

# A Dynamic Approach for Longitudinal Network Analysis: the Stochastic Actor-Oriented Model

Emily Shobana Muller (emily@aims.ac.za)  
African Institute for Mathematical Sciences (AIMS)

Supervised by: Dr Franck Kalala Mutombo  
AIMS South Africa, UCT South Africa, University of Lubumbashi Democratic Republic of Congo

18 May 2017

*Submitted in partial fulfillment of a structured masters degree at AIMS South Africa*

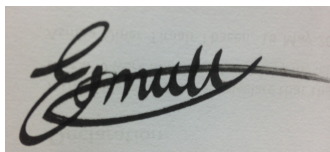


# Abstract

The Stochastic Actor-Oriented Model is implemented on two observations of the AIMS 2016/17 social network. The model estimates which network effects are significant determinants of social network evolution. Network evolution is assumed to be the result of individual friendship selection. Selection is based on maximising utility relative to the effects and is dependent only on the current network (Markovian). A Monte Carlo Markov Chain simulates the evolved networks based on the fitted model and these simulations are subject to difference tests. A method of testing the difference in network structure is defined over a set of network structure metrics. The simulated networks are found to underestimate clustering in the network. Other metrics return good fit.

## Declaration

I, the undersigned, hereby declare that the work contained in this research project is my original work, and that any work done by others or by myself previously has been acknowledged and referenced accordingly.

A handwritten signature in black ink, appearing to read 'Emull', is shown on a light-colored background.

---

Emily Shobana Muller, 18 May 2017

# Contents

<b>Abstract</b>	<b>i</b>
<b>1 Introduction</b>	<b>1</b>
<b>2 Graphs and Networks</b>	<b>2</b>
2.1 Graphs . . . . .	2
2.2 Network Effects . . . . .	3
2.3 Network Effects Example . . . . .	6
2.4 Network Topology . . . . .	7
2.5 Network Topology Example Continued . . . . .	10
2.6 Network Difference . . . . .	11
<b>3 Continuous-Time Markov Process</b>	<b>12</b>
3.1 Assumptions . . . . .	12
3.2 Continuous-Time Markov Process . . . . .	12
3.3 Example . . . . .	14
<b>4 The Model</b>	<b>15</b>
4.1 Estimation . . . . .	16
<b>5 Results &amp; Discussion</b>	<b>18</b>
5.1 Data and Cross-Sectional Interpretations . . . . .	18
5.2 SAOM Results . . . . .	20
5.3 Univariate Network Differences . . . . .	21
5.4 Multivariate Network Differences . . . . .	23
<b>6 Conclusion</b>	<b>25</b>
<b>A Appendix</b>	<b>26</b>
A.1 Powers of Adjacency Matrix . . . . .	26
A.2 Mahalanobis Distance . . . . .	26
A.3 Proof of Multinomial Logistic Regression . . . . .	26
A.4 Jaccard Coefficient . . . . .	27
A.5 Rao Score Test . . . . .	27
A.6 Network Effects . . . . .	27
A.7 Univariate Nonparametric Tests . . . . .	28
A.8 Multivariate Nonparametric Tests . . . . .	28
<b>References</b>	<b>32</b>

# 1. Introduction

The present work considers the social network of 41 AIMS 2016/17 Structured MSc students (recorded by questionnaire) at two time points in their academic year. The social networks are modelled as directed graphs,  $G(t_0)$  and  $G(t_1)$ , with edges representing friendship relations. The Stochastic Actor-Oriented Model (SAOM) is applied to the panel data. The network,  $G(t_0)$ , is the initial state for a Markov process. It is assumed that the network changes continuously in the time interval  $[t_0, t_1]$  and that each change is dependent only on the current network. The assumption that a social network is a stochastic process with Markovian properties was first suggested by Holland and Leinhardt (Holland and Leinhardt, 1981). Under these assumptions, Wasserman proposed a dynamic model incorporating only the effect of reciprocity — a measure of relationships between two individuals — for which parameters are estimated using maximum-likelihood methods (Wasserman, 1980). Snijders further developed models including higher order network effects (beyond two individuals) and a Monte Carlo Markov Chain (MCMC) method of parameter estimation (Snijders, 2013). The SAOM model (Snijders, 2005) is unique in its ability to capture high order dependency and provide a method for parameter estimation. The purpose of SAOM model is to determine which structurally dependent network effects lead individuals to select and de-select friendship nominations, thus the name: Actor-Oriented Model. The selection choice is captured by a utility function (the objective function): a linear combination of network effects,  $\rho_s$ , plus random residual influence. The probability of selection is a Discrete Choice model (Maddala, 1983). The coefficients of network effects are estimated using a converging stochastic approximation algorithm based on iterative updates of the selective Markov process. This is an MCMC method.

A statistical approach is taken to determine the significance of effects. Reciprocity, transitive triplets and homophily related to country of origin are returned as significant effects of the AIMS social network. These effects are common features of social networks (Wasserman and Faust, 1994).

The limitation of this model is the volume of data considered. The AIMS social network contains 41 individuals, and inference to the greater population is not viable based on only 41 individuals. An alternative is to collect and analyse social network data online. Such data is considered more accurate, since relations are based on actual event data, and more reliable, since measurements can be taken on large volumes. However, the data may not be valid for addressing certain research questions pertaining to the social sciences. Therefore, the inference power of the SAOM lies in its ability to predict, and the ability to predict requires an adequate model. Model adequacy is tested in this work by performing goodness of fit measures.

The fitted model is used to simulate  $N = 1000$  networks, which, based on the estimation procedure and a well fitted model, should not be structurally different from the observed network structure,  $G(t_1)$ . A statistical method of testing network difference, relative to structural network metrics, is defined. The null hypothesis is that the metric average over the 1000 simulations is not different from the observed metric for  $G(t_1)$ . The AIMS data fails to reject the null hypothesis for all metrics except clustering. Clustering is underestimated by the simulated networks. Possible explanations are explored as to how this trend could arise.

The importance of a dynamic social network model is derived from its ability to test certain sociological theories. For example, the contribution of social selection and social influence to academic performance (Lomi et al., 2011). This differs from the “gold standard” statistical approach to sociological studies which assumes independence of individuals (Carolan, 2014). It differs too from cross-sectional social network analysis, which is used to describe and understand the network at a single point in time, and is unsuitable for addressing research questions pertaining to the mechanisms of social evolution.

## 2. Graphs and Networks

The following section outlines the principles of graph theory and important applications and interpretations in social network analysis.

### 2.1 Graphs

**2.1.1 Definition (Graph).** A (directed) graph  $G$  is a pair  $G = (V, E)$ , where  $V$  is a set of vertices or nodes, and  $E \subseteq V \times V$ , a set of edges between nodes. If vertex  $v_1$  is connected to  $v_2$ , then there exists an edge  $(v_1, v_2) \in E$ .  $E$  can be seen as a relation on the set  $V$ .

**2.1.2 Definition.** The edges  $E$  can have the following properties:

1.  $E$  is anti-reflexive if  $(v, v) \notin E$  for all  $v \in V$ .
2.  $E$  is symmetric if  $(v_1, v_2) \in E \iff (v_2, v_1) \in E$  for all  $v_1, v_2 \in V$ .

The properties above have the following translation to graphs (networks):

- (i) The graph  $G$  is undirected if  $E$  is symmetric.
- (ii) The graph  $G$  is simple undirected if  $E$  is symmetric and anti-reflexive.
- (iii) The graph  $G$  is directed if  $E$  is not necessarily symmetric.
- (iv) The graph  $G$  is simple directed if  $E$  is not necessarily symmetric and anti-reflexive.

Note that the definition of  $E$  prevents multiple edges (more than one edge  $(v_1, v_2) \in V$ ) and the definition of a simple (directed) graph prevents self-loops (reflexive edges).

All the information of a graph structure can be stored in a matrix.

**2.1.3 Definition (Adjacency Matrix).** Let  $G$  be a simple graph and  $i, j \in V(G)$ . The adjacency matrix  $A = (A_{ij})$  is an  $n \times n$  matrix with entries such that

$$A_{ij} = \begin{cases} 1 & \text{if there exists an edge between } i \text{ and } j \\ 0 & \text{otherwise} \end{cases}, \quad (2.1.1)$$

for  $i, j = 1, \dots, n = |V(G)|$ .

The adjacency matrix is related to network structure in the following way:

1. If  $G$  is undirected,  $A_{ij}$  is symmetric.
2. If  $G$  is simple undirected,  $A_{ij}$  is symmetric with 0 diagonal.
3. If  $G$  is directed,  $A_{ij}$  is not-necessarily symmetric.
4. If  $G$  is simple directed,  $A_{ij}$  is non-symmetric with 0 diagonal.

A weighted graph,  $W$ , with adjacency matrix  $A$  has entries  $A_{ij} \in \mathbb{R}$  for all  $i, j \in V(W)$ .

Let  $m$  be the number of edges in a network. For a simple undirected graph the total number of possible edges is  $m_{tot} = n(n-1)/2$ . For a simple directed graph the total number of possible edges is  $m_{tot} = n(n-1)$ .

**2.1.4 Definition** (Complete Graph). A complete graph,  $G$ , has all possible edges:  $m_{tot} = |E(G)|$ .

For the rest of this work, every graph is simple directed. More on networks detailed in (Newman, 2010).

## 2.2 Network Effects

The application of graphs to social network analysis introduces new terminology. Vertices are people in the network, called *actors*, and edges represent relationships between actors, called *ties* (Carolan, 2014). The network of actors and ties together make a group. A group is defined as "a finite set of actors who for conceptual, theoretical, or empirical reasons are treated as a finite set of individuals on which network measurements are made" (Wasserman and Faust, 1994). A network may identify more than one type of node (multi-nodal), for example, students and teachers, and furthermore, more than one type of relation, for example, friendship and collegial. Analysis of multi-nodal networks has been carried out using multilevel statistical analysis (Lazega and Snijders, 2015). This work considers friendship networks with only one type of node (uni-modal).

The uni-modal friendship networks treated in this work are modelled as simple directed graphs (simple digraphs). This is reasonable in the context of friendship networks. Relationships are directed towards individuals and the meaning of this direction can be linked to important sociological theory. Self friendships and multiple edges are excluded. The latter is an assumption of this work and need not always be satisfied. Consider, for example, giving each friendship relation a weight.

The following section introduces network effects that are of particular interest in social network analysis and provides context to sociological theory. The endogenous and exogenous network effects outlined below are of most relevance to the AIMS social network. Further effects are detailed in (Ripley et al., version March 21, 2017).

Prior to the introduction of the network effects, it is useful to briefly consider the SAOM (see Section 4). The model assumes each tie to be under the control of the sending actor. Each possible network configuration provides a level of utility for the actor, and the choice to make or break a tie with one of the  $n - 1$  remaining actors is based on maximising an objective function. The objective function is a generalised linear model of network effects plus random residual influences to account for unexplained utility (i.i.d type 1 extreme value). The purpose is to establish which effects are significant in determining the mechanisms by which a social network evolves.

**2.2.1 Endogenous Effects.** Endogenous effects capture the dependence on network structure relative to the individual actor embedded in the network (Snijders, 2005). Let  $G$  be a simple directed graph and  $A_{ij}$  be the corresponding adjacency matrix. A network effect,  $\rho_s$ , takes as input a graph  $G$  and vertex  $i \in V(G)$  and returns a positive real. For  $s \in S = \{1, 2, \dots, 15\}$ , the effects  $\rho_s$  are defined below.

**2.2.2 Definition** (Out-Degree). Out-degree,  $\rho_1(G, i)$ , measures the number of edges that actor  $i$  sends to other actors  $j$  in the network:

$$\rho_1(G, i) = \sum_{j=1}^n A_{ij} = A_{i+}. \quad (2.2.1)$$

Within a social network, an actor with a high out-degree can be said to be more sociable than a contemporary with a low out-degree (Carolan, 2014). In social networks the number of outgoing ties is generally low compared to the possible number of ties (density  $< 0.5$ ) and in the context of utility, this indicates a greater cost than benefit of sending a tie (Snijders et al., 2010).

**2.2.3 Definition (In-Degree).** In-degree,  $\rho_2(G, i)$ , measures the number of edges actor  $i$  receives by each other actor  $j$  in the network:

$$\rho_2(G, j) = \sum_{i=1}^n A_{ij} = A_{+j}. \quad (2.2.2)$$

In a social network, an actor with a high in-degree is interpreted as more popular (Carolan, 2014). The dispersion of in-degrees and out-degrees can indicate a networks tendency to social hierarchy. Hierarchical structures are developed further in this section.

