Selective Revelation of Public Information and Self-Confirming Equilibrium

Zacharias Maniadis

First Draft: April 2006. This Draft: July 2009

Abstract

Why is public information revealed? We endogeneize public information by modeling aggregate information release, in games of large populations with anonymous matching, as a rational, utility-maximizing choice. We assume that there is a "planner", who knows and selectively reveals aggregate information, in order to maximize his objective function. we show that public information can affect individuals' behavior, even if it is revealed after a self-confirming equilibrium (SCE) has been reached. By selectively revealing information, the planner may upset a given SCE, in order to achieve a better outcome for him. Hence, some SCE are "unstable" to public information release. It is shown that only equilibria supported by heterogeneous beliefs can be information-unstable. Hence, unitary SCE, including Nash, are robust to public information release. Two policy applications of the concept of revelation instability are considered, illustrating how concealing information (self-censorship) and inducing experimentation with untested actions (affirmative action) can be socially beneficial. We present several real examples of endogenous public information revelation, where our theory could be useful.

1 Introduction

Social interactions among strangers can be modeled as games of large populations with anonymous matching.[1] The choices of an individual who is matched against an opponent are based on the player's expectations concerning the "average" behavior of the opponent's population. However, people rarely have enough interactions with members of other populations in order to form accurate expectations about the behavior of all other social groups. The notion of self-confirming equilibrium (SCE) of Fudenberg-Levine (FL)(1993*a*) describes a state where people optimize given their beliefs about other groups, but individuals' beliefs

[1] There is a large debate concerning the degree of sophistication of agents, since evolutionary models consider players "naive learners". See Mailath (1998), who offers support for this hypothesis of evolutionary theories against its criticisms. We believe that relatively weak assumptions about the sophistication of players are enough to justify our results. We shall further discuss the degree to which naivet needs to be invoked in our model.

need not be correct about groups they do not personally interact with.[2] Further, members of the same population may have different experience, hence different beliefs. Individuals do not necessarily share the knowledge that other members of their group have acquired by interact with other social groups.

However, governments and special interests often have access to aggregate data about the behavior of social groups. By revealing their special information, they may correct the wrong beliefs of some individuals regarding the behavior of other populations, and possibly change the formers' actions. Therefore, selective information revelation of aggregate data can become a powerful policy tool, that the possessors of information can use to manipulate behavior. This is especially relevant in modern societies, where the media easily convey public information. This information need not necessarily be exogenous, because the availability of aggregate data depends on the incentives of those who have them to disclose them. Thus, possessors of public information can choose what types of mistaken beliefs survive in a long-run equilibrium. Accordingly, a given self-confirming equilibrium is plausible as the long-run state of the economy only if the possessors of aggregate information cannot "choose" a more preferred equilibrium for them, in the sense we shall define bellow.

For a specific example, we ask the reader to look at Figure 1. Assume that there are two social groups, investors and officials. The investors move first, deciding whether to enter (invest) or not, and then officials choose whether to cooperate or not.[3] The investment is profitable only if the official cooperates. The numbers in the brackets show the fractions of the social groups making each action in the specific "state" of the dynamic system we are considering. 20% of the investors have taken the risk of investing before, and they have leaned the truth: that the officials are upright, and they always cooperate without asking for a bribe. However, 80% of investors choose to refrain from investing, holding strong prior beliefs that the officials are corrupt. This state of affairs, being a SCE, is stable in the sense of FL.

We claim that this equilibrium is implausible. Although 80% of investors are better off not investing given their priors, they would change their behavior if they knew the true behavior of officials. However, if the government possesses the data that reveal this behavior and wishes to maximize social surplus, it ought to reveal this information. Knowing the true data about corruption, it may announce the true behavior of officials through the media. Accordingly, the behavior of investors may change, if, by observing the true data, their beliefs change. Clearly, revealing the fact that officials are honest will induce all individual investors to enter, upsetting the equilibrium. The new profile, where all investors enter and all officials behave honestly, is also a steady state, because it is a SCE. Moreover, the government prefers this steady state than the previous one, so it has an incentive to reveal this information.

In our model, the basic theoretical tool employed is the notion of SCE.[4] The key idea is that if people do not experiment enough, aggregate play need not result in Nash equilibrium outcomes.[5] The model assumes a framework similar to that of FL (1993), and we simply add

[2]For example, people of one ethnic group may be brought up having strong prior beliefs that the members of another ethnic group hate them. Consequently, they avoid interacting with that group. If this belief is wrong, it cannot be falsified, and hence is never corrected.

[3]When officials do not cooperate, they illegally try to expropriate rents from the investor.

[4]See Fudenberg and Levine (1993*a*), Hahn (1997) and Kalai and Lehrer (1993).

[5]The main question that can be asked about SCE is: why would agents fail to experiment to learn the

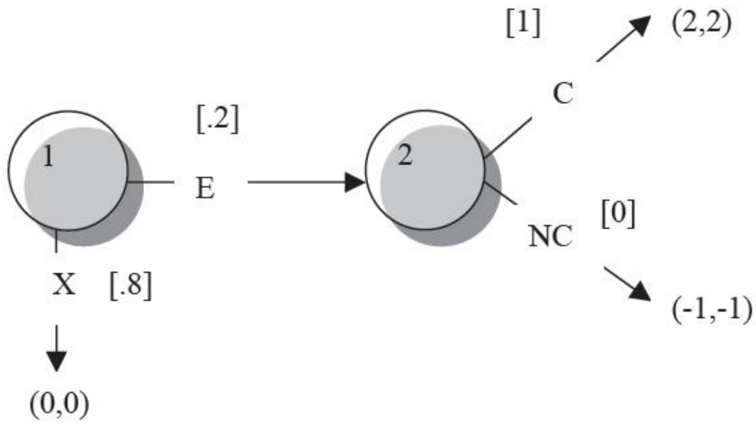

Figure 1: The modified cooperation game.

the existence of a "planner", an agent who knows and selectively reveals public information to maximize his objective function. Individuals do not know anything more about the behavior of other social groups than what personal experience teaches them, unless the planner reveals information, which is always perceived to be true. Our key insight is that, deciding whether a particular self-confirming equilibrium with non-Nash outcomes is a plausible rest point for the dynamic social interaction, one should look at the incentives of the possessors of public information.[6] This is because selective information release by the planner may upset a given self confirming equilibrium and lead the system to a different one. Equilibria that cannot be upset in such a manner are defined as revelation-robust. we show that all SCE supported by unitary beliefs (that is, where all individuals in a given population have the same beliefs, like in Nash equilibrium) are revelation-robust. Considering policy applications, we assume a "benevolent social planner" is the possessor of public information, and we examine optimal information release. We show that except from a special class of games, self-censorship (not revealing all public information that could be revealed) can be socially beneficial. Furthermore, it is shown that information-generating "affirmative action" may be useful. Finally, our framework has a wide variety of applications in Industrial Organization, Political Economy, Public Policy and other fields.

In the paper which is closer to our spirit, Esponda (2006) has a theoretical model that

true behavior of others? One reason might be that many decisions in life do not permit experimentation. For example, if some action of one social group results in the death of agents belonging to another group, these agents are not very likely to experiment with the social interaction. Moreover, in many decisions in real life, each player can only move once and for all. For example, a person decides only once whether to attend law school. Experimentation is not possible here without high cost, and the person's priors play the major role here. Finally, a non-Nash SCE can be reached if there is a very large number of possible actions and a finite life span. For example, no customer has a comprehensive knowledge of which products in a given supermarket satisfy her best, because the customer cannot try all of them. For this reason, selective information revelation is widely used by advertisers.

[6]We take the knowledge of aggregate statistics by the planner as given. Our setting can easily take into account the cost of aggregate information acquisition as well as multiple "planners"po with possibly conflicting interests.

focuses on a specific type of games, namely first price auctions. He asks whether the equilibrium feedback policy, which in most cases may be decided by the auctioneer, may affect equilibrium outcomes.[7] Here we generalize this approach to abstract extensive-form games. The literature on herding behavior and information cascades also raises the issue of aggregate information management. Bikhchandani, Hirshleifer and Welch (1992) argue that fads that are due to information cascades are sensitive to aggregate information revelation, because agents use very few of the available signals. Jackson and Kalai (1997) examine "recurring" games, where each player plays only once, but the same game is repeated with different players every time. Information revelation of aggregate play has substantial effects here, because each player learns something about herself when she gets information about the history of her group. Their conclusions regarding the benefits of "affirmative action" are similar to ours.

The experimental literature has also addressed the issue of whether reveealing aggregate information matters and whether expectations can be manipulated. Roth and Schoumaker (1983) and Harrison and McGabe (1996) directly manipulated subjects' expectations about others' play in an ultimatum game, with significant and lasting effects. Berg, Dickhaut and McCabe (1995) and Ortmann, Fitzgerald and Boeing (2000) performed experiments of one-round trust games,[8] and found some support for the notion that information revelation of aggregate data can push the economy to desirable equilibria.[9] Similar results were found in Maniadis (2007), Frey and Meier's field experiment (2003), Dufwenberg and Gneezy (2002) and Hargreaves-Heap and Varoufakis (2002).

The remainder of the paper is organized as follows. In part two we introduce the model, following Fudenberg-Levine (1993a) and define Nash and self-confirming equilibrium. In part three we introduce the planner, define the notion of revelation unstable self-confirming equilibria and provide examples illustrating the definitions. Part four we define strict instability relative to planner's preferences. In part five we characterize equilibria that cannot be improved upon with information revelation. Part six presents two particular policy applications. Further examples and a general discussion follows in part seven. Part eight concludes.

2 The Model

The model examines the steady states of a system with dynamic, anonymous interaction and learning. There is repeated, anonymous matching of individuals with other individuals, each of whom belongs to a different "population-role". In the model, we make a first step in endogneizing the information which players get as play evolves. My point of departure is Fudenberg and Levine's approach ($1993a, 1996$) in assuming that players see only the result

[7]He thus provides a very specific example of a "planner" and how he selectively reveals information about the aggregate data to maximize his objective value.

