The Role and Evolution of Central Authorities*

Paul Frijters† and Alexander F. Tieman‡

July 14, 1999

Abstract

In this paper we argue that authorities aid cooperation by means of direct coordination or the enforcement of pre-commitment devices such as contract laws.Credible threats of violence allow this role. In a local interaction model, an authority forms if mutually connected individuals with sufficient combined punishment potential have signalled their willingness to form such an authority, conditional upon the willigness of others to do so. Given a specific timing of decisions, we analyse the conditions under which authorities arise and under which they evolve into a stationary state with only one or several remaining authorities.

Keywords: *Central Authorities, Cooperation, Evolution, Externalities, Local Interaction.*

JEL-code: B25, C7, D62, D70, H1, H4.

---

*The authors would like to thank Gerard van der Laan and Frans van Winden for valuable comments upon an earlier draft of the paper.

†Corresponding Author. Department of Economics, Free University, De Boelelaan 1105, 1081 HV Amsterdam, The Netherlands. E-mail: pfrijters@econ.vu.nl. Ph: +31-20-4446155.

‡Tinbergen Institute and Department of Econometrics, Free University, Amsterdam, NL. URL: http://www.econ.vu.nl/medewerkers/xtieman, E-mail: xtieman@econ.vu.nl, Ph: +31-20-4446022, Fax: +31-20-4446020.

1. Introduction

Central authorities both coordinate some actions directly and facilitate coordination between individuals. As examples of direct coordination by authorities we can think of the provision of some public goods such as an army, a pasport administration, and some infrastructural projects. Besides directly coordinating action, a central authority makes pre-commitment possible in interactions between agents and hence allows individuals to coordinate and reach higher expected payoff strategies. Central authorities can for instance enforce contract laws and criminal laws. That coordination often involves some sort of authority seems uncontroversial: individuals with similar interests often set up an embryonic authority, for instance when individuals choose committees for some issue they want to raise, form or choose political parties, set up (military) headquarters, set up a board of directors, set up a police force, a union, or vigilante groups, etc. We use the term central authority (c.a.) in a broad sense in order to capture all the above examples. We consider the question what the defining features are of a central authority and how such authorities arise evolutionary in a local interaction setting.

Since Weber (1922), a defining feature of a central authority is that it monopolizes violence. Through this monopoly it simply punishes the perpetrators of rules and thereby allows coordination to take place peacefully (cf. Aumann (1989)). How does a central authority mobilize many individuals in order to punish a single individual that does not comply with the rules set by the c.a. and why do other (groups of) individuals not have this mobilizing potential? We argue that a central authority can directly communicate with each individual in the authority because it has an increasing returns to scale ad-

vantage in gathering and spreading information. A direct line of communication with all individuals allows it to control expectation formation on the side of individuals, which allows it to establish a monopoly of violence. Consider the benefits of having all individuals communicate with a single place in stead of having all individuals communicate with each other: when N persons are all capable of exchanging information directly with everyone else, a central authority requires $2N$ informational exchanges in order to make all information available to everyone: from each individual to one central authority, and feedback to each individual. In the absence of a central authority, it would require $N(N-1)$ informational exchanges for each individual to know the interests of every other individual and would require N times (i.e. for each individual) the processing costs of calculations for any decision to be reached collectively. Although we do not model information costs, we do implicitly assume that they lead to the monopoly of violence.

In our evolutionary model, central authorities form when neighbouring individuals notice that they advocate a non-conflicting action in an economic stage game: in the beginning each individual acts according to his own highest payoff action. Noticing negative effects of the actions of others, individuals start advocating to the individuals they interact with that they are willing to play another action if (some of) the other players are also willing to play another action. If individuals who advocate non-conflicting actions are neighbours and there are enough connected individuals to force any individual to play a different action, they form a coalition, whereby each neighbour who also advocates a non-conflicting action will join. Such a coalition starts enforcing an internal

discipline and starts expanding by forcing non-members to comply. We call such a coalition a central authority. Once at least one central authority has emerged, after a certain period, all individuals are members of a central authority, and depending on the rules of engagement, either one or many central authorities remain. The central authority itself is then a highly stylised a-personal entity in which the interests of all individuals belonging to the authority have equal weight for the maximand of the rules of the authority. As is discussed in the concluding remarks, we could have personalised the central authority by taking the a sub-group of individuals within the bounds of the authority to control the decision making of the central authority, in which case the personal interests of this sub-group determine the rules enforced in the whole authority. Because this would distract from the central mechanisms in the model though, this paper models the central authority to be free of special interest group considerations.

In section 2, we present a short survey of the literature on the evolution of cooperation and the available historical and anthropological evidence on the formation of early authorities. In section 3, we build a descriptive model in which an authority is defined and where the rules governing the evolution of central authorities are laid down. In Section 4 the conditions under which authorities arise and the outcome of the interaction between several authorities are derived. In Section 5 a particular departure from the model in the previous sections is examined under which multiple central authorities may arise as a stable steady state outcome. The implications of allowing for random mutations for this case are discussed. Section 6 concludes.

2. Literature

How do individuals solve coordination problems in prisoners' dilemmas or public good games? One possibility is that individuals use communication and conventions without physical institutions. If individual's interests do not conflict, Potters and van Winden (1996), Austen-Smith (1994) and Farrell and Rabin (1996) all show how sequences of communication may lead to coordination. Kahneman, Knetsch, and Thaler (1986) discuss how individuals develop notions of fairness, which guide them in particular situations of human interaction with conflicting interests. 'Fairness' is then a social institution not enforced by an authority. Similarly, some authors argue that coordination eventually takes place in repeated prisoner dilemma's through building a credible reputation for punishing deviation of other players (see e.g. Osborne and Rubinstein (1994) and references therein). In a similar vein, Axelrod (1987), Eshel, Samuelson, and Shaked (1998), Tieman, Van der Laan, and Houba (1997), Tieman, Houba, and Van der Laan (1998) and Karandikar, Mookherjee, Ray, and Vega-Redondo (1998) have used evolutionary arguments to show how cooperation and social norms may develop between individuals, either when individuals are boundedly rational or follow some behavioral rule.

Informal means of coordination can be wittnessed when individuals tip waitresses, adhere to notions of fairness, trust one another on their word, believe statements made by strangers, etc. In such instances, there is no formal penalty for defaulting but only implicit threats of future isolation for the perpetrators of the implicit conventions. In many instances however, coordination is achieved by a credible threat of violence in case an individual does not comply with a set of announced rules. Tax authorities

do not merely trust on generosity when they expect individuals to pay taxes, but also credibly threaten non-payers with punishment. Goverments usually do not merely rely on patriottic fervour when they expect soldiers to face mortal dangers, but have rules allowing disobedient soldiers to be shot on site in times of crisis. Less extreme examples abound.

Institutions, and more particularly authorities, therefore also solve coordination problems by use of credible pre-announced sanctions to non-cooperators. On a micro-level, it seems from anthropological studies that already in early human societies such as hunter-gatherer societies or early agricultural societies, there were coordination structures ('big men', councils, tribe leaders, etc.) where individuals made a conscious effort to solve the coordination problems of group through institutions (see Harris (1993) and Wenke (1984)). On an intermediate level, one can think of labour market institutions, such as union or cartels which solve coordination problems on the labour markets. Based on historical study, Heap (1994), Hoel (1990), and Soskice (1990) have argued that institutions are the main way in which individuals deal with coordination problems arising in labor markets or other forms of human interaction. Though often starting as voluntary organisations, unions and cartels seem to have often relied on credible sanctions in order to ensure complience with their rules. On a more macro level, central government is the most powerful of the institutions enforcing rules by violence. Weber (1922), North (1981, 1990) and Eggertson (1990) describe in detail how authorities arose in Europe, how they solved some coordination problems, what their limitations were due to their internal structure, and how the authorities changed. These studies start their in-depth

analyses of the origins of current states from quite organised pre-state authorities, such as kingdoms, powerful cities, chiefdoms, and the like. Being most interested in the evolution of these state-like institutions, the question arises how these state-like authorities arose in the first place.