**2.2.4 Definition (Reciprocity).** Reciprocity,  $\rho_3(G, i)$ , measures the number of reciprocated ties that actor  $i$  has with its adjacent neighbours:

$$\rho_3(G, i) = \sum_{j=1}^n A_{ij}A_{ji}. \quad (2.2.3)$$

Tendency towards high reciprocity is a feature of social networks (Wasserman and Faust, 1994).

**2.2.5 Definition (Transitive Triplets).** Informally, if  $i$  is a friend of  $j$  and  $j$  is a friend of  $k$ , then  $i$  is a friend of  $k$  also. This effect counts the number of relations  $A_{ik}$  such that transitive closure is formed:

$$\rho_4(G, i) = \sum_{1 \leq j, k \leq n} A_{ij}A_{jk}A_{ik}. \quad (2.2.4)$$

This defines a triangle of connected actors, or *triad*, in one particular direction only. For each set of 3 actors  $i, j, k$  there are two ways of forming a triad relative to actor  $i$ .

The presence of transitive triplets indicate the tendency towards network clustering. Where clusters are highly connected subgroups of a network. The local and global clustering coefficients are outlined in the following section. Transitive triplets also indicate local network hierarchy. Consider a single triad; actor  $i$  sends two ties, actor  $j$  sends and receives one tie and actor  $k$  receives two ties.

**2.2.6 Definition (Transitive Ties).** Consider the vertices  $i, k$ . A transitive tie between  $i, k$ , requires that there exists at least one relation  $j$  such that  $A_{ij}A_{jk} = 1$ :

$$\rho_5(G, i) = \sum_{k=1}^n A_{ik} \max_j (A_{ij}A_{jk}). \quad (2.2.5)$$

This differs from the previous transitive triplets definition by indicating that more intermediaries do not add proportionally to the tendency towards transitive closure (Snijders et al., 2010).

**2.2.7 Definition (Number Distance 2).** Consider the same triad of actors  $i, j, k$ . Number distance 2 measures the number of incomplete triads, such that there does **not** exist the relation  $A_{ik} = 1$ :

$$\rho_6(G, i) = \#\{k | A_{ik} = 0, \max_j (A_{ij}A_{jk}) > 0\}. \quad (2.2.6)$$

This is a negative representation of transitive triplets and indicates tendency away from network clustering.

**2.2.8 Definition (3 cycles).** Consider the same triad of actors  $i, j, k$ . A 3 cycle measures the cyclic relationship:

$$\rho_7(G, i) = \sum_{1 \leq j, k \leq n} A_{ij}A_{jk}A_{ki}. \quad (2.2.7)$$

The presence of 3 cycles is representative of connectivity, but note the relations imply an egalitarian structure. Local hierarchy is indicated by a strong presence of transitive triplets and weak presence of 3 cycles (Snijders et al., 2010).

**2.2.9 Definition (Cliques).** A clique is defined to be a set of actors in a network such that each actor is connected to every other actor in the set, i.e., this subgroup of actors form a complete graph.

**2.2.10 Definition (Balance/Structural Equivalence relative to out-degree).** Balance,  $\rho_8(G, i)$ , determines how similar actor  $i$  is to their neighbours based on their choices of outgoing ties:

$$\rho_8(G, i) = \sum_{j=1}^n A_{ij} \sum_{k=1, k \neq i, j}^n (1 - |A_{ik} - A_{jk}|). \quad (2.2.8)$$

Two vertices in a network are structurally similar if they share many of the same neighbours (Newman, 2010). This helps to identify patterns of clustering (cliques) in the network.

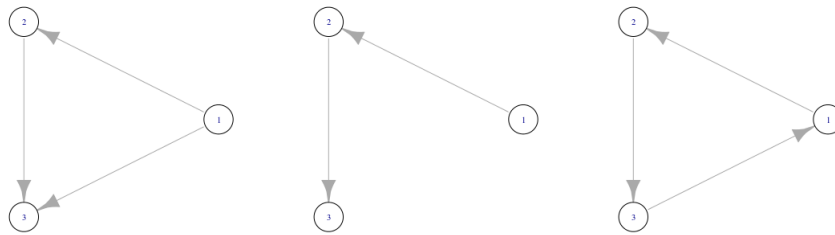


Figure 2.1: Let  $G$  be the graph of nodes  $\{1, 2, 3\}$ . An example of transitive triplets (left), number distance 2 (middle) and 3 cycles (right) is illustrated with corresponding effects:  $\rho_4(G, 1) = 1$ ,  $\rho_6(G, 1) = 1$  and  $\rho_7(G, 1) = \rho_7(G, 2) = \rho_7(G, 3) = 1$ , respectively.

**2.2.11 Definition (In-degree Popularity).** In-degree popularity,  $\rho_9(G, i)$ , measures the extent to which actors with high in-degrees are friends with other actors with high in-degrees:

$$\rho_9(G, i) = \sum_{j=1}^n A_{ij} A_{+j}. \quad (2.2.9)$$

**2.2.12 Definition (Out-degree Popularity).** Out-degree popularity,  $\rho_{10}(G, i)$ , measures the extent to which actors with high out-degrees are friends with other actors with high in-degrees:

$$\rho_{10}(G, i) = \sum_{j=1}^n A_{ij} A_{+j}. \quad (2.2.10)$$

**2.2.13 Definition (Out-degree Activity).** Out-degree activity,  $\rho_{11}(G, i)$ , measures the extent to which actors with high out-degrees send ties:

$$\rho_{11}(G, i) = A_{i+}^2. \quad (2.2.11)$$

**2.2.14 Definition (In-degree Activity).** In-degree activity,  $\rho_{12}(G, i)$ , measures the extent to which actors with high in-degrees send ties:

$$\rho_{12}(G, i) = A_{i+} A_{+i}. \quad (2.2.12)$$

The tendency towards global hierarchy is indicated by a strong positive effect ( $\beta$  coefficient) of both in-degree related popularity and out-degree related activity, combined with a strong negative effect of both out-degree related popularity and in-degree related activity (Snijders et al., 2010).



## 2.3 Network Effects Example

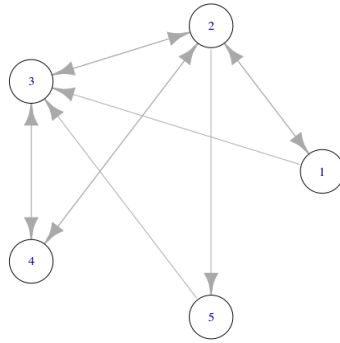


Figure 2.2: Simple digraph,  $G$ , with  $V(G) = 5$  and  $E(G) = 11$ .

Figure 2.2 is a simple digraph,  $G$ , and the corresponding network effects, for each  $i \in V(G)$ , are calculated in Table 2.1.

Table 2.1: Network effects,  $\rho_s(G, i)$ , for example in Figure 2.2.

$\rho_s \backslash V(G)$	1	2	3	4	5
$\rho_1$ : Out-degree	2	4	2	2	1
$\rho_2$ : In-degree	1	3	4	2	1
$\rho_3$ : Reciprocity	1	3	2	2	0
$\rho_4$ : Transitive Triplets	2	4	2	2	0
$\rho_5$ : Transitive Ties	1	1	1	1	0
$\rho_6$ : Number dist. 2	2	0	2	2	2
$\rho_7$ : 3 cycles	1	4	4	2	1
$\rho_8$ : Balance	2	4	2	2	1
$\rho_9$ : In-degree Pop	7	8	5	7	4
$\rho_{10}$ : Out-degree Pop	6	6	6	6	2
$\rho_{11}$ : Out-degree Act	1	16	4	4	1
$\rho_{12}$ : In-degree Act	2	12	8	4	1

Table 2.1 shows that vertex 2 sends the most ties, and vertex 3 receives the most. Vertex 2 has the greatest number of reciprocated and transitive triplet relations whilst 5 has none. Transitive triplets indicates subgroup connectivity and this is confirmed by 2 having the highest value for balance. A clique is identified between vertices  $\{2, 3, 4\}$ . The activity and popularity degree related effects are not very informative for a small network.

Note that these effects are not normalised. Therefore, it cannot necessarily be inferred that 2 is more *transitive* than node 1. Node 1 has 100% transitive relations.

**2.3.1 Exogenous Effects.** Exogenous effects capture the dependence of individual attributes on the network structure. For the application of graphs in network analysis, each vertex  $i \in V(G)$  holds

attributional information such as: age, sex etc. Such attributes are referred to as covariates and stored in the vector  $v_i$ . Each covariate is indexed by  $v_i^c$ , e.g.  $v_i^{sex}$ .

**2.3.2 Definition (Ego Activity).** Ego activity,  $\rho_{13}^c(G, i)$ , measures an actors activity relative to covariate  $c$ :

$$\rho_{13}^c(G, i) = v_i^c A_{i+}. \quad (2.3.1)$$

**2.3.3 Definition (Alter Activity).** Alter activity,  $\rho_{14}^c(G, i)$ , measures an actors popularity relative to covariate  $c$ :

$$\rho_{14}^c(G, i) = \sum_{j=1}^n A_{ij} v_j^c. \quad (2.3.2)$$

**2.3.4 Definition (Homophily).** Homophily,  $\rho_{15}^c(G, i)$ , measures tendency for ties to occur between actors with the same covariate  $c$ :

$$\rho_{15}^c(G, i) = \sum_{j=1}^n A_{ij} I(v_i^c = v_j^c). \quad (2.3.3)$$

where  $I(v_i^c = v_j^c) = 1$  if  $v_i^c = v_j^c$  and 0 if not.

Homophily within a social network, is most prominent between race and ethnicity, followed closely by age, religion, education and occupation (McPherson et al., 2001). Therefore, it is important to control for individual characteristics in the network evolution model (Lomi et al., 2011).

Note here that nominal covariates (categorical) are not considered for the network effects related to activity. Unless of dimension 2, in which case they may be centered.

This section introduced fundamental network structures relative to the actor. These effects are tested on the AIMS network data in Section 5.2. Further network effects are detailed in (Ripley et al., version March 21, 2017).

## 2.4 Network Topology

Network topology is the arrangement of vertices and edges in a network resulting in its structure. This section defines local and global structural metrics. Structural measurements are important since they allow for network comparison. Furthermore, one can make comparisons and deduce universal mechanisms which give rise to a particular network structure (Estrada and Knight, 2015).

Network difference is the fundamental principle for parameter estimation in Section 4. The method of analysis is detailed in Section 4.1. First, the need for this comparison is explored. The model estimates which network effects are significant determinants of social evolution, for a simple directed network  $G(t)$ , evolving from  $G(t_0)$  to  $G(t_1)$ . The choice of effects to include,  $\rho_s$ , are guided by theory,  $G(t_0)$  is given, a condition on running time,  $T_1$ , is imposed (see Section 4.1.1) and estimates,  $\hat{\theta}$ , are updated iteratively in a converging stochastic approximation algorithm (see Section 4.1.2). The final parameter estimates,  $\hat{\theta}$ , along with  $G(t_0)$ ,  $T_1$  and  $\rho$  become inputs for a MCMC simulation of  $N = 1000$  networks,  $\{H_i(T_1) | 1 \leq i \leq N\}$ , where  $H_i(T_1)$  are simple directed graphs and  $V(H_i) = V(G)$  for all  $i \in N$ . The goodness of fit of the model is assessed by measuring the difference in network topology of the observed network  $G(t_1)$  and the simulated networks  $\{H_i(T_1) | 1 \leq i \leq N\}$ .

Prior to defining a statistical measure of difference, network metrics are introduced. The global metrics,  $\mu_q$ , and local metrics,  $\gamma_r$ , are important measures of network topology and further details can be found in (Newman, 2010) and (Estrada and Knight, 2015).

**2.4.1 Global Metrics.** The following metrics,  $\mu_q$ , take as input a simple digraph  $G$ , with  $|V(G)| = n$ ,  $|E(G)| = m$  and corresponding adjacency matrix  $A_{ij}$ , and return a positive real. For  $q \in Q = \{1, 2, 3, 4\}$ , the metrics  $\mu_q$  are defined below.

