Restless Contracting

Can Urgun*

November 2019

Abstract

I explore how a principal dynamically chooses among multiple agents to utilize for production without full commitment. The principal chooses at most one agent to utilize in every period affecting the states of the agents. A utilized agent changes its state because it is utilized, but the nonutilized agents do not remain at rest: they also change their state. The analysis requires a novel methodological approach: the agency problem that the principal faces with each agent is shown to be an appropriately designed restless bandit, thus creating a multiarmed restless bandit. The principal's optimal contract is characterized by an index rule for the restless bandit problem.

JEL-Classification: D21, D86, L14, L24

*Princeton University, e-mail: curgun@princeton.edu. I am grateful to Alvaro Sandroni, Niko Matouschek, Bruno Strulovici, Dan Barron, Mike Powell, Ehud Kalai, Jin Li, Elliot Lipnowski, Nick Buchholz and Doruk Cetemen for helpful comments and suggestions. I am also thankful to seminar participants at Northwestern, Chicago, Princeton, UCSD, London Business School, Indiana, Columbia, UIUC and Rochester. The usual disclaimer applies.

1 Introduction

Firms often maintain relationships with multiple trading partners to outsource production. To manage complex production needs, firms rely on both "just-in-time" spot contracts and informal promises of future business across multiple partners. The fear of losing future business or the threat of a trading partner going rogue can motivate both the outsourcing and the outsourced parties to keep their promises.

When selecting a trading partner, it is commonsensical that a firm considers the state of their potential trading partner, where the state might capture the benefits of immediate trade, the outside option of the partner, the potential loss the partner can inflict on the firm or any combination of the three. In addition, the act of trading or not trading might have an impact on the future state of the trading partner since the partner could potentially be more experienced, more tired or more valuable to the firm. The case of the allocation of many trading partners for future business entails an additional tradeoff as a promise to one partner is made at the expense of others. Hence, even if the relationships appear to be bilateral, they necessarily become intertwined.

This paper explores how a firm (principal) can dynamically choose which trading partner (agent) to utilize for production when utilization affects all the trading partners. In this paper, a principal repeatedly interacts with multiple agents, and the variation in the states of the agents across time is partially controlled by the utilization decisions of the principal. The principal chooses at most one agent to utilize in every period. The chosen agent changes his state in a particular fashion because he is utilized, while all the nonutilized agents change their state in a different fashion, which captures various economic phenomena, such as recovery from exhaustion, organizational forgetting or catching up to other obligations. In the general framework, a change in state affects the benefits from immediate utilization, the

outside option of the agent and loss to the principal should the agent leave the relationship in a given state. Shutting down different channels in this framework captures different economic phenomena. For example, if just the benefit of immediate trade changes, it is possible to capture learning or exhaustion from utilization, whereas nonutilization might capture resting or organizational forgetting. If only the outside option of the agent changes, it is possible to capture an employee obtaining better outside options by experience. Finally, if the loss to the principal is changing, it is possible to capture an upstream firm utilizing the experience from working for the downstream firm to learn the inner workings and trade secrets of the downstream firm and use those against the firm.

A more commonly explored approach with endogenous control of state transitions occurs when the states change only when an agent is utilized. In such a framework, nonutilized agents do not change their state, simplifying the problem, as only one agent changes his state while the remaining agents remain at "rest". However, such a framework cannot capture an agent getting tired and recovering their productivity, or a case of not working for the principal diminishing the outside option of the agent as he becomes more "rusty". When the agents change state in different forms based on utilization and nonutilization, they are never at rest; that is, they are *restless*. Note that restlessness can capture different economic phenomena than exogenous state transitions in every period as restlessness allows for *deliberate* choices; for example allowing an agent to recuperate, as opposed to assuming said agent recovers in every period with a fixed probability despite working or not working. On the other hand, models with exogenous state transitions are more suitable for *random shocks* to productivity or outside options and are also usually more tractable.

Despite the complexities in these economies, the principal optimal utilization schedule is achieved by a simple index rule and an accompanying payment rule. The index of an agent depends only on the current state of

the agent and captures the shadow value of utilizing the agent *whenever* his state is equal to the current level. The payments play a dual role of satisfying incentive constraints and being *embedded* into the index.

The simplicity of this policy reveals striking characteristics about the optimal contract. When making a utilization decision, the principal could potentially rely on many factors, including the entire history of the relationships, all the informal promises she made, or even calendar time. At the very least one could expect an elaborate scheme that depends on the states of *all* the agents. However, the index does not depend on these factors: it simply depends on the current state of an agent and the mechanics, i.e., the underlying law of motion governing the states of the agent in question.

The effect of utilization decisions changing the state without full commitment already poses some challenges, even with a single agent, as endogenous state transitions in a repeated interaction inherently change the so-called "promise keeping" constraints in equilibria. Since such constraints involve the "future", the standard methodologies that use dynamic programming and Markovian behavior are not readily exploitable. In general settings, additional state and co-state variables are necessary to obtain a recursive formation and recursive Lagrangians, as in Marcet and Marimon (2011) or Pavoni, Sleet, and Messner (2018); thus, an index solution, despite being intuitive at first, is actually surprising considering the constraints. The Markovian behavior observed here is an unexpected consequence of surplus maximization subject to dynamic enforcement constraints, which are common in many single-agent relational contracting problems beyond the current study, as shown in results in Rustichini (1998) and (Gozzi, Monte, and Tessitore 2018). Notably, even though the optimality of some Markovian behavior can be established, a characterization is an entirely different challenge as the problem is no longer a standard problem.

To characterize the optimal Markovian behavior, I show that a principal optimal contract in this game can be identified by index policies and the

principal's problem is a relaxed version of a nonstandard bandit problem where I build upon the Whittle (1988) index. Despite the various incentive frictions and complex relationships, the indices in this paper share some of the characteristics of the Gittins (1979) index, which was celebrated for its surprising simplicity. Indeed, the index here captures the time-normalized marginal returns to changing a policy, whereas the Gittins index captured the time-normalized average returns.

The framework relies on "bidirectionality" of state transitions that are dependent on utilization, that is, utilizing an agent and not utilizing an agent have opposite effects. Despite the reliance on bidirectionality, due to the freedom provided on the states, the methodology is broadly applicable to other scenarios, such as persistent capital investments, liquidity constraints that are tied to performance, and reputation build-up in different markets, and can be altered more in the case of a single agent. In the case of multiple agents, the restless bandit approach imposes greater difficulties as, unlike the Gittins index, the Whittle index for restless bandits is generically optimal only in a relaxed version of the problem. In that vein, one advantage of a contracting setup is the fact that payments form an integral part of the index calculations and optimality of the index policy can be achieved.

