Identity and Fragmentation in Networks*

Pritha Dev
ITAM
Working Paper

Abstract

This paper looks at the role of identity in the fragmentation of networks by incorporating the choice of commitment to identity characteristics, into a noncooperative network formation game. The Nash network will feature divisions based on identity, moreover, it will have layers of such divisions. Using the refinement of strictness, I get stars of highly committed players linked together by less committed players. Next, I propose an empirical methodology to deduce which dimensions of identity cause the fragmentation of a given network. I propose a practical algorithm for the estimation and apply this to data from villages in Ghana.[1]

Keywords: Identity, Network formation, Community Structure

JEL Codes: D85, C45, Z13

*I am thankful to Alberto Bisin, Prabal De, Kaushal Kishore, Nicola Persico, Debraj Ray, Guido Ruta, Julia Schwenkenberg and Joerg Stoye for helpful comments and suggestions. I would also like to thank seminar participants at NYU, ITAM and Banco de Mexico.

[1] I would like to thank Chris Udry and Markus Goldstein for providing free access to data collected by them from four villages in Eastern Region of Ghana.

1 Introduction

This paper addresses the question of how identity leads to the fragmentation of networks; as well as which dimensions of identity are important in such fragmentation. Identity in this paper is defined as the set of characteristics/attributes attached to each person.[2] It is a well documented fact that different aspects of identity have been important in dividing society at different points of time. For instance, in India, religion is an important line of division, race is an important line of division in the US, religious identity supersedes national identity in many populations, etc. These divisions have important economic, social, and, political consequences; and it is important to know how these divisions come up as well as to know exactly which division currently prevails. To that effect, this paper suggests a theoretical mechanism via which divisions along identity dimensions arise endogenously in networks. The next logical question is, which dimensions could be leading to fragmentation in actual networks. As an answer, I give an empirical strategy to estimate which dimensions of identity are important in fragmenting a network as well as which groups along these dimensions does each person ascribe to.

This paper makes a theoretical contribution to the literature on network formation by incorporating identity into a network formation model, and in allowing players to *choose* which aspects of their identity will be important in the network thus formed. In this theoretical model, players have identity along multiple dimensions but they can choose how much they wish to commit to different aspects of their identities. More concretely, I define identity as being defined along different dimensions, where each dimension is composed of a fixed set of discrete characteristics. Each individual's identity vector consists of one characteristic from each one of these dimensions. Commitment to a characteristic is a measured by the variable, $\theta \in [0, 1]$, where a higher θ indicates higher commitment to the characteristic. For instance, a person with a West Indian and black identity, might choose to identify more with the West Indian identity while not caring much for his black identity and that would be an instance where he chooses a high commitment (θ for region close to one) for region of origin and a low commitment for race (θ for race close to zero).[3]

To see exactly how identity and commitment have an impact on fragmentation, I think of the links and groups in society as an outcome of a network formation game. Making a connection is costly and requires some investment in terms of time and effort, but the greater is the network of your connections (people connected to you directly or indirectly), the greater are the benefits.

[2] For a detailed discussion on the uses of the word 'identity' look at Brubaker and Cooper (2000)

[3] example taken from Waters (1999).

For instance, the costs of making a professional contact would be the time spent socializing with them, sending them holiday greetings to stay in touch. The benefits from knowing a person could be informational, social, or psychological; and these benefits are increasing in the number of people you know directly or through other people. The benefit of being connected to more people can be seen, for instance, in job search models; where, the greater is your network the more likely it is that you hear of job openings. The introduction of identity makes an individuals decision two-fold:
- how much to commit to his characteristics,
- which links to form.

Identity enters the network formation game by changing the cost of making connections. The profit from a link would now depend on the identities and commitments of the two people, where the closer the identity and higher the commitments, the more profitable is the link formation and vice versa. I first consider the simpler case where there is only one dimension of identity. In this case the Nash network will be either empty with no links being formed and commitment choices indeterminate; or, sort players by their identity and all players choose high commitment levels; or finally, all players are connected and have lower commitment levels. To get a better sense of what the exact structure of these networks might be, I use the commonly used refinement of strictness, where I restrict attention to those Nash equilibria where each player strictly prefers his link strategy to any other link strategy. Using this refinement, the structure of the connected network consists of each identity type having a core star (the players who are a part of the star involve one central player having direct links to all other players) of highly committed individuals, and these stars are linked together by less committed players. As players are allowed to have identities along many dimensions, the Nash network has the interesting feature that it incorporates layers of divisions. It first divides the population based on some set of dimensions. Within the elements of the partition generated, there might be further subdivisions based on added identity dimensions and this chain of subdivisions might continue. This is very similar to real world situations where for instance, we might observe division along religion and within each religion there is further subdivision based on sects. Using the refinement of strictness, the important structure that emerges again has stars of highly committed players being linked together by the less committed players.

I will now outline the empirical methodology proposed in this paper. The theoretical model explains the existence of multiple equilibria that we observe in the real world. But to identify which of the possible equilibria a given group of players is coordinating at, we need empirical methods. The theoretical model uses some stark assumptions like no decay, each link has the same value and it does not allow for any error in the formation of links. Under these assumptions, the model led to partitioning of the network into completely segregated components. In a more realistic setting without these assumptions, we will see not clear divisions, but several communities within the network. A community can be thought of as group of highly interconnected individuals with few links across communities. In other words, the stark theoretical result of no links across communities is replaced by the empirical fact that intensity of links within a community are higher than across communities. From the insight gained by the theoretical model in this paper, we know that the partition into communities and probabilities of interactions depend on identity. With that in mind, I allow the probabilities of interactions to depend on identities and at the same time restrict possible community structures to those based on identity. For each fixed partition and probabilities, I find the likelihood of observing the data. I then select the partition and probabilities, which maximize the likelihood of the observed data.

The estimation procedure has good large sample properties. It is consistent and so as the sample size grows, the estimated partition and probabilities converge to the true data generating process. If we search over some initial set of dimensions and add another dimension, which belongs to true set of dimensions, the likelihood will strictly increase. These results suggest a simple search algorithm, where starting with one dimension of identity at a time, we select the dimension that gives the maximum likelihood. Then with this as one of the dimensions, we add other dimensions and then pick the pair of dimensions that maximize the likelihood. We keep adding dimensions until the likelihood stops increasing. Moreover, when searching over some k given dimensions, we first begin with the finest partition and find the likelihood under that. Then we progressively make it coarser till the likelihood stops increasing.

The data I use was collected by Chris Udry and Markus Goldstein [4] over the course of two years and fifteen modules in four village clusters in Eastern Region of Ghana. In each village 60 couples/triples were questioned. Each respondent was questioned about his links to seven randomly selected (without replacement) matches from the entire sample and three focal village residents. In general, I found that religion and clan were important determinants of the communities. More importantly, we do see evidence of multiple equilibria in this data with the four villages dividing along different dimensions of identity.

Related Literature

The study of social identity has been long incorporated into the social sciences through the pioneering works of Mead (1934), Stryker (1968), Tajfel and Turner (1979) and Stryker and Burke

[4] See Goldstein and Udry (1999) for a detailed discussion of the data. Also see Conley and Udry (2004), and, Conley and Udry (2005).

(2000). In economics, there is a large body of work trying to evaluate the impact of membership to identity based groups on economic outcomes. An important theoretical contribution incorporating identity, is the Akerlof and Kranton (2000) model, which allows the self image (derived from identity) to affect the utility function, where they take as given which dimension of identity is salient. Fryer and Jackson (2002) incorporate identity into a person's decision making problem differently, by looking at how people form expectations/judgments of other people based on the other person's identity. Sergio Currarini and Pin (2008) explores the formation of friendships with people of different types and benefits of friendship being type dependent. Sen (2006) is an excellent exposition on the relationships between identity and violence. The work of Bisin and Verdier (2000), looks at the evolutions of a child's identity choice. They look at the impact of the parent's socialization choice on the identity of the child and thus, the child's identity is determined endogenously. Esteban and Ray (1994) look at the polarization of society by attribute and they use the concept of 'identification', which is how much a person relates to similar people. The paper by Dev (2009) is very related to the present paper as it links choice of identity with network formation model. However, the focus of that paper is towards the emergence of identity groups, whereas this paper focuses on a theoretical and empirical investigation of partitions given existing identity characteristics. Among work not directly incorporating identity into a decision makers problem, social interaction models (beginning with Schelling (1971)) evaluate the impact of membership to groups on socio-economic outcomes. Since this paper focuses on how networks partition with the introduction of identity, it is also linked to the vast literature on club formation. Though most of this literature is not concerned with the how the networks evolve within a club, the paper by **?** bridges that gap.

The literature in economics on network formation follows two main strands - one follows Jackson and Wolinsky (1996) and the other follows Bala and Goyal (2000a) and Bala and Goyal (2000b). [5][6] This paper falls into the second strand. Bala and Goyal (2000a) propose that the network formation game is a result of a non-cooperative one-shot simultaneous move game between players. They further assume homogenous players and that the cost of a link is bourne by the initiator. Galeotti, Goyal, and Kamphorst (2005), Hojman and Szeidl (2008), Sarangi, Billand, and Bravard (2006), Galeotti (2006) and Gilles and Johnson (2000) relax the homogeneity assumption, but

[5] The book by Jackson (2006) as well as Dutta and Jackson (2003) provide an excellent review of the literature.

[6] The recent paper by Page Jr. and Wooders (2009) unifies the two strands by suggesting a common framework with which to view all network games.

unlike this paper they fix the level of heterogeneity. [7] [8]

The work to extract the salient identity characteristics is linked in spirit (though not methodology) to the vast literature on constructing algorithms for partitioning network. (see Newman (2004) for an overview). It is also linked to work on finding clusters in sociology, statistics, computer science and physics. Seminal work on block models and positional models by Lorrain and White (1971) and White, Boorman, and Breiger (1976) established the concept of structural equivalence which says that two nodes are structurally equivalent if they have the same relationships with all other nodes. Nodes belonging to the same class/block would be structurally equivalent. More recent and closely related work is by Nowicki and Snijders (2001), who, assume that unobserved latent classes affect the probabilities of link formation. They do not account for homophily or that people with similar characteristics are more likely to form links with each other. Tallberg (2005) extended this model to represent homophily on attributes by allowing latent class membership to depend on attributes. But this does not allow for any possible heterogeneity within latent classes due to individual characteristics. In particular, this paper is most closely related to Copic, Jackson, and Kirman (2009), who propose a maximum likelihood approach to rank community structures. They assume that network generating process is some underlying community structure which partitions the nodes, probability of forming links within a community and a probability of forming links across communities. Assuming the probability of forming the link is strictly greater if the two nodes belong to the same community, they find the community structure and probabilities which maximize the likelihood of observing the network.

The rest of the paper is organised as follows. Section 2 explains the theoretical model. Subsection 2.1 look at the case where identity characteristics are along a single dimension. Subsection 2.2 look at the case where identity characteristics are along a multiple dimensions. In both these cases, we characterize the Nash and Strict Nash networks. Section 3 presents the empirical counterpart of including identity in a network. Subsection 3.1 presents the methodology being used in detail.

[7] Other important theoretical extensions of network formation models include Jackson and Dutta (2000), Watts (2001), Deroan (2003), Feri (2004), Kranton and Minehart (2001), Goyal and Joshi (2003), Goyal and Vega-Redondo (2005), Slikker and van den Nouweland (2001), Gilles and Johnson (2000), McBride (2006), Bramoulle and Kranton (2007), etc.

[8] Empirical investigation into the formation of networks includes work by Conley and Udry (2004), **?** and Santos and Barrett (2004). Empirical work using network ties to explain risk sharing starting with Townsend (1994) includes De Weerdt and Dercon (2006), De Weerdt (2004), Fafchamps and Gubert (2007), Fafchamps and Lund (2003) and Grimard (1997). Empirical studies using networks to explain technology adoption include Foster and Rosenzweig (1995), **?**, Conley and Udry (2005) and Bandiera and Rasul (2002). Granovetter (2005) provides an excellent overview of the relationship between social structures and outcomes. works using networks to predict various economic outcomes include Fafchamps (2002), **?**, Fafchamps, van der Leij, and Goyal (2006), Patacchini and Zenou (2008) and Munshi and Rosenzweig (2006).

Subsection 3.2 presents the data and the results. Section 4 presents the conclusion. All proofs, graphs, and, results are collected in the Appendices.

2 Identity and Network Formation

The players are denoted by the set $\mathbf{N} = \{1, 2, ..., n\}$. Identity is defined along m dimensions where the $d-th$ dimension is denoted by D_d. A dimension D has κ_d characteristics $D = \{c_1, ..., c_{\kappa_d}\}$ and each person has exactly one of these characteristics. The set $\mathbf{DIM} = \{D_1, ..., D_m\}$ collects all the identity dimensions. Each person $j's$ identity is an m-dimensional vector $I_j = \{i_{j1}, ..., i_{jm}\}$, where $i_{jd} \in D_d$ for each d. The identity profile of the population is contained in the $n \times m$ matrix $\mathbf{ID}$.

I define a 'block' as a group made up of completely homogenous players who have all the same characteristics, given $\mathbf{DIM}$. More formally,

Definition 1 *Given identity dimension* $\mathbf{DIM} = \{D_1, ..., D_m\}$, *a block* $B \subseteq N$ *is a collection of individuals such that if* $l, k \in N$, *then* $I_{ld} = I_{kd}$ *for all* $d \in \{1, .., m\}$. *Block*$(\mathbf{DIM})$ *is the set of all blocks given* $\mathbf{DIM}$.

More generally, given any set of dimensions, $\mathbf{DIM'} \subseteq \mathbf{DIM}$, let *Block*$(\mathbf{DIM'})$ denote the set of all blocks given $\mathbf{DIM'}$. For $\mathbf{DIM'} = \phi$, let *Block*$(\phi) = \{\mathbf{N}\}$.

Consider an example with three dimensions of identity, $\mathbf{DIM} = \{D_1, D_2, D_3\} = \{$Colour, Height, Gender$\}$. Further, within the dimension of Colour we have two characteristics of {Red, Blue}, within the dimension of Height we have two characteristics of {Tall, Short}, and finally, within the dimension of Gender we have two characteristics of {Male, Female}. So we now have,

$$\begin{aligned}\mathbf{DIM} &= \{D_1, D_2, D_3\} \\ &= \{\text{Colour, Height, Gender}\} \\ &= \{\{\text{Red, Blue}\}, \{\text{Tall, Short}\}, \{\text{Male, Female}\}\}\end{aligned}$$

A person $j's$ identity vector, I_j, will consist of one characteristic each of Colour, Height and Gender. For instance, $I_j = \{i_{j1}, i_{j2}, i_{j3}\} = \{$Red, Tall, Male$\}$. A block in this scenario will consist of players who have the same Colour, same Height and same Gender. *Block*$(\mathbf{DIM})$ will be the set of all possible blocks, in this case twelve. Figure 7, illustrates all possible blocks under these three dimensions. Suppose $\mathbf{DIM'}$= $\{D_2, D_3\}$ = {Height, Gender}. The we get

$Block(\mathbf{DIM}') = \{\{\text{Tall,Male}\},\{\text{Tall,Female}\},\{\text{Short,Male}\},\{\text{Short,Female}\}\}$, where {Tall, Male} for instance includes all players of all two colours who are Tall and Male.

I next define the 'similarity index' for two individuals, which is an m-dimensional vector taking the value of 1 in those dimensions in which the two individuals have the same characteristic.

Definition 2 *The **similarity index** is denoted by $S_{lk} \in \{0,1\}^m$, where the d^{th} element $S_{lk}^d = 1$ if $i_{ld} = i_{kd}$ and zero otherwise. Let $S_l = \{S_{l1}, .., S_{ln}\}$ collect all the similarity indices for the payer l.*

For example if $I_l = \{\text{Red, Tall, Male}\}$ and $I_k = \{\text{Blue, Tall, Male}\}$, then l and k differ only in the dimension of Colour and they have the same characteristic under Height and Gender. So for l and k we get $S_{lk} = \{0,1,1\}$.

Each person in the game has two choices:

Commitment: Each person j chooses his commitment along each dimension-d, which is denoted by $\theta_{jd} \in [0,1]$. A higher commitment to any characteristic will make linking with people with the same characteristic more profitable but make links less profitable with people who don't share this characteristic. The commitment choice for an individual j is given by $\theta_j = \{\theta_{j1}, \theta_{j2}, .., \theta_{jm}\}$. Let the $n \times m$ matrix $\mathbf{\Theta}$ denote the commitment profile of the population. Given the identity profile of the population as well as the commitment choices, I define $B_\theta = \{k \in B | \theta_k = \theta\}$ as the subset of a block $B \in Block(\mathbf{DIM})$, all members of which choose commitment of θ. We use this definition to club together all individuals who are homogenous in not only their characteristics but also their commitments. In particular, $B_1 = \{k \in B | \theta_k = \mathbf{1}\}$ and $B_0 = \{k \in B | \theta_k = \mathbf{0}\}$.

Links: Each person also chooses his links, $g_j \in \{0,1\}^{n-1}$ where each element $g_{jk} \in \{0,1\}$ of $g_j = \{g_{j1}, .., g_{j(j-1)}, g_{j(j+1)}, .., g_{jn}\}$ denotes his decision to form a link ($g_{jk} = 1$) or not ($g_{jk} = 0$) with agent k. The links are undirected and $g_{kl} = 1$ will allow k to access l's information and vice versa, even though the cost of the connection is borne by k.

