VIABLE NASH EQUILIBRIA: FORMATION AND SUSTAINABILITY

EHUD KALAI

In memory of John Nash

ABSTRACT. To be viable, a Nash equilibrium should deter the defection of groups of players. For a newly proposed equilibrium to be viable, it should deter defections but also have a reasonable likelihood of formation. To assess the viability of equilibria, we study a *defection-deterrence index* D that describes the size of defection groups that an equilibrium is sure to deter, and a dual *index of formation difficulty*, $F \ \ (= n - D)$, that describes the critical size of groups that guarantee equilibrium formation. These viability indices are compatible with observations in behavioral economics and with functioning social systems, and offer a tool for assessing the performance of proposed equilibria.

JEL Classification Codes: C0, C7, D5, D9.

1. OVERVIEW

For several decades, behavioral economists have been questioning the viability of Nash equilibrium. The opinions are mixed since some Nash equilibria consistently fail to perform in laboratory tests and in field experiments, while

Date: March 9, 2018, this version August 13, 2019.
Key words and phrases. Normal form games, Nash equilibrium, Stability, Fault tolerance, Behavioral Economics.
The author thanks the following people for helpful conversations: Nemanja Antic, Sunil Chopra, Vince Crawford, Kfir Eliaz, Drew Fudenberg, Ronen Gradwohl, Yingni Guo, Adam Kalai, Fern Kalai, Martin Lariviere, Eric Maskin, Rosemarie Nagel, Andy Postlewaite, Larry Samuelson, David Schmeidler, James Schummer, Eran Shmaya, Joel Sobel and Peyton Young; and seminar participants in the universities of the Basque Country, Oxford, Tel Aviv, Yale, Stanford, Berkeley, and Stony Brook. This paper replaces earlier versions of similar titles.

other Nash equilibria play critically successful roles in functioning social systems.[1]

For a better understanding of discrepancies between the theory and observed play, this paper makes a subtle but important departure from the standard view of Nash's theory: as in standard Nash models, the players in this paper are all rational and consider playing an equilibrium of the strategic game under consideration; but unlike in the standard Nash model, the players are concerned that their opponents may fail to play, or may defect from the equilibrium strategies. Relaxing the assumption about opponents' behavior leads to a richer output from a best-response analysis – an output that is useful for assessing the viability of the equilibrium.

Concerns about players' defections from a Nash equilibrium may be due to a large variety of reasons. For one, players may be concerned about defections that are not addressed by the standard Nash analysis, such as rational defections by coalitions of players. More broadly, players may be concerned that the game description has left out critical aspects of the environment. For example, opponents may be threatened or bribed by agents outside the game, defections may be motivated by options not modelled in the game, and even miscalculations and mental instability may lead to defections from the equilibrium.

Precise Bayesian modelling of all such concerns is impractical, and it is hard to predict how a player would respond to such concerns. But there are important exceptions, namely, cases in which the player's equilibrium strategy dominates all possible responses to her concerns. In such cases we may view her equilibrium strategy as more viable.

Our discussion of equilibrium viability distinguishes between two types of equilibria: equilibria that are focal in the minds of the players (e.g., those due to past play of similar games in the environment), and newly proposed equilibria of new games. Both types of equilibria have to pass a *sustainability test*: once they are focal in the minds of the players, they must survive the

[1]The literature on this subject is too large to survey here. Some key examples include Smith (1982), Erev and Roth (1998), Crawford (1998), Kahneman and Tversky (2000), Goeree and Holt (2001), and Camerer (2003).

period of examination and deliberation prior to the play of the game. But in addition, newly proposed equilibria have to also pass a *formation test*: they must have a realistic chance of becoming focal.

The two theoretical indices discussed in this paper were chosen to assess the sustainability and the likelihood of formation of Nash equilibria of n-person strategic games. Examples discussed in the paper illustrate that equilibria observed in functioning social systems are highly viable according to these theoretical indices, while low-viability equilibria are not observed in social systems and fail in laboratory experiments.

To assess sustainability we study a *defection-deterrence index* $D(\pi)$ that describes, among other things, the largest size of groups of players who are individually sure to (weakly) lose by defecting from an equilibrium π. A high $D(\pi)$ value correlates positively with the ability to sustain an equilibrium π. To assess the likelihood of formation, we study a *formation-difficulty index* $F(\pi)$ that describes, among other things, the critical number of π-committed players needed to assure the remaining players that playing π will result in gains to their individual payoffs. A low value for $F(\pi)$ correlates positively with the likelihood of successful formation of π. Accordingly, one may anticipate that (i) focal equilibria with high D-values, and (ii) newly proposed equilibria with high D-values combined with low F-values, are likely to be viable.

The viability indices D and F are dual and, in particular, $D(\pi) + F(\pi) = n$ for any equilibrium π of an n-person strategic game. But in different applications either D or F may be the more natural primitive. This duality means that statements about one of them entail meaningful dual statements about the other, as we can see below.

The defection-deterrence levels $D(\pi) = 0, 1, ..., n$ present a progression in the players' thinking about Nash equilibria: at the extremes, $D(\pi) = 0$ for the strategy profiles π that are not Nash equilibria, and $D(\pi) = n$ for the profiles π that consist of (weakly) dominant strategies. The dual views are: $D(\pi) = 0$ means that π cannot become a focal point equilibrium, and $D(\pi) = n$ means that π should become a focal point equilibrium despite any concerns about about opponent defections.

The intermediary levels, $D(\pi) = 1, 2, ..., n-1$, classify all remaining Nash equilibria in progressively increasing levels of defection deterrence. For example, $D(\pi) = 1$ implies that some π_i is a strictly inferior play in profiles in which only one opponent defects from π; and the dual view is that the commitment of $n-1$ players is needed to assure the formation of π. On the other hand $D(\pi) = n-1$ means that every π_i is almost a dominant strategy, i.e., it is an optimal play in all profiles in which at least one opponent plays π. And the dual view is that the commitment of one player is sufficient to make all the players jump on the bandwagon of playing π.

1.1. **Illustrative examples.** The first two examples contrast a high D-value equilibrium with one of low D-value.

Example 1. ***A (language) matching game.***

Simultaneously, each of $200M$ *players selects one option (say, a language) from a set of possible choices. For any choice* X*, the payoff of a player who chooses* X *equals the number of opponents she matches, i.e., the number of other players who also choose* X*.*

Consider the profile eE in which every player chooses E. It is easy to see that eE is a Nash equilibrium because it "deters single-player defections": a unilateral defection to a different choice would lower the payoff of a defecting player. But going beyond the Nash condition, eE deters defections of groups of players. For example, any player who is one of $1M$ potential defectors would compute that by staying with E, she would match at least $199M$; whereas by defecting from E, she would match at most $1M-1$. So even in the presence of the $1M$ potential defectors, defection is certain to lower her payoff.

In the terminology of this paper, eE *deters defection* of groups that consist of $1M$ players. As the reader can easily verify, the defection deterrence of eE is actually true for groups of sizes up to $100M$ potential defectors, but not for groups of size $100M+1$. For this reason we say that the defection-deterrence index of eE, $D(eE)$, is $100M$.

In contrast to the large defection-deterrence value $D = 100M$ above, the defection deterrence of the next equilibrium is only $D = 1$, barely enough to be classified as a Nash equilibrium.

Example 2. ***A confession game*** *(a stag-hunt game that only sounds like a prisoners' dilemma).*[2]

Twelve partners in a crime are interogated by the police, simultaneously, but in separate rooms. If none of the suspects confesses, everyone will be released with no penalty. However, if one or more confess, then every suspect will be sentenced to ten years in jail, except for the confessors, who will be sentenced to only three years instead of ten.

