Chapter Three
Research Methodology

> The natural contexts of new media may limit how faithfully traditional research designs and methods may be applied… the nature of new media themselves may create limitations, as well as new opportunities (Rice and Rogers, 1984).

1. Introduction

As outlined in Chapter One, it is a primary aim of this study to investigate the structure of the *Africa* category of the *Departure Lounge* on the *Thorn Tree*. In this social network, ties exist between actors because of their differentiated participation in discussion threads, manifesting in patterns. This suggests that a structural approach could throw light on certain aspects such as the nature of social ties. A CD-Rom at the back of this thesis contains the messages that are included in this dataset. It also contains other information relevant to the computer methods followed in this study, such as programming relevant to acquiring data.

In this chapter, methodological aspects are outlined. An explanation of the choice of methodology, namely social network analysis is given. The sub-section below explains the reason for choosing network analysis, but also places network research as applied to cyberculture into perspective. However, by incorporating certain standard statistical tools, it is possible to reveal a number of aspects regarding the data. This is used, *inter alia*, to verify the consistency of the data. Entitled “descriptive statistical analysis” it considers measurements such as duration

of threads, number of replies, ratio of replies versus views, message length and survival analysis amongst others.

Identifying and describing concepts related to the methodology of social network analysis used in this study are also covered in this chapter. Like Internet studies, network analysis is characterised by its own concepts and terms; likewise, with regard to the case study, terms to refer to its structural properties such as branches and categories are specific to the *Thorn Tree*; a sub-section below outlines these and other relevant concepts.

Notably, only messages on the *Africa* category of a branch called the *Departure Lounge* are used for analysis and not all messages on all branches of the whole *Thorn Tree*. This raises questions such as how a decision was reached with regard to the population, boundary specification and sampling. In addition, what measurements associated with network analysis were considered for analysis purposes? These and other methodological questions are attended to in a sub-section that deals with issues of measurement. A separate section deals with the methods used in this study, for example how network analysis techniques were applied and which calculations were done to obtain specific measurements.

2. Studying online communities: choosing a methodology

2.1 Background

While Jones (1999: xiii) states that Internet research is not easy, he also emphasises that the Internet should not be seen in isolation from the “off-line” world that has created it. Indeed, as mentioned in Chapter Two, the online world mirrors

what happens in the real world. Wakeford claims that one of the most confusing aspects of doing research about the Web, as with any other media form, is that it can be understood at many levels. Wakeford furthermore points out that web pages are a number of things simultaneously: computer code, cultural representations, material objects for consumption and the fruits of skilled labour. Studying the Internet requires clarity with regard to the ways in which it can be done. Wakeford also notes that there currently is no standard technique for studying the Web. The Internet, cyberculture and other distinct features can be included here. This holds true for communication studies and allied social science disciplines (2000: 31).

The question whether we care to study cyberculture could very well be dismissed (Jones, 1999: ix). Yet, as a medium of communication that intersects in new and remarkable ways with everyday life, understanding the Internet and its effects can only be achieved through thorough research. In the words of Wakeford,

> *we cannot presume in advance that the cultural significance of the Web can be read off its current popularity. The relationship between the Web and the rest of the social world cannot be presumed, but must be investigated* (2000: 31).

The question remains how to go about studying the Internet and its related social impact. Wakeford claims that in the absence of any clear standard technique, studying the WWW and Internet culture has become a case of

plundering existing research for emerging methodological ideas which have been developed in the course of diverse research projects, and weighing up whether they can be used or adapted for our own study *(2000: 31).*

Jones questions whether the changes in methods used to study the Internet's convergence and influence on modern life are enough. In accordance with Wakeford who calls for new methods, Jones too states:

…applying existing theories and methods to the study of Internet-related phenomena is not a satisfactory way to build our knowledge of the Internet as social medium *(1999: x).*

This approach excludes methods that have been traditionally used to study other media and social phenomena and are now being used to study the Internet also. When considering which methodological frameworks scholars have at their disposal to study the WWW and cyberculture, cognisance has to be taken of the fact that what is considered a legitimate methodology is itself in flux.

Questions of research design, sample or participant selection, choice of website or Internet-related phenomena to study, methods of data collection and analysis, ethical practice and the use of theoretical frameworks are just as relevant in the study of the Internet and cyberculture and "cannot be sidestepped however 'virtual' the data collection" (Wakeford, 2000: 33). However, from a methodological point of view, what is different about the study of the Internet? The answer lies largely in the ease with which data about Internet users can be captured, the manner in which researchers can become part of their object of study, and the speed and (seemingly) accuracy with which large amounts of statistical data can be produced. Yet, these possibilities also bring about their own sets of methodological problems if handled without the necessary scholarly vigour.

Ongoing research and continued academic interest have brought about a refinement of existing research techniques, such as ethnography, anthropology and textual analysis amongst others. On a technical level and considering the manner in which the Internet operates, a revival in network analysis and a renewed interest in this research technique for the study of physical computer networks have implications for the study of information flow and connectivity among members of computer-mediated social networks too.

Computer-mediated social networks can be studied from a number of vantage points, involving different methodologies. This depends largely on the nature of the inquiry, the goal of the research and the study field involved, amongst an array of other influencing factors. Examples are drawn from communication studies, sociology, anthropology and psychology. However, Wellman points out that online communities are foremost social networks, which emphasises the relevance and applicability of using network analysis (1997: 179; Friedkin, 1982; Garton et al, 1999). Studying social networks involves network theory and network analysis. Applied to a specific field, namely leisure choices and travel information exchange, Stokowski (1988) uses a structural approach to explain the importance of connectivity and place in a network. In this study, travel information exchange takes place in an electronic environment across the Internet. The following sub-section outlines the premises of this study and the choice of methodology.

2.2 Premises of this study and the choice of method

With the analytical tools available through network analysis, this study investigates a number of premises about the *Africa* category:

- Messages attract unequal numbers of replies resulting in clear patterns of communication flow. Using graphs, the structure and pattern become visible.
- Like human interaction in the physical world, some members (called actors in network terms) are more active than others. Using an asymetric matrix, network analytical calculations reveal information about the dynamics of the network, such as the size of the network, levels of connectivity (degree), density, centrality and direction of ties.
- This premise is tested that in a large network such as this, moreover a computer-mediated network, ties are weak but it does not implicate inefficiency in communication flow.
- Threads differ markedly in the number of messages (replies) they contain. The threads with the least and the most responses are calculated respectively in order to obtain the mean, average and mode. The premise is held that this could be indicative of the levels of information exchange among actors and the extent to which individual actors are willing to exchange information.
- With the aid of an asymmetrical matrix, the direction of communication flow between actors can be ascertained. Against the background of social relations, willingness to reply and not only post requests considers the extent to which members are "sources" (that is, they have a tendency to send more than to receive), "sinks" (that is, they have a tendency to receive more than send) or "transmitters" that both send and receive, but to different others.
- The lifespan of a thread (survival analysis) can be calculated and tests the premise that “conversation dies” once information requests have been fulfilled. This tests the premise that *Thorn Tree* members on the *Africa* category treat this electronic discussion board in a similar fashion as real world travellers do. They exchange information and move on.

2.3 What is social network analysis?

According to Emirbayer and Goodwin (1994: 1413), network analysis proceeds from certain basic theoretical presuppositions and premises that are acceptable to most, if not all, of its practitioners. It holds to a set of implicit assumptions about fundamental issues in sociological analyses. Examples they refer to include relationships between an individual and society, the relationship between "micro" and "macro" and the structuring of social action by objective and "supra-individual" patterns of social relationships. The point of departure for network analysis is what they call the "anticategorical imperative". This imperative rejects all attempts to explain human behaviour or social processes solely in terms of the categorical attributes of actors, whether individuals or collectives. This is supportive of the notion that network analysis rejects explanations of social behaviour as the result of individuals' common possession of attributes and norms – instead, it is resultant from their involvement in structured relations.

Social network analysis is indeed a distinct research perspective within the social and behavioural sciences because it is based on an assumption of the importance of relationships among interacting units. Furthermore, the social network perspective encompasses theories, models and applications that are expressed in terms of relational concepts or processes. Relations defined by linkages among units are a fundamental component of network theories (Garton et al, 1999: 78).

Growing interest in the network perspective, together with the increased use of network analysis have resulted in a consensus about the central principles underlying the network perspective. These principles distinguish social network analysis from other research approaches. In addition to the use of relational concepts, the following are important. First, actors and their actions are viewed as

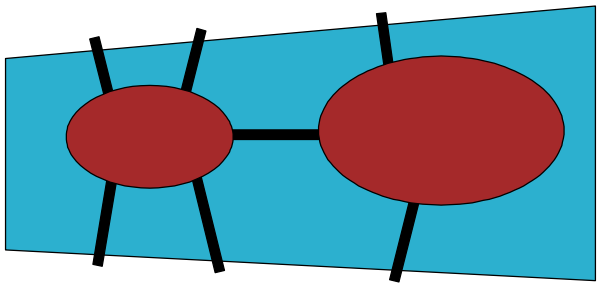
interdependent rather than independent, autonomous units. Second, relational ties (linkages) between actors are channels for transfer or "flow". Thirdly, network models focussing on individuals view the network structural environment as providing opportunities for or constraints on individual action. Lastly, network models conceptualise structure (social, economic, political) as lasting patterns of relations among actors (Wasserman and Faust, 1994: 4). Actors can be almost anything, including organisations, people and so forth.