[8]Each "sender" had 10$ that he could send to the receiver. The amount sent tripled, and then the receiver decided how much money to send back to the sender.

[9]They played the game once with some students, and subsequently they showed the data about the actions chosen to different students that were about to play the game on a different date. They found that information revelation about the same game played by different subjects does affect behavior in ways that increase social surplus.

of play in their own matches.[10] Taking as given the main results of this research, especially the possibility of the game settling in a self-confirming equilibrium with non-Nash outcomes, we shall examine how the "planner" can convey the aggregate information he has, in the best possible way, in order to change the equilibrium outcome. The planner can be interpreted as the possessor of reliable data about the actions of the public. we assume that individual agents do not understand the existence of the planner, and that the planner can reveal information only once, after play has settled to a steady state.[11] Even in this very simple setting, we get interesting results. It is shown that some self-confirming equilibria are not plausible in the presence of the planner, because by selectively - but truthfully - revealing aggregate data, the planner can move the system to a different state. In the presence of the planner, only some forms of wrong beliefs survive in the long run.

2.1 The Extensive-Form Dynamic Game

There is a given extensive-form game with I players, with no chance moves. By J we denote the set of all players. The game is played repeatedly among anonymous agents randomly matched with each other.[12] Although each individual agent plays a pure strategy, each population as a whole randomizes across strategies, since individuals in the same population may be choosing different pure strategies. Each individual knows the extensive form of the game, the realized terminal nodes of their games after each match, and their payoffs at all terminal nodes, but not the payoffs of other individuals.[13]

The extensive-form game is as follows. There is a game tree X with finitely many nodes $x \in X$. The terminal nodes of the tree are $z \in Z \subset X$. Information sets, which are a partition of $X \backslash (Z \cup 0)$, are denoted by $h \in H$, and the subset of information sets where player i has the move by $H_i \subset H$. We denote the set of feasible actions for player i at information set h_i by $A(h_i)$, and all possible actions of player i by $A_i \equiv \bigcup_{h_i \in H_i} A(h_i)$. Denote the player who moves at node x by $\iota(x)$. The function ℓ assigns for each noninitial node x the last action taken to reach it.

A pure strategy for player i is a mapping $s_i : H_i \to A_i$ satisfying $s_i(h_i) \in A(h_i)$ for all

[10]An underlying assumption we use is that a fictitious play process describes the evolution of learning.

[11]Although this set of assumptions is restrictive, we can still get useful results. The assumption that agents do not know that the planner exists captures the fact that the media never admit that they are manipulating information, hence efficient planners tend to "hide" their existence. Thus, it can be argued that we are modeling a successful manipulator, rather than an unsuccessful one. The assumption that the planner only reveals information once, is necessary in order to simplify the model significantly. However, because of this assumption, our results should be viewed as suggesting a "lower bound" of what the planner can achieve with information revelation.

[12]Notice that when we shall refer to a an "individual" or an "agent", we will mean a particular person who belongs to some population. On the contrary, the word "player" will denote a whole player-role (corresponding to a population of individuals).

[13]This captures the fact that social interactions are anonymous. A individual does not know, and does not have strong preferences about, the payoff functions of other individuals that belong in any population (even her own population). For example, an individual official in our introductory example does not know if other officials share his preferences. He might believe that other officials are corrupt, hence that they get higher payoffs if they do not cooperate. Once more, we emphasize that we model agents who learn by observing others' behavior (or by credible revelation of the observations made by the planner), rather than by introspection.

$h_i \in H_i$. Let $S_i \equiv \times_{h_i \in H_i} A(h_i)$ be the set of all such strategies. A strategy profile specifies a pure strategy for all players, and we denote it by $s \in S = \times_{i \in J} S_i$. A mixed strategy for player i is a probability distribution over pure strategies, $\sigma_i \in \Delta(S_i)$, and a profile of mixed strategies is denoted by $\sigma \in \times_{i \in J} \Delta(S_i)$. The payoff for each player depends on the terminal node. So, for players $i = 1, 2, \ldots, I$, the payoff function is $u_i : Z \rightarrow \Re$. Let $z(s)$ denote the terminal node reached when profile s is played, and $p[z/\sigma]$ the probability that terminal node z is reached under profile of mixed strategies σ.

$H(s_i)[Z(s_i)]$ denotes the subset of all information sets [terminal nodes] reachable when agent i plays s_i. Similarly, $\overline{H}(\sigma)$ denotes the set of information sets that are reached with positive probability under σ, and $\overline{Z}(\sigma)$ denotes the set of all terminal nodes that are reached with positive probability under σ. Let $\pi_j(\widehat{h_j/\sigma_j})$ denote the distribution of actions at information set h_j induced by mixed strategy σ_j for player j. [14]

Absent information revelation by the planner, players do not know the true distribution of play, so there is strategic uncertainty. Each player has beliefs over the aggregate distribution of play. These beliefs are described by a probability measure μ_i on S_{-i}, the set of profiles of strategies of other players. we assume these beliefs are independent across i's opponents, hence $\mu_i(\times_{j \neq i} \overline{S}_j) = \prod_{j \neq i} \mu_i(\overline{S}_j)$, where each $\overline{S}_j$ is a measurable subset of S_j. Given player i's beliefs about other players' strategies, the probability that terminal node z is reached when player i chooses pure strategy s_i is $\mu_i(\{s'_{-i}\})$, where $z = z(s_i, s'_{-i})$. Accordingly, the expected utility of an agent with beliefs μ_i when she plays strategy s_i is $u_i(s_i, \mu_i) = \sum_{s_{-i} \in S_{-i}} [u_i(z(s_i, s_{-i}))\mu_i(s_{-i})]$.

In this environment, it is worthwhile to explicitly define Nash equilibrium it terms of players' beliefs about their opponents. A Nash Equilibrium is a profile of mixed strategies σ such that for all i and for all $s_k \in supp(\sigma_i)$ there exists beliefs μ_{i,s_k} such that:[15]

1. s_k maximizes $u_i(\cdot, \mu_{i,s_k})$
2. $\pi_j(\widehat{h_j/\mu_{i,s_k}}(j)) = \pi_j(\widehat{h_j/\sigma_j})$ for all $h_j \in H_{-i}$

Where $\mu_{i,s_k}(j)$ is the probability distribution on j's strategies that subgroup s_k expects according to beliefs μ_{i,s_k}. Thus, a Nash equilibrium is the profile consisting of the best responses of agents to their beliefs about the aggregate distribution of play, where these beliefs are correct for every information set of opponents. However, if players do not experiment enough, they may never get to know true play in all information nodes. They may end up in a situation where as far as they can tell, their actions are optimal, but without a necessarily correct assessment of play in information nodes that they do not reach given their strategies.

This is captured by the following equilibrium notion: a self-confirming equilibrium is a mixed strategy profile σ such that for all i and all $s_k \in supp(\sigma_i)$ there exists beliefs μ_{i,s_k} such that:

[14]we assume perfect recall, so by Kuhn's theorem, every mixed strategy profile generates a unique profile of equivalent behavior strategies.

[15]The notation μ_{i,s_k} emphasizes the fact that the beliefs held by each "subgroup" of population i could differ. Each "subgroup" is identified with a different pure strategy $s_k \in S_i$, played in equilibrium. The reason is that different pure strategies cause different information sets of opponents to be reached, hence each pure strategy may be associated with different experience. Note that for this particular definition, this notation does not make a difference, since all agents have the same (correct) beliefs. However, it matters in the definitions which follow, as we shall see.

1. s_k maximizes $u_i(\cdot, \mu_{i,s_k})$
2. $\pi_j(\widehat{h_j/\mu_{i,s_k}}(j)) = \pi_j(\widehat{h_j/\sigma_j})$ for all $j \neq i$ and $h_j \in \overline{H}(s_i, \sigma_{-i})$

This means that in a self-confirming equilibrium, a specific individual who belongs to population i and follows strategy s_k in equilibrium, must hold correct beliefs about the behavior of opponent groups only at nodes that are reached with positive probability given this agent' pure strategy and the mixed profile of i's opponent populations. Thus, an individual who belongs to population i may have wrong beliefs about the distribution of opponents' play, at an information set reached by other individuals who belong to i. This may happen if these individuals choose a different strategy than the given individual. In a SCE, only agents with the same "experience" in equilibrium are required to have the same beliefs.

3 Revelation-Unstable Self-Confirming Equilibria

we shall show that selective information revelation can "direct" the economy away from specific self-confirming equilibria. In this section we introduce the planner, who maximizes his payoffs $U^{PL}(\sigma)$ that depend on the long-run "state" σ. The planner, who knows the true distribution of actions at each information set, can announce it, fully or partially. His announcements are true and are always perceived as such.[16] Information can only be announced after a steady state has been reached, and only once. This assumption allows us to consider only static equilibrium notions, simplifying the analysis. Since the planner can always choose to reveal information according to this rule, we are deriving only a "lower bound" of what the planner can achieve. In addition, it is assumed that individual players do not know the existence of the planner, so they just observe information and update their beliefs. Also note that the planner has generic payoffs. For example, the auctioneer, who chooses the level of information feedback in an auction, wishes to maximize his revenue; the "benevolent government", who maximizes social welfare, etc. In our motivating examples $1, 2$ and for our main results we will focus on the "benevolent government" interpretation. The main idea here is that if the planner can achieve a better social result than a given self-confirming equilibrium with aggregate information revelation, then this equilibrium is implausible.[17]

[16] This can be though as a benchmark case for analysis. Our key insights would not change if we assume that a given fraction α of each subgroup believes the planner's announcements, and another fraction $1-\alpha$ ignores the announcements. Clearly, the quantitative results depend on the parameter α, but the qualitative ones carry over if we assume that only some people believe the planner, so that α is not zero. This assumption is more convincing in some real economies, such as advanced democracies, than others, such as totalitarian regimes. Note that by always selectively revealing true information, the planner can also develop a reputation for truth-telling.