This paper is organized as follows. Section 1 continues with a short literature review. Section 2 describes the general framework and provides a preliminary characterization of contracting frictions. Section 3 delivers the optimal contract and the indices in a single agent setup and then extends the single-agent analysis to the full setup. Finally, section 4 concludes. All proofs not provided in the main text are in the appendix.

1.1 Related Literature

This paper builds on a large number of relational contracting papers, a vast literature that I do not survey here. Malcomson et al. (2010) provides an excellent survey.

The brief analysis in the single-agent setup addresses mainly endogenous state transitions in a relational contracting setup. The canonical reference is Levin (2003), although Thomas and Worrall (1988), Ligon, Thomas, and Worrall (2002) and Kwon (2016) also consider persistent states in a relational contracting environment. The main difference between these papers and the current one is the endogenous versus exogenous state transitions. As highlighted before, such an extension inherently captures different economic phenomena and requires different approaches.

A related strand of literature is on relational contracts with persistent private information. Such problems also inherently have persistence of states as there is information revelation through contracting. However a distinct feature is that the learning dynamics generate a particular and one-directional transitions. Thus the control, if present is limited to the speed of learning implied by the contract. Moreover the main focus is about separation or pooling of the persistent private information. Furthermore due to the learning dynamics there is always a fixed set of states (or single true knowledge) that the processes converge to regardless of the actions and that set of states is irreducible. Notable contributions in this strand include Halac (2012), Yang (2013) and Malcomson (2016). In contrast the restless structure here is intended to capture direction as well as speed of state transitions, there is no private information and the states are directly payoff relevant.

The full model explores dynamic work allocation across multiple agents. The classic reference for the multiagent contracting model is Levin (2002). However, the model is a complete contracting setup; thus, utilization is not fully dynamic. The effect of commitment is severe in multiple-agent setups and has been highlighted in Cisternas and Figueroa (2015). The references to fully dynamic work allocation are Board (2011) and Andrews and Barron (2013), which feature multiple agents in a relational contracting setting. The main difference is that control via utilization is absent in those settings, and the optimal contract is history-dependent in both. The full model with

multiple agents delivers the principal-optimal contract in a dynamic work allocation setup, highlighting the effect of control and its effect on history independence while simultaneously introducing bandits as a potential and tractable tool to analyze such settings.

The critical problem for the principal in both settings is to find an optimal utilization schedule despite the lack of an inherent recursive structure in the game. Bandit problems are also scheduling problems; thus, I build upon techniques in the bandit literature. From a methodological perspective, approaching the principal's problem as a bandit problem is different from the approach of canonical papers in relational incentive contracting, such as Levin (2003), Baker, Gibbons, and Murphy (2002), and Malcomson et al. (2010). Most of the literature utilizes the inherent recursion in repeated games, which provides a recursive characterization of the payoff space. The main advantage of a bandit approach is that it allows for an easily implementable policy in the absence of inherent recursion while still delivering the principal optimal behavior.

Forward-looking constraints with endogenous state transitions pose a technical challenge and usually require additional co-states and recursive Lagrangians. Marcet and Marimon (2011) and Pavoni, Sleet, and Messner (2018) provide the most general framework, although such an approach usually complicates the problem. For more particular setups, Rustichini (1998) and Gozzi, Monte, and Tessitore (2018) establish the existence of Markovian solutions, the former promising broader applications to relational contracting with a single agent.

Since this paper utilizes bandits, a technically close strand of literature is the experimentation literature. This is a vast literature that I do not survey here, but some notable contributions are Bolton and Harris (1999), Bergemann and Välimäki (1996), Keller, Rady, and Cripps (2005), Rosenberg, Solan, and Vieille (2007), Strulovici (2010) , Klein and Rady (2011), and Fryer and Harms (2017). A large portion of this literature uses standard

bandits (a notable exception that also utilizes restless bandits is Fryer and Harms (2017)) to answer questions of when to make a switch from experimentation to exploitation in various settings with beliefs about a project being the deciding factor of experimentation. This paper interprets the arms of a bandit as the agents themselves and thus introduces bandits as a potential framework for dynamic contracting. The "arms" have their own incentive constraints that must be satisfied, and the state reflects the commonly known state of an agent.

Within the bandit literature, this paper builds upon restless bandit problems. Gittins, Glazebrook, and Weber (2011) provides an excellent treatment of this literature, and Nino-Mora et al. (2001), Glazebrook, Nino-Mora, and Ansell (2002), Nino-Mora (2002), Glazebrook, Ruiz-Hernandez, and Kirkbride (2006), and Glazebrook, Hodge, and Kirkbride (2013) are notable contributions. Restless bandits are bandit problems where even the arms that are not operated continue to provide rewards and to change states, albeit at different rates. The pioneering work in this literature is Whittle (1988), where a heuristic index is derived based on a Lagrangian relaxation of the undiscounted problem. Papadimitriou and Tsitsiklis (1999) showed that general restless bandits are intractable, and even the indexability of the problem is difficult to ascertain. However, I show that, in this special case of bidirectional restless bandits, the intractability can be bypassed using a smaller and more tractable space of policies to calculate indices in closed form. In order to do accomplish this task, I build upon the work of Glazebrook, Hodge, and Kirkbride (2013).

Finally, as a generalization of bandit problems, this paper also utilizes some general existence results on Markov decision problems. Markov decision problems have a long tradition with an established literature that encompasses multiarmed bandits as a subfield. Notably, Blackwell (1965) is an important pioneering work, and Puterman (2014) provides a comprehensive treatment of the literature.

2 Model

2.1 Basic Setup

Suppose there are $N+1$ players, player 0, the principal (she), who interacts with N agents (he) in time periods $t \in \{0, 1, 2 \ldots\}$. In each period, the principal needs a single good that can either be supplied by one of the agents or produced by the principal herself. Producing the good by herself is normalized to a payoff of 0 for the principal. Each agent i has a state that effects their relationship with the principal in period t, denoted by $s_i^t \in S_i$, for a finite set $S_i \subset \mathbb{R}$. Let $\prod_{i=1}^{N} S_i = S$ and $s^t = (s_1^t, s_2^t \ldots s_N^t)$. All agents discount future payoffs with a common discount factor $\delta \in (0, 1)$. In each period t, the following events unfold:

1. Agents' states are realized $(s_1^t, s_2^t, \ldots s_N^t)$ and become publicly known.

2. The principal publicly offers a spot contract that specifies a set of payments $p^t = \{p_i^t\}_{i \in \{1,2,\ldots,N\}} \in \mathbb{R}^N$ and a source of production, either utilizing one of the agents or producing herself. Formally, let $\{I_i^t\}_{i \in \{0,1,2,\ldots,N\}} \in \{0,1\}^{N+1}$ denote the vector describing the principal's utilization choice with $\sum_{i=0}^{N} I_i^t = 1$ and $I_0 = 1$ denoting the principal producing herself.