Of critical importance for the development of methods for social network analysis is the fact that the unit of analysis in network analysis is not the individual but an entity consisting of a collection of actors and the linkages among them. The methods of network analysis provide explicit formal statements and measures of social structural properties that might otherwise be defined only in metaphorical terms. Phrases such as webs of relationships, closely knit networks of relations, social role, social position, group, clique, popularity, isolation, prestige, prominence and so on are given mathematical definitions (Wasserman and Faust, 1994: 17). Network methods focus on dyads (two actors and their ties), triads (three actors and their ties) or larger systems (subgroups of individuals or entire networks). At the simplest level, a tie is nothing more than an instance of a social relation between actors. Ties are sources of social capital. Depending on the research focus, ties can be indicative of relationships and roles, cognitive/perceptual and affective aspects, types of interactions, and types of affiliation.

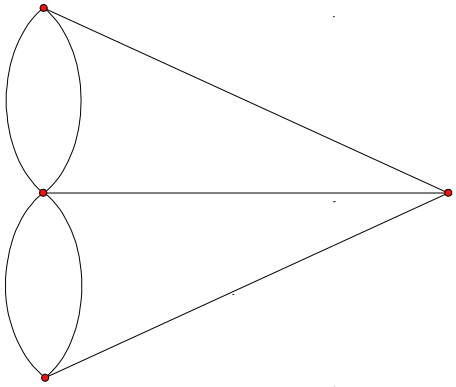
Network analysis is known for its use of graphs and matrices. While graph theory and matrix operations have served as the foundations of many concepts in the analysis of social networks (Wasserman and Faust, 1994: 92), graphs have proved useful to present information, since they graphically display nodes and the links between them. However, in a large dataset when there are many actors and/or

many kinds of relations, graphs can become too visually complicated, inhibiting interpretation.

Scholars consider the first paper to appear on graph theory to be that of the Swiss mathematician Leonhard Euler and appeared in 1736. In this paper he solved the problem of the Königsberg bridges. Königsberg, now called Kaliningrad, is situated on the banks of the river Pregel (Pregolya) and on two islands in the river. The land masses are connected by seven bridges and the people at the time discussed whether a round trip would be possible that crossed each bridge without crossing any bridge twice. A rough sketch of Königberg is shown in the following figure.



Euler reduced the problem to its essentials by representing each land mass by a dot and each bridge by a line that connected the dots. The diagram that Euler drew is shown in the following figure.



The diagram drawn by Euler is a generalisation and therefore applicable to similar problems with relationships that could be drawn in this manner. A point and line diagram of this nature consists of primary concepts and relationship between the primary concepts.

The primary concepts are first that there is some non-empty, finite collection of points, P. Second, there is a collection of lines, L. The relationships between the primary concepts are: First, every line of L meets two and only two points of P. Second, between every pair of distinct points of P there is at least one path consisting of one or more lines.

With these humble beginnings, graph theory grew to a vast, respected and interesting topic in mathematics. The mathematical theory and presentation of concepts associated with Graph Theory is somewhat daunting and has, perhaps, retarded its acceptance and use in other fields. A typical impression of graph theory may be obtained by viewing the electronic copy of the book *Graph Theory* by Reinhard Diestel that is contained on the CD-Rom inserted at the back of this thesis.

In an assessment of the visualisation of networks, Freeman claims that for most reviewers, visualisation plays an important part in the development of almost every field of science. He furthermore holds the view that this is certainly true of social network analysis where, from the beginning, visual images (particularly those grounded in graph theory) have been central to its success. Although network analysis has produced a number of computer tools to aid in the analysis of data (Freeman, 1988), it has been slower to develop computer tools designed to produce visual images. However, the most widely used tools are *Netdraw* and *Krackplot* that work in conjunction with *UCINET* (Garton et al, 1999: 96).

Although graphical representations are possible and indeed handled by the software application mentioned above, it is necessary to represent information in the form of a matrix. Representing data in this way eases the application of mathematical and computer-based statistical tools to summarise and find patterns. In the most common type of matrix, the number of rows and columns are both equal to the number of actors. The elements represent the ties between the actors. Binary notation simplifies and increases the usefulness of matrix data. A one (1) signifies the presence of a tie; if there is none, a zero (0) is entered. Other options to denote more information about the nature of ties are possible, namely relational qualifications. Two properties are important for understanding relation measurement and for categorising appropriate methods, i.e. directional relations and dichotomous relations.

In directional relations there is an origin and a destination for the tie, i.e. from one actor to the other in the pair. Often, direction is indicated with an arrow. Undirectional relations do not indicate the direction of a tie. Dichotomous relations are coded as either present or absent for each pair of actors; indicated in binary form, 1=present, 0=absent. Valued relations, on the other hand, can take on a

range of values, i.e. strength, intensity, or frequency of the tie between each pair of actors (Wasserman and Faust, 1994: 44).

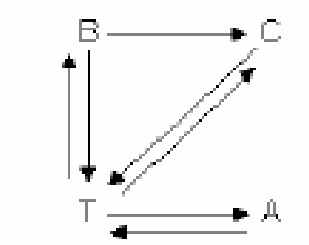
A simple matrix consists of a rectangular arrangement of a set of elements, although matrices that are more complicated can display more than two dimensions. The number of rows of elements and columns of elements describe the size of the rectangle. A "3 by 6" matrix has three rows and six columns; an "i by j" matrix has i rows and j columns. Elements of a matrix are identified by their "addresses". Element 1,1 is the entry in the first row and first column; element 13,2 is in the 13th row and the 2nd element of that row. Matrices are often represented as arrays of elements surrounded by vertical lines at their left and right, or square brackets at the left and right.

Like most other social network research outputs, this study dispenses with the mathematical convention of naming the matrix. Instead, it simply shows the data as an array of rows and columns. In some cases, matrices have labels. The labels are not part of the matrix, but are included for clarity of presentation.

The matrix below for example, is a 4 by 4 matrix, with additional labels. This kind of matrix is a starting point for almost all network analyses and is called an "adjacency matrix" because it represents who is next to or adjacent to whom in the social space mapped by the measured relations (Hanneman, ca1999: 19).

	Bob	Carol	Ted	Alice
Bob	—	1	0	0
Carol	1	—	1	0
Ted	1	1	—	1
Alice	0	0	1	—

By convention, in a directed graph, the sender of a tie is the row and the target of the tie is the column. The directed graph of friendship choices in this example among Bob and his friends looks like this:



From the matrix outlining the links between Bob and his friends, a host of characteristics, detailing the nature of their relations and the network itself can be deduced using relevant network statistical calculations. These and other relevant aspects are outlined in the sections below.

3. Methodological concepts

3.1 Network analysis: concepts and the area of study

In this section, reference is made to concepts that are closely associated with social network analysis and computer-mediated communication (CMC). These broad categories are associated with the key concepts outlined in Chapter Two that also form the boundaries of the literature review relevant to the main areas covered by this study. It is subsequently necessary to outline a number of key concepts and definitions.

An electronic bulletin board is usually dedicated to a specific topic, to which all messages sent by users are accessible to be read. Threaded discussion groups (also called computer conferencing) store text-based comments on a central server. Comments or messages are organised by topics, which is known as threads. Participants can access the discussion group and read comments and post responses. Depending on the management of the electronic bulletin board, messages can stay visible and active until such time that a discussion thread is closed or messages removed. In some cases, threads are kept for very long periods.

Asynchronous communication refers to discussions occurring independently of time or location. Using computer-mediated communication, participants send messages to a central location (discussion forum on an electronic bulletin board) where they are archived for later retrieval by other participants. An example of asynchronous communication is email. Live chat rooms and real time forms such as face-to-face communication are examples of synchronous communication.

As described in Chapter Two, “cyberspace” is a common term used to describe the digital environment that the Internet creates with all the different services it provides. Understandably, particular attention is given in cyberculture studies to theoretical and philosophical considerations regarding this term and its relevance to an understanding of Internet culture.

As mentioned in Chapter Two, the concept “virtual community” refers to a meeting place for people on the Internet. As a digital domain, it facilitates interaction and collaboration among people who share common interests and needs. Online communities can be open to all or restricted to members only. A community like the *WELL*™ is an example of a digital community that requires subscription fees in order to gain access. Digital domains may or may not offer moderator tools that can be enforced by means of terms of use, thereby restricting users and the type of exchanges. Moderators have the right to block certain users, remove or censor messages.