[17] In all the analysis that follows, we will use the following tie-braking rule: each agent adheres, as much as possible, to what he was doing before public information was revealed. If, after the information revelation, an individual is indifferent between several strategies, she continues doing what she was doing before, provided that one of these strategies was played. If a whole population is indifferent between two optimal strategies both before and after information release, the same distribution of strategy remains after the information release. Finally, if, after the information release, an agent is indifferent between two optimal strategies he was not using before, he chooses randomly among them.

3.1 The Full Information Revelation Setting

We shall assume that for the equilibria we are discussing in this setting, $\overline{H}(\sigma) = H$. For full revelation, information about play in every information set should be available. Intuitively, if the planner wants to reveal the aggregate distribution of play at all information sets, then there must be data available for him to disclose. If, in a specific self-confirming equilibrium, an information set h_j is never reached, there is nothing to be announced about the behavior of player j's at this set. If this condition does not hold, then we can only have partial information revelation.

Definition 1a. A self-confirming equilibrium σ, with full support, is full revelation-unstable, if there exists a mixed profile σ^* such that:

1. For all i and for all $s^*_k \in supp(\sigma^*_i)$, s^*_k maximizes $u_i(\cdot, \mu^*_{i,s_k})$, where μ^*_{i,s_k} satisfies

$$\pi_j(\widehat{h_j/\mu^*_{i,s_k}}(j)) = \pi_j(\widehat{h_j/\sigma_j}) \tag{1}$$

   for all $h_j \in H_{-i}$

2. σ^* is a self-confirming equilibrium profile, which, for each individual, is supported by beliefs μ^* for each information set he does not reach given σ^*.

3. $u_i(s^*_i, \mu^*_{i,s_k}) > u_i(s_i, \mu^*_{i,s_k})$ for some i, some $s_k \in supp(\sigma_i)$, and some $s^*_k \in supp(\sigma^*_i)$.

This definition says that a self-confirming equilibrium is "full revelation-unstable" if an announcement of the true distribution of actions, at all information sets of the game, leads to a different self-confirming equilibrium.[18] Since the planner's information revelation is always truthful, agents' beliefs μ^* after the planner's full information revelation assign probability 1 to the revealed distribution, induced by the "old" mixed profile σ. The best responses to these beliefs generate a "new"profile σ^*. The key part in this definition is that the best-responses to the old distribution of play are also best-responses for the distribution which occurs after the information revelation takes place, that is, $\forall i, \sigma^*_i$ is a best response to σ^*_{-i}. Hence, the change in the state of the dynamic system following an information announcement is sustainable. Condition 3 ensures that at least one player has a strict incentive to change her behavior.[19]

In applications, planner's payoffs are fixed, hence the definition of "full information revelation instability" should take the specific incentives of the planner into account. Hence, we need the following definition:

Definition 1b. A SCE σ is full revelation-unstable relative to the planner's preferences U^{PL}, if there exists a mixed profile σ^* such that conditions $1-3$ of definition $1a$ are satisfied, and, in addition, $U^{PL}(\sigma^*) > U^{PL}(\sigma)$.

[18] Note that this definition is restricted for self-confirming equilibria with full support. It does not simply mean that all available information is revealed.

[19] This condition ensures that the definition captures appropriately the main intuition, that an equilibrium is unstable if the planner can change people's behavior in a predictable manner. In the absence of this condition, even some mixed Nash equilibria would qualify as unstable, but it is not clear why aggregate information would change behavior in this case.

Example 1. We shall illustrate definition 1*b*, showing how a self-confirming equilibrium can be undone by information revelation that leads to a better outcome for the planner. Consider the social interaction between investors and officials presented in the introduction (Figure 1). We will analyze more strictly the arguments here. Note that the game is similar to a trust game, but here the subgame perfect equilibrium outcome (E, C) is good for society. If an individual player 1 believes that agents in the population of player 2 tend to cooperate, the best-response is "enter", whereas if he thinks individual 2's do not cooperate, he should refrain from entering. We assume that there is a benevolent government, the objective of which is to maximize social welfare, which depends on the terminal nodes of the game, and the frequency with which each terminal node is reached. Accordingly, $U^{PL}(\sigma) = \sum_{z \in Z}\{p[z/\sigma] \sum_{i \in J} u_i(z)\}$ is the planner's objective function, as a function of the "state", the mixed strategy profile σ.

Assume that the state of the economy is described by the specific profile of mixed strategies σ, which is illustrated in Figure 1, where only one fifth of population 1 enters, believing that player 2's cooperate with probability one, and four fifths exit, believing that 2's cooperate with probability zero. In fact, player 2's always cooperate. So, the initial self-confirming equilibrium is $\sigma = \{(0.8X, 0.2E); C\}$.[20] Assume that the planner announces the true aggregate distribution of actions in all decision nodes. If individual player 1's simply best-respond to their beliefs about individual 2's play, and they regard the information revelation as truthful, then they all enter after the announcement since they expect that 2's will cooperate. The behavior of individual 2's does not change following the announcement, since their best response to "enter" is unique.

The new state of the game, profile $\sigma = \{E; C\}$, is compelling as a steady state, despite the fact that agents best-respond to correct beliefs about play in the *previous* period, which assigns probability one to that period's distribution of play. The reason is that σ^* is a Self-Confirming Equilibrium (actually also a Nash Equilibrium), so players best-respond to the current distribution of play as well. The planner prefers σ^* to the old profile, because more profitable transactions take place, and thus has the incentive to fully reveal the aggregate information. Hence, σ is full-revelation unstable.

It should be emphasized that full information revelation need not, in general, lead to a Nash equilibrium, not even a self-confirming equilibrium. This is because agents need not have correct beliefs about the distribution of actions that follows the information release.[21] One may also wonder if, for a given σ with $\overline{H}(\sigma) = H$, full information revelation could lead to a self-confirming equilibrium that is not Nash. The following example shows that this is possible, provided that some player i is indifferent between two different strategies s_k, s_l given the state σ.

Example 2. Consider the four-player game of Figure 2. Players $1, 2$ and 3 may play "right" or "down", and player 4 may play "up" or "down". Let the actions "down" and "right" be denoted (D_i, R_i) for player $i = 1, 2, 3$, and (l, r) denote the actions of player 4.

[20] Note that this is just one of infinitely many self-confirming equilibria in this game. Any mixed strategy of population "one" coupled with fraction 1 of population "two" playing C is a self-confirming equilibrium. We chose this specific fraction for illustrative reasons.

[21] Example 3, the "beneficial superstition game", illustrates a "state" where full information revelation does not lead to any equilibrium.

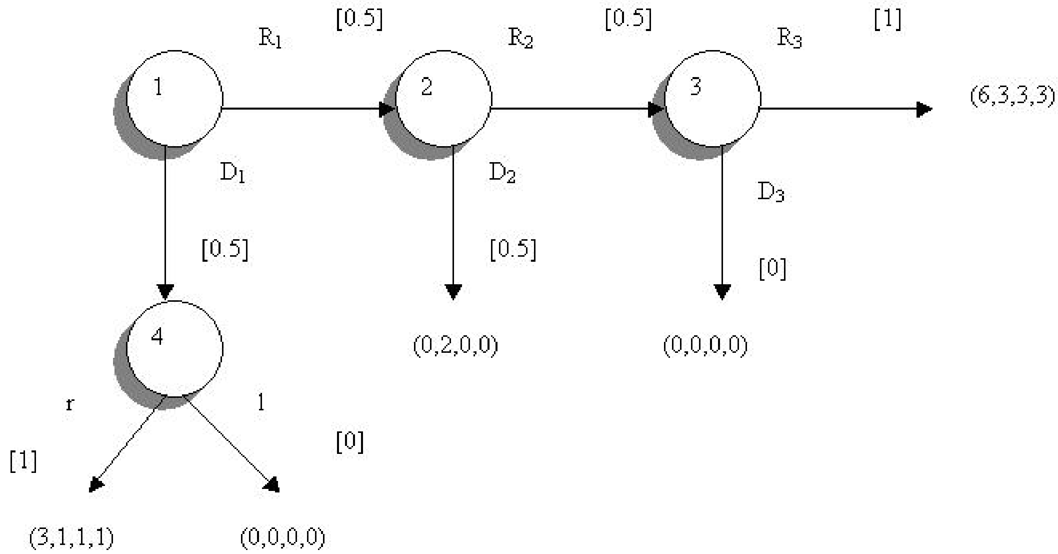

Figure 2: A SCE with an "indifferent" player.

The profile $\sigma = [(0.5R_1, 0.5D_1); (0.5R_2, 0.5D_2); (R_3); (l)] \equiv [(0.5, 0.5); (0.5, 0.5); (1, 0); (1, 0)]$ is a SCE profile, where agent 1's who play D_1 have correct beliefs about the behavior of 2's and 3's,[22] and agent 1's who play R_1 have correct beliefs about the behavior of 4's. Moreover, agent 2's who play D_2 believe that the fraction of D_3 is 0.8. Since no individual 1's have an incentive to change their behavior, full information revelation in this case will lead to the new profile $\sigma = [(0.5, 0.5); (0.5, 0.5); (1, 0); (1, 0)]$. This is a new SCE, but not a Nash equilibrium, since individual 1's who play "down" would not choose this action if they had correct beliefs. It will be shown that when no "indifference" for some player i exists, full information revelation leads to a SCE only if this equilibrium is equivalent to a Nash equilibrium.

Definition 2. Fix a self-confirming equilibrium σ such that $\overline{H}(\sigma) = H$. For each i, define B_i as the set of best responses of player i to σ_{-i}. Let b denote a profile of best responses. We say that σ has a "unitary full revelation outcome" if $z(b)$ is the same terminal node for every $b \in \times_{i \in J} B_i$.

Thus, given a σ with a unitary full revelation outcome, if full information is released, then a unique terminal node is reached when all agents play their best response to the revealed information.

Proposition 1. Let σ be a self-confirming equilibrium with a unitary full revelation outcome. Then, if there is a σ^* that satisfies conditions $1-3$ of definition $1a$, σ^* is equivalent to a Nash equilibrium (it has the same distribution over outcomes).