These two choices together define the strategy of an individual i as $s_i = \{\theta_i, g_i\} \in [0,1]^m \times \{0,1\}^{n-1}$. The two decisions are taken simultaneously to capture the fact that the choice of commitment determines which connections to make and at the same time what connections we make determine how strongly we want to be committed to any characteristic.

The strategy for links generates a network denoted by g, where $g = \{g_1, ..., g_n\}$. Define $\overline{g} = cl(g)$ where an element of $\overline{g}$ is $\overline{g}_{kl} = \max\{g_{kl}, g_{lk}\}$ for all $l, k \in N$. We say a *path* exists between agents k and l if either $\overline{g}_{kl} = 1$ or there exist $j_1, ..., j_m$ such that $\overline{g}_{kj_1} = ... = \overline{g}_{lj_m} = 1$. A *path* is denoted by $k \overset{\overline{g}}{\longleftrightarrow} l$. A component within a graph g is $C(g) \subseteq N$ such that all agents within the component have a path connecting each other and there are no link going from any player in $C(g)$ to any player not in $C(g)$. A component is said to be *minimal* if deleting any link will lead to it not being a component anymore and a network is called *minimal* if all its components are minimal. A network is said to be *connected* if it has only one component made up of all players. A network is said to be *empty* if no player makes any links.

Neighbourhood for agent k are the agents with whom k forms links and is defined by function $N^d(k;g) = \{l \in N | g_{kl} = 1\}$. The set of all agents to whom k is linked, directly or indirectly, is given by $N(k;g) = \{l \in N | k \overset{g}{\longleftrightarrow} l\}$. The payoff function for $k-th$ individual is given by $\Pi_k : \mathbf{\Theta} \times G \to R$,

$$\Pi_k(\mathbf{\Theta}, g) = \pi(N(k;g), N^d(k;g), S_k, \Theta)$$

We use the following assumptions:

- **A1**: $\pi()$ is strictly increasing in the size of $N(k;g)$.
- **A2**: $\pi()$ is strictly decreasing in the size of $N^d(k;g)$. Further, for each $l \in N^d(k;g)$, the marginal impact is smaller for larger commitments by k/l for identity dimensions they have the same characteristic in; and, it is larger for larger commitments by k/l for identity dimensions they *don't* have the same characteristic in.
- **A3**: Adding a link which accesses players k is not connected to, where this link is profitable if it were the only link made by k, will increase k's profits. Moreover, only such profitable links increase profitability.

The first two assumptions imply that payoffs are increasing in the number of people one is linked to but is decreasing in the number of links formed. The second assumptions also implies that the cost of forming a link depends both on the identity of the players as well as their commitments. For two players k and l the more their identities agree, the higher is the profit. Moreover, if agent k and l along dimension-d have the same characteristic or $(I_{kd} = I_{ld})$, then $\pi()$ is increasing the more committed they are or higher is θ_{kd} and θ_{ld}; but if $I_{kd} \neq I_{ld}$ then $\pi()$ is decreasing in θ_{kd} and θ_{ld}. The last assumption implies that given a player's θ, if any link strategy is profitable on its own, it must be profitable added to the current link strategy of the player. Moreover, each individual link in a player's strategy must be profitable.

Nash, Strict' Nash and Efficient Structures: A strategy profile s is said to be Nash equilibrium if

$$\Pi_k(\mathbf{\Theta}, g) \geq \Pi_k((\theta'_k, \mathbf{\Theta}_{-k}), (g'_k, g_{-k}))$$

for all $k \in N$, where $\mathbf{\Theta}_{-k}$ is the commitment profile of all but the $k-th$ player, g_{-k} is the link strategy of all but the $k-th$ player and $s'_k = \{\theta'_k, g'_k\}$ is any other strategy of player k.

A refinement often used in the literature is that of Strict Nash, where each player strictly prefers his link strategy. In this paper where we look for Strict' Nash networks, (g, Θ) formed by strategy profile s that satisfies strictness on the link strategy but not necessarily on the commitment strategy:

$$\Pi_k(\mathbf{\Theta}, g) > \Pi_k(\mathbf{\Theta}, (g'_k, g_{-k}))\text{for all } g'_k \neq g_k$$

In other words we are looking at strategies such that for each individual, the link strategy given $\mathbf{\Theta}$, is strictly better than any other link strategy. But they could have an indifferent strategy which involves a change in commitment only. I choose this form of strictness because any individual not forming links is always going to be indifferent about his commitment - and there will be always be at least one such individual in equilibrium.

An efficient outcome maximizes welfare over all possible commitment profiles and networks; where welfare is measured by $W : \mathbf{\Theta} \times G \to R$,

$$W(\mathbf{\Theta}, g) = \sum_{k \in N} \Pi_k(\mathbf{\Theta}, g)$$

2.1 One Dimensional Identity

I begin by assuming that people have just one dimension of identity. For instance only colour sorts people into groups. Within this one dimension of identity there are κ characteristics, for example, within color, people could be either Blue, Green or Red. With a single dimension of identity, a block collects all individuals with the same characteristic. In the previous example, there would be three blocks; a block of Blue people, a block of Green people and a block of Red people. In particular, B_i denotes the block which collects all players with the $i - th$ characteristic and we let n_i denote the number of people in B_i.Though the identity characteristic is given exogenously in the model, each person chooses his commitment. The commitment choice of player k is given by the variable $\theta_k \in [0, 1]$, where a $\theta_k = 1$ denotes the strongest level of commitment to his characteristic and $\theta_k = 0$ denotes least commitment to his own type.

The following result shows that the equilibrium Nash structure will be empty, connected or will have sorting by characteristic. Moreover, in a divided society, commitment levels will be very high, whereas in a connected society commitment levels will be lower.

Proposition 1 ***Nash Networks:*** *Under A1, A2 and A3, Nash Networks will be one of the following:*

1) ***Empty****, with no connection being formed and choice of commitment is indeterminate.*

2) ***Separated****, where for each characteristic i; either, B_i will form a minimally connected component and commitments are 1, or, everyone in B_i is a singleton.*

3) ***Minimally Connected****, with at least one player choosing commitment less than 1.*

The propositions follows directly once we establish that all players with the same characteristic must either belong to the same component or be singletons and, moreover, if any two different blocks are linked, then it must be that everyone is linked. Suppose in a Nash network g; $k, k' \in B_i$ and of them k belongs to some component $C(g)$. If k belongs to $C(g)$ then either k himself makes a profitable link or some other player player profitably links to k. By choosing an commitment $\theta_{k'} = 1$, k' can form a link with k which is cheaper or as costly as the the link k makes/receives and he gets linked to everyone in $C(g)$. If $k' \in C'(g)$, and if either k or k' were not making any links, the one not making the link would have the incentive to add a profitable link to the other. In case both are making links, then the one making the less profitable links would prefer to switch to adding a more profitable link to the other. In other words, all players in a block will either belong to the same component or none of them makes any links. Next, if it is profitable for a player $k \in B_i$ to form a link with $l \in B_j$, then the same link is also profitable for any player from some third block B_a. Suppose B_a formed a separate component. By minimality B_a must have at least one player who makes no links within B_a and this player could choose $\theta = 0$ and form a link with k. And so a Nash network must either be connected or have no links between blocks.

The three types of Nash networks are depicted in Figures 1, 2 and 3. The three figures assume that the only dimension of identity is colour and within this dimension there are three possible characteristics: Blue, Green and Red. Figure 1 shows the empty network with no links at all. Figure 2 shows sorting by colour, where there are no links between different blocks. It shows that the Green block is internally connected and forms a component and so does the Red block; the Blue block on the other hand forms no links and everyone with the characteristic Blue is a singleton. Figure 3 shows a minimally connected network where everyone is connected to everyone else.

An important thing to note about the Nash Networks is that they allow for multiple equilibria. Given the cost structure, both a minimally connected network and a network separated by identity might be possible Nash Equilibria.

Corollary 1 *If $\pi(\{k\}, \{k\}, S_{lk} = 0, \theta_k = 1, \theta_l) = -\infty$ and $\pi(\{k\}, \{k\}, S_{lk} = 0, \theta_k = \theta_l = 0) > 0$, then this society will either be minimally connected with lower average commitments or fragmented with a high average commitment.*

We can think of this as the case where people choosing $\theta = 1$ have militant identities and any player with a different identity can not have links with him. As long as $\pi(\{k\}, \{k\}, S_{lk} = 0, \theta_k = \theta_l = 0)$ is low enough, we get multiple equilibria - one where at least some people from all identity blocks choose lower commitment and the network is connected, or, another where everyone chooses high commitment and the network is fragmented by identity.

The Nash networks will be numerous given any cost structure. One refinement often used in this literature is strictness, which allows us to focus on the more stable Nash equilibria where each player has only one best response strategy. In the network games without identity, the important Strict Nash structures is the center-sponsored star or the empty network. A star network has one central player and all players are linked directly only to him. If the the star is center-sponsored, then the central player forms a link with each player and no other links are formed. Given heterogenous players, another important structure is the interconnected star, which in this framework would involve each characteristic block connected within itself as in a star and then the different stars (blocks) form links to each other. Interestingly, in this model with identity, structures very similar to interconnected stars will emerge as one of the important Strict' Nash structures. In previous work by Galeotti, Goyal and Kamphorst,(2003), a structure that emerged was the generalised center-sponsored star. Within the framework of this paper, the generalised center-sponsored star will consist of one central block making all the links. Within this central block there will be a central player forming all the links within the block. Anyone from the central block who makes outside links only will choose $\theta = 0$ but those central types who form no links will be indifferent to the choice of θ.

To simplify the discussion of the Strict' Nash, I restrict attention to discrete commitment choices, in particular where $\theta \in \{0, 1\}$. This limited choice can be viewed as the decision of whether or not to give any weight to the characteristic.

One structure that emerges is a generalised version of the interconnected star here called the 'interconnected electron star'. In this structure, members of any characteristic block choosing $\theta = 1$ will be part of a center-sponsored star and this star will sponsor links to members with the same characteristic choosing $\theta = 0$. The $\theta = 0$ players are like the electron to the main star and they make/receive links with the other blocks.

Definition 3 *A structure is called **interconnected electron stars** if:*

(1) for each characteristic i, B_i is internally connected with players choosing $\theta = 1$ being the part of a center-sponsored star. This star forms links to any other members choosing $\theta = 0$.

(2) all the blocks are interlinked with all these external link being sponsored by the same block and only players choosing $\theta = 0$ receiving links from other blocks.

Figure 5 illustrates an interconnected electron star. Each block has a core made up of the players choosing $\theta = 1$ and forming a center-sponsored star. These center-sponsored stars then form the link to player from their own block who chooses $\theta = 0$. The Red player choosing $\theta = 0$ makes the links to the Blue player and Green player choosing $\theta = 0$. The generalised center-sponsored star is illustrated in Figure 4, where the Red block forms a center-sponsored star consisting of Red players choosing $\theta = 1$. The center-sponsored star then sponsors links to other Red players choosing $\theta = 0$, who then sponsor links to all the Blue and Green players.

Given the blocks $B_1, ..., B_\kappa$, for notational ease, let us assume that $\pi(B_1, \{k \in B_1\}, S_{lk} = 0, \theta_k = \theta_l = 0) \leq ... \leq \pi(B_\kappa, \{k \in B_\kappa\}, S_{lk} = 0, \theta_k = \theta_l = 0)$. In other words, if a player could form an external link to any of the blocks, his profits would be higher by linking to B_m than to to B_{m-1}.

Proposition 2 *The Strict' Nash under the assumption A1-A2 and $\theta \in \{0, 1\}$ will be the following:*

- ***Empty Network*** *if,*

$$\pi(\{i\}, \{i\}, S_{ij} = 1, \theta_i = \theta_j = 1) < 0; \forall i, j \textit{ such that } I_i = I_j$$

- ***Unconnected Center-Sponsored Stars*** *for each block in the set $\{B_{x1}, ..., B_{xy}\}$, if,*

$$\begin{aligned} \pi(\{i\}, \{i\}, S_{ij} = 1, \theta_i = \theta_j = 1) &\geq 0; \forall i, j \in \{B_{x1}, ..., B_{xy}\} \textit{ such that } I_i = I_j \\ \pi(B_\kappa, \{k \in B_\kappa\}, S_{lk} = 0, \theta_k = \theta_l = 0) &< 0; \forall l, k, \kappa \textit{ such that } k \in B_\kappa, l \notin B_\kappa \end{aligned}$$

- ***Interconnected Electron Stars*** *or Unconnected Center-Sponsored Stars if,*

$$\begin{aligned} \pi(\{i\}, \{j\}, S_{ij} = 1, \theta_i = \theta_j = 1) &\geq 0; \forall i, j \textit{ such that } I_i = I_j \\ \pi(\{k\}, \{k\}, S_{lk} = 0, \theta_k = \theta_l = 0) &< 0; \forall k, l \textit{ such that } I_k \neq I_l \\ \pi(B_2, \{k \in B_2\}, S_{lk} = 0, \theta_k = \theta_l = 0) &\geq 0; B_2 \textit{ such that } k \in B_2; l \notin B_2 \end{aligned}$$

- ***Generalised Center-Sponsored Star*** *or Interconnected Electron Star or Unconnected Center-Sponsored Stars*

$$\begin{aligned} \pi(\{i\}, \{i\}, S_{ij} = 1, \theta_i = \theta_j = 1) &\geq 0; \forall i, j \textit{ such that } I_i = I_j \\ \pi(\{k\}, \{k\}, S_{lk} = 0, \theta_k = \theta_l = 0) &\geq 0; \forall k, l \textit{ such that } I_k \neq I_l \end{aligned}$$

The proposition says that the Strict' Nash network will be empty for very high cost ranges, unconnected center-sponsored stars if it is cheap to link within the same block but links are costly between different blocks, interconnected electron stars if links between different blocks are feasible and finally generalised center-sponsored stars are possible under very low costs. In fact, the generalised center-sponsored star can only emerge if all costs are very low, or in other words if identity does not have a significant impact on the profits. The proof focuses on the case when identity does have an impact on profits or when $\pi(\{k\}, \{k\}, S_{lk} = 0, \theta_k = \theta_l = 0) > 1$. The intuition for the proof is in three facts explained below. The first is that all those choosing $\theta = 1$ within a block, if internally linked, will form a center-sponsored star. This is so because these players are homogenous not only in their identity but also in their commitment. Consider three players, k, l and m, from the same block and all of them choose $\theta = 1$. If k forms a link with l, then in a Strict' Nash network, they will receive no links since any player choosing to link to them will be indifferent amongst linking to either. Also under strictness l can not form links with anyone else from the same block choosing the same identity. But since k and l are linked, we must have that m belongs to same component. And that is only possible if k forms a link with m, in other words, or, if they are arranged as in a star. If this star exists for B_i, we call it $B_{i1} - star$ and this star will then sponsor links to all other members of B_i choosing $\theta = 0$.

The next two facts work under the assumption that links across blocks are not possible to access a single player, or $\pi(\{k\}, \{k\}, S_{lk} = 0, \theta_k = \theta_l = 0) < 0$. The important thing to note with $\pi(\{k\}, \{k\}, S_{lk} = 0, \theta_k = \theta_l = 0) < 0$ is that no external link will be made to a single player or in other words, if an external link is made to a player, it must be because he is linked to others. Linking to a player with a different characteristic and an commitment of 1, is the costliest link for any player. If $g_{kl} = 1$ for k and l belonging to different characteristic blocks, and $\theta_l = 1$, then because $\pi(\{k\}, \{k\}, S_{lk} = 0, \theta_k = \theta_l = 0) < 0$, l must have other links, say with l', but for k linking to l' will either be cheaper or as costly as linking to l. This immediately leads to the conclusion that in a Strict' Nash network, any player receiving an external link must choose $\theta = 0$. Which also means that B_{i1} does not receive any external links and so it must be internally connected. Moreover, the only way for B_i to be externally linked is if B_{i0} and B_{i1} both receive external links and we already know that B_{i1} cant receive external links and so B_{i1} will be internally linked to B_{i0}. This now gives us for each block receiving links, the structure of a core star formed by the players choosing $\theta = 1$ and this star sponsoring links to the player choosing $\theta = 0$.

We now move to the question of efficiency, where in general, efficient networks will not be easy to pinpoint in this setting. Efficient networks will allow for blocks being partially connected, as well as for some blocks being connected but others not. But given that direct and indirect links are of equal value, all efficient networks will be minimal. Also given that internal links are cheaper than external links, any connected efficient network must have the minimum number of external links. Note that the star network need not be the most efficient way to connect all players within a block.

Let us look for how our Strict' Nash networks compare to the efficient networks. For the rest of this section, let us assume that the star network is in fact the efficient structure within each block. We see that Strict' Nash networks will approximate the efficient network whenever it is an interconnected electron star. And even when the Strict' Nash is unconnected center-sponsored stars; for some range of profit functions, it is efficient. But if forming links with other blocks is very cheap, then we know that Strict' Nash networks might be generalised center-sponsored stars which are not efficient.

In general an important reason why Strict' Nash networks are inefficient for many profit functions, is because in undirected networks, only one person bears the cost of the link while both benefit from the link. In a star network, moreover, only one person bears the cost for all the links while everyone benefits equally. We require that linking to a single player be profitable for the Strict' Nash network to be nonempty, whereas, the efficient network will be nonempty as long as linking to everyone through one link is better than having no links and observing no one.