As explained in Chapter Two the concept “virtual reality” is sometimes used interchangeably with cyberspace. It refers to the immersion of one or more individuals in a digital or so-called virtual environment, with the aim of achieving the illusion that they are in a place, time or situation different from their actual real-world location and/or time. Virtual Reality or VR can also refer to 3D computer-generated worlds that necessitate special hardware.

3.2 Concepts and the case study

Related to the *Thorn Tree*, the following concepts are applicable. First, branches and categories: The asynchronous message board is divided into distinct sections

or branches, i.e. *Departure Lounge*, *The Lobby*, *News Stand*, and *Tree House*. These branches are subdivided into categories: countries in the case of the *Departure Lounge.* Second, message threads, i.e. Posts, Replies: The postings made by users of the *Thorn Tree*. Original postings are posts, while subsequent answers to the original post are replies. Posts and their replies are grouped together as threads. Threads can be viewed as a series of linked, related (electronic text) messages.

3.3 Social network properties and the research technique

When drawing conclusions from data sets resulting from statistical experiments, facts concerning the population are postulated. In social network analysis the population is the sample thereby making inferences a moot point. Therefore, the standard extrapolation of the research to a larger population is not undertaken. In its place, the associations between actors and their affiliations become the focus of interest.

Social network analysis is concerned with understanding the linkages among social entities and the implications of these linkages (Garton et al, 1999: 78; Wasserman and Faust, 1994: 17). Concepts and terminology additional to statistical terminology are required to understand and interpret the interactions that take place. Some of these terms and their definitions are described below. The terms are intentionally ordered such that later entries depend upon earlier ones. Note that not all terms may be used in this study. However, they may be required to understand subsequent terms.

3.3.1 Actors, ties and nodes

An actor is a discrete individual, corporate or collective social unit. The term actor does not imply that these entities necessarily have volition or the ability to "act". With regard to a relational tie, actors are linked to one another by social (relational) ties. The defining feature of a relational tie is that it establishes a linkage between pairs of actors (Marsden, 1990: 437). A dyad is a relational tie between two actors. A triad consists of the (potential) relational ties between three actors. A subgroup is any subset of actors and all the relational ties between them. The term "subgraph" is the graph theoretic terminology used for a subgroup. A group on the other hand is the collection of all actors on which relational ties are to be measured. The term "graph" is the graph theoretic terminology used for a group. The collection of ties of a specific kind among members of a group is called an elation.

A social network consists of a finite set or sets of actors and the relation or relations defined on them. A social network is a set of nodes (or actors) connected to each other. Nodes can be anything and the connection can be any attribute. Actor, node and connection network data are defined by actors and by relations (or nodes and ties). Many ties are directional. A relation is directional if the ties are oriented from one actor to another. A directed graph is often referred to as a digraph (Hanneman, ca1999: 23).

3.3.2 Subgraph and cliques

A clique in a graph is a maximal complete subgraph of three or more nodes. A clique consists of a subset of nodes, all of which are adjacent to each other, and there are no other nodes that are also adjacent to all of the members of the clique.

A clique is a collection of actors all of whom “choose” each other, and there is no other actor in the group who also “chooses” and is “chosen” by all of the members of the clique. Cliques in a graph may overlap.

Structural variables measure ties of a specific kind between pairs of actors, for example, business transactions between corporations. Composition variables are measurements of actor attributes. Composition variables are also called actor variables. Composition variables are the standard variables used in social and behavioural sciences and are defined at the level of individual actors (gender, race, ethnicity, geographic location).

The ties between all members in a clique are maximal and complete in both the common use as well as the mathematical use of the words. In graph theory, cliques consist of the maximal subsets of points in which each point is in a direct and reciprocal relation with all others. Note that the directions of the ties between actors in a clique are usually ignored and that any member of the clique may be placed in any position. The remaining five 5-member cliques have the same relationship shown in Figure 3 each is connected to the remaining four members.

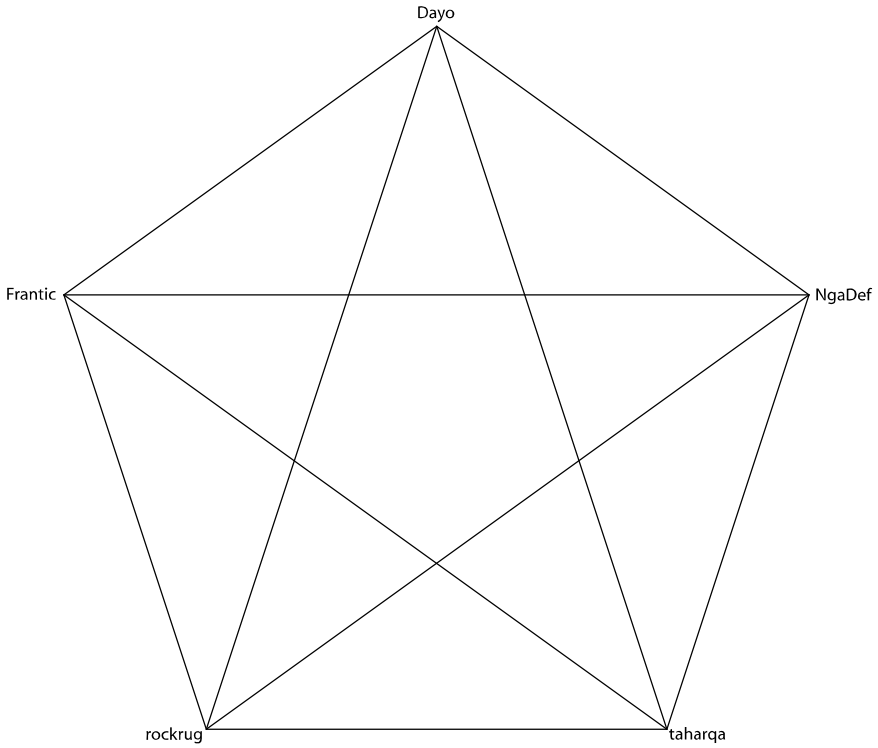


Figure 3 The adjacencies between five actors in a 5-member clique

A 4-member clique is represented in Figure 4. Note that in this case each member is connected to the other three members of a clique. Like the 5-member clique, any member of a 4-member clique may be placed on any vertex. The relationships between the actors are the same as those in any clique: they are maximal and complete. The map of a 3-member clique is simply a triangle. With the visualization of the relationships between the members of a clique, it is apparent that cliques are defined to contain at least three nodes (members). This is intentional in order to exclude mutual dyads as cliques.

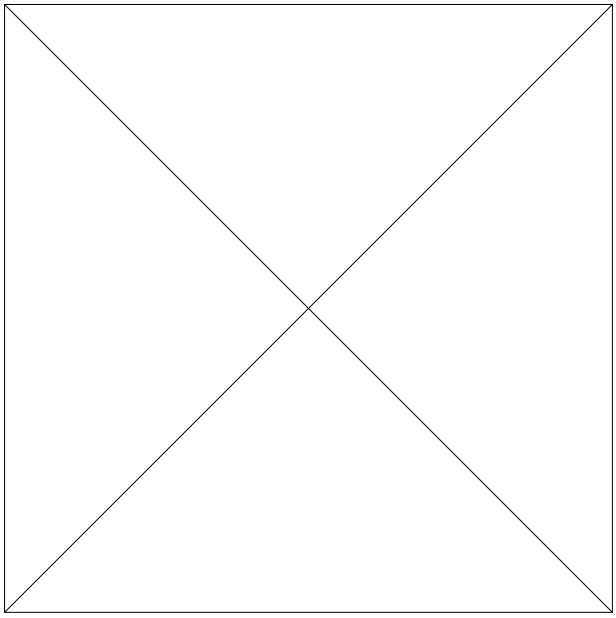


Figure 4 The adjacencies between four actors in a 4-member clique

A clique is a very strict definition of a cohesive subgroup because all ties between actors must exist and all actors are the same distance away from each other: one geodesic distance. These actors "communicate" directly with one another. There are no "friends of friends" in a clique. A clique is a special case of an *n*-clique, where *n* represents the maximum geodesic distances allowed between actors. In graph theory, a 1-clique is simply termed a *clique* and this terminology is used throughout this study. A 2-clique would allow a maximum geodesic distance of 2 that corresponds with the "friend of a friend" concept. Formally, an *n*-clique is a subgraph with node set N_S, such that

$$d(i,j) = n \quad '' \quad n_i, n_j \hat{I} N_S.$$

This equation may be read as follows with the appropriate portions of the equation embedded in the description between square brackets: The geodesic distance

between node *i* and node *j* is less than or equal to *n* [d(i,j) = *n*], for all n_i and n_j [" n_i, n_j] that are elements of the subgraph N_s [Î N_s], where N is the graph of actors (nodes) in the set {n_1, n_2, …, n_g}. In addition, there are no additional nodes that are also distance *n* or less from all nodes in the subgraph.

Simply stated, 2-cliques are subgraphs in which all members need not be adjacent (a friend of a friend may be included) however, each member is reachable through at most one intermediary. Because *n*-cliques are defined for geodesic paths that can include any nodes in the graph, two problems occur: the subgraph that makes up an n-clique may have a diameter greater than *n*; and, the *n*-clique may be disconnected. These problems occur because *n*-cliques are not as cohesive as cliques and therefore are not appropriate when researching cohesive subgroups.

