Appendix C The Role of Theory: Reputation-Based Institutions

Theory is unavoidable in positive institutional analysis. Implicitly or explicitly, a student of institutions resorts to theory to guide the selection of issues and to identify relevant factors and causal relationships. Theoretical assertions about the importance of exchange, polities, and the harnessing of coercive power direct the investigation of the institutional foundations of agency relationships, property rights security, impersonal exchange, and the mobilization of resources for collective action. The investigation itself is directed by a concept of institutions central to which are intertransactional linkages and the associated institutional elements, self-enforceability, and the nature of institutional development as a historical process. Game theory tells us what to look for in considering and evaluating the self-enforceability of institutional elements in a given environment.

Theory also makes another important contribution. By pointing to the general principles that underpin the operation of institutions that can lead to a particular outcome, theory indicates that institutions—and the history they induce—are not random. Context and contingency are important, but institutions generating similar behavior in the same central transactions are subject to the same forces and have to respond to the same considerations regardless of the particularities of time and place. Institutions that achieve the same outcomes have to mitigate the same problems that are implied by the inherent attributes of the central transaction under consideration and by the general context. Hence theory is useful in directing our search for evidence that facilitates forming a conjecture about and evaluating whether a particular institution prevailed at a particular time and place.

This appendix delineates the forces that shape the general attributes of reputation-based private-order economic institutions prominent in the historical examples analyzed in this book. It emphasizes the generic implications of these forces in directing a context-specific analysis aimed at identifying the relevant institution. The discussion highlights the distinction between a game-theoretic and an institutional analysis: game theory considers possible equilibria in a given game; institutional analysis considers the man-made, nonphysical factors that generate regularities of behavior while being exogenous to each individual whose behavior they influence.

Consider a situation in which the inherent characteristics of the central transaction can be captured as a version of the one-period repetition of a prisoners' dilemma or one-sided prisoner's

dilemma game (both games are described in Appendix A). In such games a weakly dominant strategy is for at least one player to take an action to which the other player's best response causes the game to reach a Pareto-inferior outcome. These actions are usually referred to as "cheating" and the alternatives that lead to better outcomes as "cooperating" or "playing honest." Such situations are everywhere in the economic, political, and social spheres. In the economic sphere, they are inherent in voluntary exchanges of goods and services (see Greif 2000) and involuntary exchanges, such as the poaching of a firm's workers by another firm (see Kambayashi 2002). They are inherent in the relationship between the government and economic agents (see Kydland and Prescott 1977), as well as the relationships between owners of common resources (see Ostrom 1990). In short, such situations are central to what we model as voluntary or involuntary exchange, agency relationships, collective action, and free-riding problems. The theory behind this simple game can therefore be generalized to study multiple real-world situations.

In the absence of exogenous enforcement, can an individual be nevertheless induced to take actions that are not in his short-term economic interest? In the game-theoretic formulation, how are individuals motivated to take action off the equilibrium path of the (unique) one-period game? Why would one cooperate or be honest despite the fact that cheating is economically rational if the game is repeated just once?

Two lines of analysis consider ways through which the social norms of cooperation and honesty can be sustained when at least some individuals care only about their material well-being.[1] The first examines situations in which there is *asymmetric information* regarding the propensity of various agents to cheat. In the current context, in these situations there is a probability that a player is "good," in the sense that he would not cheat (e.g., under any circumstances), despite economic temptation. Whether a particular individual is "good," however, is private information. Each player knows if he is "good," but the others do not. Cooperation can be curtailed by what is referred to as *adverse selection*; one's decisions depend on her privately held information in a manner that adversely affects those who are uninformed.

[1] For surveys of these lines of analysis and reputation-based institutions, see Greif and Kandel (1995); Klein (1996); Greif (2000); Hart (2001); and Dixit (2004). For important contributions and insights, see Milgrom and Roberts (1982); Shapiro and Stiglitz (1984); Kreps et al. (1982); Kreps (1990a); Williamson (1985); Joskow (1984); Nelson (1974); Klein and Leffler (1981); Shapiro (1983); and Akerlof and Yellen (1986).

The analysis thus focuses on why and how individuals can be motivated to cooperate in aspiring to gain reputations as players of the "good" type.

The second line of analysis examines situations in which there is a moral-hazard problem in which there are only "bad" agents who always maximize their material well-being. The focus of the analysis is on why and how the expectation of future interactions can motivate such individuals to cooperate. In either line of analysis, a player's reputation is defined as a function from the history of the game to a probability distribution over his strategies.

These two lines of analysis are not mutually exclusive, but the distinction between them is analytically useful. The game-theoretic analysis of reputation-based private-order institutions in these situations focuses mainly on particular intertransactional linkages—the same transaction over time or the same transaction among different individuals. Accordingly, this is also the focus of the subsequent discussion, although I note that focusing on these particular linkages highlights the potential roles of various others, through social exchange, organizations, and the use of violence.

C.1 Adverse Selection: Incomplete Information

Incomplete information models are useful in studying situations with asymmetric information and adverse selection. It is assumed that at least one of the interacting individuals knows his type, while the others do not. Nature moves first and selects with some probability the types of the various players. The ex-ante probability distribution over types is common knowledge but the selection itself is private information. The game is repeated for a finite or infinite number of periods. In this case, even if the actual number of "good" players is very small, they can nevertheless have a large impact on the equilibrium behavior of all agents. Particularly, cooperation can often be achieved even if many of the players are bad type. (Kreps et al. 1982.)