   (a) Each agent simultaneously decides to accept ($d = 1$) the principal's offer or reject the offer ($d = 0$) and take their outside option, ending their relationship with the principal $d_k^t \in \{0, 1\}$, $k \in \{1, 2, \ldots, N\}$.

   (b) The principal pays all the agents as contractually obligated.

   (c) If agent j is chosen for utilization and agent j accepted, he produces a good that has value v to the principal; the principal covers the production cost $c_j(s_j^t)$.

(d) If the agent chosen for utilization has taken his outside option, then the principal produces by herself at a normalized payoff of 0.

(e) Each agent i that has taken their outside option leaves the game forever, earning a payoff $\rho_i(s_i^t)$.

(f) Each agent i that has taken their outside option inflicts a loss of $\gamma_i(s_i^t)$ on the principal.

3. Any agent who has produced (chosen and accepted) changes his state as if he is utilized, all other agents that are still in the game change their state as if they are not utilized.

4. The principal sends a public and costless signal $y_t \in S \times \mathbb{R}^N \doteq Y$ that is not payoff relevant.

5. Move on to $t+1$.

Throughout, I assume c_i, ρ_i, γ_i are real-valued functions. The public signal y_t serves no purpose other than simplifying the description of off path behavior for equilibria and can be completely dispensed with.

The principal's set of achievable payoffs depends on which agents remain, in addition to their states; hence, it useful to keep track of when and if an agent has decided to break off. Let T_i be defined as $T_i = \inf\{t \geq 0 : d_i^t = 0\}$ with the convention that $T_i = \infty$ if agent i never breaks off, and let $\mathbf{1}_i^t$ be the indicator function for T_i not having occurred by t. That is, $\mathbf{1}_i^t = 1 \Leftrightarrow T_i \not\leq t$. Given the setup, the payoffs in period t are given by:

$$u_0^t = \sum_{k=1}^{N} \mathbf{1}_k^t \left[\sum_{k=1}^{N} I_k^t \left[d_k^t \left[v - c_k(s_k^t) \right] \right] - \sum_{k=1}^{N} d_k^t p_k^t - \sum_{k=1}^{N} (1 - d_k^t) \gamma_k(s_k^t) \right]$$

$$u_k^t = \mathbf{1}_k^t \left(d_k^t p_k^t + (1 - d_k^t) \rho_k(s_k^t) \right)$$

The timeline identified here captures the scenario where a principal is deciding between outsourcing to a single agent or producing in-house. Based

on the choices of the functions c_i and ρ_i, the framework can capture different incentive frictions that might arise in different setups. Making the c_i functions constant while having variable ρ_i's captures cases where the agent states can capture how good/bad the agents are at diverting funds or how good they are at holding up the principal. Making the ρ_i functions constant while letting c_i's have a monotone structure could enable focusing on the scheduling aspect from the principal's perspective, where she faces agents having different levels of tiredness and loss of efficiency due to being tired. Letting both functions vary with some monotonicity in state can capture cases of learning via doing/organizational forgetting and being able to utilize the experience against the principal. Nonetheless, for the general solution, I put no such monotone structures on the functions themselves.

2.2 States and Transitions

One important part of the interaction above is the transitions of agents' states and how they are tied to whether the agent produces or not. In most relationships, it easy to imagine that producing and not producing have different effects on an agent. Agents might become tired, thereby increasing their costs, or they might be getting better at their jobs and hence decreasing their costs. Alternatively, agents might be getting more familiar with interaction with the principal, making it easier for them to subvert funds or, in an opposing case, the principal might be getting better at understanding their behavior and limiting their ability to subvert funds. Regardless of the particular change, it is commonsensical to tie such changes to utilization by the principal rather than leaving it exogenous. The law of motion across states and their interaction with the functions ρ_i and c_i are convenient tools to model different kinds of economic phenomena within this general framework. I first describe the sets of states and their laws of motion and then introduce some necessary assumptions.

Assumption 1 (Sets of States)**.** *For each i, $S_i = \{s_{i,1}, s_{i,2}, \dots s_{i,N_i}\}$ for some $N_i < \infty$.*

Assumption 1 is fairly self-explanatory. Each agent has a finite state space. Without loss of generality, I assume that the state space is ordered according to the "less than or equal to order" with $s_{i,1} \leq s_{i,2} \dots \leq s_{i,N_i}$. If another ordering is desired, it can be achieved simply by renaming the states and applying appropriate transformations of the functions ρ_i, c_i and γ_i since only finitely many values exist due to the discrete and finite state space.

Assumption 2 (Laws of Motion and Restlessness)**.** *For each agent i, there are two transition matrices that identify the laws of motion across states: $\mathbf{P}_i^a$ and $\mathbf{P}_i^p$ with $\mathbf{P}_i^a \neq \mathbf{P}_i^p$. If an agent i is chosen to be utilized for production ($I_i^t = 1$) and undertakes ($d_i^t = 1$) production, agent i changes states according to the transition matrix $\mathbf{P}_i^a$. If an agent i is not utilized for production ($I_i^t = 0$), agent i changes states according to the transition matrix $\mathbf{P}_i^p$.*

Assumption 2 has multiple implications describing the relationship between the principal and agents. First, observe that the assumption implies that by choosing who to utilize, the principal effectively controls the state transitions of all the agents since $\mathbf{P}_i^a \neq \mathbf{P}_i^p$. If the matrices were identical, the scenario would correspond to the case of exogenous state transitions. Note that the utilization decision and the following transition after utilization resemble a bandit problem, where an agent i could be viewed as corresponding to an "arm" that changes states according to $\mathbf{P}_i^a$ when utilized. However, unlike a standard bandit, instead of remaining in their current state (a.k.a. *resting*), agents that are not utilized also continue to change their states according to $\{\mathbf{P}_j^p\}_{\{j:I_j^t=0\}}$; hence, these agents are called *restless*. However, in contrast to the restless bandit problem, the rewards from pulling an arm are endogenously defined through the payments in this setup. Additionally, even though agents are analogous to arms, again unlike any bandit problem, pulling an arm has to be incentive compatible for the arm itself, which again must be accomplished by a suitable selection of payments.

From here onward, an agent is called *active* if he is chosen to be utilized and accepts the contract. Similarly an agent that is not chosen for utilization is called *passive* in that period.