Note that 2-cliques are generally larger and more numerous than cliques. All *n*-cliques, excluding 1-cliques, are considered weak cliques. In the case when messages are posted on a bulletin board, *n*-cliques are inappropriate. Both posters and repliers connect to one another through a thread and not via an intermediary unless one considers the thread as the intermediary. It is possible that an actor might receive the contents of a message through an intermediary but this action is not recorded by the bulletin board system and therefore cannot be analysed. In addition, the physical geographic distance separating many actors obviates an intermediary.

3.3.3 Mode

Mode refers to a distinct set of entities on which the structural variables are measured. In a one-mode network, structural variables measure a single set of

actors (friendships among residents) and are therefore termed one-mode networks. On the other hand, a network data set containing two sets of actors is termed a two-mode network. A special type of two-mode network that consists of one set of actors and one set of events is termed an affiliation network (Hanneman, ca1999). Relevant to one-mode networks, an ego is a focal actor and an alter is a non-focal actor. An ego-centred network consists of an ego and a set of alters who have relational ties to ego and among themselves. An ego-centred network is often termed a local network.

3.3.4 Graphs and visualisation

Graph theory is used in social network analysis for various reasons: it provides a concise and precise vocabulary used to label and denote social structure properties; it provides mathematical operations and ideas with which many quantities can be quantified and measured; and, it provides the ability to prove theorems about graphs and hence about representations of social structure (Hanneman, ca1999: 21-23; Garton et al, 1999: 96).

As noted earlier in a preceding sub-section, graphs of large networks often become meaningless. Matrices offer a solution since they also allow for complex statistical calculations. A matrix is nothing more than a rectangular arrangement of a set of elements. Rectangles have sizes that are described by the number of rows of elements and columns of elements that they contain. Connectivity between actors is indicated in binary form where a 1 symbolises a connection and a 0 indicates the absence of any connection (Hanneman, ca1999: 26-29).

Graphs and sociograms are the visual display of a network. The sociogram is the primary matrix used in social network analysis. It is based on graph theory and often involves complex mathematical calculations and sophisticated software. One reason for using mathematical and graphical techniques in social network analysis is to represent the descriptions of networks compactly and systematically. This also enables us to use computers to store and manipulate the information quickly and more accurately than we can by hand. There are a number of variations on the theme of sociograms, but they all share the common feature of using a labelled circle for each actor in the population being investigated; a line between pairs of actors represents the observation that a tie exists between the two (Hanneman, ca1999: 35).

3.3.5 Walks, trails and geodesic distance

A walk is a sequence of nodes and lines, starting and ending nodes, in which each node is incident with the lines following and preceding it in sequence. A trail is a walk in which all of the lines are distinct though some node(s) may be included more than once. A path is a walk in which all nodes and all lines are distinct. (Note that every path is a trail and every trail is a walk.) Walks and paths are used to calculate the distance between two nodes. A closed walk is a walk that begins and ends at the same node. A tour is a closed walk in which each line in the graph is used at least once. A cycle is a closed walk of at least three nodes in which all lines are distinct and all nodes, except the beginning and ending node, are distinct. Paths, tours and cycles are used to define geodesic distance, diameter and eccentricity (Hanneman, ca1999: 47-48).

The shortest path between two nodes is referred to as a geodesic. There may be more than one geodesic between two nodes. The geodesic distance is defined as the length of a geodesic between nodes (the shortest path). For both directed and undirected data, the geodesic distance is the number of relations in the shortest possible walk from one actor to another. For each actor, we could calculate the mean and standard deviation of their geodesic distances to describe their closeness to all other actors. For each actor, that actor's largest geodesic distance is called the eccentricity — a measure of how far an actor is from the furthest other. The distance between two nodes is the length of any shortest path between them. If no such path exists between two nodes, they are unreachable and the distance is considered infinite (Hanneman, ca1999: 50).

The eccentricity or association number of a node is the largest geodesic distance between that node and any other node. Eccentricity summarises how far a node is from the node most distant from it in the graph. Several measures of centrality (centre and centroid) are based on the eccentricity of the nodes.

The diameter of a connected graph is the length of the largest geodesic between any pair of nodes. This is equivalent to the largest nodal eccentricity. The diameter of a graph quantifies how far apart the furthest two nodes in the graph are. For example, consider the transmission of a message. If messages always take the shortest route (geodesic) then we are guaranteed that a message can travel from any actor to any other actor over a path no greater than the diameter of the graph. Also note that the diameter of a subgraph is the length of the largest geodesic within the subgraph (Hanneman, ca1999: 50-53).

Usually the size of a network is indexed simply by counting the number of nodes. In any network there are (*k* * (*k*-1)) unique ordered pairs of actors (that is AB is different from BA, and leaving aside self-ties), where *k* is the number of actors.

3.3.6 Degree (indegree and outdegree), centrality and closeness

In a graph the degree of a node is the number of nodes adjacent to it. This is also equivalent to the number of lines incident with that node. In a digraph a node can be either adjacent to or adjacent from another node depending upon the direction of the line or arc. Therefore, indegree and outdegree are treated separately. The indegree of a node consists of the number of nodes adjacent to other nodes. This means that this node is a receiver of information. The number of indegrees may be considered measures of receptivity or popularity. Outdegree (out degree) of a node consists of the adjacent from this node. This means that this node is a source of information. The number of outdegrees may be considered measures of expansiveness (Hanneman, ca1999: 61).

In undirected data, actors differ from one another only in how many connections they have. With directed data, however, it can be important to distinguish centrality based on indegree from centrality based on outdegree. If an actor receives many ties, they are often said to be prominent, or to have high prestige. That is, many other actors seek to direct ties to them, and this may indicate their importance. Actors who have an unusually high outdegree are actors who are able to exchange with many others, or make many others aware of their views. Actors who display high outdegree centrality are often said to be influential actors.

An isolate node has an indegree and an outdegree of zero. An isolate is not connected to any other node. A transmitter node only exists in a digraph and has an indegree of zero and an outdegree greater than zero. A receiver node only exists in a digraph and has an indegree greater than zero and an outdgree of zero. A carrier or ordinary node only exists in a digraph and has an indegree greater than zero and outdegree greater than zero. The neighbourhood of an actor consists of all connections to that actor (ego) regardless of their direction of connection. In other words, the neighbourhood consists of all actors in the adjacency matrix who have a "1" in the same row or column as the ego (Hanneman, ca1999: 63).

Fully saturated networks (i.e. one where all logically possible ties are actually present) are empirically rare, particularly where there are more than a few actors in the population. It is useful to look at how close a network is to realizing this potential, i.e. to examine the density of ties, which is defined as the proportion of all ties that could be present to those that actually are present.

Prominent actors are those that are extensively involved in relationships with others and this involvement makes them more visible to the other. The indegree and outdegree of a prominent actor is of no concern. What is important is simply that the actor is involved. A central actor is a prominent actor, that is, an actor with many ties.

Concerning actor prestige, in a digraph a prestigious actor is one who is the object of extensive ties thus focussing solely on the actor as a recipient. In the case of group centrality and group prestige, group-level measures are used to compare different networks easily.

Actor closeness centrality is a function of the geodesic distances the actor has with the other actors. As the geodesic distances increase in length the centrality of the actor involved decreases. This type of centrality depends not only on direct ties but also on indirect ties especially when two actors are not adjacent.

The term closeness is used to mean actor closeness centrality. The measure of prestige among actors is the indegree of each actor. The idea is that actors who are prestigious tend to receive many nominations or choices. A relative indegree is the proportion of actors that choose a specific actor: the larger the index, the more prestigious is the actor.

Reciprocity is a measure of the answer to a simple question: “How strong is the tendency for one actor to ‘choose’ another if the second actor chooses the first?” Reciprocity, trust and trustworthiness are also concerns in the study of social capital, as discussed in Chapter Two.

3.4 Software applications

Virtually all analyses performed in this study are performed by the commercially available computer program *UCINET* by Borgatte, Freeman and Everett*. UCINET* bundles additional software tools that are useful for additional analysis or presentation of the social networks: *Mage 3D Visualization, Netdraw* and *Pajek*. It contains a number of standard multivariate analysis tools and routines such as multi-dimensional scaling (MDS), correspondence analysis and hierarchical clustering. The software application performs matrix symmetrization, row and column selection and other transformations. Features specific to network analysis are as follows:

- Connectivity functions include geodesic distances, node accessibility and path studies. Built-in algorithms give path lengths, link strengths and costs of links.
- Centrality functions include degree, closeness, betweenness, centrality and power-scoring on matrix-specific vectors.
- Subgroup search functions include maximal and *n*-cliques, *n*-clans, lambda sets, factions, *k*-plex, *k*-cores and identify graph components and calculate densities for each.
- Role and position analysis functions include searches for regular, structural, automorphic and other equivalence searches.