Proof/ Since σ has a unitary full revelation outcome, any best response profile σ^* to the beliefs specified in condition 1 generates a unique path to some terminal node z'. Hence,

[22]Of course these are not the only beliefs that rationalize the action D_1, but we want to emphasize the fact that both strategies of player 1 are best responses to the true beliefs.

condition 2 implies that, for all i, each agent in population i has correct beliefs about opponents' play on the equilibrium path $\overline{H}(\sigma^*)$.[23] Condition 2 additionally implies that, for all information sets h_j $(j \neq i)$, outside the equilibrium path, each individual i expects the "old" distribution of actions induced by σ. Hence, player i's beliefs μ_i that support the self-confirming equilibrium σ^* must satisfy:

$$\pi_j(\widehat{h_j/\mu_i}(j)) = \pi_j(\widehat{h_j/\sigma_j^*}), \forall h_j \in H_{-i} \cap \overline{H}(\sigma^*) \tag{2}$$

and

$$\pi_j(\widehat{h_j/\mu_i}(j)) = \pi_j(\widehat{h_j/\sigma_j}), \forall h_j \in \{H_{-i} - \overline{H}(\sigma^*)\} \tag{3}$$

Consider any mixed strategy profile σ^N that generates the distribution of actions in each node specified by the above beliefs, for all players. we claim that profile σ^N is a Nash equilibrium profile. By definition, σ_i^* is a best response to σ_{-i}^N. Notice that, for each player i, the distributions of actions generated by mixed strategies σ_i^* and σ_i^N differ only in information sets that are reached with probability zero given $(\sigma_i^*, \sigma_{-i}^N)$. Hence, $u_i(\sigma_i^*, \sigma_{-i}^N) = u_i(\sigma_i^N, \sigma_{-i}^N)$. Thus, σ_i^N is a best response to σ_{-i}^N, and we showed that σ^N is a Nash equilibrium. Clearly, σ^N reaches node z' with probability one. *QED*

It should also be noted that, although the property of having a "unitary full revelation outcome" seems rather specific, there is an interesting class of self-confirming equilibria that satisfies it, in the special case of full support ($\overline{H}(\sigma) = H$). This class consists of self-confirming equilibria that cannot be supported by unitary beliefs for any i. Hence, the hypotheses of Theorem 1 hold in the interesting case where there are no "indifferences" after the full information release.

Definition 3. A self-confirming equilibrium σ is called "strictly heterogeneous" if, for all players i, different strategies in the support of σ_i must be supported by different beliefs.[24]

Corollary. Let σ be a strictly heterogeneous self-confirming equilibrium such that $\overline{H}(\sigma) = H$. Then, if there is a σ^* that satisfies conditions $1 - 3$ of definition $1a$, σ^* is equivalent to a Nash equilibrium.

Proof/ If $\overline{H}(\sigma) = H$, "strict heterogeneity" implies that that, for each i, there is a unique best response to σ_{-i}. If this were not true, then, for some i, some strategies s_k and s_l in S_i would be best responses to σ_{-i}, hence both s_k, s_l would be supported by the correct beliefs about σ_{-i}. Thus σ has a unitary full revelation outcome. *QED*

3.2 Partial Information Revelation

Here we assume that not all information sets need to be reached, so it is possible that $\overline{H}(\sigma) \neq H$. Moreover, we assume that the planner may announce only partial information. For simplicity, we restrict our analysis to independent beliefs across opponents: each

[23]This is true because each information set on the equilibrium path is reached by all individual agents of all populations.

[24]Thus, two different strategies s_k and s_l in the support of σ_i cannot be supported by the same beliefs μ_i, for any μ_i that is correct in $\overline{H}(s_k, \sigma_{-i}) \cup \overline{H}(s_l, \sigma_{-i})$.

player has a distribution $\mu_i(j)$ over each opponent player j's pure strategies, and information about one opponent population does not affect expectations about the behavior of other player-populations. Of course, the planner may only reveal information about behavior at information sets reached with positive probability under σ, otherwise there is nothing to announce. Consequently, the planner may reveal the distribution of play at a subset of the family of all information sets reached with positive probability under σ. Hence, if we denote by H^A any set of information sets, for which the planner reveals the distribution of moves given σ, the following must hold:

$$H^A \subseteq \overline{H}(\sigma) \tag{4}$$

We also require that the planner reveals information for a subset of players, thus he is constrained to reveal information for all or none of the information sets of each player. For concreteness, denote by J_{H^A} the subset of J associated with the specific information revelation set H^A:

$$H^A = \bigcup_{j \in J_{H^A} \subset J} H_j \tag{5}$$

Definition 4. A set H^A which satisfies (4), (5) given a profile σ, is called an "information revelation set on σ".

Now, fix a SCE σ supported by beliefs μ. Since the information revelation of the planner is truthful, following the announcement of the planner, the beliefs of all players must be consistent with the distributions he announces.

Definition 5. We say that an information revelation set H^A on a SCE profile σ, supported by beliefs μ, generates "transition beliefs" μ^* if for all i and for all $s_k \in supp(\sigma_i)$ the beliefs μ^*_{i,s_k} satisfy:

$$\pi_j(\widehat{h_j/\mu^*_{i,s_k(j)}}) = \pi_j(\widehat{h_j/\sigma_j}), \forall h_j \in H^A \tag{6}$$

$$\mu^*_{i,s_k}(j) = \mu_{i,s_k}(j), \forall j \in \{J - J_{H^A}\} \tag{7}$$

Since agents do not know the payoff functions of others, they do not understand the strategic behavior of the planner, nor do they evaluate changes in others' behavior following the announcement. They simply believe the information announcement and adjust their play accordingly, believing everything else is the same. This idea is captured by "transition beliefs".

Definition 6. Let σ be a SCE supported by beliefs μ. Fix an information revelation set H^A on σ and the associated transition beliefs for each population and each subgroup, μ^*_{i,s_k}, as above. We say that σ^* is a profile supported by H^A if, for all i, there is a mapping g_i from strategies in the support of σ_i to strategies in S_i, satisfying $g_i(s_k) \in argmax(u_i(., \mu^*_{i,s_k}))$, such that:

$$\forall i, \forall s^*_k \in supp(\sigma^*_i), \sigma^*_i(s^*_k) = \sum_{\{s_k \in supp(\sigma_i): g_i(s_k) = s^*_k\}} \sigma_i(s_k) \tag{8}$$

The beliefs μ^*_{i,s_k} are the transition beliefs generated by H^A.

In other words, an information revelation set supports a profile σ^* if the transition beliefs it generates support σ^*. Since each subgroup $s_k \in supp(\sigma_i)$ could have different transition beliefs generated by H^A, not all of these groups need to have the same optimal strategy given these beliefs. Thus, the probability of each strategy s_k^* in the support of σ_i^* is determined by the sum of all probabilities of the subgroups corresponding to each $s_k \in supp(\sigma_i)$ that find s_k^* optimal given transition beliefs. The function g simply selects one optimal strategy for each population subgroup, given its beliefs. Hence it ensures that the mass of some subgroup is not counted twice.[25] Note that a given σ^* may be supported by multiple transition beliefs, but a specific information revelation set H^A generates unique transition beliefs. Now we may introduce our second main definition:

Definition 7a. A self-confirming equilibrium σ, supported by beliefs μ, is partial revelation-unstable if there exists an information revelation set H^A on σ, and a mixed profile σ^* such that the following hold:

1. σ^* is a profile supported by H^A.
2. σ^* is a self-confirming equilibrium, which $\forall i, \forall s_k^* \in supp(\sigma_i^*)$ is supported by the transition beliefs μ_{i,s_k}^* for all $h_j \in \{H_{-i} - \overline{H}(s_k^*, \sigma_{-i}^*)\}$.
3. $u_i(s_k^*, \mu_{i,s_k}^*) > u_i(s_k, \mu_{i,s_k}^*)$ for some i, some $s_k \in supp(\sigma_i)$, and some $s_k^* \in supp(\sigma_i^*)$.

This means that if all agents simply update their beliefs assigning probability 1 to the planner's announcements, and they keep their old beliefs in the opponents' information sets for which there is no revelation, then their best responses to the new beliefs form a self-confirming equilibrium profile. Again, this self-confirming equilibrium is compelling as the new steady state of the system, because if this profile is played, agents update information only in the information sets in $\overline{H}(s_k^*, \sigma_{-i}^*)$, hence they want to continue their chosen actions since this profile is a self-confirming equilibrium. In information sets outside $\overline{H}(s_k^*, \sigma_{-i}^*)$, agents maintain their old beliefs, and they do not have reason to update them in the absence of active learning.

As before, to complete the information revelation story, revelation instability must be defined in terms of the interests of the planner.

Definition 7b. A self-confirming equilibrium σ, supported by beliefs μ, is partial revelation-unstable relative to the planner's preferences if there exists an information revelation set H^A on σ, and a mixed profile σ^* such that conditions $1-3$ of definition $7a$ hold, and in addition $U^{PL}(\sigma^*) > U^{PL}(\sigma)$.

It is important to emphasize that this definition of information revelation allows much discretion for the planner. Let $H_C^A \equiv \bigcup H^A$, where the union is taken over all information revelation sets of σ, be called the "complete revelation set" of σ. The planner may opt to choose no revelation, complete revelation, or incomplete revelation, by releasing information in nonempty strict subset of H_C^A.

Example 3. We shall show that with incomplete revelation, the planner can achieve more than what he can achieve with complete revelation, even when $\overline{H}(\sigma) = H$. Assume a "social

[25] In other words, all individuals in each subgroup are required to choose the same unique strategy among its best-responses.