Written records usually only appeared long after an authority had arisen. Therefore, only vague indications are available of what might have happened in the early years of authorities (cf. Harris (1993, pg. 165), Wittfogel (1963, pg. 21)). The current 'best-guess' is that authorities arose in densely populated areas with intensive sedentary agriculture as a response to coordination problems involving the use of water (cf. Harris (1993), Wenke (1984) and Postgate (1992)). Wittfogel (1963), whose massive study incorporates an account of the rise of dozens of known civilisations, depicts a stylised initial situation without an authority as one in which 'protofarmers' are each tied on their own patch of land, unaware of many possibilities to increase their production by coordination. Being tied to a patch of land of which they cannot substantially change the soil and given a climate they cannot influence, the one element for which cooperation makes sence is the control of water. If a group of individuals could poole their labour in order to make a well, or build a reservoir, or dam a river, they could increase their yields. When a group of protofarmers recognise such an opportunity, Wittfogel (1963) argues 'They must work in cooperation with their fellows and subordinate themselves to a directing authority' (pg 18). This embryonic authority then spreads as it, either through force or persuasion, invloves the labour of others and expands the system of irrigation or the system of dams it build. These systems are then themselves a means of

establishing lines of communication with all parts of the authority. As to the eventual outcome, Wittfogel remarks that 'The pioneers of hydraulic agriculture, like the pioneers of rainfall farming, were unaware of the ultimate consequences of their choice. Persuing recognised advantage, they initiated an institutional development which led far beyond their starting point' (pg. 19). It is this account of how authorities arise that we will model formally and expand on.

Sticking as close as possible to the account above, individuals will hence be modelled to be fixed in a (social) space and are taken to have limited information about non-neighbours.[1]

Before presenting the model, there is one aspect of central authorities which makes the technical analysis non-standard: coordination by an authority is deliberate. Authorities spend a great deal of time and effort in thinking about and revising rules which are designed to be 'optimal' in some sense. Also, authorities try to overcome the informational and physical constraints that bound the behavior of individuals. If an authority is assumed to be in possession of greater abilities than individuals however, these extra abilities have to be explicitly modelled. Furthermore, the strategic interactions between authorities, individuals, and other authorities lead to many complications. As a result, the model developed in the next section has a lot of structure and is more a descriptive model of how authorities might have arisen, rather than an analytical model in which results are derived from a few first principles.

[1]Most importantly, this is true in a setting with little trade: most individuals in these communities will have been geographically tied to one plot of land and will have had little interaction with anyone outside his own community.

3. The Model

We present the specific parts of the model in the following order. First, we present the fixed (social, geographic or product) space in which individuals are situated as a graph and introduce all relevant notation. Second, we introduce the possibility of violence and discuss the threatening and punishment of players. Then, we focus how individuals learn, what their pay-off structure is, followed by a detailed desciption of the sequence of events during a time period. We end the section with a preliminary result on the formation of central authorities.

Each of the N (N large) vertices of the finite graph K is the address of one player. Every vertex $s \in K$ is directly connected to a (finite) number of other vertices by the edges of the graph K. No vertex is connected to itself, i.e. the graph K is irreflexive. We assume that all vertices have the same number $0 < m << N$ of edges, i.e. that all vertices are connected to exactly m other vertices. Further, we assume that the graph K is connected, i.e. that from any vertex any other vertex is reachable in a finite number of steps through intermediary vertices. The set of vertices directly connected to vertex s is the neighbourhood of s. This nonempty set is called V_s. For any set of vertices $R \subset K$, the boundary of R is the set $\partial R = \{\cup_{r \in R} V_r\} \setminus R$. This neighbourhood relation is symmetric: If r is a neighbour of s, the s is a neighbour of r. AS a result, there is an edge connecting s and r only if $r \in V_s$ and then also $s \in V_r$. Players are referred to by their address: player s is the player at vertex s, $s = 1, 2, \ldots, N$. A k-clique C_k is a set of k players in K which are all mutual neighbours, i.e. for all $s, r \in C_k$ it holds that $r \in V_s$ and then also that $s \in V_r$.

Objects of choice for all players are actions in the set $\Gamma = \{A, B\}$. The configuration of the population at time t is a function $\phi^t : K \to \Gamma^N$. A configuration describes the action choices of the player population: $\phi^t(s)$ is the action employed by player s at time t. Initially (at $t = 0$) each player is assigned an action at random with probability $\frac{1}{2}$ on each of the actions A and B, i.e. each possible configuration $\phi^0 \in \{A, B\}^N$ has probability $\frac{1}{2^N}$ of being selected as initial configuration. For any $R \subset K$ and configuration η, $\eta(R)$ denotes the restriction of η to R. For any set $R \subset K$ of players, let $X(R)$ denote the set of configurations of vertices in R. Let $\phi(-s)$ denote the configuration $\phi(K \setminus \{s\})$, and similarly $\phi(-R)$ denotes the configuration $\phi(K \setminus R)$. For a given configuration ϕ^t and action a, let $\phi^t_{s,a}$ denote the configuration identical to ϕ^t except that player s is using action a. Finally, for a given configuration ϕ^t and action a, let ϕ^t_a denote the number of players playing action a, $a = A, B$, i.e., $\phi^t_a = \sum_{s \in K} I\left(\phi^t(s) = a\right)$, where

$$I\left(\phi^t(s) = a\right) = \begin{cases} 1, & \text{if } \phi^t(s) = a, \\ 0, & \text{otherwise} \end{cases}$$

and let $\phi^t_a(R)$ be the number of players in the set R playing action a under ϕ^t, i.e., $\phi^t_a(R) = \sum_{s \in R} I\left(\phi^t(s) = a\right)$.

At each round of play $t = 0, 1, 2, \ldots$, each player s signals to each of its neighbours $r \in V_s$ a conditional strategy $\psi^t(s, r)$ and a punishment $\nu^t(s, r)$. A conditional strategy $\psi^t(s, r) \in \Gamma$ consists of an action player s would be willing to play, conditional on his neighbours also playing this action. The punishment $\nu^t(s, r) \geq 0$ is an amount with which player s threatens to reduce the payoff of player r through punishment, if

this player does not comply with certain demands of player s, which will be specified below. Each player s is equipped with a maximum punishment potential $\theta(s)$, which is constant over time. The values $\theta(s)$, $s = 1, \ldots, N$, are independent realizations of a random variable θ with distribution Θ, which has its support on a subset of $[0, \theta^{\max}]$. We label the distribution function of Θ by $f(.)$. In order to be credible, a player cannot threaten any of his neigbors with more than his punishment potential, i.e.$\nu^t(s,r) \leq \theta(s)$, $\forall r \in V_s$, $\forall t$. We assume that violence hurts the person being punished more than it costs the punisher, i.e. having an effect of -1 on the payoff of a neighbour through punishment costs the punisher $\frac{1}{c} < 1$, $c > 1$.

In each round of play t, $t = 1, 2, \ldots$, with probability $p \in (\underline{p}, 1)$, $\underline{p} > 0$, each individual gets the possibility to update his action and conditional strategy, a so-called *learning draw*. All players that get a learning draw choose an action and choose a signal to send out to their neighbours.