To grasp the intuition, consider a one-sided prisoner's dilemma game in which agents and merchants are randomly matched each period and past actions are observed by all players. A "bad" agent may find it optimal to mimic the behavior of a "good" agent for a period of time and refrain from cheating. Cheating in the first period implies losing gains from future cooperation or the ability to cheat again (because merchants will update their beliefs about that agent's type and not rehire him). A strategy of acting like a "good" type for some periods and then cheating later implies gaining from cooperation for some period as well as from cheating. But because cheating

is postponed, this strategy implies a higher payoff only when the agent's time discount factor and the gain from cooperation are sufficiently high relative to the gain from cheating. If this is the case, a "bad" type finds it optimal to mimic, at least for a period, the behavior of a "good" type. Given this behavior, it is optimal to interact with him, although, with some probability, he may cheat in the future. Incomplete information and the conditioning of future behavior on past conduct implies the possibility of a "pooling equilibrium" in which "bad" types and "good" types behave the same way—honestly—for many periods. Indeed, even if the probability that an individual is honest is very low, if the game is played for a sufficiently large number of periods, the extent of cooperation progressively approaches the first best (a situation in which a Pareto optimal outcome is achieved by everyone behaving honestly).[2]

By highlighting the exact way that cooperation can prevail and the conditions required for this to be the case, theory facilitates the evaluation of its relevance to the institution we seek to identify. Among the questions that the theory highlights are the following: Does the broader historical context reflect such factors as religious beliefs and a culture of guilt that could have led people to believe that some agents are inherently good? How are expectations for future interactions generated? If a player plays the stage game with different partners in different periods, how do future partners learn about his past conduct? Why are agents who cheated unable to assume a new identity (as is done in modern economies when an owner changes the name of his firm), allowing him to reestablish agency relationships despite having cheated in the past?[3]

Similarly, we can evaluate a conjecture about the relevance of an incomplete-information, reputation-based institution by looking for the generic implication of this theory. In the absence of other considerations, individuals should cheat in their old age. Is this the case? That merchants update their beliefs about each agent's type implies an economic payoff to an agent who acted in a manner that caused merchants to update their beliefs favorably. Do we see agents attempting to signal their types by taking such costly actions as contributing to charity or

[2] A similar result follows when we assume that all agents are trustworthy in the sense that each incurs some intrinsic psychic cost if he cheats. The distribution of these costs is common knowledge, but one's intrinsic cost is private information. Cooperation can be sustained on the equilibrium path as low-cost ("bad") agents mimic the behavior of high-cost ("good") agents for some periods to acquire reputations that they will eventually exploit by cheating (Hart and Holmstrom 1987).

[3] In Tadelis's (1999, 2002) model of a firm's reputation, described later, in equilibrium agents who change their identities earn a lower income.

acting as though they were religious? In a pooling equilibrium, a player conditions his actions toward another player only on the other player's past conduct, not on other considerations, such as his ethnicity. Is this actually the case? Is it the case that agents who cheated in the past are never rehired?

Asserting that reputation reflects incomplete information is intuitively appealing. But a general theoretical insight highlights the inherent difficulty of empirically substantiating the relevance of cooperation based on it. Incomplete-information models are very sensitive to the specification of incomplete information, but the researcher cannot observe the details and nature of incomplete information in a particular setting. Hence we can usually account for particular behavior, as well as its absence, as reflecting some unobserved diversity of types in the population.[4]

C.2 Moral Hazard: Complete Information

When all agents are "bad" types, motivating them to be honest can be achieved based on the lure of future reward. Conditioning future reward from cooperation on past conduct is used to motivate behavior. The basic theoretical insight highlights the importance of increasing the reward for honesty and decreasing the payoff following dishonest behavior. The larger the discrepancy between the two, the more honest behavior can be generated.

The basic intuition is captured in the Folk theorem of repeated games, which can be illustrated by considering an infinitely repeated prisoners' dilemma game (see Appendix A, section A.3). Assume that for each player the net present value if both players cooperate in every future period is higher than the net present value for a player if he cheats in the current period (while the other cooperates) and subsequently gets the payoff associated with cheating by both players every period. If this is the case, there is an equilibrium in which cooperation is achieved. In this equilibrium, future cooperation is made conditional on past conduct: following cheating by either player, both players' strategies call for cheating forever. The threat of cheating is credible—it is part of a subgame perfect equilibrium—because cheating is each player's best response to the expected behavior of cheating by the other.

That the promise of future reward can potentially support cooperation is the starting point rather than the conclusion of institutional analysis. Theory informs us about the conditions

[4] For a discussion and an example, see Hart (2001).

required for cooperation based on the long hand of the future. Institutional analysis is about the particularities of the institutional elements that made, or failed to make, these conditions a reality at a particular time and place. It is concerned with understanding how expectations for repeated interactions were generated and among whom. Why did an individual stand to gain more from cooperation than from cheating? Why could someone who cheated one partner not establish an equally profitable cooperative relationship elsewhere? Why, and how, was a cheated individual motivated and able to circulate this information? How were those supposed to respond to cheating to gain this information? What made the threat of punishment credible?

Generic theoretical insights provide a valuable guide when attempting to identify the relevant institutional elements and other factors. The following discussion considers such insights about the endogenous construction of future rewards. It then analyzes the credibility of maintaining relationships and the credibility of threats of future bilateral and multilateral punishment and renegotiation following cheating. It also provides insights into the generation and distribution of information, imperfect monitoring, the cost of reputation-based institutions, and endogenous intertransactional linkages and organizations.[5]

C.2.1 The End-Game Problem

Conditioning future reward on past conduct influences decisions about current behavior. Such conditioning, however, requires generating the belief that a future reward will be forthcoming. The distinction in the equilibrium set between finite- and infinite-horizon games reveals a fundamental difficulty in doing so.

Consider a stage game with a unique equilibrium, such as the prisoners' dilemma game. The Folk theorem establishes that if the game is repeated an infinite number of periods and the players are sufficiently patient (i.e., they place a high weight on future periods' rewards), cooperation can be sustained by conditioning future cooperation on past cooperation. Cheating implies gaining today but losing all gains from future cooperation.

