Control of Parametric Games

Carmel Fiscko Brian Swenson Soummya Kar Bruno Sinopoli

***Abstract*— This work studies a class of multi-player games in which the players' decisions can be influenced by a superplayer. We define a game with n players and parameterized utilities $u(\cdot, \alpha)$ where the superplayer controls the value of α. The regular players follow Markovian repeated play dynamics that encompass a wide class of learning dynamics including strict best response. The objective of the superplayer is to control α dynamically to achieve a desired outcome in the game-play, which in this work we define as the realization of target joint strategies. We introduce the class of parametric games and reformulate the superplayer control problem as a Markov decision process (MDP). Reachability criteria are developed, allowing the superplayer to determine which game-play may occur with positive probability. With a reachable goal joint strategy, a *cost*-optimal policy can be computed using standard tools in dynamic programming. A sample MDP reward function is presented such that a reachable target joint strategy is guaranteed to be played almost surely. Finally, an application in a cyber-security context is provided to illustrate the use of the proposed methodology and its effectiveness.**

I. Introduction

Game theory has become an effective tool for analyzing cooperative or competing multi-agent systems [1]. The framework provides a structure for modeling decision making and understanding how collective group actions influence the outcome of each individual.

Games have been used to model and analyze a wide range of applications from computer security to biological systems, social networks, and economics. In computer security the tension between hackers, malware, and networks can be understood and influenced through accurate game modeling [2]. Other security-based applications include game-centric modeling of attacks on power grids and critical infrastructure to develop defense mechanisms [3]. Recently the field of social networks has emerged as a new area of game theoretic study, as researchers are interested in modeling interesting phenomena such as opinion dynamics and the impact of leaders in media networks [4], [5].

In many of these settings, players can be unsophisticated or limited by bounded rationality. Here players can adaptively "learn" to play an advantageous or optimal (e.g., Nash equilibrium) joint strategy by repeatedly playing the game, modifying their strategy over time, and observing the results [6], [7]. Beyond identifying the equilibria, a game designer may want to influence the specific steady state reached by such an adaptive game-theoretic learning process. Work has been done to construct games whose equilibria have desirable properties, such as converging to an equilibrium that minimizes local or global costs [8], [9]. Beyond constructing games, games have been viewed from a control theoretic perspective as in [10], where the authors use a structured pricing scheme to place a game's unique equilibrium point at some desired location.

Despite the significant interest in game-play equilibria, little work has been done to control game-play to some arbitrary joint strategy. This could be interpreted as guiding game-play to one specific equilibrium or seeking repeated play at some non-equilibrium strategy.

In this paper our goal is to understand how to achieve game-play within some target set of joint strategies given a limited set of controls. Our approach focuses on games whose player utilities can be expressed as $u(\cdot, \alpha)$ with a dependence on some parameter α. This class of games we refer to as **parametric games**. Such a parameter could represent factors such as tolls, resources, or an impact from the occurrence of an event. In our work this parameter is controlled by some third-party **superplayer** who dynamically controls the value of α such that the resulting game-play is guided to within their desired set of joint strategies. Furthermore if the target set includes a Nash equilibrium strategy, then we will show when continuous game-play within the desired set can be guaranteed.

Games with parameterized utilities have been previously used for modeling and learning purposes. These parameters may represent real-world varying utilities [11], or hidden values to be learned through repeated play [12]. From a control perspective, pricing mechanisms have been used within utilities [10], and we build upon this idea.

From a broad goal perspective, this work holds similarities to mechanism design, in which player incentives are engineered to achieve desired game-play objectives [13]. However, mechanism design is conceptually different from our formalism of parametric games and associated control objectives, in that we deal with dynamic processes on games. We assume players are endowed with some adaptive learning dynamics, and we characterize the ability of a superplayer to control the *steady state* by altering the players' learning strategy; this in turn dynamically evolves the equilibria based on the current value of the parameter α. We draw inspiration from the goals of mechanism design, but concentrate on analyzing and exploiting the control capabilities of the superplayer based on the utility structure.

Our investigation focuses on parametric games played with repeated play Markovian dynamics. As the superplayer changes the parameter α, the players' incentives are altered to dynamically favor different actions. We show that under broad assumptions, the game-play can be controlled via the parameter α to converge to a desired Nash equilibrium or other desired joint strategy. The Markovian assumption allows the learning dynamics to be studied via an MDP,

which yields optimal policies for the superplayer to drive the game-play to the desired outcome. This structure will equip the superplayer with necessary tools to both analyze the feasibility of their goal, and precisely calculate the controls to realize that goal.

The paper is organized as follows: Section II introduces game notation, demonstrates that games with repeated play Markovian dynamics can be viewed as Markov chains, and defines parametric games. Section III relates parametric games to MDPs. Section IV contains the main results, which (1) detail a reachability analysis on the goal, and (2) present an MDP reward function such that optimal policies guarantee game-play at the desired state almost surely. Finally Section V presents an example in the setting of a denial of service attack.

II. Game Model

A. Game Theory Preliminaries

A finite normal form game is defined here as the tuple $G = (N, S_{p_{p\in N}}, u_{p_{p\in N}}(\cdot))$.