Assumption 3 (Bidirectionality and Skip Free)**.** *For each i and each $k, l \in \{1, 2, \dots N_i\}$, the matrices $\mathbf{P}_i^a$ and $\mathbf{P}_i^p$ satisfy the following:*

1. $p^a_{i,(kl)} = \begin{cases} q_i(s_{i,k}) \in (0,1] \text{ if } l = k+1, l < N_i \\ 1 - q_i(s_{i,k}) \in [0,1) \text{ if } l = k, < l < N_i \\ 1 \text{ otherwise} \end{cases}$

2. $p^p_{i,(kl)} = \begin{cases} 1 \text{ if } l = k-1, l > 1 \\ 1 \text{ if } l, k = 1 \\ 0 \text{ otherwise} \end{cases}$

Assumption 3 puts a bidirectional structure on restlessness. Bidirectionality means that an agent being active or passive causes transitions in opposite directions of the state space. Notably, bidirectionality is defined with respect to an order (weakly smaller than, in this case) but does not impose substantial restrictions on the incentive frictions. In particular, the ordering impacts the state transitions, but the functions ρ_i and c_i need not be monotone with respect to the states.

The first part of the assumption states that an active agent either remains in the current state or goes *up* in step sizes of at most one, that is, without *skipping* any states, with a probability that depends on the state and the agent. The skip-free assumption is analogous to a differentiability (in one direction) assumption in a continuous-state setup.

The second part of the assumption states that a passive agent goes *down* with certainty, again with a step size of one. Such a state transition is relatively restrictive but is still able to capture a broad range of economic phenomena. For example, the states could capture the tiredness level of an agent, where an active agent becomes increasingly more tired and a passive

agent becomes less tired in a gradual fashion. Similarly, the transition could capture dynamics such as learning by doing and organizational forgetting, where only an active agent can become more experienced, and a passive agent will suffer from organizational forgetting. As stated, Assumption 3 is a fairly strong assumption to eventually accommodate the presence of multiple agents. In the case of a single agent, the return to lower states could be relaxed to accommodate downward jumps of any size but would require a more complicated payment scheme.

The three assumptions together enforce a structure on how each agent transitions through states by the utilization decisions, yet no restrictions are applied jointly. The structure allows each agent to have a completely unique state space and different state transitions while respecting bidirectionality, and the functions c_i and ρ_i could be completely different for each agent.

Assumption 4 (No-Strings at Initial States and Loss From Breaking Off)**.** *For each* i $\gamma_i(s_{i,1}) = \rho_i(s_{i,1}) = 0$ *and* $\gamma_i(\cdot) \geq \rho_i(\cdot) \geq 0$

Assumption 4 implies that, in their initial states, the agents have a normalized outside option of 0 and cannot inflict any costs on the principal by leaving. That is, there are no proverbial strings attached in the initial states. This rules out cases where all the agents quit immediately and is meant to capture the scenario where having been utilized for production provides a positive outside option for the agent, as well as holding some intrinsic value for the principal. The loss to the principal being larger than the gain to the agent means that an agent breaking off and taking his outside option represents a loss in surplus, which is sufficient to rule out entry but not necessary. The cases where the latter part of the assumption is violated is also of interest but would introduce optimal stopping into the surplus maximization problem which would cause significant technical challenges and have to be tackled via quasi-variational inequalities as opposed to the simpler bandit approach.

2.3 Strategies and Equilibria

Letting $\{s_i^t\}_{i\in\{1,2,\ldots,N\}}$ denote the states of each agent at the *end* of period t (that is, after the state transition), the history for period t, h_t, which is observed by all players, is given by

$$h_t = \{\{I_i^t\}_{i\in\{0,1,\ldots N\}}, \{p_i^t\}_{i\in\{1,2,\ldots,N\}}, \{d_i^t\}_{i\in\{1,2,\ldots,N\}}, \{s_i^t\}_{i\in\{1,2,\ldots,N\}}, y_t\}.$$

Let $h^t = \{h_n\}_{n=0}^{t-1}$ be a history path at the beginning of period t, and let $h^0 = \{\{s_{i,1}\}_{i\in\{1,2,\ldots,N\}}\}$ be the initial history with all agents starting in their respective smallest states. Let $H^t = \{h^t\}$ be the set of histories until time t, and let $H = \cup_t H^t$ denote the set of histories. At the beginning of each period t, conditional on h^t, the principal decides on $\{p_i^t\}_{i\in\{1,2,\ldots,N\}} \in \mathbb{R}^N$ and $\{I_i^t\}_{i\in\{0,1,2,\ldots,N\}} \in \{0,1\}^{N+1}$ with the restriction that $\sum_{i=0}^{N} I_i^t = 1$. The principal's choice is publicly observed. Conditional on h^t and the principal's action in period t, each agent decides on d_i^t. The principal's strategy is a sequence of mappings from histories to her set of feasible actions, denoted by $\{\sigma_0^t\}_{t\in\mathbb{N}} : H^t \to \mathbb{R}^N \times \{0,1\}^{N+1} \times Y$: the full sequence is denoted by σ_0. Agents' strategies are sequences of mappings from histories and the principal's actions to their sets of feasible actions, denoted by $\{\sigma_i^t\}_{t\in\mathbb{N}} : H^t \times \mathbb{R}^N \times \{0,1\}^{N+1} \to \{0,1\}$ for $i \in \{1,\ldots,N\}$. Again, the full sequence is denoted by σ_i. Mixed strategies are defined in the usual manner. Denote Σ_0 and Σ_i, $i \in \{1,\ldots,N\}$ as the sets of strategies. Let $v_0(\sigma_0, \{\sigma_i\}_{i\in\{1,2,\ldots,N\}}|h^t)$ and $v_i(\sigma_0, \{\sigma_i\}_{i\in\{1,2,\ldots,N\}}|h^t)$, $i \in \{1,\ldots,N\}$ denote, respectively, the principal's and agents' expected utilities following a history h^t conditional on a profile of strategies $\sigma = (\sigma_0, \sigma_1, \ldots, \sigma_N)$. A profile of strategies and the one step transition matrices identified induce a probability measure P with expectation operator $\mathbb{E}$ so at the beginning of any period t, the expected payoffs to the principal (0) and agents $i \in \{1,2,\ldots,N\}$ over the infinite horizon are given by:

$$v_0^t = \mathbb{E}\left[\sum_{\tau=t}^{\infty} \delta^{\tau-t}\left(\sum_{k=1}^{N} \mathbf{1}_k^\tau \left[\sum_{k=1}^{N} I_k^\tau \left[d_k^\tau \left[v - c_k(s_k^\tau)\right]\right] - \sum_{k=1}^{N} d_k^\tau p_k^\tau - \sum_{k=1}^{N}(1-d_k^\tau)\gamma_k(s_k^\tau)\right]\right)\right]$$

$$v_i^t = \mathbb{E}\left[\sum_{\tau=t}^{\infty} \delta^{\tau-t}\left(\mathbf{1}_k^\tau \left(d_k^\tau p_k^\tau + (1-d_k^\tau)\rho_k(s_k^\tau)\right)\right)\right]$$