If such a stage game is repeated a finite number of periods, cooperation based on the promise of future reward from cooperation cannot be sustained—it is not an equilibrium outcome. Intuitively, the best one can do in the last period is to cheat. After all, a player cannot

[5] Within an institution, all these considerations are interrelated. Although discussing them sequentially focuses attention on each, it comes at the cost of commenting only on these interrelations.

be punished for cheating in the future if the game ends. Expecting cheating, the other player will not cooperate either. Anticipating that neither player will cooperate in the last period, the best each player can do is to cheat in the next-to-last period. Following this logic, the equilibrium continues to unravel backward in this manner, implying that cooperation in any period is not an equilibrium outcome. This is the end-game problem.

In general, uncertainty can mitigate the end-game problem, because the infinite-horizon game is analytically equivalent to a game with an uncertain final period. Specifically, when there is a constant or sufficiently low per period probability that the repeated game will be terminated at the end of each period, the repeated game is analytically equivalent to one in which the stage game is infinitely repeated. The only impact of the uncertainty is to decrease the time discount factor (Telser 1980). In this case, although the game will certainly end at some point, uncertainty about the final period implies that there are always (expected) gains from future cooperation that can be lost due to cheating.

The possible importance of this factor in sustaining cooperation based on the long hand of the future notwithstanding, individuals' horizons tend to become shorter in their old age—and old age is difficult to conceal. Hence the end-game problem becomes relevant when we model interactions among individuals. Institutions based on the future have to guarantee that the future is long enough. Understanding an institution based on future reward therefore requires identifying why there is still enough of a future reward to motivate honesty whenever one has to decide whether to cheat.

Theory suggests several ways to achieve this. The first, relevant particularly to one-sided prisoner's dilemma games, is altering the time profile of gains from cooperation.[6] The strategy specifies that, if one does not cheat, his share in the gains from cooperation will increase as time goes by, or he will get a bonus upon retirement. If commitment to such payments can be made, this strategy can be an equilibrium with cooperation. Such an endogenous alteration of the division of gains from cooperation can be done either by distributing the gains in the transaction under consideration (such as through wage payment in agency relationships) or by linking it to other transactions (such as social exchange).[7] In late medieval Genoa, for example, noble

[6] This can also be done in asymmetric prisoners' dilemma games with transferable utilities.
[7] Technically, we study such linkages using finitely repeated games with complete information in which (unlike the prisoners' dilemma game) there are multiple equilibria in the stage game. For a classical analysis, see Benoit and Krishna (1985).

merchant families rewarded agents who had served them for years with marriage into the nobility.

Another way to mitigate the end-game problem in either one-sided prisoner's dilemma or prisoners' dilemma games is by endogenously linking the reputational considerations of individuals from different generations. Intertemporal linking of utility streams can create the equivalent of entities with infinite life-spans, or at least entities whose per period probability of survival is high enough to allow cooperation. When others condition their behavior on an individual's past conduct and the individual's welfare depends on this behavior, it is possible to motivate the individual to cooperate even in his last period. Families, dynasties, family firms, and other innate social units served as entities with infinite life-spans that mitigated the end-game problem for their members in many historical episodes. Among the Maghribi traders, a reputation-based institution was based on intergenerational linkages that took advantage of an individual's concern about his descendants' well-being despite his own finite life-span. In modeling the relationships between merchants and rulers in Chapter 4, I assumed that the ruler had an infinite horizon in order to capture the dynastic nature of the state during the period.

In modern economies, other endogenous entities, such as firms with identities distinct from that of their owners, play a similar role. Tadelis (1999, 2002) uses a model combining moral hazard and adverse selection to explore how organizations that separate identities from entities can motivate individuals to cooperate in their old age. He assumes that a firm's reputation reflects the past ability and actions of its owner, who can sell the firm without the knowledge of the firm's clients. When buyers of the firms' products are willing to pay more for a product from a reputable firm than from another, reputation is valuable. Hence an owner of a reputable firm can find it optimal not to cheat in his old age, because the loss of reputation would decrease the value of the firm's name, an asset he can sell. The analysis also implicitly highlights the role of auxiliary organizations and the associated beliefs, such as those ensuring that a firm cannot adopt the name of another.

The end-game problem is an issue only with respect to players who can cheat. If a player who can cheat lives a long time and is sufficiently patient, there can be an equilibrium with cooperation even if the other players have short life-spans. Cooperation is based on the players with short lives conditioning their behavior on the other players' actions toward their predecessors.

To understand this causal effect, consider a game between a firm with an infinite life span and its workers. Each worker is known to live for several periods, after which he dies and is replaced by another. This game is a version of the one-sided prisoner's dilemma game. In each period the workers can first decide whether to provide their labor as an input to the firm. If they do, the firm can decide whether to pay the promised wages. When each worker knows only his private history, the end-game problem implies that there is no equilibrium with the provision of labor input and wage payments, because the firm's optimal strategy is to not pay a worker in his last period. An equilibrium with the provision of labor and wage payment exists, however, if the firm's past conduct is public information among the workers and if the firm's expected future gains from production (after paying wages) is sufficiently high. In this equilibrium the threat of future workers punishing the firm (by not working if it ever fails to pay a worker) motivates the firm to pay (see Bull 1987; Cremer 1986; Kreps 1990b; and Tadelis 1999, 2002).

Intergenerational links within an organization composed of overlapping generations of members can also mitigate the end-game problem in relationships among its members. Consider a situation in which individuals have a predetermined life-span. Every year the organization recruits a new member to replace one who just died. Members of the organization interact in a prisoners' dilemma type of situation by either contributing effort or not. Actions are observable. The best a member can do is to provide no effort in the period just before retiring. But in the last period, the member can still be rewarded or punished based on his effort. Hence a strategy in which young members work hard and older ones do not but are nevertheless compensated can support some (albeit not an optimal) level of cooperation. Younger workers are motivated to work hard and reward older workers who do not contribute effort because otherwise the play of the game will revert to no cooperation (Cremer 1986).