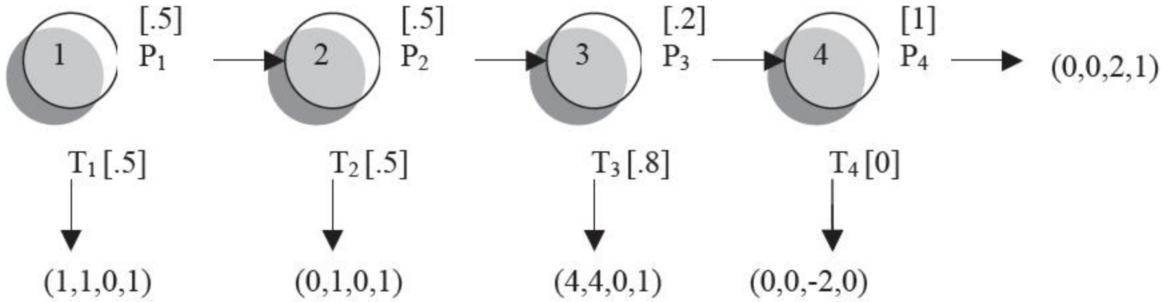

Figure 3: The beneficial superstition game.

planner", as in example 1. Consider the self-confirming equilibrium presented in Figure 2, which is the profile $\sigma = \{(0.5P1, 0.5T1); (0.5P2, 0.5T2); (0.2P3, 0.8T3); P4\}$.

The pure strategies for all players are "pass" (the horizontal move) or "take" (the vertical move). Half of individual 1's and half of individual 2's do not pass, although it would clearly be optimal for them to do so given the behavior of individual 3's. The beliefs supporting this self-confirming equilibrium are as follows. Agent 3's who "take" believe that 4's "take" with probability $\alpha > \frac{1}{2}$ and individual 2's who "take" believe that 3's "pass" with probability 1. Finally, agent 1's who take believe that 2's "take" with probability $\frac{3}{4}$ and 3's pass with probability 1. Of course, all agents have correct beliefs about all the other nodes.

The best outcome for society is $(4, 4, 0, 1)$. There are many possible announcements that may increase the frequency of this outcome. If the planner announces the aggregate play of 3's, she can induce agent 1's and 2's to enter. However, if she were to announce also the play of 4's, all individual 3's would pass, and the outcome would be $(0, 0, 2, 1)$, which is clearly worse for the planner.[26] In this example, complete revelation would not work, because some players have a "superstition" (wrong beliefs) that is beneficial for society and should be maintained. Player 3's who "take" have this "beneficial superstition".

So, assume that the planner announces aggregate behavior at node 3. Best responding to the new beliefs leads to $\sigma^* = \{(P1; P2; (0.2P3, 0.8T3); P4\}$. Note that this profile describes the best response of all populations and all individuals, with each individual having his old beliefs for all nodes except node 3 (this follows from the independence of beliefs). For example, half of individual 1's who pass believe that 2's take with probability $\frac{3}{4}$ and the other half believe that 2's take with probability $\frac{1}{2}$. However, this profile is also a self-confirming equilibrium: individual 1's who pass best-respond to the actual distribution of play σ^* as well. Player 3's believe that 1's and 2's pass with probability $\frac{1}{2}$, but still their action is optimal given the true distribution of play in nodes 1 and 2 and their beliefs about node 4. Therefore, when these players update their beliefs as they observe moves on the equilibrium path, this only reinforces their choices given their (fixed) beliefs for the nodes they never reach.

Clearly, $U^{PL}(\sigma^*) > U^{PL}(\sigma)$, since a greater mass of the population achieves $(4, 4, 0, 1)$

[26] Also note that this is not an equilibrium. Moreover, it should be emphasized that if player 1's realized that aggregate play is common knowledge, and in addition could think strategically given others' payoffs, they would not pass. However, here we assume that players do not know the payoffs of their opponents.

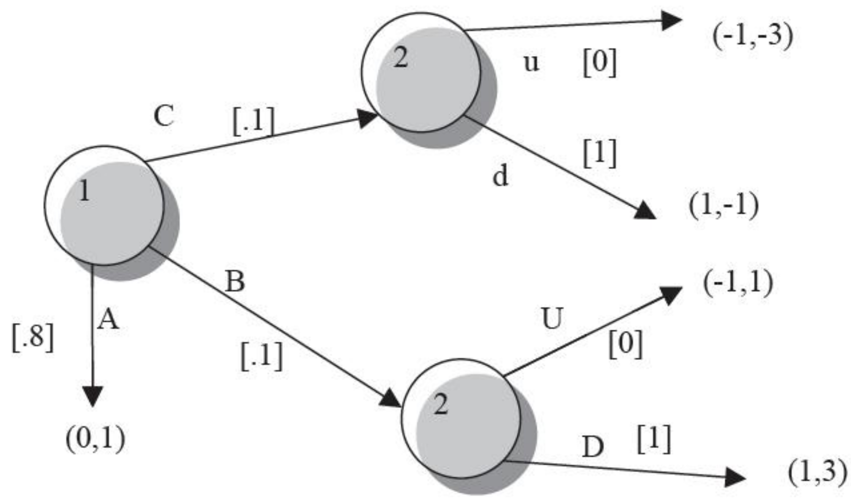

Figure 4: A non-strictly unstable equilibrium.

under σ^*, and this result cannot be achieved with full information revelation.

Note that this showed the existence of a subset of players, whose information sets are reached with positive probability under σ, and which, if revealed, lead to a better self-conforming equilibrium for the planner. There are other subsets J that could achieve this result, such as $\{2, 3\}$.

4 Strict Revelation Instability Relative to Planner's Preferences

Definition 7*b* captures the idea that a self-confirming equilibrium is "unstable" if the planner can change public behavior in a predictable way in order to achieve a new equilibrium, preferable for him. In the following examples, we show a weakness of this definition and we introduce a "strictness" requirement to correct it.

Example 4. As before, assume that the planner's utility function is such that he maximizes social welfare. The game is illustrated in Figure 4. The equilibrium described by the numbers in brackets is full revelation unstable. The problem is that after information about the behavior of player 2's is revealed, player 1's are indifferent between actions B and C.

In particular, individual 1's that choose A believe that a fraction $p_1 > \frac{1}{2}$ of individual 2's choose u and a fraction $p_2 > \frac{1}{2}$ of 2's choose U(given that the respective nodes are reached, of course). If the planner were to reveal the fact that, in both their nodes, 2's choose the action that gives high payoffs to individual 1's, then 1's would not play A. But given the fact that they are now indifferent between choice B and choice C, it is not clear how they will play following the information release. In other words, their transition beliefs support multiple profiles. For information revelation to lead to a better social outcome, hence in order to satisfy definition 7*b*, it is necessary that individual 1's choose action B, not action C. There is no obvious reason why these agents would choose this. Hence, the planner cannot guarantee that he will achieve higher payoffs with information revelation. Hence, the idea that this equilibrium is unstable relative to the planner's preferences,is not as compelling as in our previous examples. Therefore, we define the following concept:

Definition 8. A self-confirming equilibrium σ, supported by beliefs μ, is strictly partial revelation unstable relative to the planer's preferences, if there exists an information revelation set H^A on σ, such that for all profiles σ^* supported by H^A, the following hold:

1. σ^* is a self-confirming equilibrium, which $\forall i, \forall s_k^* \in supp(\sigma_i^*)$ is supported by the transition beliefs μ_{i,s_k}^* for all $h_j \in \{H_{-i} - \overline{H}(s_k^*, \sigma_{-i}^*)\}$.

2. $U^{PL}(\sigma^*) > U^{PL}(\sigma)$.

3. $u_i(s_k^*, \mu_{i,s_k}^*) > u_i(s_k, \mu_{i,s_k}^*)$ for some i, some $s_k \in supp(\sigma_i)$, and some $s_k^* \in supp(\sigma_i^*)$.

Information revelation can unambiguously lead to a better SCE for the planner in this case, regardless of the tie-breaking rule, because all possible new profiles are self-confirming equilibria and they are preferable for the planner. Note that, in our example, some player was indifferent between two actions reached with positive probability. We need to emphasize the fact that, perhaps counter-intuitively, the failure of strict instability relative to the planner's preferences can occur even in the absence of such ties in the payoffs of two best-responses.

Proposition 2. There exist strictly heterogeneous self-confirming equilibria in perfect information games, which are partial revelation-unstable relative to the planner's preferences, and yet fail to be strictly partial revelation unstable relative to the planner's preferences (Example 5).

Example 5. The reason the proposition is true is that in a self-confirming equilibrium, we allow different subgroups of each population to have different beliefs about information sets that are not reached in equilibrium. Consider the perfect information game of Figure 5 and the self-confirming equilibrium described by the fractions in the brackets. In this equilibrium, a large fraction of agent 1's chooses action t, resulting in a social surplus of 5, whereas the planner would like to push the players towards the outcome $(3, 3, 3, 3)$, which gives a social surplus of 12. As usual, we need to specify the beliefs of each subgroup for the information sets of opponents which are not reached by this subgroup. Individual 1's who play t believe that agent 2's play T with probability 0.9, and also correctly believe that individual 4's, in both their nodes, as well as agent 3's in their node, play left with probability 1. Individual 1's who play strategy pM, as well as agent 2's, believe that 4's play right with probability 0.9 in both their nodes.

If the planner reveals the behavior of agent 2's in their node, all individual 1's will now choose p. Clearly, this self-confirming equilibrium is partial revelation unstable relative to the planner's preferences, (since for $H^A = \{$ "2's information set"$\}$, there is a best-response to the transition beliefs generated, pL, which achieves the good outcome for the planner with probability 0.9). However, individual 1's who played t will have two best responses, pL and pR, and once more it is not clear enough why they would choose pL. The heterogeneity in the beliefs of population 1 about 4's behavior, off the equilibrium path, was enough to create the problem of non-strictness in this case.

