Small world networks with segregation patterns and brokers

Edoardo Gallo*

February 2009

Abstract

Many social networks have the following properties: *(i)* a short average distance between any two individuals; *(ii)* a high clustering coefficient; *(iii)* segregation patterns; the presence of *(iv)* brokers and *(v)* hubs. *(i)* and *(ii)* define a *small world* network. This paper develops a strategic network formation model where agents have heterogeneous knowledge of the network: cognizant agents know the whole network, while ignorant ones are less knowledgeable. For a broad range of parameters, all *pairwise Nash* (PN) networks have properties *(i)-(iv)*. There are some PN networks with one hub. Cognizant agents have higher betweenness centrality: they are the *brokers* who connect different parts of the network. Ignorant agents cause the emergence of *segregation patterns.* The results are robust to varying the number of cognizant agents and to increasing the knowledge level of ignorant ones. An application shows the relevance of the results to assessing the welfare impact of an increase in network knowledge due to, e.g., improved access to social networking tools.

Keywords: network, cognitive network, small world, broker, segregation.
JEL: C72, D85.

*Address: University of Oxford, Nuffield College, New Road, Oxford OX1 1NF, UK. Email: *edoardo.gallo@economics.ox.ac.uk*. I would like to thank Meg Meyer for her invaluable help and guidance throughout this project. Thanks to Misha Drugov, Aytek Erdil, Andrea Galeotti, Sanjeev Goyal, Daniel Hojman, Rahmi Ilkiliç, Zia Khan, Philippos Louis, Rocco Macchiavello, Marco Marinucci, Andrea Patacconi, Paolo Pin and Peyton Young for helpful comments and suggestions. I am also grateful to seminar/conference audiences at the University of Oxford, DMM 2008 (University of Montpellier), the 2008 Summer Meeting of the European Economic Association (Bocconi University) and the 2009 Coalition Theory Network Workshop (Maastricht University). Any remaining errors are my own.

1 Introduction

The structure of social networks is an important determinant of economic outcomes in a wide spectrum of settings including labor markets, provision of informal insurance, the generation and spread of innovations, disease epidemics, organizational performance and financial markets. In order to harness benefits from the network, individuals strategically form and break connections to acquire an advantageous position in the social structure. Understanding how these strategic choices shape the emergence of social networks is of fundamental importance to explain the main structural properties of networks observed in empirical studies and to investigate the economic implications of these networks for single individuals and the society as a whole.

A critical characteristic of social networks is their complexity and the consequent difficulty individuals face in building a correct knowledge of the intricate pattern of social relations. For instance, a group of just 10 identical individuals can form $11,716,571$ different connected network architectures. The incompleteness and heterogeneity in individuals' knowledge of their social network may play an important role in determining which network architectures emerge in equilibrium. This is particularly relevant in recent years because the introduction of social networking websites/tools allows some individuals to significantly increase their knowledge of the social structure they are embedded in.

Empirical and experimental studies in the sociology and psychology literatures have investigated the accuracy of an individual's *cognitive network*, i.e. her perception of the surrounding social network. Kumbasar et al. [1994] map the cognitive friendship networks of the members of a professional group: they show that individuals' perception of the real network is inaccurate and displays systematic and heterogeneous biases. Janicick and Larrick [2005] confirm some of these findings in an experimental setting. Moreover, Krackhardt [1990] shows that accuracy in perception of a social network gives concrete advantages: the accuracy of an individual's perception of the advice and friendship networks in a 34-person organization is positively correlated with her influence in the organization as ranked by her colleagues.[1]

In the last decade, studies in several disciplines, including sociology, physics, computer science and economics, have collected large amounts of evidence that the majority of social networks share common structural properties. Girvan and Newman [2002] highlight five main properties:

(i) Short average distance: the average distance between any two individuals in the network is small

(ii) High clustering coefficient: a high proportion of individuals with a common connection are also connected to each other

[1] Several other studies confirm and extend these findings. Casciaro [1998] and Bondonio [1998] show that cognitive network accuracy is correlated with personality traits, demographic factors, hierarchical status, and the individual's position in the informal social structure. Other studies include Krackhardt [1987], Freeman [1992] and Krackhardt and Kilduff [1999].

(iii) *Segregation patterns:* individuals tend to gather into richly connected, close-knit communities with few links across communities

(iv) *Presence of brokers:* there is a small number of individuals, called brokers, who connect across communities

(v) *Presence of hubs:* there is a small number of individuals, called hubs, with a very high number of connections

A network with properties *(i)* and *(ii)* is called a small world network. Milgram [1967] was the first to investigate *(i)* with a famous experiment where, by using a letter chain, he showed that on average it took 5.2 intermediaries to connect two randomly chosen individuals in Nebraska and Boston. Recent studies on large networks have confirmed the ubiquity of small world networks.[2]

Burt [1992] presents extensive empirical evidence of *(iii)* and *(iv)*. He shows that brokers receive higher benefits from their connections because they are able to access non-redundant information and control information flows in the network.[3] The seminal paper by Barabási and Albert [1999] was among the first to point out *(v)* in many human and non-human networks. In the ensuing years the number of studies finding power law or similar distributions of connections for a variety of networks increased at a dramatic pace.[4]

This paper develops a model of strategic network formation to investigate the importance of individuals' incomplete and heterogeneous knowledge of the network for the emergence of structural properties *(i)-(v)*. The model is as follows. There are n agents exogenously partitioned into k communities. The cost of connecting for a pair of agents belonging to different communities is constant, and greater than the constant cost of connecting for a pair of agents in the same community. There are two types of agents: network cognizant (NC) agents, with complete knowledge about the network, and network ignorant (NI) agents, with partial knowledge about the network. Specifically, NI agents are only able to see connections involving at least one agent from their own community, and they know the distribution of connections of a sample of the agents in other communities.

The network formation process is modeled as a one-shot game which was first informally defined in Myerson [1991]: the agents independently announce the links they wish to form and links are only formed under mutual consent. Agents derive their utility from

[2]Goyal et al. [2006] analyze the co-authorship network in economics journals: they find that it is a small world and that the average distance between individuals has decreased over time. Similarly, Newman [2001] finds that co-authorship networks in the physical sciences are also small world networks. Kossinets and Watts [2006] analyze the dynamic evolution of a large email network at an American university. One of the findings is that the network is a small world, and that this is persistent over time.

[3]Burt [2004] maps the idea generation network of the supply chain of a large electronics company with 673 managers, and shows that managers who are brokers also have better job evaluations, faster career tracks, higher bonuses, and they are more likely to have good ideas. Granovetter [1995] shows that white-collar workers who are brokers are more likely to find a better job, and to find it faster.

[4]See Newman et al. [2006] for a comprehensive review.

the network, and the payoff structure is the *distance-based* utility introduced by Bloch and Jackson [2006a]: links are costly, but direct and indirect connections bring benefits that decay with distance in the social network. The new component of the payoff structure introduced here is that individuals derive benefits from their cognitive network, which can differ from the real network structure. This change breaks the symmetry built into the payoffs of previous models, and it allows the emergence of richer network structures in equilibrium.

For a broad range of the parameters, the unique pairwise Nash networks constitute a family of network architectures that are structurally richer and more realistic than equilibrium networks previously characterized in the strategic network formation literature. Figure 3 illustrates graphically three examples of these networks, which have two salient features. First, they have a *kernel*: a densely connected part of the network formed by the subset of NC agents who connect across communities. Second, in all these networks the NI agents connect exclusively with other agents in their own community. The communities are either complete networks, if the cost of links within the community is low, or $d2$ networks if the cost is high.[5] The proof involves two key steps: proving the claim that NI agents only form links within their community and showing that the game is isomorphic to a simpler network formation game involving only the NC agents.

All these pairwise Nash networks have properties *(i),(iii)* and *(iv)*, plus property *(ii)* as long as there is a non-negligible number of communities with low cost of link formation. The NC agents are the *brokers* in the network: they strategically position themselves in the social structure to connect across communities thereby *shortening the average distance* between any two individuals in the network. The *high clustering coefficient* is guaranteed by the presence of a non-negligible number of low cost communities. Finally, the NI agents connect exclusively within their own community causing the emergence of *segregation patterns*. For higher costs of link formation across communities, there are also some pairwise Nash networks with a star kernel with properties *(i)-(iv)* and *(v)*: the center of the star kernel is a *hub*. Two extensions of the model show that these results are robust to varying the number of NC agents in each community and to increasing the knowledge that NI agents have about the network.

It is important to point out that *both* the heterogeneity in the cost of forming links *and* the heterogeneity in the knowledge of the network are needed for these architectures to emerge in equilibrium. If only cost heterogeneity is present then non-trivial small world networks exist solely for the limit case where all communities have low cost of link formation.[6] Moreover, an example shows that there are pairwise Nash networks without segregation patterns. If only knowledge heterogeneity is present then non-trivial small world networks do not exist. Moreover, an example shows that there are pairwise Nash

[5] A network belongs to the family of $d2$ network architectures if all agents are at most at distance 2 from each other and there is at least a pair of agents who is not directly connected. See section 2 for a more formal definition.

[6] A "trivial" small world network is a network that satisfies the properties of a small world just because it is "overconnected." An example is the complete network where each agent is directly linked to all the others.

networks without segregation patterns and brokers.

The results of the model qualify Burt's claim that being a broker brings an economic advantage: whether the NC agents, i.e. the brokers, earn higher payoffs than other agents in their community depends on the social structure of the community itself. A broker receives higher payoffs when she belongs to a poorly connected community because this gives her privileged and almost exclusive access to the benefits from the rest of the network. On the other hand, a broker receives lower payoffs when she belongs to a dense community because most of the other members of the community are directly connected with the broker and can free ride on the benefits.

An application of the model provides insight on the welfare impact of an increase in network knowledge due to, for instance, the possibility that some agents gain access to social networking tools. The first finding is that if these social networking tools increase network knowledge then they also (weakly) increase welfare. The second finding is that the same level of welfare attained in the pairwise Nash networks where all agents have access to social networking tools can be attained even if only some agents have access to these tools.

This paper is a contribution to the network formation literature. Models of network formation are of two types: stochastic, mainly developed by physicists, and strategic, mainly developed by economists. Stochastic models of network formation have been very successful in reproducing the structural regularities found in large networks. For instance, the seminal paper by Watts and Strogatz [1998] shows that small world networks emerge for a broad set of parameters starting from a regular lattice and rewiring a few links with a probability p. A large literature has developed in physics to build richer models to explain the emergence of small world networks and other structural regularities such as power law degree distributions and correlations among nodes' degrees.[7] However, stochastic models have a major drawback: while they are very good at explaining *how* these structural properties emerge, they are silent on the *why*.

Strategic models of network formation, such as the one in this paper, are therefore complementary to the stochastic approach: they explain *why* network structures emerge as a result of the decisions of utility maximizing individuals. A major drawback of most of these models has been that equilibrium networks tend to be very basic structures which are hardly representative of the structural characteristics of real social networks. Even the most complex equilibrium networks found in the literature still have a high degree of structural regularity, e.g. a number of star networks connected with each other. Moreover, essentially all the models in the literature are based on the assumption that all agents have complete and homogeneous knowledge about the network structure.

The paper that comes closest to the results in this work is Jackson and Rogers [2005]. They examine a special case of the model in this paper: a truncated version of the connec-

[7] Barabási and Albert [1999] is the seminal paper on the emergence of power law degree distributions. Dorogovtsev et al. [2000] generalize Barabási and Albert [1999] and find an exact solution to their model. Newman et al. [2006] is a comprehensive review of the work in physics on stochastic models of network formation. Jackson and Rogers [2007] is an example of a stochastic model of network formation in the economic literature.

tions model with heterogeneous costs.[8] The pairwise stable networks in their model are a small subset of the equilibrium networks found in this work, and they have properties *(i)-(iv)*. However, their equilibrium networks are very regular structures: completely connected "islands" with a few links across islands, so the main reason for having properties *(i)-(ii)* is that they are "overconnected." Moreover, their model does not say why any particular agent would be the one building the link across islands, i.e. it is silent on why certain individuals act as brokers.

Galeotti et al. [2006] and Hojman and Szeidl [2008] build models with heterogenous costs of link formation, and they find that interlinked stars is one of the possible equilibria. Interlinked stars have properties *(iii)-(v)*: each star is a community where the agents at the center of each star have a higher number of connections and act as brokers who connect to the center of other stars. However, these equilibrium networks, too, have a high degree of regularity: each community is a star with the link across communities generating from the center of the star. Moreover, interlinked stars lack properties *(i)* and *(ii)* of a small world network. McBride [2006] and McBride [2008] are the only strategic network formation models that investigate the role incomplete knowledge of the network plays in determining equilibrium network outcomes. However, the equilibrium networks that emerge in these papers lack properties *(i)-(v)* commonly found in real social networks.

The first contribution of this paper is to construct a strategic network formation model where, for a broad range of the parameters, all equilibrium networks have a rich structure with properties *(i)-(iv)*. Moreover, property *(v)* is present in some equilibria. This is the first model of strategic network formation that derives equilibrium networks with the above properties *and* where the presence of these properties is not just due to a high degree of structural regularity or an overconnected network. If at least a non-negligible fraction of the communities have low costs of link formation, *small world networks* naturally emerge because there are a few individuals that connect across otherwise separate communities. *Segregation patterns* emerge because the majority of agents connect only within their own community, with the *brokers* providing the only connections across communities.

The second contribution is to show that incompleteness and heterogeneity in individuals' knowledge of the network plays a key role for the emergence of these properties. The agents with complete knowledge of the network are the *brokers* who connect across communities *shortening the distance* between any two individuals in the network. The NI agents connect exclusively within their own community leading to the emergence of *segregation patterns*.

The third contribution is to provide the first economic analysis of the welfare impact of an increase in agents' knowledge of a network. This is particularly relevant because of the development in recent years of a variety of social networking websites, tools and applications that allow those individuals with access to them to increase their knowledge of the social environment in which they are embedded. An increase in knowledge of the network has a (weakly) positive welfare impact, and it suffices that some individuals in a community have complete knowledge of the network to match the welfare level of

[8] "Truncated" means that indirect benefits are cumulated only up to a distance D.

a community where all individuals are knowledgeable. The latter finding offers some guidance on how to use information from social network analysis to improve organizational performance.

The rest of the paper is organized as follows. Section 2 presents the model. Section 3 characterizes the equilibrium networks. Section 4 shows that the equilibrium networks have properties *(i)-(v)*. Section 5 shows that the main findings are robust to varying the number of agents with complete knowledge and to increasing the knowledge of ignorant agents. Section 6 uses the model to assess the impact of social networking websites and/or applications. Section 7 concludes. Appendix A contains all the proofs.

2 The Model

This section presents the main elements of the model: the network notation and terminology, the payoffs, the network formation game, and the assumptions on the knowledge agents have about the network.

Networks. Consider a set of n agents $N = \{1, ..., n\}$, and partition it into $k \geq 2$ subsets such that $N_M = \{M_1, M_2, ..., M_k\}$, where $M_i = \{1, ..., m_i\}$. Assume that $m_i \geq 3$ to avoid trivial subsets. The term *community* will be used to denote the M_is subsets. A *network* is represented by a symmetric matrix $g \in \{0,1\}^{n \times n}$, with $g_{ij} = 1$ denoting that i and j are connected. The *cognitive network* of an agent i is represented by a symmetric matrix $g^i \in \{0,1\}^{n \times n}$, with $g^i_{jk} = 1$ denoting that agent i *perceives* that j and k are connected. an agent's cognitive network is therefore the agent's perception of the underlying network structure.