**2.4.2 Definition (Density).** Density,  $\mu_1(G)$ , measures the number of edges as a fraction of total possible edges:

$$\mu_1(G) = \frac{m}{n(n-1)}. \quad (2.4.1)$$

**2.4.3 Definition (Reciprocity).** Reciprocity,  $\mu_2(G)$ , measures the fraction of reciprocated relations:

$$\mu_2(G) = \frac{\sum_{1 \leq j, k \leq n} A_{ij} A_{ji}}{m}. \quad (2.4.2)$$

In order to define global clustering and global mean distance, the shortest path must first be defined.

**2.4.4 Definition (Geodesic Path).** The geodesic path (or shortest path),  $d_{ij}$ , between two nodes  $i, j \in V(G)$ , is the smallest value  $r$  such that  $[A^r]_{ij} > 0$  (See Appendix A.1 for more details).

**2.4.5 Definition (Global Clustering).** Global clustering,  $\mu_3(G)$ , measures the number of transitive triplets as a fraction of possible transitive occurrences:

$$\mu_3(G) = \frac{\# \text{transitive triplets}}{\# \text{ of connected triplets of vertices}}, \quad (2.4.3)$$

where  $0 \leq \mu_3(G) \leq 1$ , and a connected triplet of vertices,  $i, j, k$ , is such that  $A_{ij} A_{jk} > 0$  for all  $i \neq j \neq k \in V(G)$ .

**2.4.6 Definition (Local Mean Shortest Path).** The mean shortest path,  $l_i$ , from actor  $i$  to every other actor  $j$  is given by

$$l_i = \frac{1}{n-1} \sum_{j=1, j \neq i}^n d_{ij}. \quad (2.4.4)$$

**2.4.7 Definition (Global Mean Shortest Path).** The average shortest path over all possible pairs of vertices is defined as

$$\mu_4(G) = \frac{\sum_{i=1}^n l_i}{n}. \quad (2.4.5)$$

**2.4.8 Local Metrics.** The following metrics,  $\gamma_r$ , take as input a simple digraph  $G$ , with corresponding adjacency matrix  $A_{ij}$ , and vertex  $i \in V(G)$  and return a positive real. For  $r \in R = \{1, 2, 3, 4\}$ , the metrics  $\gamma_r$  are defined below.

**2.4.9 Definition (Out-degree distribution).** Out-degree is defined as per Equation 2.2.1:  $\rho_1(G, i) = A_{+i}$ . The vector  $\rho_1 \in \mathbb{R}^{1 \times n}$  (with  $\rho_{1i} = \rho_1(G, i)$ ) is converted to a vector,  $\gamma_1 \in \mathbb{R}^{1 \times m}$ , representing the frequency distributions over  $m$  levels.

**2.4.10 Definition** (In-degree distribution). In-degree is defined as per Equation 2.2.2:  $\rho_2(G, i) = A_{i+}$ . The vector  $\rho_2 \in \mathbb{R}^{1 \times n}$  (with  $\rho_{2i} = \rho_2(G, i)$ ) is converted to a vector,  $\gamma_2 \in \mathbb{R}^{1 \times m}$ , representing the frequency distributions over  $m$  levels.

Before local clustering is defined, neighbourhood is introduced.

**2.4.11 Definition** (Neighbourhood). The neighbourhood,  $K_i$ , of vertex  $i \in V(G)$  is defined as

$$K_i = \{j \in V(G) | A_{ij} = 1 \text{ or } A_{ji} = 1\}, \text{ and let } k_i = |K_i|. \quad (2.4.6)$$

**2.4.12 Definition** (Local Clustering). Local clustering is a measure of how well connected actor  $i$ 's friends are to one another. It counts the number of pairs of neighbours of  $i$  that are connected and divides by the total number of pairs of neighbours:

$$\gamma_3(G, i) = \frac{\sum_{1 \leq l, m \leq k_i} A_{lm}}{k_i(k_i - 1)}, \quad (2.4.7)$$

where  $0 \leq \gamma_3(G, i) \leq 1$ . Actors tend to cohere into smaller subgroups, and higher clustering coefficients indicate some level of connectivity within a subgroup (Newman, 2010). Local clustering coefficients over all actors are stored in the vector  $\gamma_3 \in \mathbb{R}^{1 \times n}$ , where  $\gamma_{3i} = \gamma_3(G, i)$ .

Measures of centrality are interested in determining which node is most important in the network. The concept of importance can be framed in more than one way and closeness centrality is introduced below.

**2.4.13 Definition** (Closeness Centrality). In social network analysis, the closeness centrality of a node  $i$  is defined as

$$\gamma_4(G, i) = \frac{1}{l_i} = \frac{n - 1}{\sum_{j=1, j \neq i}^n d_{ij}}, \quad (2.4.8)$$

where  $d_{ij}$  is the geodesic distance between vertex  $i$  and  $j$ . Closeness measures how important each vertex  $i$  is, in terms of reachability to all other actors. The closeness centralities over all actors are stored in the vector  $\gamma_4 \in \mathbb{R}^{1 \times n}$ , where  $\gamma_{4i} = \gamma_4(G, i)$ .

Return to the definition of geodesic path. If there exists no path of outgoing edges connecting vertex  $i$  to  $j$  then  $d_{ij} = \infty$  for all  $i \neq j \in V(G)$  (Newman, 2010). A graph with this property is said to have more than one component (related to out-degree). For such a vertex  $i$ , it follows that  $l_i$  is infinite and therefore the closeness centrality is zero,  $\gamma_4(G, i) = 0$ . This is undesirable, because indeed vertex  $i$  has a measure of closeness for vertices  $j$  in the same component. This can be rectified by considering only  $d_{ij}$  for  $j$  in the same component as  $i$ . This however, is bias towards vertices in small components, as it will return a higher centrality score which does not reflect well the connectedness of a small component. Therefore, the metrics  $\gamma_4$  and  $\mu_4$  are redefined as follows:

**2.4.14 Definition** (Harmonic Mean Closeness (Newman, 2010)).

$$\gamma'_4(G, i) = \frac{1}{n - 1} \sum_{j=1, j \neq i}^n \frac{1}{d_{ij}}. \quad (2.4.9)$$

Hence, if  $d_{ij}$  is infinite for  $i$  and  $j$  in different components, the term becomes zero. Furthermore, it gives greater weight to vertices closer to  $i$  than further away (Newman, 2010).

**2.4.15 Definition** (Harmonic Mean Distance (Newman, 2010)).

$$\mu'_4 = \frac{n}{\sum_{i=1}^n \gamma'_4(G, i)}. \quad (2.4.10)$$

Again, this definition removes any contributions from vertex pairs with  $d_{ij} = \infty$  (Newman, 2010).

It is assumed that local metrics are independent of the vertex identifiers,  $i$ . Therefore, the vectors,  $\gamma_q, q \in \{1, 2, 3, 4\}$  (including  $\gamma'_q$ ), are unchanged by a permutation of vertex identifiers:  $\gamma_q(\pi(G), \pi(i)) = \gamma_q(G, i)$  for all  $i \in V(G)$ , where  $\pi(G) = (\pi(V(G)), \{(\pi(i), \pi(j)) | (i, j) \in E(G)\})$ . This is a graph isomorphism. The SAOM determines which network effects lead to an overall network topology and therefore is not dependent on vertex identifiers. Therefore, for goodness of fit analysis, the vectors are sorted in ascending order for ease of analysis of distribution.

This section can be extended further by including metrics such as assortativity (degree-degree correlation), returnability (4-cyclic, 5-cyclic,...) (Estrada and Knight, 2015) and further centrality measures such as eigenvector (degree centrality based on neighbours neighbours) and betweenness (the extent to which vertex lies on paths between other vertices) (Newman, 2010).

## 2.5 Network Topology Example Continued

Recall the network in Figure 2.2. Table 2.2 and 2.3 calculate the global and local metrics of  $G$ .

Table 2.2: Global network metrics,  $\mu_q(G)$ , for example in Figure 2.2.

$\mu_q$	$G$
$\mu_1$ : Density	0.55
$\mu_2$ : Reciprocity	0.73
$\mu_3$ : Global clustering	0.50
$\mu_4$ : Global mean shortest path	1.50
$\mu'_4$ : Harmonic mean distance	1.30

Table 2.3: Local network metrics,  $\gamma_r(G, i)$ , for example in Figure 2.2.

$\gamma_s \backslash V(G)$	0	1	2	3	4
$\gamma_1$ : Out-degree distribution	0	1	3	0	1
$\gamma_2$ : In-degree distribution	0	2	1	1	1
$\gamma_s \backslash \pi V(G)$	1	2	3	4	5
$\gamma_3$ : Local clustering	0.33	0.42	1	1	1
$\gamma_4$ : Closeness centrality	0.50	0.67	0.67	0.67	1
$\gamma'_4$ : Harmonic Closeness	0.58	0.75	0.75	0.75	1

The average shortest path length between any nodes  $i$  and  $j \in V(G)$  is 1.5 ties. The network has density greater than 0.5 and a high proportion of reciprocal relations. 3 out of 5 nodes belong to highly clustered neighbourhoods. Note here that a value of 1 for local clustering does not indicate a clique, since by definition, the neighbourhood of vertex  $i$  is defined by incoming **or** outgoing edges. 3 out of 5 vertices share the same value for closeness and harmonic closeness centrality. Note here that both centrality measures shows similar trend.

## 2.6 Network Difference

This section defines a statistical method to test whether the distribution of each network metric is equal to the observed metric. The distribution is the image of  $\mu_q$  and  $\gamma_r$  over the domain  $\{H_i(T_1) | 1 \leq i \leq N\}$ . The observed metrics are  $\mu_q(G(t_1))$  and  $\gamma_r(G(t_1))$ , for each  $q \in \{1, 2, 3, 4\}$  and  $r \in \{1, 2, 3, 4\}$ .

This is a two-sided one sample test with null hypothesis: mean of metric distribution is equal to observed metric (can also be considered a two-sample test with non equal sample sizes). The null hypothesis is rejected at a 5% confidence level. The left column of Table 2.4 describes a simple one-sample students t-test for univariate data, whilst the right column describes the one-sample Hotelling's  $T^2$  test for  $p$ -variate data (Oja and Randles, 2004) which uses Mahalanobis distance (see Appendix A.2).

Table 2.4: Statistical Tests for Network Difference

	Univariate Analysis	Multivariate analysis
Metric	$\mu_q, q \in \{1, 2, 3, 4\}$	$\gamma_r, r \in \{1, 2, 3, 4\}$
Sample Space	$\mu_q(H_1), \dots, \mu_q(H_N)$ where each $\mu_q(H_i)$ is univariate	$\gamma_r(H_1), \dots, \gamma_r(H_N)$ where each $\gamma_r(H_i)$ is of dimension $p$ .  $p = m$ for $\gamma_1$ and $\gamma_2$  and $p = n$ for $\gamma_3$ and $\gamma_4$
Sample Mean	$\bar{\mu} = \frac{\sum_{i=1}^N \mu(H_i)}{N}$	$\bar{\gamma} = \frac{\sum_{i=1}^N \gamma(H_i)}{N}$
Sample Variance	$s^2 = \frac{\sum_{i=1}^N (\mu(H_i) - \bar{\mu})^2}{N - 1}$	$S = \frac{\sum_{i=1}^N (\gamma(H_i) - \bar{\gamma})^T (\gamma(H_i) - \bar{\gamma})}{N - 1}$
Actual Metric	$\mu(G) = \mu^{obs}$	$\gamma(G) = \gamma^{obs}$
Hypothesis	$H_0 : \bar{\mu} = \mu^{obs}$	$H_0 : \bar{\gamma} = \gamma^{obs}$
Assumptions	<ul style="list-style-type: none"> <li>• Independence of sample</li> <li>• Sample normally distributed</li> </ul>	
Distance	$d = \bar{\mu} - \mu^{obs}$	$d^2 = (\bar{\gamma} - \gamma^{obs})^T (S)^{-1} (\bar{\gamma} - \gamma^{obs})$
Test Statistic	$t = \frac{d}{\frac{s}{\sqrt{N}}}$	$T^2 = Nd^2$
Test Distribution	$\sim t_{N-1}$	$\sim \frac{N-p}{Np} F_{p, N-p}$

## 3. Continuous-Time Markov Process

### 3.1 Assumptions

Let  $G(t)$  be a time-dependent simple directed graph representing a social network. Suppose two observations are made at  $t_0$  and  $t_1$ . The aim of the SAOM is to determine which network effects are significant in the evolutionary process of  $G$  from  $G(t_0)$  to  $G(t_1)$ . The following assumptions are made:

1. **Continuous Time.** Whilst discrete observations are made at  $t_0$  and  $t_1$ , time,  $t$ , is continuous and the network  $G(t)$  changes continuously in the interval  $t_1 - t_0$ .
2. **Conditional Change Independence.** In a small interval of time,  $\Delta t$ , only one tie can change in the network.
3. **Markov Property.** Change in network  $G(t)$  to  $G(t + \Delta t)$  by one outgoing tie from actor  $i$  to actor  $j$  (denoted  $G(i \rightsquigarrow j)$ ) is dependent on the current network configuration only.
4. **Discrete Choice Model.** Actors control change in outgoing ties (detailed in Section 4).

