The role and evolution of central authorities.

Paul Frijters*and Alexander F. Tieman†

February 8, 1999

Abstract

In this paper we consider the way in which authorities arise in response to the need for coordination. In a model of local interaction, an authority is understood as a self-enforcing coordination selection structure, where the threat of violence ensures compliance. Such authorities form 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 steady situation with only one or several remaining authorities.

1. Introduction

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, enforce criminal laws, contribute to the solution of public-goods problems, etc. The notion that coordination often involves some sort of authority, seems quite plausible: individuals with similar interests often set up an embryonic authority who specializes in information gathering and coordinates action, 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 or vigilante groups, etc. In this paper, we use the term central authority (c.a.) in a very broad sense and it thus includes all the above examples. We consider the question what the defining features are of a central authority and how such authorities may arise evolutionary in a local interaction setting.

Weber (1922) argued that a central authority can coordinate actions because it monopolizes violence and simply punishes perpetrators of rules. Similarly, Aumann (1989) argued that we should be grateful for the state for allowing coordination to take place

*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. E-mail: xtieman@econ.vu.nl. Ph: +31-20-4446022, Fax: +31-20-4446020.

peacefully. Therefore, we introduce a central authority which has the capability of enforcing any set of rules and which has a high potential for punishment because it can mobilize many individuals in order to punish a single individual that does not comply with the rules set by the c.a. and can coordinate who is being punished. The defining feature of a central authority that we assume allows for this coordination is that it can directly communicate with each individual in the authority because it has an increasing returns to scale advantage in gathering and spreading information. Consider for example the situation that N persons are all capable of exchanging information directly with everyone else. A central authority requires $2N$ informational exchanges: 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. Although we do not model information costs, we do assume that the institution of a central authority is driven by costly transfer of information.

In an evolutionary model, central authorities form in an evolutionary setting when neighboring 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 neighbors and there are enough connected individuals to force any individual to play a different action, they form a coalition, whereby each neighbor 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.

In section 2, we make links to the literature on the evolution of cooperation. 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 arise as a stable steady state outcome. The implications of allowing for random mutations for this case are discussed. Section 6 concludes.

2. Literature

The problem we look at is how individuals can coordinate on dominated strategies. Such problems arise in prisoners' dilemmas or, more general, in public good games. Different approaches have been taken to this question in the literature. One approach uses game theory and experiments to find out how communication and conventions are used by individuals to solve coordination problems without physical institutions. If individ-

ual'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 dilemmas through building a credible reputation for punishing deviation of other players (see e.g. Osborne and Rubinstein (1994) and the 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.

Institutions and more particularly authorities also provide a solution to the coordination problem apart from social norms. 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. More recently, 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. From anthropological studies it seems 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)).

There are four features of authorities which make it's evolution and operation distinct from that of social norms and which are taken account of in our model. Firstly, 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. This puts the evolution and operation of an authority apart from the literature on the evolution and operation of social norms. It also makes the modelling of the evolution of an authority very difficult: if an authority is assumed to be rational and in possession of greater abilities than individuals, 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.

Secondly, authorities use institutionalized violence to enforce rules (cf. Weber (1922)). An essential feature of violence is that it represents a net loss: both the individual who is punished and the individual who punishes loose out in the sense that the payoff of both decreases.

Thirdly, authorities probably arose to overcome inefficiency problems caused by (lo-

cal) externalities: there is some evidence to suggest that the first authorities arose in relatively densely populated areas with non-specialized sedentary agricultural communities in response to conflicts over the control of water (Harris (1993) and Wenke (1984)). Most importantly, this means an absence of 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. The conflicts may have arisen if the use of water by one individual has effects on the payoffs of the nearest neighbors of the individual and on all the other individuals in the area. These effect on the payoff of neighbors are called local externalities and the effects on the other individuals are called global externalities. We give two examples of what these effects might have been. If an individual applies irrigation to his own patch of land, this will reduce the amount of water available to his neighbors, which will affect their payoff negatively. There may then also be a negative effect on the payoff of his neighbors due to envy. The production of that individual will increase however, which in turn may have a positive effect on the payoff of individuals over the whole area, for instance because it allows an increased degree of specialization. One can also imagine a situation in which the attempt of an individual to control his water supply affect the payoff of his neighbors and the whole area in the opposite direction: if an individual builds a dike on his section of a river for instance, then his own payoff will increase, and perhaps the payoff of his nearest neighbors also because they are also likely to benefit somewhat from the reduced risk of flooding in that neighborhood. Other communities will however see their probability of flooding increase because less superfluous water will be drained at the site of the dike. This would imply local positive externalities and distant negative externalities of building a dike.

Fourthly, in the social networks described above, it is unlikely that each individual has complete information on each other individual. It is more likely that each individual has some information on the people located nearby in terms of the network, but that he does not know much about 'distant' individuals.

Frankly, it is not known exactly what the circumstances were when the earliest authorities arose, but a model in which individuals are fixed in a (social) space, have limited information about non-neighbors, and whose actions affect the payoff of his neighbors in a different way from that of others in the area, seems appropriate.

3. The model

In this section, we will define the stage game that is being played between neighbors. We will show how small, embryonic authorities can form. Subsequently, we describe the behavior of authorities and show in section 4 that this behavior can lead to the expansion of some authorities and the contraction of others.