Even though the efficient and Nash networks are connected, some of the resulting interconnected electron star structures might be inefficient because it allows for the possibility of players choosing $\theta = 1$ to form external links, it allows all internal links to be formed by a player choosing $\theta = 0$ and moreover if a single block sponsors all the external links, it allows that block to have more than one player choose $\theta = 0$. There are a few additional assumptions which could get rid of the first two causes of inefficiency. One is if we required players to make only one kind of link, either external or internal. Another is assuming links between players with different characteristics are prohibitively expensive if one player chooses $\theta = 1$. We could think of players choosing $\theta = 1$ as those with militant identities who want to have links only within their characteristic and moreover to repel any links initiated by any player with a different characteristic. Under both assumptions we would see the interconnected electron star with all external links formed by players choosing $\theta = 0$, each block would have a $B_{1i} - star$ and one block would sponsor all the external links. The only remaining inefficiency would occur because the block sponsoring the external links might have more than one player choosing $\theta = 0$. In that block there might be some players choosing $\theta = 0$ and making no links, this is purely because of the inherent inefficiency of undirected links, where, if a player bears no costs, he is indifferent in his commitment choice. There might also be more than one player choosing $\theta = 0$ and making external links. But its important to keep in mind, that for each such inefficient interconnected electron star, there is one that is efficient.

Unconnected center-sponsored stars might emerge as the only Strict' Nash networks where the efficient network is minimally connected. The inefficiency is again due to the fact that only one player bears the cost of the link. The cost of linking the two blocks is borne by one person, though the benefits are shared by all the members of those two blocks. When the Strict' Nash network allows for the possibility of interconnected electron stars, the unconnected center-sponsored stars might occur if in at least $m-1$ blocks every player chooses to be strongly committed at $\theta = 1$. To convert this inefficient unconnected center-sponsored star to the efficient network would require 1 player in each block to switch to choosing $\theta = 0$ and one of those players to add links to the other $\theta = 0$ players.

2.2 Multi-Dimensional Identity

We now allow players to have identities along more than one dimension. The possible number of dimensions of identity is large and, in fact, if we allowed enough dimensions of identity, we could map each person to a unique set of characteristics. Most of these characteristics would not play a role in determining costs of connections, for instance the difference in the size/shape of the nose would have no bearing on the cost of connections even though it would be characteristic in the identity vector. We think of **DIM** as collecting only those dimensions of identity, which due to some historical/sociological reasons, actually have a bearing on the costs of connection.

The next definition defines a concept very crucial to the structure of the Nash equilibrium. Consider a set of dimensions $\mathbf{DIM'} \subseteq \mathbf{DIM}$ and the blocks generated by using only $\mathbf{DIM'}$ collected in the set $Block(\mathbf{DIM'})$.

Definition 4 $\mathbf{DIM'}$ *are said to be* ***Separating Dimensions*** *in a network g; if, g is such that there are no links across the different $B_i \in Block(\mathbf{DIM}^{'})$. If no such dimensions exist, and, the network is neither connected nor empty, we say the separating dimensions are ϕ.*

Definition 5 $\mathbf{DIM}^1$ *are said to be* ***Minimal Separating Dimensions****, if they are separating dimensions and there do not exist any $\mathbf{DIM'} \subset \mathbf{DIM}^1$ such that $\mathbf{DIM'}$ are also separating dimensions.*

Suppose, continuing a previous example with $\mathbf{DIM} = \{D_1, D_2, D_3\} = \{\text{Colour, Height, Gender}\}$ $= \{\{\text{Red,Blue}\},\{\text{Tall, Short}\}, \{\text{Male, Female}\}\}$, that there is a network that has two components, the first collecting all Male's and the second collecting all Female's. In this case clearly, the network is divided along the dimension Gender. Consider another variation where there are five components, the network first divided along Gender and then further Males are divided by Colour and Females by Height. In this case now each component is a strict subset of $Block(\mathbf{DIM'}) = \{\{\text{Male}\}, \{\text{Female}\}\}$ and the minimal set of dimensions is still {Gender}.

Proposition 3 *Under A1 - A3, a Nash network g, will feature layers of separation -*

1) At the level of the entire population, the network is either connected, empty or there exists a unique minimal set of separating dimensions $\mathbf{DIM}^1 \subseteq \mathbf{DIM}$.

2) At the next level of separation, we consider the subpopulation within each $B_i \in Block(\mathbf{DIM}^1)$, for which there will be some unique minimal separating dimensions $\mathbf{DIM}^{1,B_i} \subseteq \mathbf{DIM}/\mathbf{DIM^1}$.

3) This recursive process will continue till we reach a level where all subpopulations are either connected, empty or the separating dimensions are ϕ.

To establish the result for the Nash network, we use a series of lemma's presented in the appendix. The first lemma shows the uniqueness of minimal separating dimensions. The next lemma states that the Nash network must be minimal. The next proves that all members of a block B, where $B \in Block(\mathbf{DIM})$, will either have no links at all or they will all belong to the same component. This is similar in spirit to the one-dimensional case. A component in a Nash network will then consist of some of these blocks forming links with each other. While this is true, there might be a minimal set of dimensions, $\mathbf{DIM}^1$, which is a subset of $\mathbf{DIM}$, such that under g there are no links across $B_i \in Block(\mathbf{DIM}^1)$. Existence of such a set of dimensions is guaranteed, because any network g will always define a partition over $Block(\mathbf{DIM})$ or ϕ.

Figure 8 shows some possible Nash network components under the previous example. It uses the same dimensions as the previous example, $\mathbf{DIM} = \{D_1, D_2, D_3\}$ = {Colour, Height, Gender} = {{Red, Blue}, {Tall, Short}, {Male, Female}}. In the first figure, the network forms three components, {Red, Tall}, {Red, Short} and {Blue}. This network is first divided by Colour and then further within Red, it divides by dimension Height. The next network in Figure 8 defines a partition over the blocks generated by $\mathbf{DIM} = \{D_1, D_2\}$ = {Colour, Height}. The particular partition combines the blocks of {Red, Tall}, {Blue, Tall} and {Blue, Short} into one component and the other component consists of the block {Red, Short}. The last network in Figure 8, first the network defines a partition over $Block(\{Colour\})$ and then within Red, the network subdivides based on Height and within Blue, the network subdivides based on Gender.

Notice that with multiple dimensions, the possibility of multiple equilibria is even larger.

Corollary 2 *Let us add the assumption that for any dimension $d \in \mathbf{DIM}$, if $S^d_{lk} = 0$ and either $\theta^d_l = 1$ or $\theta^d_k = 1 \Rightarrow \pi(\{k\}, \{k\}S_{lk}, \theta_l, \theta_k) = -\infty$. Let us also assume $\pi(\{k\}, \{k\}, S_{lk}, S_{lk}, S_{lk}) \geq 0$. Under these added assumptions, any partition based on any identity dimensions can be supported as an equilibrium network.*

As in the one dimensional case, we can think of anyone choosing an commitment of 1 along as any dimension as a fanatic and links with players with a different identity along that dimension are impossible. With costs low enough, we can get any partition as an equilibrium.

Next I look at the Strict' Nash networks. Before doing that I use the restriction, as in the one dimensional case, that the choice of $\theta_{kd} \in \{0, 1\}$ for all individuals and for all dimensions. The restriction is the similar to what we used in the one dimensional case and it just means that each individual has only two commitment choices for each dimension - whether to commit to it or not; he doesn't take the qualitative decision of how much to commit to it.

I will now impose some further assumptions, which serve to greatly simplify the analysis of the Strict' Nash networks.

A.M1 : Members in block $B_i > \kappa_1 \times ... \times \kappa_m$ for all $i \in \{1, .., \kappa_1 \times ... \times \kappa_m\}$

A.M2 : $\pi(\{k\}, \{k\}, S, ., .) > 1$ for all $S \neq \mathbf{1}$.

A.M1 says that each block has more members than the total number of blocks. If we are considering only those dimensions of identity which do have a bearing on costs, then this assumption sounds not unreasonable. A.M2 serves to rule out uninteresting cases in which identity does not have an important bearing on costs or when all costs are very low. The assumption also effectively rules out structures similar to the generalised star.

An important structure that emerges in the Strict' Nash networks is the interconnected tail stars which involves numerous clusters of highly committed players linked together by less committed players. Each block B_i from $Block(\mathbf{DIM})$ which belongs to an interconnected tail star, has a core consisting of players who choose to commit to all their characteristics; these players form a center-sponsored star. From these center-sponsored stars emanate tails made up of players not choosing to commit to all their characteristics. Note that a tail for the $B_{i1} - star$ might include players from blocks other than B_i. Along the tails, the player closer to the center of the star forms the link. Each of these tails is then used to establish links with other center-sponsored stars and their tails.

Definition 6 *A structure is called interconnected tail stars if*

(1) if $B_i \in Block(\mathbf{DIM})$ belongs to it, then B_{i1} is nonempty and forms a center-sponsored star.

(2) if $l \in B_i$ chooses $\theta \neq \mathbf{1}$, then it belongs to a tail or there exists some $\{k_1, .., k_l\}$ such that $g_{k_1 k_2} = ... = g_{k_l l} = 1$ and $\theta_{k_1} = \mathbf{1}$.

(3) each tail belonging to any B_{i1} makes a connection leading to some B_{j1}.

Proposition 4 *The Strict' Nash equilibrium, under A1-A3 and the additional assumptions of $\theta_{kd} \in \{0, 1\}$, A.M1, and, A.M2; if it exists, is either empty or is such that each non-empty component is an interconnected tail star.*

I will now give an intuition for the proof of each component being an interconnected tail star given in the appendix. The first lemma says that all people within a block who choose to commit to all their characteristics will be arranged in the form of a center-sponsored star, this is similar is spirit to the proof in the one dimensional case. The next lemma establishes that each block from $Block(\mathbf{DIM})$ must have some internal links if connected, which together with the first lemma and the assumptions implies that each block has a star made up fully committed players. Assuming that all external links are costly means that external links can't be made to one person alone, and that is the driving reason behind the fact that there must be some blocks with internal links. The assumption that each block has more members than the total number of blocks is useful in simplifying the analysis by ensuring that each one of the blocks will have internal connections. The third lemma effectively says that each $B_{i1} - star$ collects a tail (or more) with the first person in the tail being some player from B_i but not B_{i1}. This is a generalisation of the electron star concept, where instead of just one electron we now have a tail. These tailed stars will now form links with each other using any member of a tail. For analogy, in the electron star, the electron was used to link to other blocks. The intuition for the tail comes from the fact that now there are many different commitments that each player can choose and someone not choosing $\theta = \mathbf{1}$ need not necessarily be a part of the tail of his own block, he has more freedom and could get attached to any tail for any star in his component.

Figure 9 shows an interconnected tail star, where the tail star with the identity {Blue, Short, Male} is shown in detail. This blocks has two tails, one leading upto the {Blue, Short, Female} - star and another going to {Blue, Tall, Male} - star.

Efficient networks in the multidimensional identity case as in the one dimensional identity case will have the minimum possible number of external links. Again, let us assume that the profit function is such that a star network is efficient within a block. Moreover only the external links selected will be those which are the cheapest. For instance to connect x blocks, there will be $x - 1$ external links. Each block can be linked to the rest of the blocks using $x - 1$ possible ways/links. Of these $x-1$ possible links, the efficient network will select the cheapest possible link. Of course, in a Strict' Nash network, this externality is not always taken care of. Another source of inefficiency of the Strict' Nash networks is the existence of the longer tail because efficiency requires that no more than one person be in a tail. Again in Strict' Nash networks there is centrality in the sense that all link within the component which are similar must be made by the same player or the star he belongs to. For instance within a $B_{i1}-star$ the central player makes all the links. If from a tail of the $B_{i1}-star$ a player, say l, makes a link with similarity index S, then any other link which again has similarity index S (or more) with l , must be accessed by the same player or another tail of $B_{i1}-star$. Coupling this centrality with the fact that the cost of any link is borne by the initiator of the link, Strict' Nash networks loose a lot of efficiency. Link between blocks are possible in Nash networks when link costs are substantially below those needed in efficient networks. Though the structure of the efficient networks will still be interconnected stars, which is a special case of the interconnected tail star, so there will be cost ranges when Strict' Nash networks are efficient.

3 Identity and Community Structures

We now have an explanation about why networks would be partitioned. We know that different choices of commitment would lead to different partitions within the same set of players. What we do not know, is how to deduce the actual partitions given the data on links and identity characteristics. We hardly ever expect to find clear divisions as in the Nash networks. What we would like to find is which dimension of identity seems to be important in dividing society. To exemplify, look the the network in Figure 10, where players have identity along the dimension of Colour (white/black) and Shape (square/triangle). Looking at this figure its not clear which (if any) dimension of identity is more important in the partition. In the next two figures, we rearrange the network data once by Colour and next by Shape, and here we see that visually it is clear that the Shape is more important in generating the link data.

To see the role of commitments in the empirical strategy, keep in mind that commitments and links are chosen simultaneously, and the choice of one affects the other; knowing either would give a good idea of what the other would be. If we knew the commitment choices, we would have a natural way of ordering the data. For instance, in the example above, the network would be possible given that the commitment to Colour would be very low for all players, but the commitment for Shape should be high for most players. On the other hand, given that we know the partition is more likely Shape, we know that commitments for Shape would in general be higher than commitments for Colour. In other words, determining one should be sufficient to impute the other.

We now build the estimation strategy which is a generalisation of the ideas presented in the in Figures 10, 11 and 12. What we try to do is build an estimation strategy based on attaching likelihood numbers to the various possible partitions. And we pick the partition which maximises this likelihood. In building the estimation strategy we will incorporate the qualitative results of the theoretical model, but leave out quantitative predictions which arise from assumptions which can not be expected to hold in the data. One such prediction is that in a separated network there will be absolutely no links between two components and a network will be connected with just one link between two groups. Under more realistic assumptions, e.g. there is some error term in the payoffs, linking to some players gives higher values, etc, we would not expect complete separation. Allowing for error in the payoffs or allowing for mixed strategies, we model the link strategy as the probability of linking to another player based on identities. Another assumption that would possibly not hold in the data is the no decay assumption, relaxing which would lead to more links being formed than predicted by the Nash networks. Relaxing these assumptions, what we could in fact observe in the data would be what are called "communities". A community is a collection of people, such that each member of the community is more likely to have links with someone from the community than with someone outside of the community. A community structure is then the collection of all such communities in a population. An important insight that we keep from the theoretical model, is that communities will be built along identity dimensions and that the probability of forming links will depend on the identity of the two persons.

3.1 Identifying Community Structure given Agents' Identity

The data we expect to observe is a random sample of all possible interactions, as well as identities. What we would like to find out is the community structure and the probabilities of interaction. The method proposed here involves selecting the community structure and probabilities of interaction which maximise the likelihood of observing the data. I will now outline the likelihood strategy in detail.

Definition 7 *For two players with similarity index S, p^S_{in} is the probability that they link within the same community and p^S_{out} is the probability that they link while belonging to different communities.*

Given our assumption that community structures are based on identities, we know that the only possible community structures are the ones which have divisions along the dimensions.

Definition 8 $\mathbf{\Pi}^{DIM}$ *is the set of all community structures which involve divisions along dimensions of identity included in* **DIM**.

Note that the community structure is defined by the dimensions of identity only and it does not depend on the number of individuals in the community. $\mathbf{\Pi}^{DIM}$ effectually defines a partition over identity blocks.

Let p_{in} be the set which collects all possible p_{in}^S and p_{out} the corresponding set collecting all p_{out}^S. Let $\mathbf{P}^{DIM}$ denote the space of all feasible (p_{in}, p_{out}) given **DIM**. Let $\pi \in \mathbf{\Pi}^{DIM}$ denote a partition from $\mathbf{\Pi}^{DIM}$ and let $c_\pi(i)$ denote the component which contains i. Let g_{ij} denote the number of independent interactions between i and j in the network g and let h_{ij} denote the maximum possible such independent interactions between i and j in any network. The likelihood of observing the data is given by:

$$\begin{aligned} L_{h;g}(\pi, p_{in}, p_{out}) &= C \times_{i \in N} [(\times_{j \in c_\pi(i)} (p_{in}^{S_{ij}})^{g_{ij}} (1 - p_{in}^{S_{ij}})^{(h_{ij} - g_{ij})} \\ &\quad (\times_{j \in N \setminus c_\pi(i)} (p_{out}^{S_{ij}})^{g_{ij}} (1 - p_{out}^{S_{ij}})^{(h_{ij} - g_{ij})}] \end{aligned}$$

Then the likelihood approach will be to:

$$\text{Choose } \{\pi, p_{in}, p_{out}\} \text{ to } \max L_{h;g}(\pi, p_{in}, p_{out})$$

$$\begin{aligned} \text{such that } \pi &\in \mathbf{\Pi}^{DIM}, \\ p_{in}^S &> p_{out}^S \text{ for all } S \end{aligned}$$

For the next few propositions let $(\mathbf{DIM}^*, \pi^*, p^*)$ denote the true data generating process. The next proposition proves that this method is consistent.

Proposition 5 *Let n^t be a sequence of population size; such that the network size $n^t(n^t - 1) \to \infty$. Generate g^t using $(\mathbf{DIM}^*, \pi^*, p^*)$. Let π^t, p^t be the maximisers of the likelihood for network g^t. Then as $t \to \infty$, $\pi^t \to \pi^*$ and $p^t \to p^*$.*

The next proposition says that the likelihood will strictly increase as we add dimensions of identity which are part of $\mathbf{DIM}^*$.