A set of players $R \subset K$ forms a central authority if all players $r \in R$ have agreed to voluntarily join the authority at some time in the past. Different central authorities are disjunct, i.e. players can not be member of two central authorities at the same time. We denote the union of all central authorities present on the graph at time t by W^t, i.e. $W^t = \cup_{\text{all c.a.s } R} R$. A central authority (c.a.) is referred to by the set of players R it encompasses. The defining feature of a central authority is that it can communicate directly with all its members and has the added ability of transferring the punishment potential of any individual who agrees to this to any member.[2] Hence, whereas an

[2]In essence this assumes free transport of punishment potential within the borders of the authority, whereas individuals do not allow such transports if they do not belong to an authority. The reason for

individual $r \in R$ can only punish its neighbours in V_r, a central authority R can direct the combined punishment potential of its members, $\sum_{r \in R} \theta(r)$, to any of the members $r \in R$ of the central authority or to any of the players on the boundary ∂R of R.

A central authority R advocates a, possibly empty, set of rules to all players in $\cup_{r \in R} V_r = R \cup \partial R$. A set of rules prescribes actions to individual players and contains a punishment scheme for players that do not follow the prescribed action, when they could have done so. An empty set of rules is interpreted as absence of prescribed rules and therefore as absence of punishment, whatever the player chooses to do. As to the method by which rules are chosen, we follow Rawls (1971), by assuming that each period all members of a central authority are able to choose a set of rules under a complete veil of ignorance, i.e., with each proposed set of rules all individuals know the distribution of expected utilities next periods but not which utility is theirs. Following Harsanyi (1985), this means that the chosen set of rules will maximize the combined total expected payoff of the current members.[3] The central authority then makes these rules common knowledge within the authority and to the players on its borders. In this sense, the central authority is no more than a strategy selection device with an information advantage and the ability to transfer punishment potential on its territory.

At any time players can indicate they want to form or join a central authority. However, they will only do so if this seems profitable for them at the time of joining.

this is that allowing free transport means putting oneself in a vulnerable position, which one will only do if a central authority ensures no disadvantage is taken of this vulnerability.

[3]This way of choosing a set of rules is rather crude. It essentially assumes that there is an "honest broker", such as a computer, which, given the combined knowledge of all constituents, computes the expected utility of each possible set of rules, after which the constituents choose one. Obviously, the social choice literature discusses many other different rule-choosing mechanisms (see Pardo and Schneider (1996) for a review) which could be pursued in future work.

Thus, a group of players will form a new c.a. if they foresee profit from this and they know that the other potential members of the c.a. also foresee this, i.e., that they will also join.

The total payoff $\Pi\left(s,\phi^t\right)$ to player s at time t consist of his economic payoff from the stage game $\pi\left(s,\phi^t\right)$ from which the punishment which is administered to him and the costs of punishing others are deducted, i.e.

$$\Pi\left(s,\phi^t\right)=\pi\left(s,\phi^t\right)-\sum_{r\in V_s}\nu^t\left(r,s\right)I_{r,s}-\frac{1}{c}\sum_{r\in V_s}\nu^t\left(s,r\right)I_{s,r},$$

where

$$I_{r,s}=\begin{cases}1, & \text{when player } s \text{ does not comply with the conditions set by player } r,\\ 0, & \text{otherwise}\end{cases}.$$

The payoff $\pi\left(s,\phi^t\right)$ from the stage game depends on the entire configuration ϕ^t through distant and local externalities. An action $\phi^t\left(s\right)=A$ yields a direct payoff $\alpha>0$ to player s, but it imposes a negative externality $-\lambda<0$ upon all players $r\in V_s$, while it yields an externality μ to all players $r\in K\backslash\{s\}$. An action $\phi^t\left(s\right)=B$ results in a payoff $0<\beta<\alpha$ to player s and yields no externalities. Thus we have that

$$\pi\left(s,\phi^t\right)=\begin{cases}\alpha+\left(\phi_A^t-1\right)\mu-\phi_A^t\left(V_s\right)\lambda, & \text{if } \phi^t\left(s\right)=A,\\ \beta+\phi_A^t\mu-\phi_A^t\left(V_s\right)\lambda, & \text{if } \phi^t\left(s\right)=B\end{cases}. \tag{3.1}$$

Note that that the difference in economic payoff from playing either A or B for an

individual s is $\alpha - \beta$, since ceteris paribus $\left(\phi_A^t | \phi^t(s) = A\right) = \left(\phi_A^t | \phi^t(s) = B\right) + 1$. Thus action $\phi^t(s) = A$ is a dominant action for a player, given $\phi^t(-s)$. This action hurts all players $r \in V_s$ and affects all players $r \in K \setminus (V_s \cup \{s\})$. We assume that players have information on local externalities being administered to them (specifically they know by which players these externalities are caused), but do not realize that distant externalities also influence their payoffs.

We restrict attention to the parameter range in which $\alpha - m(\lambda - \mu) < \beta$, i.e. the range in which an individual player is better off when no negative externalities are levied on him than when all his neighbours levy negative externalities on him. Outside this range, individuals cannot benefit from cooperating with their nearest neighbours and no coordination arises. Within this range we focus on two cases. In the first case, the total economic payoff is maximized by the configuration with $\phi^t(s) = A$, $\forall s$, i.e. we set the parameters such that $\alpha - m(\lambda - \mu) < \beta < \alpha + (N-1)\mu - m\lambda$. The second case is the one in which the total economic payoff is maximized by the configuration with $\phi^t(s) = B$, $\forall s$, i.e. we set the parameters such that $\alpha - m(\lambda - \mu) < \alpha + (N-1)\mu - m\lambda < \beta$. In both cases a player s with $\phi^t(s) = A$ is willing to play action B whenever (some of) his neighbours credibly indicate that when he switches action, they will do the same thing. Note that the second case can be viewed as the standard public goods problem, in which contributing to the public good is a dominated action for the individual, but where the situation in which all individuals contribute yields the highest overall payoff. We present three examples to motivate these cases. The first two examples closely mirror the settings decribed by Wittfogel (1963) and the third is of our own.

Example 3.1. *Suppose that action A represents preventing the rainfall falling on one's land to flow to that of others by channelling the rainfal to irrigate one's own patch of land. Take action B to be building an irrigation system that benefits the neighbours also (or participating in the building of a whole irrigation system for the area). Action A increases the pay-off of the individual the most. It will reduce the amount of water available to his neighbours though and the increased production of the individual may make them envious, which implies a negative local externality of* λ*. The effects on the rest of the population of choosing A rather than B may be positive or negative. If the cooperation between neighbours (action B) reduces the available water for the rest, one can view this as a positive externality* μ *in case the individual chooses A. If the total production of the individual plus neighbours increases under B though, this may allow increasing specialisation for the whole population, implying a negative distant externality for option A. Now think of a group of neighbours who all irrigate their own lands individually but would all like the others to start building a large irrigation system that benefits himself and his neighbours. The members of this group of neighbours are better of when they all play action B and all participate in the building such an irrigation system. So, when theycooperate, they will play B. But then, when all individuals in the population act this way, the distant externalities are lost. Depending on the size of the distant externalities, the result is an inferior outcome for the population as a whole (case 1) or a superior outcome for the population as a whole (case 2).*

Example 3.2. *Another example is to think of option A for an individual as not building a dike on his section of a river and option B as building the dike on his section of the*

river. Building a dike will not increase the pay-off of the individual because the reduced risk of flooding does not outweigh the effort, but will increase the payoff of his nearest neighbours also because they are also likely to benefit somewhat from the reduced risk of flooding in that neighbourhood. Other communities will however see their probability of flooding increase if the individual builds a dike because less superfluous water will be drained at the site of the dike. This can be modelled by taking A to have local negative externalities λ (the neighbours do not enjoy the reduced risk of flooding) and positive distant externalities (the rest of the population has less risk of flooding. The members of this group of neighbours are better of when they all play action B and all participate in the building of a canal. So, when they sit together and cooperate, they will play B. But then, when all individuals in the population act this way, the total reduction in risk of flooding is not as great as before and may perhaps not outweigh the effort of building the dikes.

