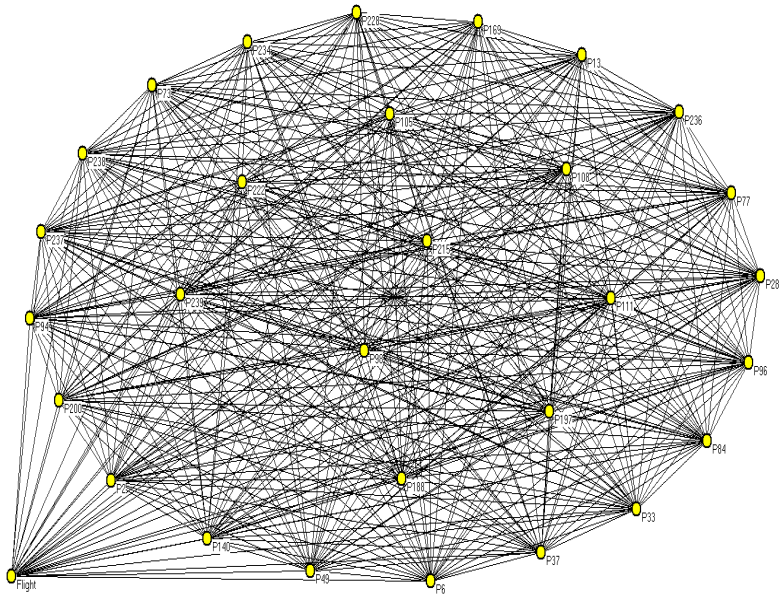

Fig. 2 The 'Artists Group' Community Structure

2. *Freescale Semiconductor*: 20 top management employees of Freescale Semiconductor, confirmed to be 12 Malaysians and 8 Chinese nationals were supposed to attend a course in China for one month (although some sources mentioned that they were even more than 20). The Freescale Semiconductor Company develops sensors, microprocessors, and stand-alone semiconductors. Just the day before the plane disappeared, the company had launched a new electronic warfare gadget for military radar systems. The company declined to reveal the names of its employees who were onboard. Some overall network measures can be found in Table 3.

Table 3 Features of the Group of the Freescale Semiconductor Community Structure

Metric	Value
Number of vertices (Dimension)	16
No. of edges	225
No. of Loops	1
Network Density	1.75390625
Average Degree	28.12500000
Connected Components	1

Even though this group consisted of 20 individuals, we were only able to recognize some of them: P_{208}, P_{43}, P_{44}, P_{52}, P_{54}, P_{70}, P_{71}, P_{115}, P_{120}, P_{132}, P_{141}, P_{153}, P_{158}, P_{161}, and P_{202}.

Fig. 3 shows how the Freescale group members are connected to the main node 'Flight'.

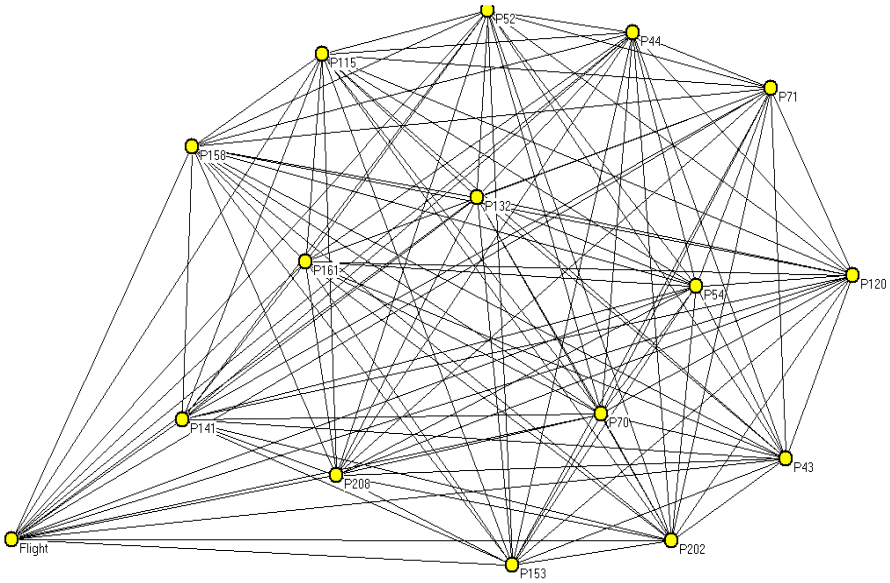

Fig. 3 The 'Freescale Semiconductor' Community Structure

3. *Aircraft Crew*: 12 of Malaysian aircraft crew members (two pilots and ten attendants) where onboard. Let's have a look at some of the network overall measures as in Table 4.