3.1. The Stage Game.

We consider a large population of N players located on an undirected connected graph. Each player is uniquely determined by a vertex of the graph, and labelled as $x \in$

$\{1,\ldots,N\}$. Each vertex x has exactly $m < N-1$ undirected distinct edges[1], that connect player x with his m distinct direct neighbors. Each player only communicates with his direct neighbors. Furthermore, each player belongs to one set of k $(2 < k \leq m)$ *mutually connected* neighbors, whereby each individual in the set of k mutually connected neighbors is connected to every other individual in that set. Denote the set of k mutually connected neighbors to which i belongs as mc_i and member j of that set as mc_{ij}. Thus, for all $j, h \in mc_i$ it holds that j and h are neighbors.

In the model each player i plays an action a_i^t, $a_i^t = A, B$, in each of the infinite rounds of play $t = 0, 1, 2, \ldots$, and sends out a signal, consisting of a conditional strategy together with a punishment potential. Such a conditional strategy consists of an action player i would be willing to play, conditional on his neighbors also playing this action.

Each individual i is equipped with a maximum punishment potential (pp_i) each period t, i.e. a maximum amount with which he is able to reduce the payoff of his neighbors through punishing some of these neighbors. The values pp_i, $i = 1, \ldots, N$, are independent realizations of a random variable θ with distribution Θ. We assume the support of Θ to be on a subset of $[0, \theta^{\max}]$. Thus the punishment potentials are a source of heterogeneity in the population. We assume that having an effect of -1 on the payoff of a neighbor through punishment, costs the punisher $\frac{1}{c} < 1$, which reflects the assumption that violence hurts the person being punished more than it costs the punisher.

From the game at time t in which all players set an action, each player i gets an economic (direct) payoff of π_i^t. The *total* payoff of an individual in a period depends on his *economic* payoff and the punishment he receives and hands out: $\Pi_i^t = \pi_i^t - punr_i^t - \frac{puna_i^t}{c}$, where $punr_i^t$ denotes the amount of punishment received at time t by player i and $puna_i^t$ denotes the amount of punishment administered at time t by player i. The constraint on administered punishment is that a player in each period t can at most apply his entire punishment potential to punishment, i.e. $puna_i^t \leq pp_i$.

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. In each subsequent round of play, with probability $p \in \left(\underline{p}, 1\right)$, $\underline{p} > 0$, 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 send out a signal (consisting of a conditional strategy and a punishment potential) only to their neighbors. Of the players who get a learning draw, those that are not yet member of any c.a. can decide to form a new c.a. with other players that are not yet member of any c.a. Players that are not a member of any c.a. and do not get the learning draw keep playing the action they were playing at time $t-1$ at time t and sending the same signal they sent at time $t-1$ again at time t. They cannot join any c.a. at time t. A players that does not get a learning draw, but is already members of a c.a. simply takes the rules the c.a. tells him to follow as given and follow these rules. However, he can decide to become a member of another c.a. and leave his current c.a., when more than one c.a. prescribes its set of rules to this player.

The choice of an action a_i^t yields a stage game. An action $a_i^t = A$ results (at time t) in a (economic) payoff $\alpha > 0$ to individual i, but it imposes a negative externality $-\lambda < 0$

[1] An edge from a to b is distinct from an edge from c to d if $(a \neq c$ or $b \neq d)$ and $(a \neq d$ or $b \neq c)$.

upon i's direct neighbors, while it yields a non-negative externality of $0 \leq \mu < \lambda$ to all other $N - m - 1$ players. Individual players observe neighbors administering negative externalities to them , but do realize that there is also a positive externality coming from distant neighbors. An action $a_i^t = B$ results in a payoff $0 < \beta < \alpha$ to player i and imposes no externalities on the other players. From this we immediately see the dilemma the individual player faces. The individual players prefer action A to action B. However, if a substantial number of their neighbors also play A, both this individual player and (some of) his neighbors would be better off if all were to play B. This dilemma drives the evolution of authorities, because individual players will set up authorities to overcome the dilemma.

Now, if we define the number of players in the population playing action A at time t to be $n^t \in \{0, 1, \ldots, N\}$, and the number of neighbors of player i playing action A at time t as $n_i^t \in \{0, 1, \ldots, m\}$, $n_i^t \leq n^t$, we get the payoff $\pi_i^t(A)$ to player i at time t from playing action A

$$\pi_i^t(A) = \alpha + \left(n^t - n_i^t - 1\right)\mu - n_i^t\lambda = \alpha - \mu + n^t\mu - n_i^t\left(\lambda + \mu\right).$$

and the payoff $\pi_i^t(B)$ to player i at time t from playing action B

$$\pi_i^t(B) = \beta + \left(n^t - n_i^t\right)\mu - n_i^t\lambda = \beta + n^t\mu - n_i^t\left(\lambda + \mu\right).$$

We investigate two cases. In case 1, the total payoff to the entire population is larger when all players choose action A, i.e. $\alpha - m\lambda + (N - m - 1)\mu > \beta$ (and thus $m\lambda - (N - m - 1)\mu < \alpha - \beta < m\lambda$). In this case, playing A is the dominant strategy for the individual player, since he does not realize that he receives a positive externality from distant players. However, the individual is willing to switch to playing action B whenever (some of) his neighbors (credibly) indicate that if he switches, they will do the same thing.

We can think of this case as action A representing 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 your 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 (μ). However, restricting access creates a direct negative externality (λ) to the neighbors, since they are no longer allowed to walk across the property. Now think of a group of neighbors who have all put up a fence around their properties and are all frustrated that they cannot access the other's properties. In case 1, the members of this group of neighbors are better of when they all play action B and give each other access to their properties. So, when they sit together and cooperate, they will play B. But then the positive influences from specialization are lost, yielding an inferior outcome for the population as a whole.