The set of players is N, and there are $|N| = n$ standard players. Each player $p \in N$ has a strategy set S_p. The set of all joint strategies is $S = \prod_{p=1}^{n} S_p$, and $s \in S$ denotes a specific joint strategy. For $s \in S, p \in N$, let $s_{\bar{p}}$ be the strategy tuple of all players except for player p. Each player $p \in N$ also has a utility (or payoff) function $u_p(\cdot) : \cdot \to \mathbb{R}$.

a) Dynamics: A game-theoretic learning algorithm is a set of decision rules (possibly randomized) governing the behavior of the strategy sequence $\{s_t\}_{t\geq 0}$. We will consider repeated play learning dynamics in this paper. Suppose that a group of players repeatedly play some fixed game G in stages $t = 1, 2, \ldots$. Let $s_t = (s_{1,t}, \ldots, s_{n,t}) \in S$ denote the joint strategy played in round t.

In this paper we consider broadly the class of Markovian learning dynamics. More precisely, we suppose that $\{s_t\}_{s_t\geq 0} \subset S$ is a stochastic process satisfying the following assumption.

Assumption 1: The game dynamics are Markovian, i.e., $P(s_{t+1} = s|s_t, s_{t-1}, \ldots, s_0) = P(s_{t+1} = s|s_t)$.

This setup could be extended to finite memory learning dynamics, i.e., $P(s_{t+1}|s_t, s_{t-1}, \ldots, s_0) = P(s_{t+1}|s_t, \ldots, s_{t-m})$, but it will not be investigated here.

An example of dynamics that satisfy A1 is best response dynamics, where $s_{p,t+1} \in \arg\max_{s_p} u_p(s_p, s_{\bar{p},t}), \forall p \in N$. If there is some deterministic tie-breaking rule between actions that yield equal utilities (such as always picking the action with the lowest index), then we note that the best response dynamics are in fact deterministic. If there is a stochastic tie-breaking rule (such as uniformly selecting between equally good actions), then the best response dynamics are stochastic. Another example of learning dynamics is fictitious play [14], as players must maintain and react to an empirical history of their opponents' actions. Fictitious play is not Markovian with respect to the previously played actions. Other non-deterministic dynamics that do fulfill A1 such as aspiration dynamics [15] can be used for our analysis.

Here, we will also restrict our dynamics to those that are static with respect to time, i.e. the same learning rule is used at each stage of play.

b) Equilibria: As equlibria, we adopt the most widely used concept, namely the Nash Equilibrium.

Definition 1: A pure strategy **Nash Equilibrium** (NE) is a joint strategy profile $s \in S$ such that no single player can obtain a higher payoff by unilaterally deviating from this profile, i.e.,

$$u_p(s) \geq u_p(s'_p, s_{\bar{p}}) \quad \forall p \in N, \forall s'_p \in S_p. \tag{1}$$

In this paper, we restrict attention to pure strategies and pure NE. In what follows, for brevity, we will refer to pure NE simply as NE. We note that, in general, pure NE may not exist; however, existence can be guaranteed by assuming, for example, that G is a potential game [16].

B. Game-Play as a Markov chain

To study game properties such as convergence to NE, we note that game-play fulfilling assumption A1 evolves as a Markov chain (MC).

Definition 2: A discrete time stochastic process $\{x_t\}_{t\geq 0}$ with finite state space X, where $t = 0, 1, \cdots$, denotes a discrete index, is said to be **Markovian** if the future is independent of the past when conditioned on the present, i.e.,

$$P(x_{t+1} = x|x_t, x_{t-1}, \ldots, x_0) = P(x_{t+1} = x|x_t), \forall x \in X. \tag{2}$$

Let the set of states X be equal to the set of joint strategies S. Denote by $P(s_{t+1} = s|s_t), \forall s \in S$, the transition probabilities dictating the probability of the next played joint strategy given the current joint strategy. Under A1 only the current joint strategy needs to be considered when analyzing state transitions. The defined states and the transition probabilities (that are determined by the particular game-play dynamics in force) formalize the MC model.

For example, consider the two player coordination game $G1$ below where the entries indicate the utilities $(u_{p_1}(s), u_{p_2}(s))$ for playing the associated actions.

G1	P2 action 1	P2 action 2
P1 action 1	(2, 2)	(0, 0)
P1 action 2	(1, 1)	(0, 0)

Suppose that players use strict best response dynamics. Under these dynamics the chosen strategies will follow the transitions shown in Fig. 1 with each transition occurring with probability 1. It can be verified that the joint strategy $(1, 1)$ with payoff vector $(2, 2)$ is the only pure NE.

Through repeated play, the players will always settle on joint strategy (1, 1), and will never switch to any other action profile.

C. Parametric Games

Continuing from the previous example, suppose now that there is a **superplayer**, an external entity as far as the game is concerned, observing the game, for whom it is beneficial for players to play specific joint strategies. For example,

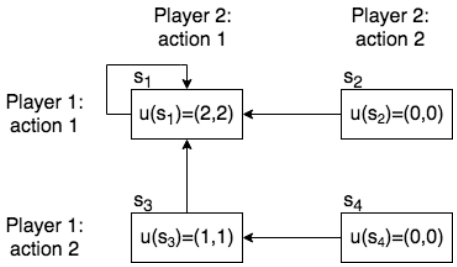

Fig. 1. Game $G1$ as a Markov Chain

suppose that the superplayer prefers players play the steady-state action profile $(1, 2)$.

A solution is adaptively incentivize desired game-play by changing the utilities based on the current game-play. Broadly speaking, in this paper, we will assume that the superplayer may select actions from a finite strategy set A to affect the utilities of the other players, thus potentially altering the game in question. To do this we will restate the utilities parametrically as a function of the superplayer's chosen action $\alpha \in A$. To formalize this, we introduce the notion of parametric games below.