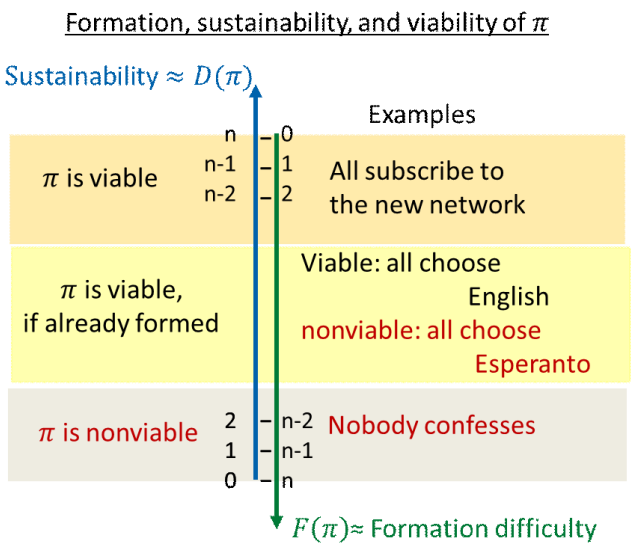
The cooperative profile nC in which nobody confesses is a Nash equilibrium, because it deters with certainty single-player defections: a unilateral confession worsens a player's outcome from no jail time to three years in jail. However, nC fails to deter (with certainty) two-player defections. For example, if a fellow potential defector confesses, then by confessing yourself, you *improve* your outcome from a ten-year sentence to a three-year sentence. For this reason we say that nC deters single-player defections but does not deter two-player defections and that the defection-deterrence index of nC is $D(nC) = 1$.

Next, we consider newly proposed equilibria and contrast one with a high F-value with one with a low F-value. For the high F-value, consider the language-choice game above, but with an equilibrium in which everybody chooses Esperanto, $eEsp$. Just like the eE equilibrium in the game, $D(eEsp) = 100M$, and by the duality of the two indices, $F(eEsp) = 200M - D(eEsp) = 100M$. So while $eEsp$ would be highly sustainable in populations where it is already played, if it were introduced as an option in a new population (for example, where people speak English), it would require a critical mass of $100M$ committed Esperanto choosers to make Esperanto a dominant choice for the rest of the players - a highly unlikely event.

In contrast to the $eEsp$ equilibrium above, the next game illustrates an equilibrium that would be considered viable, even if proposed in a new population that has never played it. In particular, it has a high defection-deterrence value D accompanied by a *low* formation-difficulty value, F.

Example 3. ***A new communication network*** *is offered for subscription in a population of* $200M$ *potential users, at an individual annual cost of* \$9.99.

[2]The author thanks Adam Kalai for suggesting this example.



The payoff of a subscriber is $k - 9.99$*, where* k *is the number of opponents who subscribe. The payoff of a nonsubscriber is zero.*

It is easy to see that if nine players subscribe, subscribing becomes a dominant strategy for each of the remaining players. Moreover, since 9 is the minimal integer with this property, we say that the formation-difficulty value of the everybody-subscribes equilibrium, $eSub$, is $F(eSub) = 9$. From the dual view of the two indices we infer that $D(eSub) = 200M - 9 = 199,999,991$. We conclude that the equilibrium $eSub$ should seem viable to the network provider, because $eSub$ is easy to form and to sustain.

The interrelationships of equilibrium formation, sustainability, and viability are summarized in the chart titled "Formation, Sustainability, and Viability." It should be clear that the values of $D(\pi)$ and $F(\pi)$ are objectively determined by the game under consideration. However, the viability assessment is subjective and depends on the environment in which the game is played. For example, $eEsp$ in a language-choice game played tomorrow morning would be considered viable in an environment in which Esperanto is regularly used, but would be considered nonviable in environments where it is never used.

1.2. **Related game-theory notions.** The stability of Nash equilibria is addressed in the literature on equilibrium refinements.[3] Equilibrium refinements are rigid, since they simply disallow behavior that is inconsistent with more demanding levels of rationality. But the viability of an equilibrium is often subjective, and it may depend on the broader context in which the game is played and on the role of the equilibrium in this broader context. Since they are multivalued, and not just binary, the viability indices allow for additional flexibility that goes beyond the existing equilibrium refinements. This paper presents examples of equilibria that pass most refinement tests, but are considered nonviable in the context in which they are played.

Formally, the D index is defined to measure the level of subgame perfection (a la Kalai and Neme 1992) in a two stage decomposition of the game. In the first stage the players publicly announce how they intend to play the game; and in the second stage, with complete knowledge of the announced intentions, they actually play the game. Announcing π in the first stage and playing the second stage with no revisions is k subgame perfect iff $k < D(\pi)$.

In applications, the D index has its roots in the literature on distributive computing in computer science, see for example Ben-Or et al. (1988), with later adaptations to the literature on faulty- player implementation in algorithmic game theory. The F index is new to this paper.

As we discuss in the section on faulty-player implementation (section 7), the D index is closely related to the notion a *k-fault-tolerant equilibrium* used by Eliaz (2002), and it coincides with the notion of *resilience* used by Abraham et al. (2006) in their applications to secret sharing games and distributive computing.[4] However, while defections in this paper can describe the behavior of faulty players, the analysis presented here deals with a broader class of rational players who are presented with incentives to defect. For this reason, the applicability of this paper's findings is significantly broader.

[3]A general survey of stability in game theory would be too large to discuss here. Some key examples include Aumann (1959), Selten (1975), Myerson (1978), Basu and Weibull (1991), Kreps and Wilson (1982), Kalai and Samet (1984), Kohlberg and Mertens (1986), Bernheim, Peleg, and Whinston (1987), Young (1993), Kandori, Mailath, and Rob (1993), Moreno and Wooders (1996), and Myerson and Weibull (2015).

[4]See also Gradwohl and Reingold (2014) for implementation in large games and for additional references from computer science.

The fundamental role of Nash equilibria in game theory is the reason that this paper focuses on the viability of Nash equilibria, and not on the viability of Nash equilibrium refinements or other solution concepts. As a result, the departure of the viability indices from the best-response behavior modeled by Nash is minimal. Positive consequences are that the viability indices studied here (1) remain applicable in a broad range of applications, and (2) require computations that do not exceed the best-response reasoning done by most players.

At the same time, minimal departure from Nash equilibrium leads us to relatively crude viability indices – ones that can be too crude for some applications. In the concluding part of the paper we discuss such examples and suggest more refined indices and associated theories of equilibrium formation. However, different such refinements may require different and inconsistent modifications to the notion of Nash equilibrium (as is the case in the Nash equilibrium refinements literature). The study of such refinements is left for future research.

1.3. **The underlying game.** This paper focuses on a strategic n-person game $\Gamma = (N, A = \times_{i\in N} A_i, u = (u_i)_{i\in N})$. N is a finite set of n *players*. Subsets of N are called *groups* or *coalitions* of players and N is the grand coalition. Elements of A_i are the *strategies of player* i, and strategy *profiles*, or *profiles* for short, are functions of the form $\alpha = (\alpha_i)_{i\in N} \in A \equiv \times_i A_i$ which assign to every player i an element $\alpha_i \in A_i$. We assume that Γ has no *dummy players*, i.e., the number of strategies available to every player i, $|A_i| \geq 2$.[5]

For a coalition D and a profile α, define $\alpha_{_D} = (\alpha_j)_{j\in D}$. The profile in which the players in D defect from α to a profile β is defined by: $(\alpha_{N\setminus D} : \beta_{_D})_j = \alpha_j$ for $j \notin D$, and $(\alpha_{N\setminus D} : \beta_{_D})_j = \beta_j$ for $j \in D$. A strategy α_i of player i is a best response to a profile β, if $u_i(\beta_{N\setminus i} : \alpha_i) \geq u_i(\beta_{N\setminus i} : \chi_i)$ for any strategy χ_i of player i.[6]. A profile α is a best response to a profile β if every α_i is a best response to β. A strategy α_i is dominant if α_i is a best response to any profile. The notions of best response and domination are weak, in the sense that they are defined by weak inequalities.

[5]|S| denotes the number of elements in a set S.

[6]When it is clear from the content, we sometime omit the brackets and replace $\{i\}$ by i .

Throughout the rest of the paper, $\pi = (\pi_i)_{i \in N}$ denotes one *arbitrary fixed focal profile*. The implicit assumption that π is focal explains why we may think of π as an equilibrium, in cases in which best-response functions involve indifference. Similarly, the definitions of the indices below are motivated by this implicit assumption.