Table 4 Features of the Aircraft Crew Group Community Structure

Metric	Value
Number of vertices (Dimension)	13
No. of edges	142
No. of Loops	0
Network Density	1.68047337
Average Degree	21.84615385
Connected Components	1

The aircraft crew members were: P_{116}, P_{209}, P_3, P_{31}, P_{36}, P_{41}, P_{65}, P_{112}, P_{124}, P_{146}, P_{148}, and P_{162}. Fig. 4 shows the graphical representation of this group connected to vertex 'Flight'.

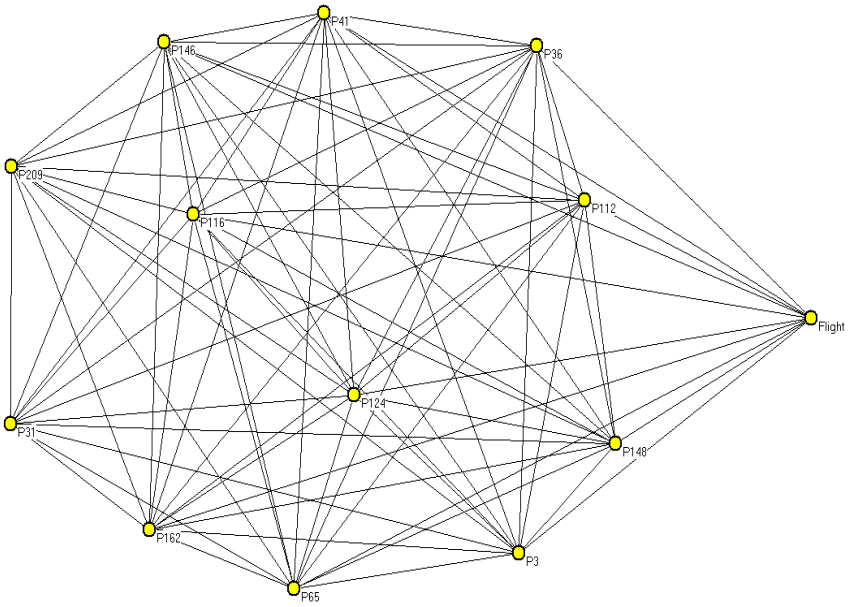

Fig. 4 The 'Aircraft Crew' Community Structure

From the previous figures, we can see that the three larger groups are fully-connected, while they are connected to the rest of other sub-communities by weak ties through node 'Flight'. This reflects the importance of the weak ties that connect communities that are otherwise become unreachable [23]. These ties are significant in that they enable users to get in touch with each other and efficiently disseminate information.

It emerges that, in our case, not only the smaller communities are well connected but also the three larger communities as well.

Also, edges connect communities of smaller size, and also connect communities of smaller size to those of larger size. However, edges in the first case are more common than they are in the second case, since the number of smaller-size communities is much greater than the number of bigger-size communities. This will help us understand the degree at which links connect communities of different size. Smaller components include:

- *Six-People Family from China*: P_{72}, P_{109}, P_{110}, P_{117}, P_{189}, and P_{225}.
- *Five-People Family from China*: P_{21}, P_{62}, P_{63}, P_{170}, and P_{171}.
- *Four-People Family from Malaysia*: P_9, P_{24}, P_{106}, and P_{119}.
- *French Family*: P_1, P_{45}, P_{69} and P_{232}.
- *Australian Family*: P_{11}, P_{107}, P_{130}, and P_{131}.
- *Nepal-Trip Group*: 9 out of 23 elderly people, who were travelling in Nepal together, were on Flight MH370. Most of those on the tour were retired Chinese academics. On their flight home, they were split into two divisions - one division returning from Kuala Lumpur and the other from Guangzhou. We were able to count only 4 of them: P_{17}, P_{27}, P_{137}, and P_{199}.
- *Singapore-Career Group*: According to Chinese media reports, at least nine Chinese workers, worked in Singapore, mainly in construction, were onboard (not necessarily were friends). We were able to recognize the following passengers; P_7, P_{23}, P_{32}, P_{93}, P_{90}, P_{139}, P_{178}, P_{179}, P_{201}, P_{204}, and P_{230}.
- More than a hundred Buddhists (not necessarily friends) were on flight MH370, returning from Kuala Lumpur to Beijing after taking part in a Buddhism conference on 2nd of March as more than 30,000 people from around the world attended this religious festival.

- Toddlers: five toddlers were onboard: P_{50}, P_{72}, P_{117}, P_{170}, and P_{233}.
- The rest of passengers are mainly married couples and alone travelers. However, we still have about 40 other passengers with no reachable information.

6 Conclusions and Challenges

The analysis of Flight MH370 community structure focused on investigating the mesoscopic and macroscopic features of the passenger community in regard to some characteristics of the network (such as network density, diameter, average degree…) and how these features may reflect organization patterns in the network.

We started our study by giving an introduction to the different community detection paradigms followed by examples from related work. Then, we moved to manifest the basic concepts in graph theory; the cornerstone of social network analysis followed by an introduction to social network analysis which included: a historical background, the fields where researchers use social network analysis in order to solve different problems, the main metrics that are usually used during analysis and the modeling tools that are used to visualize social networks. In section five, we addressed Flight MH370 community structure: data source, data description, the software that we used in our analysis, and finally the community structure analysis, where we studied the overall properties of the network and the different groups that form the aircraft network.

Our findings show that there were three larger community structures onboard: the *Artists Group*, consisting of 29 members, the *Freescale Semiconductor* group, consisting of 15 members and the *Aircraft Crew* consisting of 12 members. Other smaller groups include six-people and five-people families from China, four-people family from Malaysia, a French family, an Australian family and others.

The analysis also gave us 8 nodes (two pilots, five toddlers and one disabled elderly woman) that are not directly connected to node 'Flight' which means they were reachable only through some other network nodes.

We also noticed the significance of *Weak Ties* which connect communities that are otherwise become unreachable (in our case the larger ones to the smaller ones). Moreover, not only the smaller communities of Flight 370 show high connectivity, but also the larger communities. The findings of this study reflect the nature of air travels and can be further applied to similar cases.

The biggest issue that we encountered in this study was the lack of information available and at sometimes, the disagreement between different sources which makes it difficult for us to make a choice. Thus, and since we were concerned with the analysis of the publically accessible profiles, our results only provides an incomplete picture of that flight passenger communities, which could be slightly

different if we were to reach all passenger profiles (at the time of writing, we were missing the profiles of 40 passengers).

Also, we gathered our information with the help of online sources (governmental websites, newspapers and blogs). However, it's beyond dispute that the credibility of online sources is becoming more and more of an issue and their status as accurate sources is not fully established, especially when living the age of big data, which makes it very difficult to find a trustable information source.

Acknowledgements

This work is supported by Universiti Teknologi Malaysia under the Flagship Project: *UTM E-Learning Big Data Analytics*. The authors would like to thanks Research Management Centre (RMC), Universiti Teknologi Malaysia (UTM) for the support in R & D, and *Soft Computing Research Group* (SCRG) for the inspiration in making this study a success. The authors would also like to thank the anonymous reviewers who have contributed enormously to this work.

References

[1] Fortunato, S., "Community detection in graphs," *Phys Rep* 486(3-5), (2010), pp. 75-174.

[2] Newman, M. E. J., "Detecting community structure in networks," *Eur Phys J B* 38, (2004), pp. 321–330.

[3] Newman, M. E. J. and Girvan, M., "Finding and evaluating community structure in networks," *Phys Rev E*, Vol. 69, No. 2: 026113, 2004.

[4] Traud, A., Kelsic, E., Mucha, P., and Porter, M., "Comparing community structure to characteristics in online collegiate social networks" *SIAM Rev* 53, (2011), pp. 526-546.