Note that the full capability of *UCINET* is not required for the analyses performed in this study. Only those features of *UCINET* that are applicable to this study are used.

3.5 Descriptive statistical methods: concepts and this study

Formulas and equations have been intentionally left out of this study as they may be found in any good reference covering the material described in this section. It is to be noted that the theoretical detail of statistics does not form part of this study. However, the use of these techniques is sufficient to draw conclusions or describe the findings and arguments put forth in Chapter Four. This also applies to additional explanations not mentioned in Chapter Three to clarify or indicate a point that may or may not be obvious to the reader.

The following terms are referred to in this study either directly or indirectly. With the exception of those terms relating to survival analysis (see “3.5.12 Survival analysis”) all terms are standard statistical terminology and are provided to facilitate the reader with limited exposure to statistical analysis. The order of the terms has been explicitly selected since terms lower in the list are generally dependent upon previous definitions.

3.5.1 Population, census

While a population is the entire group of objects about which information is required, a sample is part of or a subset of the population used to gain information about the whole. A census is a sample consisting of the entire population. Furthermore, a unit is any individual member of the population. A variable is a characteristic of a unit that is to be measured for those units in the sample.

3.5.2 Measurements

A measurement of a property has a nominal scale if the measurement tells only what class a unit falls in with respect to the property. Gender and race are examples of measurements on a nominal scale. The measurement has an ordinal scale if it also tells when one unit has more of the property than does another unit. Sizes of small, medium, large and extra large are examples of measurements on an ordinal scale. The measurement has an interval scale if the numbers tell us that one unit differs by a certain amount of the property from another unit. Temperatures in degrees Celsius are measurements on an interval scale. The measurement has a ratio scale if in addition the requirements of an interval scale the measurements

indicate that one unit has so many times as much of the property as does another unit. Length and mass are measured on a ratio scale.

3.5.3 Mean

The mean of a set of *n* numbers is the arithmetic average; it is the sum of the observations divided by the number of observations, *n*. The mean is a measure of the centre of a data set. The mean makes sense only when used with the interval/ratio data because it requires addition of the measurements. In practice, means are frequently computed for ordinal variables as well.

3.5.4 Median

The median is the typical value; it is the midpoint of the observations when they are arranged in increasing order. The median is the 50^{th} percentile. The median is a measure of the centre of a data set. The median makes sense only for interval/ratio and ordinal variables.

3.5.5 Mode

The mode is the most frequent value; it is any value having the highest frequency among the observations. The mode is a measure of the centre of a data set. The mode is the only measure that makes sense with nominal variables.

3.5.6 Proportion

A proportion is defined as a part (fraction) of a whole. A percentage is obtained by multiplying a proportion by 100. The n^{th} percentile of a set of numbers is a value such that *n* percent of the numbers fall below it and the rest fall above it. The lower quartile is the 25^{th} percentile. The upper quartile is the 75^{th} percentile.

3.5.7 Minimum and maximum

The minimum number is the smallest value in a set of numbers. The maximum number is the largest value in a set of numbers. The five-number summary of a set of numbers consists of the minimum value, the lower quartile, the median, the upper quartile and the maximum value. The five-number summary is a method to measure the centre and spread of a set of numbers.

3.5.8 Deviation

The deviation is the arithmetic difference between two values. The variance is the mean of the squares of the deviations of the observations from their mean. The standard deviation is the positive square root of the variance. The three-number summary of a set of numbers consists of their mean, standard deviation (or variance) and total number of elements in that data set. The three-number summary is a method to measure the centre and spread of a set of numbers.

3.5.9 Charts, histograms and other graphic illustrations

A chart or graphic is any illustration or drawn design. A bar chart is a graphic representation of data usually in the form of solid vertical or horizontal bars. A bar chart is usually used to illustrate and compare several nominal variables.

Histograms look like bar graphs but they differ in several respects: the bars are always vertical; the base scale is always marked off in equal units; the widths of each bar are identical and represent the same range; the height of each bar is proportional to the whole. (If counts are used for the heights of the bars, this chart is often called a frequency chart or frequency distribution.) Line graphs show the behaviour of a variable over time. Time is marked on the horizontal axis and the variable being plotted is marked on the vertical axis. The horizontal axis is commonly referred to as the x-axis and the vertical axis as the y-axis.

Scatterplots are used to analyse the cause and effect relationship between two variables where two sets of data are plotted on the horizontal and vertical axes, respectively. Each observation is represented by a single point with the horizontal coordinate equal to the value of the first variable and vertical coordinate equal to the value of the second variable. If the scatterplot shows regression then the explanatory variable is always plotted on the horizontal axis and the response variable is always plotted on the vertical axis. Scatterplot graphs bivariate data when both variables are measured in an interval/ratio or ordinal scale. The horizontal axis is commonly referred to as the x-axis and the vertical axis as the y-axis.

A regression line is a straight line that describes how a response variable (vertical axis) changes as an explanatory variable (horizontal axis) changes. Regression is

often used to predict the value of a response variable for a given explanatory value. Regression, unlike correlation, requires an explanatory variable and a response variable. The horizontal axis is commonly referred to as the x-axis and the vertical axis as the y-axis. The least-squares regression line is the line that makes the sum of the squares of the vertical distances of the data points from the line as small as possible. The horizontal axis is commonly referred to as the x-axis and the vertical axis as the y-axis.

3.5.10 Correlation

Correlation is a measurement of association between two variables and is referred to as *r* or *R*. The correlation measures the strength and direction of the linear relationship between two quantitative variables. Correlation only makes sense when both variables have an interval/ratio scale but is sometimes used with ordinal scales as well.

The square of the correlation (r^2) is the fraction of the variation in the values of response variable that is explained by the least-squares regression of the response variable on the explanatory variable. The value, r^2, is a measure of how successful the regression is in explaining the response. (If $r^2 = 0.75$, this means that 75% of the variation in the response variable is accounted for by the linear relationship between the two variables and the remaining 25% is caused by another unknown source.)

3.5.11 Censored observations

Censored observations contain only partial information and they may or may not have survived the period of study. Those surviving the total time of the study or

those who became disengaged ("lost") during the study represent censored observations: information on their status is not available. Survival means that the event of interest (termination) has not occurred.

3.5.12 Survival analysis

The Kaplan-Meier Product-Limit Method estimates the survival function directly from the continuous survival or failure times of the observed units. This estimate of the survival function is known as the product-limit estimator. The advantage of the Kaplan-Meier Product-Limit method over the life table method for analysing survival and failure time data is that the resulting estimates do not depend on the grouping of the data into a certain number of time intervals. The Kaplan-Meier Product-Limit method is often used to estimate the probability of survival over a given time period.

Perhaps the technique of survival analysis must be further explained as this technique is not universally understood or used by many researchers. In effect, survival analysis tries to estimate the length of time a unit "survives", "endures" or lasts. The variable to be studied is the time delay until the occurrence of an event (death, disease, treatment outcome). This time delay corresponds to survival duration (the difference between the beginning study date and the event date). Survival analysis techniques were primarily developed in the medical and biological sciences but they are also widely used in the social and economic sciences as well as in engineering. Social scientists study the survival of marriages, high school dropout rates (time to drop-out) and turnover in organizations; economists study the survival of new businesses or the survival times of products; and, quality control engineers study the survival of parts under stress (failure time analysis).

What distinguishes survival analysis from most other statistical methods is the presence of “censoring” for incomplete observations. For example, in a study of survival following two different treatment regimens, analysis of the trial typically occurs well before all the patients have died. For those still alive at the time of analysis, the true survival time is known only to be greater than the time observed to date. Such an observation is said to be “censored”. There are two other sorts of incomplete observation: the “lost to follow-up” (patient missing during the study duration) or the appearance of an event other than the event being studied. These observations are also considered censored.

In this study the duration of threads is investigated in order to estimate its longevity. Some threads were active when the study began, some were active when the study ended and some threads were determined to be incomplete owing to various factors. Survival analysis enabled a precise estimation of the duration of a thread that informs the researcher in selecting data in order to draw a conclusion.

The so-called hazard function contributes to findings about the survival of threads. The overall shape of a typical hazard function is often referred to as the “bathtub curve” and is used to describe the probability of failure of components in a product. At the beginning of the life cycle component failure is usually high and is often referred to as “infant mortality failures”. This period is followed by a relatively “quiet” period of random failures when the failure rate (termination rate of threads) is relatively low and constant. Then after some time of operation the failure rate (termination rate of threads) begins to increase until all components or devices have failed or terminated. In this study, a so-called bathtub curve is superimposed over the results of the hazard function in order to illustrate this point. The bathtub curve is not a result of a mathematical calculation and may not be the best fit. It is rather an illustrative exercise to elucidate its expected shape.