Given the fixed profile π, for any profile β define the set of π-defectors at β, or *defectors* for short, by $D_\pi(\beta) = D(\beta) = \{j \in N : \beta_j \neq \pi_j\}$; and the set of π-loyalists at β, or *loyalists* for short, by $L_\pi(\beta) = L(\beta) = N \backslash D(\beta) = \{i \in N : \beta_i = \pi_i\}$.

For a profile χ and a coalition C, it is useful to consider the subgame induced on the members of C, when the players outside C are fixed at their χ strategies.

Definition 1. *The* subgame played by C under χ, Γ_C^χ *, is described as follows: the set of players is C; the strategy set of every $i \in C$, A_i, is the same as in Γ; and the payoff of every player $i \in C$ at any profile α_C is the same as her payoff in Γ at the concatenated profile $(\chi_{N \backslash C} : \alpha_C)$.*

Below is an example of an asymmetric game, used to illustrate the new concepts defined in the next section.

Example 4. ***Mismatch the opposition (MMOP):*** *Simultaneously, each of three Democrats and five Republicans selects one of two choices, E or F. The payoff of a player is the number of opposite-party players whose choice she mismatches. We will consider the completely* divisive equilibrium*, $cDIV$, in which all the Democrats choose F and all the Republican choose E.*

2. Viability Indices

2.1. **Basic definitions and properties.** The confidence that player i has in the optimality of her chosen strategy π_i has a natural turning value, $PA_i(\pi)$: when the number of potential defectors, d, is smaller than $PA_i(\pi)$, staying with her π_i is a dominant strategy; but for any d at or above $PA_i(\pi)$, there are defections that make π_i an inferior strategy. Put differently, $PA_i(\pi)$ is the minimum number of defectors that can push player i away from playing π_i.

Definition 2. *The push-away value of player i, $PA_i(\pi)$, is the minimal number of players d for which the following is true: π_i is a (strictly) suboptimal*

response to some d-defector profile $\delta_{i,d}$. Define $PA_i(\pi)$ to be n if such a number d does not exist.

Notice that the minimality condition in the definition above implies that player i is not one of the d defectors in the profile $\delta_{i,d}$, because whether or not she herself defects does not affect the optimality of her response. Also $PA_i(\pi) = 0$ iff π_i is an inferior response to π, and $PA_i(\pi) = n$ iff π_i is a dominant strategy.

Definition 3. *A player v is* most vulnerable *if $PA_v(\pi) = \min_i PA_i(\pi)$.*

In the completely divisive equilibrium, $cDIV$, in which the three Democrats choose F and the five Republicans choose E, each Democrat can be pushed away from her F strategy iff three or more Republicans choose F, so $PA_D(cDIV) = 3$ for each Democrat D. For any Republican, $PA_R(cDIV) = 2$, since a Republican can be pushed away from his E strategy iff two or more Democrats choose E. So in this game the Republicans are the most vulnerable players. As defined next, the defection deterrence of $cDIV$ is the push-away value of the most vulnerable players, i.e., $D(cDIV) = \min_i PA_i(cDIV) = 2$.

Definition 4. *The* defection deterrence *of π, $D(\pi) \equiv \min_i PA_i(\pi)$.*

Under the definition above, we may view $D(\pi)$ as the minimal player-power needed to undo π as an equilibrium. The next definition describes the player-power needed to start π as an equilibrium.

We consider the same structure as above, but replace the number of defectors in a profile by the number of loyalists, i.e., $l = n - d$, to get a dual view of the turning value. And while this dual view is "mathematically parallel," it offers a related useful interpretation. More specifically, the D index is useful for someone who thinks of the equilibrium as ongoing, and is concerned with the incentives of players to defect. On the other hand, the dual view is useful for someone who thinks about the process of forming a new equilibrium, and is concerned with the incentives of newcomers to join.

Definition 5. *The* join-resistance *value of player i, $JR_i(\pi)$, is the maximal integer l that satisfies the following: π_i is an inferior response to some l-loyalist profile $\delta_{i,l}$. Define $JR_i(\pi)$ to be 0 if such a number l does not exist.*

The maximality condition in the definition of $JR_i(\pi)$ implies that player i is always counted as one of the l loyalists in the profile $\delta_{i,l}$ that player i would prefer not to join. Notice also that $JR_i(\pi) = 0$ iff π_i is a dominant strategy, and that $JR_i(\pi) = n$ iff π_i is an inferior response to π. Also, the definition implies the complementarity condition: $JR_i(\pi) = n - PA_i(\pi)$.

Definition 6. *The* formation difficulty *of* π, $F(\pi) \equiv \max_i JR_i(\pi)$.

As should be clear from the definitions above, the duality of PA_i and JR_i implies the duality of the two indices, and that the players who are easiest to push away from the equilibrium are the ones who are most resistant to joining the equilibrium.

Proposition 1. ***Duality****:* $F(\pi) = n - D(\pi)$, *and*
$\arg\max_i JR_i(\pi) = \arg\min_i PA_i(\pi)$.

Proof. For any player i and any profile δ, consider the condition A_d: δ has exactly d defectors that do not include player i, and π_i is an inferior response to δ. Now compare condition A_d with condition B_l: player i is one of exactly l loyalists, and π_i is an inferior response to δ. It is easy to see the equivalence of the two conditions, any profile δ satisfies A_d iff it satisfies B_{n-d} .

The assertions in the propositions are a direct consequence of the equivalence above. □

The definitions of D and F above describe *individual assurance levels* for players who contemplate playing their π_is: D assures every player that continuing to play her π_i is optimal, as long as the number of defectors is smaller than D; and F assures every player that adopting the play of her π_i is optimal, as long as at least F players adopt their π strategies. Next, we discuss coalitional interpretations of D and F that could serve as alternative definitions of the two indices.

Definition 7. *Coalitional deterrence:*

(1) The profile π *deters defection of a coalition* D if π_i is a dominant strategy in the game Γ_D^π for every player $i \in D$.
(2) The profile π *deters* d *defectors*, if it deters the defection of any coalition D of d players.

Proposition 2. ***Defection Deterrence****: The profile π deters d defectors iff $d \leq D(\pi)$.*

Proof. Assume the left-hand side (lhs) of the proposition. To show that $d \leq D(\pi)$, it suffices to show that $d \leq PA_i(\pi)$ for every player i. Consider any profile δ with a set of defectors D such that $i \in D$ and $|D| \leq d$. The lhs implies that π_i is a best response to δ. This shows that $d-1$ players cannot push player i away from π, which implies that $d-1 < PA_i(\pi)$, or that $d \leq PA_i(\pi)$.

Next, assume the rhs of the statement and consider any set of players D with $|D| \leq d$. For any player $i \in D$, the rhs implies that π_i is a best response to the profiles δ in which the number of defectors does not exceed d. So π_i is a dominant strategy for any game Γ_D^{χ} in which $i \in D$ and $|D| \leq d$. □

Definition 8. *Coalitional formation:*

(1) *A* coalition L forms π *if π_i is a dominant strategy in the game $\Gamma^{\pi}_{N\backslash L}$ for every $i \in N\backslash L$.*
(2) *For any l, l* players form π*, if any l-player coalition forms π.*

Lemma 1. *The profile π deters d defectors iff $n-d$ form π.*

Proof. As is clear from the definitions of the two concepts, any coalition L forms π iff π deters the defection of the coalition $N\backslash L$. This implies that any l-player coalition forms π iff π deters defection of any $(n-l)$-player coalition. □

Proposition 3. ***Equilibrium Formation****: l players form π iff $l \geq F(\pi)$.*

Proof. $l \geq F(\pi)$ iff $n-l \leq n-F(\pi)$, i.e., iff $n-l \leq D(\pi)$. By Lemma 1, the last condition is equivalent to l players forming π. □

The next proposition presents a sustainability progression in Nash equilibria, starting with profiles that are not Nash equilibrium, moving through Nash equilibria in increasing levels of defection deterrence, and ending with profiles that consist of (weakly) dominant strategies. The Ride sharing game discussed in the sequel shows that all the levels of defection deterrence, 0,1,...,n, are obtained in simple games that require only reasonable levels of computational ability.