Proposition 6 *Let $L_{h;g}(\mathbf{DIM})$ denote the maximised likelihood when searching over dimensions* **DIM**. *Let D be a dimension such that $D \in \mathbf{DIM}^*$ but $D \notin \mathbf{DIM}$ and let $\mathbf{DIM'} = \{\mathbf{DIM}, D\}$. Then as $n_t(n_t - 1) \to \infty$, $L_{h;g}(\mathbf{DIM'}) > L_{h;g}(\mathbf{DIM})$.*

The above propositions suggest the following search algorithm:

- Layer 1: Begin with one identity dimension, and find the highest likelihood. Repeat this for all other dimensions. Find the dimension (as well as the partition and probabilities) which maximises the likelihood.
- Layer 2: Use the dimension from layer 1, as the primary identity dimension. Combine that dimension with a second identity dimension and find the pair which maximizes the likelihood.
- Layer k: Use the dimensions which maximised likelihood for layer k-1 as the primary set of identity dimensions as the fixed dimensions of identity. Repeat stage 2 using the new fixed dimensions of identity.
- Within a layer, for any set of dimensions, start with the finest community structure (all blocks separate) and keep making it coarser (by combining blocks) until the likelihood is maximized.

Hypothesis Testing: Once we have the maximised likelihood for any layer, we want to check and see if this likelihood is significantly different from likelihoods at layers lowers than this. Because at each layer we are adding another dimension of identity to the last layer, we will be interested in knowing if for $\mathbf{DIM}' = \{\mathbf{DIM}, D\}$, the maximised likelihood using $\mathbf{DIM}'$ is significantly more than the maximised likelihood using $\mathbf{DIM}$. Let $\pi(\pi')$ be the community structure which maximises likelihood if dimensions are $\mathbf{DIM}(\mathbf{DIM}')$. The form of the likelihood function will be different depending on whether we use π' or π. We would like to pick the partition and probabilities which maximise the likelihood, and by introducing a new variable $\lambda \in \{0, 1\}$ we can think of the problem as being:

$$M(\lambda, p) = \max_{\lambda, p}\{\lambda(L_{s;g}(\pi, p)) + (1-\lambda)(L_{s;g}(\pi', p))\}$$

The maximised likelihood under $\mathbf{DIM}$ is maximum of $M(\lambda, p)$ when we constrain $\lambda = 1$ and restrict $p \in P^{\mathbf{DIM}}$. Applying standard LR techniques we can check if the maximised likelihood under $\mathbf{DIM}'$ is significantly greater.

We want to compare the case where there is no partition to the maximised likelihood under layer 1. If there is no partition, then the entire sample is one community and there is only the probability of making links within the community. Under layer 1, the data is (potentially) partitioned using one dimension of identity. In this case, there are three probabilities of making links: {probability of being in the same community with same characteristic, probability of being in the same community with different characteristic, probability of being in different community with different characteristic}.[9] By allowing the data to be partitioned along one dimension, we add 3 degrees of freedom over the case with no partition - two for the added dimensions of probability and one for λ. Similarly, by allowing a layer of two, we add seven degrees of freedom over the case with no partition. The layer of two adds five degrees of freedom over the layer of one.

3.2 Estimating Community Structure in Ghana

The data was collected by Chris Udry and Markus Goldstein over the course of two years and fifteen modules in a four village clusters in Eastern Region of Ghana. In each village 60 couples/triples were questioned. The network data used here was collected by asking each individual in the sample about seven randomly selected (without replacement) from the sample and three focal village residents. The questions asked were:

Could you go to ___ if you had a problem with unhealthy crops?

Could you go to ___ for advice about when to apply a new kind of fertilizer?

Could you go to ___ if you wanted to discuss changing your method of planting?

Could you go to ___ if you wanted to find a buyer for any of your crops?

If we think of the village residents as the population participating the network formation game, then the randomly selected 60 couples and further their links with randomly selected seven individuals from within that sample, allows us to see a randomly selected portion of the network. Analysing the structure of connections within this portion of the network would give us a good idea of the actual network.

I also use data on identity and this includes information on the respondent's religion, clan, gender, if they are the first of their family to reside in that village, and the crops grown.

3.2.1 Characterization of Data

The network data I use here looks at four related information networks which look at information flows on unhealthy crops, fertilizers, methods of planting and buyers. Table 1 gives the summary statistics for the link variables and it turns out that each respondent on average contacts approximately three from his sample of ten for information on unhealthy crops, fertilizers and methods of planting, and for information on buyers. From Table 2, we can see that the four kinds of links are highly correlated. In fact looking at the data it turns out that for many respondents, if they ask their matched respondent about any one unhealthy crops, fertilizers and methods of planting,

[9] We constrain probability of being in different community with same characteristic to being equal to zero, since we assume the blocks move together.

then they ask about the other two as well. For this reason, the rest of the analysis will take into consideration only the link indicated by the first question on unhealthy crops.

We need to be able to sort people into groups along different dimensions of identity. The summary statistics for the identity variables used for the all the respondents are presented in Table 3. The variables used are whether the respondent is the first of family in the village, the religion of respondent[10], whether the respondent grows pineapple or not[11], respondent's clan[12] and gender. For each of the identity variables, I construct another variable which take the value 1 if both the respondent and his match have the same characteristics (or belong to the same group) under that identity dimension. The summary statistics for these similarity variables are presented in Table 4.

The correlation structure of the links with the identity variables is presented in Table 5. The variables are such that they take a value of 1 if both the respondent and the match share the same characteristic in that identity dimension and 0 otherwise. As can be seen, some of the correlations are negative, implying that links are more probable when the characteristics is not the same and that there might be gains to having links with individuals with different characteristics. Another explanation might be that different identities have different pieces of information, and the respondent values more the information possessed by someone he does not share the identity characteristic with.

3.2.2 Community Structures in the Four Villages

Table 6 shows the results when we search over layer 1. For each village and each dimension of identity I report the log likelihood corresponding to the best partition of that village along that dimension. For the maximised log likelihood, I also report whether this likelihood is significantly different from the baseline likelihood of assuming no partition and the probability of the link not depending on identity. '-Inf' indicates the fact that no feasible partition exists along that dimension of identity. We see that a for most of the villages there is in fact no feasible division along the variable 'Firsthere', implying that there a lot of links across those who are the first of the village here and those who are not. We find that villages 1 and 3 divide along clan, village 2 divides along pineapple growers and village 4 divides along religion. But of these divisions only the divisions for village 1 and 4 are significantly different from the baseline assumption of no division.

Next we look at partitions along layer 2 in Table 7. We keep one dimension fixed (at the one

[10] I keep only the religions which had at least 5 members

[11] pineapple was a relatively new crop at the time of the survey and we would expect those who did crop pineapple to want to share information with each other

[12] again I keep only the clans which have at least 5 members

which maximised likelihood at layer 1) and to this we add the other 4 dimensions and report the maximising likelihoods along the two dimensions. We see that all these likelihoods are significantly different from the baseline likelihood. Village 1 divides along clan and religion, village 2 divides along religion and pineapple growing, village 3 divides along clan and religion; and village 4 divides along gender and religion.

To get a better understanding of the community structures, we present four graphs show the community structures for village 3 and 4 for layer 1 and 2 (Figures 13, 14, 15, and, 16). The most interesting results are for village 4, which divides along religion when searching over layer 1. It shows three religions combining to form one community and the Pentecostal's forming a separate community. This is contrary to our result for Nash equilibrium with one dimension of blocks being all connected or separated. Then this kind of division points to the presence of another dividing dimension of identity. In the next figure, Figure 13, we see that in fact, this village shows layers of divisions. It first divides along gender and then further subdivides the females by religion. Religion is in fact a very strong dividing line for women, who in many cases are more likely to link to other men than to women with a different religion. The division along gender is more difficult to explain, but it might be the case that women just participate less in information networks.

4 Conclusion

This paper presented looked at the impact of identity on networks. We saw a theoretical model of network formation which allowed for the choice commitments to identity simultaneously with the choice of links. The Nash networks arising in this framework exhibited partition along identity, and, interestingly, these partitions are not unique. In other words, they allow for the fact that populations with similar identity profiles might be partitioned very differently. If we restricted attention to those Nash equilibria where players strictly preferred their link strategy to any other, the network structures that emerged, featured center sponsored stars of strongly committed players linked together by less committed players.

Given that the Nash networks could have many different partitions, the empirical section of the paper proposes and implements a methodology to extricate the salient identity dimensions and partition given network data. Applying the methodology to network and identity data from four villages in Ghana, we see that the four villages featured different partitions. In other words, the multiplicity of Nash equilibria is bourne out in the data.

These results point to the fact that partitions in societies along a particular identity dimension might be seen as a coordination problem - players could as well have coordinated to partition along some other dimension. It also points out that the population as a whole chose to partition along that dimension rather than choose any other dimension or none. Since, over time, these partitions seem to change even though the underlying population identity profile does not, future work could focus on the understanding the evolution of these changes.

A Network Formation Proofs

A.1 Proof for Nash Networks in One-Dimensional Case

Lemma 1 *If any $k \in B_i$ belongs to a non-singleton component $C(g)$, then $B_i \subseteq C(g)$.*

Proof. Suppose $k \in B_i$ belongs to a component $C(g)$ and assume to the contrary $k' \in B_i$ and $k' \in C'(g)$. Both k and k' must be receiving or forming some links, and these must have non-negative payoffs. Lets consider the possible scenarios:

1) If either k' or k forms no links then the player who does not form links will wish to deviate to choosing $\theta = 1$ and form a link with the other.

2) If both form links, and assume that the links of k are more profitable than those of k', then k' could profitably deviate to choosing $\theta_{k'} = 1$ and form a single link with k. Since k' will then be accessing all of the links of k using a single link, he will make higher profits than k. ■

Lemma 2 *If $g_{kl} = 1$ for $k \in B_i$ and $l \in B_j$ where $i \neq j$, then the network is connected.*

Proof. Let k and l belong to the component $C(g)$. From the previous lemma it must be that $B_i \subseteq C(g)$ and $B_j \subseteq C(g)$. Since k and l are connected, it must be that this link if profitable. Again from the previous lemma we know that either B_a, where $a \notin \{i, j\}$, has no link or it is connected. If it has no links then any member of B_a can do better by setting $\theta = 0$ and forming a link with k or l. If on the other hand B_a is connected and belongs to some other component $C'(g)$, then it must be that $C'(g)$ is minimally connected, and so there must be at least one player within $C'(g)$ who does not form any links with any other player in $C'(g)$. This player can then set $\theta = 0$ and form a link with k or l. ■

A.2 Proof for Strict' Nash Networks One-Dimensional Case

Lemma 3 *All player's who choose to make links only within their characteristic will choose $\theta = 1$ while all player's who choose to make links with players outside their characteristic will choose $\theta = 0$*

Proof. Since we restrict attention to $\theta \in \{0, 1\}$, this must be so. Though choice of θ for players not forming any links is uncertain. ■

Lemma 4 *If $g_{kl} = 1$ where $k, l \in B_{i1}$ then it must be the case that $g_{km} = 1$ for all $m \neq k, m \in B_{i1}$ and there will be no other links within B_{i1} (This structure will be referred to as the $B_{i1} - star$)*

Proof. Suppose $g_{kl} = 1$ where $k, l \in B_{i1}$ and let $k, l \in C(g)$. It must be that $B_{i1} \in C(g)$. Let $m \in B_{i1}$ where $m \neq k, l$. Since forming a link with k or l costs m the same and has the same benefits, in a strict nash m will not form either link. Moreover, m will also not form a link with anyone linked to k or l, because linking to k or l is weakly better than that. If $g_{lm} = 1$ then k is indifferent between being linked to l or switching to a link with m, hence l and m must not have a direct link. Since m must be connected to k, l in some way, the only possibility besides $g_{km} = 1$, is one where $g_{k'm} = 1$ for some $k' \notin B_{i1}$ such that $k' \overset{g}{\leftrightarrow} k$. Suppose wlog $\overline{g}_{k'k} = \max\{g_{kk'}, g_{k'k}\} = 1$, but $g_{kk'} = 1$ is not possible because k would get the same benefits by linking to m and $g_{k'k} = 1$ is not strict nash because k' is indifferent between linking to k or to l. Hence it must be that $g_{km} = 1$.
■

Lemma 5 *In a Strict' Nash, if $B_{i1} - star$ exists and is not a singleton, it forms links with all $l \in B_{i0}$.*

Proof. Suppose $l \in B_{i0}$ and l does not receive a link from $B_{i1} - star$. Since in any Nash network, all of B_i must belong to the same component, there must be some k such that $l \overset{g}{\leftrightarrow} k$ and $k \overset{g}{\leftrightarrow} l'$ for some $l' \in B_{i1}$. Wlog assume $\overline{g}_{kl} = 1$ and $\overline{g}_{kl'} = 1$. Since k will be indifferent amongst linking to different member of B_{i1}, in a Strict' Nash network it must be that $g_{l'k} = 1$. But then either $k \in B_{i1}$ or l' would weakly prefer linking to l and so in Strict' Nash is must be that $g_{l'l} = 1$.
■

Lemma 6 *In a Strict' Nash, if, $\pi(\{k\}, \{k\}, S_{lk} = 0, \theta_k = \theta_l = 0) < 0$ for all k, l s.t. $S_{lk=0}$; then, only players with $\theta = 0$ can receive a direct outside links. Moreover, the player receiving an outside link forms/receives no other link with any other player who also has $\theta = 0$.*

Proof. Suppose not and $k \in B_i$ and $l \in B_j$ such that $g_{kl} = 1$ and $\theta_l = 1$. Since solely linking to l is not profitable, it must be that l has some other links and let l' be such that $\overline{g}_{l'l} = 1$. But then it must be that the cost of a link between k and l' can not be greater than the cost of a link between k and l. And so in a Strict' Nash, if $k \in B_i$ and $l \in B_j$ such that $g_{kl} = 1$ then it must be that $\theta_l = 0$ and l can not be linked to any other player with $\theta = 0$. ■

Lemma 7 *In a Strict' Nash, if $\pi(\{k\}, \{k\}, S_{lk} = 0, \theta_k = \theta_l = 0) < 0$ then any block B_i is internally connected*

Proof. Let $l, l' \in B_i$. Similar to the last lemma, wlog, assume there is some $k \notin B_i$ such that $g_{kl} = 1$ and $g_{kl'} = 1$. Since $\pi(\{k\}, \{k\}, S_{lk} = 0, \theta_k = \theta_l = 0) < 0$ for the $g_{kl} = 1$ and $g_{kl'} = 1$ to be sustainable, l and l' must have other links. But linking to l or linking to l' is the same in terms of cost and benefits, implying that neither can receive any links. So they both must be making links and moreover only to players choosing $\theta = 1$. If l or l' made links to someone from B_{i1},then they would want to switch to $\theta = 1$. So they must be making links to B_{j1} where $i \neq j$. But from the previous lemma, an external link to any player choosing $\theta = 1$ is not possible. And so in a Strict' Nash network, B_i must be internally connected. ■

Lemma 8 *In a Strict' Nash, if* $\pi(\{k\}, \{k\}, S_{lk} = 0, \theta_k = \theta_l = 0) < 0$ *and* $g_{kl} = 1$ *for* $k \in B_i$ *and* $l \in B_j$, *then it must be that every other member of* B_j *belongs to the* $B_{1j} - star$.

Proof. Since $\pi(\{k\}, \{k\}, S_{lk} = 0, \theta_k = \theta_l = 0) < 0$; from an earlier lemma we know that B_j must be internally connected. From another lemma we know that l cannot have any links with l' if $\theta_{l'} = 0$. Which means that the rest of B_j must choose $\theta = 1$. l will not sponsor any links to any member of B_{1j} because then he would want to switch to $\theta = 1$. Implying that B_{1j} is linked within itself and then sponsors a link to l. And so we must have a $B_{1j} - star$. ■

Lemma 9 *In a Strict' Nash, if* $\pi(\{k\}, \{k\}, S_{lk} = 0, \theta_k = \theta_l = 0) < 0$ *and* $g_{lk} = 1$ *for* $l \in B_i$ *and* $k \in B_j$ *and* $\theta_k = 0$, *then all external links are sponsored by* B_i.

Proof. Suppose not and some $x \notin B_i$ makes an external link to y. The network must be connected and so suppose wlog, l forms a links with y. If $\theta_x = \theta_y = 0$ then l will be indifferent between linking to x or y. So suppose $\theta_x = 1$. But in that case, x will be indifferent to linking to y, k. ■

Lemma 10 *In a Strict' Nash, if* $\pi(\{k\}, \{k\}, S_{lk} = 0, \theta_k = \theta_l = 0) > 1$ *and* $g_{lk} = 1$ *for* $l \in B_i$ *and* $k \in B_j$ *and* $\theta_l = 0$, *then either* $B_{i1} - star$ *exists and sponsors a link to* l *or* l *forms the center of the star for* B_{i1}.