The neighborhood of i in g is $L_i(g) = \{j \in N | j \neq i, g_{ij} = 1\}$. A *path* $p_{ij}(g)$ between i and j in a graph g is a set of links $p_{ij}(g) = \{g_{ii_1}, g_{i_1 i_2}, ..., g_{i_p j}\}$ such that $g_{ii_1} = g_{i_1 i_2} = ... = g_{i_p j} = 1$. The *length* of a path is $|p_{ij}(g)|$, if there is no path between i and j then the length is infinite. The *geodesic distance* $d_{ij}(g)$ between i and j in g is the minimum number of links that need to be used along some network path to connect i and j. If there is no such path, then $d_{ij}(g) = \infty$. The diameter $D(g)$ of a network g is the maximum geodesic distance in g. A network is *connected* if there is a path of finite length between any two nodes i and j. A network g is *minimal* if $g - g_{ij}$ for any $i, j \in N$ $(i \neq j)$ is such that $|d_{ij}(g - g_{ij})| = \infty$.

The term *network architecture* refers to the geometric properties of a graph, and permutations of agents do not generate different architectures. There are special network architectures that will frequently arise. The complete network is the network $g^C = \{g | g_{ij} = 1 \ \forall i, j \in N\}$. The empty network is the network $g^\emptyset = \{g | g_{ij} = 0 \ \forall i, j \in N\}$. A star network g^* is a network where, for some agent i, all agents $j \neq i$ are connected to i and there is no other link in the network. Analogously, a community M_i is a *complete community* if all the agents in M_i are directly connected to each other. A community is a *star community* if, for some agent $i \in M_i$, all agents $j \neq i$ $(j \in M_i)$ are connected to i and there is no direct link between any other two agents in M_i. The set $G = \{g | g \subseteq g^C\}$ is the set of all

possible networks.

The shorthand notation g^{INDEX} denotes a family of network architectures with similar structural properties. For instance, $g^{d2} = \{g | d_{ij}(g) \leq 2 \ \forall i, j \in N$ and $\exists i, j \in N$ such that $d_{ij}(g) = 2\}$. In words, $g \in g^{d2}$ if any two nodes in g are at a maximum geodesic distance of two and at least a pair of nodes is not directly connected. Clearly, $g^* \in g^{d2}$ and $g^C \notin g^{d2}$. Analogously, a community M_i is a *d2 community* if, considering only paths involving agents in M_i, any two agents in M_i are at a maximum geodesic distance of two and at least a pair of agents is not directly connected.

Payoffs. Agents derive utility from a network according to the *distance-based* payoff structure first introduced by Bloch and Jackson [2006a]:

$$u_i(g^i) = \sum_{j \neq i} b_i(d_{ij}(g^i)) - c_{ij} \cdot g_{ij} \tag{1}$$

where $b(.)$ is a non-increasing function and c_{ij} is the cost to agent i of linking with j. Agents receive a benefit and pay a cost for direct connections, and they also receive benefits (weakly) decaying with distance from indirect connections. For instance, consider a network formed by three agents i, j, k such that $g_{ij} = g_{jk} = 1$ and $g_{ik} = 0$. Assuming all the agents have complete knowledge of the network, the payoffs are: $u_k(g) = u_i(g) = b(1) - c + b(2)$ and $u_j(g) = b(1) - c$. Clearly, the connections model and the truncated connections model in Jackson and Wolinsky [1996] are special cases of these payoffs.

The only departure from the literature is that the utility for agent i depends on g^i, i.e. i's cognitive network, and not the real network g. Thus, agents with less knowledge about the network will be able to extract less benefits from it. This has an intuitive interpretation. Imagine i and j are connected because they do the same job and they usually communicate on work-related issues. Also, j happens to be connected to k because they were roommates in college, and k's family friend l is a renowned producer on Broadway. However, i does not know that k and l are connected. Agent i would love to go to the première of the much awaited new Broadway show of the season. If she knew that $g_{kl} = 1$ then she would subtly mention it to j, hoping that he would ask k whether his friend l has any spare VIP invitations. However, $g^i_{jk} = 0$ so she would never think of raising the subject with her colleague j, and therefore i will have to miss the première and the VIP party. This short story illustrates that there are *some* benefits from the network that need to be "prompted," and that an individual cannot have access to them unless she knows the network structure.

Knowledge. First of all, a note on terminology. In this paper *cognitive knowledge* of a network should be interpreted as the individual's "mental picture" of the network that she is embedded in. This knowledge is exogenously given and it should be interpreted as an individual's cognitive ability to perceive the pattern of connections surrounding her.

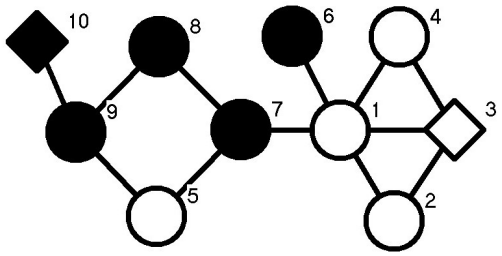
Assume there are two types of agents in the network with complete and partial knowledge about the underlying network structure. Specifically, the types of agents are:

(i) *Network Cognizant (NC)*: an NC agent i has complete knowledge about all the nodes and links in the network g, i.e. $g^i \equiv g$. Let the *kernel* $K \subset N$ be the subset of NC agents in the network.

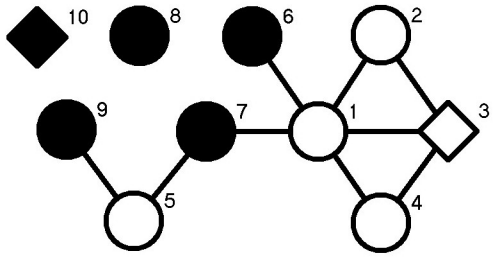
(ii*) *Simple Network Ignorant (NI)*: an NI agent $i \in M_p$ has complete knowledge about any link g_{jk} where $j \in M_p$ and/or $k \in M_p$, but she is not able to see any link g_{jk} such that $j, k \notin M_p$. Moreover, she assumes that any link that she is not able to see does not actually exist.

Clearly *(ii*)* assumes an extreme form of network ignorance and bounded rationality: the NI agent is unable to perceive any intra-community link in other communities and she assumes that agents that are not connected to anyone in her own community are social isolates. This extreme case of network ignorance is for expositional purposes only. An extension in section 5.2 analyzes the same game with more "sophisticated" NI agents who, in addition to the above, know the degree distribution of a sample of agents in each of the other communities. A parameter regulates the size of the sample, allowing the investigation of increases/decreases in the knowledge of the NI agents. However, as section 5.2 shows, the results do not change, so for clarity of exposition the next section investigates the model with the NI agents defined as in *(ii*)*.

Figure 1 illustrates the cognitive networks of NC and NI agents. The network in Figure 1(a) is the real network, which is also the network perceived by the NC agents. The network in Figure 1(b) is the cognitive network g^i of one of the NI agents i belonging to the white community: she knows of the existence of all the connections involving at least one agent from the white community, but she is unable to see the connections involving only agents from the black community. Moreover, she assumes that the connections that she cannot perceive do not exist and therefore she believes that most of the agents in the black community are social isolates.



(a) The real network g, which is also the cognitive network of the diamond-shaped NC agents.



(b) The cognitive network of the circle-shaped NI agents in the white community.

Figure 1: A network $N = \{M_1, M_2\}$ with $|M_1| = |M_2| = 6$. Individuals in M_1 are color-coded in white, and individuals in M_2 are color-coded in black. Round-shaped nodes are NI agents, and diamond-shaped nodes are NC agents.

It is useful to discuss a practical example to see how realistic the assumptions on agents' knowledge are. Imagine Figure 1(a) describes the friendship network in a firm

with two departments: marketing (white nodes) and IT (black nodes). Node 3 is the diamond-shaped NC employee in marketing, and 2 is an NI employee in the marketing community. Both 2 and 3 know who is connected to whom in the marketing department. Both of them also know that 1 and 5, in marketing, are connected to 6, 7 and 9 in IT. Moreover, 3 knows who is connected to whom in IT as shown in 1(a). On the other hand, 2 sees no other connection in the IT community and she thinks that 8 and 10 are anti-social and have no friends.

Network formation game. All agents $i \in N$ simultaneously announce the links they want to form. An agent i's strategy is a vector $s_i \in S_i = \{0,1\}^{n-1}$, where S_i is the set of pure strategies for agent i. The strategy vector $s_i = \{s_{i1}, ..., s_{in-1}\}$ is such that $s_{ij} = 1$ if i wants to form a link with j, and $s_{ij} = 0$ otherwise. An undirected link g_{ij} forms if and only if $s_{ij} \cdot s_{ji} = 1$. A strategy profile $s = (s_1, ..., s_n) \in S$ determines an undirected network $g(s)$, where $S = S_1 \times ... \times S_n$ is the space of pure strategies.

Equilibrium. A strategy profile s is a Nash equilibrium strategy profile of the game if and only if $u_i[g^i(s)] \geq u_i[g^i(s'_i, s_{-i})]$ for all agents $i \in N$ and all strategies $s'_i \in S$. Note that a Nash equilibrium strategy for agent i is defined as a best response by i to the other agents' strategies as they are *perceived* by i. The same observation applies to the stability/equilibrium notions defined below.

The main equilibrium concept used in this paper is pairwise Nash equilibrium:

Definition 1. The network $g(s)$ is a *pairwise Nash equilibrium network* if and only if:

(i) s is a Nash equilibrium strategy, and

(ii) for all $g_{ij} \notin g$, if $u_i(g^i) < u_i(g^i + g_{ij})$ then $u_j(g^j) > u_j(g^j + g_{ij})$

In a pairwise Nash network there is no agent who wants to sever one or more of her links, and there is no pair of agents who both (at least one of them strictly) want to form a new connection.

A network g is *efficient* if the sum of the payoffs of the agents in g is (weakly) higher than the sum of the payoffs the same agents could achieve if they were connected by any other network $g' \neq g$. Formally, let $V(g) = \sum_{i=1}^{n} u_i(g)$. Then g is efficient if and only if $V(g) \geq V(g')$ for all $g' \in G$, $g' \neq g$.

3 Equilibrium Analysis

This section characterizes the pairwise Nash networks of the game. Section 3.1 lists and motivates some simplifying assumptions to the set-up of the model for both expositional and tractability purposes. Section 3.2 characterizes all the pairwise Nash networks for a broad range of the parameters of the model. Section 3.3 investigates the economic implications of the equilibrium networks for the different types of agents.

3.1 Assumptions and motivation

As the literature review in section 1 discusses, the characterization of equilibrium networks for network formation games with heterogeneous agents is not easily tractable without some simplifying assumptions. Moreover, it is important to reduce the heterogeneity in the model to zoom in on the role of heterogeneity in cognitive knowledge of the network. Thus, for tractability and expositional purposes, assume the following:

(a) *Local uniqueness of the NC agent:* in each community M_i there is one and only one NC agent y_i.

(b) *Simple NI agents:* see *(ii*)* in section 2.

(c) *Homogeneous community size:* the partitions M_is have equal cardinalities $|M_i| = m, \forall i = 1, ..., k$.

(d) *Homogeneous benefits:* the benefit function is the same for all $i \in N$, i.e. $b_i(.) \equiv b(.)$

(e) *Cost structure:* if $i, j \in M_k$ then the cost to form a link g_{ij} is equal to $c_k \in [\underline{c}, \overline{c}]$; if $i \in M_k$ and $j \in M_p$ with $k \neq p$ then the cost to form a link g_{ij} is constant and equal to $C > \overline{c}$. A community M_i is said to have high costs of link formation if $c_i > b(1) - b(2)$.

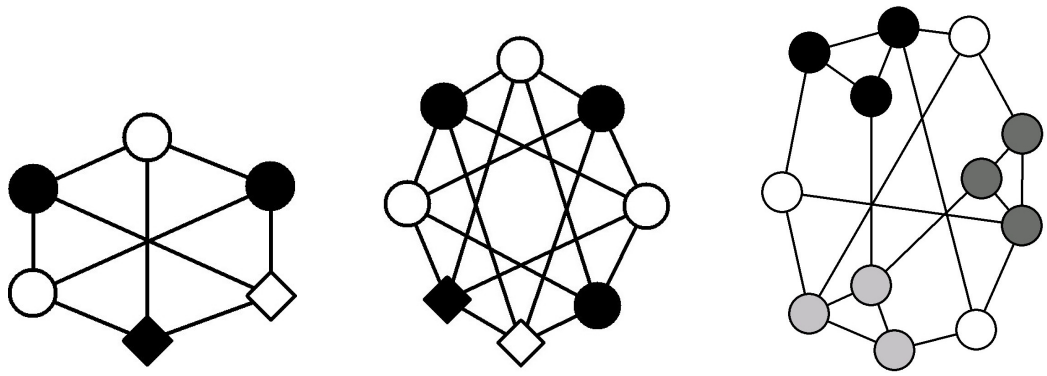
Assumptions *(a)* and *(b)* are for expositional purposes only. Section 5 will show that they are not necessary for the main results to hold. Assumptions *(c)*, *(d)* and *(e)* make the model more tractable. The cost structure in *(e)* has an intuitive interpretation. The internal costs of connections are homogeneous within a community, but they vary depending on the type of community. Moreover, costs of connections across communities are more costly than the internal ones within a community. This captures the fact that a bond between two individuals in the same community is easier to establish because people in the same community are similar, while it requires more effort to bond with someone from another community who has different characteristics.

Why is there the need of introducing heterogeneity both in costs and in the knowledge that agents have about the network? Intuition may suggests that either heterogeneity in costs or in knowledge should suffice to ensure that in any equilibrium network agents connect mainly with other agents in their own community, i.e. the network has *segregation patterns*.[9] The following examples show that this intuition is not true: if only one form of heterogeneity is present then there are equilibrium networks without segregation patterns.

Example 1 - Heterogenous knowledge with homogeneous costs
Assume *(b)-(d)* hold, but costs are homogeneous, i.e. $c_{ij} \equiv c$, $\forall i, j \in N$. It is relatively easy to construct equilibrium networks where there is maximal separation of agents from their communities, i.e. each agent is connected with all the individuals in other

[9]Section 4 gives a formal definition of segregation patterns, but for now think of a network with segregation patterns as a network where most links are *within* instead of *across* communities.

communities and with none in her own. Figures 2(a) and 2(b) are examples of such networks. It is straightforward to check that they are pairwise Nash networks in the $b(1) - b(2) < c < b(1) - b(3)$ range. Architectures with maximal separation are equilibria because they help NI agents in perceiving the network: there is no black-black (or white-white) link so every white (or black) NI agent can perceive any link in the network since any link involves one white (or black) agent. Thus, if one takes away the cost differential there are equilibrium networks without segregation patterns.



(a) A network $N_M = \{M_1, M_2\}$ with $|M_1| = |M_2| = 3$.

(b) A network $N_M = \{M_1, M_2\}$ with $|M_1| = |M_2| = 4$.

(c) A network $N_M = \{M_1, M_2, M_3, M_4\}$ with $|M_1| = |M_2| = |M_3| = |M_4| = 3$.

Figure 2: Nodes with the same color belong to the same community. Round-shaped nodes are NI agents, and diamond-shaped nodes are NC agents. In (a) and (b) there is maximal separation of agents from their communities in both the white and black networks. In (c) there is maximal separation of agents in the white community.