The first three assumptions are formalised in the following section.

### 3.2 Continuous-Time Markov Process

Let  $G(t)$  be a time-dependent simple directed graph representing a social network.  $G(t)$  is identified by its  $n \times n$  adjacency matrix  $A(t) = (A(t)_{ij})$ . Assume the actors,  $V(G)$ , remain constant through time whilst the ties,  $E(G)$ , are time dependent. Although  $G(t)$  is measured at discrete-time points,  $t_0$  and  $t_1$ , it is assumed that the network evolution follows a continuous-time Markov Process.

**3.2.1 Definition** (Stochastic Process). A stochastic process is a set of all possible random variables,  $\{X(t), t \in T\}$ , where  $T$  is the index set of time.

**3.2.2 Definition** (Continuous-Time Markov Process). Let  $\mathcal{G} = \{G(t), t \in T\}$  be the state space of all stochastic processes  $G(t)$ , where  $T$  includes discrete time observations,  $T = \{t \in \mathbb{R}^+ | t_0 \leq t \leq t_1\}$ . We call such a stochastic process a continuous-time Markov process (CTMP) if for all  $t \in T$

$$Pr[G(s+t) = j | G(s) = i, G(u) = g(u), \text{ for all } 0 \leq u < s] = Pr[G(s+t) = j | G(s) = i], \quad (3.2.1)$$

for  $i, j \in \mathcal{G}$ . In words, the next state is dependent on the current state only. The probabilities are called transition probabilities, denoted  $P_{ij}(t, s)$ , and represent the probability of moving from  $i \in \mathcal{G}$  to  $j \in \mathcal{G}$ . The state space  $\mathcal{G}$  is of order  $\mathcal{N} = 2^{n(n-1)}$  and is the set of all possible permutation of relations, 0 to  $n(n-1)$  edges over  $V(G)$  nodes. It is therefore, finite, though computationally exhaustive for large  $n$ .

**3.2.3 Definition** (Conditional Change Independence (Wasserman, 1980)).

$$P_{ij}(t, s) = Pr[G(s+t) = j | G(s) = i] = \prod_{k,l} P\{G_{kl}(s+t) = j_{kl} | G(s) = i\} + o(t), \quad \text{as } t \rightarrow 0. \quad (3.2.2)$$

For small time intervals,  $s+t$ ,  $t \rightarrow 0$ , the transition probability from  $i$  to  $j$  is the product of individual changes. This implies changes in ties are statistically independent, thus the probability of two ties changing simultaneously is essentially zero.

**3.2.4 Definition** (Time-homogeneity). A CTMP is time-homogeneous if the conditional probability is independent of time

$$P_{ij}(s, t) = P[G(s + t) = j | G(s) = i] = P[G(t) = j | G(0) = i] = P_{ij}(t), \quad (3.2.3)$$

for all  $i, j \in \mathcal{G}$ ,  $t, s \in T$ . Note this is not the same as a stationary CTMP, which requires that  $P_{ij}$  is independent of  $t$  for all  $t \in T$ .

Transition probabilities at time  $t$  are given by the transition matrix  $P(t) = (P_{ij}(t))_{i,j \in \mathcal{G}}$ . The transition matrix satisfies  $\sum_{j=1}^N P_{ij}(t) = 1$ , and is therefore a stochastic matrix.

**3.2.5 Definition** (Regular CTMP). A CTMP is regular if each state is reachable from every other state, i.e.,  $P^k > 0$  for some  $k$  (recall the Definition 2.4.4 and see Appendix A.1).

**3.2.6 Definition** (Transition Rates). Given a time-homogeneous and regular CTMP, the right derivative at  $t = 0$  is defined as

$$Q_{ij} = \left. \frac{d}{dt}(P_{ij}(t)) \right|_{t=0^+} = \lim_{t \rightarrow 0^+} \frac{P_{ij}(t) - P_{ij}(0)}{t} \quad \text{and} \quad Q_{ij} = \begin{cases} -q_{ii} & \text{if } i = j \\ q_{ij} & \text{if } i \neq j \end{cases}. \quad (3.2.4)$$

For small time  $t$  the transition probabilities are as follows:

$$P[G(t) = j | G(0) = i] = q_{ij}t + o(t), \quad (3.2.5)$$

$$P[G(t) = i | G(0) = i] = 1 - q_{ii}t + o(t). \quad (3.2.6)$$

Therefore the probability of exiting from state  $i$  is  $q_{ii}t$ , and the probability of transitioning to state  $j$  is  $q_{ij}t$  for all  $j \in \mathcal{G}$ . Therefore,  $q_{ii} = -\sum_{j=1}^N q_{ij}$ .

The waiting time, or sojourn time, in state  $i$  is exponentially distributed with parameter  $-q_{ii}$  (Van Miegheem, 2009). It then exits and transitions to state  $j$  with probability  $p_{ij} = \frac{q_{ij}}{q_{ii}}$ ,  $i \neq j$ . Again, it waits in state  $j$  for a time that is exponentially distributed with parameter  $-q_{jj}$  and so the process continues. These successive transitions from states  $\in \mathcal{G}$  form a Markov process. Provided there are no *absorbing states*, such that  $p_{ii} = 1$  and  $p_{ij} = 0$  (i.e., regular CTMP), the process will continue indefinitely. Hence a condition on running time is required (given in Section 4.1).

**3.2.7 Definition** (Infinitesimal Generator). The matrix  $Q = (q_{ij})$  is called the generator matrix of the Markov process or the infinitesimal generator.  $Q$  has the following properties for all  $i \in \mathcal{G}$ :

1.  $-\infty < q_{ii} < 0$
2.  $q_{ij} \geq 0 \quad i \neq j$
3.  $\sum_{j \in \mathcal{G}} q_{ij} = 0$

The transition rates are defined as

$$q_{ij} = \begin{cases} q_{ii}p_{ij} & i \neq j \\ -\sum_{i \neq j} q_{ii}p_{ij} & i = j \end{cases}, \quad (3.2.7)$$

(Serfozo, 2012). Such a formulation is made for the SAOM (see Section 4.0.3).

This section has defined a continuous-time Markov process and outlined important assumptions of the SAOM. Further information is detailed in (Taylor and Karlin, 1998) and (Van Miegheem, 2009). Prior to introducing the model, an example is explored.



### 3.3 Example

The directed graph in Figure 3.1 represents a CTMP with 5 states,  $\{1, 2, 3, 4, 5\}$ , given by the graph vertices. The direction and weight of the graph edges represent the transition probabilities.

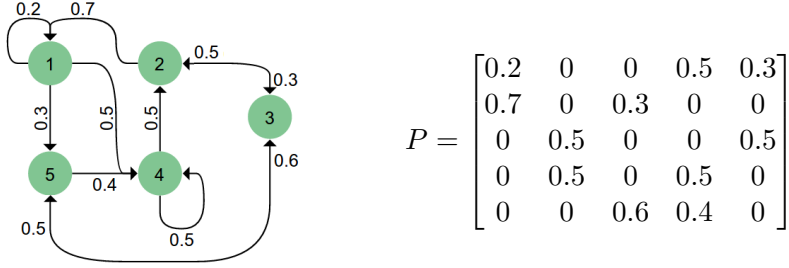


Figure 3.1: Example of CTMP with 5 states (left) and corresponding transition matrix,  $P$  (right).

Consider the system initialised to state 2. The probability of transition to state 1 is  $p_{21} = 0.7$ , and to state 3 is  $p_{23} = 0.3$ .

Whilst the transition matrix  $P$  (or embedded Markov chain) completely determines the probabilistic nature of a discrete-time Markov chain, it does not consider the continuous-time rates of transition. Let each of the states,  $\{1, 2, 3, 4, 5\}$ , remain in their respective states for a time exponentially distributed with mean 10, 10, 10, 5, 5 units of time, respectively. The generator matrix (as per Section 3.2.7) is constructed as

$$Q = \begin{bmatrix} -\sum_{i \neq j} q_{11}p_{1j} & q_{11}p_{12} & q_{11}p_{13} & q_{11}p_{14} & q_{11}p_{15} \\ q_{22}p_{21} & -\sum_{i \neq j} q_{22}p_{2j} & q_{22}p_{23} & q_{22}p_{24} & q_{22}p_{25} \\ q_{33}p_{31} & q_{33}p_{32} & -\sum_{i \neq j} q_{33}p_{3j} & q_{33}p_{34} & q_{33}p_{35} \\ q_{44}p_{41} & q_{44}p_{42} & q_{44}p_{43} & -\sum_{i \neq j} q_{44}p_{4j} & q_{44}p_{45} \\ q_{55}p_{51} & q_{55}p_{52} & q_{55}p_{53} & q_{55}p_{54} & -\sum_{i \neq j} q_{55}p_{5j} \end{bmatrix}.$$

Where  $q_{ii}$  is the exponential parameter of waiting times. Hence,  $\frac{1}{q_{ii}} = [10, 10, 10, 5, 5]$  is the mean waiting time, for  $i = \{1, 2, 3, 4, 5\}$ . Therefore

$$Q = \begin{bmatrix} -0.08 & 0 & 0 & 0.05 & 0.03 \\ 0.07 & -0.1 & 0.03 & 0 & 0 \\ 0 & 0.05 & -0.1 & 0.05 & 0 \\ 0 & 0.1 & 0 & -0.1 & 0 \\ 0 & 0 & 0.12 & 0.08 & -0.2 \end{bmatrix}.$$

Where  $Q$  satisfies the properties in Equation 3.2.7.

## 4. The Model

The SAOM is implemented in R (Muller, 2017) and the results are detailed in Section 5.2. This section outlines the model approach with regards to 2 observed time points. The model can be extended further for up to 10 time points (Snijders et al., 2010). In this case, the assumption of time-independent coefficients of network effects needs to be tested.

Let  $G(t)$  be a time-dependent simple directed graph with Markovian properties as per Section 3.1. Observations are made at  $G(t_0)$  and  $G(t_1)$ . Recall the vertices  $V(G)$  are independent of time.

**4.0.1 Rate Parameter.** The rate parameter,  $\lambda(G, i)$ , is the rate at which each actor is expected to change their friendship ties. Snijders proposes a rate parameter that is a function of covariates and network effects, including out-degree, in-degree and reciprocated relations (Snijders, 2005). For example, actors with a high out-degree may change their friendship ties at a faster rate based on wanting to know a lot of people. However, for simplicity, this model considers a constant rate function  $\lambda(G, i) = \rho_0 > 0$  for all  $i \in V(G)$ .

**4.0.2 Objective Function.** The objective function  $f(i, G(i \rightsquigarrow j))$  is a linear combination of network effects,  $\rho_s$ . It determines the associated value actor  $i$  gives to the network configuration  $G(i \rightsquigarrow j)$  for all  $j \neq i$ . That is, the network  $G$ , subject to the following change:  $A_{ij} = 1 - A_{ij}$ . This is the perceived future benefit of making or breaking a tie. For  $i = j$ , define  $G(i \rightsquigarrow i) = G$ . Actor  $i$  makes the choice to change to  $j$  based on maximising the function

$$f(i, G(i \rightsquigarrow j)) = \sum_{s=1}^L \beta_s \rho_s(G(i \rightsquigarrow j), i), \quad (4.0.1)$$

where  $L$  denotes the number of effects in the model,  $\rho$  is the vector of network effects with elements  $\rho_s$  (as per Section 2.2.1), and  $\beta$  is the vector of network effect coefficients with elements  $\beta_s$ .

It is important to note that each  $\beta_s$  is the same for all actors. Thus, the model does not determine the network effects relative to the actor, but rather relative to the network structure as a whole. Conclusions are thus made on the whole network, and statements referring to individuals in the network cannot be deduced. The simulated networks are thus attempts at graphs isomorphic to  $G(t_1)$ , and the network difference (Section 2.4) is a statistical measure of how close the simulations are to isomorphic.

