Selective Revelation of Public Information and Self-Confirming Equilibrium [*]

Zacharias Maniadis[†]

October 8, 2011

Abstract

We model aggregate information release, in a dynamic setting with random matching, as a conscious, preference-driven choice. Starting from the environment of Fudenberg and Levine (1993a), we introduce a "planner", who possesses and selectively reveals aggregate information. Aggregate information is gathered slowly, by taking small samples from the population, and can only be revealed after the dynamic process has stabilized. By selectively revealing information, the planner may upset a given self-confirming equilibrium, in order to achieve a preferred outcome for him. Hence, some self-confirming equilibria are "unstable" relative to public information release. We show that only equilibria supported by heterogeneous beliefs can be information-unstable. We provide several real-life examples of manipulation by means of public information, showing the relevance of the theoretical analysis.

1 Introduction

Social interactions among strangers can be modeled as games of large populations with anonymous matching.[1] In belief-based models, the optimal choice of an individual, who is

[*] This paper is based on Chapter 2 of my doctoral dissertation at UCLA. I am particularly grateful to David K. Levine and Bill Zame for their support and guidance during this project. I greatly benefited from discussions with Jean-Laurent Rosenthal and Paolo Battigalli. Seminar participants at UCLA, U Bocconi, U of Maastricht, U of Granada and the University of Athens, as well as participants at the La Pietra-Modragone workshop if Florence, contributed with insightful comments. Comments by two anonymous referees and an associate editor greatly contributed to improving a previous version. The author is also grateful to the Greek State Scholarships Foundation (IKY) and the Cournot Center for Economic Research for financial support.

[†] Department of Decision Sciences and Carlo Dondena Center, Bocconi University

[1] In such games, the steady states of the dynamic recurring interaction, with no strategic links across repetitions of the interaction, have a close relationship with the equilibria of the static game (Fudenberg and Levine, 1993b).

matched against an individual opponent, depends on the former's expectations concerning the behavior of the opponent's population as a whole. However, people rarely have enough interactions with all other social groups (which are modeled as populations of "opponents" in the game) in order to form accurate expectations about the behavior of their members. The notion of self-confirming equilibrium of Fudenberg and Levine (1993*a*) (henceforth referred to as FL) describes a state where people optimize given their beliefs about other groups, but individuals' beliefs need not be correct about groups they do not personally interact with.[2]

However, governments and special interests often have access to aggregate data about the behavior of social groups. By revealing their private 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.[3] 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, consider Figure 1, which describes the interaction between two social groups, investors and officials. Investors move first, deciding whether to enter (invest) or not, and then officials choose whether to cooperate or not.[4] The investment is profitable only if the official cooperates. Each number in the brackets represents the fraction of the particular population that follows each action, in the specific "state" of the dynamic system we are considering.[5] In the state illustrated in Figure 1, 20% of investors "enter" and thus know the truth: that officials are upright,[6] and they always cooperate without asking for a bribe. However, 80% of investors refrain from investing, holding strong prior beliefs that officials are corrupt. This state of affairs, being a self-confirming equilibrium, is stable in the sense of FL. We claim that this equilibrium is implausible. Investors who do not enter would

[2] The word "personally" is very important here, since 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 interacting with other social groups. Battigalli (1987), and Kalai and Lehrer (1993) also introduced concepts similar to self-confirming equilibrium, but they ruled out this type of heterogeneous beliefs within populations.

[3] This is especially relevant in modern societies, where the media can easily convey public information. This information need not necessarily be exogenous, because the availability of aggregate data could depend on the incentives of those who have them.

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

[5] Strictly speaking, this is not the state of the system, because a specification of beliefs is also necessary, in order to determine the future evolution of the system. We will use the same convention as Fudenberg and Levine (1998), calling σ "the state".

[6] This is captured by officials' negative payoffs from expropriating rents. Their pure monetary payoffs might be positive, but their overall utility is negative.

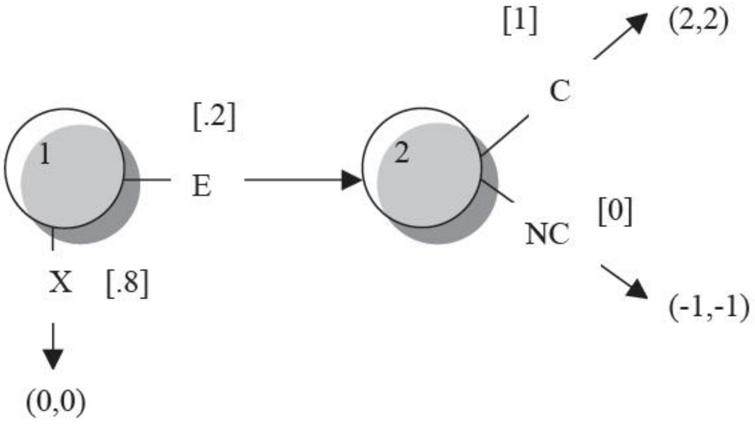

Figure 1: The modified cooperation game.

change their behavior if they knew the true behavior of officials. Moreover, if the government has collected credible data about officials' behavior and wishes to maximize the social surplus, it ought to publicize the relevant information in the media. By revealing the fact that officials are honest, the government will change investors' beliefs (except of those who already "enter", and have correct beliefs) and induce them 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 self-confirming equilibrium. Moreover, the government prefers this steady state than the previous one, so it has an incentive to reveal this information.