Definition 3: Let $G = (N, S_{p,p\in N}, u_{p,p\in N}(s, \alpha))$ be a finite n-player game whose utility functions are parameterized by $\alpha \in A$, where A is a finite set. We say that G is a **parametric game**.

We assume the superplayer may control the value of $\alpha \in A$. The superplayer can be thought of as player $n + 1$, but their goal is to steer the normal players to play some chosen joint strategies. The superplayer will have a different utility function and follow a different set of dynamics from the rest of the players, which will be defined in Section III-D.

For any fixed α the resulting player utilities take on values that generate a probability distribution over the next chosen joint strategy. We will extend the assumption A1 to reflect this dependence on α.

Assumption 2: The game dynamics satisfy the following Markovian property:

$$P(s_{t+1} = s|s_t, s_{t-1}, \ldots, s_0, \alpha_{t+1}, \alpha_t, \ldots, \alpha_1) = P(s_{t+1} = s|s_t, \alpha_{t+1}), \forall s \in S.$$

The parameter α may represent a toll, artificial congestion, availability of resources/infrastructure, or other reasonably changeable specification. For now we assume that the dynamics rules are fixed and time-invariant.

In a parametric game, different joint strategies may become NE of the game resulting from a fixed α. We will say that a joint strategy s is a **feasible NE** if there exists an α such that (1) holds mutatis mutandis.

As an illustrative example, consider the parametric game $G2$ below.

G2	P2 action 1	P2 action 2
P1 action 1	$(\alpha + 2, \alpha + 2)$	(0, 0)
P1 action 2	$(\alpha + 1, \alpha + 1)$	(0, 0)

Note that, setting $\alpha = 0$ yields $G1$, where the pure NE is (1, 1) and best-response dynamics lead to a steady-state realization of that strategy. If instead the superplayer sets $\alpha = -3$, the resulting Markov chain for α fixed will have completely different transition probabilities as shown in Fig. 2, where each possible transition (denoted by an arrow) occurs with probability 1.

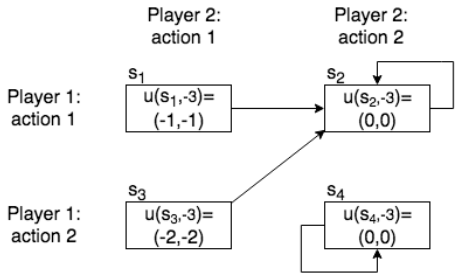

Fig. 2. $G2, \alpha = -3$

We can see that $(1, 2)$ and $(2, 2)$ are NE for $\alpha = -3$, and can therefore illustrate that in the parametric game $G2$ the set of feasible NE are $\{(1, 1), (1, 2), (2, 2)\}$ for the set $A = \{-3, 0\}$.

This example further suggests that it may be possible to actively change α based on the current state to achieve game-play at some desired state. For example, suppose $\alpha = -3$, and note that $s_4 = (2, 2)$ is a NE for this choice of α. Suppose, moreover, that the superplayer has some predilection for the strategy profile $s_2 = (1, 2)$, and would prefer this strategy profile be played in the long run. In order to accomplish this, α must be changed to 0 for the system to leave s_4, after which α can be changed back to -3 for game-play to reach the NE at s_2. This result shows that in a parametric game, the superplayer can target specific joint strategies and change the played NE by exploiting this utility structure.

From the superplayer's perspective, we can see that a parametric game with fixed learning dynamics looks like a Markov decision process with time invariant transition kernels dependent on α. This suggests that the goal of choosing the played NE can be viewed as an optimal control problem that is solved via dynamic programming with the superplayer's strategy α acting as the control.

III. Parametric game as a Markov decision process

Given the parametric game G with fixed learning dynamics, we wish to study policies for the superplayer that select α in order to drive game-play to a desired strategy. We have seen that a parametric game can be viewed as a Markov chain with transition kernels controlled by α, so it is a natural step to restate the problem as an MDP and solve for optimal policies.

A. Markov Decision Problem

Definition 4: A finite state-action space **Markov decision process** is a tuple $M = (X, A, T, r)$ consisting of: discrete set of states X; discrete set of control actions $A \supseteq \alpha$; transition matrices $T(\alpha)$ on X where $T_{ij}(\alpha) = P(x_{t+1} =$

$j|x_t = i, \alpha_{t+1})$; and a reward function $r : (X, \alpha) \to \mathbb{R}$ that assigns an instantaneous reward for being in each state. The reward function will be used to model our objective of reaching a goal state.

In what follows, to achieve desired equilibrium behavior and/or to optimize other objectives of the superplayer, we will focus on stationary policies $\pi : X \to A$ that map states to control actions [17].

B. Relating decision-making in a parametric game to an MDP

We will now recast the superplayer's decision-making in a parametric game $G = (N, S_{p,p\in N}, u_{p,p\in N}(s, \alpha))$ with fixed Markovian dynamics as a MDP $M = (X, A, T, r)$.

a) States: Unless otherwise stated, the states of the MDP are equal to the set of pure joint strategies, i.e. one state represents each possible joint strategy.

$$X = \left\{ \begin{bmatrix} s_{p_1} & \dots & s_{p_n} \end{bmatrix} \mid s_{p_i} \in S_{p_i} \right\}, \quad \forall s_{p_i} \in S_{p_i}. \quad (3)$$

Note that $|X| = \Pi_{i=1}^{n} |S_{p_i}|$.

b) Transition matrix: A transition matrix for each α must be specified to define the probability of transitioning from $i \to j$ for all states $i, j \in X$. Here the transition matrices are time-homogeneous as the game dynamics are fixed with respect to time. Denoting by $\{x_t\}$ and $\{\alpha_t\}$ the stochastic processes representing state and action evolution respectively, as x_t evolves as a controlled Markov process, the transition probabilities will only depend on x_{t-1} and α_t. Building upon the previous assumptions A1 and A2, we can restate the Markovian assumption as follows.