Another example is building an irrigation canal for one's own field, which increases the production of the individual and hence generates wealth for all players in the population, but which reduces the amount of water flowing to neighboring fields substantially and

hence imposes negative local externalities. The alternative action is not building the canal. For coalitions of individuals larger than some substantial minimum size the effect of the positive distant externalities nullify the large negative local externalities at which point every group member playing A is better than having every group member playing B.

In case 2, the total payoff to the entire population is larger when all players choose action B, i.e. $\alpha - m\lambda + (N-m-1)\mu < \beta$ (and thus $\alpha - \beta < m\lambda - (N-m-1)\mu < m\lambda$). In this second case, A is still the dominant action for the individual. Still, groups of players can gain from cooperating by coordinating on action B. When they do so, this is also optimal for the population as a whole. This case represents the standard public goods problem.

We assume that each individual knows the individual payoff to himself of any actions he and his neighbors might play in the stage game. Furthermore, each player i knows the set mc_i : he observes the conditional strategy and punishment potential of his neighbors. He also sees which of his neighbors are handed the learning draw at time t. Thus, a player records which of his neighbors are not playing in accordance with their conditional strategy. Furthermore, he can infer which players deviated because they were not yet handed a learning draw, which players needed to change their action to be in accordance with the conditional strategy, and which players were handed a learning draw but still didn't change their action to be in accordance with their conditional strategy. The individual is not aware of the existence, actions played, and payoffs of anyone else. He plays adaptive given the knowledge and information he has.

3.2. The Behavior of an Authority.

The defining feature of an 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 individual can only punish its nearest neighbors, a central authority can use the combined punishment potential of its members on any of the members of the central authority or on any of the neighbors of the members.

As to the set of rules that a central authority advocates to its members (and possibly to non-members that are neighboring members), 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

[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 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

makes these rules common knowledge within the authority. 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. The sequence of decisions and payoffs each period is then as follows below. Note that all players make decisions simultaneously, i.e. they are not informed of the actions of others in the current period before they play themselves.

1. As stated above, an individual that gets a learning draw and is not a member of any c.a. can set up a c.a. with one or more neighbors that also receive learning draws as long as they are also not members yet of any c.a.. Note that all players in mc_i for a certain i have to get the learning draw simultaneously in order to actually set up a c.a.
   Each central authority decides upon a new set of self-confirming rules to be adhered to next period and announces this set of rules to all its members and their direct neighbors. The set of rules is determined by the members of the authority by mean of voting under a complete veil of ignorance. Alternatively, this process can be thought of as the c.a. setting rules *as if* its members vote under a complete veil of ignorance. We denote the set of individuals who are members of the central authority s at time t by S^t. Furthermore, we denote the set individuals that are not members of s, but are located adjacent to a member of s at time t by S^t_{NBS}.

2. At time t, all individuals in the set $S^t \cup S^t_{NBS}$ observe the rules that central authority s proposes. The individuals in S^t respectively S^t_{NBS} that get a learning draw can decide whether to remain (become) a member of his current c.a. or whether to become (become or remain) a member of another c.a. whose rules he observes or not to be a member of any c.a. The set of individuals choosing to be a member of the central authority s at this point in time is then denoted by S^{t+1} and the individuals in this set automatically remain a member of s until this stage next period.
   As stated above, the individuals that do not get a learning draw and are not yet a member of any c.a. cannot join a c.a.. The individuals that do not get a learning draw but are a member of a c.a. can only decide on whether to remain with their current c.a. or whether to switch to another c.a..

3. Individuals take an action in the stage game. Those that do not receive a learning draw, are members of a c.a. and do not switch to another c.a., take the action prescribed by the rules of his c.a. in that period. Those who do receive a learning draw choose an action.

4. Each central authority s at time t observes all information observed by all the individuals in the set S^{t+1}. Based on all this information, the c.a. decides which player(s) will be punished in accordance with its rules.

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.

5. Each central authority s gives instructions to the individuals in the set S^{t+1} as to which player to punish and how severely to punish this player in accordance with its rules.

In order to enforce its announced rules whatever these rules are, a central authority has to find an enforcement mechanism. We focus on enforcement mechanisms in the context of the equilibrium concept of self-confirming equilibrium, as defined by Fudenberg and Levine (1993). One 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 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 1 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. Results.

We want to characterize the (self-confirming) equilibrium of the total model.

Consider first the circumstances under which a c.a. will form. Individual i, being rational, but playing adaptively, is willing to set up a c.a. with a subset of his neighbors when this increases his expected payoff that period. Individual i can only know that a c.a. will increase his expected payoff if he knows what the effect of his own action will be on the payoff of the all the individuals with which he may want to form a c.a. If he does not have this information, he does not know what set of rules will be optimal for everyone and will not participate in the formation of a c.a. (he could not trust the information given by his neighbors because he would then not have any way of knowing whether they have an incentive to lie. Hence that information could be cheap talk). Because of the condition that player i has to be aware of the effect of his actions on the payoff of all other individuals who may want to form a c.a. together with i, a c.a. can only be set up by the set of mutually connected individuals of player i, mc_i. Moreover, all individuals in mc_i have to get the learning draw at the same time t, in order to form a c.a. s with $S^t = mc_i$. We now have the following result:

Theorem 4.1. *Rank the k mutually connected neighbors of i from high (mc_{i1}) to low (mc_{ik}) according to their punishment potential. When the combined punishment potential of the neighbors labelled 2 to m is greater than $\alpha - \beta$, there is positive probability*

that a central authority around player i will form whenever $k \geq \left\lfloor \frac{\alpha-\mu-\beta}{\lambda+\mu} \right\rfloor + 2$. When $k < \left\lfloor \frac{\alpha-\mu-\beta}{\lambda+\mu} \right\rfloor + 2$, no c.a. will emerge.