The main contribution of this paper is conceptual. We derive a refinement of self-confirming equilibrium, the properties of which we illustrate with a number of examples and some simple propositions. The model uses the framework of FL (1993*a*). The novel element is that we examine how the existence of a "planner" - an agent who knows and selectively reveals public information - alters the predictions about long-run behavior. Our key insight is that, deciding whether a particular self-confirming equilibrium is a plausible rest point for the dynamic social interaction, one should look at the incentives of the possessors of public information.[7] 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 self-confirming equilibria supported by unitary beliefs (that is, where all individuals in a given population have the same beliefs, as in Nash equilibrium) are revelation-robust. Our framework has a wide variety of potential applications in industrial organization, political

[7] We take the costless knowledge of aggregate statistics by the planner as given. Our setting can easily take into account costs of aggregate information acquisition.

economy, public policy and other fields.

In the papers which are closer to our spirit, Esponda (2008) and Jehiel (2011) have independently worked on theoretical models, which address the issue of manipulation by means of selective information release. They also employ notions similar to self-confirming equilibrium, and they focus on a specific type of games, namely auctions. They ask whether the equilibrium feedback policy, which in most cases may be decided by the auctioneer, may affect equilibrium outcomes.[8] Our paper deals with abstract extensive-form games and provides a general refinement of self-confirming equilibrium. The advantage of working in a general setting is that our approach models a wide array of interesting real-life phenomena of manipulation by means of aggregate information revelation.[9]

The remainder of the paper is organized as follows. In part two we introduce the model and we define Nash and self-confirming equilibrium. In part three we introduce the planner, define our main concepts, and provide some simple propositions and examples that illustrate the definitions. A general discussion follows in part four. Part five concludes.

2 The Model

The model examines the steady states of a system with dynamic, anonymous interactions and learning. In each period, all individuals in a given population-role are randomly matched with other individuals, each of whom belongs to a different population-role. Our point of departure is Fudenberg and Levine's approach ($1993a, 1998$), which assumes that players see only the result of play in their own matches. We take as given the main results of their research, especially the possibility of the game settling in a self-confirming equilibrium with non-Nash outcomes. Our objective is to examine how the "planner" can selectively convey public information, in order to change the equilibrium outcome. We assume that the planner can reveal information only once, after play has converged to a steady state, because garnering aggregate information is a slow process. We show that some self-confirming equilibria are not plausible in the presence of the planner, because by selectively revealing aggregate data, the planner can move the system to a different state.

[8] They thus provide a specific example of a "planner" and how he selectively reveals information about the aggregate data to maximize his objective value.

[9] The experimental literature has also addressed the issue of whether revealing aggregate information matters, and whether expectations can be manipulated. Roth and Schoumaker (1983) and Harrison and McCabe (1996) directly manipulated subjects' expectations about others' play in an ultimatum game, with significant and lasting effects. Berg et al. (1995) and Ortmann et al. (2000) performed experiments of one-round trust games, and found some support for the notion that information revelation of aggregate data can push the economy to desirable equilibria. Similar results were found in Frey and Meier's field experiment (2004).

2.1 The Extensive-Form Dynamic Game

A given extensive-form game is played repeatedly among anonymous agents randomly matched with each other.[10] Each individual knows the extensive form of the game, the realized terminal node after each match, and her payoffs for each terminal node, but not necessarily the payoffs of other individuals.[11] The extensive-form game is as follows. The set of players is $J = \{1, 2, \ldots, I\}$. The game tree X has finitely many nodes $x \in X$. Terminal nodes of the tree are denoted $z \in Z \subset X$. Information sets, which form a partition of all nonterminal nodes of the tree, 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 $A_i \equiv \bigcup_{h_i \in H_i} A(h_i)$. 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 $h_i \in H_i$. 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 each player, and we denote it by $s \in S = \times_{i \in J} S_i$. To avoid unnecessary complications, in the remainder of the paper, we will bundle equivalent pure strategies together, hence S_i will denote the set of reduced normal form strategies for player i. Let $\Delta(A)$ denote the set of probability measures over some set A. A mixed strategy for player i is a probability distribution over pure strategies,[12] $\sigma_i \in \Delta(S_i)$, and a profile of mixed strategies is denoted by $\sigma \in \times_{i \in J} \Delta(S_i) \equiv \Sigma$. Let $p(x/\sigma)$ be the probability that node x is reached under the profile of mixed strategies σ. The payoff for each player depends on the terminal node. So, for players $i = 1, 2, \ldots, I$, the payoff function is $u_i : Z \to \mathbb{R}$.

$H(s_i)[Z(s_i)]$ denotes the subset of all information sets [terminal nodes] reachable when player 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. A behavior strategy π_i for player i is a map from the set H_i, the family of all information sets where this player has the move, to probability distributions over moves. That is, $\pi_i(h_i) \in \Delta(A(h_i))$. Denote the set of all such strategies for player i by Π_i and denote by $\pi \in \times_{i \in J} \Pi_i$ a profile of behavior strategies. Let also Π_{-i}

[10] By the term "individual" or "agent", we shall refer to 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).

[11] This captures the fact that our social interactions are anonymous. A individual need not know, and need not have strong beliefs 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 need not know whether other officials share his preferences. He might believe that other officials are corrupt, so they get a higher payoff by not cooperating. In general, learning models emphasize the fact that agents form beliefs by observing others' behavior, rather than by introspection.