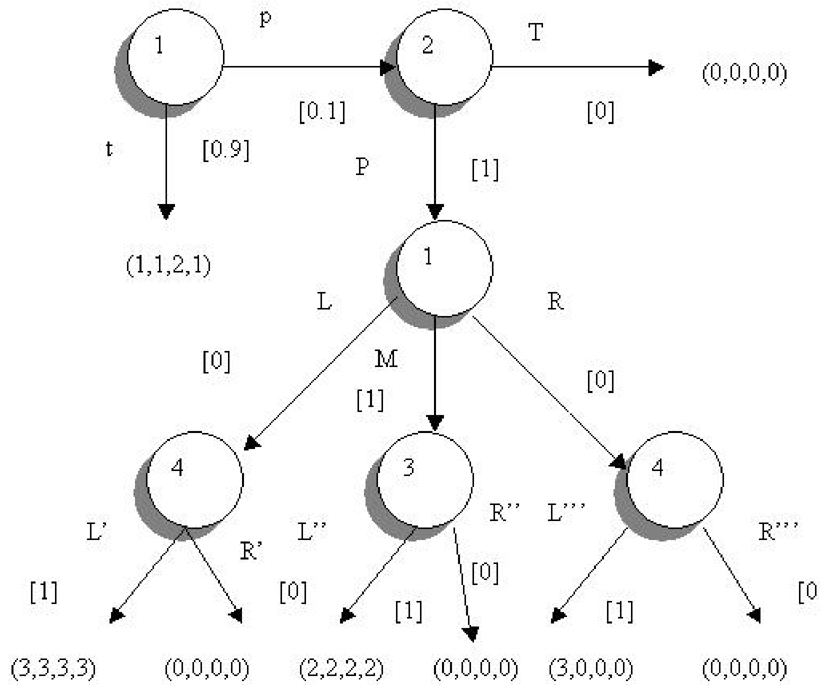

Figure 5: Failure of strict instability even in a strictly heterogeneous SCE.

5 Revelation-Robust SCE

After having introduced full and partial revelation instability we can now introduce our main concept:

Definition 9. A self-confirming equilibrium σ is called "revelation-robust" if it is not partial revelation-unstable.

Beliefs μ are called unitary if, for all i, $\pi_j(\widehat{h_j/\mu_{i,s_k}}(j)) = \pi_j(\widehat{h_j/\mu_{i,s_l}}(j))$ for any two strategies s_k, s_l, in the support of σ_i, for all $j \neq i$, and for all $h_j \in H_{-i}$. In other words, in a self-confirming equilibrium supported by unitary beliefs, the same beliefs are used to rationalize all pure strategies of a given mixed strategy.

Definition 10. Beliefs μ are called on-path unitary if, for all i, $\pi_j(\widehat{h_j/\mu_{i,s_k}}(j)) = \pi_j(\widehat{h_j/\mu_{i,s_l}(j)})$ for any two strategies s_k, s_l in the support of σ_i, for all $j \neq i$, and for all $h_j \in H_{-i} \cap \overline{H}(\sigma)$.

Proposition 3. All self-confirming equilibria supported by on-path unitary beliefs are revelation stable.

Proof/ Let σ be such a self-confirming equilibrium. For all players i and all opponents $j \neq i$ the following holds: if $h_j \in \overline{H}(\sigma)$, then $h_j \in \overline{H}(s_k, \sigma_{-i})$ for some $s_k \in supp(\sigma_i)$. Thus, Condition 2 in the definition of a self-confirming equilibrium implies that for all players i, $\pi_j(\widehat{h_j/\mu_{i,s_k}}(j)) = \pi_j(\widehat{h_j/\sigma_j})$ for all $s_k \in supp(\sigma_i)$ and all $h_j \in H_{-i} \cap \overline{H}(\sigma)$. Hence, the initial beliefs of all individual i's, μ_i, must be correct for all $h_j \in \overline{H}(\sigma)$, and for all $j \neq i$. It follows that for any information revelation set H^A, the transition beliefs μ^* generated by H^A are the same as the initial beliefs μ. Clearly, then, there is no $\sigma^* \neq \sigma$ such that condition 3 of definitions 1a and 7a hold. *QED.*

Corollary. All unitary self-confirming equilibria (thus, all Nash equilibria) are revelation-stable.

Proposition 4. Every finite game has a revelation robust self-confirming equilibrium.

Proof/ This simply follows from Nash theorem and the corollary above.

6 Two Economic Policy Applications

In this chapter we shall address information revelation from the policy point of view. Accordingly, we assume that the holder of information is always the social planner, who has utilitarian preferences, and hence wishes to maximize the sum of welfare across all members of society. First, we examine policies that promote experimentation with untested strategies, which may generate socially beneficial information. This is defined as affirmative action. Secondly, we will show how by revealing less information than possible, (a practice that we call self-censorship) the social planner may achieve the optimal equilibrium. we also specify a class games where self-censorship cannot be optimal.

6.1 Affirmative Action

It will be shown by example that proposition 3 is important for economic policy, because it provides a justification for "affirmative action". By this term, we mean providing agents with incentives to try novel strategies that will "test" the ability of members of unrepresented social groups to perform well in activities that they are expected to fail. The proposition shows that prejudice that totally prevents certain social groups from interacting with other groups is the most dangerous. Inducing individuals to experiment against their priors could generate socially desirable information, which, combined with selective information release, may facilitate reaching a better social equilibrium. This result is similar with that of Jackson and Kalai (1997) who, in a setting where agents always observe aggregate information, argue that socially valuable information cannot be generated if people's priors are such that they never try a certain action. Hence, these authors also argue that incentives should be given for "experimentation" even against one's priors.

Example 6. Consider an employment decision by Player 1 (the employer), who chooses whether or not he hires Player 2, a member of a minority group. The game is illustrated in Figure 6. If player 1 chooses not to employ player 2, we assume that both get zero. If he hires the minority member, then the latter may either choose to work hard, in which case the employer gets a payoff of 3, or she may choose to shirk, and his employer gets -1. The employer does not know if the potential employee is honest (or efficient) or not, so it is unknown whether her payoff from working hard exceeds the payoff from shirking. If employers believe that minority workers tend to shirk, the long-run state of the economy may end up

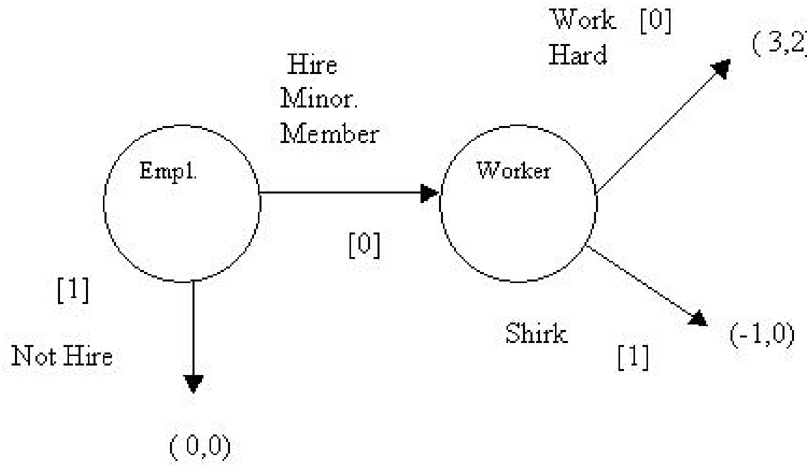

Figure 6: The Employment Game with a Unitary Equilibrium.

as shown in Figure 7. The state is characterized by the self-confirming equilibrium [Not hire, Shirk] where workers in fact are honest, but they never get the opportunity to prove it (note that "shirk" is optimal given the fact that employers never hire). Here, the equilibrium is supported by unitary beliefs, hence it is revelation-stable. Public information is not effective here. However, consider a policy that subsidizes the employment of minority members (for example, by offering an additional payoff of 2 to every employer who chooses "hire", regardless of workers' choice). Even if the program is "small-scale", and only a low fraction α of employers choose to hire, this is enough to generate the socially valuable information that minority workers are efficient. Public information then may shift the equilibrium to the socially efficient [hire, work hard] with social payoffs equal to 5.

6.2 Self-Censorship

It is worth considering conditions, under which, the planner may not want to reveal all available information given a SCE σ. We define "self-censorship" as the practice of not revealing all available information regarding the aggregate data. This issue is important for economic policy, with the recent debates regarding the role of the media in the proliferation of mass-murders and consumer panic. As far as we know, economic theory has not explicitly addressed the issue of self-censorship.[27] In the following paragraphs we shall try to characterize cases where "self-censorship" improves social welfare.

Formally, let σ be an information unstable self-confirming equilibrium supported by beliefs μ.

Definition 11. A (self-confirming equilibrium) profile $\sigma^* \neq \sigma$, which satisfies the conditions of definition 7b, relative to the preferences of the social planner, is called an "information

[27]The main arguments in the social debate regarding self-censoring are philosophical. There are major philosophical questions involved, related to ethical values such as freedom, but we argue that game theory can contribute to this debate as well, regardless of the importance of the philosophical questions.

dominant" (self-confirming equilibrium) profile over σ.

Let K_σ be the set of all information dominant self-confirming equilibria over the SCE σ (this can be of course, empty). Define $U^{PL}_{max} \equiv max_{\sigma' \in K_\sigma} U^{PL}(\sigma')$.

Definition 12. A self-confirming equilibrium σ is called "incomplete revelation improvable" if $U^{PL}_{max} > U^{PL}(\sigma'')$ for all σ'' supported by H^A_f.

In other words, a given SCE is "incomplete revelation improvable" if the optimal information revelation, given σ and μ, entails concealing some aggregate information that could be revealed. In the following, when we say that self-censorship is socially beneficial given a SCE σ, we shall mean that the particular SCE is incomplete revelation improvable.

The beneficial superstition game (Figure 3), is an example where complete revelation of the existing information may be socially detrimental. In games like this, there is a social group whose welfare is maximized at a bad social outcome, and the interests of different social groups are conflicting. Roughly speaking, this special social group corresponds to "criminals" who appropriate the material payoffs of others. Example 3 reveals that "criminals" should not be fully informed. This agrees with common sense, which dictates that it is not a good idea to reveal information that shows that "crime pays". We next show that, even in cases where the interests of all social groups are aligned, self-censorship may be useful.

Example 7. The following example shows that even if a strictly Pareto superior outcome exists, and it is reached with positive probability under a self-confirming equilibrium σ, self-censorship may still be optimal. Consider an a decision of a high-level public official (Player 1) whether to assign a lower level official (Player 2) to hire a contractor to build a school (S), or a park (P). The second official may choose to withdraw the project (W), or assign it to the final contactor (A). The contractor (there are Park Contractors and School contactors) could choose not to cheat and to do a good job (N.CH), or he could choose to cheat and exert low effort (CH). In fact, all contactors are honest, and everybody is better off when the school is built. Similarly, everybody benefits when the park is built, but to a lesser degree. Here there is no clash of interests in society, since everyone want the same outcome. As we shall see, heterogeneity in beliefs can still make self-censorship optimal.