Snijders presents the choice model as a *myopic stochastic optimisation rule* where actor  $i$  changes tie to  $j$  based on maximising

$$\max_{j \in V(G)} (f(i, G(i \rightsquigarrow j)) + U(j)), \quad (4.0.2)$$

where  $U(j)$  is an unexplained variable for  $i$ 's attraction to  $j$  not captured by the model.  $U(j)$  is assumed to be Gumbel distributed (extreme value type 1) with mean 0 and scale parameter 1. Under this assumption, the probability that actor  $i$  changes tie with actor  $j$  can be modelled as multinomial logistic regression (see Appendix A.3) with probability given by (Snijders, 2005)

$$p_{ij}(G) = \frac{\exp(f(i, G(i \rightsquigarrow j)))}{\sum_{h=1, h \neq i}^n \exp(f(i, G(i \rightsquigarrow h)))}, \quad j \neq i, \quad (4.0.3)$$

$$p_{ij}(G) = \frac{\exp(f(i, G(i \rightsquigarrow j)) - f(i, G))}{\sum_{h=1, h \neq i}^n \exp(f(i, G(i \rightsquigarrow h)) - f(i, G))}. \quad (4.0.4)$$

**4.0.3 Markov Process.** A Markov Process is completely defined by the space of all possible states  $\mathcal{G}$ , the initial state and the transition rate matrix  $Q$ . The evolution of the social network,  $G(t)$ , has transition rates

$$q_{ij}(G) = \lambda p_{ij}(G), \quad (4.0.5)$$

where  $\lambda$  is analogous to  $q_{ii}$  (see Equation 3.2.7). Note that the model is dependent on unknown parameters  $\hat{\theta} = (\hat{\lambda}, \hat{\beta})$ . These are estimated inputs, along with the initial state,  $G(t_0)$ , and selected network effects,  $\rho$ , for the Markov process:

1. Initialise:  $t = 0, G(t_0), \rho$  and  $\hat{\theta}$ .
2. Sample  $i$  from uniform distribution over vertices  $V(G)$  (since constant  $\lambda$ ).
3. Given actor  $i$ , sample  $j$  with probability  $p_{ij}(G)$ .
4. Let  $t = t + \Delta t$  for  $\Delta t$  sampled exponential random variable with parameter  $\hat{\lambda}_+ = n\hat{\lambda}$ .
5. Change network  $G(t)(i \rightsquigarrow j)$  and denote  $H(t)$ .
6. Repeat step (b) until  $t = T_1$  (see Section 4.1.1).

Denote the final output  $H(T_1)$ .  $H$  is therefore dependent on  $T_1, G(t_0), \rho$  and  $\hat{\theta}$ .

The Markov process is used in the estimation of parameters  $\hat{\theta} = (\hat{\lambda}, \hat{\beta})$ . The following section outlines the method of estimation.

## 4.1 Estimation

The model uses methods of moments estimation (Snijders, 2001) which is described in the following section. The basic principle is to find a relevant  $Z$  statistic that captures the variability in the data for the given parameters, and equate the expectation of the estimated values to the observed values from  $G(t_1)$

$$E[Z(H(t, \hat{\theta})) | G(t_0), \rho] = z^{obs}. \quad (4.1.1)$$

The chosen statistics are  $Z = (C(t), \mathbf{P}(t))$ , where  $C$  is a measure of distance between two networks. Let  $H$  and  $G$  be simple digraphs with corresponding adjacency matrices  $X$  and  $Y$ , and let  $V(G) = V(H) = n$ . The distance between  $H$  and  $G$  is (Snijders, 2005):

$$\|H - G\| = \sum_{1 \leq i, j \leq n} |X_{ij} - Y_{ij}|. \quad (4.1.2)$$

The vector  $\mathbf{P}$  contains elements  $P_s$  for all  $s \in [1, L]$ . The  $Z$  statistics are

$$C = \|H(t) - G(t_0)\|, \quad P_s = \sum_{i=1}^n \rho_s(H(t), i), \text{ for } s \in [1, L], \quad (4.1.3)$$

$$c^{obs} = \|G(t_1) - G(t_0)\|, \quad p_s^{obs} = \sum_{i=1}^n \rho_s(G(t_1), i), \text{ for } s \in [1, L]. \quad (4.1.4)$$

Small case letters are used to denote observed statistics,  $c$  and  $\mathbf{p}$ , and capitals to denote simulated,  $C$  and  $\mathbf{P}$ . The statistic  $c$  is the distance between two networks and measures the number of changes in edges. This is reasonably related to the rate parameter since the number of changes divided by number of actors, gives the average number of changes per actor, which is proportional to the rate of change.

The statistics  $\mathbf{P}$  are a means of measuring the network similarity based on structure. The method of moments requires that the summation of each network effect is the same for the simulated and observed networks. Since the  $\beta$  parameters determine the structure of the simulated network,  $\mathbf{P}$  is sensitive to changes in  $\beta$ . Section 5.2.1 investigates how well these parameters account for the overall network structure of the AIMS network.

The moment equations are

$$E[C(H(t, \hat{\theta})|G(t_0), \rho)] = c, \quad (4.1.5)$$

$$E[\mathbf{P}(H(t, \hat{\theta})|G(t_0), \rho)] = \mathbf{p}. \quad (4.1.6)$$

**4.1.1 Conditional Moment Estimation.** Snijders modifies the method of moments by conditioning on the parameter  $c$  in the Markov Process (Snijders, 2001). Let  $T_1 = \min\{t|C(t) \geq c\}$ . The Markov process then loops until  $C \geq c$ , i.e., the simulated network should not have less changes than the observed network. The equation to solve is now reduced to

$$E[\mathbf{P}(H(t, \hat{\theta})|G(t_0), \rho, C)] = \mathbf{p}, \quad (4.1.7)$$

which is independent of  $\hat{\lambda}$  and can therefore be written as

$$E[\mathbf{P}(H(t, \hat{\beta})|G(t_0), \rho, C)] = \mathbf{p}. \quad (4.1.8)$$

The conditional expectation cannot be calculated explicitly and therefore a stochastic approximation is implemented (Snijders, 2005).

**4.1.2 Stochastic Approximation.** Snijders uses an updated version of the Robbins-Monro method to iteratively update the parameters

$$\hat{\beta}_{N+1} = \hat{\beta}_N - a_N D_0^{-1}(\mathbf{P}(H(t, \hat{\beta}_N)) - \mathbf{p}), \quad (4.1.9)$$

where  $N$  is an index for steps in the Markov process,  $a_N$  is a series that slowly converges to 0, and  $D_0$  is the diagonal of the derivative matrix:  $D_\beta = \frac{\partial E[P_s(\hat{\beta})]}{\partial \hat{\beta}_s}$ . This is a Newton method of gradient descent for the non-linear function  $E[\mathbf{P}(H(t, \hat{\beta}_N))] - \mathbf{p} = 0$  (Equation 4.1.8).

For further reading please refer to (Snijders, 2013), (Snijders, 2001) and (Robbins, 1951).

**4.1.3 Simulation Investigation for Empirical Network Analysis (SIENA).** The implementation of this model is done in R using the package RSiena. The documentation provides methods of ensuring convergence (Ripley et al., version March 21, 2017). Important assumptions to note are:

1. Rate parameter is assumed constant
2. Method of Moments has been used for stochastic approximation (as described above, other methods available in the package).

## 5. Results & Discussion

### 5.1 Data and Cross-Sectional Interpretations

The data collected is from the AIMS 2016/17 Structured Masters cohort. 50 students live together in permanent residence for a period of 11 months, where they eat, sleep and learn together. Amongst the students are tutors, lecturers and resident researchers who do not remain permanently in the building in the year. Consider this group of students to be a closed group of actors (Wasserman and Faust, 1994).

The data was collected in December 2016,  $t_0$ , 3 months after enrolment, and again in April 2017,  $t_1$ . The networks are digraphs,  $G(t_0)$  and  $G(t_1)$ , with vertices  $i \in V(G)$  representing students, and the directed edges  $E(t)$  represent a friendship nomination. Each vertex has associated a covariate vector  $v_i^c = [sex, country]$ , with discrete variables  $sex \in [0, 1]$  and  $country \in [1, 15]$ .

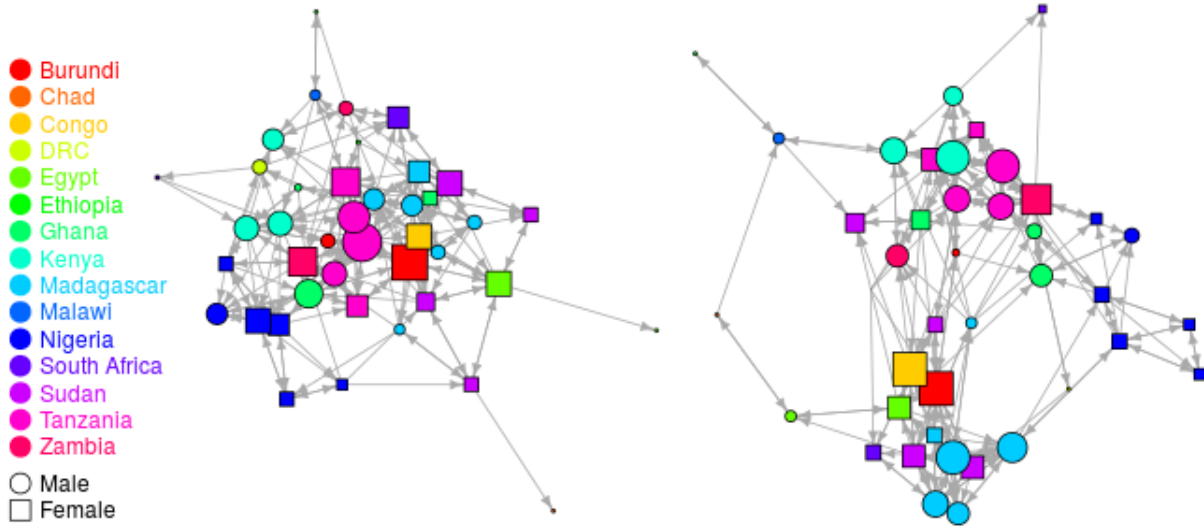


Figure 5.1: AIMS Network: December (left) and April (right). The size of each node is proportional to in-degree. The shape of the node represents the actors sex, and the colour of nodes represent *country*.<sup>1</sup>

The students were required to nominate those other students, in the AIMS cohort, whom they considered their closest friends. A friend was exemplified as a person one wishes to spend time with outside of school, and/or share details of their life with. The number of friendship nominations was not restricted in the questionnaire, and for the purpose of analysis, only the first 8 nominations are considered. The underlying assumption is that the most important relations are captured in the first responses of the individual (de Nooy et al., 2005).

The total number of AIMS students in the cohort is 50 and the AIMS friendship network contains  $n = 41$  actors. 9 of the students chose not to participate in the study. These actors have been treated by removing them from the data-set and thus assuming the 41 actors to be the complete network.

<sup>1</sup> Plotted in R using `layout_nicely()` in `plot.igraph` package.

Table 5.1: Descriptive Statistics including Global Metrics for AIMS Network

	December	April
Number of Nodes	41	41
Number of Edges	212	203
Average Degree	5.17	4.95
Density $\mu_1$	0.13	0.12
Reciprocity $\mu_2$	0.58	0.58
Global Clustering $\mu_3$	0.41	0.62
Harmonic Mean Distance $\mu'_4$	2.24	2.43

Table 5.1 shows that the AIMS network has increased in clustering and distance from December to April, and that reciprocity has remained constant. Density has decreased, due to a total of 11 fewer ties. Increase in clustering and distance suggests clustering within groups. This is illustrated in Figure 5.1 by the appearance of two subgroups in the April network.

The density plots in Figure 5.2 allow for comparison of local metrics on the AIMS network at each observation. Out-degree shows a change in dispersion, with most actors choosing to nominate few, less than 4, or many, more than 4, friends. In-degree has fewer nominations of degree 7, otherwise, it has not changed significantly. A relationship between local clustering and harmonic closeness is observed. It appears that actors with high closeness centrality correspond to actors with low local clustering. This is reasonable, since an actor who has friends who are less connected to one another, but connected to different actors, is more likely to reach a greater proportion of the network and hence have a greater closeness centrality. In April, there is a decrease in the most frequently observed values for clustering and closeness and the values become more dispersed. Particularly, there exists nodes with large values of local clustering in April.