Since there is no hidden information, the equilibrium concept is subgame perfect equilibrium (SPE). A SPE is a strategy profile σ such that the strategy profiles following any history form a Nash equilibrium following that history. In particular, a strategy profile is an SPE, where for each history h^t,

$$\sigma_0 \in \arg\max_{\tilde{\sigma}_0 \in \Sigma_0} v_0(\tilde{\sigma}_0, \{\sigma_i\}_{i\in\{1,2,\ldots,N\}}|h^t),$$

$$\sigma_i \in \arg\max_{\tilde{\sigma}_i \in \Sigma_i} v_i(\tilde{\sigma}_i, \sigma_0, \{\sigma_j\}_{j\neq i\in\{1,2,\ldots,N\}}|h^t) \text{ for all } i$$

I denote the set of achievable SPE payoffs by $\mathcal{E}$. It is important to emphasize that $\mathcal{E}$ depends on the initial states of the agents; however, in this setup, I suppress the dependence since I assume that all agents i start at their respective first state $s_{i,1}$. Furthermore, observe that any strategy profile σ induces a sequence of controlled Markov chains over S_i for each i. However, due to assumption 3, these controlled Markov chains are not necessarily irreducible; hence, without restricting the strategy space, there is no a priori guarantee of a recursive representation of the set of continuation payoffs.

2.4 Early Analysis and Constraints

A relational contract is a profile of strategies σ that constitute an SPE of the repeated game. An *optimal relational contract* is a relational contract that maximizes the principal's payoff at the beginning of the game. Given assumption 4, I restrict attention to contracts where no agent breaks off on path, that is, $d_i^t = 1$ for all i and for all t on path, notice that an agent

can still be non-utilized indefinitely without breaking off. Since the signal sent by the principal is payoff irrelevant and the principal can convey any information credibly by her offers, on the equilibrium path I will assume that $y_t = (s^t, p^t)$ that is she just repeats what is publicly known and all players ignore the signal on the equilibrium path.

In order to proceed with a characterization of an optimal relational contract, it is important to pin down the punishment payoffs for each player. Since the agents do not directly interact with each other, I assume the agents cannot cooperate to punish the principal; that is, the punishments are bilateral.

The agents do not interact directly other than observing the public history; hence, I also assume that the agents cannot punish each other. In particular, the only interaction an agent has is whether to remain with the principal or take their outside option at any point in time; hence, the incentive constraint of an agent i takes the following form:

Lemma 1. *An optimal relational contract is incentive compatible for agent i if*

$$\mathbb{E}\left[\sum_{\tau=t}^{\infty} \delta^{\tau-t} p_i^{\tau}\right] \geq \rho_i(s_i^t) \text{ for all } t \qquad (IC_i)$$

Furthermore observe that in case the principal deviates against a single agent, then she can just keep relationships with the other agents unchanged by replacing the utilization of the agent with self utilization which has a normalized payoff of 0. Given the bilateral punishment and the ability to utilize herself, the principal's incentive constraint for not deviating reduces to the following:

Lemma 2. *An optimal relational contract is incentive compatible for the*

principal if

$$\mathbb{E}\left[\sum_{\tau=t}^{\infty}\delta^{\tau-t}I_i^\tau\left[v-c_i(s_i^\tau)\right]-p_i^\tau\right]\geq-\gamma_i(s_i^t)\ \textit{for all } t\ \textit{and for all } i\quad(IC_0^i)$$

Notice that since the principal always has the ability to produce herself at a normalized utility of 0, any offer that is expected to be turned down can just be replaced by the principal offering to produce herself. Thus without loss I will restrict attention to equilibria in which no offer of the principal is turned down on the equilibrium path. With the restriction that no offer of the principal is turned down on the equilibrium path, an optimal relational contract is a solution to the following problem, which I call the principal's problem:

$[Principal's\ \ Problem]$

$$\max_{\{I_i^t\},\{p_i^t\}}\mathbb{E}\left[\sum_{\tau=0}^{\infty}\delta^\tau\left(\sum_{k=1}^{N}I_k^\tau\left[v-c_k(s_k^\tau)\right]-\sum_{k=1}^{N}p_k^\tau\right)\right]$$

subject to

$$\sum_{l=0}^{N}I_k^t=1\text{ for all }t$$

$$IC_0^i\text{ for all }i$$

$$IC_i\text{ for all }i$$

3 Analysis

In this section, I first explore a setting in which there is a single agent. I characterize the payment scheme and transform the single-agent contracting problem to a single-arm restless bandit problem using the payment scheme. Then, I utilize the results of Glazebrook, Hodge, and Kirkbride (2013) for the optimality of the index policy. After the analysis of the single agent,

I strengthen assumption 3 and consider the multiagent setup. In the multiagent setup, I first use a relaxation and show that the relaxed problem decouples into single-agent problems and then show that the index policy from the single-agent problem is feasible in the nonrelaxed version to show the optimality of the index policy.

3.1 Single-Agent Analysis

Before delving into the full problem, first let me characterize the solution when there is a single agent i. In this case, only two incentive constraints exist, and the principal's problem reduces to the following:

$$
\begin{aligned}
&[\textit{Principal's Single Agent Problem}] \\
&\max_{\{I_i^t\},\{p_i^t\}} \mathbb{E}\left[\sum_{\tau=0}^{\infty} \delta^\tau \left(I_i^\tau \left[v - c_i(s_i^\tau)\right] - p_i^\tau\right)\right] \\
&\text{subject to} \\
&\qquad \mathbb{E}\left[\sum_{\tau=t}^{\infty} \delta^\tau I_i^\tau \left[v - c_i(s_i^\tau)\right] - p_i^\tau\right] \geq -\gamma_i(s_i^t) \text{ for all } t \\
&\qquad \mathbb{E}\left[\sum_{\tau=t}^{\infty} \delta^{\tau-t} p_i^\tau\right] \geq \rho_i(s_i^t) \text{ for all } t
\end{aligned}
$$

Definition 1 (Surplus of i-Dyad and Continuation Payoffs with a Contract)**.** *For any incentive compatible relational contract $\{I_i^t\}, \{p_i^t\}$ the i-dyad surplus at period t is defined as:*

$$\mathcal{S}_i^t = \mathbb{E}\left[\sum_{\tau=t}^{\infty} \delta^\tau \left(I_i^\tau \left[v - c_i(s_i^\tau)\right]\right)\right]$$

The principal's total payoff from i in period t is defined as

$$\mathcal{U}_{0,i}^t = \mathbb{E}\left[\sum_{\tau=t}^{\infty} \delta^\tau I_i^\tau \left[v - c_i(s_i^\tau)\right] - p_i^\tau\right]$$

The agent's total payoff in period t is defined as

$$\mathcal{U}_i^t = \mathbb{E}\left[\sum_{\tau=t}^{\infty} \delta^{\tau-t} p_i^\tau\right]$$

The i-dyad surplus is the sum of the utilities of the principal and the agent arising from a relational contract starting from period t, the principal's total payoff from i is the portion of the profits of the principal from her relationship with agent i, and the agent's total payoff is just his continuation payoff under the relational contract.