Organizations can also mitigate the end-game problem in interactions between organizations, as the analysis of the community responsibility system illustrates. Intergenerational relationships within communities were part of an institution that enabled these communities to commit to act as if they had infinite life-spans, even though they were concerned only with the welfare of their finitely lived members.

C.2.2 Endogenous Payoffs

A necessary condition for the promise of future reward to foster cooperation is that the net present value of the gain from cheating and the implied utility stream in the following periods be less than the net present value from cooperating. Understanding a reputation-based institution requires identifying the way in which this condition has been fulfilled endogenously. Reputation-based institutions manipulate one's gains from various actions and outside options to enable cooperation. Theory suggests various ways in which payoffs can be endogenously manipulated and the relationship between these payoffs and the environment.

Consider a situation in which employers and employees are randomly matched to play a one-sided prisoner's dilemma game. Past actions are private information, and there is some exogenous probability that the relationship between any employer and employee will terminate at the end of each period, even if the employee was honest. Hence in each period some merchants randomly hire agents from the pool of unemployed agents. Because there are more employees than employers, one can remain unemployed for some periods before being rehired.

In equilibrium with cooperation in which an employer fires an employee who cheated, each employer has to pay workers a wage that is high enough that the gain from cheating and then joining the pool of unemployed agents is lower than the expected wage from being honest and continuing to receive the wage. Wages and the unemployment rate are thus endogenously adjusted to create the right incentives. In equilibrium some employees are involuntarily unemployed, in the sense that they are willing to work for less than the equilibrium wage but are nevertheless not hired (Shapiro and Stiglitz 1984). Organizations that distribute information about employees' past conduct, however, can alter employees' outside options (following cheating in a particular relationship) by reducing the probability that a worker who cheated in the past will be hired (Greif 1989, 1993).

If the environment is such that there are more employers than employees and if wage contracts are legally enforceable, an employer cannot punish an employee by firing or not paying him. Because wages are legally enforceable, as long as past conduct is private information, an unemployed employee will be hired. In such cases an equilibrium with cooperation requires a different manipulation of utility streams. One option is to pay employees bonuses rather than wages (MacLeod and Malcomson 1989).

Another option is to create an endogenously sunk cost by "building relationships" among the two players interacting in a prisoners' dilemma game. Various means, such as posting bonds

and exchanging gifts, can be used ex ante to increase the ex post cost of cheating implied by the need to establish new relationships.[8] While theory affirms this intuition, it also highlights the important role that incomplete information plays in making an investment in building relationships an equilibrium outcome.

To see why this is the case, suppose that the endogenous sunk cost of building relationships is achieved in the following way. Once two particular players are matched, they can choose whether to play a high-payoff or low-payoff prisoners' dilemma game in each period. In the high-payoff game, a player can lose more if cheated. The players can thus invest in their relationships by playing the low-payoff game for some period. After these periods of reduced utility to both players, they begin to cooperate to the fullest possible extent. If the players' strategies call for such investment whenever new relationships are formed, cheating entails having to invest in building a relationship with another player.

These intuitive strategies are not part of an equilibrium, because two newly matched agents have an incentive to forgo paying this bond, given that everyone else in the population requires it. After all, it is the need to pay the bond in the next new relationship following cheating that contributes to deterring cheating in the present relationship. But because this is true for everyone, no one has the incentive to post the bond. Hence there is no equilibrium with an endogenous cost of building relationships. This problem disappears, however, if there is a sufficiently high probability that an individual is a "bad" type, who will cheat in either game. If one's type is unobservable, this uncertainty motivates each player to first verify the other's type by playing the lower-payoff game (Kranton 1996; see also Ghosh and Ray 1996 and Watson 1999).

Organizations also play a part in endogenously altering payoffs. In the late medieval period, nonrefundable entry fees to merchant and other guilds, which had a monopoly over certain trade and crafts, arguably enabled intraguild cooperation that otherwise would not have been possible. Regulations for entry and exit play a similar role in modern economies. In modern economies, organizations manipulate the ownership of resources to enable them to commit to provide high-quality service. This is possible when this ownership fosters the ability of the

[8] Note, however, that posting a bond creates a one-sided prisoner's dilemma situation. Once an employee posts a bond, the employer can expropriate it and hire another agent. In many cases bonds are placed in the hands of a third party (such as an escrow company), whose actions are disciplined by either the legal system or reputational concerns.

organization's clients to punish it when necessary. Hotel chains, for example, purchase independent hotels, thereby increasing their clients' ability to punish them if they fail to provide good service. Following mediocre performance by one hotel in the chain, the client can refrain from using the other hotels in the chain (Ingram 1996).

More generally, manipulation of payoffs can be achieved by linking the central transaction—modeled as a prisoners' dilemma or a one-sided prisoner's dilemma game—with other transactions. Social exchange, norms, and violence often play a role in achieving this. Social, psychological, and physical harassment of a cheater can be a means to alter payoffs to deter cheating.[9]

The details of the underlying transaction have another important ramification for the manipulation of payoffs required for cooperation. The preceding discussion implicitly assumed that cheating in one period does not directly influence an individual's utility or possible actions in future periods. In particular, it was implicitly assumed that a cheater "consumes" the gains from doing so at the end of the period in which he cheats. But cheating often implies obtaining an investment good that can be used to change one's payoffs in subsequent periods. Among the Maghribis, for example, an agent who cheated gained capital, which he had the ability, knowledge, and opportunity to invest in future periods. Reputation-based institutions supporting cooperation in such situations therefore have to ensure that honesty is profitable, despite the higher gain from cheating. The Maghribis did so by having agents invest their own capital through other agents, who, in turn, were not expected to be punished for cheating an agent who had himself cheated in the past.