Proof. If l forms a link with some $l' \in B_{1i}$, then l must form links with all other member of B_{i1} by strictness. In other words, l should either be the center of the $B_{i1} - star$ or it will receive a link from them. ■

A.3 Proof for Nash Network in Multi-Dimensional Case

Lemma 11 *The Minimal Separating Dimensions* $\mathbf{DIM}^1$ *from a network* g *are unique.*

Proof. Suppose not and that there exist $\mathbf{DIM'} \not\subset \mathbf{DIM}^1$ such that $\mathbf{DIM'}$ are also minimal separating dimensions. But if they are both separating dimensions, then there are no links across different $B_i \in Block(\mathbf{DIM'})$, neither across different $B_j \in Block(\mathbf{DIM}^1)$. Which would imply that the set of dimensions $\mathbf{DIM}^1 \cup \mathbf{DIM'}$ are also separating dimensions. In other words, $\mathbf{DIM}^1$ cannot be the minimal separating dimensions to begin with. ■

Lemma 12 *A nash network must be minimal.*

Proof. If it was not, then some links could be deleted without impacting connectivity. ■

Lemma 13 *In any nash network g, if any $k \in C(g)$ and $k \in B$ where $B \in Block(\mathbf{DIM})$ and $|C(g)| = x > 2$, then $B \in C(g)$.*

Proof. Similar to the one-dimensional case, suppose $k' \in B$ and $k' \in C'(g) \neq C(g)$. Since $k \in C(g)$, it must be receiving/forming some links within $C(g)$ which we assume are (wlog) more profitable than the links of k'. For any link structure, it will be profitable for k' to set $\theta_{k'} = \mathbf{1}$ and add a link to k. ■

Lemma 14 *Either the network is connected/emppty or there exists a set of minimal separating dimensions.*

Proof. If the network is neither connected nor empty' then it must be that either $\mathbf{DIM}^1 = \mathbf{DIM}$ or ϕ will work as separating dimensions. We have earlier proven, if separating dimensions exist, the minimal separating dimensions are unique. ■

Lemma 15 *Suppose along any $\mathbf{DIM}^1 \subset \mathbf{DIM}$, there is some $B_i \in Block(\mathbf{DIM}^1)$ which is not connected within itself and no member of B_i is connected to any member outside B_i. Then it must be that either B_i is connected or empty, or there exists some $\mathbf{DIM}^1 \subset \mathbf{DIM}^{1,B_i} \subseteq \mathbf{DIM}$, such that g allows no links across the blocks of $Block(\mathbf{DIM}^{1,B_i})$ within B_i.*

Proof. Such a partition must exist because $\mathbf{DIM}^{1,B_i} = \mathbf{DIM}$, or ϕ will definitely work. We need to show that $\mathbf{DIM}^1 \subset \mathbf{DIM}^{1,B_i}$, which follows from the fact that since all characteristics along $\mathbf{DIM}^1$ are the same for all members of B_i, it must be that any link will be cost minimized if its along $\mathbf{DIM}^1$ and more dimensions. ■

A.4 Proof for Strict' Nash in Multi-Dimensional Case

Lemma 16 *In a strict nash network, if $l, k \in B_{i1}$ and l forms the $lk - link$ then the members of B_{i1} form a unique center-sponsored star with l in the center (henceforth called a $B_{i1} - star$)*

Proof. Same as in the one-dimensional case. ■

Lemma 17 *Assuming $\pi(k, k, S_{lk}, \Theta) \leq 0$ for all $S \neq \mathbf{1}$, each block within a component of the nash network must have internal connections. Using A.M1, each block, B_x, must have a $B_{x1} - star$.*

Proof. Suppose $\{B_1, ..., B_x\} \in C(g)$ and suppose contrary to the assumption, B_1 does not have any internal connections. But then all members of B_1 must be participating in external links. Since external links are expensive enough, each external link must end with a cluster of internally connected individuals. If the members in B_1 are greater than the number of blocks, as implied by A.M1, then this in not possible.

If AM.1 does not hold, then the block could be scattered among the tails of other stars. But even then, we must have at least two blocks for whom there are internal links.

■

Lemma 18 *If any $l \in B_x$ forms an external link, then he must either receive a link from $B_{x1} - star$ or be the part of the tail of some other $B_{y1} - star$ where $B_x \neq B_y$ and this tail should have no links to the $B_{x1} - star$*

Proof. Suppose $l \in B_x$ does not receive a link from $B_{x1} - star$. He will not link to the $B_{x1} - star$, because he is indifferent to linking to any one of them. He must be forming an external link which leads to some $B_{y1} - star$.. The only way for l to be connected to the $B_{x1} - star$ is for him to receive an external link which would indirectly link him the $B_{x1} - star$. ■

Lemma 19 *If $g_{lk} = 1$, where $l, k \in C(g)$, $l \in B$ and all other $p \in C(g)$ such that $S_{lp} \geq S_{lk}$ or $S_{lp} < S_{lk}$, then all $k' \in C(g)$ such that $S_{lk'} = S_{lk}$ either receives a direct link from l' where $l' \overset{g}{\longleftrightarrow} l$ and $S_{ll'} > S_{lk}$ or is $k'' \overset{g}{\longleftrightarrow} k'$ and $S_{k''k'} > S_{lk}$*

Proof. If $g_{lk} = 1$, then $\theta_l = \theta_k = S_{lk}$. The only way they could be different is by choosing commitment less than S_{lk} and the only reason θ_l or θ_k could be different is if they made some other links, but there are no other links to be made which are less than $S_{lk.}$

Since $\theta_l = \theta_k = S_{lk}$, noone would form an S_{lk} link with either of them due to strictness. Infact all S_{lk} links must originate with l or some l' who is linked to l and $S_{ll'} > S_{lk}$. ■

B Community Structure Proofs

B.1 Alternative Representation

An alternative representation using log-likelihood instead of likelihood is also possible. For that we define a few terms

$$\begin{aligned} In(S,\pi) &= \{ij \mid S_{ij} = S,\ j \in c_\pi(i)\} \\ Out(S,\pi) &= \{ij \mid S_{ij} = S, j \notin c_\pi(i)\} \end{aligned}$$

$In(S,\pi)$ collects all the pairs which have similarity index S and which belong to the same community, while $Out(S,\pi)$ collects all the pairs which have similarity index S and which belong to the different communities. Let

$$T(S,g) = \sum_{S_{ij}=S} g_{ij} \text{ and } T(S,h) = \sum_{S_{ij}=S} h_{ij}$$

denote the total links for all those pairs who have the same similarity index, S, under g and the total possible links for the same pairs.

$$T^{In(S,\pi)}(S,g) = \sum_{ij \in In(S,\pi)} g_{ij} \quad \text{and } T^{In(S,\pi)}(S,h) = \sum_{ij \in In(S,\pi)} h_{ij}$$

$T^{In(S,\pi)}(S,g)$ is the total links observed under g for all pairs belonging to $In(S,\pi)$ and $T^{In(S,\pi)}(S,h)$ is the maximum possible links we could observe in any network for the pairs in $In(S,\pi)$. $T^{Out(S,\pi)}(S,g)$ and $T^{Out(S,\pi)}(S,h)$ can be similarly defined. Using these definitions and ignoring the constant, we get the log-likelihood:

$$\begin{aligned} l_{s;g}(\pi, p_{in}, p_{out}) &= \log(L_{s;g}(\pi, p_{in}, p_{out})) \\ &= \sum_S \left\{ \begin{array}{c} (T^{In(S,\pi)}(S,g) * \log(p_{in}^S)) + \\ ((T^{In(S,\pi)}(S,h) - T^{In(S,\pi)}(S,g)) * \log(1 - p_{in}^S))) \\ +(T^{Out(S,\pi)}(S,g) * \log(p_{out}^S)) \\ +((T^{Out(S,\pi)}(S,h) - T^{Out(S,\pi)}(S,g)) * \log(1 - p_{out}^S)) \end{array} \right\} \end{aligned}$$

Defining $k_1^S = \log(p_{in}^S/(1-p_{in}^S)), k_2^S = \log(1-p_{in}^S), k_3^S = \log(p_{out}^S/(1-p_{out}^S)), k_4^S = \log(1-p_{out}^S)$, we get:

$$l_{s;g}(\pi, p_{in}, p_{out}) = \sum_S \left\{ \begin{array}{c} k_1^S * T^{In(S,\pi)}(S,g) + k_2^S * T^{In(S,\pi)}(S,h) \\ +k_3^S * T^{Out(S,\pi)}(S,g) + k_4^S * T^{Out(S,\pi)}(S,h) \end{array} \right\}$$

B.2 Proofs of Proposition 6

Proof. Firstly, we fix the dimensions as the largest possible, and next, if we fix the partition, the likelihood will be maximised at the probabilities given by:

$$\widehat{p}^S_{in}(\pi) = \frac{T^{In((S,\pi)}(S,g)}{T^{In((S,\pi)}(S,h)}$$

$$\widehat{p}^S_{out}(\pi) = \frac{T^{Out((S,\pi)}(S,g)}{T^{Out((S,\pi)}(S,h)}$$

Given this result, we get consistency by showing that as size becomes larger, it must be that for any $\pi \neq \pi^*$,

$$L_{h;g}(\pi^*, \widehat{p}(\pi)) > L_{h;g}(\pi, \widehat{p}(\pi))$$

There are countably many ways in which π^* and π can differ, but the manner of difference will not affect result. I consider one particular way in which they are different to illustrate the proof. Suppose that the only difference is that within all links with similarity index S; π has more pairs within the same component, or:

$$\begin{aligned} In(S,\pi) &\supset In(S,\pi^*) \\ Out(S,\pi) &\subset Out(S,\pi^*) \end{aligned}$$

and

$$Out(S,\pi^*) \cap In(S,\pi) = K$$

Under this particular π and π^*, the differences in log-likelihood using $\widehat{p}(\pi)$ is:

$$\begin{aligned} L_{h;g}(\pi, \widehat{p}(\pi)) - L_{h;g}(\pi^*, \widehat{p}(\pi)) \quad = \quad & \left(\sum_{ij\in K} g_{ij}\right) * \log(\widehat{p}^S_{in}(\pi)) \\ & + \left(\sum_{ij\in K} h_{ij} - g_{ij}\right) * \log(1 - \widehat{p}^S_{in}(\pi)) \\ & - \left(\sum_{ij\in K} g_{ij}\right) * \log(\widehat{p}^S_{out}(\pi)) \\ & - \left(\sum_{ij\in K} h_{ij} - g_{ij}\right) * \log(1 - \widehat{p}^S_{out}(\pi)) \end{aligned}$$

Taking the limit of the derivative with respect to $\widehat{p}^S_{in}(\pi)$, we get:

$$\lim_{n^t(n^t-1)\to\infty} \left(\frac{\partial(L_{h;g}(\pi, \widehat{p}(\pi)) - L_{h;g}(\pi^*, \widehat{p}(\pi)))}{\partial \widehat{p}^S_{in}(\pi)}\right) = \lim_{n^t(n^t-1)\to\infty} \left(\frac{\sum\limits_{ij\in K} g_{ij}}{\widehat{p}^S_{in}(\pi)} - \frac{\sum\limits_{ij\in K} h_{ij} - g_{ij}}{1 - \widehat{p}^S_{in}(\pi)}\right) < 0$$

The last inequality follows because as $n^t(n^t-1) \to \infty$,

$$\frac{\sum\limits_{ij \in K} g_{ij}}{\sum\limits_{ij \in K} h_{ij}} = p^S_{out} < \widehat{p}^S_{in}(\pi)$$

(Since $\widehat{p}^S_{in}(\pi)$ in the limit will be some convex combination of p^S_{in} and p^S_{out}). ■

B.3 Proof of Proposition 7

Proof.

Let (π, p_{in}, p_{out}), $(\pi', p'_{in}, p'_{out})$ and $(\pi^*, p^*_{in}, p^*_{out})$ be the community structure and the probabilities of interaction which maximise the likelihood restricting dimensions to **DIM**, **DIM**$'$ and **DIM***.

Now, if **DIM*** are the true partitioning dimensions, it must be that the true partition and probabilities $((\pi^*, p^*_{in}, p^*_{out}))$ must be different from both $((\pi, p_{in}, p_{out}))$. Also, since **DIM**$'$ includes one more dimension than **DIM**, it must lead to a partition or probabilities closer to the true partition/probabilities. Hence, using similar methodology as in the previous proposition, we must have the maximised likelihood under **DIM**$'$ strictly better than under **DIM**$'$.

■

C Network Formation Graphs

(For all the following figures: Each box represents a player. It lists the identity and then the corresponding commitment levels in the brackets)