Figure 7 illustrates the game, which has with a Pareto dominant outcome, where all players earn 5. As usual, the numbers in the brackets show the fractions of each population following each strategy in the equilibrium σ. There are two subgroups in the population of official 2's. Seventy percent of individual official 2's choose strategy [A,a] and thirty percent choose [W,a]. Note that the payoff-dominant terminal node is reached with positive probability. This equilibrium is supported by the following beliefs. Individual official 1's who play S believe that official 2's play w with probability $p > 0.9$, and Park Contractors always cheat. Individual official 1's who play P have correct beliefs about the behavior of individual official 2's and School Contractors. Finally, official 2's who play pure strategy [W,a] believe that School Contactors cheat with probability 0.8.

Now, complete revelation will make the outcome $(3, 3, 3, 3)$ be reached with probability equal to one. The reason is that, given the behavior of individual 2's, official 1's had better choose P. However, if the planner only announced the behavior of school contractors, then the optimal outcome would be achieved 90% of the time, which is clearly better for society. In this case, the optimality of self-censorship is caused by the heterogeneity of beliefs across

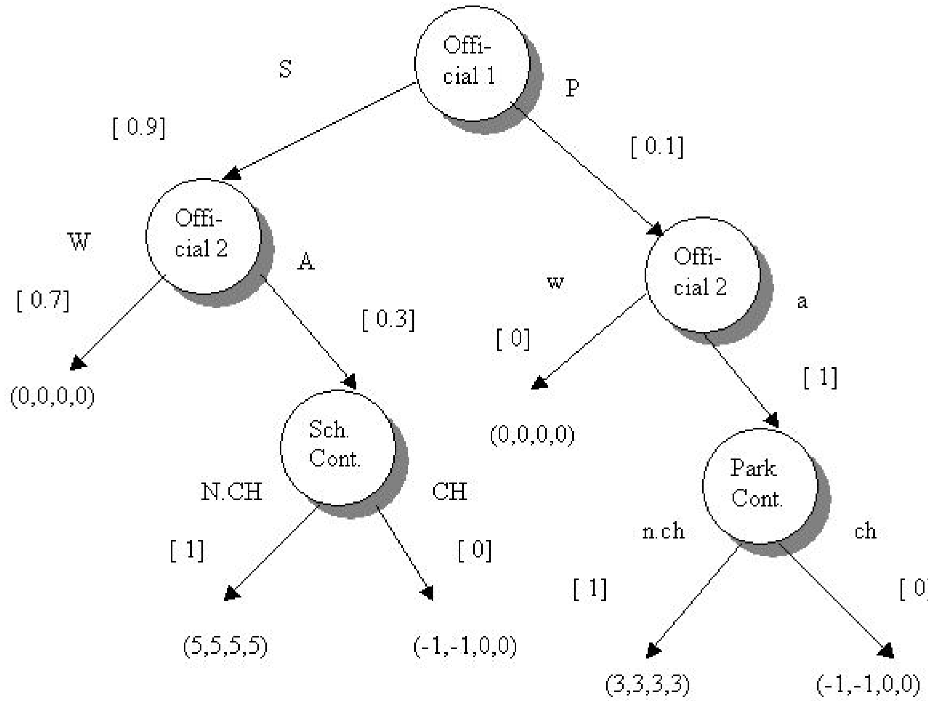

Figure 7: The Public Expenditure Game

two different populations, official 1's and official 2's.[28]

We shall specify a particular class of games for which self-censorship does not improve welfare.

Let G be the set of all extensive-form games that have a terminal node ψ with the following property: ψ is the unique *argmax* of $\sum_{i\in J} u_i(z)$ and also the unique *argmax* of $U_{\iota(y)}$, where y is the immediate predecessor of ψ. Note that $\Theta \subseteq G$.

Definition 13. A game Ξ is game of "monoambiguous choices", if for all players i and for all $h_i \in H_i$, there is at most one $\alpha' \in A(h_i)$ such that some $x \in \ell^{-1}(\alpha)$ is a decision node for some player.

That is, for each player i and for all information sets $h_i \in H_i$, a terminal node immediately follows all actions, except (possibly) one. An example of such a game is the "beneficial superstition" game, where each player had at most one action that was followed by some decision node. Note that if Ξ is a game of monoambiguous choices, all players have perfect information of other's moves.

Proposition 5. Let $\Xi \in G$ be a game of "monoambiguous choices". Let σ be a self-confirming equilibrium of Ξ such that $\psi \in \overline{Z}(\sigma)$ and ψ is not reached with probability one given σ. Then, σ is not incomplete revelation improvable.

Proof/ Clearly, all information sets in the game are singletons. Let $\widehat{\alpha}$ be the path of

[28]Of course, the fact that this SCE is revelation unstable still depends critically in the heterogeneity of beliefs of a single population, that is, the population of official 1.

actions that leads to ψ, which is indexed by the precedence relation of the tree. Let $\iota(t)$ be the player that moves at the t-th step of the path, T be the total number of steps, and $\widehat{\alpha}(t)$ denote the action at the t-th step of the path. Let also $h^t_{\iota(t)}$ be the information set of player $\iota(t)$ where action $\widehat{\alpha}(t)$ is available. Notice that $\pi_{\iota(T)}(\widehat{h^T_{\iota(T)}/\sigma})(\widehat{\alpha}(T)) = 1$ by the definition of G. Consider the set of all information sets of player i reached in the path to ψ, $H^{\widehat{\alpha}_i}$. If this set in nonempty, then the pure strategy $s_i[\widehat{\alpha}]$ that prescribes the choice of actions in $\widehat{\alpha}$ for all $h_i \in H_i^{\widehat{\alpha}}$ is optimal given beliefs that assign probability 1 to the true distribution of actions induced by σ. The reason for this is that, since ψ is reached with positive probability given σ, there are some player i's that choose pure strategy $s_i[\widehat{\alpha}]$. These players know the true σ, since because of the form of the game, the information sets on the path to ψ are the only ones reached with positive probability under σ. By "monoambiguous actions", it follows that they also know the exact payoffs they would get following any other strategy. Hence, for all players that have a decision in the path to ψ, $s_i[\widehat{\alpha}]$ is the optimal strategy when they know σ. Hence, following full information revelation, the welfare-maximizing node ψ is reached with probability one. Clearly this outcome is preferable to the planner than any other. *QED*

Corollary: Let Ξ be a game of "monoambiguous actions". Let there be a terminal node ψ such that ψ is the strictly Pareto superior outcome. Let σ be a self-confirming equilibrium of Ξ such that node ψ belongs to $Z(\sigma)$. Then, σ is not partial revelation improvable.

These results are interesting because we have pinpointed two different reasons why self-censorship may be socially beneficial. The first is conflict: some social groups may win when the rest lose. The second is the fact that some players have more than one actions with ambiguous payoffs (meaning that expected payoffs of these actions depend on other's behavior).

7 Discussion and More Examples

There are important implicit assumptions behind our basic model that should be defended. First of all, it seems that our agents are "naive" in the sense that they do not understand that other populations will change their behavior after the announcements. We have already underscored the fact that sophisticated agents who do not know the payoffs of other agents (including the planner) will behave in this manner as well. Secondly, it has been pointed out in seminars that it seems easier for the planner to reveal agents' utility functions, rather than their actions. We believe that this impression is simply wrong. Many of our important examples involve uncertainty about the moral incentives of agents, which are not directly observable. The notion that the planner reveals the utility function of officials in Example 1 seems nonsensical, but he may reveal their behavior.

Moreover, the informational requirements for the planner appear too strong. How does the planner know the moral payoffs in Example 1? Our answer to this question is based on revealed preference. If the planner can see in the aggregate data that all officials cooperate, he can infer their preferences. A seemingly stronger assumption is that the planner knows agents' beliefs. We argue that much can be inferred from the aggregate data about beliefs as well. In Example 2, there is a specific range of beliefs about opponents' actions that rationalizes the choices of agent 1's and 2's who play "take". To sum up, although some of our assumptions seem excessively strong, they are many important cases where they need not be so.

We shall now discuss the issue of management of public information more broadly, providing examples that may or may not be fully captured by our simple model. In fact, we believe that there are other theoretical reasons to expect that aggregate information revelation can direct the behavior of the public, such as preferences for conformity. A large literature in psychology explains where this type of preferences stems form. This paper makes only a first step in the direction of endogneizing the release of public information. It is certainly worthwhile extending the model to capture such issues as conformity and reciprocity preferences, but this is beyond the scope of this paper.

Governments may follow implicit strategies of selective information revelation. [29] An obvious example is information release in times of war and national emergency. It is clear that a government will not want to reveal public information about the actions of deserters who flee the country. Even in less extreme circumstances, we argue that the media may tend to conform to the government's view on what information should be revealed. In most countries, the State typically do not provide accurate data about the amount of people who escape capture. This practiced is explained by our model and illustrated in the "Beneficial Superstition Game" of Figure 2. 1's and 2's are two populations of investors who get a high benefit if they cooperate $(P1, P2)$. The group of player 3's are potential thieves who can grab part of the surplus (move $T3$) or not (move $P3$). Players 4's are police officers who may catch the criminal (move $T4$) or shirk (move $P4$). As we have seen, the planner should not reveal the distribution of actions of police officers here. This is an example where society is better off when certain agents, whose actions entail significant externalities on others, are ignorant of the true distribution of actions. A potential criminal may have the erroneous belief that the probability of punishment is high, an error which is socially beneficial.