Proposition 4. ***Classification of Nash equilibria****: π is not a Nash equilibrium iff $D(\pi) = 0$ (alt. $F(\pi) = n$); π is a dominant-strategy equilibrium iff $D(\pi) = n$ (alt. $F(\pi) = 0$); and the intermediary values $D(\pi) = 1, ..., n-1$ describe all other Nash equilibria at increasing levels of defection deterrence.*

Proof. From the definition of D, π is not a Nash equilibrium iff $D(\pi) = 0$, and thus π is a Nash equilibrium iff $D(\pi) > 0$. Also from the definition of D, π is a dominant-strategy equilibrium iff $D(\pi) = n$. This means that the remaining intermediary values must be assigned to all non-dominant-strategy Nash equilibria. □

The next proposition states that in all subgames played by coalitions above some critical minimal size, π is a Nash equilibrium no matter what the players outside the coalition play. Equivalently, no matter what strategies are played by groups of defectors of limited size, π is a Nash equilibrium for the remaining players. As discussed later in the paper, this defection-tolerance property is important in problems of implementation.

Definition 9. *For any coalition C, π is a* uniform Nash equilibrium of C*, if π is a Nash equilibrium of the game Γ_C^χ for every profile χ.*

Proposition 5. ***Nash defection tolerance****: π is a uniform Nash equilibrium of any coalition C with $|C| \geq F(\pi) + 1$.*

Proof. Let C be a coalition with $|C| \geq F(\pi) + 1$. We first show that for any profile χ, $F_{\Gamma_C^\chi}(\pi) \leq |C| - 1$. Since $|C| - 1 \geq F(\pi)$, for every $i \in C$, $C \backslash i$ forms π in Γ. But due to the use of domination in the definition of formation, this implies that every such $C \backslash i$ forms π in the game Γ_C^χ , which in turn implies that $F_{\Gamma_C^\chi}(\pi) \leq |C| - 1$. With $F_{\Gamma_C^\chi}(\pi) < |C|$, we conclude from the equilibrium classification of $F_{\Gamma_C^\chi}$ that π is a Nash equilibrium of the game Γ_C^χ . □

The proposition above motivates the following definition:

Definition 10. *The* Nash critical mass *of π, $NCM(\pi) \equiv F(\pi) + 1$.*

For example, in the divisive equilibrium $cDIV$ of the MMOP game (in which the theree Democrats choose F and the five Republicans choose E), it is a Nash equilibrium for any $7 (= F(cDIV) + 1)$ players to "follow the party

line," even if they are not sure who the eighth excluded player is and what she may choose. But in a 6-player game played by all the Republicans and one Democrat following the party line is not a Nash equilibrium. For example, the Republicans are not best responding if the two excluded players are Democrats who choose E.

2.2. **Elaboration on the definitions.**

The process of formation and sustainability. The equilibrium π is a strategy profile of the game Γ, which is to be played once. One may decompose the play of Γ into stages, and derive the index D formally from earlier notions of subgame perfection.[7] But for more intuitive explanations, we restrict ourselves to the direct definitions given in the previous section.

The formation of π is the process in which individual or group thinking leads to the perception of π as a focal point in the minds of the players (see Schelling 1960). The term sustainability refers to the ability of such a focal point π to survive through the examination and deliberation period that leads to the play of the game.[8]

On the domination in the definition of deterrence. The viability indices are applicable in a large variety of situations that involve defections and equilibrium formation, such as those described below.

Rational coordinated defections take place when a group of players, who know each other, coordinates a switch to different individual strategies to benefit the members of the group. One example would be a group of senators in the US Senate who jointly defect to a vote that contradicts their party's policy, but which changes the outcome of the vote to benefit their respective constituents.

---

[7]For example, in stage one every player i declares the strategy (pure or mixed) ς_i that she *intends* to play. In stage two, with full knowledge of the entire profile of intentions ς, she chooses her *actual* strategy $\sigma_i = \sigma_i(\varsigma)$. The payoff of the two-stage game is $u_i(\varsigma, \sigma) = u_i(\sigma)$, if her $\varsigma_i = \pi_i$; otherwise $u_i(\varsigma, \sigma) = w$, where w is the lowest possible payoff of any player in Γ. Consider the strategy profile $\overline{\pi} = (\pi, \pi^e)$ in which $\varsigma = \pi$, and $\pi_i^e(\varsigma) = \varsigma_i$. It follows directly from the definitions that π is k-perfect (see Kalai and Neme 1992) iff $k < D(\pi)$.

[8]Notice that high sustainability is different from being ex-post Nash (see Kalai (2004)), which requires that no player regrets the choice of her strategy after seeing the realized outcome of the game (including the realized values of mixed strategies).

Anonymous defections take place when individuals who may not know each other defect, trusting that others like themselves will do the same. An example would be a person who believes in the principles of the "pink movement" who decides to attend an illegal demonstration, trusting that others with similar values will attend.

Other defections from an equilibrium may result from *threats* or *bribes* from outsiders who are not formal players of the game. And defections that may be irrational can simply result from *miscalculations* on the part of strategy choosers.

The definition of the viability indices targets the mindset of a rational player who thinks that her opponents may defect for a variety of reasons, such as the ones above. Even if she assumes that the number of potential defectors is bounded, she may still be unsure about the identity of the potential defectors, what motivates them, and to what strategies they may defect. A player who faces such uncertainty is likely to have more reasons to defect than a Nash player who is certain about the strategies of the opponents. Moreover, using a Bayesian method to describe all her possible uncertainties may be too overwhelming, even to a rational player.

The use of the concept of domination in the definition of defection deterrence means that a player who is deterred from defection has incentives to stay with the equilibrium despite this broader set of concerns about her opponents.[9] This is an advantage over "Bayesian rationality," but it also means that the defection-deterrence index errs on the conservative side: Having a high defection-deterrence value is a reliable strong support to the sustainability of an equilibrium, but a low defection-deterrence value may be a result of unjustfiable fears.[10]

From a technical viewpoint, the domination property also gives rise to the monotonicity properties used in the definition of the indices.

[9]The standard exception to the use of dominant strategies applies here too; for example, a player may have binding agreements with other potential defectors.

[10]The last section of the paper discusses applications that require less demanding notions of deterrence.

Staying close to Nash. Key modeling choices in the definition of the indices are conservative, keeping philosophical discrepancies with Nash equilibrium to a minimum. The following are some important similarities:

(1) The domination condition that prevents coalitional defections is a minimal expansion of Nash's condition that prevents individual defections. As stated above, the use of domination in the definition of deterrence means that an equilibrium of a high deterrence value, beyond the level 1 required by Nash, must be a truly exceptional Nash equilibrium.

(2) A Nash equilibrium is binary (i.e., a profile either is or is not an equilibrium) and the viability indices take on a finite number of integer values $0, 1, ..., n$. Just like Nash equilibrium, the indices do not take into consideration cardinal levels of losses from defections, but consider only whether a loss occurs.

(3) Nash equilibrium is anonymous in that it only verifies that every player best responds to the others, regardless of players identities and their positions in the game. Similarly, the viability indices are anonymous in that they only count the number of players who best respond to others, disregarding players' identities and their positions in the game.

(4) Like Nash equilibrium, best responses considered by the indices are limited to immediate gains, ignoring longer-term possible gains due to consequential follow-up defections.[11]

The concluding part of the paper offers a brief discussion of alternative indices needed for more refined applications, although comprehensive discussion of these issues is left for future research.

Subjective uses of the indices. The indices $D_\Gamma(\pi)$ and $F_\Gamma(\pi)$ are objectively determined by the game Γ and the equilibrium π. But since they assume many values, they offer more flexibility in comparison to to standard binary equilibrium refinements. However, this paper does not specify bounds on the values of the indices that make an equilibrium viable. Viability is a subjective determination that is left to the judgment of the user. This is similar to the use of the standard-deviation index in probability theory. The standard deviation

[11]See Chwe (1994)

is objectively computed from the distribution of a random variable, but it is only used to aid a decision maker who determines subjectively whether the random variable is too risky.

3. More examples and illustrations

3.1. **Highly sustainable examples.** In the eE equilibrium of the Language Matching game in section 1.1, it is easy to see that pushing any player i away from the choice of E requires a minimum of $100M$ defectors, so, $PA_i(eE) = 100M$ for every i, and $D(eE) = \min_i PA_i(eE) = 100M$.