[5] Lancichinetti A., Kivela M., Saramaki J. and Fortunato, S., "Characterizing the Community Structure of Complex Networks," *PLoS ONE*, Vol. 5, No. 8, e11976, 2010.

[6] Tibély, G., Kovanen, L., Karsai, M., Kaski, K., Kertész, J., and Saramäki, J., "Communities and beyond: mesoscopic analysis of a large social network with complementary methods," *Phys Rev E*, Vol. 83, No. 5:056125, 2011.

[7] Ahn, Y., Bagrow, J. and Lehmann, S., "Link communities reveal multiscale complexity in networks," *Nature* 466(7307), (2010), pp. 761-764.

[8] Grabowicz, P., Ramasco, J., Moro, E., Pujol, J. and Eguiluz, V., "Social features of online networks: the strength of intermediary ties in online social media," *PLoS ONE* 7:e29358, 2012.

[9] Raju, E. and Sravanthi, K., "Analysis of Social Networks Using the Techniques of Web Mining," *International Journal of Advanced Research in Computer Science and Software Engineering*, Vol. 2, Issue 10, (2012), pp. 443-450.

[10] Kothari, A., Hamel, N., MacDonlad, J. A, Meyer, M., Cohen, B. and

Bonnenfant, D., "Exploring Community Collaborations: Social Network Analysis as a Reflective Tool for Public Health," *Syst Pract Action* Res 27, (2014), pp. 123–137.

[11] Burt, R. S., Kilduff, M. and Tasselli, S., "Social Network Analysis: Foundations and Frontiers on Advantage," *Annu. Rev. Psychol*. 64, (2013), pp. 527–547.

[12] Scott, J., "Social network analysis: developments, advances, and prospects," *SOCNET*, Vol. 1, (2011), pp. 21–26.

[13] Campbell, W. M., Dagli, C. K. and Weinstein, C. J., "Social Network Analysis with Content and Graphs," *Lincoln Laboratory Journal*, Vol. 20, No. 1, (2013), pp. 62-81.

[14] Borgatti, S. P., Mehra, A., Brass, D. J. and Labianca, G., "Network Analysis in the Social Sciences," *Science*, Vol. 323, (2009), pp. 892-895.

[15] Kadry, S. and Al-Taie, Mohammed Z., "*Social Network Analysis: An Introduction with an Extensive Implementation to a Large-Scale Online Network Using Pajek,*" Bentham Science Publishers, (2014).

[16] Hawe, P., Webster, C. and Shiell, A., "A glossary of terms for navigating the field of social network analysis," *J Epidemiol Community Health*, Vol. 58, (2004), pp. 971–975.

[17] O'Malley, A. J. and Marsden, P. V., "The Analysis of Social Networks," *Health Serv Outcomes Res Methodol*, Vol. 8, No. 4, (2008), pp. 222–269.

[18] Apostolato, I. A., "An overview of Software Applications for Social Network Analysis", *International Review of Social Research*, Vol. 3, Issue 3, (2013), pp. 71-77.

[19] Al-Taie, Mohammed Z. and Kadry, S., "Applying Social Network Analysis to Analyze a Web-Based Community," *International Journal of Advanced Computer Science and Applications*, Vol. 3, No.2, (2012), pp. 29-41.

[20] Mrvar, A. and Batagelj, V., "Pajek and Pajek-XXL Programs for Analysis and Visualization of Very large Networks" *Reference Manual*, (2014).

[21] Rotta, R. and Noack, A., "Multilevel local search algorithms for modularity Clustering," *ACM J. Exp. Algor*. (2011), Vol. 16(2).

[22] Ferrara, E., "A large-scale community structure analysis in Facebook", *EPJ Data Science*, (2012), pp. 1:9.

[23] Granovetter, M., "The strength of weak ties," *Am J Sociol* 78, (1973), pp. 1360-1380.

[24] Sarina Sulaiman, Siti Mariyam Shamsuddin and Ajith Abraham, "Implementation of Social Network Analysis for Web Cache Content Mining Visualization," *Computational Social Networks: Mining and Visualization*, Springer-Verlag London (2012), pp. 345-376.