Example 2 - Heterogeneous costs with homogeneous complete knowledge
Assume *(c)-(e)* hold, but all agents have complete knowledge about the network. It is possible to construct networks where there is maximal separation of agents from their community. Consider a network of four communities $N_M = \{black, white, dark, light\}$ with three agents each, and such that $c_{white} < b(1) - b(2) < c_j < b(1) - b(3)$ where $j \neq white$. It is not difficult, and left to the reader, to check that the network in Figure 2(c) is pairwise Nash in the $b(1) + b(2) - 2b(3) < C < b(1) + 2b(2) - 2b(3) - b(4)$ range. Clearly in this network all agents in the white community are separate from each other. Thus, if one takes away the knowledge differential, then there are equilibrium networks without segregation patterns.

These examples give us a hint that the presence of both types of heterogeneity leads to different equilibrium predictions than the presence of either type alone. After characterizing the equilibrium networks in the following section, it will be enlightening to come back to these examples and compare them to the equilibria that are supported when both types of heterogeneity are included.

3.2 Pairwise Nash networks

The following lemma is a key step in characterizing the pairwise Nash networks: it shows that the combination of heterogeneity in costs and in knowledge of the network leads the less knowledgeable agents to connect exclusively within their own community.

Lemma 1. *Assume (b)-(e). Let* $C > b(1) + b(2) - b(3)$ *and* $c < b(1) - b(3)$. *No NI agent forms links with agents in a different community.*

The intuition of the proof is as follows. If costs of connections c_i within communities are low, i.e. $c_i < b(1) - b(2)$ for $i = 1, ..., k$, then the result is straightforward because all communities are complete networks and NI agents cannot get any indirect benefits from inter-community links. If costs c_i are higher, i.e. $b(1) - b(2) < c_i < b(1) - b(3)$ for $i = 1, ..., k$, then the proof is *ad absurdum*. Assume that there exists a link between two agents i and j belonging to different communities and that i is NI. In order for i not to want to sever this link at least one agent k belonging to i's community has to be connected with j. Moreover, there cannot be another agent l directly connected to both i and k, otherwise i would not need the link with j to gain benefits from the indirect connection with k. The proof shows that it is not possible to construct an equilibrium network satisfying these requirements, and therefore the link g_{ij} cannot exist in equilibrium.

The lower bound C on the cost of inter-community links excludes the possibility of small loops that could help sustain a link between i and j. Consider a circle network with 5 agents such that: agents 1-4 belong to community U_i, agent 5 belongs to community U_j, agent 4 is NI and $g_{45} = 1$. The net benefits for 4 to remove the link with 5 are $C - b(1) - b(2) + b(3)$, so if the lower bound in the statement of the lemma does not hold then NI agent 4 would like to keep a link with an agent 5 from another community. The upper bound c on the cost of intra-community links ensures a minimum level of cohesion within communities: no agent is more than two links away from another agent belonging to her community. Without this minimum level of cohesion there might be equilibrium networks where NI agents "infiltrate" another community due to its very sparse structure.

Some more notation and terminology before characterizing the equilibria. Let λ be the proportion of communities with high cost of link formation. Let $\overline{B}(i) \equiv b(i)+(m-1)b(i+1)$ and $\underline{B}(i) \equiv b(i) + b(i+1) + (m-2)b(i+2)$. $\overline{B}(1)$ is the total direct and indirect benefits for the NC agent i to connect to agent $j \in M_j$ if M_j is a complete community. $\underline{B}(1)$ is the total direct and indirect benefits for i if M_j is a star community with j at the periphery. Intuitively $\underline{B}(1)$ and $\overline{B}(1)$ are the lower and upper bound to the benefits from connecting to another community. Clearly, $\underline{B}(i)$ and $\overline{B}(i)$ are the lower and upper bound from being $i-1$ links away from an agent that is directly connected to another community. Recall that the *kernel* K is the subset of NC agents in N. The following proposition fully characterizes the equilibrium networks for $b(1) < C < \underline{B}(1) - \underline{B}(2)$.

Proposition 1. *Assume (a)-(e), and let* $b(1) + b(2) - b(3) < C < \underline{B}(1) - \underline{B}(2)$. *The unique pairwise Nash network architectures are:*

(i) a complete kernel with complete communities, denoted g^{CKC}, *if* $\overline{c} < b(1) - b(2)$

(ii) a complete kernel with mixed complete and d2 communities, denoted g^{CKM}*, if* $\underline{c} < b(1) - b(2) < \overline{c} < b(1) - b(3)$.

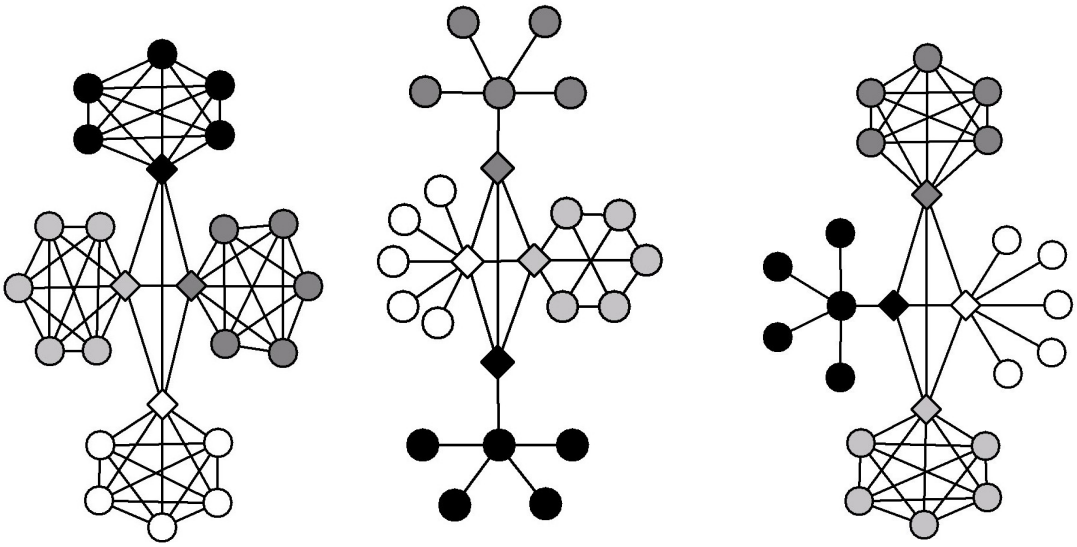
(iii) a complete kernel with d2 communities, denoted g^{CKd2}*, if* $b(1) - b(2) < \underline{c} < \overline{c} < b(1) - b(3)$.

The proof involves several lemmata. The first two lemmata are technical and they show that it is possible to use pairwise stability, a simpler equilibrium concept than pairwise Nash, to characterize the pairwise Nash networks. Lemma 3 proves that the utility defined in (1) is superadditive in own-links. This superadditivity condition means that the marginal utility to an agent i from having a subset of links is higher than the sum of the marginal utilities from each of the links separately. Lemma 4 shows that this condition matters because if $u(.)$ is superadditive then the set of pairwise stable and pairwise Nash networks coincide. Computationally, pairwise stability is a much easier condition to prove because it only requires to check for one-link deviations: unilateral in case of link severance and bilateral for link formation.

By lemma 1 the NI agents do not form any link with agents outside of their community. Using this result it is easy to see that communities M_i with low costs of link formation $c_i < b(1) - b(2)$ are complete communities and communities M_j with high costs of link formation $b(1) - b(2) < c_j < b(1) - b(3)$ are *d2* communities. Finally, the last step is to characterize the structure of the kernel K of NC agents that form links across communities. Given that communities are either g^C or g^{d2}, the benefit of forming a link with an NC agent is in the $[\underline{B}(1), \overline{B}(1)]$ range. Since $C < \underline{B}(1) - \underline{B}(2)$ then the kernel is a complete network and this completes the characterization of the equilibria. Hereafter, the pairwise Nash networks in proposition 1 will be denoted as g^{CK}, or *pairwise Nash networks with complete kernel*.

Figure 3 is a graphical illustration of g^{CK} networks. For instance, Figure 3(c) shows a g^{CKM} network architecture for $\lambda = 0.5$. Note in Figure 3(b) the variety of structures that the *d2* communities can have: the dark gray and black communities are stars with the NC agent at the periphery, the white community is a star with the NC agent at the center, and the light gray community is a hexagon with diagonals. Adopting, for now, an informal definition of a network with segregation patterns as a network where most links are *within* instead of *across* communities, it is clear that the g^{CK} networks have *segregation patterns*. In order to pinpoint the role played by the two types of heterogeneity present in the model, it is useful to return to examples 1 and 2 in the previous section.

Example 1 - Heterogenous knowledge with homogeneous costs
Consider the same set-up as example 1 in section 3.1: assume *(b)-(d)* hold and homogeneous costs $b(1) - b(2) < c < b(1) - b(3)$. The networks without segregation patterns in Figures 2(a) and 2(b) are then pairwise Nash networks. Moreover, it is not difficult to show that the g^{CK} networks in proposition 1 are *not* pairwise Nash networks with homogeneous costs in this parameter range. Thus, if one takes away the heterogeneity in costs then g^{CK} architectures are not equilibrium networks and there are equilibrium



(a) Complete kernel of NC agents with complete communities g^{CKC}.

(b) Complete kernel of NC agents with d2 communities g^{CKd2}.

(c) Complete kernel of NC agents with mixed complete and $d2$ communities g^{CKM} with $\lambda = 0.5$.

Figure 3: Pairwise Nash network architectures with complete kernel for a set of 4 communities $N_M = \{M_1, M_2, M_3, M_4\}$ with $|M_i| = 6$ for all i. Round-shaped nodes are NI agents, and diamond-shaped nodes are NC agents. Nodes with the same color belong to the same community.

networks without segregation patterns.

Example 2 - Heterogeneous costs with homogeneous complete knowledge
Consider the same set-up as example 2 in section 3.1: assume *(c)-(e)* hold, all agents have complete knowledge about the network and $c_{white} < b(1) - b(2) < c_j < b(1) - b(3)$ where $j \neq white$. The network without segregation patterns in Figure 2(c) is pairwise Nash if $b(1) + b(2) - 2b(3) < C < b(1) + 2b(2) - 2b(3) - b(4)$. Moreover, it is not difficult to show that the g^{CK} networks in proposition 1 are *not* pairwise Nash networks in this parameter range if all agents are network cognizant. Thus, if one takes away the heterogeneity in knowledge then g^{CK} architectures are not equilibrium networks and there are equilibrium networks without segregation patterns.[10]

These two examples show that both heterogeneity in costs and in knowledge are necessary to sustain the g^{CK} pairwise Nash networks. Without either type of heterogeneity g^{CK} networks are not equilibria and there are equilibria that do not have segregation

[10]Note that there is one limit case where g^{CK} architectures are equilibrium networks for a narrow parameter range without the knowledge differential. This is when all internal community costs c_i are so low that all communities are complete communities. The case with two communities is analyzed in Jackson and Rogers [2005].

patterns. The cost differential ensures that there are higher net benefits from direct connections with agents in one's own community. The knowledge differential leads the network ignorant agents to underestimate the indirect benefits they would obtain by connecting with agents in other communities, and therefore NI agents connect exclusively within their community leading to the emergence of segregation patterns. Section 4 will show that g^{CK} networks are quite interesting because they have, at least in stylized form, some of the main properties observed in real social networks.

The characterization of equilibria for higher costs $C > \underline{B}(1) - \underline{B}(2)$ of inter-community links is harder because there are multiple equilibrium architectures for the kernel. This multiplicity of equilibria is a well-known result since Jackson and Wolinsky [1996]. If $C < \underline{B}(1) - \underline{B}(3)$ then the kernel will be a $d2$ network, but if $C > \underline{B}(1) - \underline{B}(3)$ then different architectures can be sustained in equilibrium. Among these equilibria there are some with a star kernel that are interesting because, as the following section will discuss, they have one agent with a much higher degree than all the others. Corollary 1 characterizes these networks.

Corollary 1. *Assume (a)-(e), and let* $\overline{B}(1) - \overline{B}(2) < C < \underline{B}(1)$. *The following are pairwise Nash network architectures:*

(i) a star kernel with complete communities, denoted g^{SKC}, *if* $\overline{c} < b(1) - b(2)$

(ii) a star kernel with mixed complete and d2 communities, denoted g^{SKM}, *if* $\underline{c} < b(1) - b(2) < \overline{c} < b(1) - b(3)$.

(iii) a star kernel with d2 communities, denoted g^{SKd2}, *if* $b(1) - b(2) < \underline{c} < \overline{c} < b(1) - b(3)$.

The proof follows closely the one of proposition 1, and it provides no new intuition. The network architectures are similar to the ones in Figure 3 with the only difference being that the diamond-shaped NC agents form a star instead of a complete kernel.

3.3 The advantages and disadvantages of knowledge

A simple inspection of the equilibrium networks in proposition 1 highlights that agents occupy different positions in the network according to their knowledge. More knowledgeable NC agents position themselves at the center of the overall network structure connecting with each other to form the "kernel" of the network. Less knowledgeable NI agents remain at the periphery, insulated within their own community. By comparing the payoffs of NC and NI agents it is therefore possible to explore what are the advantages (and disadvantages) of network knowledge.

Before proceeding, an important caveat. In the equilibrium analysis conducted in the previous sections the NI agents base their linking decisions on their cognitive network. As discussed after equation (1) in section 2, the cognitive network is what matters for *some* of the payoffs an agent receives, e.g. i is not able to get the VIP invitation unless she knows the path leading to l. However, other types of indirect network benefits flow in the network without any need of being 'prompted.' Thus, in the g^{CK} pairwise Nash networks

the NI agents will receive (weakly) higher payoffs from the network than the payoffs they compute given their knowledge of the network. There are different potential explanations of why they do not realize this in equilibrium and change their linking decisions. One of these explanations is that benefits that flow in a network are very 'intangible' and NI agents are not able to pin down the payoff discrepancy. Another one is bounded rationality: they are not able to 'see' where the benefits come from and they are not able to infer it. If this is the case then the payoffs of the NI agents from the actual network can be a useful upper bound on the effective payoffs they receive. These are the payoffs that we will consider for the analysis in this section.

Let g^{CKSP} denote *complete kernel networks with star communities with peripheral NC agents*, i.e. g^{CK} networks where all communities are star networks with the NC agent located at the periphery of the star. The following corollary to proposition 1 compares the payoffs NC and NI agents receive in the g^{CK} equilibrium networks.

Corollary 2. *Consider NC agent $i \in M_i$ and NI agent $j \in M_i$ and a network $g \in g^{CK}$, then:*

(i) $u_i(g) < u_j(g)$ *if* $g \in g^{CKC}$

(ii) $u_i(g) > u_j(g)$ *if* $g \in g^{CKSP}$

The g^{CKC} and g^{CKSP} network architectures are the two extreme cases that give the highest benefits to the NI and NC agents respectively. g^{CKC} networks favor NI agents because all of them are directly linked to NC agents thereby *minimizing* the geodesic distance from NI agents to agents in other communities and *maximizing* the costs that NC agents have to pay to be linked to their own community. On the other hand, g^{CKSP} architectures favor NC agents because each one of them is linked to one NI agent in her community thereby *maximizing* the geodesic distance from NI agents to agents in other communities and *minimizing* the costs that NC agents have to pay to be linked to their own community. In the other g^{CK} pairwise Nash architectures the difference in payoffs between NC and NI agents falls in between these two extremes: star communities with the NC agent at the center and $d2$ communities balance the trade-offs above; g^{CKM} networks are clearly combinations of g^{CKC} and g^{CKd2} networks.