[12] Note the specific interpretation of mixed strategies here. Each individual is assumed to play a pure strategy, but each population as a whole randomizes across strategies, since individuals in the same population may be choosing different pure strategies.

be the space of behavior strategies for players other than i. We assume perfect recall, so by Kuhn's theorem, every mixed profile induces an equivalent profile of behavior strategies. Let $\pi(\widehat{h_j/\sigma_j})$ denote the distribution of actions at information set h_j induced by mixed strategy σ_j for player j. Let also $p(x/\pi)$ $[p(h/\pi)]$ be the probability that node x [information set h] is reached under the profile of behavior strategies π. Moreover, let $z(s)$ denote the terminal node reached when profile s is played.

Absent information revelation by the planner, players do not know others' 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 Π_{-i}, the set of profiles of behavior strategies of other players. Given player i's beliefs about other players' behavior strategies, the probability that terminal node z is reached when player i chooses pure strategy s_i is $p(z/s_i, \mu_i) = \int_{\Pi_{-i}} p(z/\pi_{-i}, s_i)\mu_i(d\pi_{-i})$. Accordingly, the expected utility of a player with beliefs μ_i, when she plays strategy s_i, is $u_i(s_i, \mu_i) = \sum_{z \in Z(s_i)} u_i(z)p(z/s_i, \mu_i)$.

In this environment, it is worthwhile to explicitly define Nash equilibrium in 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 $s_i \in supp(\sigma_i)$ there exists beliefs μ_i such that:

1. s_i maximizes $u_i(\cdot, \mu_i)$

2. $\mu_i[\{\pi_{-i} \in \Pi_{-i} : \pi_j(h_j) = \pi_j\widehat{(h_j/\sigma_j)}\}] = 1$ for all $h_j \in H_{-i}$

In words, a Nash equilibrium is a profile consisting of players' best responses to their beliefs about others' 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 sets of opponents. 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 sets 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 $s_i \in supp(\sigma_i)$, there exists beliefs μ_i such that:

1. s_i maximizes $u_i(\cdot, \mu_i)$

2. $\mu_i[\{\pi_{-i} \in \Pi_{-i} : \pi_j(h_j) = \pi_j\widehat{(h_j/\sigma_j)}\}] = 1$ for all $j \neq i$ and $h_j \in \overline{H}(s_i, \sigma_{-i})$

Consider a specific individual, who belongs to population i, and whose equilibrium strategy is s_i. The above definition means that this agent is required to hold correct beliefs about opponents' actions, only at information sets reached with positive probability under s_i and the profile of mixed strategies of "opponent" populations. Thus, an individual, who belongs to population i, may have wrong beliefs about the distribution of opponents' actions at an information set reached by other agents, who belong to population i. This may happen if these agents choose a different equilibrium strategy than the given individual. Therefore, beliefs held by each "subgroup" of population i, corresponding to a different pure strategy s_i, could differ. In a self-confirming equilibrium, 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 certain self-confirming equilibria. In this section we introduce the planner. The planner should be thought of as any institution or special interest which has control over public information, and also has an incentive to use it to affect the behavior of the public. The most natural example is a benevolent government, which collects aggregate statistics and would like to reveal information to induce pro-social behavior and economic growth. For another example, an auctioneer, who chooses the level of information feedback in an auction, wishes to maximize his own revenue (see Esponda, 2008, and Jehiel, 2011). The main idea is the following: if the planner, with aggregate information revelation, can achieve a new long-run outcome, which he prefers to that of a given self-confirming equilibrium, then the latter equilibrium is implausible.

The planner has payoff function $U^{PL} : \Sigma \to \mathbb{R}$, so his payoffs depend on the behavior of the populations that play the game. Although we shall not directly address the issue of optimal behavior by the planner, we should emphasize that the planner's payoffs matter for our analysis. The set of "possible payoffs" of the planner is not completely arbitrary, but constrained by the nature of each application. Since our definitions rule out self-confirming equilibria that the planner prefers to upset, this set of possible payoffs plays an important role.

The planner, who knows the true distribution of actions in each information set reached, can announce it for a subset of $H(\sigma)$. His announcements are true, and are always perceived as such.[13] It is important to emphasize that we are trying to capture a story where the planner can only gather information gradually, by sampling from the population. Since

[13] This can be thought 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. The assumption that the planner is credible 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.

many periods are required in order to garner the information, only steady-state information can be thus accumulated.[14] Hence, information can only be announced after a steady state has been reached.

We further assume that the planner has a conservative "maxmin" approach: he only reveals information if this unambiguously leads to a better equilibrium for him. It is worth explaining what this implies. As we shall see below, public information may change beliefs in such a way that several profiles of best responses exist. Some of these profiles might constitute self-confirming equilibria, and other profiles might not.[15] With respect to the equilibria, the planner may prefer only some of them to the "initial" equilibrium. Our basic hypothesis is that, because of the planner's conservative approach, he will reveal information only when all of the induced profiles of best responses constitute a better self-confirming equilibrium for him. In other words, the planner reveals public information only when this will certainly lead to higher payoffs for himself.[16]

3.1 The Full Information Revelation Setting

We define "full information revelation", as information revelation about the distribution of actions in every information set of the game. This section concerns only totally mixed self-confirming equilibrium profiles σ, so that $\overline{H}(\sigma) = H$. 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 in all information sets, there must be data available for him to disclose.