Lemma 3. *Suppose there exists a relational contract that generates a surplus $\mathcal{S}_i^t \geq \rho_i(s_i^t) - \gamma_i(s_i^t)$ for all t. Then, any pair of total payoffs $\mathcal{U}_{0,i}^t, \mathcal{U}_i^t$ such that $\mathcal{U}_{0,i}^t \geq -\gamma_i(s_i^t)$ and $\mathcal{U}_i^t \geq \rho_i(s_i^t)$ with $\mathcal{U}_{0,i}^t + \mathcal{U}_i^t = \mathcal{S}^t$ can be implemented in a relational contract.*

Proof. Consider the relational contract that generates surplus $\mathcal{S}_i^t$ at period t but delivers payoffs $\tilde{\mathcal{U}}_{0,i}^t, \tilde{\mathcal{U}}_i^t$. Without loss of generality, assume $\tilde{\mathcal{U}}_{0,i}^t > \mathcal{U}_{0,i}^t$ Since the contract is relational, it must be the case that $\tilde{\mathcal{U}}_{0,i}^t \geq -\gamma_i(s_i^t)$ and $\tilde{\mathcal{U}}_i^t \geq \rho_i(s_i^t)$, but keeping the rest of the contract as is and increasing p_i^t by $\tilde{\mathcal{U}}_{0,i}^t - \mathcal{U}_i^t$ does not affect any future incentives compared to the original contract that generates $\mathcal{S}_i^t$ and remains incentive compatible for both the principal and the agent at period t and, hence, is a relational contract that delivers $\mathcal{U}_{0,i}^t, \mathcal{U}_i^t$. □

Lemma 3 is similar to the results in Levin (2003) and Kwon (2016) in an even simpler incentive setting. Since the payments are contractual, the principal can freely transfer utility from herself to the agent or vice versa by

adjusting the payment p_t. The lemma is used in an identical fashion to Levin (2003) and Kwon (2016) to restrict attention to surplus maximization.

Observe that the two incentive constraints can be combined to obtain the dynamic enforcement constraint:

$$\mathbb{E}\left[\sum_{\tau=t}^{\infty}\delta^{\tau-t}\left(I_i^{\tau}\left[v-c_i(s_i^{\tau})\right]\right)\right] \geq \rho_i(s_i^t)-\gamma_i(s_i^t) \tag{DE_i}$$

Due to lemma 3, the single-agent problem is equivalent to maximizing the $i-dyad$ surplus subject to the dynamic enforcement constraint DE_i. Hence, the problem is equivalent to:

$$\max_{\{I_i^t\}} \mathbb{E}\left[\sum_{\tau=0}^{\infty}\delta^{\tau}\left(I_i^{\tau}\left[v-c_i(s_i^{\tau})\right]\right)\right] \tag{SP_i}$$

subject to

$$\mathbb{E}\left[\sum_{\tau=t}^{\infty}\delta^{\tau-t}\left(I_i^{\tau}\left[v-c_i(s_i^{\tau})\right]\right)\right] \geq \rho_i(s_i^t)-\gamma_i(s_i^t) \text{ for all } t \tag{DE_i}$$

Note that DE_i is a forward-looking constraint with endogenous state transitions, so we cannot proceed along the lines of Kwon (2016).

Proposition 1. *The surplus maximization problem SP_i subject to the dynamic enforcement constraint DE_i is solved optimally by a Markovian policy.*

Proof. The proposition is a straightforward application of (Rustichini 1998) for incentive-constrained problems where the incentive constraint is reliant on the continuation utility being greater than a state-dependent value. Observe that the set of available actions including mixtures is $[0,1]$, which is independent of the state at every point in history and is compact valued. The transition probability is continuous with respect to the action. The aggregator for the utility is the discounted sum; hence, it is separable after the first period, first-period separable, stationary and strictly increasing in

future utility, and biconvergent. Finally, $\rho_i - \gamma_i$ does not depend on the action taken. Hence, theorem 3.6 in Rustichini (1998) applies to deliver the existence of an optimal Markovian policy. □

According to proposition 1, there is an optimal solution $I_i^t(s_i^t)$ that is dependent on only the current state of the agent.

3.1.1 Markovian Behavior in Single-Agent Problems

Note that due to the binary nature of the utilization decision, any Markov policy is simply a partitioning of the state space, where in one partition, the agent is active and in the other partition, the agent is passive.

Proposition 2. *Under assumptions 1, 2 and 3, any Markovian policy is equivalent to a threshold policy identified by a threshold state $\bar{s}_i$. Only the threshold and the state immediately below are recurrent; all other states are transient.*

Note that under assumption 3, if the agent is passive in the initial state, he remains passive forever, and the initial state is the threshold. If the agent is active in the initial state, he will continue going up in states until he reaching the smallest element of the passive set. Once the passive set is reached, the agent will become passive and return to the last active set and cycle between these two states forever.

This threshold structure is one of the investigated sufficient conditions for the indexability of restless bandits ((Niño-Mora 2006), (Glazebrook, Hodge, and Kirkbride 2013)) since the necessary conditions are not generally known. In the scenario of contracting, these small cycles around the threshold serve a secondary purpose of pinning down the payment structure for incentive compatibility.

3.1.2 Active-Passive Payments

Note that for any Markov utilization policy, since there are no limited liability constraints and payments are allowed even when agents are not utilized, there are multiple ways to maximize the principal's profits. Furthermore, the specific threshold nature can be utilized to shape the payment schedule as desired. Below, I introduce a particularly useful method that identifies a sequence of payments that are optimal for any threshold policy.

Definition 2 (Active-Passive Payments)**.** *For any agent i and any state $s_{i,k} \in S_i$, active-passive payments are defined as*

$$p_i^a(s_{i,k}) = (1 - \delta(1 - q_i(s_{i,k})))\rho_i(s_{i,k}) - \delta q_i(s_{i,k})\rho_i(s_{i,k+1})$$
$$p_i^p(s_{i,k}) = \rho_i(s_{i,k}) - \delta\rho_i(s_{i,k-1})$$

With the convention that $p_i^p(s_{i,1}) = 0$.