Assumption 3: The state transitions are conditionally Markovian, i.e.,

$$P(x_{t+1} = x|x_t, x_{t-1}, \dots, x_0, \alpha_{t+1}, \alpha_t, \dots, \alpha_1) = P(x_{t+1} = x|x_t, \alpha_{t+1}), \forall x \in X.$$

This will allow us to define the transition matrices $T(\alpha)$,

$$T_{ij}(\alpha) = P(x_{t+1} = j|x_t = i, \alpha). \quad (4)$$

c) Control actions: In an MDP the control actions are equivalent to the superplayer's actions. It may be useful to only allow a subset of actions be available at each state, i.e., the set of permissible actions is a subset $A(x) \subseteq A$.

d) Reward: The reward function must be specified to successfully drive the system to the desired NE or desired subset of states. This can be adjusted for criteria such as avoiding certain states (zero reward or high cost), minimizing the number of α transitions, or including a cost on α. Reward functions will be discussed further in Section IV-B.

C. Desired States

Before solving the MDP for the optimal superplayer policy, the superplayer will need to define their goal states as a set $D \subset X$. If their desired game-play is a NE for some α, then the desired set is the state associated with that NE. Recall that as there can be multiple NE for a given α, game-play at any NE cannot be guaranteed by fixing a single α, so it is necessary to specify the desired NE. If the superplayer instead wants game-play within some subset of states, then $D = \{x_j, \dots, x_k\}$. In either case α must be actively adjusted to encourage game-play within the desired set.

Given a goal set, a simple reward function can be an indicator for if the state x is in D,

$$r(x, \alpha) = \mathcal{I}(x \in D). \quad (5)$$

Here $\mathcal{I}(\cdot)$ stands for the indicator function.

D. Evaluating the MDP

Given the determined MDP, the superplayer needs a method of selecting α. This is done by evaluating the MDP for an optimal policy $\pi^* : X \to A$ that assigns an α^* for the superplayer to choose at each state. For this setup we will only consider stationary Markovian policies as they will assign an α to each state independent of any previous game-play. The MDP solves for these policies by maximizing the discounted reward, which will act as the superplayer's utility function,

$$V(\pi^*, x) = \max_{\pi} E_x^{\pi} \left[\sum_{t=0}^{\infty} \gamma^t r(x_t, \alpha_t) \right], \ x \in X. \quad (6)$$

Here $\gamma \in (0, 1)$ is the discount factor. For discrete state space, the policy π^* may be found through standard policy or value iteration.

Remark 1: Note that in the above MDP formulation, the class of admissible control policies is restricted to stationary Markovian policies. However, given our finite time-homogeneous state-action space model, stationary Markov policies are optimal even if the admissible class of policies are extended to include non-stationary or more generally history dependent policies [17].

IV. Main Results

The main results are organized within two sections (1) Reachability and (2) Game-play. A method for determining reachability of desired states will be presented in Theorem 1; this will inform the superplayer of which joint strategies could ever be played given any initialization and choices of α. In Theorem 2, a sample reward function will be presented such that the MDP solves for policies that guarantee game-play at a reachable goal set almost surely.

A. Reachability Analysis

First we will focus on determining the reachability of the system; this will be done by identifying the states that can eventually be played given any initialization. We will define reachability in terms of α.

Definition 5: If for every $i \in X, i \notin D$ there exists an integer t and some sequence $\{\alpha_t\}_{t\geq 1}$ such that $P(x_t \in D|x_0 = i, \alpha_1, \dots, \alpha_t) > 0$, we say that the set D is α-**reachable**. We also say that a state j is α-reachable from a specific state x if $P(x_t = j|x_0 = x, \alpha_1, \dots, \alpha_t) > 0$ for some t and $\alpha_1, \dots, \alpha_t$.

We will use the following concept to help determine α-reachability.

Definition 6: Define the composite adjacency matrix $\widetilde{T}$ as follows:

$$\widetilde{T} = \mathcal{B}\left(\sum_{\alpha} T(\alpha)\right). \tag{7}$$

Here the operation $\mathcal{B}$ is defined element-wise as $(\mathcal{B}(X))_{ij} = 1$ if $X_{ij} > 0$ and zero otherwise. Note that an entry T_{ij} of $\widetilde{T}$ is zero iff $T_{ij}(\alpha) = 0$ for all $\alpha \in A$, otherwise $T_{ij} = 1$.

Remark 2: If $j \in D$ is α-reachable in t steps from $i \in X$, then there exists a sequence $(x_0 = i, \alpha_1), \ldots, (x_t = j, \cdot)$ such that

$$P(x_1|x_0, \alpha_1) \times \cdots \times P(x_t = j|x_{t-1}, \alpha_t) > 0. \tag{8}$$

This implies that $(\widetilde{T})^t_{ij} > 0$.

Given the above, we now characterize if a desired set D is α-reachable. Theorem 1 presents a metric based on $\widetilde{T}$ that indicates the reachability of each state.