Definition 1. A self-confirming equilibrium σ is called "full revelation-unstable relative to the planner's preferences", if for every profile σ^* that satisfies condition 1 bellow, conditions $2-4$ are also satisfied.

1. For all i and for all $s_i^* \in supp(\sigma_i^*)$, s_i^* maximizes $u_i(\cdot, \mu_i^*)$, where μ_i^* satisfies

$$\mu_i^*[\{\pi_{-i} \in \Pi_{-i} : \pi_j(h_j) = \widehat{\pi_j(h_j/\sigma_j)}\}] = 1 \tag{1}$$

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

[14] In other words, only a sample gathered over many periods can be released, but if behavior changes in every period, such a sample cannot be representative of true behavior. Therefore, only information that describes behavior in a steady-state may be released.

[15] If the best responses do not constitute a self-confirming equilibrium, it is not clear where the system will stabilize.

[16] In our analysis, we will use the following assumption. If, after public information is released, some individual is indifferent between his pre-revelation strategy and a different one, she adheres to what she was doing before the information was revealed.

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. $u_i(s_i^*, \mu_i^*) > u_i(s_i, \mu_i^*)$ for some i, some $s_i \in supp(\sigma_i)$, and some $s_i^* \in supp(\sigma_i^*)$.

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

This definition simply says that an announcement of the true distribution of actions, in all information sets of the game, unambiguously leads to a better self-confirming equilibrium for the planner. Agents' beliefs μ^*, after the planner's full information revelation, assign probability 1 to the revealed distribution, induced by the "old" mixed profile σ. Best-responding to these beliefs will generate some "new" profile σ^* (note that profile σ^* need not have full support). Consider any such profile. By condition (2), the best-responses to the old distribution of actions are also best-responses, given the ensuing beliefs, for the profile that follows information revelation. Hence, the change in the state of the dynamic system following the planner's announcement is sustainable. Condition 3 ensures that at least one subgroup of some population has a strict incentive to change its behavior.[17]

EXAMPLE 1. We shall illustrate definition 1, 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 this example more formally here. If an individual player 1 believes that agents of population 2 tend to cooperate, the best-response is "enter", whereas if he believes that individual 2's do not cooperate, he should choose "exit". We further assume that the planner maximizes social welfare, so his payoffs are $U^{PL}(\sigma) = \sum_{z \in Z} \{p[z/\sigma] \sum_{i \in J} u_i(z)\}$.

Consider $\sigma = \{(0.8X, 0.2E); C\}$, the initial self-confirming equilibrium, illustrated in Figure 1.[18] This equilibrium is supported by the following beliefs. Individual 1's who exit believe that 2's cooperate with probability q, which is less than 0.2. All other subgroups, in both populations, have correct beliefs about the behavior of their opponents. Assume that the planner announces the true aggregate distribution of actions, in all information sets. Individual 1's simply best-respond to their beliefs about individual 2's distribution of actions, and they regard the information revelation as truthful. Accordingly, all individual 1's choose "enter" after the announcement, since they expect that 2's will cooperate. The behavior of individual 2's does not change following the announcement.

[17] In the absence of this condition, one could find examples of mixed Nash equilibria that would qualify as unstable, where it is not clear why aggregate information would change anybody's behavior.

[18] Note that this is just one of infinitely many self-confirming equilibria in this game. Any purely mixed strategy of player 1, coupled with pure strategy C, is a self-confirming equilibrium profile (and so is the pure profile $\{E, C\}$). We use this specific numerical example only for concreteness.

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 actions. The reason is that σ^* is a self-confirming equilibrium (actually also a Nash equilibrium), so players best-respond to the current distribution of actions as well. The planner prefers σ^* to the old profile, and thus has an incentive to fully reveal the aggregate information. Therefore, σ is full-revelation unstable relative to the planner's preferences.

We should clarify a few things about the implications of condition (1), which describes profiles that ensue from full information release. It should be emphasized that full information revelation need not, in general, lead to a Nash equilibrium, or even to a self-confirming equilibrium profile. This is because agents need not have correct beliefs about the distribution of actions that follows the information release. One may also wonder if, for a given totally mixed σ, 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 "left" or "right". Let the actions "down" and "right" be denoted (D_i, R_i) for player $i = 1, 2, 3$, and let (l, r) denote the actions of player 4. The profile $\sigma = [(0.5R_1, 0.5D_1); (0.5R_2, 0.5D_2); (R_3); (l)]$ is a self-confirming equilibrium profile, supported by the following beliefs. Individual 1's who play D_1 believe that 2's choose D_2 with probability 0.9, and have correct beliefs in other nodes. Individual 2's who play D_2 believe that the fraction of D_3 is 0.8, and have correct beliefs elsewhere. All other subgroups in all populations have correct beliefs. Individual 1's do not have an incentive to change their behavior after they learn the true behavior of opponents. Accordingly, information revelation in this case will lead to the new profile $\sigma = [(0.5R_1, 0.5D_1); (R_2); (R_3); (l)]$. This is a new self-confirming equilibrium, but not a Nash equilibrium, since individual 1's who play "down" would not choose this action if they had correct beliefs. We shall show that, when no "indifference" for some player i exists, full information revelation leads to a self-confirming equilibrium only if this equilibrium is equivalent to a Nash equilibrium.