Proof.
The difference in payoff for an individual in playing either A or B equals $\alpha - \beta$. When an authority is able to administer a punishment of this magnitude to any player that deviates, an authority can form because it can more severely punish an invidual than an individual can benefit from deviating. This *pp* has to be generated by the $k-1$ least powerful members (i.e. the players with the lowest *pp*) of the authority with m members. This way, the coalition of these k players is able to punish deviation by the most powerful player severely enough to make it unprofitable.

To form an authority with player i participating, (some of) i's neighbors have to indicate that they are willing to play the same action that player i is indicating he would like to play. Since $\alpha > \beta$ any player that gets the learning draw will play action A in the next round of play. What action he will signal depends on the values of the parameters α, β, λ and μ. A player compares his current payoff from playing A, $\pi_i^t(A) = \alpha - \mu + n^t\mu - n_i^t(\lambda+\mu)$ with his payoff from playing B, $\pi_i^t(B) = \beta + n^t\mu - n_i^t(\lambda+\mu)$. He will signal that he is willing to play B if enough of his neighbors that are currently playing A are also willing to switch to B. Equating

$$\alpha - \mu + n^t\mu - n_i^t(\lambda+\mu) = \beta + n^t\mu - \widetilde{n}_i^t(\lambda+\mu),$$

with $\widetilde{n}_i^t$ being the variable for which to solve, results in $\widetilde{n}_i^t = n_i^t - \frac{\alpha-\mu-\beta}{\lambda+\mu}$. Thus, player i signal that he is willing to form a B playing central authority, if at least $n^* := \left\lfloor \frac{\alpha-\mu-\beta}{\lambda+\mu} \right\rfloor + 1$ [4] of his A-playing neighbors also signal that they are willing to form a B playing c.a. The player knows that if enough players switch to action B, a central authority with enough punishment potential to punish deviators will be formed. Thus he sees he is better off than by advocating to form a c.a. if enough of his neighbors are willing to join in, than by just playing A forever.

Note that, depending on the values of α, β, λ and μ, it is possible for n^* to take any integer value. Because each individual who agrees to set up a c.a. has to observe the conditional play of all others wanting to set up the c.a., the maximum number of individual who can set up a c.a. equals k. Therefore, we focus on the inequality $k \geq n^* + 1$ which yields $\left\lfloor \frac{\alpha-\mu-\beta}{\lambda+\mu} \right\rfloor + 2 \leq k$. Thus, at value of k below this lower bound, no central authority will emerge. At values of k above this lower bound, c.a.'s emerge when the prospective c.a. can generate enough pp, as mentioned above. For this range of parameter values, individual players will play A and signal that they're willing to play B if $n^* + 1$ of the k mutually connected neighbors are also willing to do so. Then each individual in the set of mutually connected individuals around i knows that his payoff will increase by setting up a c.a. because he knows that all the members of that c.a. would play B. This c.a. will then be set up as soon as at least n^* mutually connected neighbors of i who have sufficient punishment potential all obtain a learning draw together with i himself. □

[4] $\lfloor . \rfloor$ denotes the entier function, i.e. $\lfloor z \rfloor = \max\{x \in \mathbf{Z} \mid x \leq z\}$.

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 factors pp_i are random draws from a specific class of distribution.

Corollary 4.2. *When the population is large enough and when the heterogeneity in the population with respect to the punishment factor c_i is such that the distribution Θ of the punishment factor puts positive weight on values above $\frac{\alpha-\beta}{(k-1)}$, and $k \geq \left\lfloor \frac{\alpha-\mu-\beta}{\lambda+\mu} \right\rfloor + 2$, at least one central authority will emerge.*

Proof.
When there is a strict positive probability that there are some player i in the population with $pp_i > \frac{\alpha-\beta}{(k-1)}$, there is a positive probability that the $k-1$ weakest members of the set mc_i have a combined punishment potential of at least $\alpha - \beta$. (Note that by assumption every player belongs to exactly one set of mutually connected neighbors.) Thus, when the population grows large, there will always be a set of mutually connected neighbors with enough pp present somewhere in the population, i.e.

$$\lim_{N\to\infty} \Pr\left(\exists\ mc_i \mid \sum_{j=2}^{k} pp_{mc_i j} > \alpha - \beta \right) = 1,$$

satisfying the first condition of Theorem 4.1 for a large enough population. When the second condition of Theorem 4.1 is also satisfied, i.e. when $k \geq \left\lfloor \frac{\alpha-\mu-\beta}{\lambda+\mu} \right\rfloor + 2$, the proof of this Theorem leads to the above result. □

This Corollary says that in a population that is very heterogeneous w.r.t. the punishment potential, a central authority will emerge when the population is large enough. In a population that is homogeneous w.r.t. punishment potential, a similar argument shows that a c.a. will emerge if the homogeneous punishment potential is high enough, i.e. is $pp_i > \frac{\alpha-\beta}{(k-1)}\ \forall i$.