Example 3.3. *A final example (of case* 1*) is when action A represents the possibility to turn one's land into property inaccessible for other individuals, while action B is not restricting access. The direct profit to the individual of restricting access to one's land is higher than that of letting everybody walk across one's land and thus disrupting your use of the land. Thus action A dominates action B. Moreover, restricting access allows for specialization of the labor force to take place and this way has a positive effect on the payoff of all other individuals in the population ($\mu > 0$). However, restricting access creates a direct negative externality (λ) to the neighbours, since they are no longer allowed to walk across the property. Again, a (small) group of neighbours which want*

to cooperate, agree to all play B, yielding an inferior outcome for the population as a whole when all individuals in the population act in the same way.

In case 1, small authorities advocate a set of rules which yields global inefficiencies. However, for coalitions of individuals larger than some substantial minimum size, the effect of the positive distant externalities nullify the large negative local externalities. Thus, when an authority grows above this minimum size, it realizes that advocating to every group member to play A is better than having every group member playing B and it thus changes it rules accordingly. Large authorities therefore do not yield the global inefficiencies depicted in the above examples of case 1.

All players have limited knowledge about the graph K and the stage game. Each player s knows which players are in the set V_s and in the sets V_r $\forall r \in V_s$. Thus, players can asses whether they are part of a clique and authorities can asses whether players on their borders are also on the border of some other authorities. As to the stage game, players know about the payoff difference $\alpha - \beta$ between the different actions. They also observe the payoff effects λ and μ of actions taken by their neighbours and are able to trace down from which neighbour the externalities originate. However, they are unable to see which non-neighbours caused the global externality μ to be levied on them. Players observe the signals of their neighbours, i.e. they know the punishment potential and the conditional strategy each of their neighbours has committed to. Players also observe which of his neighbours receive a learning draw. Thus, players can infer which of their neighbours do not comply with their conditional strategy while they were handed the learning draw and which of their neigbors do not comply with their conditional strategy

because they were not (yet) handed a learning draw. Players are not aware of the action played by players other than their neighbours, or for that matter, we can assume that they do not even know about the existence of players other than their neighbours.

We now specify the sequence of events in a time period in detail.

1. Players get the opportunity to form a central authority-in-formation. Such an authority can already set rules in step 2, but players only make the final decision on joining in step 3.

2. Each central authority R decides upon a set of rules to be adhered to by all players $r \in R \cup \partial R$. It communicates the set of rules to all these players.

3. Learning draws are handed out. Players r ($r \in R \cup \partial R$ for all c.a.'s R) that get a learning draw decide whether to join R or remain in their current situation (member of another c.a. or non-member of any c.a.). Players that do not get a learning draw and are currently not member of any c.a. cannot join a c.a. Players that do not get a learning draw, but are member of a c.a. evaluate the rules set by this c.a. against the rules set by any c.a. from which they received a signal (i.e. of which they are a neighbour). Subsequently they decide whether to remain member of their 'current' c.a. or to become member of one of the c.a.'s from which they received a signal.

   Once a member of a c.a., players remain member of that c.a. at least until step 3 next period.

4. Players play an action from Γ.

5. Each central authority R observes all information observed by all players $r \in R$. Based on this information the c.a. decides on administering punishment to the players in $R \cup \partial R$ which did not comply with the rules set in step 2. It orders individual members $r \in R$ to carry out this punishment and subsequently transfers punishment potential if necessary.

From this sequence of moves and the information structure, we can directly infer the following corollary.

Corollary 3.4. *A central authority can only be formed by a subset of players from one clique.*

Proof

Facing the decision whether or not to participate in a c.a.-in-formation, a player weighs the potential profit from joining against the potential costs of being exploited by others who claim they will join, but do not do so in the end. A c.a.-in-formation can set its rules in step 2 above such that players who choose not to join in step 3 are punished severely. If the total punishment potential of the c.a. is large enough, this can make not-joining in step 3 harmful. Thus, a necessary condition for a player to join is that he knows that enough others will join to give the c.a. enough punishment potential. In a clique, all players know that all other players in the clique face the same decision and will thus also join. Such a necessary common knowledge information structure is only present within a clique of players, which yields the necessary condition in the corollary. □

In order to enforce its announced rules whatever these rules are, a central authority has to find an enforcement mechanism. A possibility for an enforcement mechanism of rules is then that the central authority announces the rules and then compiles a list of individuals to be punished for non-compliance or for failure to punish when instructed to do so. Because the first one on the list will expect to be punished by the others if he does not comply, he will comply. Hence the first person will comply, and, through a repetition (forward induction) of this argument, the second person will comply, and so forth. The notion of a list that enforces discipline is similar to the notion of a 'matrix of discipline' of Kuhn (1962). As to the punishment for non-compliance, the only requirement of the severity is that it outweighs the possible benefit of deviation. Because individuals cannot coordinate on strategies without forming a central authority, a complete break-down of the central authority will not occur, and no c.a. can be formed within another c.a.

An important point is that the announced strategy cannot be altered until step 2 next period, which implies that a central authority can credibly pre-commit on its own rules for one period at a time. One could therefore interpret a period as the length of time it takes between decision rounds. Because of the possibility of revising the rules each period, there is a collective time-inconsistency problem in the sense of Asheim (1997).

4. Basic Results

In this section we characterize the (self-confirming) equilibrium of the total model.

Consider first the circumstances under which a c.a. will form. Corollary 3.4 states that we need only to focus on players in a clique. We now state a sufficient condition on

the combined punishment potential of the potential members of a c.a. By $\lfloor . \rfloor$ we denote the entier function, i.e. $\lfloor z \rfloor = \max \{x \in \mathbb{Z} \mid x \leq z\}$.

Theorem 4.1. *Consider player $s \in C_k$, with*

$$k \geq \left\lfloor \frac{\alpha - \beta}{\lambda - \mu} \right\rfloor + 2 \text{ and } \sum_{r \in C_k \cap V_s \cup \{s\}} \theta(r) - \max_{r \in C_k \cap V_s \cup \{s\}} \theta(r) > \alpha - \beta.$$

Then there is a strict positive probability that a central authority will evolve in this clique. When

$$k < \left\lfloor \frac{\alpha - \beta}{\lambda - \mu} \right\rfloor + 2 \text{ or } \sum_{r \in C_k \cap V_s \cup \{s\}} \theta(r) - \max_{r \in C_k \cap V_s \cup \{s\}} \theta(r) < \alpha - \beta.$$

the probability of emergence of a central authority in C_k is zero.

Proof.

From (3.1) we have that the difference in economic payoff from playing either A or B for an individual s is $\alpha - \beta$, since ceteris paribus $\left(\phi_A^t | \phi^t(s) = A\right) = \left(\phi_A^t | \phi^t(s) = B\right) + 1$. Thus, when even the most powerful player in a c.a.-in-formation can be punished by an amount $\alpha - \beta$, no individual player can benefit by deviating from the rules prescribes by this c.a. This is the case when $\sum_{r \in C_k \cap V_s \cup \{s\}} \theta(r) - \max_{r \in C_k \cap V_s \cup \{s\}} \theta(r) > \alpha - \beta$. Since a c.a. is the only coordination device present, individual players are unable to form a coalition of deviators. Thus no player will deviate and the c.a. will form.