C2.3 Credibility

Understanding the effectiveness of a reputation-based institution requires understanding how the promise and threat of various actions are made credible. Unless the (implicit) promise to continue hiring an honest agent is credible, the best thing for the agent to do is cheat. Symmetrically, if the agent cannot commit to refrain from cheating and establishing new relationships, no merchant will hire him. Most of the generic theoretical insights about how the

[9] See Wiessner (2002) for the role of gossip among African bush women of low social rank in disciplining high-ranked men who deviate from the groups' norms.

credibility of continuing relationships is achieved were discussed earlier in connection with the endogenous manipulation of payoffs.

As we have seen in the case of the Maghribis, understanding this credibility is an integral part of the analysis. Among the Maghribis, merchants could have committed to continue hiring intragroup agents, because the collective punishment entailed that the wage premium required to keep an agent honest was lower within the group than outside. Arguably, agents could have committed to retain their affiliation with the group because of the higher expected income from agency relationships (due to the higher probability of being employed) and the capital premium.[10]

Game theory is very useful in identifying the conditions under which the threat of punishment following cheating is credible, because it highlights the distinction between the Nash equilibrium and the subgame perfect equilibrium. A subgame perfect equilibrium is a Nash equilibrium that satisfies the additional condition that it is a Nash equilibrium in every proper subgame. In particular, for threats and promises to be credible, behavior off-the-equilibrium-path has to constitute a Nash equilibrium (see Appendix A, section A.3).

A generic insight of game theory is that punishment is credible if the players' strategies entail a transition to an equilibrium in the stage (one-period) game in the case of punishment.[11] In the case of a prisoners' dilemma game, this is the (unique) equilibrium, in which both players cheat. The credibility of a promise to be honest can also be fostered by the nature of the goods exchanged. Indeed, in contemporary international trade, barter is commonly used for exactly this purpose (Marin and Schnitzer 1995).

C.2.4 Credibility and Multilateral (Third-Party) Punishment

Of particular interest and importance to institutional analysis is the credibility of punishments and rewards in reputation-based institutions in which punishments and rewards are provided by a third party, namely, an individual who is not a party to the central transaction the institution governs. Such reputation-based institutions are usually able to support more cooperation than bilateral relationships, as we have seen in the case of the Maghribi traders. Multilateral

[10] As noted in Chapter 3, this assertion cannot be empirically substantiated, but the theoretical possibility that this was the case increased confidence in the identification of the coalition.

[11] More generally, the punishment is credible and can deter cheating if it entails a transition to an equilibrium with a lower payoff for one who is to be punished.

punishment usually implies a harsher punishment than a bilateral one, enabling cooperation in a wider range of parameters.[12]

The problem of credibility of punishment is more severe in cases of multilateral punishment. Why would one punish an individual who had not hurt him? How is a threat of collective punishment made credible? Without denying the possible importance of such motivational factors as contempt, disgust, and desire to punish one who acted unfairly toward others, game theory draws attention to additional factors. In the case of incomplete information, one is motivated to participate in collective punishment because cheating reveals that an individual is a "bad" type. An employer would not hire a worker who had already revealed himself as a "bad" type, because he would expect the worker to cheat him as well. When collective punishment is based on incomplete information, individuals are motivated to acquire information about who has cheated in the past.

Complete-information models reveal other ways to motivate individuals to participate in collective punishments. In prisoners' dilemma games, individuals can be motivated to participate in punishing individuals who did not cheat them by the threat that failing to do so will invoke punishment from others. The equilibrium strategy is not to cooperate with a player who has either cheated in the past or has failed to punish someone who cheated in the past. This "second-order punishment" has to be supported by yet higher punishment orders for cheating someone who failed to punish someone who failed to punish and so forth.

Second-order punishment is not effective in one-sided prisoner's dilemma games, which—unlike the prisoners' dilemma game—have an asymmetric structure. In a one-sided prisoner's dilemma game, there are two types of players (e.g., merchants and agents), and matching is always between individuals of different types. Hence in the merchant-agent game, a merchant always plays with an agent. A merchant therefore cannot directly punish another merchant by refusing to cooperate with him.

Multilateral punishment in such situations can be achieved in two other main ways. The first is by not punishing an agent who cheated a merchant who failed to punish an agent. The second is by linking the basic transaction, which we capture in the one-sided prisoner's dilemma game with another transaction. The merchant guild provides a historical example of this strategy

[12] For an exception, see Bendor and Mookherjee (1990). When a player is simultaneously involved in many identical bilateral games, if all games are identical, multilateral punishment cannot support cooperation if it cannot be supported in each of the separate games based on bilateral punishment.

and linkage. A merchant who did not participate in punishing someone who did not respect the merchant's property rights abroad was excluded from using the guild's ships for transporting his goods; another merchant who carried the excluded goods in these ships as if they were his own was subject to fine. Theory thus reveals the relationships between the features of the underlying central transaction and the feasibility and nature of a reputation-based institution based on collective punishment.

Other strategies can also be used to make the threat of collective punishment credible. A difficulty in inducing collective punishment in prisoners' dilemma games (without relying on second-order punishment) is that punishment based on reverting to the stage-game equilibrium in which both parties cheat is costly to the one who inflicts the punishment. One way to mitigate this problem is through a strategy in which an individual participates in his own punishment (Kandori 1992; Ellison 1994). In such a strategy, an individual who cheated in the past is supposed to cooperate with the one who punishes him by cheating. Hence the one who punishes is motivated to do so because it is profitable. Punishing entails receiving the payoff associated with cheating while the other cooperates. But why would a cheater cooperate in his punishment rather than continue to cheat? Motivation can be provided by making the punishment phase finite in length. After participating in his own punishment for a while, a cheater is "forgiven," and the players' strategies call for cooperating with him as if he had never cheated. He is induced to participate in his own punishment by the expected gains from future forgiveness. Others are motivated to participate in punishing him because they directly benefit from doing so, as they cheat while he cooperates in the punishment phase.