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 will have a set of self-confirming rules that maximize the combined payoff in period t for the individuals in S^t. The effect of action A by individual i on the total payoff of the individuals in S^t equals $\alpha I_{i\in S^t} - \lambda n_i^{S^t} + \mu(|S^t| - ne_i^{S^t} - I_{i\in S^t})$, where

$$I_{i\in S^t} = \begin{cases} 1, & \text{if } i \in S^t, \\ 0, & \text{otherwise,} \end{cases}$$

, $|S^t|$ denotes the cardinality of the set S_t^t and $ne_i^{S^t}$ denotes the number of individuals in the set S^t that are neighbors of i. The effect of action B equals $\beta I_{i\in S^t}$. Denote the action

with the largest total effect on payoff as $\widetilde{a}_i^t$, i.e. $\widetilde{a}_i^t = A$, if $\alpha I_{i\in S^t} - \lambda n_i^{S^t} + \mu(|S^t| - ne_i^{S^t} - I_{i\in S^t}) > \beta I_{i\in S^t}$ and $\widetilde{a}_i^t = B$, if $\alpha I_{i\in S^t} - \lambda n_i^{S^t} + \mu(|S^t| - ne_i^{S^t} - I_{i\in S^t}) < \beta I_{i\in S^t}$. In case of ties, $\widetilde{a}_i^t$ is either A or B with equal probability. The c.a. s can determine the optimal action $\widetilde{a}_i^t$ for each individual $i \in S^t \cup S_{NBS}^t$. If s can find a set of rules under which each individual in the set $S^t \cup S_{NBS}^t$ will play $\widetilde{a}_i^t$ without having to order any punishment nor having any of its members being punished in period t, then that is a set of rules that maximizes the combined total payoff of all its members in period t. We now discuss the different sets of rules that are optimal under different circumstances. We will identify the rules of s under two different circumstances: i) when none of the players in S_{NBS}^t is a member of another c.a. at stage 1 in period t, and ii) when there is a player in S_{NBS}^t who is a member of another c.a. at stage 1 in period t.

1. For each individual $i \in S^t$ the rules are to play $\widetilde{a}_i^t$. There are several possible punishment schemes to enforce this rule. Since decisions are taken simultaneously, an example of such a punishment scheme is to punish the individual highest on a list of individuals (with an arbitrary ranking of the individuals). Since an authority s has formed, it is the case that the combined punishment potential of all the members except the deviator is at least as great as $\alpha - \beta$, which means the punishment is greater than the possible benefit to the potential deviator on top of the list. Thus the potential deviator will not deviate from playing $\widetilde{a}_i^t$. The second player on the list can infer that the first player will not deviate and, hence, will not deviate himself. Iterating this line of reasoning leads to the conclusion that each member of s will comply (and will therefore remain a member) with the rules and punishment scheme set by s. Hence, the believes of the players are confirmed on the equilibrium path of play and we have a self-confirming equilibrium.
   Similarly, each neighbor $j \in S_{NBS}^t$ is told to follow $\widetilde{a}_j^t$ by a c.a. denoted as s unless j does not receive a learning draw that period (because then it cannot comply and punishment is only payoff decreasing), or unless j joins another c.a. in that period whose combined punishment potential is greater than the combined punishment potential of s in that period. These rules ensures that, in case there is no other c.a. than s connected to any member or any neighbor, all members and neighbors of s that receive a learning draw play the actions that maximize the total payoff of the individuals in S^t. No punishment is administered in equilibrium.

2. Focus on a c.a. s for which it holds that $j \in S_{NBS}^t$ is a member of another c.a. and that the combined *pp* of all individuals in the other c.a. is smaller than that of the individuals in s (whenever multiple bordering c.a.'s are present, one can always order the c.a.'s such that this holds). Then, under the optimal set of rules, j is given instructions to join s and to play $\widetilde{a}_j^t$ advocated by s in the stage game. The threatened punishment (by s) for non-compliance is a minimal fraction bigger than the maximum *pp* of the other c.a. Because j knows that s has more *pp* than his own c.a., j will indeed comply and become a member of s. The other c.a., knowing that j will in this case choose to join s whatever punishment it threatens j with, will not punish j for complying with the rules of s, because it knows that it has to enforce this punishment, which only leads to a loss of combined total payoff to its members

(which includes the payoff of j), whilst not affecting the payoff of the individuals in S^t. Conversely, each member $i \in S^t$ will know that if only he deviates by joining the other authority, he will be punished by his own authority by at least $\alpha - \beta$. Each member $i \in S^t$ also knows that if he does not deviate, he will not be punished by the other authority and hence will receive no punishment at all. This will see to it that i remains a member of s. This means that the stronger authority, in this case s, can get all its members and all the neighbors of its members that belong to another, weaker authority, to play the action $\tilde{a}_i^t$ it prescribes to them, whilst not having to punish anyone. Therefore these rules and punishment scheme maximizes the combined payoff of all individuals in S^t. The optimal punishment scheme for the weaker authority is then not to administer any punishment on its members that are neighbors of individuals belonging to stronger authorities, and to prescribe an arbitrary action $\tilde{a}_i^t$ to them. To the individuals in the weaker authority that are not neighboring a stronger authority the c.a. prescribes the set of rules and the punishment scheme mentioned in 1.

Consider the outcome of this interaction between central authorities and individuals. Denote the c.a. with the largest combined punishment potential at time t by F^t, i.e.

$$F^t = \arg\max_{S^t} \sum_{i \in S^t} pp_i.$$

Denote the punishment potential of F^t by pp^{F^t}. Following the above mentioned rules, F^t will force all neighbors not belonging to a c.a. yet and receiving a learning draw to become a member of this c.a. next period and will force any neighbor belonging to a c.a. to become a member next period, whilst no member of F^t will leave in period t. This means, that as long as F^t does not encompass all individuals, The punishment potential of the c.a. indicated by F^t will not diminish between time t and $t+1$. Thus, it cannot be that $pp^{F^{t+1}} < pp^{F^t}$. Moreover, there is a strictly positive probability that $pp^{F^{t+1}} > pp^{F^t}$. Note that it does not have to hold that $F^{t+1} = F^t$. These fact on the power of the most powerful c.a., sees to it that the model will not exhibit cyclic behavior. The only ultimate stable outcome is then a single remaining c.a. of which all individuals are members. □

We add several comments to these results. First, as mentioned in the proof, enforcement of the set of rules within a c.a. becomes easier the larger authorities get, since in large c.a.'s there will always be a substantial number of individuals that cannot leave the c.a. at a given time, because they do not have a learning draw. If this group of players is large enough, it generates enough pp to threaten the player on top of the list of 'players to be punished when deviating' severely enough to keep him from deviating. Inferring this, the second player on the list will not deviate and so forth. Insuring compliance with the rules of the authority will then be easier.