4. This study: issues of measurement

In this section, issues that pertain to the population, boundary and sampling are discussed. In Chapter One it is stated that the *Africa* category of the *Departure Lounge* is taken as a case study of travel information exchange in a computer-mediated social network. In this discussion forum, nature and frequency of participation differ among actors because actors make different types of contributions (described by action) to different threads and post unequal numbers of messages. If messages are taken as the reason for a tie between actors, a structure transpires. By investigating the structure of this network it is possible to describe the nature of the social ties and to comment on the impact this structure has on the flow of resources (information and other exchanges) through the network. In network analysis research, like any other social sciences research, it is necessary to specify the population, demarcate the boundaries and where necessary, work with a sample. In the sub-sections below, these issues are clarified.

4.1 Boundary specification

The selection of one category of all the categories that are available on the *Thorn Tree* deserves attention since it affects the boundary specification. The selection of the *Thorn Tree* as case study deserves attention as much as the selection of one of the categories amongst all others on a specific branch. With reference to the study of Wang and Fesenmaier (2004), *Lonely Planet* is only mentioned as an example of a company website that uses the community building features of the Internet to facilitate information sharing among travellers. This leaves room for scholarly research about the *Thorn Tree*. Moreover, the *Thorn Tree* is exemplary of

computer-mediated communication and with a focus on the role of the Internet in facilitating computer-mediated information dissemination; *Lonely Planet's* website offers a usable example representative of others. On a technical level, the manner in which the *Thorn Tree* is managed together with the data that is retrievable for analysis purposes made it a successful contender as a case study of an electronic bulletin board with a shared interest in travel.

Three reasons are foremost in selecting the *Africa* category. First, this study was undertaken on African soil. Second, the researcher is knowledgeable about the tourism industry in South Africa. Lastly a personal trip to Morocco at the outset of this study gave the researcher first-hand experience, especially regarding the need for travel information. More importantly, it gave the opportunity to compare travel information on the *Thorn Tree* about Morocco with some of the limited personal experiences while travelling in Morocco.

Some investigators have stipulated inclusions of rules in terms of two or more of the three definitional foci outlined in the above. They state that:

> *…[w]hile this may lead to theoretically elegant definitions of membership, it also has a major weakness, in that it reduces the number of problematic features to be explained given knowledge of network structure* (Lauman et al, 1989: 69).

In this study, the focus falls upon the participative approach since the participation in a thread determines whether someone is included in the dataset or not. People who merely read a thread and don't respond in any way are not recorded. (The number of times a message has been viewed is displayed on the *Thorn Tree*

webpage, however, the viewer remains unidentified and cannot be used in this study).

Lauman et al (1989: 70) produced an eightfold typology of boundary specification strategies using the distinction between nominalist and realist views and cross-tabulating that on the ontological status of social phenomena. It is summarised in Table 1 in the Annexure. If the meta-theoretical perspective to this study is assumed to be a realist approach, it would mean that Strategy V on the typology of boundary specification for delimiting actors within this network is appropriate. This complies with the nature of the data and the focus of this study.

Strategy V entails a realist and participative approach (Lauman et.al, 1989: 72). An actor's inclusion in a network is defined in terms of participation or interest in one or more events, activities or concerns. This is the primary alternative to Strategy I from the realist perspective. Strategy I deals with tightly bound groups, where the inclusion rule for actors refers to socially defined and recognised group memberships. Can the *Thorn Tree* be defined as a “tightly bound” group? Seemingly not, since membership is unrestricted, although necessary for posting and/or replying to threads. Ties appear to be weak while the network as such is sparse. A participatory approach is therefore more appropriate since membership is open to anyone anywhere with no excluding criteria other than general terms of use. After all, inclusion in the dataset depends on participation in a thread, as outlined above. Moreover, participation levels suggest that large numbers of members are inactive because they contribute infrequently to discussions. This is outlined in more detail in Chapter Four.

4.2 Sampling

Statistical studies may be classified in three categories: producing data, organising and analysing data and drawing conclusions from data. When producing data the researcher determines the type of sampling required and designs the experiment accordingly. In this study, and as is usual with most studies involving social network analysis, the population is used and a census is taken.

Although a census is a sample consisting of the entire population, in fact, it is an attempt to sample the entire population. In any census there are units of the population that are not obtained for one reason or another. The missing units must be estimated and the reasons for their absence identified in order to form a judgment of their effect on the result. This is covered in the sub-section dealing with data integrity and eliminating errors.

Reminiscent of sampling in more traditional social research, for the purposes of this study, measurement is limited to the *Africa* category. However, in network analysis sampling is not undertaken since it will culminate in errors.

4.3 Reliability and validity

Mouton and Marais (1996: 79) state that the core consideration of validity "concerning the process of data collection is that of reliability". In essence, this means that the application of a valid measuring instrument to different groups under different sets of circumstances should lead to the same observations. This concurs with Reinhard (1994: 240) who states that "validity is the consistency of a measure with a criterion".

However, a further statement (Reinhard, 1994: 240) is important for any research, namely that it:

> *is possible to have a reliable test without having a valid one…but you cannot have a valid measure without it first being reliable. Thus, validity presumes reliability.*

According to Garton et al (1999: 92), gathering data electronically replaces issues of accuracy and reliability with issues of data management, interpretation and privacy. Indeed, as is the case with this study, electronic monitoring can routinely collect information on whole networks. By closely following the accepted practices of this research technique, care was taken to ensure reliability. For example, by employing basic statistical calculations on the acquired data, errors that arose during the data capturing process were eliminated, thereby increasing the level of data integrity and reliability in the study of this whole network. It is important to note that two standard software packages available as Open Software from the Free Software Foundation were used to capture the data: *wwwoffle* and *wget*. This is discussed in more detail in sub-section “5.1 Acquiring data”.

The type of data and the manner in which data capturing was handled meant that some of the factors that are referred to as “nuisance variables” could be avoided. Considering what Mouton et al (1996: 81-82) refer to as “researcher effects”, it is necessary to note that at no point prior, during or even after the research period for this study was it made known to members of the *Thorn Tree* that the nature and extent of information exchange resulting from their participation in a computer-mediated social network were being studied.

A factor that relates closely to reliability concerns the biographical information of *Thorn Tree* members. When a member applies, an online form requests biographical information. The validity of the analysis relies on the detail people reveal in their profiles when they register and whether profiles are visible or not. Relying on this to compile a comprehensive profile of members proved problematic since members might either be supplying real/truthful information or not, might not have completed all fields, or might either be serious or joking, which means that data becomes questionable. This means that almost no means exist to compare stated identities with real ones. Subsequently, any attempt to work with biographical information in order to assist with the interpretation and analysis of network data was discarded.

Another factor influenced the completeness of the dataset, namely the way in which *Lonely Planet* manages the *Thorn Tree.* This involves the amount of time a particular thread stays active and therefore visible on the *Thorn Tree* or the possibility that messages from a thread can be censored and thus deleted. This is outlined in a thread from the *All About the Thorn Tree* category on *The Lobby* branch. On 19 March 2004 at 11:56, *mauriziogiuliano* posted a message with the subject “deletions from TT”:

> *Dear Roman,*
>
> *I wonder, do posts on TT get deleted sometimes and how ? Maybe you delete them when the question has been answered fully, and / or when the question is no longer relevant ? Or when something else is wrong ?*
>
> *I see for example some deletions...*

> *On the "all about LP" branch, I had put a thread on the Falklands guidebook. I received a reply (from you I think), and then the thread disappeared, so maybe you deleted it because it had been replied ? I had also put a thread about "new suggested guidebooks" and it was deleted too, maybe because so many other threads cover the same thing ? On the Africa branch, my Somalia thread was deleted, maybe due to lack of replies.*
>
> *Any policy ideas would be appreciated. In any case, I think it would be good to tell the thread's author when these get deleted.*

In a reply, *Lan* posted a message at 12:28 on the same day:

> *Most of the branches have an automatic expiry set, if there's no posts in a thread it'll usually vanish after two weeks - some are longer, some I think are shorter. Although they could notify someone it might get tedious for some - I've posted more than 10000 times, I certainly don't want a notification each time a message expires ;-)*

In an effort to assist even more, *dlutzy* posted the following at 18:56:

> *If you click "Subscribe to this thread" you'll get a copy of any/all replies sent to your email account. So you can keep the information forever if you wish.*

The following reply by *mauriziogiuliano* at 19:01 reveals much about the uses of the *Thorn Tree*:

> *Dear Dlutzy, thanks. Yes true, but the point off TT is to inform others and not just myself. So I put information on there, and if no-one replkies* [sic] *within a couple of weeks it will be deleted ? Could I send a weekly message to keep it alive ?*
>
> *I am not really clear. There is even the old TT. So I thought thr point was to keep messages and replies for ever. Wrong ?*

The following message posted by *hunwagner* at 21:42 heeds a relevant warning:

> *Wrong - if they kept all messages forever, the TT would grow enormous, full of outdated info. Some particularly good or popular threads are made "kept" though, and these never expire.*
> *You will find such kept threads at the last page of each forum. You can try keeping your post alive by sending weekly replies to yourself, bit I guess that will just bore people...*

Revealing something about the management of the *Thorn Tree*, *montyman* reminds everyone of moderators' roles:

> *some moderators behave like Major Major in catch 22 deleting whole lines and obscure words sometimes whole threads and also they use nom de plumes like washington Irvine to hide their crimes. Often they*

> *leave a thread devoid of meaning which is a form of Trolling in itself. They are very dangerous and when they do it they may not be the actual moderator you think did it. They will then go into hiding until they feel it is safe to venture out cut up another threadless victim obsequious to their own ideology*

Since moderators remove some threads, it means that the sociogram of the *Africa* category reflects those messages that were available at the time data capturing was done. If this is regarded as a shortfall, no method exists to overcome it. It is impossible to reinstate all the threads made to any category since the inception of the *Thorn Tree* or determine a specific period during which no threads are deleted. Since the data was obtained from *Lonely Planet's* website in an automated fashion, none of the usual factors associated with participants played a role in the reliability of the data, i.e. memory decay, the omnipresent syndrome, interview saturation, role selection, response patterns or the necessity to motivate participants to participate.