Intuitively, an agent knowledgeable about the network receives higher payoffs in a society formed by communities that are sparsely connected internally. In this type of society the benefits that the NC agent brings in to her own community from the rest of the society are not shared effectively due to the lack of internal cohesion of the community. On the other hand, the NC agent receives relatively lower payoffs in a society formed by many close-knit communities. In this type of society the NI agents have direct access to the benefits the NC agent brings in and they free-ride on her investment to bridge to other parts of the network.

As section 4.2 will discuss more formally, the NC agents are the *brokers* in the pairwise Nash networks in proposition 1. The statement in corollary 2 therefore qualifies Burt's claim that being a broker is beneficial. A broker receives higher payoffs due to her position in the network structure if she is part of a sparsely connected community and/or network.

On the other hand, being a broker in a densely connected community/network may not provide higher payoffs than the ones accruing to an agent who is directly connected to a broker.

4 Structural properties of equilibrium networks

This section shows that equilibrium networks have properties *(i)-(v)* listed in the introduction. Section 4.1 shows that essentially all pairwise Nash networks with a complete kernel are small world networks. Section 4.2 shows that all pairwise Nash networks with a complete kernel have segregation patterns and that the brokers are the network cognizant agents. Section 4.3 shows that some of the pairwise Nash networks have one hub.

4.1 Small world networks

Before giving the formal definition of a small world network, it is necessary to introduce two new concepts.

The first one is the *clustering coefficient* of a network. In the network literature the clustering coefficient of a node of a network is the fraction of its neighbors that are also directly linked to each other. Following Watts [1999], let us focus on the *average clustering coefficient* $C(g)$ of a network g:

$$C(g) = \frac{1}{n} \sum_i \frac{\sum_{j \neq i; k \neq j,i} g_{ij} g_{jk} g_{ik}}{\sum_{j \neq i; k \neq j,i} g_{ij} g_{ik}} \tag{2}$$

The average clustering coefficient of a network is simply the average of the clustering coefficients of all its nodes.[11]

The second one is a particular type of networks called *random networks*, which have been the subject of extensive study in the graph theory literature. It is not difficult to generate a random network: following the seminal papers by Erdős and Rényi [1959, 1960], let $G_{n,p}$ be the set of all networks consisting of n vertices where each pair is connected together with uniform probability p. In order to generate a network sampled uniformly at random from $G_{n,p}$ follow this process: take n initially unconnected vertices and go through each pair of them, joining the pair with an edge with probability p.[12]

In the Erdős-Rényi model, and in many other random graph models, when the number of nodes is large the average geodesic distance has approximately the same magnitude as

[11] Note that this is not the only clustering coefficient metric in the network literature. Apart from measures for directed networks, another popular clustering coefficient measure is the *total clustering coefficient* $C^T(g)$. The latter is the overall fraction of "triangles" in the network, i.e. $C^T(g)$ is not computed node by node and then averaged out, but directly for the whole network. Newman [2003] shows these measures can be very different for certain networks. For the rest of this paper, "clustering coefficient" will mean $C(g)$. However, the results will apply to $C^T(g)$ as well, except for a few special cases of network structures.

[12] Several generating processes for random networks have been the subject of extensive study and for a broad spectrum of generating processes the resulting networks share the same general properties in terms of average geodesic distance and clustering coefficient. See Bollobás [2001] for an extensive review.

the ratio of the logarithm of the total number of vertices in the graph to the logarithm of the average degree of a node. Mathematically, $\bar{d}(g^{random}) \approx log(n)/log(z)$ where $z = \sum_{i=1}^{n}\sum_{j=1}^{n} g_{ij}/n$ is the average degree of a node in g. This approximation holds as long as the average degree of a vertex is greater than one and it is significantly smaller than n, i.e. as long as the number of links is not so small that the network is disconnected in many small components and as long as the number of links is not so large that the network is close to be a complete network. Moreover, the clustering coefficient $C(g^{random})$ of a random graph with n nodes tends to zero as n becomes very large. The interested reader can refer to Watts [1999] for a short derivation.

Following Watts [1999], the formal definition of a small world network is as follows.

Definition 2. A network g is a *small world* if:

(i) it has *short average geodesic distance*. More precisely, the average geodesic distance is similar to the one of a randomly generated network with the same number of nodes and the same average degree, i.e. $\bar{d}(g) \approx \bar{d}(g^{random}) \approx log(n)/log(z)$

(ii) it has a *higher clustering coefficient* compared to the one of a randomly generated network. More precisely, $lim_{n\to\infty} C(g) > lim_{n\to\infty} C(g^{random}) = 0$

The following proposition shows that the majority of g^{CK} pairwise Nash networks are small world networks.

Proposition 2. *For any g^{CK} pairwise Nash network in proposition 1 we have that:*

(i) the maximum diameter $D(g^{CK})$ is five

(ii) if $\lambda < 0.95$ and $n < 10^5$ then $\overline{d}(g^{random}) \approx \overline{d}(g^{CK}) < 5$

(iii) a lower bound on the clustering coefficient is $lim_{m,k\to\infty} C(g^{CK}) = 1 - \lambda$.

The derivation of the results is by inspection of the g^{CK} architectures. Statements *(i)* and *(ii)* show that g^{CK} networks have the first property of a small world, i.e. an average geodesic distance comparable to the one of a random graph with the same number of nodes and links. Statement *(iii)* shows that as long as there is a non-negligible fraction of complete communities, i.e. $\lambda < 1$, then g^{CK} networks have the second property of a small world: a non-negligible clustering coefficient. These results show that small world networks are prevalent in the broad parameter range where the equilibrium networks are g^{CK} networks.

It is important to point out that, even though g^{CK} networks have a rich and varied structure, they are still "stylized" small world networks. This is evident in statements *(i)* and *(ii)*: there is an upper bound of 5 to the diameter, which limits the validity of statement *(ii)* to networks with less than 100,000 nodes. As it is clear from definition 2, small world networks can be defined for any number of nodes. This is an indication that for very large networks this model starts to break down because an even more complex

structure is needed. A potential avenue to explore would be to have the communities be grouped into larger communities that interact to form even larger communities.[13]

Intuitively, all g^{CK} networks in proposition 1 that have a non-negligible fraction of communities with low costs of link formation are small world networks. The requirement to have a few low cost communities is necessary for the clustering coefficient to be bounded away from zero. In the limit, the clustering coefficient of a random network goes to zero while the clustering coefficient of a g^{CK} network does not, as long as $\lambda < 1$. Finally, note that for *(iii)* to hold it is not necessary that both parameters m and k are very large, either of them would suffice.[14]

The presence of agents with complete knowledge about the network is key for the emergence of the first property of small worlds. The NC agents provide the few links across different communities that dramatically shorten the social distance between any two individuals in the network. The presence of a few close-knit communities with low costs of link formation, where everyone is connected to everyone else, is enough for the emergence of the second property of small world networks.

To sum up, *both* heterogeneity in costs *and* heterogeneity in knowledge determine the first property of a small world network: a short average distance between individuals in the network. Moreover, heterogeneity in costs is key to have the second property: an average clustering coefficient bounded away from zero due to the presence of communities with low cost of link formation.

4.2 Segregation patterns with brokers

A prominent characteristic of social networks is that they show segregation patterns. Intuitively, individuals of the same type stick together in close-knit communities and they form very few links outside of the community.[15] A classical example is racial segregation patterns in US urban areas. Moreover, these close-knit communities are connected by a few agents, called *brokers*, who strategically position themselves in the social network structure to bridge different communities.

A basic metric to measure the extent of segregation of an individual $i \in M_p$ is the fraction of i's connection that are with members of her community M_p. Averaging the segregation of all the agents in M_p gives a measure of the overall segregation of the M_p community.[16] Formally, the *segregation index* S_p of a community M_p is equal to:

$$S_p = \frac{1}{m} \sum_{i \in M_p} \left(\frac{\sum_{j \in M_p} g_{ij}}{\sum_{j \in M_p} g_{ij} + \sum_{k \notin M_p} g_{ik}} \right) \tag{3}$$

[13] These networks would have a fractal-like structure. See Dorogovtsev et al. [2002] for a deterministic, non-strategic model of network formation that leads to a pseudo-fractal graph with many of the empirical properties observed in real networks.

[14] Specifically, one can show (see proof of proposition 2*(iii)* in appendix A) that if k is finite then $lim_{m \to \infty} C(g^{CK}) = 1 - \lambda$. Viceversa, if m is finite then $lim_{k \to \infty} C(g^{CK}) = 1 - \lambda + \frac{\lambda}{m} \geq 1 - \lambda$.

[15] As the popular saying goes, "birds of a feather flock together."

[16] See Currarini et al. [2008] for a review and discussion of this and other segregation metrics used in the literature.

The higher is this ratio, the more segregated is the community from the rest of the network.[17] Also, let us define the *segregation NI index* S_p^{NI} to be equal to (3) above with the restriction that i has to be an NI agent. The S_p^{NI} index captures the extent that network ignorant agents are segregated from other communities.

The concept of broker captures the idea of an agent that is on many paths connecting other agents. The *betweenness centrality* metric first defined by Freeman [1977] defines this notion more formally. Let $\eta[p_{jk}(g)]$ be the number of paths $p_{jk}(g)$ between j and k in the network g such that $|p_{jk}(g)| = d_{jk}(g)$, i.e. such that the path length is equal to the geodesic distance between j and k. Also, let $\eta_i[p_{jk}(g)]$ be the number of geodesic paths between j and k that pass through agent i, where $i \neq j, k$. The betweenness centrality $I_B(\eta_i)$ of an agent i is then equal to:

$$I_B(\eta_i) = A \sum_{k=1}^{n-1} \sum_{j=k+1}^{n} \frac{\eta_i[p_{jk}(g)]}{\eta[p_{jk}(g)]} \tag{4}$$

where $j, k \neq i$ and $j \in M_p$, $k \in M_q$ with $q \neq p$. A is just a normalization factor so that $I_B(\eta_i) \in [0, 1]$.[18]

Armed with these metrics it is now possible to give more formal definitions of what it means for a network to have segregation patterns and for an agent to be a broker.

Definition 3. A network g has *segregation patterns* if $S_p > \frac{1}{2}$ for every $p = 1, ..., k$ and it has *perfectly segregated NI communities* if $S_p^{NI} = 1$ for every $p = 1, ..., k$. An agent $i \in M_p$ is a *broker* if $I_B(\eta_i) > I_B(\eta_j)$, for every $j \in M_p, j \neq i$.

The first part of the definition says that a network with segregation patterns is formed by communities whose members connect mainly with each other. Moreover, in a network with perfectly segregated NI communities the NI agents do not form links with members of other communities.The second part says that the broker(s) in each community is the agent(s) with the highest betweenness centrality, i.e. the one(s) who is more crucial to give that community access to the individuals in other communities.

The following proposition shows that g^{CK} networks have segregation patterns and perfectly segregated NI communities, and that having complete knowledge about the network structure is key to be a broker.

Proposition 3. *All g^{CK} pairwise Nash network in proposition 1 are such that:*

(i) they have segregation patterns and the minimum segregation index is

$$S_p \geq 1 - \frac{k-1}{mk} > \frac{2}{3}$$

[17] Blau [1977] makes the important point that this basic metric is sensitive to differences in community size. However, this criticism can be safely ignored here because of assumption *(c)* that all communities have the same size.

[18] Note that this definition restricts the metric to the paths between agents in different communities, while Freeman [1977] defined it for any two agents in the network. The reason is that the goal here is to understand which agents connect different communities in the network.

(ii) they have perfectly segregated NI communities

(iii) the k brokers are the NC agents

The proof is by inspection of the g^{CK} network architectures. Parts *(ii)* and *(iii)* tell a clear story: in each community the NI agents only connect with each other because their limited knowledge on the network structure does not allow them to see the benefits of connecting across communities, and therefore the only brokers are the NC agents. Thus, all g^{CK} networks have segregation patterns. Moreover, the level of segregation is very high. The lower bound of $\frac{2}{3}$ only applies to the limit case of communities of only 3 agents. Unless the communities are very small, the minimum level of segregation is much higher. For instance, if $m = 10$ then the minimum segregation index is 0.9; if $m = 50$ then the minimum segregation index is 0.98.

It is important to note that *both* heterogeneity in costs *and* heterogeneity in knowledge are important to obtain the results in proposition 3. Example 1 in section 3 shows that with heterogeneous knowledge and homogeneous costs the networks in Figures 2(a) and 2(b) are pairwise Nash. These two networks have none of the properties *(i)-(iii)* above: both the black and white communities have zero segregation index and all agents are brokers. Example 2 in section 3 shows that with heterogeneous costs and homogeneous knowledge the network in Figure 2(c) is pairwise Nash: in this network the white community has zero segregation index.

4.3 Hubs

The majority of social networks have a degree distribution with a "long tail": there are *a few* agents, called hubs, who have a *much higher degree* than the other individuals in the network. Physicists have shown that this has important implications for many phenomena ranging from diffusion processes to the robustness of networks to random and targeted attacks.[19] The pairwise Nash networks in this model do not have enough degree variation, so the overall shape of the degree distribution does not provide much information. However, for illustrative purposes, it is still useful to show that some equilibrium networks in this model satisfy a stylized definition of a network with a hub.

Definition 4. An agent i is a hub in the network g if $|L_i(g)| \gg |L_j(g)|$, $\forall j \neq i$.

Clearly, this is a basic and stylized definition which is also restrictive because it rules out the possibility of having more than one hub in a network. However, it has the advantage of being very clear-cut and of being very unambiguous in identifying as a hub the agent that satisfies it. The following remark points out that a simple inspection of the equilibrium networks with a star kernel in corollary 1 reveals that all g^{SK} networks have one hub as long as the number of communities is large compared to the number of agents in each community.

Remark. If $k \gg m$ then any g^{SK} network has one hub, and the hub is an NC agent.

[19] See Newman et al. [2006] for a comprehensive review.

Proof. Let i be the center of the star kernel, then $\min_{g^{SK}}\{|L_i(g^{SK})|\} = k$. Consider any $j \neq i$, then $\max_{i\neq j, g^{SK}}\{|L_j(g^{SK})|\} = m$. Clearly, i is the hub and by definition he is an NC agent. □

There are equilibrium networks which have an agent with a much higher number of connections than all the other agents in the network. The hub is the central individual in the "super-community" of NC agents that connect different communities. To become the hub it is crucial to be one of the agents that have better knowledge about the network because only by connecting outside of one's community it is possible to have a number of links that is much larger than any other agent in the network.

Finally, note that a network can be a small world without a hub (e.g. a large number of g^{CK} networks), or, viceversa, it can have a hub but not be a small world (e.g. g^{SKS} networks with $\lambda \approx 1$), and it can also be a small world with a hub. This is in line with the findings in the empirical literature, but it is not uncommon to come across studies that erroneously assume that *any* social network must necessarily have both small world characteristics and a degree distribution with a "long tail."

5 Extensions

This section explores two extensions of the basic model. Section 5.1 shows that the properties of equilibrium networks are robust to relaxing the assumption of local uniqueness of the NC agents. Section 5.2 shows that the properties of the equilibria are robust to increasing NI agents' knowledge of the network.

5.1 Robustness to multiple NC agents

Consider the model in section 3, without assumption *(a)* of local uniqueness of the NC agents. Assume that in each community M_i there are p_i NC agents, with $0 < p_i < m - 1$. Moreover, let $P_i \subset M_i$ be the subset of NC agents in the community M_i. Let the set $K_P = \{P_1, ..., P_k\}$ be the *p-kernel*. Subsets P_i and P_j are connected, i.e. $g_{P_iP_j} = 1$, if and only if there exists *at least* one pair of agents $k \in P_i$ and $l \in P_j$ such that $g_{kl} = 1$. K_P is complete if $g_{P_iP_j} = 1$, $\forall i, j = 1, ..., k$, $i \neq j$. Finally, let p_{max} be the maximum number of NC agents in any community M_i, and let p^*_{max} be the maximum number of NC agents in any community with low cost of link formation $c_i < b(1) - b(2)$.