Many conventions and social arrangements in large populations rely on highly-sustainable Nash equilibria, as in the paragraph above. Some examples of such social arrangement are: everybody choosing the same language, Spanish, Mandarin, or any one of many other languages; everybody using dollars, euros, or any one of many other currencies; everybody obeying traffic signals; everybody using the same communication software and/or the same hardware; and the choice of market locations done by sellers and buyers as places to trade. But also dress codes, food culture, and many social mores often rely on such equilibria.

3.2. **Minimally sustainable examples.** For the no-confession equilibrium nC of the confession game in section 1.1, $D(nC) = 1$. That $D(nC) \geq 1$ is easily seen by observing that nC is a Nash equilibrium. But for every player i, if one player $j \neq i$ confesses (defects), then player i's best response is to confess too. This means nC fails to deter two defectors, and thus, $D(nC) = 1$.

We elaborate further on the low sustainability of the equilibrium nC in the next section. But before doing so, we present three additional examples of familiar games with Nash equilibria that also have the minimal sustainability level $D = 1$.

In the **beauty-contest game**, each of n judges submits a real number $r_i \in [0, 100]$. The judges whose submitted number is closest to two-thirds of the average submitted number, $\frac{2}{3}\sum_{i=1}^{n} r_i/n$, are each paid one, the others are paid zero. The Nash equilibrium in which all the judges submit the number 0, $e0$, is rarely observed in lab and field experiments (see Nagel (1995)). The reader can easily verify that its defection-deterrence index is only $D(e0) = 1$.

Mixed-strategies equilibria are often minimally sustainable, as can be seen again in the Language Matching game. Consider the equilibrium in which every one of the $200M$ players chooses English or Spanish with equal probability. If one player changes her choice probabilities to 2/3 on choosing English and 1/3 on choosing Spanish, then every player's best response is to choose English with probability one. So the mixed-strategy equilibrium has the minimal defection-deterrence level $D(eE) = 1$ (see O'Neill(1987) and follow-up papers for empirical studies of mixed-strategy equilibria).

Having everyone show up for work on **a simple production line** also has the low deterrence index $D = 1$. Consider a group of n workers who stand to receive a bonus if and only if they all report to work. A player's payoff is positive if everyone reports to work and she receives the bonus; it is zero if she does not report to work; and it is negative if she reports to work but receives no bonus (because somebody else did not report to work). Having everybody report to work is an appealing equilibrium, but with a defection-deterrence index of only 1 (since one worker not showing up incentivizes others to not show up). This low sustainability level is one of the reasons that companies such as Toyota use more sophisticated production lines with higher defection-deterrence values (see Mishina (1992)).[12]

3.3. **Information, signals, and signal duplication.** Our next example illustrates that duplicating information can be used to increase equilibrium sustainability, even if the information is common knowledge. The example is a simple version of a Crawford-Sobel (1982) sender-receiver game, in which the senders, their information, and their strategies are simply duplicated.

Example 5. ***Signalling game with duplicated senders****: A two-element set $T = \{\alpha, \beta\}$ denotes two possible states, two possible signals, and two possible actions in the game. There are three senders, of which two are orderly and one chaotic, and there are 100 receivers. Each of the senders is informed about the true state $\theta \in T$, and recommends an action $s_i \in T$. The receivers are informed of which state is the most recommended (i.e., recommended by the majority of senders), and each one selects an action $\lambda_i \in T$. The payoff of*

[12]The author thanks Sunil Chopra and Martin Larivier for providing this reference.

the each receiver and each of the orderly senders equals the number of recivers who choose θ. The payoff of the chaotic sender equals the numbers of receivers who fail to choose θ.

We consider the straightforward equilibrium, SF, in which each the orderly senders recommend the true state θ, the chaotic sender recommends the false state, and each of the receivers chooses the most recommended action. It is easy to see that $D(SF) = 1$, because a defection by one orderly sender may push R to switch to the least recommended action.

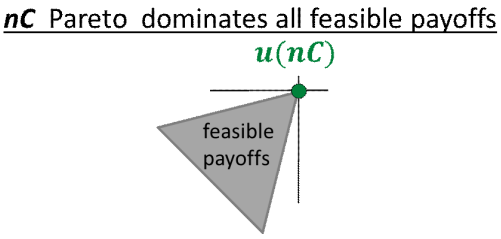
As the reader can see, if we alter the game to have four orderly senders, $D(SF)$ would be 2, because a defection by two orderly senders can push the receivers to go with the minority, yet no defection of a single player can push away any other player from the equilibrium strategy.

The fact that the SF equilibrium in the altered game is more sustainable than in the unaltered game is somewhat suprizing, since the two games are "common knowledge equivalent." It reveals a concern not captured by standard Bayesian equilibrium, namely, that one of the two senders in the unaltered game may fail to play the equilibrium for the various reasons discussed in the introduction. This concern is reduced in the altered game, where two senders have to fail. Notice also that the chaotic sender may also wish to duplicate herself, if she wishes to reduce the sustainability of the SF equilibrium.

It seems that the sustainability of the equilibrium may itself be subject to strategic considerations. This may explain why political parties send many representatives to repeat the same exact "talking points" in public presentations.

4. Comparison with equilibrium refinements

The nobody-confesses profile nC of the confession game is a Nash equilibrium with the lowest possible level of defection deterrence, $D(nC) = 1$. However, considerations of standard game theory lead one to conclude that nC is an appealing equilibrium. As one can see in the payoff graph of the two-player case, $u(nC)$ strongly and uniquely Pareto-dominates every feasible payoff in the game. Any defection – whether deterministic or probabilistic, and whether by individuals or groups – is certain to strictly decrease the expected payoffs



of all the players, including the defectors. As Aumann and Sorin (1989) argue, this should be the undisputed outcome of this game. Indeed, this equilibrium is *perfect* à la Selten (1975), *proper* à la Myerson (1978), *strong* à la Aumann (1959), and *coalition-proof* à la Bernheim, Peleg, and Whinston (1987).

It seems, however, that crime syndicates are not impressed by nC's high acclaim among game theorists. They are more concerned that somebody would confess out of fear that others might, a concern that coincides with the logic of the defection-deterrence index. Crime-syndicate remedies, such as killing confessors, change the game's payoff functions to make no-confession a dominant strategy with the maximal defection-deterrence level n. And more respectable organizations, such as high-tech and bio-research companies, demand that their employees sign no-disclosure agreements for similar reasons.

The example above also highlights our earlier point that the reasoning behind our stability index is simple enough to be understood even by players with a limited computational ability. Players simply have to fear the outcomes if other players defect.

The next example illustrates the difference between the defection-deterrence index and notions from evolutionary game theory. It illustrates that despite the evolutionary stability of an equilibrium, concerns about the environment in which a game is played may make players uneasy about playing an equilibrium.

Example 6. ***Match the Center.*** *The boss, B, and n subordinates each select one element from a given set of choices. B's payoff is 1 if he chooses E, zero otherwise; and a subordinate's payoff is 1 if her choice matches B's choice, zero otherwise.*

The game has a unique equilibrium eE, in which every player chooses E. This equilibrium has a large basin of attraction and is stochastically stable in the sense of Young (1993) and Kandori et al (1993).[13]

However, eE is only minimally sustainable, i.e., $D(eE) = 1$. All the subordinates are vulnerable because the optimality of their choice depends entirely on the choice of one player, B, and they may have concerns about B's commitment to his strategy, his ability to withstand pressure coming from outside the game, possible miscalculations, and so forth.

It is important to note that unlike the dynamic approach of evolutionary game theory, our model deals with a game that is played just once. Moreover, in many evolutionary models deviations from equilibrium are controlled by forcing mutation probabilities to approach zero. Here, on the other hand, imagined deviations from equilibrium are controlled only by bounding the number of players who may deviate, but the probabilities of their imagined deviations are unrestricted.

As a side remark about centrally controlled games, noticed that the sustainability of the eE equilibrium can be increased if the center is occupied by a group of bosses. For example, if there are three bosses, all having E as a dominant strategy, and the subrodinates all wishing to match the majority choice of the bosses then $D(eE) = 2$, and not 1 as it is for a single boss.