Theorem 1: Consider a parametric game setup that fulfills the dynamics assumption A3 and is represented as a Markov decision process. Let R be defined as follows:

$$R \triangleq \sum_{t=1}^{k-1} (\widetilde{T})^t, \tag{9}$$

where $k = |X|$. The set D is α-reachable if and only if for every $i \in X$, $i \notin D$, there exists some $j \in D$ such that $R_{ij} > 0$.

To prove this theorem, we will first present the following lemma.

Lemma 1: Let $T \in \mathbb{R}_+^{K \times K}$. For $i \neq j$, if $(T^t)_{ij} > 0$ for some $t \geq K$, then there exists a $\tau \in \{1, \ldots, K-1\}$ such that $(T^\tau)_{ij} > 0$.

Proof: First, note that by the Cayley-Hamilton Theorem, $T^K = \sum_{\tau=0}^{K-1} \beta_\tau T^\tau$, where each $\beta_\tau \in \mathbb{R}$. Thus, if $(T^K)_{ij} > 0$, then we must have $(T^\tau)_{ij} > 0$ for some $\tau \in \{0, \ldots, K-1\}$.

Now, let $t' \geq K$ and suppose (for the sake of induction) that for each $\hat{t} \in \{K, \ldots, t'\}$ we have that $(T^{\hat{t}})_{ij} > 0$ implies $(T^\tau)_{ij} > 0$ for some $\tau \in \{0, \ldots, K-1\}$. Then we claim that:

$$(T^{t'+1})_{ij} > 0 \Rightarrow (T^\tau)_{ij} > 0 \text{ for some } \tau \in \{0, \ldots, K-1\}. \tag{10}$$

To see this, note that by the Cayley-Hamilton Theorem we may write $T^{t'+1}$ as a polynomial function of lower order terms $T^{t'+1} = \sum_{\tau=t'+1-K}^{t'} \beta_\tau T^\tau$. Hence, if $(T^{t'+1})_{ij} > 0$ then we must have $(T^\tau)_{ij} > 0$ for some $\tau \in \{t'+1-K, \ldots, t'\}$. But by assumption, this implies that there exists some $\tau \in \{0, \ldots, K-1\}$ such that $(T^\tau)_{ij} > 0$. Hence, (10) holds.

Now note that since $i \neq j$, $(T^0)_{ij} = 0$, so there must exist some $\tau \in \{1, \ldots, K-1\}$ such that $(T^\tau)_{ij} > 0$. The desired result now follows by induction. ∎

We will now prove Theorem 1.

Proof: If D is α-reachable then $R_{ij} > 0$:

If D is α-reachable, then for every $i \notin D$ there exists some t and a sequence $\alpha_1, \ldots, \alpha_t$ such that $P(x_t \in D|x_0 = i, \alpha_1, \ldots, \alpha_t) > 0$. From Remark 2 we can see there exists a sequence of states $(x_0 = i, \alpha_1), \ldots, (x_t \in D, \cdot)$ such that:

$$P(x_1|x_0 = i, \alpha_1) \times \cdots \times P(x_t \in D|x_{t'-1}, \alpha_t) > 0 \tag{11}$$

This implies that $(\widetilde{T}^t)_{ij} > 0$.

By Lemma 1 this implies there exists a $\hat{\tau} \in \{1, \ldots, K-1\}$ such that $(\widetilde{T}^{\hat{\tau}})_{ij} > 0$. Since $\widetilde{T}$ is non-negative, we see that $R_{ij} = \sum_{\tau=1}^{k-1} (\widetilde{T}^\tau)_{ij} \geq (\widetilde{T}^{\hat{\tau}})_{ij} > 0$.

If $R_{ij} > 0$ then D is α-reachable:

Suppose, on the contrary, there is some $i \notin D$ and $j \in D$ such that $R_{ij} > 0$ but j is not α-reachable from i. If $R_{ij} > 0$ then there exists some $\widetilde{T}^t_{ij} > 0$ for $t \in \{0, k-1\}$; this implies that there is a sequence of α's such that

$$P(x_1|x_0 = i, \alpha_1) \times \cdots \times P(x_t = j|x_{t-1}, \alpha_t) > 0$$

But this in turn implies $P(x_t = j|x_0 = i, \alpha_1, \ldots, \alpha_t) > 0$, so j must be α-reachable from i.

Likewise if $R_{ij} = 0$, then $\widetilde{T}^t_{ij} = 0$ for all $t \in [0, k-1]$ because every $\widetilde{T}^t \geq 0$. Therefore $P(x_1|x_0 = i, \alpha_1) \times \cdots \times P(x_t = j|x_{t-1}, \alpha_1, \ldots, \alpha_t) = 0$ for all sequences of α and j must not be α-reachable from i. ∎

Theorem 1 informs the superplayer if the desired set D is reachable in the given model. If each state is associated with a specific joint strategy, then Theorem 1 additionally specifies which joint strategies could ever be played in the sequence of games.

It suffices that, assuming more than $k-1$ time steps are available, the superplayer only needs to evaluate the reachable states matrix from $t = 1, \ldots, k-1$ to characterize all possible α-reachable states. Note that if the dynamics specification includes some exploration aspect, i.e. players randomly choose an action with some probability, states may be trivially reachable.

Remark 3: Note that under strict best response dynamics, if there exists an α such that $P(x_t = x|x_{t-1} = x, \alpha) = 1$, then x represents a joint strategy that is a NE. This is exemplified in the previous example game $G2$.