The following proposition is the equivalent of proposition 1 for the case of multiple NC agents.

Proposition 4. *Assume (b)-(e), and let* $b(1) < C < \underline{B}(1) - \underline{B}(2)$. *The unique pairwise Nash network architectures are:*

(i) a complete p-kernel with complete communities, denoted g^{CpKC}, *if* $\overline{c} < b(1) - b(2)$

(ii) a complete p-kernel with mixed complete and d2 communities, denoted g^{CpKM}, *if* $\underline{c} < b(1) - b(2) < \overline{c} < b(1) - b(3)$.

(iii) a complete p-kernel with d2 communities, denoted g^{CpKd2}*, if* $b(1) - b(2) < \underline{c} < \overline{c} < b(1) - b(3)$*.*

The proof follows closely the one of proposition 1. There is one main difference between the statements of the two propositions: here the p-kernel, not the kernel, is a complete network. This means that the kernel of NC agents is not necessarily a complete network, and multiple network architectures for the kernel are possible. It is easy to see why: suppose there are two complete communities M_1, M_2 with NC agents $y_1, y_2 \in M_1$ and $y_3, y_4 \in M_3$. If $g_{y_1 y_3} = 1$ then it might be that $g_{y_1 y_4} = 0$ because y_1 is already getting benefits from its indirect connection with y_4 through $g_{y_1 y_3} = 1$.

The proposition below shows that also g^{CpK} networks are small world networks.

Proposition 5. *For any* g^{CpK} *pairwise Nash network in proposition 4 we have that:*

(i) the maximum diameter $D(g^{CpK})$ *is five*

(ii) if $\lambda < 0.95$ *and* $n < 10^5$ *then* $\overline{d}(g^{random}) \approx \overline{d}(g^{CpK}) < 5$

(iii) if p^*_{max} *is finite then a lower bound on the clustering coefficient is* $lim_{m,k\to\infty} C(g^{CpK}) = 1 - \lambda$*.*

The intuition for this result is as follows. g^{CpK} networks are structurally the same as g^{CK} networks, but with a higher number of links across communities due to the higher number of NC agents. Thus, the upper bounds on diameter and average geodesic distance in *(i)* and *(ii)* in proposition 2 apply here as well. The difference with proposition 2 is that some g^{CpK} equilibrium networks with many NC agents and with communities with low cost of link formation have a higher number of inter-community links than g^{CK} networks. However, it is straightforward to show that the main result is unchanged: the average geodesic distance of g^{CpK} networks is short and similar to the one of an equivalent random network.

Proposition 2 has shown that in the limit $C(g^{CK})$ is determined by the proportion $1-\lambda$ of completely connected communities, i.e. by the NI agents in the completely connected communities who have clustering coefficient equal to one. These NI agents will have clustering coefficient equal to one in g^{CpK} architectures as well. Thus, as long as there is enough of them, i.e. p^*_{max} is finite, then the limit of $C(g^{CpK})$ will be approximately the same as the limit of $C(g^{CK})$. It follows that, if $\lambda < 1$ and p^*_{max} is finite, the limit of $C(g^{CpK})$ will be bounded away from zero for large g^{CpK} networks. Thus, small world networks are prevalent even with multiple NC agents in each community.

The proposition below also shows that g^{CpK} networks have segregation patterns and that the role of brokers is still mainly played by the NC agents.

Proposition 6. *All* g^{CpK} *pairwise Nash network in proposition 4 are such that:*

(i) if $p_{max} \leq \frac{m}{2}$ *they have segregation patterns*

(ii) they have perfectly segregated NI communities

(iii) the brokerage role is played by NC agents:

$$\sum_{i \in M_p, i \in K} I_B(\eta_i) > \sum_{j \in M_p, j \notin K} I_B(\eta_j)$$

for any community M_p, $p = 1, ..., k$.

These statements mirror and enrich the results in proposition 3. As in proposition 3*(ii)*, the NI agents connect exclusively with each other driving the emergence of segregation patterns. Proposition 6*(i)* shows that the finding in proposition 3*(i)* is very robust: segregation patterns emerge even if up to half of the agents in each community are NC agents. Unless the majority of agents have complete knowledge of the network, the emergence of segregation patterns is an inevitable by-product of the intrinsic difficulty of perceiving the network structure.

The statement in proposition 6*(iii)* differs from proposition 3*(iii)* because here the NC agents "compete" for providing brokerage to their community and therefore the brokerage role gets diluted among them. Because of this competition in some g^{CpK} networks an NI agent may have higher betweenness centrality than everyone in her community if she is directly connected to the NC agents and therefore on the paths connecting them to everyone else in the community. For instance, imagine a star community with two NC agents i and j at the periphery of the star and an NI agent k at the center. If i and j split their inter-community links then k will have higher betweenness centrality than either of them because all the geodesic paths from other communities to the other individuals in the star community will pass through the center.

However, the overall brokerage role between any given community and the rest of the network is still mainly in the hands of the subset of NC agents belonging to that community: any geodesic path connecting two agents in different communities involves *at least* one intermediary NC agent from each community, but it involves *at most* one intermediary NI agent from each community. Thus, the sum of the betweenness centralities of all the NC agents in one community is always strictly higher than the sum of the betweenness centralities of all the NI agents in the same community.

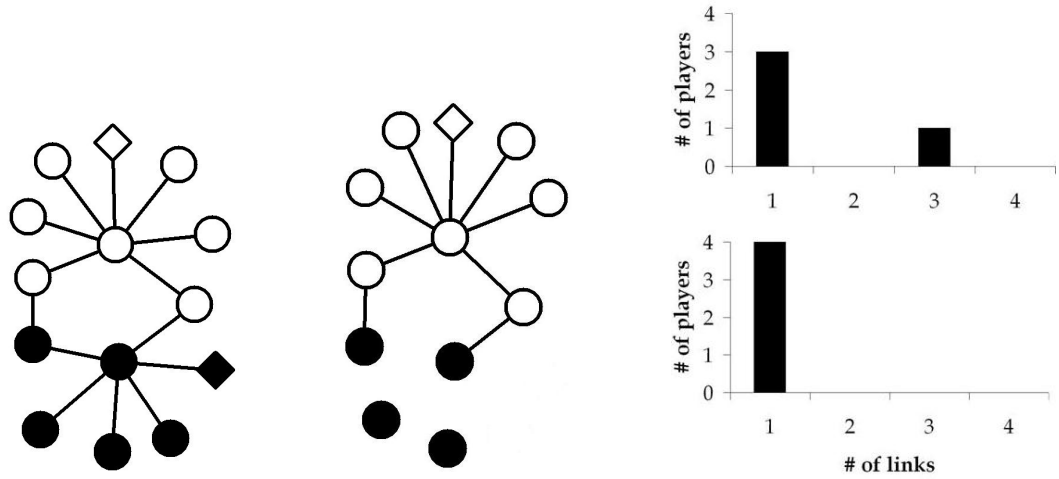
5.2 Better informed NI agents

The model in section 3 makes a strong assumption on the lack of cognitive knowledge of network ignorant agents: they do not know anything about connections that do not involve at least one agent from their group. This section relaxes that assumption and it shows that the results do not change with the more general and realistic form of network ignorance below.

(ii) *Network Ignorant (NI)*: an NI agent $i \in M_p$ has complete knowledge about any link g_{jk} where $j \in M_p$ and/or $k \in M_p$. Additionally, for each community M_q $(q \neq p)$ she knows of the existence of a randomly chosen fraction of agents $\psi_q m_q$ $(0 < \psi_q < 1)$, and for these $\psi_q m_q$ agents she knows the intra-community degree distribution

$\varphi_i(\psi_q)$. Finally, she does not have any knowledge of any inter-community link g_{lt} (where $l \in M_q$, $t \in M_s$, $s \neq q, p$) and she assumes these links do not exist.

Figure 4 illustrates the new knowledge structure. The network in Figure 4(a) is the real network, which is also the network perceived by the NC agents. The network in Figure 4(b) is the cognitive network g^i of one of the NI agents i belonging to the white community: she knows of the existence of a sample of two thirds ($\psi = \frac{2}{3}$) of the agents in the black network, but she is not able to see the links among these agents. However, i knows the degree distribution of the agents in this sample. If the central agent in the black network is included in the randomly picked sample then the degree distribution will be the one on top in Figure 4(c), otherwise it will be the one on the bottom.



(a) The real network g, which is also the cognitive network of the diamond-shaped NC agents.

(b) The cognitive network of the circle-shaped NI agents in the white community.

(c) Additional knowledge available to one of the circle-shaped NI agents (assuming $\psi = \frac{2}{3}$) in the white community.

Figure 4: A network $N = \{M_1, M_2\}$ with $|M_1| = 8$ and $|M_2| = 6$. Individuals in M_1 are color-coded in white, and individuals in M_2 are color-coded in black. Round-shaped nodes are NI agents, and diamond-shaped nodes are NC agents.

An NI agent $i \in M_i$ computes the expected benefits $EB_i(g_{ij})$ from a link g_{ij} with agent $j \in M_j$ by assigning equal probabilities that j is one of the $\psi_j m_j$ individuals in i's sample of the M_j community. Thus, the expected indirect benefits from a link with j are equal to the expected number of connections j has, given the sample degree distribution. Consider an example with the networks in Figure 1 and assume that the randomly picked distribution is the one on top in Figure 4(c). The expected benefits of connecting to j are equal to: $EB_i(g_{ij}) = b_i(1) + \frac{1}{4}(b_i(2) \cdot 3 + 3b_i(2) \cdot 1) = b_i(1) + 1.5b_i(2)$.The maximum possible expected benefits for any NI agent i connecting to an agent k in another network

M_q are clearly when M_q is a complete network. In that case:

$$EB_i(g_{ij}) = b_i(1) + \frac{1}{\psi_q m_q}[(\psi_q m_q - 1)\psi_q m_q b_i(2)] = b_i(1) + (\psi_q m_q - 1)b_i(2) \equiv EB_{max}$$

In addition to the ignorance, there is some degree of bounded rationality in the specification of the NI agents. First, the definition implicitly assumes that NI agent $i \in M_i$ is not able or does not try to figure out whether an individual $j \in M_j$ who is linked to individuals in M_i, i.e. links that i is able to perceive, is included in the randomly picked sample for M_j or not. The NI agent simply assumes that j has a number of links with other agents in M_j, which is equal to the expected number of links given the sample distribution. Second, NI agent i does not try to figure out the second order degree distribution of another community, i.e. the value of the expected indirect benefits of connections that are two links away from an agent she is considering whether to connect to. Finally, agents do not know and they are not able to figure out whether other individuals are NI or NC. Note that all these assumptions are very plausible if communities are relatively large.

It is useful to discuss a practical example to see what this knowledge structure adds to the previous one. As in section 2, imagine the network in Figure 4(a) depicts the friendship relations in a firm: the white nodes belong to the marketing department and the black ones are in IT. The diamond-shaped NC employee in marketing is omniscient: he knows all the connections in the marketing department, the two inter-community connections, and who is connected to whom in IT. As in section 2 a circle-shaped NI employee in marketing knows who is connected to whom in the marketing department and the two inter-community connections. Additionally, now a circle-shaped NI employee also knows of the existence of four IT people and she knows that one person "dominates" the IT social sphere because she has three connections, while the others have only one. This additional knowledge increases the value that she places on connections with someone in the IT department because she is able to see that this friendship would also bring additional indirect benefits.

Now consider the model in section 3 replacing the assumption *(b)* of simple NI agents with the more sophisticated form of NI agents in *(ii)* above. For expositional purposes, assume $\psi_q \equiv \psi$, $\forall q = 1, ..., k$. The following proposition is the equivalent of proposition 1 for the case of more knowledgeable NI agents.

Proposition 7. *Assume (a), (c)-(e) and that ψ is such that $EB_{max} < \underline{B}(1) - \underline{B}(2)$. Let $EB_{max} < C < \underline{B}(1) - \underline{B}(2)$. The unique pairwise Nash network architectures are:*

(i) a complete kernel with complete communities, denoted g^{CKC}, if $\overline{c} < b(1) - b(2)$

(ii) a complete kernel with mixed complete and d2 *communities, denoted g^{CKM}, if $\underline{c} < b(1) - b(2) < \overline{c} < b(1) - b(3)$.*

(iii) a complete kernel with d2 *communities, denoted g^{CKd2}, if $b(1) - b(2) < \underline{c} < \overline{c} < b(1) - b(3)$.*

The only difference from proposition 1 is that the range of values allowed for the cost C is narrower since, by definition, $EB_{max} > b(1)+b(2)-b(3)$. This makes intuitive sense: as the knowledge available to NI agents increases (i.e. ψ increases), it becomes more difficult to sustain equilibrium networks where the NC agents use their better knowledge to acquire a strategic position as brokers among different communities. Note that if EB_{max} is high enough then these equilibria may not exist.

These equilibrium network architectures are exactly the same as the networks characterized in proposition 1. Thus, the discussion of the properties of these equilibrium networks in section 4 applies here as well. To summarize, these networks are such that: *(i)* they have short diameter and average geodesic distance similar to the ones of a random graph g^{random} with the same number of links and nodes; *(ii)* they have higher clustering coefficients than g^{random}, as long as a non-negligible fraction of communities M_is are low cost, i.e. $c_i < b(1) - b(2)$; *(iii)* they have segregation patterns and the NC agent $i \in M_i$ has higher betweenness centrality than any other agent $j \in M_i$, i.e. the k NC agents are the brokers in the network.

It is interesting to discuss what happens in the $\psi \to 1$ limit. First of all, the NI agents have complete knowledge of other communities with low cost of link formation because in those communities each agent is connected to everyone else in their community, and therefore if NI agents have complete knowledge of the distribution of links then they also know the degree of each individual. Thus, none of the g^{CK} networks is an equilibrium because $EB_{max}(\psi \to 1) > \underline{B}(1) - \underline{B}(2)$ and therefore in all equilibrium networks NI agents will link with agents in other low cost communities. However, NI agents still do not have complete knowledge of the communities with high cost of link formation. Thus, the equilibrium networks in this limiting case are different from the ones illustrated in example 2 in section 3. Unlike the case where everyone has complete knowledge, here segregation patterns will still always exist in equilibrium because NI agents still do not have complete knowledge on the communities with high costs of link formation.

Finally, it is worth pointing out that there are two dimensions in the incompleteness of the knowledge that NI agents have. First, they have knowledge only on a fraction ψ of the agents in another community. Second, they only know the degree distribution, not the exact degree of each node. These two dimensions are not both required to obtain the results in proposition 7. The first would suffice: if the NI agent knows the exact network structure of a fraction ψ of the agents in another community then, for appropriate values of ψ, it is still possible to obtain that the benefits to an NI agent of a connection to any community would be lower than $C < \underline{B}(1) - \underline{B}(2)$, and therefore only NC agents would connect across communities. On the other hand, the second dimension alone would not suffice: let M_j be a complete community, if an NI agent $i \in M_i$ knows the degree distribution of M_j, then it has complete knowledge on that community and therefore the same degree of complete knowledge of the NC agent.