Individual players in principle choose the dominant action A when given a learning draw. Players are only willing to switch to action B and form a (B-playing) c.a. if

they benefit from this. This means that they will demand that enough local negative externalities will be gotten rid of by forming a c.a., i.e., they demand that enough neighbours join in the c.a., or in other words that $\phi_A^t(V_s)$ is lowered enough when they form a c.a. Thus they solve

$$\alpha + \left[\left(\phi_A^t | \phi^t(s) = A\right) - 1\right]\mu - \phi_A^t(V_s)(\lambda - \mu) < \beta + \left(\phi_A^t | \phi^t(s) = B\right)\mu - \widetilde{\phi}_A^t(V_s)(\lambda - \mu)$$

for $0 \leq \widetilde{\phi}_A^t(V_s) \leq k$, since $\lambda - \mu$ is the net negative externality a player experienced from his neighbours. Solving this yields

$$\widetilde{\phi}_A^t(V_s) < \phi_A^t(V_s) - \frac{\alpha - \beta}{\lambda - \mu}.$$

Thus a player s announces that he is willing to play B when at least $\left\lfloor \phi_A^t(V_s) - \widetilde{\phi}_A^t(V_s) \right\rfloor + 1 = \left\lfloor \frac{\alpha-\beta}{\lambda-\mu} \right\rfloor + 1$ of his A-playing neighbours also signal that they are willing to play B (and thus vote for a c.a. with a punishment scheme that enforces compliance with these conditional strategies).

The last condition for a c.a. to form is that these $\left\lfloor \frac{\alpha-\beta}{\lambda-\mu} \right\rfloor + 1$ neighbours of s and s himself are all in the same clique C_k. This ensures s that all these players also see that they will benefit from being in the B-playing c.a. This requires that the clique is large enough, i.e., $k \geq \left\lfloor \frac{\alpha-\beta}{\lambda-\mu} \right\rfloor + 2$.

When the above conditions are met for a certain clique C_k, (a subset of player from) C_k will form a c.a. once at least $\left\lfloor \frac{\alpha-\beta}{\lambda-\mu} \right\rfloor + 2$ players in this clique get a learning draw at the same time. This event happens with strict positive probability. When one of the

conditions is not met for C_k, a c.a. can never initiate in the group C_k. □

This theorem shows that one needs enough mutually connected individuals with similar interests to start a central authority with enough punishment potential. Now we show that the conditions in the Theorem are met when the punishment potentials for different players are random draws from a distribution with positive weight on values above $\frac{\alpha-\beta}{(k-1)}$.

Corollary 4.2. *Consider a large population in which a large number of k-cliques with $k \geq \left\lfloor \frac{\alpha-\beta}{\lambda-\mu} \right\rfloor + 2$ is present. Then, when*

$$\int_{\frac{\alpha-\beta}{(k-1)}}^{\theta^{\max}} f(\theta)\, d\theta > 0 \tag{4.1}$$

almost surely at least one central authority will emerge.

Proof.

Condition (4.1) ensures that in each k-clique there is strict positive probability that $\sum_{r \in C_k} \theta(r) - \max_{r \in C_k} \theta(r) > \alpha - \beta$ is satisfied and thus that the condition on punishment potential in Theorem 4.1 is met. A strong law of large numbers argument guarantees that in a large population with a large number of k-cliques almost surely the condition will be met. □

The corollary states that heterogeneity w.r.t. punishment potential facilitates the emergence of a central authority. In a population that is homogeneous w.r.t. punishment potential, a similar strong law of large numbers argument argument shows that a c.a.

will emerge if the homogeneous punishment potential is high enough, i.e. when $\theta(s) = \overline{\theta} > \frac{\alpha-\beta}{(k-1)}\ \forall s \in K$.

We now infer what happens once one or more authorities have emerged.

Theorem 4.3. *When the conditions of Theorem 4.1 are met, one single central authority, with all players in the population as its members, will be the only stable outcome.*

Proof.

By construction a central authority S at time t will have a set of rules that maximize the combined payoff of the players which are in S before new players have had an opportunity to join S. We label the action which S prescribes to $\widetilde{s} \in S \cup \partial S$ by $a^t_{\widetilde{s}}$. Suppose the action $a^t_{\widetilde{s}}$, $\widetilde{s} \in S \cup \partial S$, are the action that optimize combined economic payoff $\sum_{s \in S} \pi\left(s, \phi^t\right)$ without considering the possibility of handing out or receiving punishment.. Obviously, a set of rules leading to a situation in which all players $\widetilde{s} \in S \cup \partial S$ play according to $a^t_{\widetilde{s}}$ and no punishment is administered or received by the players $s \in S$, is optimal. We now describe such a set of rules. The rules set by S prescribe $\widetilde{s} \in S \cup \partial S$ to play $a^t_{\widetilde{s}}$. A punishment scheme which makes deviating unprofitable for all players $\widetilde{s} \in S \cup \partial S$ involves different treatment for players $s' \in \partial S$ who are not member of a c.a. (different from S) and for players $s' \in \partial S$ who are member of a c.a. (different from S). We specify these different treatments for the players in ∂S below. First we construct a possible punishment scheme for the players $s \in S$. Rank the players $s \in S$ in an arbitrary way. Announce the order on the list to all players $s \in S$ and threaten to punish the player who is first on the list (labelled s_1) by an amount $\alpha - \beta + \varepsilon$, with $\varepsilon > 0$ small. The fact that the c.a. S has formed ensures that S has a punishment potential larger than

$\alpha - \beta$ and thus that this threat is credible. Threaten to punish the second player on the list (s_2) by the same amount, if the first player does not deviate from $a^t_{s_1}$. Given the punishment potential and the threat w.r.t. player s_1, the threat to s_2 is also credible. Furthermore, s_2 will infer that s_1 will comply and thus that deviating is not profitable for s_2 (remember that s_1 and s_2 cannot form a sub-coalition within S). Then, the scheme involves a threat of $\alpha - \beta + \varepsilon$ to player s_3, if players s_2 and s_1 do not deviate and so forth. In equilibrium no player $s \in S$ deviates and the believes of the players are confirmed.

The scheme involves the following threats to players $s' \in \partial S$.

1. If s' is not member of any c.a., he is threatened by $\sum_{s \in S} \theta(s)$, the total punishment potential of the c.a. S. This punishment will only be administered if s' does receive a learning draw next period and fails to comply with $a^t_{s'}$. The c.a. S thus forces compliance by s' with the rules it set.

2. If s' is a member of a c.a. R (different from S) which is weaker than S, i.e. $\sum_{s \in S} \theta(s) > \sum_{r \in R} \theta(r)$. Then, s' is threatened by an amount $\sum_{s \in S} \theta(s)$ if he does not join S and plays $a^t_{s'}$ in the next period. The c.a. S sees that it is able to counter any threat R makes by a stronger threat and therefore it knows that it can force compliance by s' with the rules $a^t_{s'}$.

3. If s' is a member of a c.a. R (different from S) which is stronger than S, i.e. $\sum_{s \in S} \theta(s) < \sum_{r \in R} \theta(r)$. In this case, s' is not threatened by any punishment. The c.a. S sees (by noticing the strength of R) that it will not be able to force s' to comply with $a^t_{s'}$ and that it has to administer any punishment it threatens with in

case of non-compliance. Given these considerations, threatening with punishment is harmful to S.

Under this punishment scheme, all players that are not member of a c.a. which is stronger than S will recognize that the will be harmed by non-compliance and will thus comply with the rules set by S. Next period, all players s' that were not member of any c.a. will join S if they are handed the learning draw. All players $s' \in R$ with R weaker than S will join S. Finally, all players $\widetilde{s} \in \partial R \cap S$ with R stronger than S, will join R.