Furthermore, governments' policies against social discrimination may involve selective information revelation, and an implicit agreement with the media is likely to be achieved. In order to serve the public good, the media may agree to refrain from emphasizing certain types of information. Thus, it is acceptable to conceal aggregate information that reinforces social stereotypes. It might also be the case that information against these stereotypes is overemphasized. A typical example of this is the extensive media coverage of the cases where women are performing jobs that are considered "men's jobs". In our interpretation, this may be done in order to change expectations about the women's strategies in the population, and hence to change the optimal strategies when playing *against* a woman.[30] In some sense, therefore, selective information revelation is the management of self-fulfilling prophecies.[31] Another example is the management of expectations regarding the extent to which institutions (apart from police) work properly. The professional basketball leagues of the NBA and

[29] It is worth emphasizing that governments and special interests would not like to be criticized for manipulating the public's behavior and for self-censorship. Moreover, manipulation is more effective when it is covert. For these reasons, the descriptive validity of our approach becomes difficult to substantiate, because the importance of information management is typically not explicitly acknowledged. Hence, we will often cite examples where the importance of self-censorship is implicitly presupposed, but not explicitly stated.

[30] Preference for conformity within the group of women also plays an important role here.

[31] See Hargreaves-Heap and Varoufakis (2002) for strong experimental evidence for this.

Euroleague have explicit policies that punish public statements against referees. It is clear that media commentators take this into account, as they tend to conceal referees mistakes and emphasize their correct decisions.

There are also significant applications related to antisocial behavior, investor sentiments, and opinion polls in voting. The media may deliberately refrain from publicizing the behavior of the underclass, in order to avoid encouraging antisocial behavior. Since many forms of antisocial behavior depend on the non-pecuniary social rewards that the actors of such behavior may reap during the interaction with their peers, information about the extent of such antisocial phenomena is handled carefully.[32] Moreover, policies aiming to protect investor sentiments often selectively conceal information. After the Great Depression, new institutions and policies were created in many western countries in order to protect against investors' pessimism. Stock market authorities may selectively reveal aggregate data in order to avoid worrying investors and promote optimism. Furthermore, in some countries, the State restrains the use of public opinion polls during election periods. There is much evidence that voters like to vote for the wining party.[33] A specific political party and the interest groups that support this party, may want to selectively reveal polls that show that the party is winning, and conceal the ones that show that it is loosing. This may manipulate the election results, and, in many countries, restrictions on polls during election campaigns have been imposed.[34]

The previous example makes it clear that it is not only the "benevolent social planner" that may use selective information revelation to achieve specific objectives. In fact, the planner need not even be unique. The two opposing parties may both reveal poll information, each to maximize its probabilities of winning. Advertising is a major example where information revelation is selective: the publisher of a book will promptly announce that the book has sold a million copies, but this will not be the case if only a few copies have been sold. Here the planner is the firm, which has special access to data regarding its sales, and selectively reveals it.

8 Conclusions

We used a dynamic framework with anonymous interactions to examine how public information may be revealed to manipulate the behavior of individuals. We showed that the "planner", who knows the aggregate information, may affect behavior, even after an equilibrium has been reached. In particular, he can "push" the economy to his preferred equilibria by selectively revealing information. In this sense, some self-confirming equilibria are not robust to information manipulation. This "revelation-instability" critically depends on the heterogeneity of beliefs across agents of the same population. Equilibria with unitary beliefs,

[32]For an anecdote, according to a Dutch journalist, there is an implicit agreement in the Dutch press to refrain from overemphasizing the occurrences of sports violence and hooliganism, in order not to encourage "potential new hooligans".

[33]This has been supported by many studies, and it is called "the bandwagon effect". Preference for conformity seems to be a major driving force for this phenomenon.

[34]See Michalos, p. 410 and Morwitz and Pluzinski (1996), p. 53. The countries that have implemented orconsider implementing a ban on political polling during election periods include Brazil, France, Canada and Germany.

including Nash equilibria, are robust to such manipulation. Hence, information revelation-robust equilibria always exist. We presented several examples illustrating how concealing information (self-censorship) and inducing experimentation (affirmative action) can be socially beneficial in certain cases. Finally, we presented a wide range of social phenomena which fit well with our approach, including propaganda in times of crisis, prohibition of polls during campaigns, and policies of sports leagues not to emphasize the referees' mistakes.

The model could be extended in several different directions. Firstly, experimental evidence indicates that "social preferences" play a major role when aggregate information is revealed. Incorporating such preferences, especially conformity and reciprocity preferences, in the model would be a promising path of research. Secondly, using an explicitly dynamic approach would be fruitful, because it would allow us to examine the potential for many information revelations, rather than a single one. Moreover, in such an environment with multiple information revelations, it would be equally rewarding to study more sophisticated learning rules, and allow agents to predict changes in other's actions when information is revealed.

References

[1] Ball, Sheryl; Eckel, Catherine; Grossman, Philip G. and Zame, William. "**Status in Markets**". Quarterly Journal of Economics, Vol. 116, 2001, pp. 161-188.

[2] Banerjee, Abhijit. "**A Simple Model of Herd Behavior**". The Quarterly Journal of Economics, Vol. 107, No. 3, August 1992, pp. 797-817.

[3] Battigalli, P. and Dufwenberg, M. "**Guilt in Games**". AEA Papers and Proceedings 97 (2), May 2007, pp. 170-176.

[4] Berg, Joyce; Dickhaut, John and Kevin McCabe, "**Trust, Reciprocity and Social History**". Games and Economic Behavior 10, 1995, pp. 122-142.

[5] Bikhchandani, Sushil; Hirshleifer, David, and Welch, Ivo. "**A Theory of Fads, Fashion, Custom, and Cultural change as Informational Cascades**". The Journal of Political Economy, Vol. 100, No. 5, October 1992, pp. 992-1026.

[6] Camerer, Colin and Weigelt, Keith. "**Experimental Tests of a Sequential Equilibrium Reputation Model**". Econometrica 56 (1), 1988, pp. 1-36.

[7] Dufwenberg, Martin, and Gneezy, Uri. "**Information Disclosure in Auctions: an Experiment**". Journal of Economic Behavior and Organization, Vol. 48, 2002, pp. 431-444.

[8] Esponda, Ignacio. "**Information Feedback in First Price Auctions**". NYU Working paper, November 2006.

[9] Frey, Bruno and Meier Stephan. “**Social Comparisons and Pro-social Behavior. Testing ’Conditional Cooperation’ in a field Experiment**”. The American Economic Review 94(5), 2004, pp. 1717-1722.

[10] Fey, Mark; McKelvey, Richard and Palfrey, Thomas. “**An Experimental Study of Constant -sum Centipede Games**”. International Journal of Game Theory, Vol. 25 No. 4,1996, pp. 269-287.

[11] Friedman, Daniel. “**Equilibrium in Evolutionary Games: Some Experimental Results**”. The Economic Journal, Vol.106, No 434, January 1996, pp. 1-25.

[12] Fudenberg, Drew and Levine David K. “**Superstition and Rational Learning**”. 2005, Working paper.

[13] Fudenberg, Drew and Levine, David K. “**Self-Confirming Equilibrium**”. Econometrica, Vol. 61, May 1993, No. 3, pp. 523-545.

[14] Fudenberg, Drew and Levine, David K. “**Steady-State Learning and Nash Equilibrium**”. Econometrica, Vol. 61, May 1993, No. 3, pp. 547-573.

[15] Fudenberg, Drew and Levine, David K. “**The Theory of Learning in Games**”. The MIT Press, Cambridge, Massachusetts, 1996.

[16] Fukuyama, Francis. “**Social Capital and Development: The Coming Agenda**”. SAIS Review, Vol. XXXII, No. 1, Winter-spring 2002.

[17] Fukuyama, Francis. “**Social Capital, Civil Society and Development**”. Third World Quarterly, Vol. 22, No. 1, pp. 7-20, 2001.

[18] Hargreaves-Heap, Shaun and Varoufakis, Yanis. “**Some experimental Evidence on the Evolution of Discrimination, Cooperation and Perceptions of Fairness**”. The Economic Journal 112, 2002, pp. 679-703.

[19] Harrison, Glenn and McCabe, Kevin. “**Expectations and Fairness in a Simple Bargaining Experiment**”, International Journal of Game Theory 25, 1996, pp. 303-327.

[20] Jackson, Matthew, and Kalai, Ehud. “**Social Learning in Recurring Games**”. Games and Economic Behavior 21, 1997, pp. 102-134.

[21] Kalai, Ehud, and Lehrer, Ehud. “**Rational Learning Leads to Nash Equilibrium**”. Econometrica, Vol. 61, No. 5, September 1993, pp. 1019-1045.

[22] Mailath, George. “**Do people Play Nash Equilibrium? Lessons from Evolutionary Game Theory**”. Journal of Economic Literature 36(3), September 1998, pp. 1347-1374.

[23] Maniadis, Zacharias. “**An Experimental Study of Aggregate Information Revelation in the Centipede Game**”. Working paper, June 2007.

[24] McKelvey, Richard and Palfrey, Thomas. "**An Experimental Study of the Centipede Game**". Econometrica, Vol. 60 No. 4, July 1992, pp. 803-836.

[25] Morris, Stephen and Song Shin, Hyun. "**Social Value of Public Information**". American Economic Review 92(5), 2002, pp. 1521-1534.

[26] Putnam, Robert. "**Bowling Alone: America's Declining Social Capital**". Journal of Democracy, 6.1, pp. 65-78, 1995.

[27] Weibull, Jorgen. "**Evolutionary Game Theory**". The MIT Press, Cambridge, Massachusetts, 1997.

[28] Marks, Gary and Miller, Norman. "**Ten Years of Research on the False-Consensus Effect: An Empirical and Theoretical Review**". Psychological Bulletin, Vol. 102, 1987, pp. 72-90.

[29] Michalos, Alex. "**Ethical Considerations Regarding Public Opinion Polling During Election Campaigns**". Journal of Business Ethics 10, 1991, pp. 403-422.

[30] Morwitz, Vicky G. and Pluzinski, Carol."**Do Polls Reflect Opinions or do Opinions Reflect Polls? The Impact of Political Polling on Voters' Expectations, Preferences, and Behavior**". Journal of Consumer Research, Vol. 23, June 1996, pp. 53-67.

[31] Ortmann, Andreas; Fitzgerald, John and Boeing, Carl."**Trust, Reciprocity and Social History: a Re-examination**". Experimental Economics 3, 2000, pp. 81-100.