6 An application

Recent years have seen the rapid development of web-based social networking tools which have entered many people's daily lives: *facebook.com* boasts more than 150m users after only 5 years of operations, and it is now one of the major players among internet corporations.[20] Alongside generalist social networking tools, there has been the development of specialized professional networking sites that claim to provide a competitive edge by increasing the users' knowledge of their professional network. For instance, *linkedin.com* is a social network site targeted to professionals to "discover inside connections when you're looking for a job or new business opportunity."

Firms are becoming increasingly receptive to the advantages of knowing the social network within and outside the organization. For instance, *hooversconnect.com* sells a software product that maps email/contact networks within an organization to increase employees' knowledge of the professional resources that may be available to them through the network of other members of the organization. The underlying theme behind all these social networking websites, tools and applications is that an increase in network knowledge is beneficial to an employee and to an organization. The model developed in this paper provides a framework to test these claims.

Consider a simpler version of the model with two low cost communities, and where all agents are NC. Moreover, also assume for convenience that the two communities have the same size. Formally, let $k = 2$, $|M_1| = m_1 = |M_2| = m_2 \equiv m$ and $c_1, c_2 < b(1) - b(2)$. To avoid long algebraic expressions define the following quantities:

$$\begin{aligned} E_L(p) &= b(1) - b(3) + 2(m - p - 1)[b(2) - b(3)] \\ E_U(p) &= E_L(p) + b(2) - b(3) \\ P_L(q) &= b(1) - b(3) + (m - q - 1)[b(2) - b(3)] \\ P_U(q) &= P_L(q) + b(2) - b(3) \end{aligned}$$

As the following proposition states, $E_L(p)$ and $E_U(p)$ are the lower and upper bounds for the range of the inter-community costs C for which the unique efficient network has p inter-community links. Similarly, $P_L(q)$ and $P_U(q)$ are the lower and upper bounds for the range of the inter-community costs C for which the unique pairwise Nash network has q inter-community links.

Lemma 2. *Assume $k = 2$, $m_1 = m_2 \equiv m$, $c_1, c_2 < b(1) - b(2)$ and that all agents are NC:*

(i) For any value of C such that $E_L(1) < C < E_U(m-1)$, there is a unique efficient network formed by two complete communities linked by p^* inter-community links such that each agent is involved in at most one inter-community link, and where p^* is such that $E_L(p^*) < C < E_U(p^*)$ with $1 \leq p^* < m$.

[20] Other general social networking sites include *orkut.com* (operated by Google), *hi5.com*, *bebo.com*, etc.

(ii) For any value of C such that $P_L(1) < C < P_U(m-1)$, there is a unique pairwise Nash network formed by two complete communities linked by q^* inter-community links such that each agent is involved in at most one inter-community link, and where q^* is such that $P_L(q^*) < C < P_U(q^*)$ with $1 \leq q^* < m$.

Fix C so that $E_L(p^*) < C < E_U(p^*)$ and therefore the efficient network has p^* inter-community links. Then $P_U(p^*) \leq E_L(p^*)$ for all $p^* \in [1, m-1)$.Hence the unique pairwise Nash network is never efficient.

The structure of the efficient and equilibrium networks is intuitively clear. Internally the communities are complete networks because the cost c of intra-community links is low. Moreover, the first few links across communities bring very high indirect benefits which overcome their relatively high costs. As inter-community links cumulate, the indirect benefits they bring decrease until it is no longer beneficial to form these links.

There is a fundamental discrepancy between efficient and equilibrium networks: pairwise Nash networks are never efficient because they are under-connected. This is analogous to the result in Jackson and Wolinsky [1996] for the connections model: if costs are moderately higher than direct benefits then the star is the efficient network but it is not an equilibrium network. The intuition for the discrepancy is that an inter-community link brings indirect benefits to both the initiator of the link and the other members of her community, but the latter benefits are not taken into account by the initiator in her decision. In other words, there are positive externalities to link formation that lead to the discrepancy between equilibrium and efficient networks.

Now consider a society where the knowledge of the network is low, and the introduction of a social networking tool allows some agents to increase their knowledge of the network. Formally, suppose that in the initial society all agents are NI and that it is possible to turn q agents into NC by, e.g., giving them access to a social networking tool. Note also that the caveat in the second paragraph of section 3.3 applies here as well: the welfare analysis that follows considers the payoffs from the real, not the cognitive, network. The following proposition states the effects of increasing knowledge of the network in this society.

Proposition 8. *Let $P_L(1) < C < P_U(m-1)$. Consider the set-up in lemma 2, except that all agents are NI:*

(i) Increasing knowledge of the network by turning some agents into NC agents (weakly) increases welfare, and

(ii) Turning p^* agents in each community into NC agents is sufficient to match the welfare of the pairwise Nash network where all agents are NC.

The intuition for the proof is rather straightforward, and the results follow almost directly from lemma 2. The message of proposition 8 is that the impact of social networking tools is positive because they increase agents' knowledge of the network, allowing them to better exploit its benefits. However, there is no need for everyone in a community to know about the network: a few agents within each community suffice.

There are several service firms[21] that offer software packages and consulting services that involve mapping the informal network of an organization to increase employees' knowledge of the network. The standard approach is to map the network and then feedback the information to all the employees. This turns out to be a rather costly process, mainly due to the demands on employees' time. An alternative approach that is less time-consuming and equally effective would be a 'pick-the-winners' strategy: select a few individuals who will receive the feedback from the network analysis and therefore acquire better knowledge of the network. These individuals would then act as brokers, while the rest would benefit from the indirect benefits of brokerage.

Note that the results in proposition 8 focus on a very specific case of the model with only two communities and very low intra-community costs of link formation. The need of these specific assumptions is due to the well-known difficulties of characterizing efficient networks with heterogeneity in costs. However, I conjecture that the statements in proposition 8 apply to the general setting with k communities and intermediate costs of intra-community links. The intuition is that even in the general case the pairwise Nash networks will be under-connected compared to the efficient networks. This implies that forming additional links will always improve welfare because there are positive externalities to link formation. Thus, as the knowledge of the network increases, the pairwise Nash network will either stay unchanged or it will have "new" links leading to a (weakly) positive change in welfare. Moreover, the pairwise Nash network will obviously not be the complete network, so it should be possible to obtain the same pairwise Nash network with a lower number of NC agents.

7 Conclusion

This paper has presented a strategic model of network formation where agents have incomplete and heterogeneous knowledge of the network structure. For a broad range of the parameters, the unique pairwise Nash equilibrium networks are such that *(i)* the average and maximum distance between any two agents in the network are similar to those in an equivalent random network; *(ii)* the clustering coefficient is significantly higher than in an equivalent random network; *(iii)* segregation patterns are a robust equilibrium feature and *(iv)* the segregated communities are connected by the brokers who are the agents with complete knowledge of the network. Moreover, in a different parameter range, *(v)* some equilibrium networks have one hub: an agent with complete knowledge of the network with a much higher number of connections than anyone else.

The heterogeneity in the knowledge of the network breaks the symmetry of the payoff structure and, jointly with the cost heterogeneity, is a key driver of the emergence of properties *(i)-(v)*. The presence of agents with complete knowledge is key to shorten the social distance between individuals in the network, because these agents see the benefits of connecting otherwise separate communities. The presence of agents with incomplete knowledge is essential for segregation patterns to emerge. Moreover, the individuals with

[21] Examples include *orgnet.com*, *keyhubs.com*, *morphix.com*, *netminer.com*, etc.

complete knowledge are the only brokers: they strategically position themselves to be at the center of the paths that connect different parts of the network. Finally, the hub is always an agent with complete knowledge because connecting to individuals in different communities is key to cumulating a very high number of connections.

Knowledge of the network has (weakly) positive welfare consequences both at the individual and at the collective level. However, depending on the social environment it may be comparatively more beneficial to be directly connected to someone knowledgeable instead of being a knowledgeable individual. If the social network is very dense and close-knit then it is better to free ride on the benefits knowledgeable individuals bring into the community. On the other hand, if the network is sparse then it is comparatively better to be the knowledgeable person who sees the value of having prime access to other parts of the network. In order to obtain the maximum collective welfare that can be achieved in equilibrium it is not necessary that all individuals have complete knowledge of the network; a few knowledgeable individuals in each community suffice.

Organization Network Analysis (ONA) is becoming a widespread tool to analyze and improve the performance of the informal component of organizations. Cross and Parker [2004] and other studies provide extensive evidence that the social network structure and the position of key individuals matter for individual and collective performance. Cross and Thomas [2009] and a variety of firms providing tools and services to map companies' informal networks claim that increasing employees' knowledge of the social network has a positive impact on performance. An application of this model supports these claims, showing that increases in knowledge have a (weakly) positive welfare impact. Furthermore, this model suggests that it is not necessary to feedback the results of an ONA to the whole organization: a more cost-effective approach of targeting a subset of employees would suffice to obtain the same positive impact on collective welfare.

It is worthwhile to point out that this model is not only applicable to social networks. There are a variety of networks for which knowledge about the network structure varies across the entities involved and such that this knowledge matters for equilibrium outcomes. An example is a partnership network among firms. When a firm i decides whether to invest in a partnership with firm j, the knowledge on j's existing partnerships and the way they fit in the overall partnership network in the industry is relevant for i's decision. Moreover, many firms do not disclose their partnership agreements, so better knowledge about the partnership network might lead to strategic positioning as brokers and the emergence of small world networks.

An avenue for future research is to test experimentally the implications of the model. The methodology developed by Krackhardt [1987] is so far the only one available to map cognitive networks and investigate their role on determining economic outcomes. However, this methodology is only applicable to small networks and it would not be adequate to test the predictions for the structural properties of relatively large networks. There is the need to develop a new methodology to untangle the causality arrow: does better cognitive knowledge of the network lead to strategic positioning, or do certain positions in the network structure facilitate learning the network? The answer to this question could shed further light on the role that individuals' knowledge of the network plays in

shaping network structures, and it would open the path for policy interventions to improve individual and collective welfare by enhancing knowledge of the social structure everyone belongs to.

A Appendix: Proofs

This appendix contains all the proofs omitted in the main body of the paper. The following terminology and concepts will be useful for some of the proofs. Let $l_i(g) \subseteq L_i(g)$ be any subset of i's links in g. The *marginal utility* for an agent i from a set of links $l_i(g)$ in a network g is:

$$mu_i[g, l_i(g)] = u_i(g) - u_i[g - l_i(g)] \tag{5}$$

Following Bloch and Jackson [2006b], a utility function $u(.)$ is *superadditive in own-links* if:

$$mu_i[g^i, l_i(g^i)] \geq \sum_{g_{ij} \in l_i(g^i)} mu_i(g^i, g_{ij}) \tag{6}$$

for all i, g^i and $l_i(g^i) \subseteq L_i(g^i)$.

Jackson and Wolinsky [1996] first defined the notion of pairwise stability, a weaker equilibrium concept than pairwise Nash equilibrium. The graph g is *pairwise stable* if: *(i)* $\forall g_{ij} \in g$, $u_i(g^i) \geq u_i(g^i - g_{ij})$ and $u_j(g^j) \geq u_j(g^j - g_{ij})$ and *(ii)* $\forall g_{ij} \notin g$, if $u_i(g^i) < u_i(g^i + g_{ij})$ then $u_j(g^j) > u_j(g^j + g_{ij})$. Hereafter denote by $PS(u)$ and $PN(u)$ the sets of all possible pairwise stable and Nash networks respectively, when $u(.)$ is the functional form of the utility.

Proof of Lemma 1. The strategy of the proof is to show that there is no pairwise stable network g with a link g_{ij} such that i and j belong to different communities and i and/or j is a NI agent. By lemmata 3 and 4 it then follows that there is no pairwise Nash network g with the above characteristics.

First, consider a community M_p where the cost of forming links is $c < b(1) - b(2)$. Clearly, all agents $i \in M_p$ are directly connected to each other, because if agents $1, 2 \in M_p$ are not connected in g then $mu_i(g^i + g_{12}, g_{12}) = b(1) - c - b(2) > 0$ for $i = 1, 2$ and therefore the agents would form the link. Now, let $i \in M_p$ be an NI agent, and let $j \in M_q$, $q \neq p$. By definition of an NI agent, i will assume that any link g_{jk} where $k \notin M_p$ is such that $g^i_{jk} = 0$. Thus, $mu_i(g^i + g_{ij}, g_{ij}) = b(1) - C < 0$ and NI agent i will not form any link with agents in a different community.

Second, consider a community M_p where the cost of forming links is $b(1) - b(2) < c < b(1) - b(3)$, and therefore *(a)* there are no agents i, j, k such that $g_{ij} = g_{jk} = g_{ik} = 1$. First, note that *(b)* all agents $i \in M_p$ are within a geodesic distance 2 of each other. This is because if agents $1, 2 \in M_p$ are such that $d_{12}(g) > 2$ then the minimum marginal utility they would gain by linking with each other is $mu_i(g^i + g_{12}, g_{12}) = b(1) - c - b(3) > 0$ for $i = 1, 2$ and therefore they would form the link. Now, proceed *ad absurdum*. Let $i \in M_p$ be an NI agent and suppose that $g_{ij} = 1$, ($j \in M_q$, $q \neq p$). For $mu_i(g^i, g_{ij}) < 0$ to hold, the following conditions must be true: *(c)* there must exist an agent 1 such that $g^i_{j1} = 1$ because $C > b(1) + b(2) - b(3)$, and *(d)* there is no other agent l such that $g_{1l} = 1$ and $g_{il} = 1$. Moreover, *(e)* $1 \in M_p$ because i is NI. By assumption, $|M_p| = m_p \geq 3$ and therefore there is another agent $k \in M_p$ that is connected to the network. Consider two cases.

(i) Suppose that $g_{kj} = 1$. For j not to remove the links with $i, k, 1$, these agents must provide some indirect benefits. Thus, assume there are agents $2 \in M_t$ and $3, 4 \in M_r$ such that $g_{i2} = g_{k3} = g_{14} = 1$. Now there are two cases to consider: $t = p$ and $t \neq p$.

First, $t = p$. In order to satisfy *(b)* it must be that $d_{21}(g), d_{2k}(g) \leq 2$. By *(d)* we cannot have that $g_{21} = 1$ and/or $g_{2k} = 1$. Thus, in order to have $d_{2k}(g) \leq 2$, there must be one or more agents q such that $g_{2q} = g_{qk} = 1$. However, if this were the case then $mu_j(g, g_{ji}) = C - b(1) - b(2) + b(3) > 0$. Similarly, in order to have $d_{21}(g) \leq 2$, there must be one or more agents q such that $g_{2q} = g_{q1} = 1$. However, if this were the case then $mu_j(g, g_{ji}) = C - b(1) - b(2) + b(3) > 0$.

Second, $t \neq p$. For $mu_i(g^i, g_{i2}) < 0$ to hold, the following conditions must be true: *(c')* there must exist an agent 5 such that $g^i_{25} = 1$, and *(d')* there is no other agent l such that $g_{7l} = 1$ and $g_{il} = 1$. Moreover, *(e')* $7 \in M_p$ because i is NI. By *(b)*, agents 7 and k must be at most at geodesic distance 2. There are three possibilities: $g_{73} = 1$, $g_{7j} = 1$ or $g_{7k} = 1$. In any of these three links exist then $mu_j(g, g_{ji}) \geq C - b(1) - b(2) + b(3) > 0$.[22]

Thus, if $g_{kj} = 1$ then there is no possible stable network with $g_{ij} = 1$.