Now, label the strongest c.a. present in the population by $S^{t,\max} = \arg\max_{S \in W^t} \sum_{s \in S} \theta(s)$. This authority will be at least as strong at time $t+1$ as it was at time t. It will be stronger at $t+1$ with strict positive probability. Of course there is the possibility that $S^{t+1,\max} \not\supset S^{t,\max}$, i.e. the strongest c.a. at time $t+1$ emerged from a c.a. different from $S^{t,\max}$. In that case it is still the case that $\sum_{s \in S^{t+1,\max}} \theta(s) > \sum_{s \in S^{t,\max}} \theta(s)$. Thus we have show that the strength of the strongest c.a. increases with positive probability and never decreases, thus excluding cyclic behavior of the model. The only stable outcome is then a single c.a. of which all players in the population are members. □

We add two comments to these results. First, we see that, in equilibrium, punishment will never take place, implying that the total payoff obtained by any individual each period equals his economic payoff from playing the stage game. Second, although the model is limited to a stage game with two actions, we argue that the qualitative results carry over to models incorporating more general stage games. With more actions in the stage game, it is still the case that a c.a. can start if enough mutually connected and sufficiently strong individuals can all increase their payoffs by forming a c.a. that

maximizes their combined utilities. A c.a. that has started, will still expand and the number of c.a.'s remaining therefore still converges to 1. The important changes are hence the conditions under which a c.a. starts. We think the most likely setting for such a thing to happen is when actions inflict externalities on others.

As illustrations of the evolution of play, we consider the two cases presented in section 3. In both cases in populations without any c.a.'s, all players play A. Within the initial authorities that arise and that are still small, playing action B is advocated. In case 2, one c.a. advocating B to all members is also the final outcome. In case 1 however, in the long run at least one of the authorities will become large enough to see that the local negative externalities are outweighed by the distant positive externalities of having all its members play A. Consequently, such a large c.a. changes the action it advocates to its members from B to A. This result is in the following corollary.

Corollary 4.4. *Suppose a small c.a. has formed and the total economic payoff is maximized by the configuration with $\phi^t(s) = A$, $\forall s$. Then, the small c.a. will advocate playing B to its members. Only when a c.a. grows sufficiently big, it will see the benefits of playing A. Consequently, sufficiently large c.a.'s will tell their members to play A.*

An illustration of the evolution of a c.a. under the conditions of case 1 is given in figures 4.1 to 4.3. In the figures $k = m = 4$, $\alpha = 10$, $\beta = 5$, $\lambda = 4$, and $\mu = 1$. All individuals receive a learning draw at each time t. In this specific example, the sets V_s have a similar structure $\forall s \in K$.

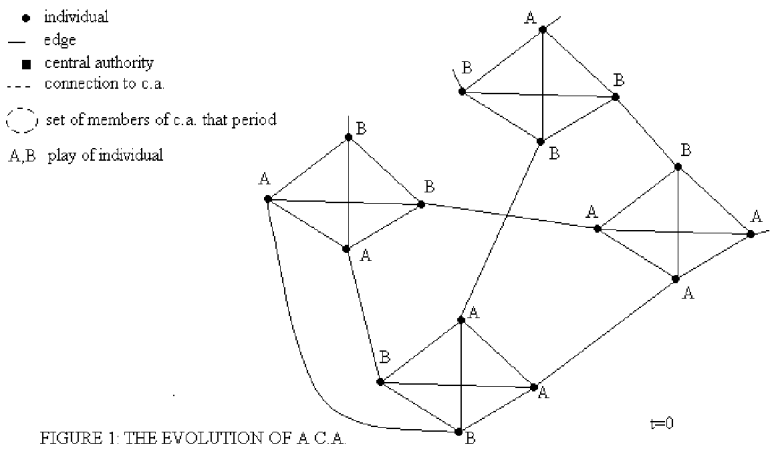

Figure 4.1: Part of the population at time $t = 0$.

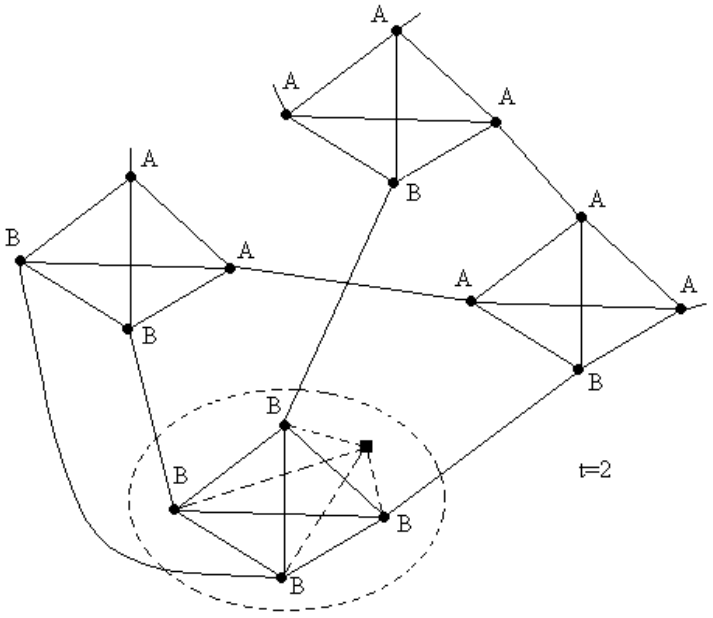

Figure 4.2: Part of the population at time $t = 2$.

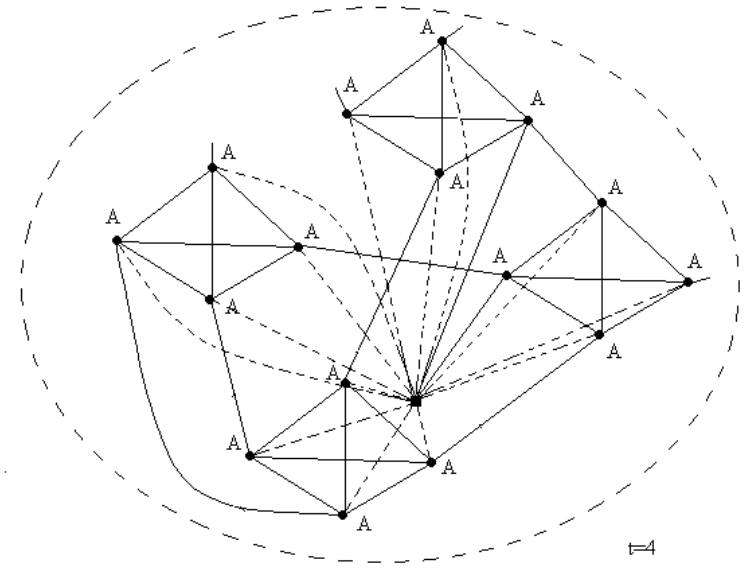

Figure 4.3: Part of the population at time $t = 4$.

The situation of the game is shown in period 0, where individuals randomly play an action. In period 1 (not shown) all individuals play A. In period 2, the bottom 4 players were strong enough to set up a c.a., that forces its members to play B and forces the neighbours to play B because the positive externality μ does not yet outweigh the negative externality λ. Finally, period 4 is shown, in which all members and neighbours are forced to play A.

5. Extensions: Asymmetric Punishment and Mutations

In this section we consider a slightly altered version of the model.

Suppose that the ability to punish non-members is a fraction $0 < \delta < 1$ of the ability to punish members of a c.a. This means that punishment potential is asymmetric as to whether the punishment is between or within central authorities. This asymmetry

reflects the argument that it is easier to punish others close to you than it is to punish others which are further away. In this setting Theorem 4.1 and Corollary 4.2 still hold. It is only when multiple c.a.'s have arisen that the model differs. This is illustrated in the following theorem.

Theorem 5.1. *Consider the model with asymmetric punishment potential. Assume that the population is large, that there are a large number of k-cliques in the population with $k \geq \left\lfloor \frac{\alpha-\beta}{\lambda-\mu} \right\rfloor + 2$ and that*

$$\int_{\frac{\alpha-\beta}{(k-1)}}^{\theta^{\max}} f(\theta)\, d\theta > 0.$$

Then, there both one single central authority or several central authorities of which not necessarily all players are a member and with possibly conflicting rules are possible long run outcomes of the model.