5. All index values are supported by simple rational reasoning

The next example shows that any integer, $0, 1, ..., n$, is a possible value of a defection deterrence level of some equilibrium. Moreover, these levels can be arrived at by simple rational reasoning, compatible with the computational ability of most players.

Example 7. ***Ride-sharing game****: Eight individuals from a small town sign up to attend an event at specific time and place in a nearby city. For transportation, they each have to sign up and commit to one of two options: riding a private taxi that costs* $80*, or sharing a ride on a bus that can comfortably take any number of them. The cost of the bus,* $180*, will be shared equally by*

[13]For further discussion and references on these topics, we refer the reader to Ellison (2000).

Defection-deterrence computations of eT

# of bus riders x	cost/rider $180/x	cost of taxi $c	Deterrence D(eT))
		$180 < c$	0
1	180	$90 < c \leq 180$	1
2	**90**	$60 < c \leq 90$	**2**
3	60	$45 < c \leq 60$	3
4	45	$36 < c \leq 45$	4
5	36	$30 < c \leq 36$	5
6	30	$25.7 < c \leq 30$	6
7	25.7	$22.5 < c \leq 25.7$	7
8	22.5	$c \leq 22.5$	8

the riders who sign up for it. Assuming that the only consideration of every rider is to minimize her transportation cost, what would they choose?

The table titled "Defection-deterrence computations of eT" illustrates the simplicity of computing the defection-deterrence index for the profile in which everybody chooses the taxi, eT, as a function of the cost of riding the taxi, c.

When $c = 80$, as in the example above, we look at the third row of the table, the case in which $60 < c \leq 90$. A single player's defection from a taxi to the bus can result only in a loss, raising her cost from \$80 to \$180, so eT is a Nash equilibrium. But what about multi-player defections?

Even in the presence of one other potential defector, a player who switches from the taxi to the bus is sure to lose: her best possible outcome after switching is a cost of \$90 (= \$180/2) for the bus, instead of \$80 for the taxi. But eT does not deter defections by groups of three or more players, because now a defection may actually improve the player's outcome, by reducing her costs from \$80 for the taxi to \$60 (= \$180/3) for the bus, if all three sign up for it. This leads us to say that eT deters two defectors but does not deter three defectors, and to conclude that the deterrence index of eT is $D(eT) = 2$. The table titled "Defection-deterrence..." shows how this reasoning applies to all values of taxi costs to generate all the levels of defection deterrence.

The eT equilibrium also illustrates the type of defections that players may consider. Do they consider individual or coordinated defections? Three

people moving from the taxi line to the bus may come about through explicit communication: for example, when a ride seeker standing in a taxi line approaches two others to coordinate a money-saving joint defection. But alternatively, a ride seeker may switch to the bus, counting on the likelihood that for similar reasons at least two others will switch. In either case, having a higher defection-deterrence index, i.e., requiring the participation of more switchers, makes such defections more difficult to bring about.

6. Forming, switching, and undoing equilibria

6.1. **An illustration of a viable equilibrium switch.** In the ride-sharing game above, having everybody take the bus, eB, is a Nash equilibrium with deterrence index $D(eB) = 6$, and thus a formation-difficulty index $F(eB) = 8-6 = 2$. In other words, two riders committing to take the bus should suffice to convince the remaining six riders, if they are rational, to take the bus too.

In general, depending on the game, there may be a variety of reasons why $F(\pi)$ of the n players would commit to their π strategies. These can be easily illustrated in the ride-sharing game. First, as discussed above, any two rational players may realize that forming the equilibrium eB would reduce their costs, from \$80 all the way down to \$22.5 (= 180/8). This may motivate them to make an agreement in which they both take the bus. Moreover, the agreement can be counted on since it is compatible with their individual incentives throughout the process of the equilibrium formation.

A second example is an initiative of the bus company, which would like to see the formation of the equilibrium eB. If they guarantee the first two bus-riding candidates that their cost will never exceed \$79, no matter what the other riders do, they can count on the rationality of the first two to take the bus and on the others to follow. Notice that this manipulation by the bus company involves little risk, due to the low formation index, $F(eB) = 2$.

While in the example above, switching the equilibrium from eT to eB seems viable, other situations that call for switching equilibria are less viable, as can be seen from theoretical considerations reinforced by observed behavior.

6.2. **Nonviable switch: changing the US measurement system.** Consider a large game of matching measurement systems, the same as the (language) matching game but where each player has to select one of two choices: the metric system, MT, or the US measurement system, US. Every player choosing MT, eMT, and every player choosing US, eUS, are both Nash equilibrium, and both have a defection-deterrence index $D = 100M$. With their complementary high formation-difficulty index, $F = 100M$, it seems that establishing either one of these equilibria in a new population would be challenging.

However, in an existing population in which the equilibrium eUS is already established, forming (switching to) the equilibrium eMT would be even more challenging. For every player i and every profile of opponents' choices, β_{-i}, the gain from choosing MT over US, $u_i(MT_i, \beta_{-i}) - u_i(US_i, \beta_{-i})$, is lower in the established population due to the associated transition costs and other such considerations. This means that in this population, defections from MT to US are easier than defections from US to MT, which means that $F(eMT) > 100M$. In other words, it is even harder to move a population in which everybody uses US to MT, than it is to guide a new population to have everybody choose MT.

The US experience with measurement systems illustrates this type of difficulty. Attempts to switch the US population from the use of the US measurement system to the metric system keep failing despite encouraging actions taken by the US government, including the Congress, in 1866, 1873, 1893, 1968, 1975, and 1988. It seems that encouraging a change does not overcome the high formation difficulty. What would be helpful – and perhaps indispensible – is a law imposing penalties for use of the US system. With sufficiently high penalties, the use of the metric system would become a dominant strategy with minimal formation difficulty.

6.3. **Formation difficulty as the natural index for viability.** The explosion of web devices gives rise to many new games and equilibria. The viability of proposed equilibria in such games is important to investors. The next example, discussed in the introduction, shows how the formation index may be useful in such situations.

Recall the example: *A new communication network is offered for subscription in a population of 200M potential users, at an individual cost of $9.99. The payoff of a subscriber is* $k - 9.99$*, where* k *is the number of opponents (other players) who subscribe. The payoff of a nonsubscriber is* zero.

The network provider may be interested in the viability of the equilibrium in which every player subscribes, eS. It is easy to see that for every player i, the resistance to joining is $JR_i(eS) = 9$: if 9 opponents subscribe, it is a (weakly) dominant strategy for her to subscribe. This implies that $F(eS) = \max_i F_i(eS) = 9$, which implies that this game has a high defection-deterrence index, $D(eS) = 200M - 9 = 199,999,991$. The relatively low formation-difficulty value, coupled with a high defection-deterrence value, suggests to the network provider that eS is a viable new equilibrium.

As the reader may see, the simple computation above can be conducted by the network provider for different subscription costs, and also for asymmetric potential subscribers. Moreover, even if the provider knew that the pool of potential users was large, but without knowing their exact number, the Nash defection-tolerance property assures him of the relevance of the computation and their payoff implications.

7. Implementation in the presence of faulty players

Going beyond Maskin (1999), Eliaz (2002) studies implementation in an environment in which k of n players may be faulty. Implementation in such an environment is difficult because (1) the identity of the faulty players is unknown, and (2) the faulty players are irrational and choose unpredictable strategies. Thus, an Eliaz implementor can rely on the rational behavior of only $n - k$ unknown players. And like the implementor, every rational player knows that she is making choices in an environment with k unknown faulty players.

To overcome these difficulties, Eliaz introduces an implementation method that makes use of an equilibrium concept he calls *k-fault-tolerant Nash equilibrium* (k-$FTNE$). A profile π is a k-$FTNE$ if playing π_i is a best response for every rational player i, when it is common knowledge that the number of faulty players is at most k. Through the use of this concept, Eliaz shows

that the implementor can accomplish his goal, provided that the social-welfare function satisfies appropriate monotonicity conditions.[14]

While the objectives of the current paper are different from the objective of Eliaz (2002), the viability indices discussed here provide a simple interpretation of Eliaz's findings. In particular, Eliaz's faulty players may be viewed as a specific type of defectors in the current paper, and for this reason his equilibrium concept may be stated through the notion of Nash critical mass presented here. More specifically, saying that "π is k-$FTNE$" in Eliaz' language is the same as requiring the number of rational players to exceed the Nash critical mass of π, i.e., $n - k \geq NCM(\pi) = F(\pi) + 1$. This means that n- $F(\pi) \geq k + 1$, i.e., that $D(\pi) \geq k + 1$. So π is k-$FTNE$ iff $D(\pi) > k$.