Second, 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.

Third, although the model is limited to a stage game with two actions, we argue that the qualitative results carry over to models incoporating 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 3.1. Remember that although in both cases action A dominates action B for the individual player, in case 1 choosing A is optimal for large c.a.'s, though it is not optimal for small c.a.'s, while case 2 represents a standard public good problem, with everybody playing B as the Pareto superior outcome. 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. *When a small c.a. has formed and $\beta < \alpha - m\lambda + (N-m-1)\mu$, i.e. when having all players play A maximizes the total payoff of the entire population, the small c.a. will advocate playing B to its members. It is only when a c.a. grows sufficiently big that it will see the benefits of playing A. Consequently, sufficiently large c.a.'s will tell their members to play A.*

Theorem 4.3 now tells us that the stable outcome in this case will be a single c.a. of which all players are members and which advocates the play of action A to its members.

An illustration of the evolution of a c.a. under the conditions of case 1 is given in figure 1. In the figure, $k = m = 4$, $\alpha = 10$, $\beta = 5$, $\lambda = 4$, $\mu = 1$. All individuals receive learning draws.

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 and send out conditional plays. In period 1 (which is shown), the bottom 4 players were strong enough to set op a c.a., that forces its members to play B and forces the neighbours to play B (the positive externality does not yet outweigh the negative externality in this case). Finally, period 4 is shown, in which all members and neighbours are forced to play A.

5. A game with multiple remaining central authorities.

In this section we focus on a altered version of the model. We assume that non-members of a coalition cannot be punished as severe as the members can be punished, i.e. the

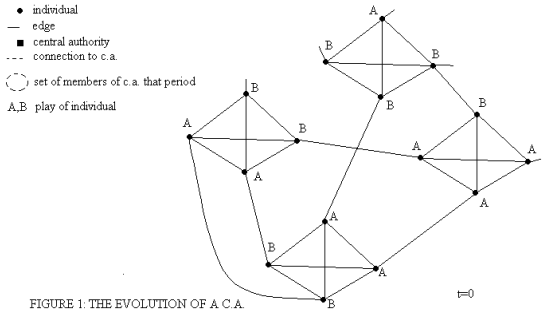

FIGURE 1: THE EVOLUTION OF A C.A.

Figure 4.1:

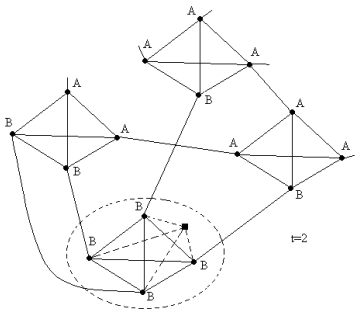

Figure 4.2:

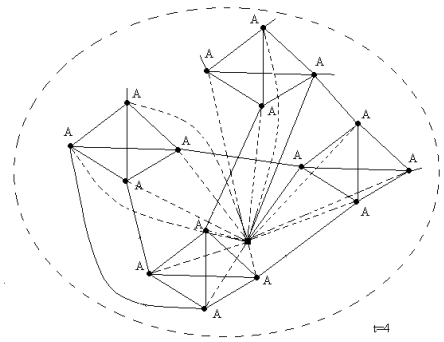

Figure 4.3:

ability to punish non-members is a fraction $0 < d < 1$ of the ability to punish members of a coalition. This means that punishment potential is asymmetric as to whether the punishment is between or within central authorities. One can interpret this asymmetry to mean that it is easier to observe and punish those you know than those you do not, as someone outside a central authority is further away and hence it may take more effort to punish a distant person equally hard as a person nearby. In this setting the following Theorem applies.

Theorem 5.1. *Consider the model with asymmetric punishment potential. When the population is sufficiently large, when the distribution* Θ *of the punishment potential puts positive weight on values above* $\frac{\alpha-\beta}{(k-1)d}$, *and when* $k \geq \left\lfloor \frac{\alpha-\mu-\beta}{\lambda+\mu} \right\rfloor + 2$, *there will almost surely be at least one central authority. In the long run either one single central authority or several central authorities with possibly conflicting rules can remain.*

Proof

The condition $k \geq \left\lfloor \frac{\alpha-\mu-\beta}{\lambda+\mu} \right\rfloor + 2$ is needed to ensure that the formation of a c.a. is profitable to a group of mutually connected neighbors (see Theorem 4.1). Because the distribution Θ of the punishment potential puts positive weight on values above $\frac{\alpha-\beta}{(k-1)d}$, there will almost surely be a mc_i that satisfies the conditions on punishment potential of Theorem 4.1. The argument is the same as that of Corollary 4.2. Thus in the long run, at least one central authority will emerge almost surely.

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.