Proof

The conditions in the theorem ensure that at least one c.a. will emerge almost surely (see Theorem 4.1 and Corollary 4.2). As to the different possible outcomes, we can restrict the proof to providing two examples, each of which leads to one of the possible outcomes mentioned in the Theorem.

First, consider a large value of δ. Suppose that, through a specific distribution of learning draws, the first c.a. S^t emerges at time t. Subsequently, at time $t+1$, only the players in ∂S^t get a learning draw. They join S^t, since this authority is able to threaten with enough punishement. Now S^t becomes $S^{t+1} = S^t \cup \partial S^t$. Distribute learning draws only to the players in ∂S^{t+1}, and so forth until all players in the population are member

of S. This specific distribution of learning draws has a low probability of happening, but there are obviously many more ways in which a single c.a. may emerge. There is thus positive probability that the model long run outcome is a single c.a. of which all players are members.

Second, suppose that, through a specific distribution of learning draws, two c.a.'s S and R have emerged and that at a certain time t, $S \cup R = K$, i.e. all players in the population are member of either S or R. Suppose, w.l.o.g. that $\sum_{s \in S} \theta(s) > \sum_{r \in R} \theta(r)$. Thus also $\sum_{s \in S} \theta(s) > \delta \cdot \sum_{r \in R} \theta(r)$, seeing to it that R cannot threaten players in $\partial R \subset S$ enough to force them to join R. However, with positive probability it can be that $\delta \cdot \sum_{s \in S} \theta(s) < \sum_{r \in R} \theta(r)$, preventing S from threatening individuals in $\partial S \subset R$ enough to force them to join S. Thus, there is a 'power balance' between S and R and a stable equilibrium exists in which both S and R threaten all their members with their maximal punishment potential and no authority threatens non-members (note that this punishment scheme is only one possible scheme). In this equilibrium, no player becomes member of a different c.a. at any time $t' > t$ and the two c.a.'s S and R remain. Depending on the exact size of S and R and the parameter values, the action the authorities prescribe to their members can be either A or B. Moreover, it is possible that the action S prescribes is different from the action R prescribes. When the value of δ is close to 0, it is even possible that S is not able to threaten a single individual enough to make him join. Therefore, there is a parameter range for which single individuals can remain non-member of a c.a. forever.

Similar arguments apply to situations in which the population is divided over more

than two c.a.'s, which imply that multiple c.a.'s may coexist forever. Moreover, not all players are necessarily members of a c.a. □

The theorem states that there are several possible limit outcomes. The initial configuration of the population, the specific parameter values and the realizations of the sequences of learning draws will determine which outcome is reached.

Multiple c.a.'s may yield global inefficiency in which all individuals in all groups are worse of than with a single c.a., since the members in each c.a. do not take account of the effect of their actions on the payoff of the members of the other authority. The possibility that groups of individuals are 'locked' into an inefficient equilibrium in which all groups loose out, is a way of modelling discrimination or conflicts between groups or regions.

We now introduce a small probability of mutations into the model, which allows us to identify which of the multiple stationary states sketched in Theorem 5.1 is the *stochastically stable state* or *long run equilibrium*. This is the equilibrium that is played 'almost all of the time' when the mutation rate goes to 0 in the limit. This concept was developed in the papers of Kandori, Mailath, and Rob (1993) and Young (1993) and surveys of this literature are given by Samuelson (1997) or Young (1998). Mutations are usually taken to represent one of three phenomena. First, mutations may represent experimentation by the players to learn about what might happen off the equilibrium path. Second, mutations may represent (computational) errors on the part of the individual players in the implementation of an action. Lastly, mutations can represent genetic mutations in that individuals' actions are 'preprogrammed' by their set of genes and sometimes

spontaneous mutations in these genes occur.

We introduce mutations as follows. At each time t every player in the population has a very small probability $\varepsilon > 0$ of mutating. When mutating, a player joins a randomly selected c.a. to which he is adjacent, with each adjacent c.a. having the same probability of being chosen. If the mutant is not adjacent to any c.a., he does not become a member of any c.a.. On top of joining an arbitrary c.a., a mutant randomly selects an action to play in the stage game. The following theorem states the selection result we obtain.

Theorem 5.2. *Consider the model with asymmetric punishment potential and mutations. Assume that the population is large, that there are a large number of k-cliques in the population with $k \geq \left\lfloor \frac{\alpha-\beta}{\lambda-\mu} \right\rfloor + 2$ and that*

$$\int_{\frac{\alpha-\beta}{(k-1)}}^{\theta^{\max}} f(\theta)\, d\theta > 0.$$

Then, the only long run equilibrium is the state in which there is only one c.a. present in the population, and in which all players are member of this c.a.

Proof.

This proof is based on Markov theory by Freidlin and Wentzell (1984). Our strategy is to show that the probability of leaving a stationary state in which multiple c.a.'s are present in the population is of a lower order of ε than the probability of leaving the stationary state in which there is only one c.a., of which all players in the population are a member. Then, the latter state is stochastically stable.

Consider a stationary state in which multiple c.a.'s are present and each player is a

member of one of these c.a.'s. Label two of these c.a.'s which have a common boarder S and R. Consider a player $s \in S \cap \partial R$, which mutates and joins R. Now, the mutation might distort the power balance (see proof of Theorem 5.1) between R and another c.a. in the population (not necessarily S). In this case, the new state is not stable and R will subsequently expand by taking over (part of) another c.a. The mutation does not necessarily distort the power balance between all c.a.'s however, in which case the new state would also be stable without the presence of mutations. We then consider another player $s' \in S \cap \partial R$ who mutates at some later time, which again leads either to a distortion of the power balance or not. When over time enough players from S mutate and become members of R, the power balance in the population will eventually be distorted and at least one c.a. will disappear. Should there still be multiple c.a.'s, we repeat the above argument, until the population is in a state with only one c.a. of which all players are members. The path leading to such a state consists of (a finite number of) single mutations occurring at different times, which occurs with a probability of order $O(\varepsilon)$, and does not take any *simultaneous* mutations.

A single mutation from a state $S = K$ will never get the system to another stable state, since a single player cannot form a c.a. all by himself that is capable of keeping up a balance of power between the individual and c.a. S. Even the most powerful mutant $s' \in S$, still has a punishment potential $\theta(s') < \delta \cdot \sum_{s \in S} \theta(s) - \theta(s')$ when the population, and thus size of S, is sufficiently large because the punishment potential of a single player can never be larger than $\theta^{\max}$. In any case, as an individual cannot credibly threaten with punishment, an individual cannot resist the threat of a c.a. Thus, $l \geq 2$ *simultaneous*

mutants are needed to upset the state $S = K$. Such an event happens with a probability of the order ε^l.

Since the probability of a mutation is very low, the move of a system not in stationary state to a stationary state is relatively fast. From standard Markov chain theory we then know that the system is in state $S = K$ a fraction $\frac{\varepsilon}{\varepsilon+\varepsilon^l}$ of the time. When the mutation rate ε is taken to 0 in the limit, the system is in state $S = K$ a fraction 1 of the time. □

Hence, although multiple equilibria are present in the altered model, introducing a small probability of random mutations serves as a selection device, selecting the stationary state with only one c.a. as the only stochastically stable state. A typical path to this stochastically stable state may look like this. From the initial state, the system moves rapidly to a stationary state with multiple c.a.'s. After some mutations that do not change the power balance between the c.a.'s, eventually a mutation does change the power balance and the system moves quickly to a new stationary state with less c.a.'s remaining. Repeating this procedure in finite time leads to the stochastically stable state being reached.