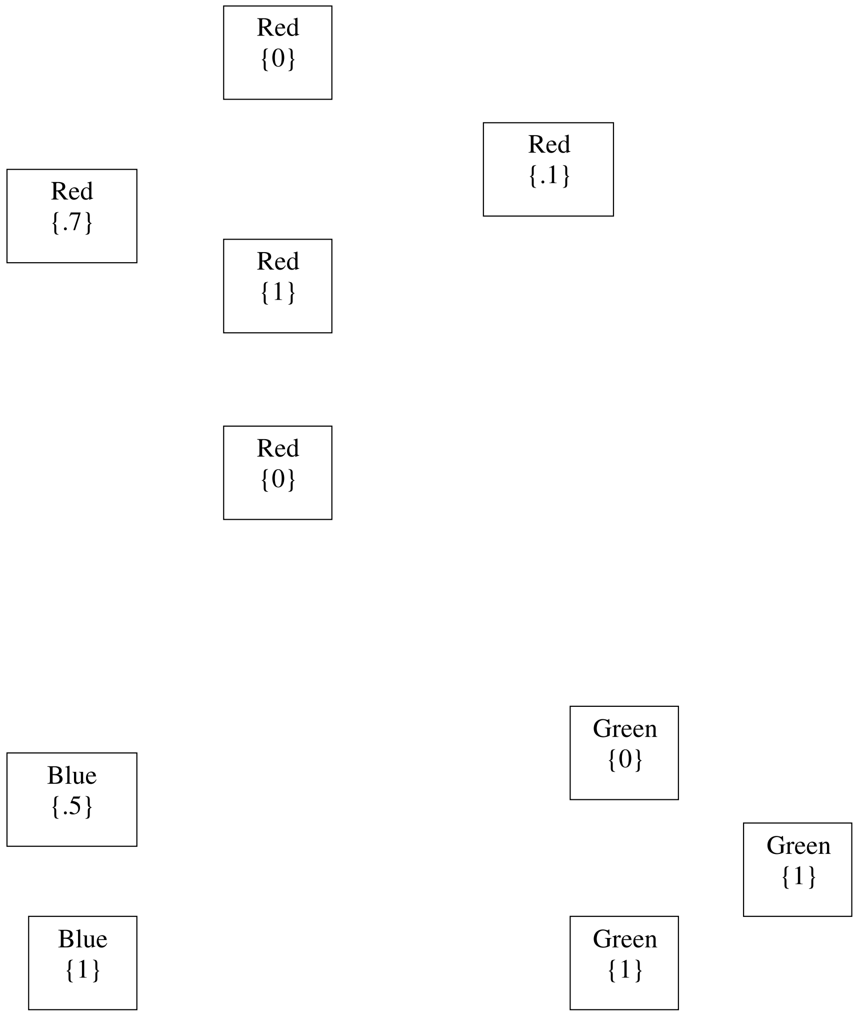

Figure 1: Unconnected Nash

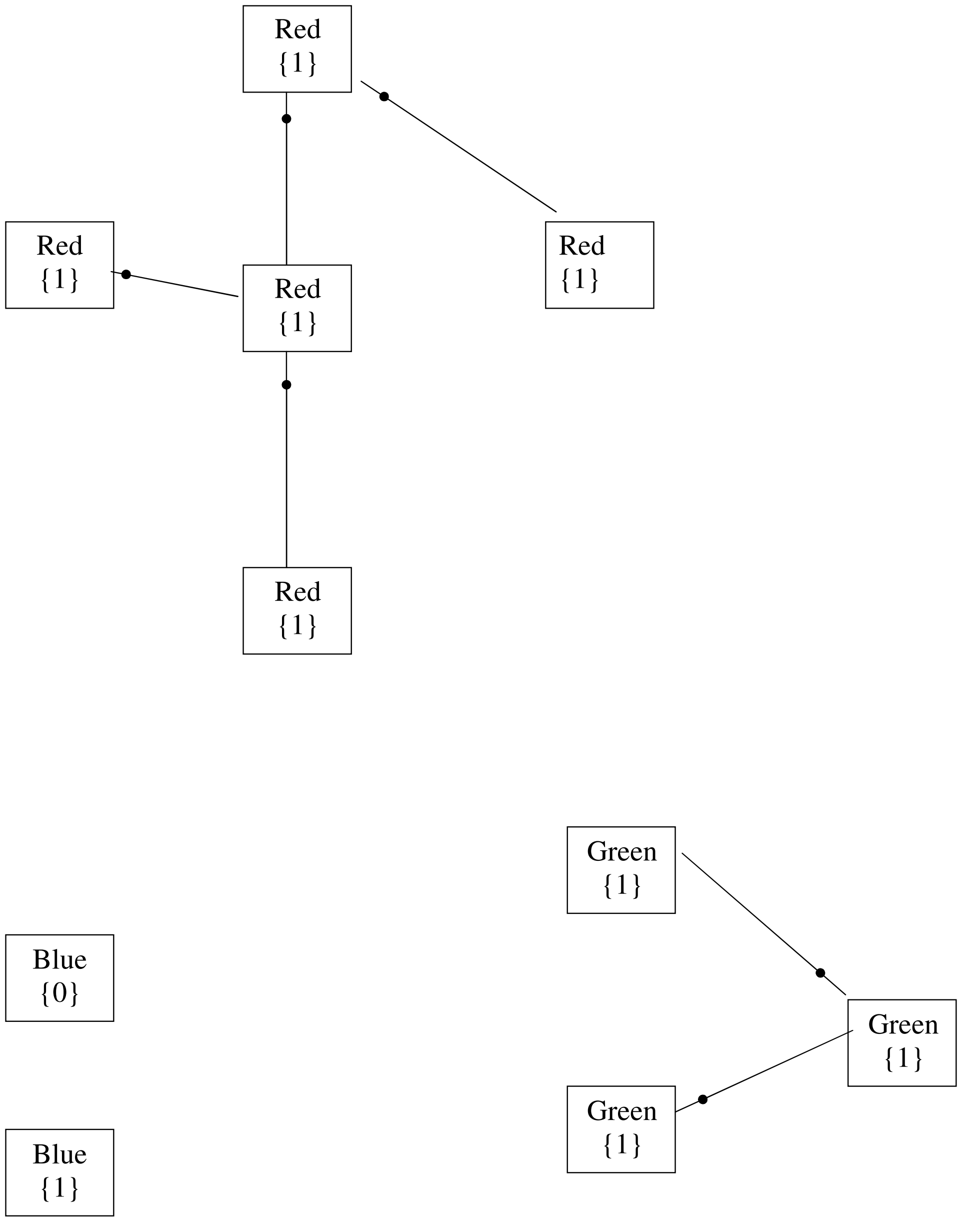

Figure 2: Separated Nash

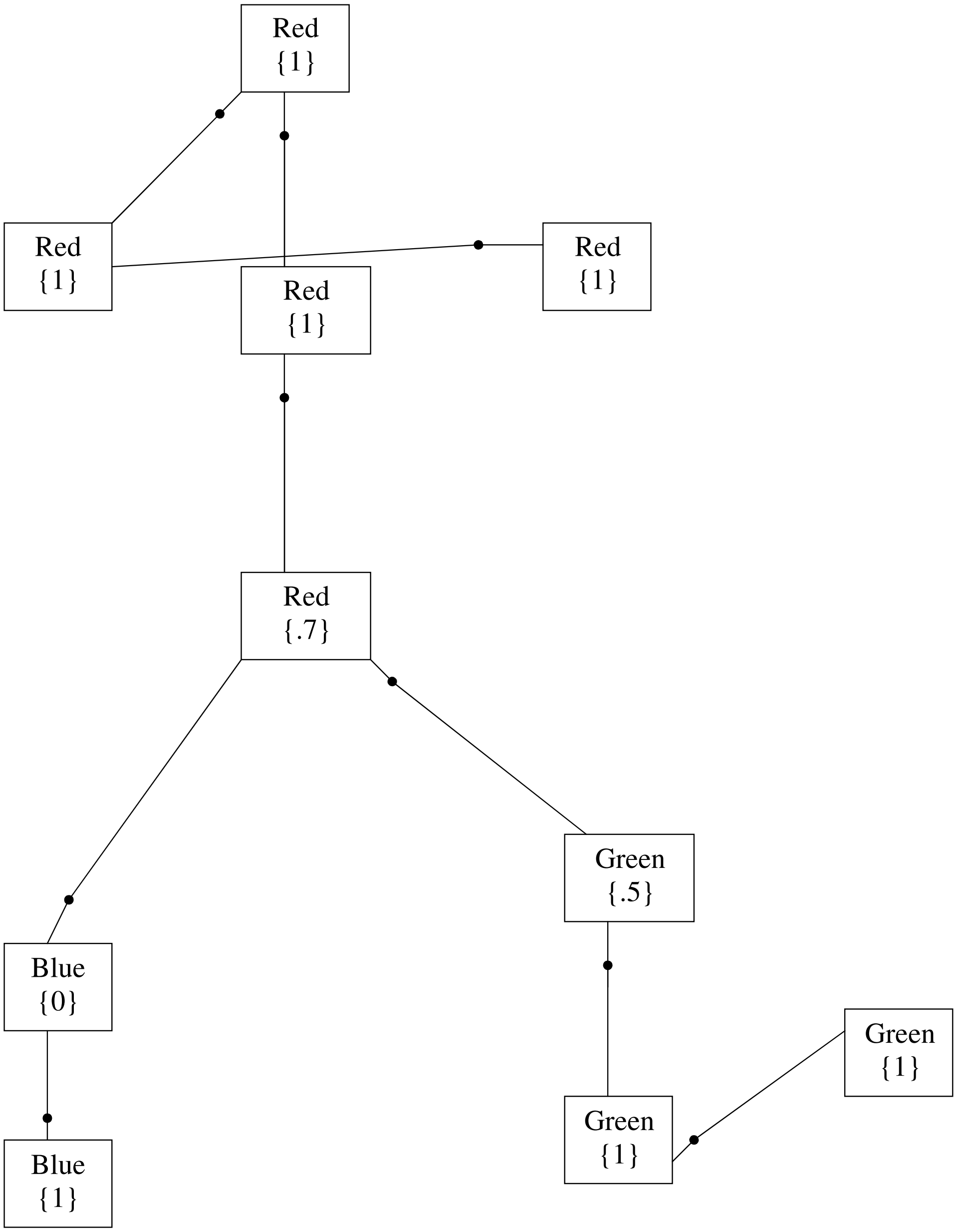

Figure 3: Connected Nash

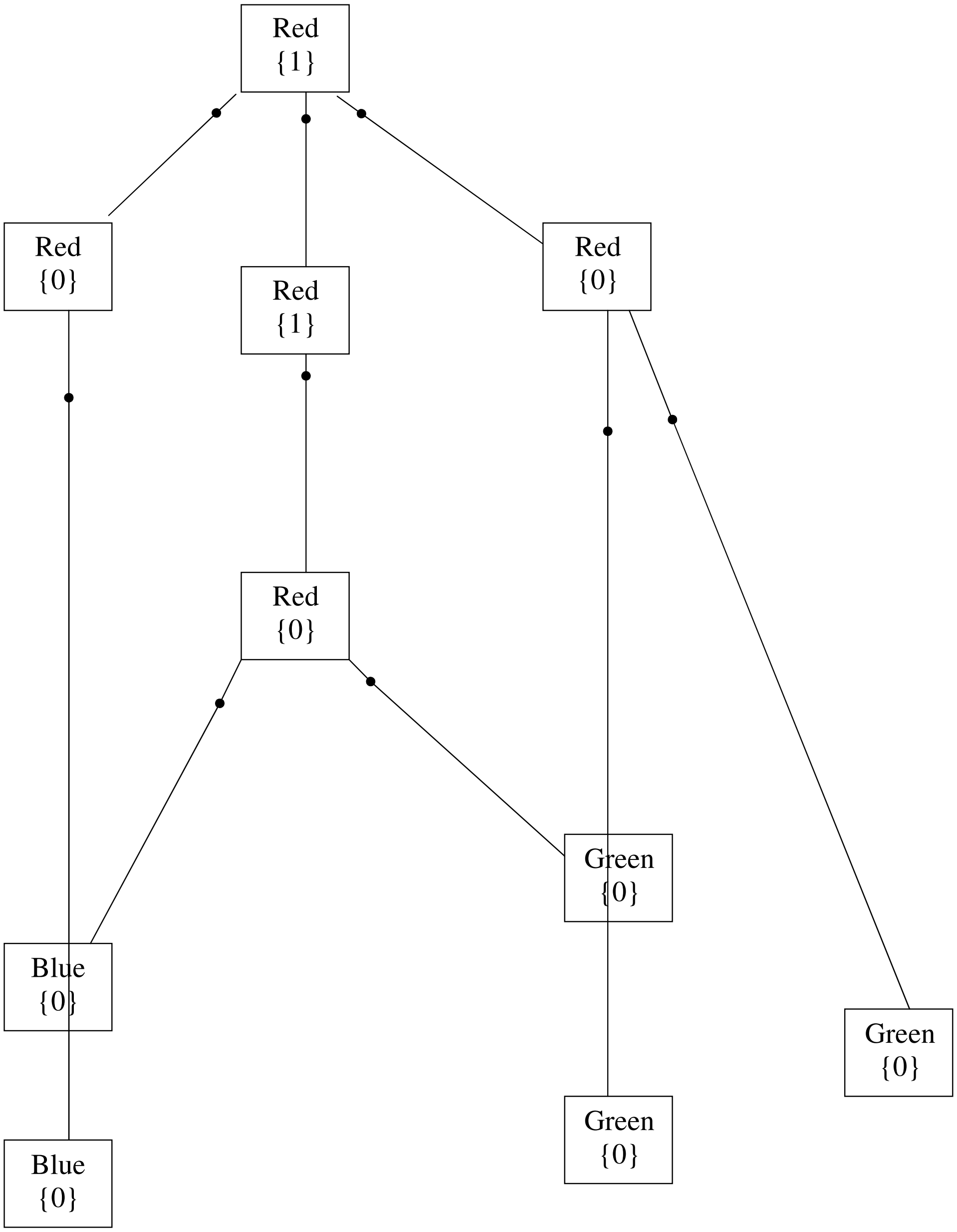

Figure 4: Strict Nash, Generalised Center Sponsored Star

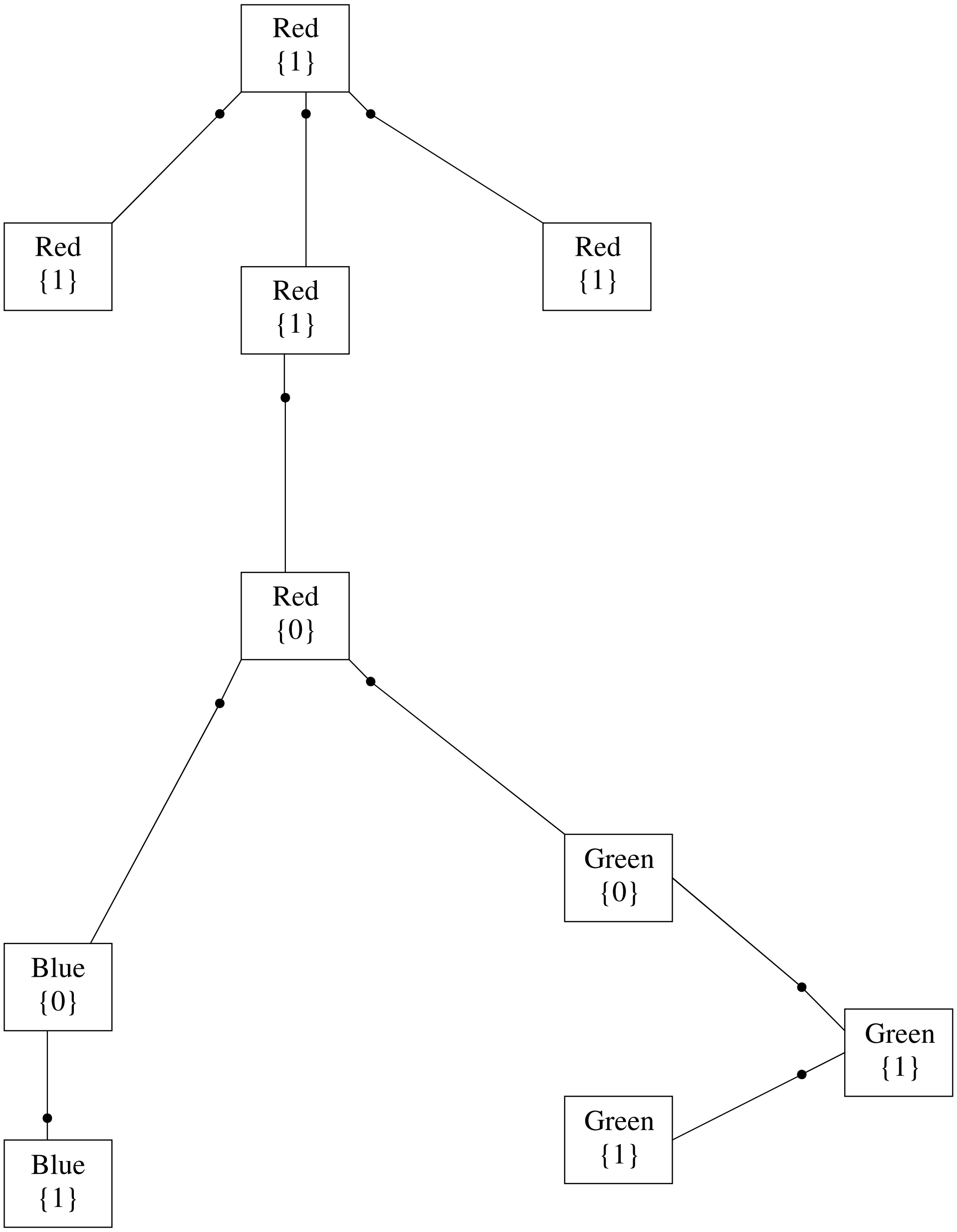

Figure 5: Strict Nash, Electron Star

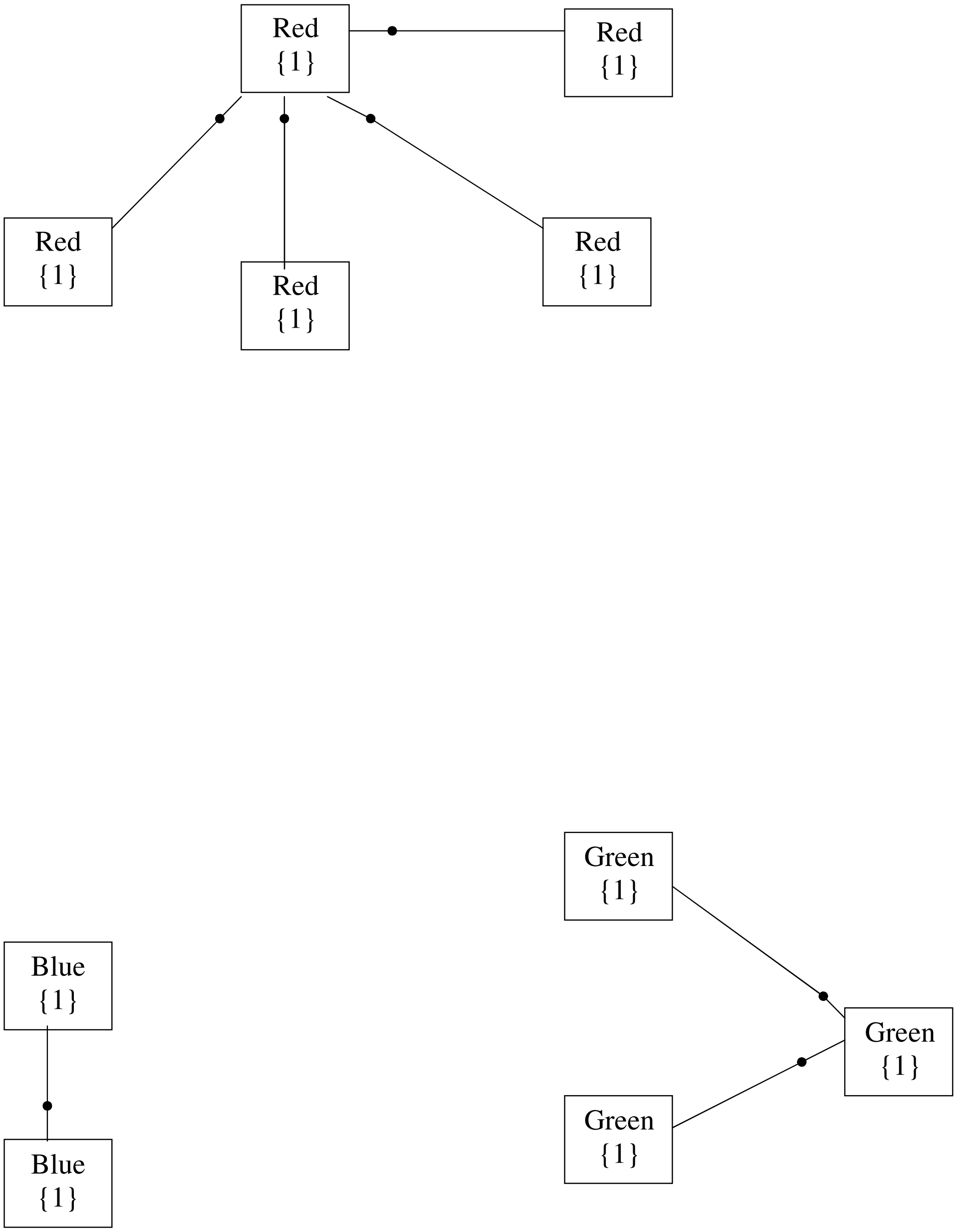

Figure 6: Strict Nash, Unconnected Center-Sponsored Stars

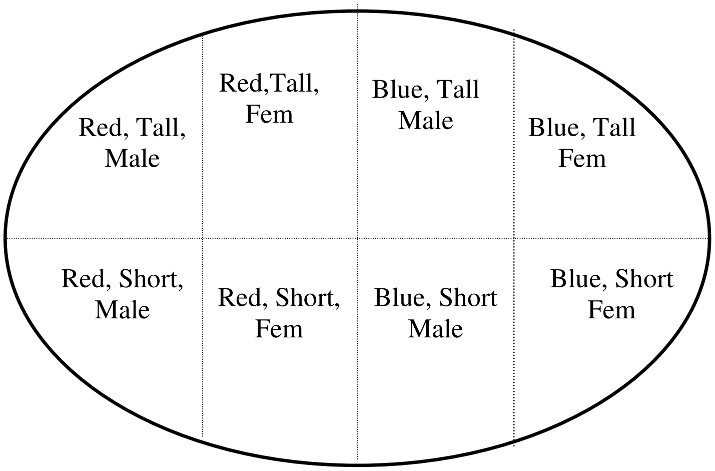

Figure 7: All possible blocks under the three dimensions of {Color, Height, Sex}

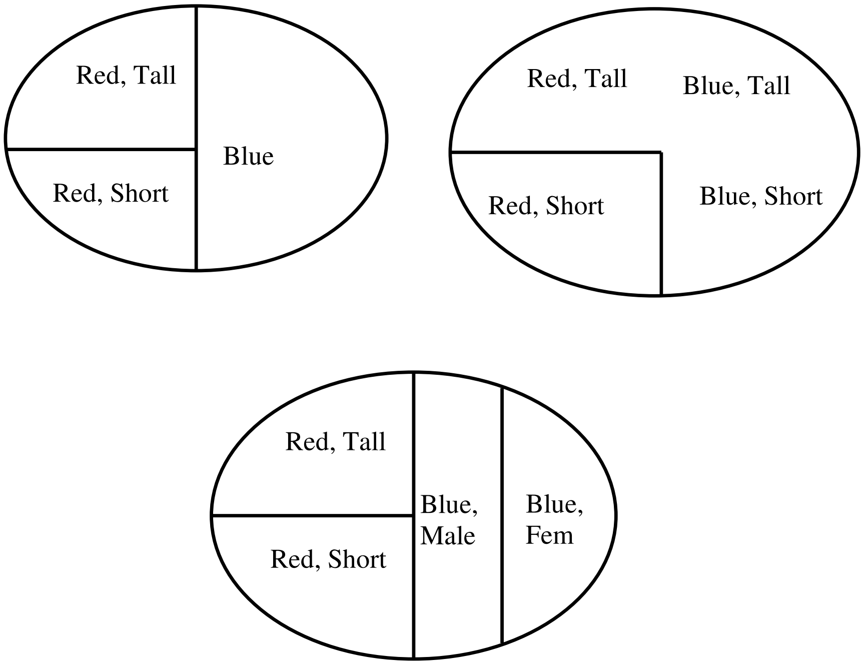

Figure 8: Three possible Nash Network components(Components are delineated by strong lines)