Remark 4: If there exists α such that $P(x_{t+1} \in D|x_t \in D, \alpha) = 1$ then we can choose α such that the system always stays within D. If not then we cannot guarantee continuous game-play in D, but can instead actively adjust α that the system will continuously return to D. We will show an example of this in Section V.

B. Game-play

After the superplayer evaluates the reachability of the states $x \in D$, the next natural question is how to control α such that the game-play will go to a joint strategy in D. If it has been determined that D is α-reachable, then the superplayer knows game-play within D is possible, but they need to establish that the policies computed by the MDP will succeed in realizing game-play within D. Theorem 2 will present an example reward function such that π^* is guaranteed to drive the system to a desired state.

Theorem 2: Let G be a parametric game setup that fulfills A3. If the set D is α-reachable and the reward for each state-action pair is defined as $r(x, \alpha) = \mathcal{I}(x \in D)$, then for any

optimal policy $\pi^* : X \to A$ there exists, almost surely, a finite time τ such that $x_\tau \in D$.

Proof: We denote by $\{x_t^{\pi^*}\}_{t\geq 0}$ the stochastic process that results when the controls are generated using an optimal policy π^*. Now, note that, since π^* is a fixed stationary policy, the associated process $\{x_t^{\pi^*}\}_{t\geq 0}$ evolves as a time-homogeneous Markov chain (see [17]). Hence, to achieve the desired assertion, it suffices to show that for any initial state $x_0^{\pi^*} = x \in X$, there exists a (deterministic) finite time t such that $P\left(x_t^{\pi^*} \in D\right) > 0$; this is because, if the above holds, by standard concepts in recurrence of Markov chains (see [18]) it would follow that the hitting time to set D is almost surely finite.

We now show that for each initial state $x_0^{\pi^*} = x \in X$, there exists a (deterministic) finite time t such that $P\left(x_t^{\pi^*} \in D\right) > 0$. Without loss of generality, we assume that the initial state $x \notin D$, otherwise the assertion follows trivially.

To this end, suppose on the contrary, for some initial state $x_0^{\pi^*} = x \notin D$ and for all $t \geq 0$, $P\left(x_t^{\pi^*} \in D\right) = 0$. Now, recall that the discounted cost optimality equation for the MDP is defined as:

$$\max_{\pi} V(\pi, x) = \max_{\pi} E_x^{\pi}\left[\sum_{i=0}^{\infty} \gamma^i r(x_i, \alpha_i)\right].$$

Here the maximization is over all possible policies π, which may include time-varying and non-Markovian policies. From Remark 1 it is known that the optimal stationary policy π^* will be optimal even in this larger class of policies, i.e.,

$$\begin{aligned}\pi^* &= \arg\max_{\pi} V(\pi, x) \\ &= \arg\max_{\pi} E_x^{\pi}\left[\sum_{i=0}^{\infty} \gamma^i r(x_i, \alpha_i)\right].\end{aligned}$$

From the definition of the reward function we know there are two cases: if $x \in D$ then $r(x, \alpha) = 1$, or if $x \notin D$ then $r(x, \alpha) = 0$.

By the contradiction hypothesis, under the optimal policy π^*, the associated state process $\{x_t^{\pi^*}\}_{t\geq 0}$ satisfies $P\left(x_t^{\pi^*} \notin D\right) = 1$ for all t. This implies, under the optimal policy, $r(x^{\pi^*}, \alpha^*) = \mathcal{I}(x^{\pi^*} \in D) = 0$ almost surely, so the optimal value $V(\pi^*, x)$ must be equal to zero. Since the optimal policy maximizes the value function $V(\pi, x)$, then all policies π have a corresponding value $V(\pi, x) = 0$.

However, since D is α-reachable, we can construct a policy (possibly non-stationary) $\tilde{\pi}$ such that, if the controls are generated according to this policy, there exists a (deterministic) finite time $\acute{t}$ with the property that the associated state process $\{x_t\}$ satisfies $P(x_{\acute{t}} \in D) = \epsilon > 0$. We will calculate the value of this policy as follows.

$$\begin{aligned}V(\tilde{\pi}, x) &= E_x^{\tilde{\pi}}\left[\sum_{i=0}^{\infty} \gamma^i r(x_i, \alpha_i)\right] \\ &\geq \gamma^{\acute{t}} E_x^{\tilde{\pi}}[r(x_{\acute{t}}, \alpha_{\acute{t}})] \\ &= \gamma^{\acute{t}}[(1)\epsilon + (0)(1-\epsilon)] \\ &> 0\end{aligned}$$

Note the policy $\tilde{\pi}$ used above may be time-varying and non-Markovian. From Remark 1 it is known that the optimal stationary policy π^* will be optimal even in this larger class of policies. Hence, the above clearly contradicts that the value $V(\pi^*, x)$ associated with the optimal policy π^* is zero. We thus conclude that for each initial state $x_0^{\pi^*} = x \in X$, there exists a (deterministic) finite time t such that $P\left(x_t^{\pi^*} \in D\right) > 0$. The desired assertion now follows by invoking standard properties of time-homogeneous Markov chains as noted above.

■

Theorem 2 offers one example of a reward function such that use of the optimal α guarantees game-play within D.

Remark 5: Let $\mathcal{R}$ be the class of reward functions that produce π^* that are guaranteed to yield game-play within a desired set. The reward function presented in Theorem 2 of $r(x, \alpha) = \mathcal{I}(x \in D)$ is contained within $\mathcal{R}$. Future work can be done to better characterize larger classes of rewards in $\mathcal{R}$.