Regarding the content of messages and discussions, no manipulation took place by the researcher, such as participating in online discussions, or creating aliases in an attempt to steer conversations in a particular direction.

5. This study: methods

This section describes the sequential steps that were taken in order to obtain the data and prepare it for analysis.

5.1 Acquiring data

In the previous section, it was outlined that the population consists of all the actors that contributed to discussion threads on the *Africa* category of the *Departure Lounge* branch of the *Thorn Tree*. Reasons were given for this boundary specification. It was also stated that in network analysis no sampling of the population is taken.

Mentioned by Rice (1994: 174), an intriguing aspect of studying computer-mediated communication systems such as the Thorn Tree is that they can be more or less unobtrusive components of research design. It is also true for this study, since data was collected without anyone being aware of it. The acquisition of the preliminary data was obtained on 27 December 2003. This attempt to obtain and view the raw data available from *Lonely Planet* indicated several potential problems: the amount of data available from the *Africa* category was large and the on-site retention of the data was approximately one month.

The size of the data indicated that automated techniques would be required to capture the data. In order to minimize the loss of data, the duration of its acquisition would have to be as short as possible. The timeframe within which data are retained on *Lonely Planet*, however, is sufficient for analysis. Two standard software packages available as Open Software from the Free Software Foundation were used to capture the data: *wwwoffle* and *wget*. The manual pages and text dumps for these two software packages may be found on the accompanying CD-Rom. The package *wwwoffle* is an off-line Web (http) reader and *wget* is a package that downloads Web (http) pages.

The package *wwwoffle* was required to read the Web pages from *Lonely Planet* because of the dynamic nature of the site. When *wget* was used it reconnected to the previously downloaded but changed pages and commenced downloading them again in order to keep them up-to-date. This repetition coupled with a dialup connection only allowed about ten pages to be downloaded before *wget* would start at the first screen again. The advantage to using *wget* was that there were no missing data, however, the disadvantage was that only about 100 threads covering less than one week were available for analysis. Therefore, it was decided to use *wwwoffle* to page through each screen manually and to accept the loss of some messages. The advantage of using this method is that more messages as well as longer threads would be available for analysis.

5.1.1 wwwoffle

The package *wwwoffle* was used in proxy mode to acquire the individual pages from *Lonely Planet* and to store them on the local computer. Owing to the lack of speed from a dialup modem and the fact that this was intentionally a manual process, the duration of capture was approximately 10 hours. During this period of capture messages were added to the current list and messages were discarded from the end of the list. The number of threads with potential errors encountered and the methods used to account for this loss are discussed in the sub-section entitled “Preparing data for analysis”. The date of capture of the *Africa* category was 5 June 2004.

5.1.2 wget

Once the entire *Africa* category was downloaded with *wwwoffle* the local copy of the site appeared static. At this stage *wget* was used to reconstruct the screens for analysis. In addition to the date of capture, eight separate days were spent trying to configure *wget* to capture the *Africa* branch without continuously repeating from the first screen. These screens have also been included in the analysis. The advantage of including these data mean that missed messages would be recovered and included in the analysis. The disadvantage of including these data means that duplicates would have to be identified and removed.

5.2 Preparing data for analysis

The result of using *wwwoffle* and *wget* produced two generic files: *categories* and *messagepost*. The *categories* files contain the high level structure of threads: title, message, poster, date, time, number of replies and number of views. The *messages* file contains the original message plus the replies: replier, title, message, date and time. Occasionally, several message files are used to contain the replies.

The name of each Web page saved by *wget* contains additional information following the file name. In fact, the filename is the name of the dynamic page created by *Lonely Planet* (*messages.cfm* or *category.cfm*) but each may be distinguished by the GET section of the screen that is now part of the filename. (The GET section of a Web address is only one of three methods used with Web pages to send and receive information. The other two methods are POST and so-called cookies. Understanding how these techniques are used is not critical to this study).

The information attached to the *categories* file is its category ID (called *catid* and is 9 for the *Africa* category) and start page number (STARTPAGE). Neither the category IDs nor the start page numbers are used in the analysis but the category ID was essential in obtaining the correct information from *Lonely Planet*. The information attached to the *messages* file, however, is used in the analysis for the purpose of checking the internal integrity of the information. There are up to eight fields attached to each *messages* file of which only the thread ID and message ID are important. The other fields were not used in analysing the messages: post action (always “reply”), category ID (9 for *Africa* category), start page, parent ID, “from” and a “showall” field. The “showall” field appears only once. The “from” field is used to sequence the reply screens and ranges from 1 to 63 in the final data set.

The “from” field contains the number associated with each screen that contains messages. In this case, the maximum screen page is 63. When using *wget* the maximum screen numbers downloaded was seven (9 May 2004) before repeated screens began appearing. Therefore, allowing the loss of some data, as described in the subsection entitled *Acquiring data*, enabled the analysis of nine times more screens and enlarges the period of observation from less than a week to about six weeks.

5.3 Extracting the data

As emphasised by Garton et al (1999: 92), using electronic means to study networks (such as computer-mediated social networks) often revolves around the ingenuity of researchers and programmers in their study design. Undertaken by *Forthwith Computers*, a simple program (*lonely.c*) was written in the *C*-programming language to capture the relevant information from the downloaded web screens. A

copy of the programme is provided on the accompanying CD-Rom. The programming technique used to acquire the information is commonly called “screen-scraping” by programmers and is a method used to acquire data from an active (live) screen. However, in this case, the screens were not active as they had been downloaded by *wwwoffle* and processed into screen image files by *wget*.

The information obtained in this fashion was stored in two tab delimited text files, namely *categoryfile* and *messagefile*. Each file has a header section that is identical and its content is derived from the filename and GET section of the file.

5.4 Internal data integrity checks

Preliminary data integrity checks consist of matching the filename with the name of the original file that was produced, thus ensuring that the category ID always refers to the *Africa* category (9). Both files contain the correct information and are therefore valid files. In addition, the message filename refers to the thread ID and the message ID. These were checked against one another and verified correct.

Since the information contained in the filename corresponded with the information taken from the screen, both files were imported into two tables in a database for further validation. Some editing of special characters was required to ensure that the data loaded correctly but other than these few changes no additional content was changed.

The *category* and *message* tables in the database consisted of 1 520 and 7 254 entries (rows), respectively. The data retrieved on 27 December 2003 was

intentionally included in the dataset to ensure that any changes occurring on *Lonely Planet* would be incorporated correctly. After it was verified that all data had imported properly, a SQL statement deleted the old data acquired on 27 December 2003. The remaining entries consist of the data that would be further validated, corrected and analysed. The remaining rows in the *category* and *message* tables were 1 500 and 6 547, respectively. Most of the following validations were written in SQL, PHP or in a combination of the two languages.

Owing to the fact that threads and messages downloaded both at the beginning and the ending of the data may contain discrepancies, two specific checks were performed: isolate threads in *category* that are not in *message* and vice-versa. The following threads were discarded because there is no possible method to link them to valid messages. First, 286 threads were in *category* but not in *message, while* eleven threads were in *message* but not in *category.* The significance of this lies in the process of data-capturing and the manner in which the *Lonely Planet* servers store data. It surely pointed at the necessity to check data for integrity and other errors.

Duplicate posts that were introduced (intentionally) were then removed. No exact duplicates were detected in the table *category*, however, 1 791 duplicates were detected in *message* and subsequently removed. The remaining valid entries totalling 4 756 were placed in a database; it forms the basis for analysis.

Three threads were found containing a total of 29 messages with corrupted titles. The corrupted titles were caused by a small software error in the *C* program and appeared only when more than one page of replies were received. As only the titles were corrupted these entries were corrected and a few duplicate messages and three threads were removed.

The results at this stage show 1 282 actors (poster and repliers) contributing toward 1 027 individual threads. There are no duplicate posters for any thread and each thread has exactly one poster. There are no duplicate titles or duplicate content for any thread. Further internal integrity checks were then possible by looking at the contents of the data and interpreting the results.