These analytical results were in games without transferable utilities—that is, in situations in which the distribution of the gains from cooperation (within a stage game) cannot be determined by the interacting individuals. These games assume that matching is random—individuals cannot choose whom they interact with and can thus not decide whether they want to be matched with someone who had previously cheated.

Greif (1989, 1993) considers a one-sided prisoner's dilemma game in which utilities are transferable and individuals have some control over whom they interact with. In addition, the analysis incorporated the assumption that the relationship between a particular merchant and agent can end exogenously even if the agent was honest. In this case, as we have seen in Chapter 3, there is yet another way to support collective punishment. In equilibrium, the wage required to

keep an honest agent is lower under the threat of collective punishment than under bilateral punishment. This is the case because the worst punishment that can be inflicted on any agent is the same: total exclusion from future interactions. But one who has been honest in the past has more to gain from future interactions. Once his relationship with the current merchant ends, he will be hired by another merchant with a positive probability, earning the equilibrium wage. Because the equilibrium wage is higher than an agent's income if he is unemployed, an agent who has never cheated in the past has more to lose from cheating. But if the wage that has to be paid to an agent who cheated someone else in the past is higher than that paid to an agent who did not cheat, every merchant has an incentive to hire an agent who has been honest in the past.

C.2.5 Renegotiation

The discussion of the credibility of punishments ignores another important theoretical insight into the nature of reputation-based institutions—renegotiation by the interacting individuals. It might intuitively be assumed that renegotiation, in which the players decide on how the game will be played after a given history, would improve welfare. In fact, theory indicates that it can undermine it. To see why this is the case, consider a prisoners' dilemma game and recall that to induce cooperation, punishment from cheating requires a transition to an equilibrium in the stage game in which the total payoff is lower than when the players cooperate. When renegotiation during this punishment phase is possible, both parties have a strong incentive to let bygones be bygones and resume cooperation. But if this is known ex ante, it decreases the punishment from cheating, implying that the original cooperative equilibrium cannot be sustained. If cooperation will be resumed after cheating, why not cheat?

Theory suggests that attention should be given to why the possibility of renegotiation did not undermine cooperation to begin with. The historical analyses illustrate two basic reasons why this can be the case. Among the Maghribis, renegotiation was not an issue for two interrelated reasons. First, because the "market" for agents was thick—many agents were active in each trade center and they were substitutes for each other—a merchant could switch agents at little cost. Second, a merchant had to pay a strictly higher wage to an agent who had cheated in the past than to an agent who had never cheated, because every merchant's strategy specified that no one would hire an agent who had cheated in the past and because agency relationships between a particular merchant and agent could have been terminated for exogenous reasons. This was the

case because an agent who did not expect to be hired by others would not expect to lose future gains from serving them as an agent in the future. Because the punishment is lower, a higher wage premium had to be paid to keep an agent honest.

The merchant guild reflects another response to the problem of renegotiation. In this case, the problem of renegotiation expressed itself as a free-rider problem, in which some merchants would trade with a ruler during an embargo. The maximum punishment that could be inflicted upon the ruler following an abuse of rights was switching to the one-stage-game equilibrium of no trade and abuse of rights if a merchant traded. But this equilibrium yields lower payoffs to both the ruler and the merchants than an equilibrium in which some merchants do trade while their property rights are secured because of the low level of trade during an embargo. This low level of trade implies that the ruler's gain from taxing merchants is sufficiently high to motivate him to respect their property rights under the threat that they would not return to trade if their rights were abused. Switching to this equilibrium, however, undermines the severity of the punishment that can be inflicted on a ruler following an abuse of rights in the optimal level of trade. The response to this problem was an organizational change that linked the ruler-merchant transaction with one among the merchants themselves. The organization of the merchant guild used coercive power to punish a merchant who traded during an embargo.

C.2.6 Endogenous Information

Theory also highlights the details of the information required for a reputation mechanism to function. Multilateral punishment depends critically on the ability of those who are supposed to punish to identify the one who is to be punished. Theory indicates that sufficient information for collective punishment can be contained in a "label" indicating whether one's status is that of one who has to be punished or not (Kandori 1992). In addition, a cheated agent must be motivated to make the cheating known and those who punish must be motivated to acquire this information, even though both actions are likely to be costly. The endogenous generation and transmission of such information and motivation is an integral part of how an institution functions.

Such information may be readily available to the interacting individuals if interactions are confined to a relatively small group, particularly if these individuals also interact socially. Throughout most of history interactions within such groups, intertransactional linkages within them, and the associated beliefs and norms provided information and provided motivation to

transmit, acquire, and act on it. But when such information is based on personal familiarity, as existed among the Maghribis, for example, cooperation is limited by the extent to and speed at which the social network can transmit information.[13]

More generally, the manner in which such information is circulated and motivation is provided influences the extent (in terms of the number of interacting individuals and the amount one is willing to entrust to the other) to which the threat of collective punishment is credible. One of the main institutional transitions in the modern, economically developed world has been the introduction of institutional elements that enabled more impersonal exchange to prevail among more individuals. The regulations of personal identities by the state, identification cards, passports, credit bureaus, and credit cards are among the institutional innovations that enabled individuals to identify themselves credibly to strangers and provide information regarding their past conduct.

For a multilateral reputation mechanism to function, individuals have to be induced to transmit information. Why would an individual who has been cheated in the past inform others that someone had cheated him? Knowing that no one would cheat on the equilibrium path, why would anyone invest in gaining access to an information network or gathering current information?