The active-passive payments identify two potential payments for each state: when the agent is active at state $s_{i,k}$, the agent is paid the active payment $p_i^a(s_{i,k})$ corresponding to that state; when the agent is passive at state $s_{i,k}$, the agent is paid the passive payment $p_i^p(s_{i,k})$ corresponding to that state. At state $s_{i,1}$, due to assumption 4, there is no immediate threat that the agent can use; similarly, there is no cost to having the agent quit. Hence, for the case where the formula does not apply, the passive payments must be exactly 0.

Proposition 3. *For any agent i and any threshold level $\bar{s}_i \in S_i$, active-passive payments are optimal.*

The point of the proposition is slightly subtle. Because of the relatively simple incentive friction, there are multiple ways to achieve optimality for a given threshold. Active-passive payments, on the other hand, are optimal for *any* threshold. In particular, for any threshold policy, the agent's incentive

constraint holds with equality at every reachable history. This particular construction is specific to the form of bidirectionality assumed in assumption 3, where jumps are skip-free in both directions. In general, due to the technical limitations associated with restless bandits, one side being skip-free has been identified as a sufficient condition for indexability in Glazebrook, Hodge, and Kirkbride (2013) and cannot be dispensed with without altering the tractability of the setup. However, just having one side skip free would require a different and more complicated payment scheme that would be difficult to generalize to multiple agents.[1]

The key ingredient in moving from a generic Markov decision problem to a restless bandit problem is appropriate choice of payments. The multiplicity of the potential payment schemes might seem like a problem but such multiplicity allows "construction" of bandits as the payments are integral part of the returns from an agent. The freedom to "construct" bandits is useful in tackling the intractability issues related with restless bandits.

3.1.3 Transforming the Single-Agent Problem into a Restless Bandit Problem

Active-passive payments satisfy the agent's incentive constraint at every point in history for every threshold policy and, therefore, every Markovian policy. Hence, the principal's problem with a single agent can be reduced to an optimal utilization problem (via an optimal threshold) assuming that the principal will have to pay the appropriate passive and active payments to the agent. To reformulate the problem in this manner, let me first introduce the returns from agent i with active-passive payments.

Definition 3 (Returns with Active-Passive Payments)**.** *The return from agent i is the net profit from agent i when the agent is paid according to*

[1]Some simple forms of one-side skip free would be easy to accommodate such as resetting to initial state when inactive, moving one step when active. However, more general laws of motion could potentially require very intricate constructions that needs to be solved on a case by case basis.

active-passive payments. In particular, for each state $s_{i,k}$, there is a passive $R_i^p(s_{i,k})$ and an active return $R_i^a(s_{i,k})$ defined as follows:

$$R_i^p(s_{i,k}) = -p_i^p(s_{i,k})$$
$$R_i^a(s_{i,k}) = v - c_i(s_{i,k}) - p_i^a(s_{i,k})$$

Now, observe that due to the presence of γ_i, a sufficient condition for the principal's incentive constraint to be satisfied is that the principal's continuation payoff is positive at every point in history. Hence, we can introduce the restless bandit formulation of the principal's single-agent problem as follows:

$$[Principal's\ \ Single\ \ Restless\ \ Bandit\ \ Problem]$$
$$\max_{\{I_i^t\}} \mathbb{E}\left[\sum_{\tau=0}^{\infty} \delta^\tau \left(I_i^\tau R_i^a(s_i^\tau) + (1 - I_i^\tau) R_i^p(s_{i,k})\right)\right]$$

Note that this problem has an optimal solution in Markovian policies, and since $I_i^t = 0$ for all t is a feasible solution that yields exactly 0 returns, it must be the case that the principal's incentive constraint is satisfied. Under assumptions 1, 2, and 3, the restless bandit in the single-agent problem is a finite-state bidirectional restless bandit that is skip-free in at least one direction. Therefore, according to Glazebrook, Hodge, and Kirkbride (2013), it is indexable without any requirement on the returns from activity or passivity and can be solved optimally by the Whittle index policy. Instead of introducing the index directly, let me introduce two related definitions to highlight the economic intuition of the index, originally introduced in Niño-Mora (2007), adapted to the threshold setting.

Definition 4 (Reward Measure with Thresholds)**.** *The reward measure with threshold $s_{i,j}$ starting from $s_{i,k}$, denoted $f_i^k(j)$, is the sum of the expected*

discounted rewards with a threshold $s_{i,j}$ when the initial state is $s_{i,k}$.

$$f_i^k(j) = \mathbb{E}\left[\sum_{\tau=0}^{\infty} \delta^\tau (I_i^\tau R_i^a(s_i^\tau) + (1 - I_i^\tau) R_i^p(s_i^\tau)) | s_i^0 = s_{i,k}; I_i^t = 1 \Leftrightarrow s_i^t < s_{i,j}\right]$$

Definition 5 (Work Measure with Thresholds)**.** *The work measure with threshold $s_{i,j}$ starting from $s_{i,k}$, denoted $g_i^k(j)$, is the sum of expected discounted utilization with threshold $s_{i,j}$ when the initial state is $s_{i,k}$.*

$$g_i^k(j) = \mathbb{E}\left[\sum_{\tau=0}^{\infty} \delta^\tau (I_i^\tau) | s_i^0 = s_{i,k}; I_i^t = 1 \Leftrightarrow s_i^t < s_{i,j}\right]$$

The first measure is the expected discounted rewards from a threshold policy. The reward measure can be identified for any activation policy and captures the profits that the principal receives subject to paying the agent with active-passive payments. For standard bandits, the reward measure with respect to the optimal stopping policy is the numerator of the celebrated Gittins index. The second measure is the expected discounted time the agent is utilized with a threshold policy. Again, in principle, this measure can be identified for any policy, not only threshold policies. For standard bandits, the work measure with respect to the optimal policy is the denominator of the Gittins index. With the two measures defined, the index of state $s_{i,k}$ is defined as follows.