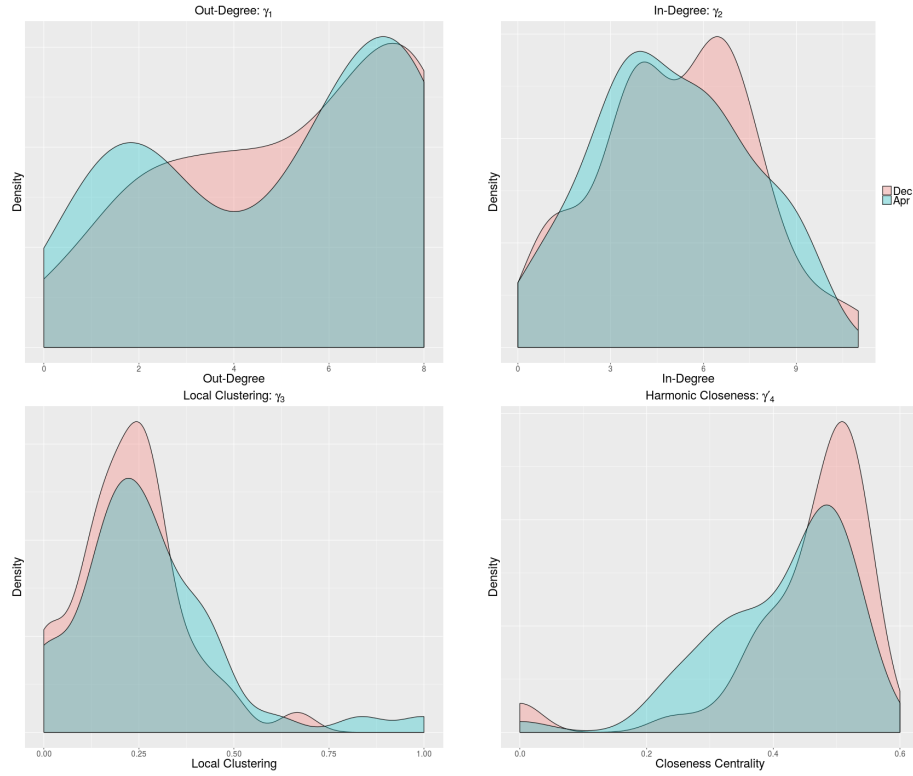


Figure 5.2: Density Plots for Local Metrics on AIMS network.

The network has changed by a total of  $c = 169$  ties from December to April. This appears a lot since there are a total of 203 ties in the April network. However, recall the distance between two networks measures the presence of new ties **and** the absence of old ties. Therefore, each tie change contributes  $+2$  to the distance  $c$ . Based on a fairly constant density, approximately 85 friendship selections have changed, which is an average of 2 per actor. The Jaccard coefficient (see Appendix A.4) for this period is 0.4, indicating sufficient changes have occurred for the model to have relevant meaning (Snijders et al., 2010).

## 5.2 SAOM Results

The data has been analysed using the RSiena package and the fitted models including effects, estimates, standard errors,  $p$ -values and  $t$ -test for convergence are given in Table A.1. The models show good convergence of the stochastic approximation method since all  $t$ -ratios  $< 0.1$  (Snijders, 2005).

The model is fitted using forward and backward selective steps, adding and removing effects, and tested for significance at each step (Snijders et al., 2010).

Model I begins with a backward step, including multiple effects guided by the theory. Model I includes the following effects: reciprocity, transitive triplets, transitive ties, 3-cycles, balance, number distance 2, and ego, alter and homophily effects corresponding to the covariate *sex*, and homophily effect only corresponding to *country* (since country is nominal). Transitive triplets indicates connectivity in the network (transitive closure) whilst also taking into account hierarchy. Transitive ties also measures transitive closure but without considering the number of triplets formed. 3-cycles also measures connectivity,

but it implies an egalitarian structure (Snijders et al., 2010). Balance detects cliques and number of actors distance two is a measure of connectivity (a negative parameter indicates the tendency towards connectivity, i.e., away from distance 2).

Each effect is tested for significance under the hypothesis,  $H_0 : \hat{\beta}_s = 0$ , with normally distributed  $t$ -statistic,  $t_s = \frac{\hat{\beta}_s}{s.e(\hat{\beta}_s)}$  (Ripley et al., version March 21, 2017), where  $s.e(\hat{\beta}_s)$  is the standard error of  $\hat{\beta}_s$ . The column corresponding to  $p$ -value, in Table A.1, shows the significance of each effect based on the  $t$ -tests. Backward selection is applied and a second model is fitted by removing all effects which are not significant. Forward selection is applied: one by one the deleted effects are tested for significance using the score-type test. The model is restricted by the tested effect (assuming it to be zero,  $\hat{\beta}_0$ ) and the restricted model is compared to the fitted model under the hypothesis:  $H_0 : \hat{\beta}_s = \hat{\beta}_0$  with test statistic  $S(\hat{\beta}_0)$  (see Appendix A.5). This is done for other network effects,  $\rho_s$ , outlined in Section 2.2.1, and further effects in RSiena (Ripley et al., version March 21, 2017). Model II includes all significant effects after backward selection, forward selection and score-type testing.

The rate parameter is an estimate of the number of opportunities each actor  $i$  receives to change an outgoing tie. This is much greater than the observed number of changes made per actor since it considers non-action and reversal of ties. It is not of importance in the model (recall conditional estimation in Section 4.1).

Model II excludes sex related effects since they do not appear to play a significant role in the evolution of the AIMS network. It confirms the significance of reciprocity and transitivity, which are known properties of social networks. The presence of transitive closure and absence of 3-cycles may be indicative of local hierarchy and therefore forward selection and score-type tests were carried out for the endogenous degree-related effects. These were not significant. The presence of balance indicates cliques and a negative parameter for number distance 2 implies transitive closure. Country related homophily is significant in the model and this is illustrated in Figure 5.1 amongst the light blue (*Malagasy*), dark blue (*Nigerian*) and pink (*Tanzanians*) vertices.

Model III has been fitted, using the same network effects, to confirm model II estimates. The objective function for Model III can be written explicitly as follows:

$$f(i, G(i \rightsquigarrow j)) = \beta_1 \rho_1 + \beta_3 \rho_3 + \beta_4 \rho_4 + \beta_6 \rho_6 + \beta_8 \rho_8^* + \beta_{15}^{country} \rho_{15}^{country} \quad (5.2.1)$$

where the  $\rho_s$  effects are as per Section 2.2.1 (modification to effects in RSiena denoted  $\rho_s^*$  and are defined in Appendix A.6). The contribution of each effect is relative to the size of parameter estimate, except that the coefficients are unstandardised (inhomogeneous variance).

**5.2.1 Differences.** Using the fitted model, the MCMC simulates  $N = 1000$  networks. Denote the sample of networks as follows  $\{H_i(T_1) | 1 \leq i \leq N\}$ . The aim is to apply the difference tests, as per Section 2.6, as a measure of goodness of fit of the model.

## 5.3 Univariate Network Differences

Recall the null hypothesis,  $H_0 : \bar{\mu} = \mu^{obs}$ .

Independence of the sample for each metric is assumed, homogeneous variance is not required since there is only one sample and normality is tested using the Shapiro-Wilk test. All metric samples reject the null hypothesis of normal distribution. Therefore, the samples have been tested using nonparametric



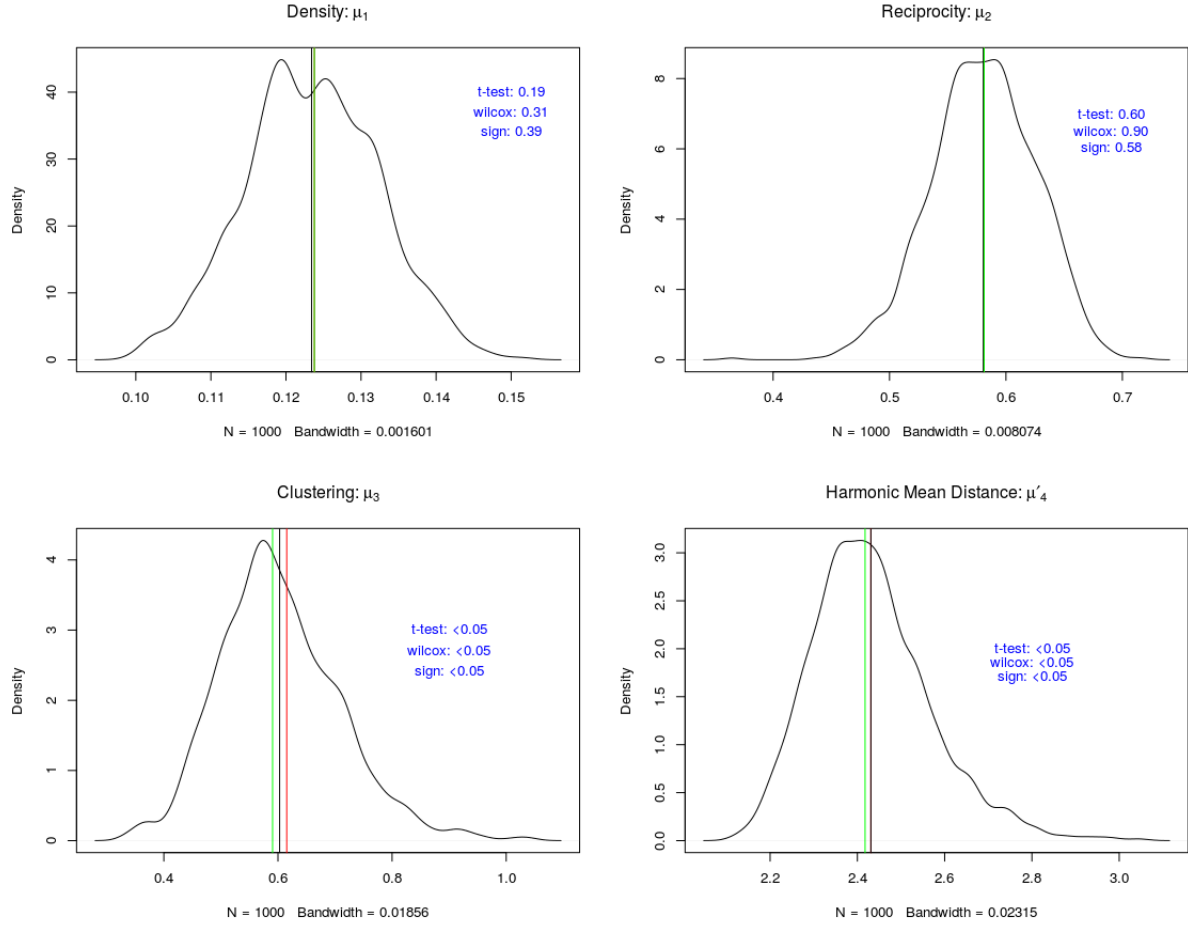


Figure 5.3: Density plots for difference metrics,  $\mu_q$ . Red line is  $\mu^{obs}$ , black line is  $\bar{\mu}$  and green line is  $\mu^{median}$ .

tests: Wilcoxon and Signed-Rank (see Appendix A.7). The null hypothesis under these tests is that the sample median does not differ from the observed metric. The assumption of Wilcoxon does not require normality whilst it does require symmetric distribution about the median, i.e.,  $\bar{\mu} = \mu^{median}$ . Signed-Rank test does not require symmetry about the median and therefore the null hypothesis is  $H_0 : \mu^{median} = \mu^{obs}$ . The power of tests is highest for  $t$ -test and lowest for signed rank test. The tests are in agreement for the metrics.

Figure 5.3 shows that density and reciprocity fail to reject the null hypothesis and harmonic mean distance and clustering reject the null hypothesis. The graph shows the data under  $\mu_4'$  is not symmetric and hence the signed-rank test is the most appropriate. However, rejecting the null hypothesis  $H_0 : \mu_4'^{median} = \mu_4'^{obs}$  does not tell us about the mean  $\bar{\mu}_4'$ . Using the graphical representation of  $\mu_4'^{obs}$  and  $\bar{\mu}_4'$ , distance is considered adequately fit. Using the same approach for clustering we observe a difference between  $\mu_3^{obs}$  and  $\bar{\mu}_3$  and conclude that global clustering is not well fit.

Recall the statistics,  $p_s$ , used in the estimation method (detailed in Section 4.1). Table 5.2 shows the values of  $p_s$  for the observed network and the average  $P_s$  for the simulated networks.