6. Discussion and concluding remarks

In this paper we considered the role and evolution of central authorities. Its role in an evolutionary sense is to prevent individuals from taking decisions with greater negative externalities than private benefits. As such, central authorities promote cooperation. The ability of a central authority to communicate directly with all its members allows it to obtain a monopoly over violence in which it can punish individuals that do not behave

as the authority wants them to do. This set-up is justified if there is a returns-to-scale advantage in the gathering and processing of information.

We showed that central authorities arise when many individuals promote the same set of rules, because these rules generate higher payoffs, but cannot act according to these rules in the absence of a (credible) commitment device. This happens in an environments in which individual actions generate externalities on other individuals. As central authorities grow, they incorporate more and more externalities and may change the set of rules they set over time. This description of the evolution of central authorities concurs with the observation that many central authorities, political parties and other organizations, start out as single-issue groups, but end up representing several interests: van Waarden (1985) for instance showed in a detailed account of the evolution of pressure groups and branch organizations in 19th century Holland, that many current institutions that incorporate the different interests of many industries actually started by representing a single interest.

Another insight of the model was that the enforcement of the set of rules within a c.a. becomes easier in large authorities, because in large c.a.'s there will always be a substantial number of individuals that blindly follow the rules the central authority sets. This is because not all individuals in every period bother to think about the alternatives to following the rules. The punishment potential of this group of individuals ensures that no single individual can benefit by deviating from the rules of the central authority, which forces everyone to observe the rules.

An important conclusion in the standard model is that ultimately, only one central

authority remains, of which all players are members. Although allowing for asymmetric punishment potential did mean that multiple authorities could co-exist in equilibrium, allowing for mutations revealed that in the stochastically stable state again only one central authority remains.

Perhaps the most important contribution of the paper is that it provides a flexible framework for the analysis of the evolution and behaviour of central authorities under alternative assumptions. One obvious assumptions to alter is that the central authority itself is a rather passive equilibrium selection device in which all members' interests have equal weight. A natural alternative would be to assume that a specific group of individuals within the bounds of the authority actually forms the control center in which rules are decided upon. One choice for the controlling group would be the individuals who set up the central authority in the first place. Though the precise rules the authority would enforce would then depend on the interests of this controlling group, the other results in this model would not change. Because reviews of the rent-seeking literature indeed suggest that special interest groups controlling or lobbying within an authority have success (Mitchell and Munger (1991) or Austen-Smith (1994)), this seems a plausible route. In this paper however, we deliberately assumed the central authority to be passive, in order to free the model and the ensuing analyses of considerations of special interest groups. As such, the model stresses the possible benefits of central authorities.

Another line of inquiry that could be taken with this model is to vary the amount of information individuals and central authorities have about the existence and strategies or actions of other individuals and authorities. This affects the strategic interactions

between authorities and individuals and seems a promising way to capture aspects of actual conflicts between central authorities, where shifting coalitions of central authorities are a common phenomenon (e.g. Burbidge, DePater, Meyers, and Sengupta (1997)).

References

Asheim, G.B. (1997). Individual and collective time-consistency. *Review of Economic Studies 64*, 427–443.

Aumann, R. (1989). Game theory. In *The New Palgrave on Game Theory*. London: MacMillan.

Austen-Smith, D. (1994). Interest groups: Money, information and influence. In D.C. Meller (Ed.), *Perspectives on Public Choice*. Cambridge, Massachussets: Cambridge University Press.

Axelrod, R. (1987). *The Evolution of Cooperation*. New York: Basic Books.

Burbidge, J.B., J.A. DePater, G.M. Meyers, and A. Sengupta (1997). A coalition-formation approach to equilibrium federations and trading blocks. *American Economic Review 87*, 940–956.

Eggertson, T. (1990). *Economic Behavior and Institutions*. Cambridge, MA: Cambridge University Press.

Eshel, I., L. Samuelson, and A. Shaked (1998). Altruists, egoists and hooligans in a local interaction model. *American Economic Review 88*, 157–179.

Farrell, J. and M. Rabin (1996). Cheap talk. *Journal of Economic Perspectives 10*,

103–118.

Freidlin, M. and A. Wentzell (1984). *Random Perturbations of Dynamical Systems.* New York: Springer Verlag.

Harris, M. (1993). *Culture, People, Nature* (6th ed.). New York: Harper Collins Publishers.

Harsanyi, J. (1985). Rule utilitariansim, equality and justice. *Social Philosophy and Policy 2*, 115–127.

Heap, H.S.P. (1994). Institutions and (short-run) macroeconomic performance. *Journal of Economic Surveys 8*, 35–55.

Hoel, M. (1990). Local versus centralised wage bargaining with endogenous investments. *scandinavian Journal of Economics 92*, 453–469.

Kahneman, D., J.L. Knetsch, and R. Thaler (1986). Fairness as a constraint on profit seeking: Entitlements in the market. *American Economic Review 76*, 728–741.

Kandori, M., G.J. Mailath, and R. Rob (1993). Learning, mutation and long run equilibria in games. *Econometrica 61*, 29–56.

Karandikar, R., D. Mookherjee, D. Ray, and F. Vega-Redondo (1998). Evolving aspirations and cooperation. *Journal of Economic Theory 80*, 292–331.

Kuhn, T.S. (1962). *The Structure of Scientific Revolutions.* Chicago, Illinois: University of Chicago Press.

Mitchell, W.C. and M.C. Munger (1991). Economic models of interest groups: An introductory survey. *American Journal of Political Science 35*, 512–546.

North, D.C. (1981). *Structure and Change in Economic History*. New York: Norton Publishers.

North, D.C. (1990). *Institutions, Industrial Change and Economic Performance*. Cambridge, Massachussets: Cambridge University Press.

Osborne, M.J. and A. Rubinstein (1994). *A Course in Game Theory*. Cambridge, MA: The M.I.T. Press.

Pardo, J.C. and F. Schneider (Eds.) (1996). *Current Issues in Public Choice*. Cheltenham, United Kingdom: Edgar Elgar.

Postgate, J.N. (1992). *Early Mesopotamia: Society and Economy at the Dawn of History*. London: Routledge.

Potters, J. and F. van Winden (1996). Comparative statics of a signalling game. *International Journal of Game Theory 25*, 329–353.

Rawls, J. (1971). *A Theory of Justice*. London: Oxford University Press.

Samuelson, L. (1997). *Evolutionary Games and Equilibrium Selection*. Cambridge, MA: The M.I.T. Press.

Soskice, D. (1990). Wage determination: The changing role of institutions in advanced industrialised countries. *Oxford Review of Economic Policy 6*, 36–61.

Tieman, A.F., H.E.D. Houba, and G. Van der Laan (1998). On the level of cooperative behavior in a local interaction model. Discussion paper, nr. TI 98-024/1 (revised version), Free University and Tinbergen Institute.

Tieman, A.F., G. Van der Laan, and H.E.D. Houba (1997). Bertrand price competition

in a social environment. Discussion paper, nr. TI 96-140/8 (revised version), Free University and Tinbergen Institute.

van Waarden, F. (1985). Regulering en belangenorganisaties van ondernemers. In F. van Halthoon (Ed.), *De Nederlandse Samenleving sinds 1815*. Assen: Van Gorcem. In Dutch.

Weber, M. (1922). *Wirtschaft und Gesellschaft*. Tuebingen: Mohr.

Wenke, R. (1984). *Patterns in Prehistory* (1st ed.). New York: Oxford University Press.

Wittfogel, K.A. (1963). *Oriental Despotism. A Comparative Study of Total Power*. New Haven, CT: Yale University Press.

Young, H.P. (1993). The evolution of conventions. *Econometrica 61*, 57–84.

Young, H.P. (1998). *Individual Strategy and Social Structure: An Evolutionary Theory of Institutions*. Princeton, New Jersey: Princeton University Press.