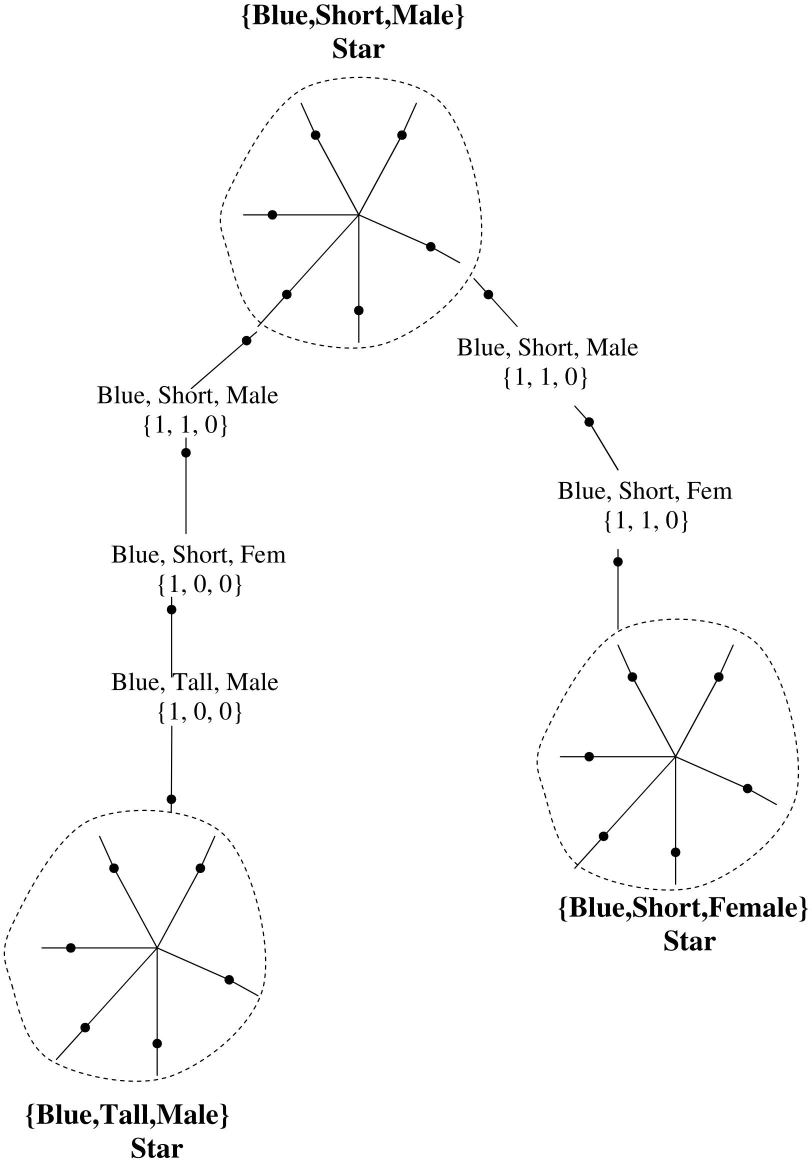

Figure 9: An Interconnected Tail Star

D Community Structure Graphs

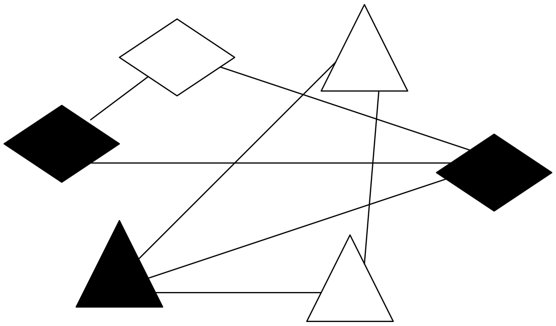

Figure 10: Raw Network Data

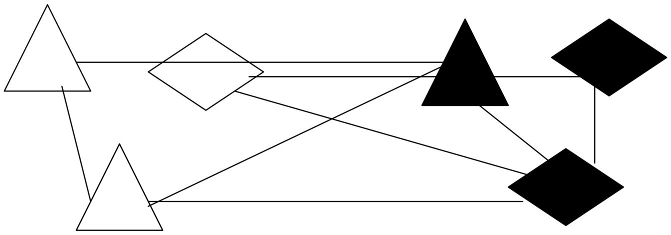

Figure 11: Sorting by Color

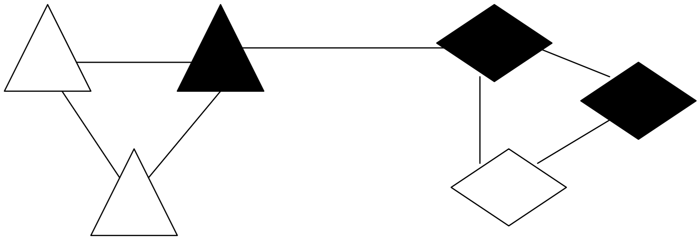

Figure 12: Sorting by Shape

E Community Structure Results

E.1 Community Structure of Village 3 and 4

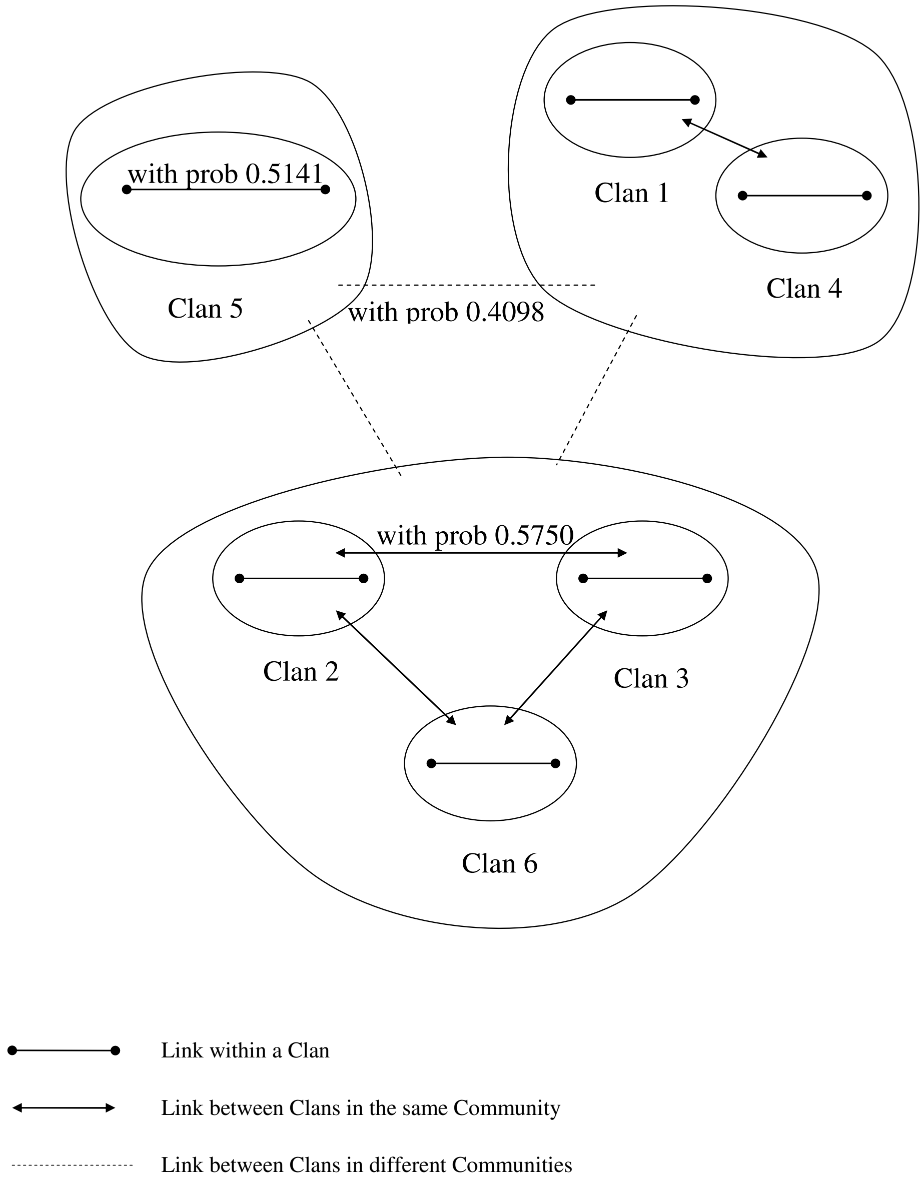

Figure 13: Village 3, Depth 1: Clan

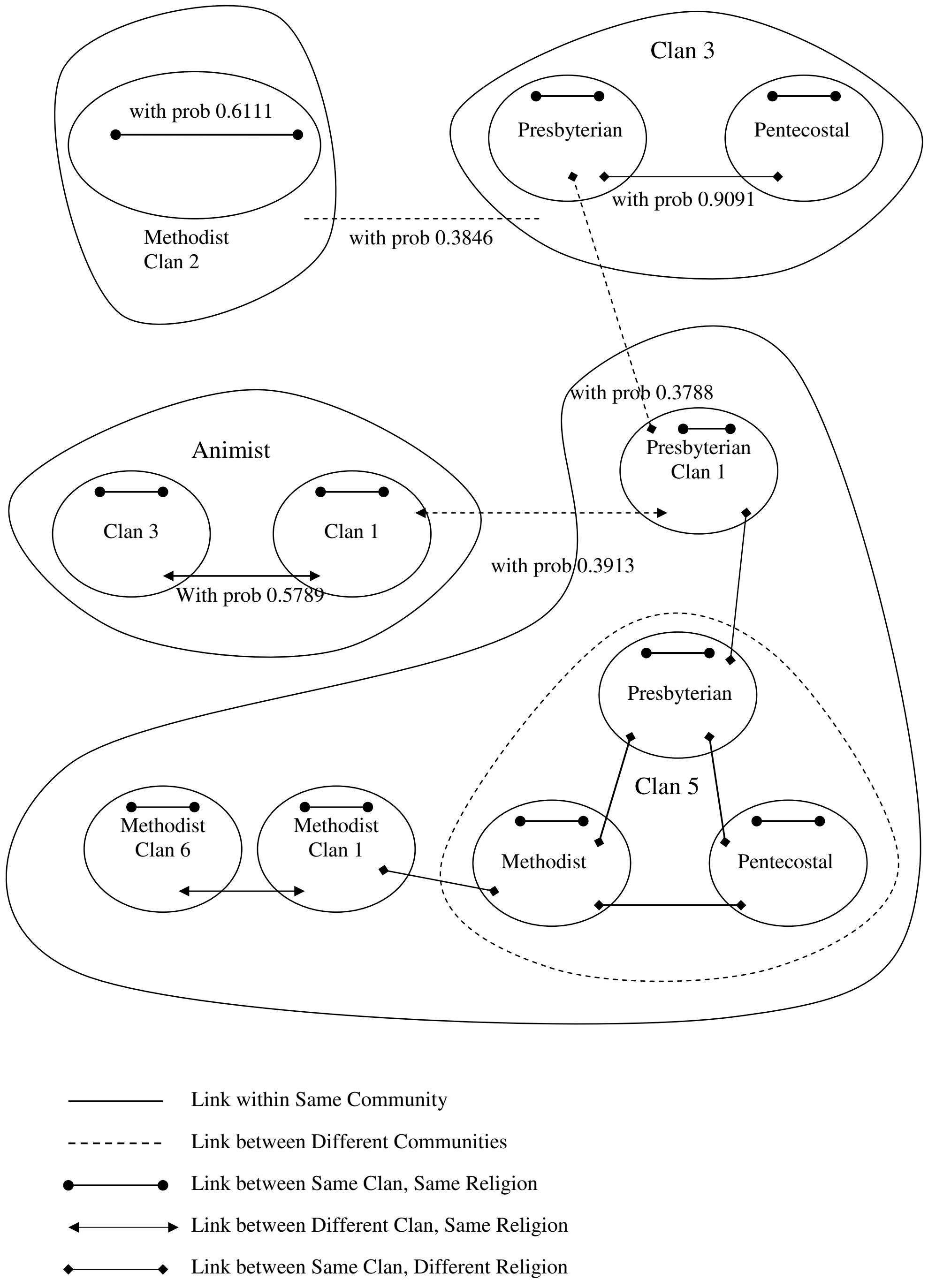

Figure 14: Village 3, Depth 2: Religion and Clan

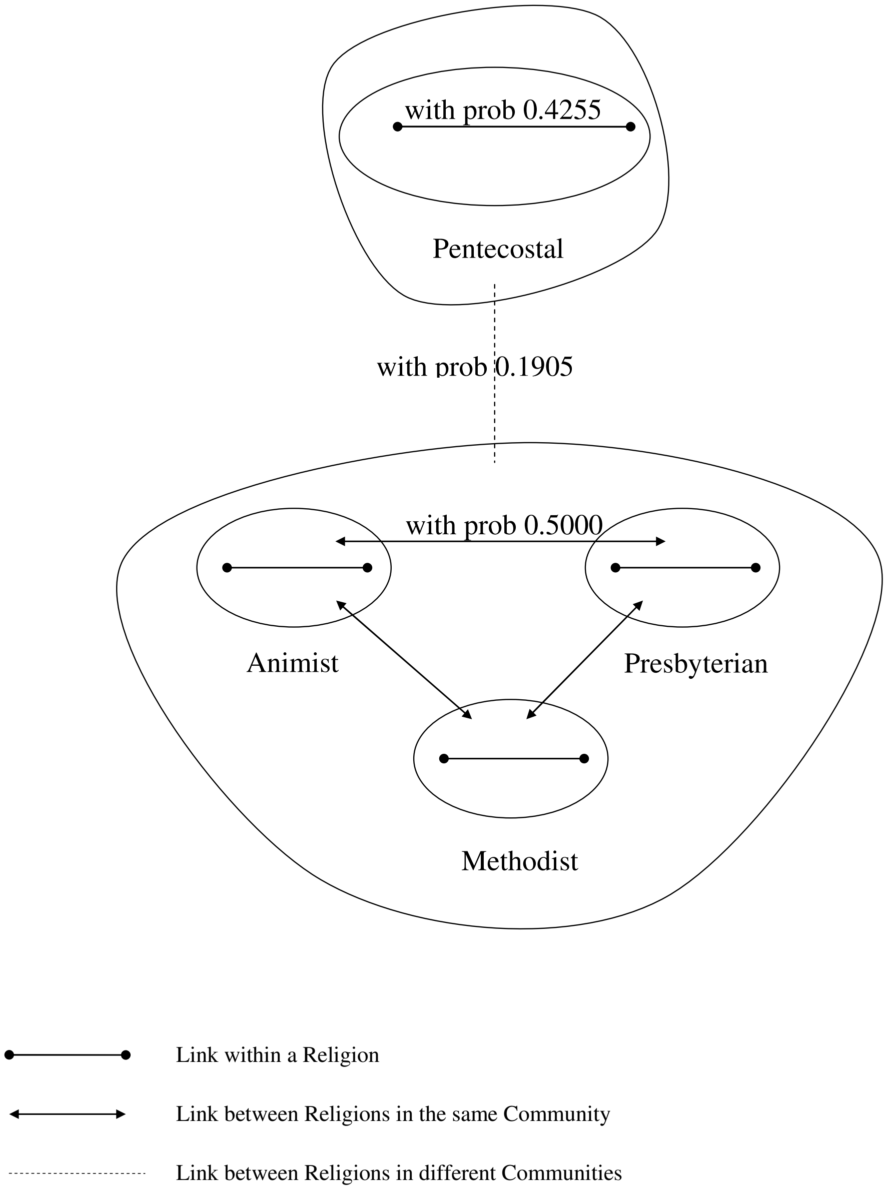

Figure 15: Village 4, Depth 1: Religion

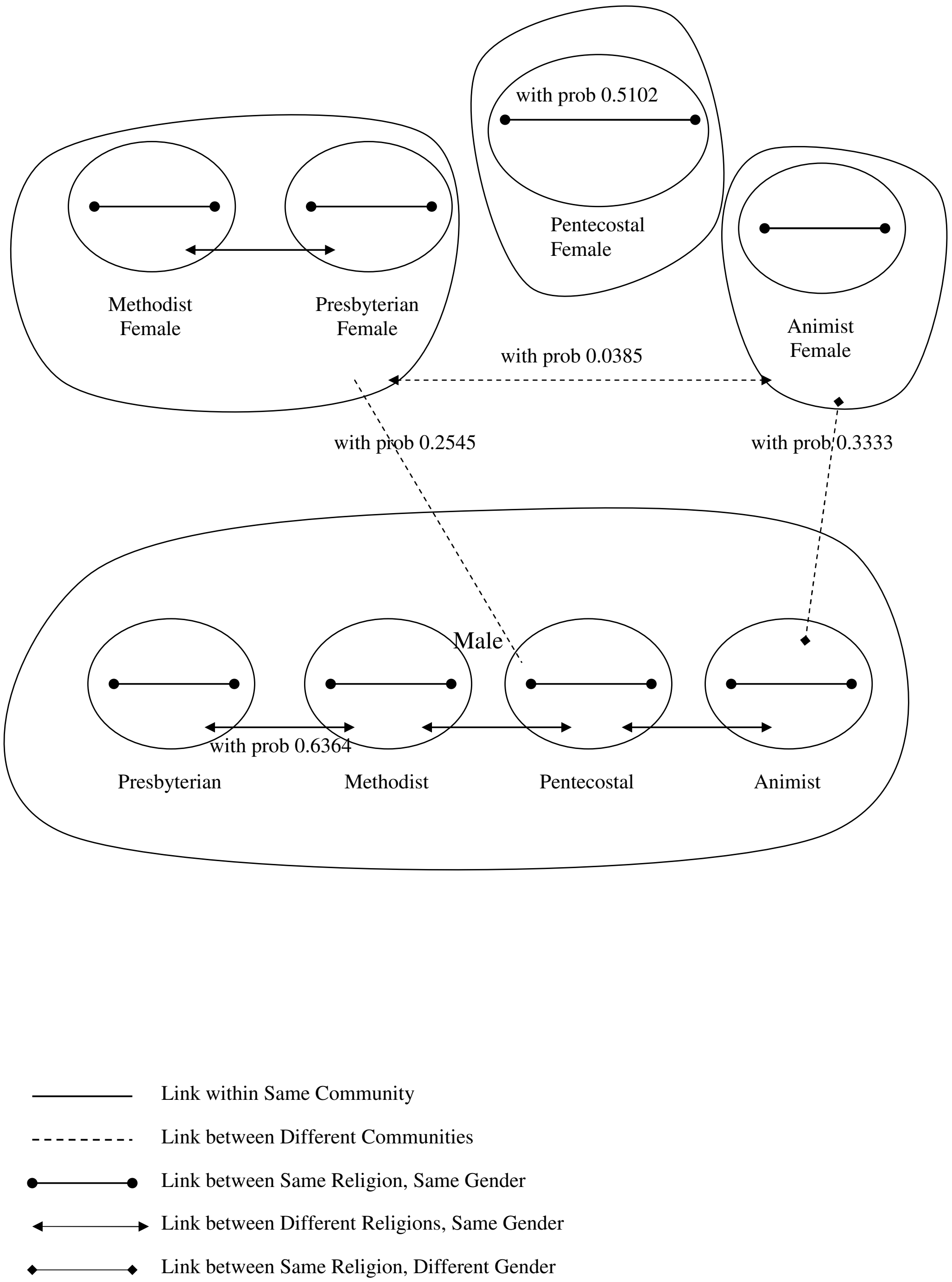

Figure 16: Village 3, Depth 2: Religion and Gender

E.2 Tables

Table 1: Variables Measuring Presence of Link

Variable	Definition	Mean (Std Dev)
Askprob	0-1 variable taking value 1 if respondent would ask match if they had a problem with unhealthy crop	0.358848 (0.47986)
Askfert	0-1 variable taking value 1 if respondent would go to match for advice on new fertilizer	0.334156 (0.471889)
Askplant	0-1 variable taking value 1 if respondent would go to match to discuss planting method	0.330041 (0.470421)
Askbuyer	0-1 variable taking value 1 if respondent would go to match for finfinf a buyer	0.316049 (0.465124)

Table 2: Correlation amongst Link Variables

	Askporb	Askfert	Askplant	Askbuyer
Askprob	1			
Askfert	0.8305	1		
Askplant	0.8834	0.8906	1	
hline Askbuyer	0.6651	0.6631	0.7088	1

Table 3: Summary Statistics of Identity Variables

Variable	Definition	Mean (Std. Dev.)
Firsthere	variable taking value 1 if respondent is not the first of their family in the village and 2 o.w.	1.174468 (0.380322)
Resprel	religions Presbyterian, Methodist, Pentacostal and Animist/Traditional are codede as 1,2,3 and 4 resp.	2.461847 (1.054727)
Pineyes	variable taking value 1 if not a pineapple grower and 2 otherwise	1.440329 (0.497451)
Clan	6 clans are coded as numbers 1-6	3.26506 (1.832179)
Gender	variable taking value 1 if respondent is female, 2 if male	1.420814 (0.494811)

Table 4: Summary Statistics of Variables Measuring Similarity

Variable	Definition	Mean (Std. Dev.)
SGender	0-1 variable taking value 1 if respondent and match have the same sex and 0 o.w	0.491282 (0.500181)
SClan	0-1 variable taking value 1 if respondent and match have the same clan and 0 o.w.	0.385185 (0.486839)
SFirsthere	0-1 variable taking value 1 if either respondent and match were both first from their families in the village, or both not the first in the village and 0 o.w.	0.749541 (0.433476)
SResprel	0-1 variable taking value 1 if respondent and match have the same religion and 0	0.395062 (0.489065)
SPineyes	0-1 variable taking value 1 if either respondent and match both have experience in pineapple, or if both don't have experience in pineapple and 0 o.w.	0.514555 (0.500002)

Table 5: Correlations between the Link and Similarity variables

	Askprob	Askfert	Askplant	Askbuyer
SGender	0.0112	-0.0028	0.0014	0.0253
SClan	0.0878	0.0858	0.0761	0.1201
SFirsthere	-0.0428	-0.0457	-0.0464	-0.0277
SResprel	0.0617	0.0291	0.0193	-0.0307
SPineyes	0.0103	-0.0054	-0.0446	-0.0186

Table 6: Maximised Likelihood at Depth 1

Dimension	Village 1 (obs = 200)	Village 2 (obs = 82)	Village 3 (obs = 339)	Village 4 (obs = 219)
Firsthere	-Inf	-Inf	-Inf	-145.234
Religion	-106.37	-51.4337	-232.831	**-137.76****
Gender	-Inf	-50.8523	-234.798	-142.191
Clan	**-103.064***	-Inf	**-232.46**	-143.232
Pineyes	-Inf	**-50.8418**	-Inf	-Inf

The maximized likelihood for depth 1 for each village is highlighted
'-Inf' denotes there was no feasible community structure along that dimension
* significant at 5%; ** significant at 1% (from no partition)

Table 7: Maximised Likelihood at Depth 2

Dimension	Village 1 (Depth 1: Clan)	Village 2 (Depth 1: Pineyes)	Village 3 (Depth 1: Clan)	Village 4 (Depth 1: Religion)
Firsthere	-99.5182	-Inf	-228.753	-Inf
Religion	**-96.6904****	**-42.5427***	**-223.328****	-
Gender	-Inf	-46.4117	-Inf	**-126.877****
Clan	-	-44.7001	-	-132.56
Pineyes	-99.274	-	-Inf	-135.845

The maximized likelihood for depth 1 for each village is highlighted
'-Inf' denotes there was no feasible community structure along that dimension
* significant at 5%; ** significant at 1% (from no partition)

References

Akerlof, G. A., and R. E. Kranton (2000): "Economics and Identity," *The Quarterly Journal of Economics*, 115(3), 715-753.

Bala, V., and S. Goyal (2000a): "A Noncooperative Model of Network Formation," *Econometrica*, 68(5), 1181–1229.

——— (2000b): "A Strategic Analysis of Network Reliability," *Review of Economic Design*, 5(3), 205–228.

Bandiera, O., and I. Rasul (2002): "Social Networks and Technology Adoption in Northern Mozambique," *CEPR Discussion Paper No. 3341.*

Bisin, A., and T. Verdier (2000): ""Beyond the Melting Pot": Cultural Transmission, Marriage, and the Evolution of Ethnic and Religious Traits," *The Quarterly Journal of Economics*, 115(3), 955–988.

Bramoulle, Y., and R. Kranton (2007): "Public Goods in Networks," *Journal of Economic Theory*, 127(1), 478–494.

Brubaker, R., and F. Cooper (2000): "Beyond "Identity"," *Theory and Society*, 29(1), 1–47.

Conley, T. G., and C. R. Udry (2004): "Social Networks in Ghana," .

——— (2005): "Learning About a New Technology: Pineapple in Ghana," *Proceedings.*

Copic, J., M. O. Jackson, and A. Kirman (2009): "Identifying Community Structures from Network Data via Maximum Likelihood Methods," *The B.E. Journal of Theoretical Economics*, 9(1), Article 30.

De Weerdt, J. (2004): "Risk-Sharing and Endogenous Network Formation," in *Insurance against poverty,*, ed. by S. Dercon, chap. 10, pp. 197–216. Oxford University Press, Oxford.

De Weerdt, J., and S. Dercon (2006): "Risk-Sharing Networks and Insurance Against Illness," *Journal of Development Economics*, 81(2), 337–356.

Deroan, F. (2003): "Farsighted Strategies in the Formation of a Communication Network," *Economics Letters*, 80(3), 343–349.

Dev, P. (2009): "Choosing 'Me' and 'My Friends': Identity in a Non-Cooperative Network Formation Game with Cost Sharing," Mimeo, ITAM.

Dutta, B., and M. O. Jackson (2003): *Networks and Groups: Models of Strategic Formation.* Springer-Verlag.