The motivation to inform others that an individual had cheated depends critically on the relationships among the players who are supposed to punish a cheater. Competition among those who are supposed to punish reduces the motivation to provide such information. The Maghribis were not in competition with one another. Because they sold their goods in competitive markets, one merchant's loss was not another's gain. Because informing others that a particular agent cheated did not lower the payoff of the merchant who informed, a merchant had nothing to lose from informing on a cheater. The thick information networks and constant business communication among the traders made the cost of supplying this information negligible. This would not be the case among producers or merchants competing with one another in a "thin" market in which a reduction in the economic activity of one is another's gain.

[13] Reputation-based institutions face a trade-off between the benefits of a larger network, which enables more benefit from cooperation, and the delay and cost of information transmission that this larger size entails. Technically, we can capture the additional information cost of the larger size by making the time discount factor a decreasing function of size: the larger the group, the more time it takes for the information about cheating to be diffused.

Similarly, for collective punishment to be credible, individuals have to be motivated to acquire the necessary information. If people do not know whom to punish, the threat of punishment is not credible. Motivating individuals to acquire information is trivial when the situation is one of incomplete information and they are motivated to acquire information about a new partner's past conduct. Motivating individuals to acquire information is more problematic in situations in which cheating is not supposed to occur on the equilibrium path or the probability of its occurring is so low that investing in information is not worthwhile.

These considerations highlight the importance of what can be called a secondary information network: an information network – namely, to which one is motivated to acquire access, irrespective of considerations about cheating. Among the Maghribis, traders were motivated to retain an information network because it was valuable to gather commerce-related information in general. Geographical proximity and constant interactions in social or religious activities are among the other reasons why an independent network may exist. Both factors are present in the case of the Jewish diamond traders of New York (Bernstein 1992).

Organizations specializing in soliciting and distributing information can also provide individuals with the incentive to acquire the information required for multilateral punishment. The article by Milgrom et al. (1990) discussed in Chapter 10 analyzes the role of such organizations. The authors consider an infinitely repeated game in which two players are matched only once to play a prisoners' dilemma game and the players do not share the social network required to make past actions known to all. They then enrich the game by introducing an organization capable of verifying past actions and keeping records of those who cheated in the past. Acquiring information and appealing to the organization is costly for each player. Despite these costs, there exists a (symmetric sequential) equilibrium in which cheating does not occur and players are induced to provide the court with the information required to support cooperation. The court's ability to activate a multilateral reputation mechanism by controlling information provides the appropriate incentives. Hence an organization can ensure contract enforcement over time even if it cannot use coercive power against cheaters by supplementing the operation of a reputation mechanism.[14]

Not all situations require information flows for the threat of multilateral punishment to be effective. Kandori (1992) and Ellison (1994) consider a situation in which players with infinite

[14] Today such organizations as credit bureaus and Verisign fulfill such functions (Greif 2000).

life-spans are randomly matched each period to play a prisoners' dilemma game. Bilateral punishment cannot sustain cooperation, and past cheating is private information. Nevertheless, cooperation may be possible based on a contagious equilibrium. The strategy in this equilibrium is for every player to cheat subsequently if he either cheated or was cheated in the past. Cheating thus leads to a total collapse of cooperation.

Equilibria constructed in this manner are not very reasonable, because any unintentional or perceived cheating or cheating by one "bad apple" leads to a transition to a punishment phase.[15] Furthermore, such equilibria do not exist in one-sided prisoner's dilemma games. For the fear of punishment to prevent cheating, a player's utility during the punishment phase has to be lower than it would have been had cooperation taken place during this phase. So why would an individual start cheating after having been cheated? In the prisoners' dilemma game, a player cheats after having cheated or having been cheated because he expects the other player to continue cheating as well; if this is the case, the best he can do is to cheat. In one-sided prisoner's dilemma games, however, only one individual can cheat.[16] Thus no individual can be motivated to continue cheating by the expectation that the other player will do so as well.

C.2.7 Imperfect Monitoring

The discussion so far has assumed perfect monitoring in which, in particular, one knows ex post with certainty the actions of the person one played against. Those who are supposed to punish a cheater can verify if cheating indeed occurred. Reality, however, is often characterized by imperfect monitoring.

Imperfect monitoring is a situation in which actions are not directly observed (see Appendix A, section A.3). One can deduct others' actions from a signal that is not perfectly correlated with these actions. If one player took a particular action, the signal indicates that it was taken with a higher probability than if it was not taken, but because the signal is only probabilistic, it can still indicate that this action was not taken. Players can thus receive a false impression about others' past behavior.[17]

---

[15] It is possible to get out of this state if everyone switches to cooperating again at some future time. This requires coordination among players who lack the ability to communicate, however.

[16] The assumption is that the one who was cheated drops out of the game.

[17] The classical work on imperfect monitoring games is Green and Porter (1984). See also Abreu et al. (1986, 1991) and Fudenberg, Levine, and Maskin (1994). For recent surveys, see Pearce (1995) and Kandori (2002) and the articles by Bhaskar, van Damme, Piccione and Ely, Valimaki, Compte, Mailath,

The basic insights of games with perfect monitoring are relevant to games with imperfect monitoring, with one important addition. On the path of an equilibrium with cooperation, although no one actually cheats, (finite) periods of punishment nevertheless occur when cheating is signaled. The intuition is that if one's strategy does not specify punishment after observing cheating, then the best response of other players is to cheat, implying that cooperation cannot be sustained. To support cooperation, after observing a signal that cheating has occurred, each player has to punish the specified player, even if it is known that he did not cheat.