(ii) Suppose that $g_{ik} = 1$. By *(b)* it must be that $d_{k1}(g) \leq 2$. By *(d)* we cannot have that $g_{k1} = 1$. Thus, it must be that there is an agent 2 such that $g_{k2} = g_{12} = 1$. However, if this were the case then $mu_j(g, g_{jk}) = C - b(1) - b(2) + b(3) > 0$. Thus, if $g_{ik} = 1$ then there is no possible stable network with $g_{ij} = 1$.

Thus, it must be that $g_{ij} = 0$, i.e. no NI agent i has any link with an agent j in another community. □

Proof of Proposition 1. Before proving the proposition, let us establish the following lemmata.

Lemma 3. *The utility defined in (1) is superadditive in own-links.*

Proof. By definition 3 I need to show that $mu_i(g, l_i(g)) \geq \sum_{g_{ij} \in l_i(g)} mu_i(g, g_{ij}) \; \forall \; i, g$ and $l_i(g) \subseteq L_i(g)$. It suffices to show the statement is true for two links $g_{ij}, g_{ik} \in g$, where $i \neq j \neq k$. Thus, the claim is that

$$mu_i(g, g_{ij} + g_{ik}) \geq mu_i(g, g_{ij}) + mu_i(g, g_{ik}) \tag{7}$$

Note that the distance $d_{ij}(g)$ between i and j in g is such that $d_{ij}(g) \leq d_{ij}(g - l_i(g))$ since any path from i to j in $g - l_i(g)$ is also present in g, but the opposite may not be true if the path goes through the deleted link $g_{il} \in l_i(g)$. Thus, it follows that for all agents $p \neq i$:

$$d_{ip}(g) \leq d_{ip}(g - g_{ij}) \leq d_{ip}(g - g_{ij} - g_{ik}) \tag{8}$$

$$d_{ip}(g) \leq d_{ip}(g - g_{ik}) \leq d_{ip}(g - g_{ij} - g_{ik}) \tag{9}$$

[22] Another possibility is that 7 is connected to a new agent 8 who is directly connected to k. However, in this case it is easy to iterate the argument in the preceding paragraph to show that if this were the case then $mu_j(g, g_{ji}) > 0$.

The marginal utility for i from deleting a link g_{ij} is the recovered cost of the link g_{ij} minus any benefit i cannot access anymore or for which it requires a longer path to have access to. For expositional convenience, define $B_{ip}(g, l_i(g)) = b(d_{ip}(g)) - b(d_{ip}(g - l_i(g)))$. By simple computations we get:

$$mu_i(g, g_{ij}) = \sum_{p \in N} [b(d_{ip}(g)) - b(d_{ip}(g - g_{ij}))] - c = \sum_{p \in N} B_{ip}(g, g_{ij}) - c$$

$$mu_i(g, g_{ik}) = \sum_{p \in N} [b(d_{ip}(g)) - b(d_{ip}(g - g_{ik}))] - c = \sum_{p \in N} B_{ip}(g, g_{ik}) - c$$

$$mu_i(g, g_{ij} + g_{ik}) = \sum_{p \in N} [b(d_{ip}(g)) - b(d_{ip}(g - g_{ij} - g_{ik}))] - 2c = \sum_{p \in N} B_{ip}(g, g_{ij} + g_{ik}) - 2c$$

Clearly, by definition if $B_{ip}(g, g_{ij}) \neq 0$ then $B_{ip}(g, g_{ij} + g_{ik}) \neq 0$. Similarly, if $B_{ip}(g, g_{ik}) \neq 0$ then $B_{ip}(g, g_{ij} + g_{ik}) \neq 0$. Furthermore, there is no p such that $B_{ip}(g, g_{ij}) \neq 0$ and $B_{ip}(g, g_{ik}) \neq 0$. To show this, suppose there exists such a p, then there are two distinct shortest paths $p_{ip}^{(1)}$ and $p_{ip}^{(2)}$ in g such that $|p_{ip}^{(1)}| = |p_{ip}^{(2)}|$, $g_{ij} \in p_{ip}^{(1)}$, $g_{ik} \in p_{ip}^{(2)}$, $g_{ij} \notin p_{ip}^{(2)}$ and $g_{ik} \notin p_{ip}^{(1)}$. But then $d_{ip}(g - g_{ij}) = d_{ip}(g)$ since $p_{ip}^{(2)}$ still exists and therefore $B_{ip}(g, g_{ij}) = 0$ which is a contradiction. Thus, there is no p such that $B_{ip}(g, g_{ij}) \neq 0$ and $B_{ip}(g, g_{ik}) \neq 0$.

But then for all p we have that

$$B_{ip}(g, g_{ij} + g_{ik}) \geq B_{ip}(g, g_{ij}) + B_{ip}(g, g_{ik}) \tag{10}$$

since $B_{ip}(g, g_{ij} + g_{ik}) \geq B_{ip}(g, g_{ij})$ from eq. (8) and $B_{ip}(g, g_{ij} + g_{ik}) \geq B_{ip}(g, g_{ik})$ from eq. (9) and there is no p such that $B_{ip}(g, g_{ij}) \neq 0$ and $B_{ip}(g, g_{ik}) \neq 0$. The inequality in (10) proves the claim in (7). □

Lemma 4. *If $u(.)$ is superadditive in own-links on $PS(u)$ then $PS(u) = PN(u)$.*

Proof: Theorem 1 in Calvó-Armengol and Ilkiliç [2006] proves that if $u(.)$ is α-convex on $PS(u)$ for some $\alpha \geq 0$ then $PS(u) = PN(u)$. Setting $\alpha = 1$, their definition of α-convexity reduces to the superadditivity condition in section 2. By lemma 3, $u(.)$ is superadditive, and this proves the statement. □

Lemma 5. *Consider a network formation game where the benefits from connections are heterogeneous, and vary in the $[\underline{b}(1), \overline{b}(1)]$ range. The pairwise Nash network architectures for the game are:*

(i) $c < \underline{b}(1) - \underline{b}(2)$ is a necessary and sufficient condition for the unique pairwise Nash network to be the complete community g^C

(ii) If $\overline{b}(1) - \overline{b}(2) < c < \underline{b}(1)$ then the star community g^* is a pairwise Nash network architecture.

(iii) If $c > \overline{b}(1)$ then the empty graph $g^{\emptyset}$ is a pairwise Nash network, and g^* is not a pairwise Nash network

Proof. By lemmas 3 and 4 it suffices to prove the above statements for pairwise stability. The proofs are then a straightforward check of the conditions for pairwise stability, and they are therefore omitted. □

Now, the **proof of Proposition 1**:

Proof. First of all, note that each community M_i is either a g^C or a g^{d2} architecture. If $c_i < b(1) - b(2)$ then M_i is a g^C network because if $1, 2 \in M_i$ are such that $g_{12} = 0$ in g then the minimum benefit to add the link g_{12} is $mu_i(g + g_{12}, g_{12}) = b(1) - c - b(2) > 0$ for $i = 1, 2$ so they will form the link. Similarly, if $b(1) - b(2) < c_i < b(1) - b(3)$ then M_i is a g^{d2} network because if $1, 2 \in M_i$ are such that $d_{12}(g) > 2$ then the minimum benefit to add the link g_{12} is $mu_i(g + g_{12}, g_{12}) = b(1) - c - b(3) > 0$ for $i = 1, 2$ so they will form the link.

By lemma 1 the only agents forming links across communities are the NC agents. Moreover, the benefits to connecting to an NC agent are in the $[\underline{B}(1), \overline{B}(1)]$ range. The maximum benefit to connecting to y_i is when M_i is a g^C network and all agents are directly connected to y_i, in that case the benefits are equal to $\overline{B}(1) \equiv b(1) + (m-1)b(2)$. The minimum benefit is when M_i is a g^* network and all, except for one, agents are two links away from y_i, in that case the benefits are equal to $\underline{B}(1) \equiv b(1) + b(2) + (m-2)b(3)$.

Let $G(\overline{B}(1), \underline{B}(1), C)$ be the game played by the kernel of NC agents, with homogeneous costs C and heterogeneous benefits in the $[\underline{B}(1), \overline{B}(1)]$ range. Now consider the network formation game $G_L(\overline{b}(1), \underline{b}(1), c)$ in lemma 5, with homogeneous costs c and heterogeneous benefits in the $[\underline{b}(1), \overline{b}(1)]$ range. By construction, $G_L(\overline{b}(1), \underline{b}(1), c)$ is isomorphic to $G(\overline{B}(1), \underline{B}(1), C)$. Thus, by lemma 5*(i)* and since $C < \underline{B}(1) - \underline{B}(2)$ the kernel of NC agents is a complete network.

Now it suffices to examine the cost structure of the communities to characterize the pairwise Nash equilibria:

(i) Since $\underline{c} \leq \overline{c} < b(1) - b(2)$ then all communities are g^C networks. Thus, the unique pairwise Nash network is a complete kernel with complete communities g^{CKC}.

(ii) There are $(1-\lambda)k$ communities M_i such that $c_i < b(1) - b(2)$, and therefore there are $(1-\lambda)k$ complete communities. The remaining λk communities M_j are such that $b(1) - b(2) < c_j < b(1) - b(3)$, and therefore they are g^{d2} networks. Thus, all pairwise Nash networks are formed by a complete kernel with mixed complete and d2 communities g^{CKM}.

(iii) Since $b(1) - b(2) < \underline{c} \leq \overline{c} < b(1) - b(3)$ then all communities are g^{d2} networks. Thus, all pairwise Nash networks are formed by a complete kernel with $d2$ communities g^{CKd2}. □

Proof of Corollary 1. The proof is exactly the same as the one of proposition 2 except for the last sentence of the third paragraph (see above) that should now read "by lemma 5*(ii)* and since $\overline{B}(1) - \overline{B}(2) < C < \underline{B}(1)$ the kernel of NC agents is a d2 network." □

Proof of Corollary 2. Consider each case separately:

(i) The payoffs for the NC agent are $u_i(g^{CKC}) = (m-1)[b(1)-c]+(k-1)[b(1)-C]+(k-1)b(2)$, while the payoffs for the NI agent are: $u_j(g^{CKC}) = (m-1)[b(1)-c]+(k-1)b(3)$. It is straightforward to show that $u_i(g^{CKC}) < u_j(g^{CKC})$ if $C > b(1)+b(2)-b(3)$.

(ii) The payoffs for the NC agent are $u_i(g^{CKSP}) = b(1)-c+(m-1)b(2)+(k-1)[b(1)+b(2)+(m-2)b(3)-C]$. In terms of position in the network structure there are two types of NI agents: at the center and at the periphery of the star community. The NI j agent at the center of the star community has payoffs: $u_j(g^{CKPS}) = (m-1)[b(1)-c]+(k-1)[b(2)+b(3)+(m-2)b(4)]$. The NI agent k at the periphery of the star community has payoffs: $u_k(g^{CKSP}) = b(1)-c+(m-1)b(2)+(k-1)[b(3)+b(4)+(m-2)b(5)]$. It is straightforward to verify that $u_i(g^{CKSP}) > \max\{u_j(g^{CKPS}), u_k(g^{CKSP})\}$. □

Proof of Proposition 2. Consider each part separately.

(i) By inspection of the g^{CK} networks, the maximum geodesic distance between any $k \in M_i$ and $l \in M_j$ is when $i \neq j$ and both k and l are NI agents in a $d2$ community who are not directly connected to the NC agent. Let $g \in g^{CK}$ be such that there are at least two communities M_i and M_j with such agents k and l. Let $x_i \in M_i$ and $x_j \in M_j$ be the agents that connect k and l respectively to the NC agents $y_i \in M_i$ and $y_j \in M_j$. Then the shortest path between k and l is $p_{kl} = \{g_{kx_i}, g_{x_iy_i}, g_{y_iy_j}, g_{y_jx_j}, g_{x_jl}\}$, so $Max\{D(g^{CK})\} = D(g) = |p_{kl}| = 5$.

(ii) By inspection of the g^{CK} networks, it is evident that $Min\{\bar{d}(g^{CK})\} = \bar{d}(g^{CKC})$, i.e. the architecture with the shortest average path length is the network with complete kernel and complete communities. The closed-form expression for $\bar{d}(g^{CKC})$ is:

$$Min\{\bar{d}(g^{CK})\} = \bar{d}(g^{CKC}) = \frac{1}{mk-1}[k(3m-2)+1-2m]$$

The random network g^{random} which is equivalent to g^{CKC} has average geodesic distance approximately equal to:

$$\bar{d}(g^{random}) \approx \frac{log(mk)}{log((m^2-m-1+k)/m)}$$

It is straightforward to verify that $\bar{d}(g^{CKC}), \bar{d}(g^{random}) \in (1,3)$ and $\bar{d}(g^{CKC}) \approx \bar{d}(g^{random})$ for any parameter values of m, k satisfying $mk = n < 10^5$.

Clearly, the average geodesic distance of a network $g^{CK}(\lambda)$ with a fraction λ of high cost communities is increasing in λ: the more high cost communities there are, the higher is the average geodesic distance because high cost communities are less connected than low cost ones. Thus, $\bar{d}(g^{CK}(\lambda)) \leq \bar{d}(g^{CKS}) = Max\{\bar{d}(g^{CK})\}$ for any λ, where g^{CKS} is the network with complete kernel and star communities with the NC agents at the periphery. The closed-form expression for $\bar{d}(g^{CKS})$ is:

$$Max\{\bar{d}(g^{CK})\} = \bar{d}(g^{CKS}) = \frac{1}{m(mk-1)}\left[mk(5m-6)-(3m^2-2m-2)\right]$$

The random network g^{random} which is equivalent to $g^{CK}(\lambda)$ has average geodesic distance approximately equal to:

$$\bar{d}(g^{random}) = \frac{log(mk)}{log((m+k-2+(1-\lambda)(m-1)^2+\lambda(m-1))/m)}$$

It is straightforward to verify that $\bar{d}(g^{CKS}), \bar{d}(g^{random}) < 5$ for any parameter values of m, k satisfying $mk = n < 10^5$.

(iii) First, note that the g^{d2} networks with the lowest clustering coefficient are star networks since $C(g^*) = 0$. Thus, in order to find a lower bound for the clustering coefficient of any g^{CK} network, assume that all g^{d2} communities are star networks. The following expression is the formula for the clustering coefficient of any g^{CK} network with $d2$ communities that are star networks.

$$Min\{C(g^{CK})\} = \left[\frac{(1-\lambda)[4+k(k-3)+m(m-3)]}{m(m+k-2)(m+k-3)} + (1-\lambda)\left(1-\frac{1}{m}\right)\right] + \tag{11}$$

$$+ \left[\frac{\mu(k-1)(k-2)}{m(m+k-2)(m+k-3)} + \frac{(k-2)(\lambda-\mu)}{mk}\right] \tag{12}$$

where μ (with $0 \leq \mu \leq \lambda$) is the proportion of communities that are a star community with the NC agent at the center.

Taking the limit of the above expression gives the result: $lim_{m,k\to\infty}[Min\{C(g^{CK})\}] = 1-\lambda$. Moreover, if m is finite then $lim_{k\to\infty}C(g^{CK}) = (1-\lambda) + \frac{\lambda}{m} > 1-\lambda$. Viceversa, if k is finite then $lim_{m\to\infty}C(g^{CK}) = 1-\lambda$. □

Proof of Proposition 3. Statement *(ii)* is a direct consequence of lemma 1. Let us prove *(i)* and *(iii)*.