Definition 2. For each i, let $b_i : \Sigma_{-i} \to S_i$ be the pure strategy best response correspondence of player i to the mixed profile of her opponents. If σ is a self-confirming equilibrium such that $\overline{H}(\sigma) = H$, we say that σ has a "unitary full revelation outcome" if whenever $s', s'' \in \times_{i \in J} b_i(\sigma_{-i}), s' \neq s''$, it holds that $z(s') = z(s'')$.

So, all profiles of best responses to the revealed information lead to the same terminal node. Note that, if each player i has a unique pure strategy best response to σ_{-i}, this is

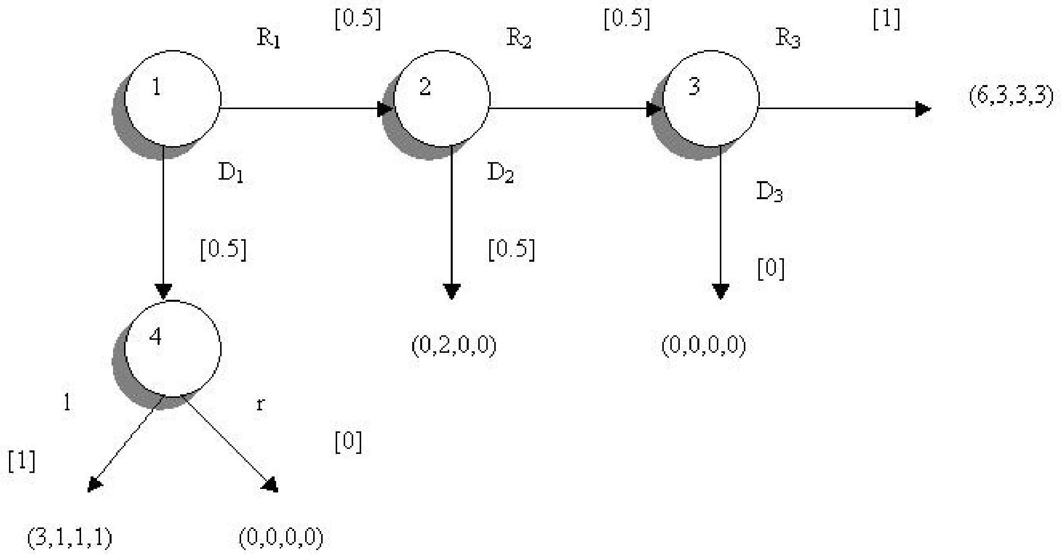

Figure 2: A self-confirming equilibrium with an "indifferent" player.

sufficient but not necessary for σ having a unitary full revelation outcome.

Proposition 1. Let σ be a self-confirming equilibrium with a unitary full revelation outcome. Then, if there is a profile σ^* that satisfies conditions $1-2$ of definition 1, σ^* is equivalent to a Nash equilibrium.

Proof/ Since σ has a unitary full revelation outcome, all best-response profiles to the beliefs specified in condition 1 of definition 1 generate a path to the same terminal node $z' \in Z$. Let a profile σ^* satisfy conditions $1-2$ of definition 1. By condition 2, this best-response profile σ^* is a self-confirming equilibrium supported by the following beliefs. Each agent in each population i has correct beliefs about opponents' play in $\overline{H}(\sigma^*)$.[19] For all information sets h_j $(j \neq i)$ outside $\overline{H}(\sigma^*)$, everybody expects the "old" distribution of actions, induced by σ.

Consider any mixed strategy profile σ^N that generates the distribution of actions, in each information set, specified by the above beliefs, for all players. We claim that profile σ^N is a Nash equilibrium profile. By condition 2, σ_i^* is a best response to σ_{-i}^N. Notice that, for each player i, the distribution of actions generated by mixed strategies σ_i^* and σ_i^N differs 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. *QED*

There is an interesting class of totally mixed self-confirming equilibria with a unitary full

[19] This is true because each information set on the equilibrium path is reached by all individuals of all populations.

revelation outcome. This class consists of self-confirming equilibria that cannot be supported by unitary beliefs. That is, for any population i, different strategies that are played with positive probability must be supported by different beliefs.

Definition 3. A self-confirming equilibrium σ is called "strictly heterogeneous" if for all i, s_k, $s_l \in supp(\sigma_i)$ implies that $u_i(s_k, \mu_i) \neq u_i(s_l, \mu_i)$ for any beliefs μ_i that satisfy the belief conditions of self-confirming equilibrium for both strategies (that is, they are correct in both $\overline{H}(s_k, \sigma_{-i})$ and $\overline{H}(s_l, \sigma_{-i})$).

Corollary. Let σ be a totally mixed, and strictly heterogeneous, self-confirming equilibrium. Then, if there is a σ^* that satisfies conditions $1-2$ of definition 1, σ^* is equivalent to a Nash equilibrium.

Proof/ Since $\overline{H}(\sigma) = H$, "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 and some strategies s_k and s_l in S_i, s_k and s_l would be best responses to σ_{-i}, hence both would be supported by the correct beliefs about σ_{-i}.) Thus σ has a unitary full revelation outcome. *QED*

3.2 Partial Information Revelation

For the self-confirming equilibrium profiles we consider in this section, it is possible that $\overline{H}(\sigma) \neq H$. We want to capture the possibility that the planner may not wish to announce all available information, but only part of it. For simplicity, here we also restrict our analysis to independent beliefs.[20] Note that, with independent beliefs, information about one population does not affect expectations about the behavior of other populations.