In example 1, denote all the groups of mutually connected neighbors mc that have the ability to form a c.a. which can threaten its neighbors by at least $\alpha - \beta$ as mc^*. By the assumption on the distribution of Θ, there will be many of these mc^*s for a sufficiently large population. An authority set up by one of these mc^*s is able to form a c.a. which can force its neighbors that are not members of a more powerful c.a. to join also. Now

suppose that in mc_1^*, all individuals get a learning draw in the first period, whereas in all other mc^*s, not all individuals belonging to that mutually connected set get a learning draw simultaneously until a finite period T. Then, the individuals in mc_1^* will form a c.a. and this c.a. will in each period force all its neighbors that receive a learning draw to become members. If all the neighbors of this c.a. in period $t < T$ receive learning draws, then this c.a. will expand 'one layer' at a time, until it incorporates all other individuals. This specific chain of events has a low probability of happening, but there are obviously many more ways in which a single c.a. may emerge. From standard Markov chain theory we then know that any positive probability to end up in a recurrent state (i.e. one single c.a.) suffices to put positive weight on this state in the limit outcome. There is thus positive probability that the model long run outcome is a single c.a. of which all players are members.

In example 2, suppose that through a similar chain of events as in example 1, only two c.a.'s with sufficient pp to force its neighbors to become members have formed. Label these two c.a.'s s and r. In addition, suppose these two c.a.'s have each expanded in such a way that currently (at time t) all individuals in the population are members of either s or r. Remember that the set of individuals that are members of c.a. q is denoted by Q^t, $q = r, s$. Suppose that $\sum_{i \in S^t} pp_i > \sum_{i \in R^t} (d * pp_i)$ and at the same time $\sum_{i \in R^t} pp_i > \sum_{i \in S^t} (d * pp_i)$. Note that such a division of all N individuals over r and s, and hence of punishment potential, can happen with positive probability. Now, neither c.a. can threaten the individuals in the other c.a. with a punishment sufficiently severe to force them to switch membership. Thus, no player switches membership and $S^{t+1} = S^t$ and $R^{t+1} = R^t$, i.e. there is a powerbalance between s and r.

In this case there are multiple optimal sets of rules and punishment schemes. A possibility is that s (r) threatens all players in $S^t_{NBS} \subset R^t$ $(R^t_{NBS} \subset S^t)$ with its maximal total punishment potential, if r (s) threatens any of the members of s (r), while s (r) will administer no punishment to individuals in S^t_{NBS} (R^t_{NBS}) if r (s) does not punish players in R^t_{NBS} (S^t_{NBS}). Depending on their exact size and the parameter values, the action the authorities prescribe to their members can be either A or B. Moreover, it is possible that both authorities prescribe different actions to their members, e.g. for $i \in S^t$, $\widetilde{a}^t_i = A$ and for $j \in R^t$, $\widetilde{a}^t_j = A$. Under these rules, one self-confirming equilibrium is for neither c.a. to threaten the members of the other c.a.. Note that this equilibrium is stable, since equilibrium play at time t does not alter the situation at time $t + 1$ and consequently, time t equilibrium play is also time $t + 1$ equilibrium play.

A similar situation is possible in case the population is divided over more than two c.a.'s. Similar arguments then lead to the conclusion that multiple c.a.'s may coexist forever. □

The Theorem states that several possible limit outcomes are possible. 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. Thus we have path dependency in this altered version of the basic model.

Multiple c.a.'s may yield global inefficiency, 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

and neither c.a. is sufficiently strong to take over the members of the other authority. One example of such an inefficient outcome is give in the proof of the Theorem. Another example is when both authorities are not large enough and both suffer from choosing to promote the action which is not globally optimal, as depicted in Corollary 4.4. Since neither authority will grow, the Pareto superior outcome will not be reached in the limit. The possibility that groups of individuals are 'locked' into an inefficient equilibrium in which all groups loose out, is one way of modelling discrimination or conflicts between groups or regions. Indeed, the possibility that discrimination is the outcome of individual groups pursuing their own interests, with each individual maximizing his own utility by maximizing that of the group as a whole, is also argued in e.g. Frijters (1998).

The possibility of multiple remaining c.a.'s depends heavily on the assumption that all c.a.'s simultaneously set their rules with the other c.a.'s, because only then is it possible to have an equilibrium in which neither c.a. forces the members of another c.a. to join it.

5.1. Selection through mutations.

Theorem 5.1 states that there are many long-run self-confirming equilibria of the altered model. We want to select among these equilibria. We do so by introducing a small probability of mutations in the model. The equilibrium that is selected in this way, is referred to as the *stochastically stable equilibrium*. It 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). Good surveys of this literature are given by e.g. Samuelson (1997) or Young (1998). In the literature, 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, there is the interpretation closest to the biological literature on evolution, which is genetic mutation, i.e. 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, positive 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. This setup allows us to select among the multiple equilibria indicated in Theorem 5.1.

Theorem 5.2. *Suppose the conditions for theorem 5.1 are met. For sufficiently large populations, the state in which there is only one central authority, of which all player are a member, present in the population is the only stochastically stable state of the model.*

Proof.

The structure of this proof is based on Freidlin and Wentzell (1984). We show that the probability of leaving an (equilibrium) 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 (equilibrium) state in which there is only one c.a., of which all players in the population are a member. Then, the latter equilibrium is stochastically stable.

Label the state in which there is only one c.a., of which all players in the population are a member z^*. For the state in which two c.a.'s are present in the population, the two c.a.'s are labelled s and r, and where the mutant is a member of s. It is straightforward to see that the results carry through to equilibria with more than two c.a.'s or where mutants occur in a different c.a. Every player is a member of either s or r and in equilibrium, neither s nor r can expand. We look at the different impact for mutants occurring at different locations. We label a mutant by i.