We conclude that in an environment with faulty-players, Eliaz implementors can implement a social welfare function that satisfies Eliaz's monotonicity condition, provided that they uses an equilibrium π with a $D(\pi)$ that strictly exceeds the number of faulty players.

It is important to note that the findings of the current paper go significantly further than the implementation results of Eliaz. While the faulty players of Eliaz are a special type of defectors in the current paper, the current paper also studies the incentives of rational (nonfaulty) players to defect or to join an equilibrium. These incentives are important, for example, for agents who wish to do such things as (1) undo an equilibrium by bribing or threatening rational players to defect, (2) convince players to join an equilibrium, counting on the rational incentives of the others to join, (3) convince rational players to switch equilibrium, or (4) reinforce the play of an equilibrium.

The current paper relates to the implementation problem studied by Abraham et al. (2006) in a way that is similar to how it relates to Eliaz (2002). Abraham et al. (2006) study implementation in secret sharing and in multiplayer computation games, and use what they call *resilient equilibria* in order to overcome difficulties due to faulty play. As stated earlier in section 1.2, the measure of defection deterrence in this paper is the same as their measure of *resilience*; as a result, the classes of games they study consist of games with equilibria having high defection-deterrence levels.

[14]These are conditions related to the ones in Maskin (1999).

Other positive results related to faulty play are presented in Gradwohl and Reingold (2014), who use results about the robustness of equilibria of large games (similar to the robustness results in Kalai (2004)) to show that such equilibria can sustain a significant number of defectors.

8. Matching in networks

Games on networks, as in Jackson and Zenou (2015), provide an understanding of the viability of Nash equilibria as determined by social connectivity. It is easy to compute the defection-deterrence index for equilibria of a *network matching game*, described as follows.

The set of players N consists of the vertices in a graph with a set of directed edges $E \subset \{(i,j) \in N \times N : i \neq j\}$. The set of (outward-directed) *neighbors* of a player i is defined by $\eta b(i) = \{j \in N : (i,j) \in E\}$. Every player selects a choice X from a set of possible choices, and her payoff is the number of her neighbors that her choice matches.

Let C denote the set of *connected* players, i.e., players i with $\eta b(i) \neq \emptyset$, then a player $v \in C$ is *most vulnerable* as defined in section 2.1, if she is minimally connected among all connected players, i.e., $|\eta b(v)| = \min_{j \in C} |\eta b(j)|$.

Proposition 6. ***Defection deterrence in network matching games****: The defection-deterrence index of the profile in which every player chooses X, eX, is*

$D(eE) = n$ if no player is connected, i.e., $C = \varnothing$; otherwise,

$D(eE) = [|\eta b(v)|/2] + 1$, where v is any most vulnerable player. In other words, it is the strict majority of neighbors of a least connected player.[15]

Proof. Clearly, no group of players can push away a disconnected player, which is the reason for the first implication. And from the definition, a connected player can be pushed away by a group of players iff the group includes a strict majority of her opponents. □

A similar analysis can be easily conducted for problems of mismatching in a network. For illustration, recall the Mismatch-the-opposition game discussed

[15]Recall that [x] is the largest integer value below a number x.

in section 1.3 of the introduction, in which we considered the completely divisive equilibrium, $cDIV$, in which all three Democrats choose F and all five Republicans choose E. This game may be described by a bipartite graph, connecting every player to all the players of the opposite party, with a player's payoff being the number of her neighbors that her choice *mismatches*. As in Proposition 6 above, every player can be pushed away from the $cDIV$ equilibrium by a majority of the opposition. This means that the Republicans are most vulnerable, since they are each connected only to the three Democrats, whereas each Democrat is connected to the five Republicans. Conducting the same analysis as above, we conclude that $D(cDIV) = [3/2] + 1 = 2$.

8.1. **Centralized vs. decentralized interaction.** The language choice game, discussed in the introduction, is an example of a network matching game based on a complete graph, i.e., every two distinct players are connected (in both directions). In this game every player has $200M - 1$ neighbors, so $D(eE) = [(200M - 1)/2] + 1 = 100M$.

This is in contrast to the low defection-deterrence level of centralized interaction, as discussed in the Match-the-Center game. That game is based on a star-shaped graph in which the boss B is the center vertex, and n subordinate players are vertices connected to him. In that game the equilibrium eE has the minimal value $D(eE) = 1$.

The contrast in defection-deterrence levels of the two games above has interesting implications in a variety of contexts. In a political context, it suggests that matching equilibria are significantly more sustainable (in the sense defined in this paper) in free societies than in dictatorships. In games of currency choices, it suggests that free-trade equilibrium is more sustainable than centralized-trade equilibrium. In supply-chain games it means that relying on a single source is risky, and backup sources for supplies are important for sustainability.

As already discussed, the low defection-deterrence value of the star-shaped graph is due to the total dependence of the subordinate players on the boss, B. If B defects from the equilibrium – due for example to threats and bribes or to miscalculations – the effect on all the subordinates may be devastating.

9. Future research

9.1. **Expanding on Nash-existence-theorem.** Nash's existence theorem provides condition on a game that assures the existence of a profile π with $D(\pi) \geq 1$. For applications in which the low level of sustainability $D(\pi) = 1$ is a concern, one would like to have an answer to the following: for integers $k > 1$, what are sufficient conditions on the game that assure the existence of an equilibrium π with $D(\pi) \geq k$?

9.2. **More refined directions.** To stay close to the Nash view, the viability indices studied in this paper consider ordinal best response to opponents, and amend it only with the count of the number of defectors that violate the ordinal best-response property. But this minimal approach cannot deal with some applications that require more subtle reasoning. We illustrate some below.[16]

Discontinuity of the D index: Consider the matching game defined by a complete graph on n players, K_n, and compare it with the following amended version K_n^+: K_n^+ includes all the players and edges of K_n but adds to it one more player, H, who is connected (in both directions) only to one of the n players in K_n. Sustainability computations point to a large discrepancy: $D_{K_n}(eX) \approx n/2$ whereas $D_{K_n^+}(eX) = 1$.

The discrepancy above is disturbing for some important applications. For example, for assessing the reliability of communication in two populations, one structured accoding to K_n and one according to K_n^+, you may want to compare the sustainability of the equilibrium eE, where E is a common language chosen in both populations. For large values of n, the difference, $n/2$ versus 1, does not reflect the fact that communication is essentially the same in both populations, with the exception of only one out of the n players.

More refined defection measures: Recall for example, the no-confession equilibrium nC of the confession game, discussed in section 1.1. The benefit, in reduced jail time, to a player who confesses is $r = 7$ years ($r = 10 - 3$ years). It is easy to see, however, that its low defection deterrence value of $D(nC) = 1$ holds for all positive levels of sentence reduction r. Under a more

[16]Substantial discussion of this subject with highly revealing examples is provided by Goeree and Holt (2001) and its follow-up papers.

careful analysis we may want to assign a higher defection-deterence value to nC as the benefit from confessing decreases. In many applications it may be reasonable to define defection-deterrence indices that assume a continuum of values, as r decreases continuously.

Structural considerations: Consider the the amended $(n+1)$-player complete graph K_n^+ discussed under the discontinuity illustration above, and compare it with the $(n+1)$ star-shaped graph C_n in which n players are each connected to one central player, discussed in section 8.1. In a language choice game in which eE is the common-choice equilibrium, $D(eE) = 1$ in both graphs. This coincidence misses the fact that a defection of the central player in C_n would have a devastating effect on the communication ability of the players, whereas the defection of any of the players in K_n^+ will have only a minor effect on the rest.

The structural insensitivities discussed in the last paragraph are due to the fact that our viability indices are anonymous in that they ignore the identity and position in the game of defecting players and of the players who incentivize them to defect.