The planner may reveal the distribution of actions in a subset of the family of all information sets reached with positive probability under σ. Hence, if we denote by H^A any family of information sets, for which the planner reveals the distribution of actions given σ, the following must hold:

$$H^A \subseteq \overline{H}(\sigma) \tag{2}$$

We also make the simplifying assumption that the planner is constrained to reveal information for all or none of the information sets of each player. More formally:

$$H^A = \bigcup_{j \in J_{H^A} \subset J} H_j \tag{3}$$

J_{H^A} denotes the subset of J, for which, information is revealed according to the set H^A.

[20] We also assume that for any population i, each individual's beliefs μ_i are such that, for all $h \in \overline{H}(\sigma) \bigcap H_{-i}$, $p(h/\mu_i, s_i) > 0$ for some $s_i \in S_i$.

Definition 4. A set H^A which satisfies (2), (3) given a profile σ, is called an "information revelation set on σ".

Fix a self-confirming equilibrium σ supported by beliefs μ. Since the information revelation of the planner is truthful, the post-revelation beliefs of all individuals must be consistent with the distributions he announces.

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

1. $\mu^*_{i,s_i}[\{\pi_{-i} \in \Pi_{-i} : \pi_j(h_j) = \pi_j\widehat{(h_j/\sigma_j)}\}] = 1$ for all $h_j \in H^A$
2. $\mu^*_{i,s_i}\{\overline{\Pi}_j\} = \mu_{i,s_i}\{\overline{\Pi}_j\}$ for all $j \neq i$ such that $j \notin J_{H^A}$, and for all measurable $\overline{\Pi}_j \subseteq \Pi_j$

Condition (2) simply says that agents' beliefs remain the same for opponents, whose behavior is not revealed in H^A. For the other opponents, (1) states that the new beliefs agree with the information revealed. Agents simply believe the information announcement, and adjust their play accordingly, believing everything else is the same. This idea is captured by "transition beliefs". Note that our notation explicitly takes into account the fact that different subgroups, who correspond to different pure strategies, may have different transition beliefs, because they might have different initial beliefs. The notation μ_{i,s_i} describes the initial beliefs of the specific subgroup of population i that plays the pure strategy s_i.

Definition 6. Let σ be a self-confirming equilibrium supported by beliefs μ, H^A be an information revelation set on σ, and μ^* be the transition beliefs it generates. We say that σ^* is a profile supported by H^A if, for all i, there is a mapping $g_i : supp(\sigma_i) \to S_i$, such that, for all $s_i \in supp(\sigma_i)$, $g_i(s_i) \in argmax[u_i(., \mu^*_{i,s_i})]$, and such that:

$$\forall i, \forall s^*_i \in supp(\sigma^*_i), \sigma^*_i(s^*_i) = \sum_{\{s_i \in supp(\sigma_i) : g_i(s_i) = s^*_i\}} \sigma_i(s_i) \tag{4}$$

In other words, an information revelation set supports a profile σ^* if the transition beliefs it generates supports σ^*. Since each subgroup, corresponding to each $s_i \in supp(\sigma_i)$, could have different transition beliefs generated by H^A, not all of these subgroups need to have the same optimal strategy given these beliefs. Thus, the probability of each strategy s^*_i in the support of σ^*_i is determined by summing up all subgroups $s_i \in supp(\sigma_i)$ that find s^*_i optimal, given their transition beliefs. The function g_i simply selects one optimal strategy (given transition beliefs) for each subgroup of population i. So, it ensures that the mass of some subgroup is not counted twice.[21] Note that a given σ^* may be supported by multiple

[21] Note that we assume that all individuals in a given subgroup choose the same best-response, even if there exist multiple best responses.

transition beliefs. Now we may introduce our second basic definition:

Definition 7. 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, such that for all profiles σ^* supported by H^A, the following conditions hold:

1. σ^* is a self-confirming equilibrium, which, for each population i, is supported by the transition beliefs μ^*, for all $h_j \in \{H_{-i} - \overline{H}(s_i^*, \sigma_{-i}^*)\}$
2. $U^{PL}(\sigma^*) > U^{PL}(\sigma)$
3. $u_i(s_i^*, \mu_{i,s_i}^*) > u_i(s_i, \mu_{i,s_i}^*)$ for some i, some $s_i \in supp(\sigma_i)$, and some $s_i^* \in supp(\sigma_i^*)$

This definition says the following. Assume that the planner reveals aggregate behavior in the information revelation set H^A. All agents simply update their beliefs, assigning probability 1 to the planner's announcements, and they keep their old beliefs in opponents' information sets for which there is no revelation. Then, their best-responses to the new beliefs will form a self-confirming equilibrium profile. Again, this self-confirming equilibrium is compelling as the new steady state of the system. For, if this profile is played, agents update their beliefs only in information sets that belong to $\overline{H}(s_i^*, \sigma_{-i}^*)$. Hence, they want to continue choosing the same actions. In information sets outside $\overline{H}(s_i^*, \sigma_{-i}^*)$, agents maintain their old beliefs, since they do not have a reason to update it.

EXAMPLE 3. We shall illustrate the fact that, in some cases, the planner may prefer not to reveal all available information. Consider the four-player game presented in Figure 3. The pure strategies for all players are "pass" (the horizontal move) or "take" (the vertical move). Assume that the planner has the same payoffs as in Example 1.