Esteban, J., and D. Ray (1994): "On the Measurement of Polarization," *Econometrica*, 62(4), 819–51.

Fafchamps, M. (2002): "Returns to Social Network Capital Among Traders," *Oxford Economic Papers*, 54(2), 173–206.

Fafchamps, M., and F. Gubert (2007): "The Formation of Risk Sharing Networks," *Journal of Development Economics*, 83(2), 326–350.

Fafchamps, M., and S. Lund (2003): "Risk-Sharing Networks in Rural Philippines," *Journal of Development Economics*, 71(2), 261–287.

Fafchamps, M., M. J. van der Leij, and S. Goyal (2006): "Scientific Networks and Co-authorship," Discussion paper.

Feri, F. (2004): "Stochastic Stability in Networks with Decay," Mimeo. University of Venice.

Foster, A. D., and M. R. Rosenzweig (1995): "Learning by Doing and Learning from Others: Human Capital and Technical Change in Agriculture," *Journal of Political Economy*, 103(6), 1176–1209.

Fryer, R. G., and M. O. Jackson (2002): "Categorical Cognition: A Psychological Model of Categories and Identification in Decision Making," Microeconomics 0211002, EconWPA.

Galeotti, A. (2006): "One-way Flow Networks: The Role of Heterogeneity," *Economic Theory*, 29(1), 163–179.

Galeotti, A., S. Goyal, and J. Kamphorst (2005): "Network Formation with Heterogenous Players," *Games and Economic Behavior*, 54(2), 353–372.

Gilles, R. P., and C. Johnson (2000): "Spatial Social Networks," *Review of Economic Design*, 5(3), 273–299.

Goldstein, M., and C. Udry (1999): “Agricultural Innovation and Resource Management in Ghana,” *Final Report to IFPRI under MP17.*

Goyal, S., and S. Joshi (2003): “Networks of Collaboration in Oligopoly,” *Games and Economic Behavior*, 43(1), 57–85.

Goyal, S., and F. Vega-Redondo (2005): “Network Formation and Social Coordination,” *Games and Economic Behavior*, 50(2), 178–207.

Granovetter, M. (2005): “The Impact of Social Structure on Economic Outcomes,” *Journal of Economic Perspectives*, 19(1), 33–50.

Grimard, F. (1997): “Household Consumption Smoothing Through Ethnic Ties: Evidence from Cote d’Ivoire,” *Journal of Development Economics*, 53(2), 391–422.

Hojman, D. A., and A. Szeidl (2008): “Core and Periphery in Networks,” *Journal of Economic Theory*, 139(1), 295–309.

Jackson, M. O. (2006): “The Economics of Social Networks,” in *Advances in Economics and Econometrics, Theory and Applications: Ninth World Congress of the Econometric Society*, ed. by R. Blundell, W. Newey, and T. Persson, vol. I, chap. 1. Cambridge University Press.

Jackson, M. O., and B. Dutta (2000): “The Stability and Efficiency of Directed Communication Networks,” *Review of Economic Design*, 5(3), 251–272.

Jackson, M. O., and A. Wolinsky (1996): “A Strategic Model of Social and Economic Networks,” *Journal of Economic Theory*, 71(1), 44–74.

Kranton, R. E., and D. F. Minehart (2001): “A Theory of Buyer-Seller Networks,” *American Economic Review*, 91(3), 485–508.

Lorrain, F., and H. White (1971): “Structural Equivalence of Individuals in Social Networks Blockstructures with Covariates,” *Journal of Mathematical Sociology*, 1, 49–80.

McBride, M. (2006): “Imperfect Monitoring in Communication Networks,” *Journal of Economic Theory*, 126(1), 97–119.

Mead, G. H. (1934): *Mind, Self, and Society.* University of Chicago Press, Chicago.

Munshi, K., and M. Rosenzweig (2006): "Traditional Institutions Meet the Modern World: Caste, Gender, and Schooling Choice in a Globalizing Economy," *American Economic Review*, 96(4), 1225–1252.

Newman, M. E. J. (2004): "Detecting Community Structure in Networks," *The European Physical Journal B - Condensed Matter and Complex Systems*, 38(2), 321–330.

Nowicki, K., and T. A. B. Snijders (2001): "Estimation and Prediction for Stochastic Blockstructures," *Journal of the American Statistical Association*, 96(455), 1077–1087.

Page Jr., F. H., and M. Wooders (2009): "Strategic basins of attraction, the path dominance core, and network formation games," *Games and Economic Behavior*, 66(1), 462 – 487.

Patacchini, E., and Y. Zenou (2008): "Ethnic Networks and Employment Outcomes," IZA Discussion Papers 3331, Institute for the Study of Labor (IZA).

Santos, P., and C. Barrett (2004): "Interest And Identity In Network Formation," Discussion paper.

Sarangi, S., P. Billand, and C. Bravard (2006): "Heterogeneity in Nash Networks," *Departmental Working Papers, Department of Economics,Louisiana State University.*

Schelling, T. (1971): "Dynamic Models of Segregation," *Journal of Mathematical Sociology*, 1, 143–186.

Sen, A. (2006): *Identity and Violence: The Illusion of Destiny.* W. W. Norton.

Sergio Currarini, M. O. J., and P. Pin (2008): "An Economic Model of Friendship: Homophily, Minorities and Segregation," *Econometrica*, 77(4), 1003–1045.

Slikker, M., and A. van den Nouweland (2001): "A One-Stage Model of Link Formation and Payoff Division," *Games and Economic Behavior*, 34(1), 153–175.

Stryker, S. (1968): "Identity Salience and Role Performance: The Relevance of Symbolic Interaction Theory for Family Research," *Journal of Marriage and the Family*, 30(4), 558–64.

Stryker, S., and P. J. Burke (2000): "The Past, Present, and Future of an Identity Theory," *Social Psychology Quarterly*, 63(4), 284–297.

Tajfel, H., and J. C. Turner (1979): "An Integrative Theory of Intergroup Conflict," in *The Social Psychology of Intergroup Relations*, ed. by W. G. Austin, and S. Worchel, pp. 33–47. Monterey, Calif.: Brooks/Cole Pub. Co.

Tallberg, C. (2005): "A Bayesian Approach to Modeling Stochastic Blockstructures with Covariates," *Journal of Mathematical Sociology*, 29, 1–23.

Townsend, R. M. (1994): "Risk and Insurance in Village India," *Econometrica*, 62(3), 539–591.

Waters, M. C. (1999): *Black Identities: West Indian Immigrant Dreams and American Realities*. The Russell Sage Foundation and Harvard University Press.

Watts, A. (2001): "A Dynamic Model of Network Formation," *Games and Economic Behavior*, 34(2), 331–341.

White, H. C., S. A. Boorman, and R. L. Breiger (1976): "Social-Structure from Multiple Networks I: Blockmodels of Roles and Positions," *American Journal of Sociology*, 81, 730–780.