First, suppose $i \in S^t \backslash R^t_{NBS}$, i.e. i is in the 'interior' of s. This mutant cannot join r. Moreover, the occurrence of this mutant is not observable by r. Therefore, unless the punishment potential of i was such that the balance of power between s and r is in fact disrupted by the mutant (meaning that it is no longer the case that $\sum_{j \in S^t} pp_j \leq \sum_{j \in R^t} (d * pp_j)$), nothing changes. The selection pressure within s will be directed against the mutant (s will threaten the mutant to change his action back to $\widetilde{a}_i^{t+1}$ at time $t+1$) and the mutant, not being able to resist this threat, will change his action to be $\widetilde{a}_i^{t+1}$ at time $t+1$.

Second, suppose $i \in S^t \cap R^t_{NBS}$, i.e. the mutant is on the border of s and is therefore a neighbor of r. Now, the mutant might remain in s and play $\widetilde{a}_i^t$, in which case the equilibrium is not distorted. The mutant might remain in s and play an action different from $\widetilde{a}_i^t$ and not distort the powerbalance between s and r, in which case he is forced by s to play $\widetilde{a}_i^{t+1}$ and the equilibrium is restored. The mutant might remain in s and play an action different from $\widetilde{a}_i^t$ and thereby distort the powerbalance between s and r. This distortion is disadvantageous to s, i.e. it no longer holds that $\sum_{i \in S^t} pp_i > \sum_{i \in R^t} (d * pp_i)$, meaning that r now acquires the power to take over individuals in $S^t \cap R^t_{NBS}$. Obviously, r will start forcing these individuals to be in R^{t+1}, thereby further increasing its total payoff and punishment potential. After some time this leads to the complete elimination of s and only one c.a. of which all players are members, remains.

The mutant might also join r and play a random action. Now, there are again two possibilities. Either the new situation with $i \in R^t$ is also an equilibrium (i.e. there is still a powerbalance between s and r), in which case the system will stay at this equilibrium. Note that in the new equilibrium the punishment potential of r will be (slightly) higher than before and that of s will be (slightly) lower than before. Or, the powerbalance is distorted in favor of r, in which the system has moved out of equilibrium and through a same scenario as above will move towards the equilibrium state in which only r remains and all players are members of r.

When a mutant has led the system to a different equilibrium, at time $t+1$ a new mutant may lead the system to yet another equilibrium in which r has acquired even more power at the expense of s. A series of such sequential mutations will see to it that eventually the balance of power between s and r is distorted in favor of r. Subsequently, r will take over the entire population. Note that this scenario will happen with a prob-

ability that is of order ε. Although it might take multiple mutations to get the system out of equilibrium, it does not take any *simultaneous* mutations.

Disrupting the equilibrium state z^* in which there is only one c.a., takes a number of simultaneous mutations: a single mutation from a state z^* will never get the system to another equilibrium, 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. r. Even the most powerful mutant i, still has a punishment potential $pp_i < \sum_{i \in R^t} (d * pp_i)$, when the population, and thus size of r is sufficiently large, since punishment potential 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 we see that $l \geq 2$ mutants are needed to upset the equilibrium z^*. 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 out-of-equilibrium to an equilibrium is relatively fast. Thus, we can neglect the time the system spends out of equilibrium. Standard Markov chain theory now suffices to deduce that the system is in state z^* a fraction $\frac{\varepsilon}{\varepsilon+\varepsilon^l}$ of the time. Taking the ε to 0 in the limit, i.e.

$$\lim_{\varepsilon \downarrow 0} \frac{\varepsilon}{\varepsilon + \varepsilon^l} = 1$$

shows that in the long run, when the mutation rate is taken to 0 in the limit, the system will be in state z^* a fraction 1 of the time. □

Hence, although multiple equilibria are present in the altered model, introducing a small probability of random mutations may serve as a selection device, selecting the equilibrium with only one c.a. as the only stochastically stable equilibrium. A typical path to this stochastically stable equilibrium may look like this. First, the system rapidly moves from the initial state to an equilibrium state, not being the stochastically stable equilibrium. Then the system remains in this equilibrium for some periods, until a mutant takes the system to a different equilibrium. Again, the system stays at this new equilibrium for some time, until another mutation occurs. After being at different equilibria in a row, and possibly sometimes moving back-and-forth, i.e. moving from equilibrium a to equilibrium b and moving back to a again, the equilibrium play is distorted and the system moves out of equilibrium. From there, it takes a relatively short period of time for the system to move to the stochastically stable equilibrium.

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 benefits. As such, central authorities promote cooperation. The defining feature of a central authority is that it is able to communicate directly with all its members whereas each individual only has a limited number of neighbors with which it can communicate. This 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 assume that central authorities arise when many individuals want to promote the same set of rules, because these rules seem to generate higher payoffs, but cannot act according to these rules in the absence of a (credible) commitment device. Authorities then arise in 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 branche organisations in 19th century Holland, that many current institutions that encorporate 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, of which all players are members, is present in the population. In a slightly altered version of the model, we see that multiple authorities can co-exist in equilibrium. However, if we select among this multitude of equilibria allowing for mutations, we find that the only ultimate stochastically stable state is the one in which there is only one central authority in the population.

Many arguments on internal decision making found in the social choice literature were not incorporated in the model. An interesting extension to the present set-up is, for instance, to allow for a more realistic internal decision-making mechanism which would allow sub-coalition formation and lobbying activities to take place. Reviews of the rent-seeking literature indeed suggests that special interest groups lobbying within an authority have some success (Mitchell and Munger (1991) or Austen-Smith (1994)). Another interesting extension would be 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. Strategic interaction between central authorities in an evolutionary model 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.

Frijters, P. (1998). Discrimination and job-uncertainty. *Journal of Economic Behavior and Organization 36*, 433–446.

Fudenberg, D. and D.K. Levine (1993). Self-confirming equilibrium. *Econometrica 61*, 523–545.

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 interset 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.

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.

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.