The profile $\sigma = \{(0.5P_1, 0.5T_1); (0.5P_2, 0.5T_2); (0.2P_3, 0.8T_3); P_4\}$, illustrated in Figure 3, is a self-confirming equilibrium. 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 that 3's "pass" with probability 1. 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. For example, if the planner announces only the behavior of 3's, she can induce agent 1's and 2's to enter. However, if he were to announce also the behavior 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.[22]

[22] Note that this profile of best responses does not constitute a self-confirming equilibrium. This example illustrates our assumption that individuals do not predict others' behavior based on a priori knowledge of

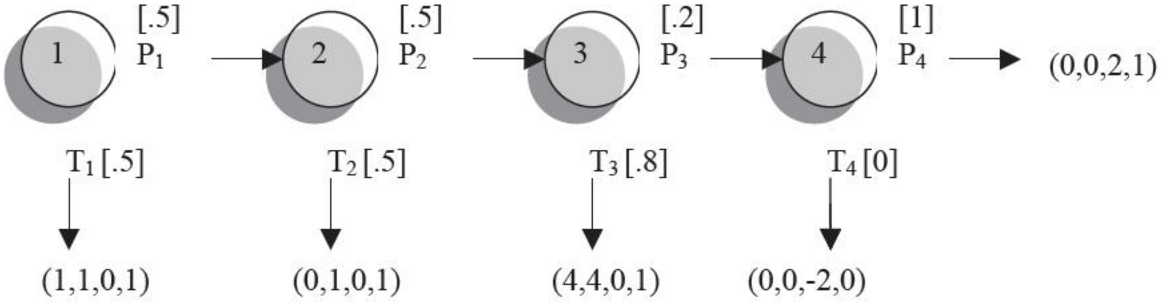

Figure 3: The "beneficial superstition game".

So, assume that the planner announces aggregate behavior at node 3. Which are the transition beliefs? Each individual keeps his old beliefs for all nodes except node 3, where his beliefs now agree with the revealed distribution of actions (notice the critical role of the assumption of independent beliefs). Best-responding to these new beliefs leads to profile $\sigma^* = \{(P_1; P_2; (0.2P_3, 0.8T_3); P_4\}$.[23] Clearly, $U^{PL}(\sigma^*) > U^{PL}(\sigma)$, since the outcome $(4, 4, 0, 1)$ is achieved with greater frequency under σ^*. This result would not be achieved if the planner revealed all available information. Note that information release in other strict subsets of J could also achieve this result, such as in the set $\{2, 3\}$.

3.3 Revelation-Robust Self-Confirming Equilibria

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

Definition 8. A self-confirming equilibrium σ is called "revelation-robust" if it is not full revelation-unstable or partial revelation-unstable.

A unitary self-confirming equilibrium is a mixed strategy profile σ such that for all i there exists beliefs μ_i such that for all $s_i \in supp(\sigma_i)$:

1. s_i maximizes $u_i(\cdot, \mu_i)$

2. $\mu_i[\{\pi_{-i} \in \Pi_{-i} : \pi_j(h_j) = \widehat{\pi_j(h_j/\sigma_j)}\}] = 1$ for all $j \neq i$ and for all $h_j \in \overline{H}(s_i, \sigma_{-i})$

their payoffs. If individual 1's and 2's knew that the distribution of actions is common knowledge, and also knew 3's payoffs, they could infer the change in 3's behavior, but they do not do so in our setting.

[23] Notice that half of individual 1's who "pass", according to this profile, 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, since this profile is a self-confirming equilibrium, all individual 1's who pass best-respond to the actual distribution of actions, generated by σ^*. Moreover, individual 3's believe that 1's and 2's pass with probability $\frac{1}{2}$. Yet, the 3's action is optimal given the true distribution of actions in nodes 1 and 2, generated by σ^*, and their beliefs about node 4. Therefore, when these players update their beliefs, as they observe moves on the path of play, this only reinforces their choices given their (fixed) beliefs for the nodes they never reach.

In other words, for such a self-confirming equilibrium, the same beliefs are used to rationalize all pure strategies of a given mixed strategy.

Proposition 2. All unitary self-confirming equilibria (thus, all Nash equilibria) are revelation-robust.

Proof/ Let σ be a unitary self-confirming equilibrium supported by beliefs μ. For all opponents j of i, if $h_j \in \overline{H}(\sigma)$, then there is some pure strategy $s_i \in supp(\sigma_i)$ such that $h_j \in \overline{H}(s_i, \sigma_{-i})$. Hence, for each population i, the initial beliefs μ_i must be correct for all $h_j \in \overline{H}(\sigma) \bigcap H_{-i}$. 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, condition 3 of definitions 1 or 7 cannot hold for any information revelation set. QED.

Proposition 3. Every finite game has a revelation robust self-confirming equilibrium.

Proof/ This simply follows from Nash's theorem and the proposition above.

4 Discussion and More Examples

There are important implicit assumptions behind our basic model, and their realism should be defended. Our agents ignore the fact that others will adjust their behavior after the planner's announcements. This is because, as we assume, they do not have strong beliefs about others' payoffs. No unrealistic degree of naivete of agents needs to be invoked. Moreover, it might seem easier for the planner to reveal agents' utility functions, rather than their actions. However, in our examples, agents often have moral incentives, which are not verifiable. 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. In summary, although some of our assumptions might seem too restrictive, they need not be so.