(i) By lemma 1 NI agents have segregation index equal to one. Each NC agent $y_p \in M_p$ is connected to at least another agent $i \in M_p$ so the minimum segregation index for an NC agent is $\frac{1}{1+k-1} = \frac{1}{k}$. The minimum segregation index of M_p is therefore:

$$S_p = \frac{1}{m}\left(\underbrace{1+...+1}_{m-1} + \frac{1}{k}\right) = 1 - \frac{k-1}{mk}$$

To obtain the lower bound on the minimum segregation index, let $m_p = 3$ so that there is the minimum possible number of NI agents:

$$S_p = \frac{1}{3}\left(1+1+\frac{1}{k}\right) > \frac{2}{3}$$

(iii) Let $i \in M_i$ be the NI agent with the highest betweenness centrality $I_B(\eta_i)$ among all NI agents in M_i, i.e. $\eta_i[p_{kl}(g)] > \eta_j[p_{kl}(g)]$, where $j \neq i, y_i$, $k \in M_p, l \in M_q, q \neq p$. Consider the NC agent $y_i \in M_i$. Clearly, in any g^{CK} network all the paths from k to l that include i have to include y_i as well because that is the only agent in M_i that forms connections with agents in other communities. Thus, $\eta_{y_i}[p_{kl}(g)] \geq \eta_i[p_{kl}(g)]$. Moreover, y_i is also on the geodesic paths that connect i to agents in other communities, and therefore $\eta_{y_i}[p_{kl}(g)] > \eta_i[p_{kl}(g)]$. Thus, $I_B(\eta_{y_i}) > I_B(\eta_j)$, $\forall j \in M_i, j \neq y_i$. □

Proof of Proposition 4. First, note that lemmas 1 and 5 apply here as well. Also, as in proposition 1, each community M_i is either a g^C or a g^{d2} network. The benefits for y_i (or y_j) from the *first* link $g_{y_i y_j}$ ($y_i \in P_i$, $y_j \in P_j$) between subsets $P_i \in M_i$ and $P_j \in M_j$ are in the $[\underline{B}, \overline{B}]$ range. The maximum benefit is when M_j (or M_i) is a g^C network. The minimum benefit is when M_j (or M_i) is a g^* network and all NC agents are at the periphery of the star.

Let $G(\overline{B}(1), \underline{B}(1), C)$ be the game played by the p-kernel K_P, with homogeneous costs C and heterogeneous benefits in the $[\underline{B}(1), \overline{B}(1)]$ range. Now consider the network formation game $G_L(\overline{b}(1), \underline{b}(1), c)$ in lemma 5, with homogeneous costs c and heterogeneous benefits in the $[\underline{b}(1), \overline{b}(1)]$ range. By construction, $G_L(\overline{b}(1), \underline{b}(1), c)$ is isomorphic to $G(\overline{B}(1), \underline{B}(1), C)$. Thus, by lemma 5*(i)* and since $C < \underline{B}(1) - \underline{B}(2)$ the p-kernel is a complete network.

The last step to characterize the pairwise Nash equilibria is to examine the cost structure of the communities. This is the same as points *(i), (ii), (iii)* in the proof of proposition 1 and it is not repeated here. □

Proof of Proposition 5. First, note that $g^{CK} \subseteq g^{CpK}$ since g^{CK} architectures are the special case $p_i = 1$, $\forall i = 1, ..., k$. Second, note that for any $g \in g^{CpK}$ there exists a $g' \in g^{CK}$ such that g has the same community structure of g' plus some additional intra-community and inter-community links. The additional links come from the additional NC agents who form additional inter-community links and, possibly, additional intra-community links since they become more valuable due to the indirect benefits they bring. Now, consider each case separately.

(i) The proof is the same as the proof of *(i)* in proposition 2.

(ii) The proof is very similar to the proof in *(ii)* in proposition 2 and it is therefore omitted.

(iii) The clustering coefficient for g^{CK} networks in the limit $m, k \to \infty$ is determined by the clustering coefficient of NI individuals in complete communities which have clustering coefficient equal to one. NI agents in complete communities in g^{CpK} architectures will clearly also have clustering coefficient equal to one. Following a similar argument to the proof of *(iii)* in proposition 2, we have that:

$$lim_{m,k\to\infty}[Min\{C(g^{CpK})\}] \approx 1 - \lambda\left(1 - \frac{p^*_{max}}{m}\right) \tag{13}$$

where p^*_{max} is the maximum number of NC agents in any community M_i with low cost of link formation $c_i < b(1) - b(2)$. If $p^*_{max} \nrightarrow \infty$ then $lim_{m,k\to\infty}[Min\{C(g^{CpK})\}] = 1 - \lambda$ as in *(iii)* in proposition 2.

The right-hand-side of equation (13) comes from the second term in equation (11) where clearly $p^*_{max} = 1$ for g^{CK} networks. Note that the equivalent of the expression in (11) will be much more complicated for g^{CpK} networks since the clustering coefficients of several nodes will be affected by the additional links. However, the clustering coefficient of all these nodes will go to zero in the limit so it is not necessary to derive the exact formula for $C(g^{CpK})$ in order to obtain the result in (13) above. □

Proof of Proposition 6. Statement *(ii)* is a direct consequence of lemma 1. Let us prove *(i)* and *(iii)*.

(i) By the definition of segregation index, a community M_q with the lowest S_q is composed by agents with the lowest possible segregation index. By lemma 1, NI agents have segregation index equal to 1. To minimize S_q, let $p_q = \frac{m}{2}$, i.e. half of agents in M_q are NC. Each NC agent $y_p \in M_p$ is connected to at least another agent $i \in M_p$ so the minimum segregation index for an NC agent is $\frac{1}{1+k-1} = \frac{1}{k}$. The minimum segregation index of M_p is therefore:

$$S_p = \frac{1}{m}\left(\underbrace{1+...+1}_{\frac{m}{2}} + \underbrace{\frac{1}{k}+...+\frac{1}{k}}_{\frac{m}{2}}\right) > \frac{1}{2}$$

(iii) Consider the agents in the M_p community. There are three types of paths connecting any $j \in M_q$ $(q \neq p)$ to an agent in $i \in M_p$: *(i)* a path with no intermediary agent $k \in M_p$ if i is an NC agent such that $i \in K$; *(ii)* a path with one intermediary NC agent k if $g_{ik} = 1$ and $k \in K$; *(iii)* a path with two intermediary agents $k \in K$ and $l \notin K$ if $g_{il} = g_{kl} = 1$ and $g_{ik} = 0$, with k and l an NC and an NI agent respectively. Note that *(i)-(iii)* exhausts all possibilities because all networks have segregation patterns and the maximum geodesic distance between two agents in the same community is two. Also note that each type of path always exists for each community M_p except for *(iii)*, which exists only if M_p is a *d2* network. Clearly, paths like *(i)* do not contribute to the betweenness centrality of any of the agents in M_p. Paths like *(iii)* contribute equally to the betweenness centrality of NC and NI agents in M_p. Finally, paths like *(ii)* contribute only to the betweenness centrality of NC agents. Thus, $\sum_{i\in M_p, i\in K} I_B(\eta_i) > \sum_{j\in M_p, j\notin K} I_B(\eta_j)$. □

Proof of Proposition 7. In the statement and proof of lemma 1, replace $C > b(1) + b(2) - b(3)$ with $C > EB_{max}$. The proof of the lemma then follows unchanged. The proof of this proposition is then the same as the proof of proposition 1. □

Proof of Lemma 2. For the proof of statement *(i)* please refer to Jackson and Rogers [2005] (proposition 2, page 624).

(ii) It is easy to see that any pairwise Nash network g is such that $g_{ij} = 1$ if i and j belong to the same community: suppose not, then $mu(g + g_{ij}, g_{ij}) = b(1) - c - b(2) > 0$ and g is not pairwise Nash. Now consider a network g with complete communities and q (with $1 \leq q < m$) inter-community links such that each agent i is involved in no more than one inter-community link. Let $a, b \in U_1$ and $c, d \in U_2$ be such that $g_{ac} = 1$ and $g_{bd} = 0$, then for g to be pairwise Nash we must have:

$$\begin{aligned} mu(g + g_{bd}, g_{bd}) = -C + b(1) - b(3) + (m - q - 1)[b(2) - b(3)] \equiv -C + P_L(q) &< 0 \\ mu(g, g_{ac}) = C - b(1) + b(3) + (m - q - 1)[-b(2) + b(3)] \equiv C - P_L(q) - b(2) + b(3) &< 0 \end{aligned}$$

Thus, if $P_L(q) < C < P_L(q) + b(2) - b(3)$ then g is pairwise Nash. It is clear that g is the unique pairwise Nash architecture: in any pairwise Nash network all agents within a community are connected and if there were a number of inter-community links different than q then the expressions above show that the network would not be pairwise Nash.

By simple substitution and computation, it is clear that $P_U(p^*) < E_L(p^*)$ if $1 \leq p^* < m-1$ and $P_U(p^*) = E_L(p^*)$ if $p^* = m-1$. □

Proof of Proposition 8. Consider each statement separately.

(i) Let g_r be the pairwise Nash network when there are r NC agents, and let g_s be the pairwise Nash network when there are $s < r$ agents. Suppose that $r - s$ NI agents in g_s are turned into NC agents. There are two cases:

- g_r and g_s have the same number of inter-community links, i.e. they have the same network architecture given that they both have complete communities. Then clearly after turning $r-s$ NI into NC agents there is no link formation/removal in g_s because g_s is the same as g_r which by lemma 2 is the unique pairwise Nash network with r NC agents. Thus, $V(g_s) = V(g_r)$.

- g_r has more inter-community links compared to g_s. If each community in g_s has at least one "new" NC agent then new inter-community link(s) will form because they have positive marginal utility for the agents involved. Otherwise if all "new" NC agents are in one community then there is no link formation because forming a link requires bilateral consent and nothing changed in one of the communities from the g_s pairwise Nash equilibrium. Clearly, there is no link removal either. Thus, $V(g_s) \leq V(g_r)$.

Thus, turning $r-s$ NI into NC agents leads to a (weakly) higher welfare for the whole network.

(ii) Fix the cost C of inter-community links. By lemma 2 the pairwise Nash network when all agents are NC is two complete communities with q links across communities. Let $p^* \equiv q$, and therefore consider a network with q NC agents in each community. Clearly, the pairwise Nash network has complete communities connected by q links across the two communities and it is the same network as the case where all agents are NC. Thus, the total welfare in both cases will be the same. □

References

A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286 (5439):509–512, 1999.

P. M. Blau. *Inequality and Heterogeneity: A Primitive Theory of Social Structure*. Free Press, New York, 1977.

F. Bloch and M. O. Jackson. The formation of networks with transfers among players. *Journal of Economic Theory,* forthcoming, 2006a.

F. Bloch and M. O. Jackson. Definitions of equilibrium in network formation games. *International Journal of Game Theory*, 34(3):305–318, 2006b.

B. Bollobás. *Random Graphs*. Academic Press, New York, 2001.

D. Bondonio. Predictors of accuracy in perceiving informal social networks. *Social Networks*, 20:301–330, 1998.

R. S. Burt. *Structural Holes: The Social Structure of Competition*. Harvard University Press, 1992.

R. S. Burt. Structural holes and good ideas. *American Journal of Sociology*, 110(2): 349–399, 2004.

A. Calvó-Armengol and R. Ilkiliç. Pairwise-stability and Nash equilibria in network formation. Working Paper, March 2006.

T. Casciaro. Seeing things clearly: Social structure, personality, and accuracy in social network perception. *Social Networks*, 20:331–351, 1998.

R. L. Cross and A. Parker. *The Hidden Power of Social Networks: Understanding How Work Really Gets Done in Organizations*. Harvard Business School Press, 2004.

R. L. Cross and R. J. Thomas. *Driving Results Through Social Networks: How Top Organizations Leverage Networks for Performance and Growth*. Jossey-Bass, 2009.

S. Currarini, M. O. Jackson, and P. Pin. An economic model of friendship: Homophily, minorities and segregation. *Econometrica,*forthcoming, 2008.

S. N. Dorogovtsev, J. F. F. Mendes, and A. N. Samukhin. Structure of growing networks with preferential linking. *Physical Review Letters*, 85:4633–4636, 2000.

S. N. Dorogovtsev, A. V. Goltsev, and J. F. F. Mendes. Pseudofractal scale-free web. *Physical Review E*, 65(066122):1–4, 2002.

P. Erdős and A. Rényi. On random graphs. *Publicationes Mathematicae*, 6:290–297, 1959.

P. Erdős and A. Rényi. On the evolution of random graphs. *Publications of the Mathematical Institute of the Hungarian Academy of Sciences*, 5:17–61, 1960.

L. C. Freeman. A set of measures of centrality based on betweenness. *Sociometry*, 40: 35–41, 1977.

L. C. Freeman. Filling in the blanks: A theory of cognitive categories and the structure of social affiliation. *Social Psychology Quarterly*, 55(2):118–127, 1992.

A. Galeotti, S. Goyal, and J. Kamphorst. Network formation with heterogeneous players. *Games and Economic Behavior*, 54:353–372, 2006.

M. Girvan and M. E. Newman. Community structure in social and biological networks. *PNAS*, 99(12):7821–7826, 2002.

S. Goyal, M. J. van der Leij, and J. L. Moraga-Gonzalez. Economics: An emerging small world. *Journal of Political Economy*, 114(2):403–412, 2006.

M. Granovetter. *Getting A Job: A Study of Contacts and Careers*. The University of Chicago Press, 1995.

D. Hojman and A. Szeidl. Core and periphery in networks. *Journal of Economic Theory*, 139(1):295–309, 2008.

M. O. Jackson and B. Rogers. The economics of small worlds. *Journal of the European Economic Association*, 3(2-3):617–627, 2005.

M. O. Jackson and B. Rogers. Meeting strangers and friends of friends: How random are social networks? *American Economic Review*, 97(3):890–915, 2007.

M. O. Jackson and A. Wolinsky. A strategic model of social and economic networks. *Journal of Economic Theory*, 71:44–74, 1996.

G. A. Janicick and R. P. Larrick. Social network schemas and the learning of incomplete networks. *Journal of Personality and Social Psychology*, 88(2):348–364, 2005.

G. Kossinets and D. J. Watts. Empirical analysis of an evolving social network. *Science*, 311:88–90, 2006.

D. Krackhardt. Cognitive social structures. *Social Networks*, 9:109–134, 1987.

D. Krackhardt. Assessing the political landscape: Structure, cognition, and power in organizations. *Administrative Science Quarterly*, 35(2):342–369, 1990.

D. Krackhardt and M. Kilduff. Whether close or far: Social distance effects on perceived balance in friendship networks. *Journal of Personality and Social Psychology*, 76(5): 770–782, 1999.

E. Kumbasar, K. A. Romney, and W. H. Batchelder. Systematic biases in social perception. *American Journal of Sociology*, 100(2):477–505, 1994.

M. McBride. Imperfect monitoring in communication networks. *Journal of Economic Theory*, 126:97–119, 2006.

M. McBride. Position-specific information in social networks: Are you connected? *Mathematical Social Sciences*, 56:283–295, 2008.

S. Milgram. The small world problem. *Psychology Today*, 2:60–67, 1967.

R. B. Myerson. *Game Theory: Analysis of Conflict*. Harvard University Press, 1991.

M. E. J. Newman. Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality. *Physical Review E*, 64, 2001.

M. E. J. Newman. The structure and function of complex networks. *SIAM Review*, 45: 167–256, 2003.

M. E. J. Newman, A.-L. Barabási, and D. J. Watts. *The Structure and Dynamics of Networks*. Princeton University Press, 2006.

D. J. Watts. *Small Worlds: The Dynamics of Networks between Order and Randomness*. Princeton University Press, 1999.

D. J. Watts and S. H. Strogatz. Collective dynamics of "small-world" networks. *Science*, 393(6684):440–442, 1998.