9.3. **Nonanonymous indices and equilibrium formation.** Nonanonymous indices may depend on the identities of defectors and the players who incentivize them to defect. Such indices may be more applicable, but may also require more demanding computations. This paper does not study nonanonymous indices, but we proceed to show how the concepts already discuss may fit into such a broader discussion.

Our formation-resistance index $F(\pi)$ is anonymous; it is the minimal integer k such that *any* coalition C with k or more players can form π. But in an analysis of equilibrium formation, we may focus on coalitions with specified (nonanonymous) players which can form the equilibrium.

Following a nonanonymous version of our equilibrium-formation index, we would say that a coalition C forms the equilibrium π if the play of π by C makes π a dominant strategy for the remaining players, outside C. It is easy to see that being π-forming defines a monotonic partial order (in containment) over the coalitions of the game (i.e., if C forms π, so does any of its supersets), with the grand coalition N being its unique maximal element.

Since our underlying game has a finite number of players, we can identify the minimal forming coalitions – the *roots*.

Definition 11. *A* π-root *is a minimal* π*-forming coalition, i.e., no strict subset of a root forms* π*.*

Clearly, a coalition forms π iff it contains a root, and a coalition is incentivized to play π iff its complement contains a root. Also, since any coalition whose size is at least $F(\pi)$ forms π, it must contain a root.

Consider for example the Mismatch-the-opposition game (MMOP), with the completely divisive equilibrium, $cDIV$, in which the three Democrats all choose F and the five Republicans all choose E. It is easy to verify that the the following two coalitions are roots of $cDIV$: (i) all the Democrats and (ii) all the Republicans.

With this in mind, we may consider the following three ways of forming the equilibrium $cDIV$:

CD: Convince the three Democrats to choose F.

CR: Convince the five Republican to choose E, and

C6: Convince *any* six players to choose their divisive strategies.

CD and CR work, because they target two roots by name. C6 is anonymous but it works because $6 = F(\pi)$.

As this example illustrates, it may be more efficient to form an equilibrium through the use of nonanonymous coalitions. But due to the multiplicity of roots, the decision of which root to target may require more information and computations. In the simple game above, for example, one would have to determine whether it is easier to convince the three Democrats or to convince the five Republicans. The direction discussed next may lead to more efficient processes of equilibrium-formation.

9.4. **Dynamic formation processes.** Dynamic formation processes may be developed by replacing the one-step formation with sequential ones. For example, in the Mismatch-the-opposition game (MMOP), one may first recruit two Democrats to choose F; this in turn will incentivize all the Republicans to choose E, which in turn will incentivize the third Democrat to choose F.

While the more refined multistage approach is useful in many applications, it involves a larger number of possible procedures, which in turn requires more complex computations. This is similar to the computations involved in sequential elimination of dominant strategies (See, for example, Gilboa et al. (1993) and Marx and Swinkels (1997)).

10. References

Abraham, I., D. Dolev, R. Gonen, and J. Halpern (2006), "Distributed computing meets game theory: robust mechanisms for rational secret sharing and multiparty computation," in Proceedings of the. 25th ACM Symposium on Principles of Distributed Computing, 53–62.

Aumann, R.J. (1959), "Acceptable points in general cooperative n-person games," *Contributions to the Theory of Games IV*, Annals of Mathematics Study 40, Princeton University Press, 287-324.

Aumann, R.J. and S. Sorin (1989), "Cooperation and bounded recall," *Games and Economic Behavior*, 1 (1), 5-39.

Basu, K. and J.W. Weibull (1991), "Strategy subsets closed under rational behavior," *Economics Letters* 36, 141-146.

Ben-Or, M., S. Goldwasser, and A. Wigderson, (1988) "Completeness theorems for non-cryptographic fault-tolerant distributed computation," in STOC '88 Proceedings of the twentieth annual ACM symposium on Theory of computing, ACM New York, 1-10.

Bernheim, B. D. , B. Peleg, and M. D. Whinston (1987), "Coalition-proof equilibria: I. Concepts," *Journal of Economic Theory*, 42, 1–12.

Camerer, C. (2003), *Behavioral Game Theory: Experiments in Strategic Interaction*, Princeton, NJ: Princeton University Press.

Chwe, M. (1994), "Farsighted coalitional stability," *Journal of Economic Theory*, 63, 299-325.

Crawford, V.P. (1998), "A Survey of Experiments on Communication via Cheap Talk." *Journal of Economic Theory*, 78, 286-298.

Crawford, V.P. and J. Sobel (1982), "Strategic information transmission," *Econometrica,* 50 (6), 1431-1451.

Eliaz, K. (2002), "Fault-tolerant implementation," *Review of Economic Studies*, 69(3), 589-610.

Ellison, G. (2000), "Basins of attraction, long run stochastic stability, and the speed of step-by-step evolution," *Review of Economic Studies*, 67 (1), 17-45.

Erev, I. and A. E. Roth (1998), "Predicting how people play games: reinforcement learning in experimental games with unique,mixed strategy equilibria," *The American Economic Review,* 88 (4), 848-881.

Gilboa, I., E. Kalai, and E. Zemel (1993), "The complexity of eliminating dominated strategies," *Mathematics of Operations Research,* 18, 553-565.

Goeree, J.K. and C.A. Holt (2001), "Ten little treasures of game theory and ten Intuitive contradictions," *American Economic Review*, 91 (5), 1402-1422.

Gradwohl, R. and O. Reingold (2014), "Fault tolerance in large games," *Games and Economic Behavior*, 86, 438-457.

Jackson, M. O. and Y. Zenou. (2015) "Games on Networks," In *Handbook of Game Theory with Economic Applications*. Vol. 4. Elsevier.

Kahneman, D. and A. Tversky (2000), *Choices, Values, and Frames*, New York : Russell Sage Foundation.

Kalai, E. and A. Neme (1992), "The strength of a little perfection," *International Journal of Game Theory*, 20 (4), 335-355.

Kalai, E. and D. Samet (1984), "Persistent equilibria in strategic games," *International Journal of Game Theory*, 13(3), 129-144.

Kalai, E. (2004), "Large robust games," Econometrica, 72 (6), 1631-1665.

Kandori, M., G. Mailath, and R. Rob (1993) "Learning, mutation, and long run equilibria in games," *Econometrica*, 61(1), 29-56.

Kohlberg, E. and J.F. Mertens (1986), "On the strategic stability of equilibria," *Econometrica*, 54 (5), 1003-1037.

Kreps, D.M. and R. Wilson (1982), "Sequential equilibria," *Econometrica,* 50 (4), 863-894.

Marx, L. M., and J. M. Swinkels (1997), "Order independence for iterated weak Dominance," *Games and Economic Behavior*, 18, 219-245.

Maskin, E. (1999), "Nash implementation and welfare optimality," *Review of Economic Studies*, 66, 23-38.

Mishina, K. (1992), "*Toyota Motor Manufacturing, U.S.A., Inc.*" Harvard Business School Case 693-019, September 1992. (Revised September 1995.)

Moreno, D. and J. Wooders (1996), "Coalition-proof equilibrium," *Games and Economic Behavior*, 17, 80–112.

Myerson, R.B. (1978), "Refinements of the Nash equilibrium concept," *International Journal of Game Theory,* 7, 73-80.

Myerson R.B. and J.W. Weibull (2015), "Tenable strategy blocks and settled equilibria," *Econometrica* 83 (3), 943-976.

Nagel, R. (1995), "Unraveling in guessing games: an experimental study," *American Economic Review*, 85 (5), 1313-1326.

O'Neill, B. (1987), "Nonmetric test of the minmax theory of two-person zerosum Game," *Proceedings of the National Academy of Sciences*," 84(7), 2106– 09.

Schelling, T.C. (1960), *The strategy of conflict* (1st ed.), Cambridge: Harvard University Press.

Selten, R. (1975), "Reexamination of the perfectness concept for equilibrium points in extensive games," *International Journal of Game Theory*, 4 (1), 25-55.

Smith V.L. (1982), "Microeconomic systems as an experimental Science," *The American Economic Review*, 72 (5), 923-955.

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

Kellogg School of Management, Northwestern University
E-mail address: kalai@kellogg.northwestern.edu
URL: http://www.kellogg.northwestern.edu/faculty/directory/kalai_ehud.aspx