There are several possible applications of our framework, and information release in times of war and national emergency is a clear example of such an application. In particular, the government would not want to reveal complete public information about the actions of deserters who flee the country. Similarly, regarding crime prevention policy, in most countries, the State typically does not provide detailed data about the amount of people who escape capture. The "Beneficial Superstition Game" of Figure 3 models such a situation.[24]

[24] Assume that groups 1 and 2 correspond to producers who could invest (P_1, P_2) or not invest (T_1, T_2),

Furthermore, the fight against social discrimination might require implicit agreements with the media, to refrain from emphasizing information that reinforces social stereotypes. A typical example of this is the extensive media coverage of the cases where women perform jobs that are considered "men's jobs".

Other possible applications include policies against antisocial behavior, the management of macroeconomic expectations, and political campaigning. The media may deliberately avoid revealing full information about the behavior of antisocial groups, which often seek attention and enjoy such publicity.[25] Moreover, policies aiming to protect investor sentiments often selectively reveal information (the literature on the management of macroeconomic expectations is vast). Political campaigning is another case in point: for, there is much evidence that voters prefer to vote for the wining party. This has been supported by many studies, and it is called "the bandwagon effect". Because of this, political parties may selectively reveal polls, which show that the party is winning. This practice may manipulate the election results, so restrictions on polls during election campaigns have been imposed in many countries.[26]

In addition to the above examples, marketing and advertising also involves selective information release. Here the planner is the firm, which has special access to data regarding its sales, and selectively reveals it. Clearly, the publisher of a book will promptly announce that the book has sold a million copies, but he will not declare that only two copies have been sold. Our framework may also capture the manipulation of expectations regarding the extent to which institutions work properly (this manipulation aims at enhancing respect for institutions). For instance, the professional basketball leagues of the NBA and Euroleague have explicit policies that punish public statements against referees. Media commentators take this into account, and they might tend to conceal referees' mistakes and emphasize their correct decisions.

group 3 to criminals who could steal (P_3) or not (T_3), and group 4 to the police who could punish crime (T_4) or shirk (P_4). The State would like to reveal the fact that crime is not rampant, but not to reveal the inadequacy of the police, which would induce more crime.

[25] 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, since hooligans are often "proud" of their violent acts.

[26] See Michalos (1991) p. 410 and Morwitz and Pluzinski (1996) p. 53. The countries that have implemented, or consider implementing, a ban on political polling during election periods include Brazil, France, Canada and Germany.

5 Conclusions

In this paper, we examined how manipulation of aggregate information may affect equilibrium behavior, where equilibrium is thought as the steady state of a belief-based learning process. Starting from an environment of dynamic interaction with anonymous matching, we added the existence of a planner, and we examined how public information may be revealed in order to manipulate the behavior of individuals. We showed that the planner 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" is caused by the heterogeneity of beliefs across agents of the same population. Equilibria with unitary beliefs, such as Nash equilibria, are robust to such manipulation. Hence, revelation-robust equilibria always exist.

The model could be extended in several different directions. Firstly, 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 to allow agents to predict changes in others' actions when information is revealed. Another possible direction of new research would be to use the idea of a "planner", in order to endogenize other types of information release. For example, the analogy classes in Jehiel (2005) could be endogenized assuming that individuals observe public information after each period, and they do not remember what they observe in their personal interactions with others.

References

Pierpaolo Battigalli. Comportamento razionale ed equilibrio nei giochi e nelle situazioni sociali. Master's thesis, Bocconi University, 1987.

Joyce Berg, John Dickhaut, and Kevin McCabe. Trust, reciprocity, and social history. *Games and Economic Behavior*, 10(1):122–142, 1995.

Ignacio Esponda. Information feedback in first price auctions. *The RAND Journal of Economics*, 39(2):491–508, 2008.

Drew Fudenberg and David K Levine. Self-confirming equilibrium. *Econometrica*, 61(3): 523–45, 1993a.

Drew Fudenberg and David K Levine. Steady-state learning and nash equilibrium. *Econometrica*, 61(3):547–573, 1993b.

Drew Fudenberg and David K. Levine. *The theory of Learning in games.* MIT Press, 1998.

Glenn Harrison and Kevin McCabe. Expectations and fairness in a simple bargaining experiment. *International Journal of Game Theory*, 25(3):303–327, 1996.

Philippe Jehiel. Manipulative auction design. *Theoretical economics*, 6:185–217, 2011.

Phillipe Jehiel. Analogy-based expectation equilibrium. *Journal of Economic Theory*, 123 (2):81–104, 2005.

Ehud Kalai and Ehud Lehrer. Rational learning leads to nash equilibrium. *Econometrica*, 61(5):1019–45, 1993.

Alex Michalos. Ethical considerations regarding public opinion polling during election campaigns. *Journal of Business Ethics*, 10:403–422, 1991.

Vicky G. Morwitz and Carol Pluzinski. Do polls reflect opinions or do opinions reflect polls? the impact of political polling on voters' expectations, preferences, and behavior. *Journal of Consumer Research*, 23:53–67, 1996.

Andreas Ortmann, John Fitzgerald, and Carl Boeing. Trust, reciprocity, and social history: A re-examination. *Experimental Economics*, 3:81–100, 2000.

Alvin Roth and Francoise Schoumaker. Expectations and reputations in bargaining: An experimental study. *The American Economic Review*, 73(3):362–372, 1983.