Obtaining the optimal policy allows for construction of a new matrix, T^*, where each row i is the transition probabilities from state x_i under $\alpha^*_{x_i}$. In effect, T^* is the effective transition probability matrix that results from always selecting α^* at every state. With the use of π^* and α^*, the MDP will appear as a Markov chain with the single transition matrix T^*.

Remark 6: It can be shown that if D is α-reachable and $P(x_{t+1} \in D | x_t \in D, \alpha) = 1$ for some value(s) of α, then the system can be controlled via π^* such that once the system enters a state within D, then the played joint strategy will always be at a state within D. If D is α-reachable but $P(x_{t+1} \in D | x_t \in D, \alpha) < 1$ for all α, then the system can be controlled via π^* such that it is infinitely often within D.

V. Hacking Example

In this section we use an academic example to illustrate the use of the parametric game model and its effectiveness.

Assume that a server is under attack by a botnet. The hacker (h) launches denial of service attacks of varying strength on the server, with the goal of disabling the server from processing any requests. The server (v) can choose to automatically reject a percentage of their received requests to defend from attacks, and their goal is to process as many requests as possible while also successfully blocking attacks.

The server is additionally monitored by the system administrator (m), who can activate a filter to block requests before they reach the server; this decreases the hack's strength at the cost of neglecting some valid requests. In this game assume the two players are the hacker and the server, and the superplayer is the system administrator. (For simplicity, assume all players have full knowledge of the system and all parameters.)

Let the hacker's action space be given by $S_h = \{0\%, 50\%, 100\%\}$, indicating the strength of their attack, where $1 - s_h$ is the percentage of requests the server is able to process when there is no defense and no filter active. Let the server's action space be given by $S_v =$

$\{0\%, 25\%, 50\%, 75\%, 100\%\}$. The cost incurred by players for each action is tabulated below.

s_v	0%	25%	50%	75%	100%
$c_v(s_v)$	0	0.05	0.11	0.18	0.32

s_h	0%	50%	100%
$c_h(s_h)$	0	0.06	0.13

We let the action of the administrator be denoted by $\alpha \in \{0, 1\}$, indicating the activation (or lack thereof) of the additional filter. This filter causes two effects: (1) the filter reduces the hacker's attack strength by some amount $b\%$, and (2) the filter erroneously blocks $e\%$ of the valid requests.

Define $w(s_v, s_h, \alpha)$ to be the proportion of requests the server processes, i.e.,

$$\begin{aligned} w(s_v, s_h, \alpha) =& \textit{server defense} \times \textit{filter error} \\ & \times \textit{attack strength} \\ =& (1 - s_v)(1 - e\alpha)(1 - (1 - b\alpha)s_h). \end{aligned}$$

For this example let the filter's blocking strength be $b = 0.66$ and the filter's error penalty be $e = 0.33$.

Note that under no attack and with no server defense then $w(s_v = 0, s_h = 0, \alpha = 0) = 1$ as the server is able to process all requests. If the server is under a strong attack without any defense, then $w(s_v = 0, s_h = 1, \alpha = 0) = 0$.

Define $y(s_v, s_h, \alpha)$ to be the proportion of the hacker's bogus requests that are blocked, i.e.,

$$\begin{aligned} y(s_v, s_h, \alpha) =& \textit{attacks blocked by filter} \\ & + \textit{attacks blocked by server} \\ =& b\alpha s_h + s_v(1 - b\alpha)s_h. \end{aligned}$$

Note that if server is blocking all requests under a strong attack, $y(s_v = 1, s_h = 1, \alpha = 0) = 1$. If instead the server blocks no attacks but the admin activates the filter, $y(s_v = 0, s_h = 1, \alpha = 1) = b$.

The server's first priority is to process requests, and its second priority is to block attacks. Reflecting this, its utility is defined as,

$$u_v(s_v, s_h, \alpha) = w(s_v, s_h, \alpha) + 0.5y(s_v, s_h, \alpha) - c_v(s_v).$$

The hacker cares about minimizing the requests processed by the server, so their utility is defined as,

$$u_h(s_v, s_h, \alpha) = 1 - w(s_v, s_h, \alpha) - c_v(s_v).$$

Define the dynamics as a modified strict best response. Each player evaluates their utility based on the last played joint action $u_p(s_p, s_{\bar{p},t-1}, \alpha_t)$; they choose the action that maximizes their utility with probability 75%, and they choose the action that yields their second highest utility with probability 25%. Note that these decisions only depend on s_{t-1} and α_t; thus these dynamics satisfy A3.

Next we will define the MDP model. We define one state for each possible joint action $X = \{[s] | \forall s \in S\}$ totaling 15 states. The set of control actions is equal to our set of α, thus $A = \{0, 1\}$. A transition kernel for each α is determined by calculating the transition probabilities from each state according to the dynamics and utilities.

Suppose that the admin's desired set D is the joint action where the server does not block any requests (0%) and the hacker does not attack (0% strength). Let this joint strategy be associated with state 1, thus $D = \{1\}$. Let the reward function be an indication on membership to D: $r(x) = 1$ if $x = 1$, else $r(x) = 0$.