5.5 Estimating uncertainty of data

Each thread on the *Lonely Planet* site contains an indication of the number of replies as well as the number of views that thread received. The number of threads recorded and the number of threads reported by *Lonely Planet* indicate the precision of the data. In total, there were 4 756 unique messages in 1 027 threads. The recording mechanism captured 46 more messages than reported by *Lonely Planet*. In addition 332 messages were missed. The difference (281) indicates that the data have an inherent uncertainty owing to incompleteness of approximately 6%. The uncertainty is not unusual nor is it excessive. As the analysis consists primarily of ratios and the amount of data is relatively large, the inherent uncertainty of results should be negligible.

6. This study: calculations and measurements

6.1 Descriptive statistical calculations and this dataset

Table 2 lists the number of threads and the number of replies each thread contains. Using descriptive statistical calculations, the median, average and mode of responses could be calculated.

Calculating ratios, considering the total number of threads (1 027) and the number of threads that received no replies (165), a ratio of 0.161 is derived at. This means, as outlined in more detail in Chapter Four, that 16% of all messages in this dataset remain unanswered.

For the purposes of determining the length of time that a thread remains active, the Kaplan-Meier Product-Limit is used to do a survival analysis. Findings are outlined in more detail in Chapter Four. Here it is necessary to state that the captured data include the date and time of messages, thus enabling the calculation of the duration of threads. The hazard function is another method to describe the duration of threads.

Based on the raw data containing the number of messages, the number of replies and the number of views, a ratio can be calculated with regard to replies versus views. A scatterplot, as explained earlier in this chapter, is used. The exact linear equation of the linear least squares (LLS) line relating the number of replies to the number of views is not relevant for this analysis. It is, however, relevant that the number of replies is positively associated with the number of views in order to gain information about the correlation between views and replies. As noted earlier in this chapter, only actors that post messages are contained in this dataset. Actors who merely view messages cannot be captured. Subsequently, the only data that reflects upon viewers is contained in a scatterplot where the number of views is correlated with the number of replies. Based on the findings of this scatterplot, the ratio of replies to views is calculated (refer to Figure 8 in Chapter Four).

A histogram was used to visualise the general length of messages. The length of messages points at actor behaviour in as far as it is vaguely indicative of the amount of information contained in a message. In a verbal conversation, the length

of a message correlated to the duration of a conversation. It is not, however, indicative of the content or the value of the content, since three words can sometimes mean more than three paragraphs if placed within context.

6.2 Network analysis, methods and this dataset

In order to study the connections between posters and repliers, social network analysis techniques are employed on a dataset that contains the binary code reflecting the ties between actors based on the messages they contributed to different threads. There are those actors who initiate threads and those who reply to messages. In those instances where an actor who initiates a thread also participates in the ensuing conversation by posting more messages, such messages are excluded from analysis. Nevertheless, in this study the data is asymmetric which means that it is possible to distinguish between ties being sent and ties being received. Later, the importance of this transpires since it enables a view of directional ties, i.e. sources or sinks.

The size of a network is often a very important factor to consider since size is critical for the structure of social relations because of the limited resources and capacities that each actor has for building and maintaining ties. As a group gets bigger, the proportion of all of the ties that could (logically) be present -- density -- will fall, and the more likely it is that differentiated and partitioned groups will emerge. Usually the size of a network is indexed simply by counting the number of nodes. In any network there are (k * k-1) unique ordered pairs of actors (that is AB is different from BA, and leaving aside self-ties), where k is the number of actors. It follows from this that the range of logically possible social structures increases exponentially with size (Hanneman, ca1999: 41-42).

In a one-mode network such as this one, as explained earlier, the names of actors in the rows are repeated in the columns[1]. In this study, some of the information derived from network analysis is verifiable with findings from the descriptive statistical calculations, i.e. the number of actors, messages and threads. Focussing first on the network as a whole, the number of actors, the number of connections that are possible, and the number of connections that are actually present are determined using *UCINET*. Although fully saturated networks (i.e. one where all logically possible ties are actually present) are empirically rare, it is useful to determine the density of ties, which is defined as the proportion of all ties that could be present and that actually are. Through this measurement, it is possible to determine whether a network is dense or sparse. It furthermore follows on the strength of ties, i.e. weak ties or strong ties. The implications of these measurements for characterising the flow of information through an exchange network such as the *Africa* category are outlined in Chapter Four.

The number and kinds of ties that actors have are a basis for similarity or dissimilarity to other actors and hence to possible differention and stratification. The number and kinds of ties that actors have are keys to determining how much their embeddedness in the network constrains their behaviour, and the range of opportunities, influence, and power that they have. These characteristics are underpinned by measurements using *UCINET* with reference to network size, degree (indegree and outdegree), centrality and reachability. Considering degree and specifically indegree and outdegree depends on whether information is looked at row-wise or column wise. In the case of the former, the outdegree or out-ties are calculated, or the extent to which actors are senders to others. In the case of the latter, the indegree or in-ties are measured which reveals the extent to which actors

[1]. This matrix (raw data) is contained in the CD-Rom inserted at the back of this thesis.

are receivers (Hanneman, ca1999: 43). Findings in this regard are outlined in Chapter Four. Exploring the neighbourhoods of actors (Table 11 and Table 12 in the Annexure), calculations derived from a one-mode network using *UCINET* allow for an ego-analysis of actors. From this follows that reciprocity and transitivity can be measured (Hanneman, ca1999: 45).

Clique analysis is calculated by *UCINET* and is used to determine which actors are more closely and intensely tied to one another. As outlined earlier in this chapter a clique is some number of actors who have all possible ties present among themselves. With reference to Table 13, findings are outlined in Chapter Four. These findings are based on the raw data, contained in Table 7 namely 1 282 actors, 1 027 threads and 6 547 messages. Algorithms to calculate cliques are contained within *UCINET*. One of the most common interests of structural analysts is in the "sub-structures" that may be present in a network. Dyads, triads, and ego-centered circles can all be thought of as substructures. Networks are also built up or developed out of the combining of dyads and triads into larger, but still closely connected structures. Many of the approaches to understanding the structure of a network emphasize how dense connections are compounded and extended to develop larger "cliques" or sub-groupings. This view of social structure focusses attention on how solidarity and connection of large social structures can be built up out of small and tight components: a sort of "bottom up" approach. Network analysts have developed a number of useful definitions and algorithms that identify how larger structures are compounded from smaller ones: cliques, *n*-cliques, *n*-clans, and *k*-plexes all look at networks this way. In this study, only cliques are identified using *UCINET's* ability to calculate this (Hanneman, ca1999: 77-80).

By using measurements from a two-mode network perspective, the results are focussed on events. In this case study, events point at messages or participation in

particular threads. In network analytical terms, this is referred to as an affiliation network, where actors are affiliated to events. Once again, *UCINET* was used to do calculations regarding specific threads. The selection of threads chosen for a two-mode perspective is based on the three threads with the highest number or replies and the actors notable as sources, sinks, and transmitters, i.e. *Dayo*, *dysfunctional*, *NgaDef*, *JayDawg* and *Micksailor*.

With reference to findings based on a one-mode network where the focus is upon actors, and relating it to findings based on a two-mode network common threads between four notable actors could be identified.

7. Final remarks and conclusion

In this chapter the methodological aspects related to network analysis have been outlined. Since results from descriptive statistical calculations are also relevant to this study for reasons outlined above, brief explanations and explanations of concepts were included.

Organising and analysing data produces many benefits of which the most important is perhaps the validity and consistency of the data with respect to further detailed analysis. Numerous techniques are used in this process. They include measurements of the centre and spread of various variables and the shape of the distribution these variables assume. Numerous graphical techniques normally present visual representations of patterns and trends in the data set to the researcher for further investigation and interpretation. In addition, numerous ratios, tables and calculated values are required for the researcher to appreciate the

information contained in the data sets. Often these somewhat mechanical techniques offer an opportunity for further investigation or explanation.

For example, if a data set contained a time element as a variable (arrival of messages) and no data appeared within a certain interval (midnight to 01:00) this would have to be investigated and explained. This time gap may have resulted in incorrect coding by the researcher, a problem with the source supplying the data or a systematic feature of the system (maintenance of the server sourcing the information).

The population for this study is limited to the *Africa* category, taken as a whole network. As such, no scientifically researched conclusions could be drawn regarding the level of connectivity, participation and reciprocity across the whole *Thorn Tree* or beyond, i.e. in real life. However, as outlined in Chapter Four, using results from calculations based on a one-mode and two-mode network, the structure of the network resulting from differentiated contributions to electronic discussions can be investigated. Moreover, the effects of structure on information exchange can be explained. Notably absent from this chapter, however, is an explanation of content analysis. Although used to a lesser extent, extracts from messages are included in Chapter Four to elucidate on the nature of exchanges, i.e. the significance of the message (Garton et al, 1999: 93).

The replicability of this study to include other categories in other branches of the *Thorn Tree* suggests a follow-up investigation to measure the extent of activity across all categories on the *Thorn Tree*. As such, replicability in network terms points to a high rate of validity and reliability concerning the research design.