C.2.8 Endogenous Intertransactional Linkages and Organizations

The preceding discussion focused on a particular intertransactional link: that among the same central transaction in different time periods. In reputation-based institutions, the interacting individuals can link other transactions, thereby changing the set of beliefs in the central transaction under consideration. This is the case, for example, when one harasses or uses violence against someone who cheated him. Organizations also play an important role in facilitating the operation of reputation-based institutions by linking transactions. Organizations—either informal ones, such as social networks and communities, or formal ones, such as credit bureaus and guilds—change the set of self-enforcing beliefs in the central transaction in various ways. We have seen that organizations representing infinite-horizon players enable individuals to commit despite their finite life-spans. Organizations can also increase the frequency of interactions and internalize the cost of cheating inflicted by one player on others. In addition, they acquire, store, and distribute information; produce and propagate the meaning of various actions; provide a uniform interpretation of past actions; and coordinate behavior by providing public signals.

Organizations can also reduce the expected cost of imposing and participating in a punishment. They can be an appropriately motivated third party required to verify past actions, to arbitrate, and to enable the players to compensate one another during disputes in a Pareto-improving manner (by avoiding costly punishment). Indeed, within an institution organizations can be relevant for the endogenous construction of future rewards and payoffs, enhancing the credibility of maintaining relationships and threats of future punishment, preventing

---

Morris, and Aoyagi in the January 2002 issue of the *Journal of Economic Theory*. For applications for institutional analysis, see Clay (1997) and Maurer and Sharma (2002).

renegotiation following cheating, generating and distributing information, and improving monitoring.

An important class of organizations not mentioned so far comprises those which serve as intermediaries with a greater ability to commit. In modern economies, credit card companies, escrow accounts, cash against document contracts, and cashier's checks are among the organizations and instruments used for this purpose. The implied enhanced ability to commit is endogenously achieved because the organization both increases the frequency of interactions and creates an infinite-horizon player. Instead of transacting with other players, each player involved in the original transaction interacts with the organization.

Consider the operation of a credit card company. The exchange between a seller and a buyer is replaced by an exchange between the seller and the credit card company and between the credit card company and the buyer. The credibility of the payment from the credit card company to the seller is based on the public institutions that enable it to commit. The credibility of the payment from the buyer to the credit card company is based partly on the company's ability to taint the buyer's credit rating.

Organizations, however, are made up of individuals. Understanding their behavior and implications therefore requires considering the motivation and ability of these individuals to take various actions (see Chapter 5). An important generic theoretical insight is that in reputation-based institutions, an organization's motivation to act in a manner that fosters cooperation may reflect its concern with its own profitability and reputation. *Consumer Reports* commits to provide dependable information, because otherwise readers would not continue to buy it. Stock exchanges are motivated to monitor the accuracy of the information provided by the firms that trade in them, because otherwise people may be less willing to purchase stock.

C.2.9 The Costs of Reputation-Based Institutions

Reputation-based institutions are not free. Their operation often depends on costly organizations, and their capacities and operation rely on and create barriers to engaging in various activities.

The following examples illustrate such costs. In an institution based on the expectation of multilateral punishment, a player will be honest, fearing the response of all members of the group. The expected length of his relationship with any particular individual within that group is thus less important than under bilateral punishment. If there are efficiency gains from frequently

changing the people with whom one interacts, these changes will occur only within the group. In contrast, in an institution based on investment in the sunk cost of establishing bilateral relationships, once these costs are sunk one would refrain from establishing new relationships, even if they were more efficient and therefore generated a larger surplus to divide. Sunk costs create a wedge between efficient and profitable relationships. If a new seller arrives offering a potential buyer the same goods at a lower price, the buyer may nevertheless refrain from establishing a relationship with him, because doing so would require making another sunk investment in establishing a relationship.[18]

The discussion here, however, is not directly concerned with the costs of reputation-based institutions. Instead, the concern is with the ability to use the observable implications of such costs, as revealed by generic theoretical insights, to help identify an institution. Indeed, the distinct behavioral implications of the costs associated with each of these two institutions fosters the ability to identify them empirically.

C.3 Concluding Comments

The preceding discussion highlights the contributions of theoretical insights in facilitating the forming and substantiating of conjecture regarding the relevance of a particular institution. The basic game-theoretic insights that cooperation, for example, is possible if interactions are of an infinite duration and the players are sufficiently patient, is the institutional analysis's initial observation rather than its conclusion. It sets the stage for evaluating whether the conditions required for the operation of this mechanism are in place and in what form. In conducting an interactive analysis aimed at such an evaluation, there is a constant feedback from evidence to theory and from theory to evidence. We use theory to delineate various possibilities and the conditions conducive to the existence and functioning of a particular institution; we use evidence to direct the analysis toward particular issues and possibilities rather than others,

In using theory to consider various possibilities, it is imperative to be attentive to the possible importance of factors outside that theory. In the case of private-order, reputation-based institutions, there are often complementarities between them and public-order (and, more generally, coercion-based) institutions. Institutions based only on reputation are particularly

[18] Fafchamps (2004) reports such behavior in contemporary Africa. For analyses of the costs of reputation-based institutions, see Kranton (1996); Kali (1999); Dasgupta (2000); and Annen (2003).

important when actions cannot be verified by the court (as was the case among the Maghribis) or when the interacting individuals involved are also those who control the court (as was the case with the merchant guild). But even in such circumstances public-order institutions can nevertheless play an important role in the operation of private-order institutions. In the case of the merchant guild, for example, a ruler's ability to control the use of violence in his domain was crucial for the operation of a reputation-based institution between him and foreign merchants. The theory of such complementarities is not well developed, however. In attempting to identify an institution generating behavior in a particular central transaction, it is therefore important to keep in mind that its institutional elements may have both private-order, reputation-based and public-order, coercion-based components. In identifying reputation-based private-order institutions in particular, it is useful to consider their possible reliance on and interactions with public-order institutions.