The estimates,  $P_s$ , of out-degree and reciprocity are the same (or similar) value as the targets,  $p_s$ . This explains why the metrics for density and reciprocity fail to reject the null hypothesis; the condition of

Table 5.2:  $P_s$  and  $p_s$  statistics for network effects,  $\rho_s$ , for Model III

Effect	$\rho_s$	Target $p_s$	Mean Estimate $P_s$
Out-degree	$\rho_1$	203	202
Reciprocity	$\rho_3$	118	118
Transitive triplets	$\rho_4$	360	355
Number Distance 2	$\rho_6$	429	429
Balance	$\rho_8$	399	405
Same country	$\rho_{15}^{country}$	65	65

zero difference for these metrics is already built in to the model. Note that clustering is dependent on transitivity, but it differs from number of distance 2 by considering nodes  $i, k$  connected by intermediary  $j$  to be a path of length 2 away, even if there exists an edge  $A_{ik} = 1$ . Therefore, clustering is not built into the model. Harmonic distance is not built into the model, however, it is well fit in the model.

## 5.4 Multivariate Network Differences

Recall the null hypothesis,  $H_0 : \bar{\gamma} = \gamma^{obs}$ .

Independence of the sample for each metric is assumed, homogeneous variance is not required since there is only one sample and multivariate normality is tested, of which each metric sample is not multivariate normally distributed. We refer to the multivariate nonparametric sign test (see Appendix A.8). The sign test assumes distribution free (nonparametric) given directional symmetry (Oja and Randles, 2004). Therefore, the null hypothesis under the sign test is  $H_0 : \bar{\gamma} = \gamma^{obs} = \gamma^{median}$ .

The Hotellings  $T^2$  is not adequate for non-normal multivariate distributions, however, the  $p$ -values have been included for comparisons to the sign test. The sign test is not adequate for non-symmetrical distributions.

Sample tests contain a certain proportion of the data; a method of treating outliers and attaining a more symmetrical distribution. In-degree and out-degree distribution,  $\gamma_1$  and  $\gamma_2$ , have variables  $m \in \{0, 1, \dots, 10\}$ , which contains 96% and 93% of the counts respectively. For local clustering,  $\gamma_3$ , and harmonic closeness,  $\gamma'_4$ , and all ordered variables  $n = 41$  are included. The violin plots shown represent this trimmed data.

All metrics fail to reject the null hypothesis for the nonparametric test, except harmonic closeness. Out-degree rejects the null hypothesis under the Hotelling's  $T^2$  test. The appropriateness of these tests is uncertain on data which cannot be assumed to be symmetrically distributed. Therefore, the main point of inference is the violin plots themselves.

In-degree appears to be well fitted to the data. The simulated out-degree distribution does not represent well the non-linear cumulative trend of the observed out-degree distribution. Harmonic closeness appears to be represented well by the simulations, except for the three least central nodes. Local clustering appears adequately fit, except for nodes with higher clustering. The underestimated value for these nodes is likely to be the leading cause for the ill-fitted metric,  $\mu_3$ . These nodes are not necessarily those that form part of a larger subgroup, but rather, nodes which have few relations, and hence a local clustering is sensitive to individual tie changes. This ties in to the observation for harmonic closeness, as the least central nodes — inevitably those with fewer relations — have underestimated centrality value

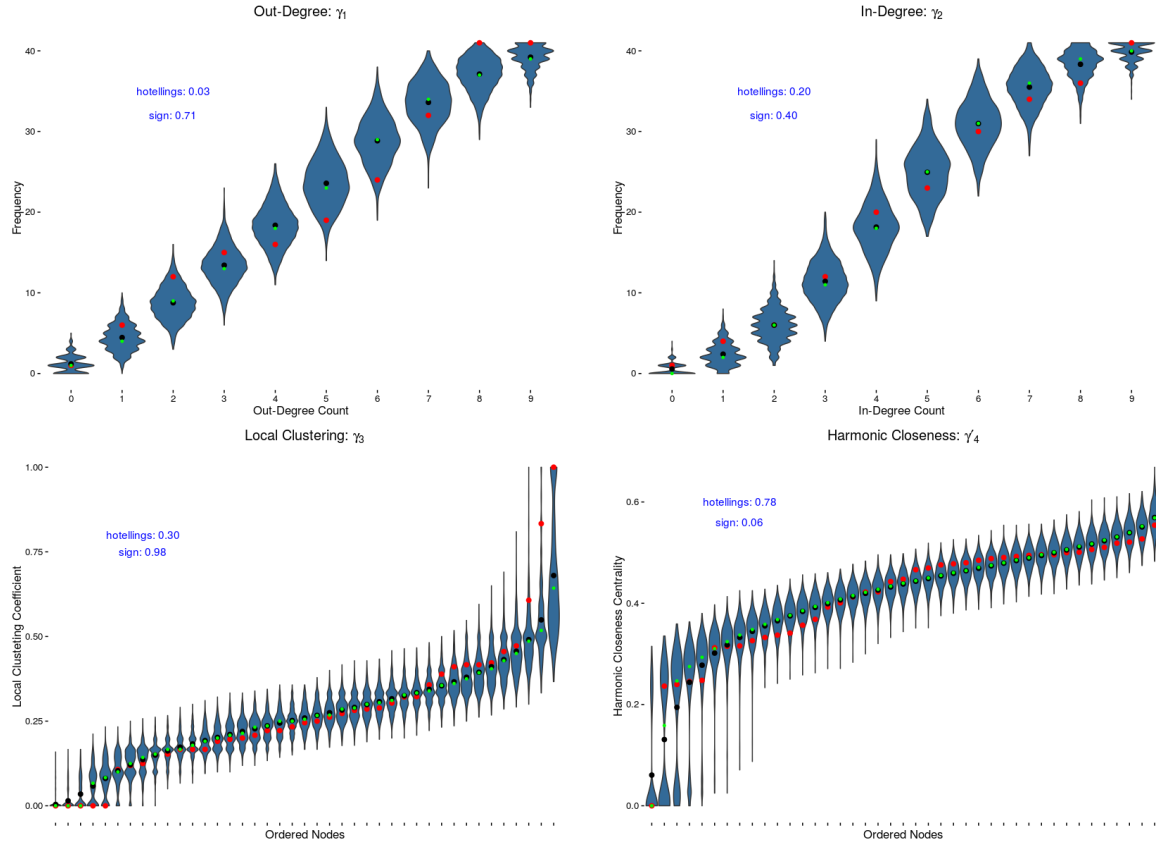


Figure 5.4: Violin Plots for difference metrics,  $\gamma_r$ , red dots are  $\gamma^{obs}$  and green dots are  $\gamma^{median}$ . The blue shaded region (violin) represents the sample data for each variable. In-degree and out-degree show cumulative degree distribution. The width of the violin is proportional to the density and the length is proportional to the range. Symmetry of the data is indicated by symmetrical violins on the horizontal and vertical axis.

since they are also sensitive to individual tie changes. Recall, the correlation observed between local clustering and harmonic closeness for the AIMS network in December and April. It is probable that the ill-fitted actors, for both local clustering and harmonic closeness, are indeed the same. Nevertheless, the trend in clustering and harmonic closeness for the simulated network is similar to the observed trend, for the majority of the data.

## 6. Conclusion

The evolution of the AIMS network has been analysed using the SAOM. Reciprocity, transitive triplets, balance, number distance 2 and homophily related to *country* have been found to be significant effects determining the evolution of the network from December,  $G(t_0)$ , to April,  $G(t_1)$ . All these effects are indicative of social grouping. The structure of grouping implied by the model has been investigated using the metrics defined. The model is considered well fitted in terms of global density, reciprocity and harmonic mean distance. However, the model underestimates global clustering in the network. The model is well fitted in terms of local in-degree. It fails to establish the non-linear trend of out-degree. Furthermore, although the trend of local clustering and harmonic closeness is well fitted, values at the extreme are underestimated.

The inference power of the model is analysed by its predictive ability. Based on the defined metrics, the model fitted to the AIMS data simulates well the evolved network, except for values at the extreme. Possible methods for treating these vertices should be considered. It must be highlighted that this work considers only 8 metrics of network topology and can therefore be extended by including further metrics.

As highlighted in Section 2.2.1, homophily effects are prominent between race, ethnicity, age, religion etc., however, the data includes only country and sex. Additional covariates may improve the fit of the model. Furthermore, data collection for additional time points will more accurately determine the evolution of the network and add to its predictive power.

This model used in this work was simple: a constant rate parameter and constant covariates. There exists possibility to extend the model to include degree dependent rate parameters, time-dependent covariates and dyadic covariates (network related covariate such as *previous friendship*) (Snijders, 2005).

This work should be extended by means of data collection. For example, the AIMS social network can be recorded periodically in the year for multiple cohorts, in a controlled manner. Additional data can be considered and an extended period of analysis can allow for more reliable predictive models. Research questions and data collection can be expanded to various organisational institutions and community structures, such as schools, universities, work environments, care homes etc.

# A. Appendix

## A.1 Powers of Adjacency Matrix

Let  $G$  be a simple directed graph with corresponding adjacency matrix  $A_{ij}$ . Recall  $A_{ij} = 1$  if there exists a path from  $i$  to  $j$ . Therefore,  $A_{ik}A_{kj} = 1$  if there exists a path from  $i$  to  $j$  via  $k$ .  $N_{ij}^2$  is the total number of paths length 2 from  $i$  to  $j$  via  $k$  for all  $k \in V(G)$ .

$$N_{ij}^2 = \sum_{k=1}^n A_{ik}A_{kj} = [A^2]_{ij},$$

which is generalisable for all paths of length  $r$  to (Newman, 2010)

$$N_{ij}^r = [A^r]_{ij}.$$

## A.2 Mahalanobis Distance

Mahalanobis distance is a measure of distance between a distribution,  $\mathbf{X}_1, \dots, \mathbf{X}_n$ , and a point  $\mu$ . It considers  $p$ -variate distribution for each i.i.d  $\mathbf{X}_i$ , i.e.,  $\mathbf{X}_i$  is a vector of dimension  $p$ . It accounts for the correlation between variables and heterogeneous variance by decorrelating and standardising the distribution:  $z = L^{-1}(\bar{\mathbf{X}} - \mu)$ , where  $\bar{\mathbf{X}} = \sum_{i=1}^n \mathbf{X}_i$ .  $L$  is the Cholesky factor such that  $S = LL^{-1}$  for covariance matrix  $S$  (Press, 2007). Note that  $S$  is required to be hermitian and positive semi-definite for such a transformation to exist, which holds for all covariance matrices. The squared Mahalanobis distance is:

$$z^T z = (\bar{\mathbf{X}} - \mu)^T S^{-1} (\bar{\mathbf{X}} - \mu).$$

Note that in the univariate case, Mahalanobis distance reduces to Euclidean distance.

## A.3 Proof of Multinomial Logistic Regression

Denote  $Y_j^*$  the level of utility associated to the choice  $j$ . Then based on stochastic myopic optimisation

$$Y_j^* = \max_j (f(i, G(i \rightsquigarrow j)) + U_j),$$

for all  $j \neq i$ .  $Y_j = 1$  when  $Y_j^* = \max_j$  and 0 otherwise. Denote  $\max_j f(i, G(i \rightsquigarrow j))$  as  $f_j$  then  $Y_j^* = f_j + U_j$ . Assume  $U_j$  are i.i.d Gumbel distributed then the cumulative distribution and probability density function can be expressed as

$$\begin{aligned} F(U_j < U) &= \exp(-e^{-U}), \\ f(U_j) &= \exp(-U_j - e^{-U_j}). \end{aligned}$$

Based on the maximising utility, it is required that  $f_j + U_j > f_k + U_k$  for all  $j \neq k$ , which can be rewritten  $U_j + f_j - f_k > U_k$ . Hence

$$\begin{aligned} Pr(Y_j = 1) &= Pr(U_k < U_j + f_j - f_k), \quad \text{for all } k \neq j \\ &= \int_{-\infty}^{\infty} \prod_{k \neq j} F(U_j + f_j - f_k) f(U_k) dU_k, \end{aligned}$$

where

$$\begin{aligned} \prod_{k \neq j} F(U_j + f_j - f_k) f(U_j) &= \prod_{k \neq j} \exp(-e^{-U_j - f_j + f_k}) \exp(-U_j - e^{-U_j}) \\ &= \exp \left[ -U_j - e^{-U_j} \left( 1 + \sum_{k \neq j} \frac{e^{f_k}}{e^{f_j}} \right) \right], \\ \implies \Pr(Y_j = 1) &= \int \exp \left[ -U_j - e^{-(U_j - \lambda_j)} \right] dU_j, \end{aligned}$$

where  $\lambda_j = \log \left( 1 + \sum_{k \neq j} \frac{e^{f_k}}{e^{f_j}} \right) = \log \left( \sum_k \frac{e^{f_k}}{e^{f_j}} \right)$ . Let  $U_j^* = U_j - \lambda_j$ , then

$$\begin{aligned} \Pr(Y_j = 1) &= \exp(-\lambda_j) \int \exp(-U_j^* - e^{-U_j^*}) dU_j^* \\ &= \exp(-\lambda_j) \\ &= \sum_k \frac{e^{f_j}}{e^{f_k}}. \end{aligned}$$

This is a discrete choice model formalised by (Maddala, 1983).