We calculate $R = \sum_{t=1}^{k-1} (\widetilde{T})^t$ and find that state 1 is α-reachable from any initialization. An optimal policy is found though value iteration with $\gamma = 0.95$. The chart below shows the α^* associated with the optimal policy π^*.

	s_h 0%	s_h 50%	s_h 100%
s_v 0%	$\alpha^* = 0$	$\alpha^* = 0$	$\alpha^* = 0$
s_v 25%	$\alpha^* = 0$	$\alpha^* = 0$	$\alpha^* = 0$
s_v 50%	$\alpha^* = 1$	$\alpha^* = 1$	$\alpha^* = 1$
s_v 75%	$\alpha^* = 1$	$\alpha^* = 1$	$\alpha^* = 1$
s_v 100%	$\alpha^* = 0$	$\alpha^* = 0$	$\alpha^* = 1$

In simulation, we find that by fixing $\alpha = 0$ the system is at the desired state about 3% of the time, and with a permanent filter $\alpha = 1$ then the system is never at the desired state; however, by adaptively changing α according to the computed optimal policy this percentage increases to 17%. While the desired state is associated with a joint strategy that is not a NE, the system admin can intelligently activate the filter to increase the time spent in the desired state.

VI. Conclusions and Future Work

This work has demonstrated a class of games whereby a superplayer can influence players' decision-making such that some desired outcome occurs. We define parametric games, where player utilities are expressed as a function of the superplayer's input. For parametric games played with Markovian dynamics, the players' learning dynamics evolve according to the superplayer's control. This phenomenon can be modeled as a Markov Decision Process. Reachability criteria were defined to identify the set of joint strategies that may be played given any initialization and appropriate sequence of actions from the superplayer. Finally we presented a reward function for the MDP such that the calculated optimal policies guarantee game-play at a reachable desired state almost surely.

A strong assumption of the paper is that the superplayer knows the utility structures of the players. This is not the case in most applications. Future work will aim at developing tools to allow the superplayer to learn an unknown model of the players' utilities by testing their action space. Additional work will tackle controlling more general processes, possibly exogenous, that are dependent on the players' joint strategy, expand the class of reward functions that guarantee desired game-play, or introduce multiple competing superplayers.

References

[1] S. Parsons and M. Wooldridge, "Game theory and decision theory in multi-agent systems," *Autonomous Agents and Multi-Agent Systems*, vol. 5, no. 3, pp. 243–254, 2002.

[2] S. Roy, C. Ellis, S. Shiva, D. Dasgupta, V. Shandilya, and Q. Wu, "A survey of game theory as applied to network security," in *System Sciences (HICSS), 2010 43rd Hawaii International Conference on*. IEEE, 2010, pp. 1–10.

[3] C. Y. Ma, D. K. Yau, X. Lou, and N. S. Rao, "Markov game analysis for attack-defense of power networks under possible misinformation," *IEEE Transactions on Power Systems*, vol. 28, no. 2, pp. 1676–1686, 2013.

[4] G. C. Chasparis and J. S. Shamma, "Control of preferences in social networks," in *Decision and Control (CDC), 2010 49th IEEE Conference on*. IEEE, 2010, pp. 6651–6656.

[5] J. Ghaderi and R. Srikant, "Opinion dynamics in social networks: A local interaction game with stubborn agents," in *American Control Conference (ACC), 2013*. IEEE, 2013, pp. 1982–1987.

[6] D. Fudenberg, F. Drew, D. K. Levine, and D. K. Levine, *The theory of learning in games*. MIT press, 1998, vol. 2.

[7] J. R. Marden, H. P. Young, G. Arslan, and J. S. Shamma, "Payoff-based dynamics for multiplayer weakly acyclic games," *SIAM Journal on Control and Optimization*, vol. 48, no. 1, pp. 373–396, 2009.

[8] U. Bhaskar, K. Ligett, L. J. Schulman, and C. Swamy, "Achieving target equilibria in network routing games without knowing the latency functions," *Games and Economic Behavior*, 2018.

[9] X. Wang and T. Sandholm, "Reinforcement learning to play an optimal nash equilibrium in team markov games," in *Advances in neural information processing systems*, 2003, pp. 1603–1610.

[10] T. Alpcan and L. Pavel, "Nash equilibrium design and optimization," in *Game Theory for Networks, 2009. GameNets' 09. International Conference on*. IEEE, 2009, pp. 164–170.

[11] R. E. Ruelas, D. G. Rand, and R. H. Rand, "Nonlinear parametric excitation of an evolutionary dynamical system," *Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science*, vol. 226, no. 8, pp. 1912–1920, 2012.

[12] A. K. Chorppath and T. Alpcan, "Learning user preferences in mechanism design," in *Decision and Control and European Control Conference (CDC-ECC), 2011 50th IEEE Conference on*. IEEE, 2011, pp. 5349–5355.

[13] M. Jackson, "Mechanism theory," 2014.

[14] N. Nisan, T. Roughgarden, E. Tardos, and V. V. Vazirani, *Algorithmic game theory*. Cambridge University Press, 2007.

[15] J. Du, B. Wu, P. M. Altrock, and L. Wang, "Aspiration dynamics of multi-player games in finite populations," *Journal of the Royal Society Interface*, vol. 11, no. 94, p. 20140077, 2014.

[16] D. Monderer and L. S. Shapley, "Potential games," *Games and economic behavior*, vol. 14, no. 1, pp. 124–143, 1996.

[17] D. P. Bertsekas, *Dynamic programming and optimal control*. Athena scientific Belmont, MA, 2005, vol. 2, no. 3.

[18] E. Seneta, *Non-negative matrices and Markov chains*. Springer Science & Business Media, 2006.