## A.4 Jaccard Coefficient

The Jaccard Coefficient measures the amount of change between two consecutive network observations (Snijders et al., 2010):

$$\frac{N_{11}}{N_{11} + N_{01} + N_{10}},$$

where  $N_{11}$  is the number of times present in both waves,  $N_{01}$  is the number of ties newly created and  $N_{10}$  the number of ties terminated.

## A.5 Rao Score Test

The null hypothesis under this test is  $H_0 : \theta = \theta_0$  for univariate parameter  $\theta$ . The power of the score-type test is that it does not require an estimate of the parameter under the alternate hypothesis. Restricting the model by the parameter to be tested is the null hypothesis. The following definition is taken from (Engle, 1984).

The test statistic is  $S(\theta_0) = \frac{U(\theta_0)^2}{I(\theta_0)}$ . Where  $U(\theta) = \frac{\partial \log L(\theta|x)}{\partial \theta}$ ,  $x$  is the data and  $I(\theta) = -E \left[ \frac{\partial^2}{\partial \theta^2} \log L(x; \theta) | \theta \right]$  is the Fischer information.  $S(\theta_0)$  has asymptotic distribution of  $\chi_1^2$  under  $H_0$ .

## A.6 Network Effects

The network effects are defined in RSiena as follows (Ripley et al., version March 21, 2017):

**A.6.1 Definition** (Balance).

$$\rho_8(G, i) = \sum_{j=1}^n x_{ij} \sum_{h=1, h \neq i, j}^n (b_0 - |x_{ih} - x_{jh}|)$$

where  $b_0$  is a constant included to reduce the correlation between this effects and the density effect.

**A.7 Univariate Nonparametric Tests**

Under the assumption of symmetrical distribution, the Wilcoxon Signed Rank Test, tests the null hypothesis:  $H_0 : \bar{\mu} = \mu^{act}$ . It does so by taking the differences,  $|\mu_i - \mu^{act}|$  and the sign,  $sgn(\mu_i - \mu^{act})$  for all  $i \in N$  observations in the sample. It orders the distances from small to large where  $R_i$  denotes the ranking. The test statistic is  $W = \sum_{i=1}^{N_r} [sgn(\mu_i - \mu^{act}) R_i]$  where  $N_r$  is the total number of observations,  $N$ , less variables with difference  $|\mu_i - \mu^{act}| = 0$ . Under the assumption of symmetric distribution,  $W = 0$  if the null hypothesis is true. Therefore, the Wilcoxon Signed Rank Test, tests the deviation of  $W$  from 0 where  $W \sim \left(0, \frac{N_r(N_r + 1)(2N_r + 1)}{6}\right)$  (Wilcoxon, 1945).

The Sign test does not assume symmetric distribution, and the null hypothesis itself,  $H_0 : \mu^{median} = \mu^{act}$  tests for symmetric distribution around  $\mu^{act}$ . Again, the differences and signs are taken  $|\mu_i - \mu^{act}|$  and two sums are computed  $W^+$  are the number of differences which are positive and  $W^-$  are the number of differences which are negative. Under the null hypothesis,  $W^+$  and  $W^-$  are binomial variables with probability 0.5.

**A.8 Multivariate Nonparametric Tests**

The following multivariate nonparametric test is detailed in (Oja and Randles, 2004) and was implemented in R using the SpatialNP package. Recall the Hotelling's  $T^2$  test for  $p$ -dimensional data from observations  $\mathbf{X}_1, \dots, \mathbf{X}_n$ . It is assumed that  $\mathbf{X}_i$  are i.i.d multivariate normally distributed. The null hypothesis  $H_0 : \boldsymbol{\theta} = \mathbf{0}$  is rejected if

$$T^2 = n\bar{\mathbf{X}}^T S^{-1} \bar{\mathbf{X}} \geq \frac{np}{n-p} F_{p, n-p},$$

where  $\bar{\mathbf{X}} = \sum_{i=1}^n \mathbf{X}_i$  and  $nS = \sum_{i=1}^n (\mathbf{X}_i - \bar{\mathbf{X}})^T (\mathbf{X}_i - \bar{\mathbf{X}})$  is  $n$  times the covariance matrix. Without loss of generality, the null hypothesis  $H_0 : \boldsymbol{\theta} = \boldsymbol{\mu}$  can be tested using the change of coordinates  $\mathbf{X}_i - \boldsymbol{\mu}$ .

A test statistic is said to be affine invariant if  $T(D\mathbf{X}_1, \dots, D\mathbf{X}_n) = T(\mathbf{X}_1, \dots, \mathbf{X}_n)$  for every  $p \times p$  non singular matrix  $D$  and  $p$ -variate data set  $\mathbf{X}_1, \dots, \mathbf{X}_n$ . A multivariate nonparametric test is constructed by considering the transformation  $\mathbf{Y}_i = S(A_x \mathbf{X}_i)$ , for  $i = 1, \dots, n$  and *spatial sign function*

$$S(\mathbf{X}) = \begin{cases} \|\mathbf{X}\|^{-1} \mathbf{X}, & \mathbf{X} \neq \mathbf{0} \\ \mathbf{0} & \mathbf{X} = \mathbf{0} \end{cases}.$$

This transformation is affine-invariant for  $A_x^T A_x = V_x^{-1}$ , where  $V_x^{-1}$  is a positive semi-definite matrix with  $\text{trace}(V_x) = p$  and  $\text{pave}\{\mathbf{Y}_i \mathbf{Y}_i^T\} = I_p$ , with corresponding test statistic

$$Q^2 = np \bar{\mathbf{Y}}^T \bar{\mathbf{Y}},$$

which, for large sample sizes, and a symmetric underlying distribution,  $Q^2 \sim \chi_p^2$ .

Table A.1: Parameter Estimates of Friendship Evolution

Network Effects	Model I				Model II				Model III			
	Estimate	S.E	p-value	Conv. t-ratio	Estimate	S.E	p-value	Conv. t-ratio	Estimate	S.E	p-value	Conv. t-ratio
0 Rate parameter	7.61	( 0.85 )			7.70	(0.86)			7.69	(0.88)		
1 . eval outdegree (density)	-1.40	( 0.24 )	3.66e-09	-0.04	-1.33	(0.17)	1.71e-15	0.02	-1.26	(0.18)	5.46e-12	-0.04
2 . eval reciprocity	1.38	( 0.22 )	7.12e-10	-0.01	1.32	(0.19)	1.18e-11	0.05	1.30	(0.21)	2.99e-10	-0.02
3 . eval transitive triplets	0.21	( 0.08 )	0.01	-0.02	0.18	(0.05)	5.37e-04	0.03	0.17	(0.05)	6.99e-04	-0.07
4 . eval 3-cycles	-0.03	( 0.13 )	0.79	0.02								
5 . eval transitive ties	-0.04	( 0.23 )	0.88	-0.01								
6 . eval balance	0.04	( 0.02 )	0.03	0.03	0.04	(0.02)	0.03	-0.01	0.05	(0.02)	0.01	0.06
7 . eval number of actors at distance 2	-0.27	( 0.07 )	1.10e-3	-0.03	-0.27	(0.07)	1.10e-04	-0.03	-0.29	(0.07)	4.68e-05	-0.03
10 . eval same country	0.58	( 0.24 )	0.01	-0.01	0.57	(0.23)	0.01	0.01	0.54	(0.24)	0.03	0.03
11. eval sex alter	0.20	( 0.16 )	0.18	0.07								
12. eval sex ego	0.04	( 0.16 )	0.81	0.06								
13. eval same sex	0.05	( 0.17 )	0.76	-0.03								



# Acknowledgements

One thanks for AIMS For tickling my brain Eyes open wide A leap in my stride.

One thanks for those Explorers before us Who laid down the laws And still left more.

One final thanks for Mom, a belated Happy Mother's Day.

# References

- B. V. Carolan. *Social Network Analysis and Education: Theory, Methods and Applications*. Sage Publications, 2014.
- W. de Nooy, A. Mrvar, and V. Batagelj. *Exploratory Social Network Analysis with Pajek*. Structural Analysis in the Social Sciences. Cambridge University Press, 2005.
- R. F. Engle. Wald, likelihood ratio, and lagrange multiplier tests in econometrics. *Handbook of econometrics*, 2:775–826, 1984.
- E. Estrada and P. A. Knight. *A First Course in Network Theory*. Oxford University Press, 2015.
- R. Eynon. The rise of big data: what does it mean for education, technology, and media research? *Learning, Media and Technology*, 38(3):237–240, 2013.
- P. W. Holland and S. Leinhardt. An exponential family of probability distributions for directed graphs. *Journal of the American Statistical Association*, 76(373):33–50, 1981.
- E. Lazega and T. A. B. Snijders. *Multilevel Network Analysis for the Social Sciences: Theory, Methods and Applications*. Methodos Series. Springer International Publishing, 2015.
- A. Lomi, T. A. B. Snijders, C. E. Steglich, and V. J. Torló. Why are some more peer than others? evidence from a longitudinal study of social networks and individual academic performance. *Social Science Research*, 40(6):1506–1520, 2011.
- G. S. Maddala. *Limited-Dependent and Qualitative Variables in Econometrics*. Econometric Society Monographs. Cambridge University Press, 1983.
- M. McPherson, L. Smith-Lovin, and J. M. Cook. Birds of a feather: Homophily in social networks. *Annual review of sociology*, 27(1):415–444, 2001.
- E. Muller. AIMS Network Data and Analysis using R. <https://github.com/emilymuller1991/AIMS-Essay>, 2017.
- M. E. J. Newman. *Networks, an Introduction*. Oxford University Press, 2010.
- H. Oja and R. H. Randles. Multivariate nonparametric tests. *Statistical Science*, pages 598–605, 2004.
- W. R. Penuel, W. Sussex, C. Korbak, and C. Hoadley. Investigating the potential of using social network analysis in educational evaluation. *American Journal of Evaluation*, 27(4):437–451, 2006.
- W. Press. *Numerical Recipes 3rd Edition: The Art of Scientific Computing*. Cambridge University Press, 2007.
- R. M. Ripley, T. A. B. Snijders, Z. Boda, A. Voros, and P. Preciado. Manual for rsiena version 4.0, version March 21, 2017. URL <http://www.stats.ox.ac.uk/siena/>.
- S. M. Robbins, Herbert. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3):400–407, 1951.
- R. Serfozo. *Introduction to Stochastic Networks*. Stochastic Modelling and Applied Probability. Springer New York, 2012.

- T. A. B. Snijders. The statistical evaluation of social network dynamics. *Sociological Methodology*, 31(1):361–395, 2001. ISSN 1467-9531.
- T. A. B. Snijders. Models for longitudinal network data. In *Models and Methods in Social Network Analysis*, pages 215–447. Cambridge University Press, 2005.
- T. A. B. Snijders. Stochastic actor-oriented models for network change. In *Evolution of Social Networks*, pages 185–200. Taylor & Francis, 2013.
- T. A. B. Snijders, G. G. van de Bunt, and C. E. G. Steglich. Introduction to stochastic actor-based models for network dynamics. *Social Networks*, 32(1):44–60, 1 2010.
- H. Taylor and S. Karlin. *An Introduction to Stochastic Modeling*. Academic Press, 1998.
- P. Van Mieghem. *Performance Analysis of Communications Networks and Systems*. Cambridge University Press, 2009.
- S. S. Wasserman. A stochastic model for directed graphs with transition rates determined by reciprocity. *Sociological methodology*, 11:392–412, 1980.
- S. S. Wasserman and K. Faust. *Social Network Analysis: Methods and Applications*. Cambridge University Press, 1994.
- F. Wilcoxon. Individual comparisons by ranking methods. *Biometrics bulletin*, 1(6):80–83, 1945.