Definition 6 (Index of State $s_{i,k}$)**.** *The index of state $s_{i,k}$, denoted by $\lambda_i(s_{i,k})$, is equal to:*

$$\lambda_i(s_{i,k}) = \frac{f_i^k(k+1) - f_i^k(k)}{g_i^k(k+1) - g_i^k(k)}$$

The index identified here is the so-called Whittle index of state $s_{i,k}$ that captures the work normalized marginal gains to increasing the threshold from $s_{i,k}$ to $s_{i,k+1}$. Nino-Mora calls the index the "Marginal Productivity Index" in

a series of works (Nino-Mora 2002, Niño-Mora 2006, Niño-Mora 2007, Niño-Mora and Villar 2011), as the index can be viewed as the "shadow price" of the policy, that is, the gains from changing a *policy* only marginally, albeit the margin is on the "thresholds". Indeed, the original interpretation of Whittle (1988) is from Lagrangian relaxation of the multiagent problem, where the indices captured are exactly the shadow prices of a policy in this relaxed problem. Another interpretation common in the literature of restless bandits, first proposed by Whittle (1988), is obtained by considering a hypothetical situation where, in addition to the passive returns $R_i^p(\cdot)$, the principal also receives a subsidy λ whenever the agent is not utilized. As the level of λ changes, it is conceivable that the optimal threshold changes. If the optimal threshold changes monotonically, then the restless bandit is indexable. The level of lambda that makes both thresholds $s_{i,k}$ and $s_{i,k+1}$ optimal is the index. Indeed, this λ subsidy problem is the basis of the multiple-agent problem, as will be shown in the next section.

Theorem 1. *Under assumptions 1, 2, and 3, in the problem with only agent i, a principal optimal contract is as follows:*

1. *Agent i is paid according to active-passive payments.*

2. *At each period t, agent i at state s_i^t is utilized if and only if $\lambda_i(s_i^t) \geq 0$.*

The optimality of the index policy in a single-arm, indexable, restless bandit problem is established in Nino-Mora et al. (2001). According to proposition 3, the incentive conditions of the agent are satisfied exactly, so the principal is giving away minimal rents. An important observation about the actual path of play is that in case of a single agent if the indices are both negative and positive then in the long run the agent will be cycling between two states due to the bi-directional law of motion. The principal will employ the agent repeatedly until the first time the index is negative then will produce herself once the index becomes negative. However, once the principal produces herself the agent will return to the previous state with the positive

index and will cycle indefinitely from that point onward. If the index is negative for all states then the agent is never utilized, if the index is positive for all states then the agent is utilized at every period on the path of play. Finally, the index being positive implies that the profits increase in every state by moving the threshold up compared to those achieved by never utilizing the agent. Therefore, the principal's profits are always positive, which implies they are larger than the outside option of $-\gamma_i(s_i^t)$.

3.2 Multiple-Agent Analysis

In order to address multiple agents, first I strengthen assumption 3 to the following assumption:

Assumption 3' (Bidirectionality and Two-Skip Free)**.** *For each i and for each $k, l \in \{1, 2, \dots N_i\}$, the matrices $\mathbf{P}_i^a$ and $\mathbf{P}_i^p$ satisfy the following:*

1. $p^a_{i,(kl)} = \begin{cases} 1 \text{ if } l = k+1, l < N_i \\ 1 \text{ if } l, k = N_i \\ 0 \text{ otherwise} \end{cases}$

2. $p^p_{i,(kl)} = \begin{cases} 1 \text{ if } l = k-1, l > 1 \\ 1 \text{ if } l, k = 1 \\ 0 \text{ otherwise} \end{cases}$

Assumption 3' removes the randomness in moving up while keeping the rest of the structure identical. In particular, despite the deterministic movements, the sets of states and the functions ρ_i, c_i, γ_i still allow for some flexibility. Assumption 3' serves a dual purpose for the analysis. First, it enables tackling the forward-looking constraints without extending the state space. Generally, when forward-looking constraints are considered, one must either take the continuation utilities as additional constraints à la Abreu, Pearce, and Stacchetti (1990) or consider recursive/dual Lagrangians with Lagrange

multipliers as additional states à la Marcet and Marimon (2011), Pavoni, Sleet, and Messner (2018). Two general results allow for addressing the forward-looking constraints without extending the state space. Rustichini (1998) requires both the forward constraint and the objective to be the same but allows for general stochasticity; Gozzi, Monte, and Tessitore (2018) allows for more general forward constraints but does not allow for stochasticity. Since the forward constraints come from a relaxation, in contrast to the case of the single agent, the latter result will be necessary. Additionaly, the deterministic law of motion is used to establish the optimality of the index policy. Multiarmed restless bandits are generally intractable (Papadimitriou and Tsitsiklis 1999), and the optimality of the index policies are established only partially in specialized settings, (Jacko 2011), (Liu and Zhao 2010), (Niño-Mora and Villar 2011), such as using optimality in relaxations or guessing and verifying. In this setup, I illustrate the optimality in the relaxed problem, which remains feasible in the original problem, to achieve the optimality of the index policy.

3.3 Lagrangian Relaxation and Markovian Behavior In Multiple-Agent Problems

Observe that since the principal always produces herself if she does not utilize the agents, the utilization constraint can be altered to rewrite the principal's

problem as follows:

$[Principal's\ \ Problem]$

$$\max_{\{I_i^t\},\{p_i^t\}} \mathbb{E}\left[\sum_{\tau=0}^{\infty}\delta^\tau\left(\sum_{k=1}^{N} I_k^\tau\left[v - c_k(s_k^\tau)\right] - \sum_{k=1}^{N} p_k^\tau\right)\right]$$

subject to

$$\sum_{k=1}^{N}(1 - I_k^t) \geq N - 1 \text{ for all } t$$

$$IC_0^i \text{ for all } i$$

$$IC_i \text{ for all } i$$

Consider a relaxation of the utilization constraint, where instead of holding at every period, the utilization constraint holds in the long-run on average.

$[Principal's\ \ Relaxed\ \ Problem]$

$$\max_{\{I_i^t\},\{p_i^t\}} \mathbb{E}\left[\sum_{\tau=0}^{\infty}\delta^\tau\left(\sum_{k=1}^{N} I_k^\tau\left[v - c_k(s_k^\tau)\right] - \sum_{k=1}^{N} p_k^\tau\right)\right]$$

subject to

$$\mathbb{E}\left[\sum_{\tau=0}^{\infty}\delta^\tau\sum_{k=1}^{N}(1 - I_k^\tau)\right] \geq \frac{N-1}{1-\delta}$$

$$IC_0^i \text{ for all } i$$

$$IC_i \text{ for all } i$$

Using a Lagrange multiplier λ for the relaxed utilization constraint and re-organizing the terms yields the following form of the relaxed problem

$$\max_{\{I_i^t\},\{p_i^t\}} \mathbb{E}\left[\sum_{\tau=0}^{\infty}\delta^\tau\left(\sum_{k=1}^{N}\left[I_k^\tau(v-c_k(s_k^\tau))\right]+\lambda(1-I_k^\tau)-\sum_{k=1}^{N}p_k^\tau\right)-\lambda\frac{N-1}{1-\delta}\right]$$

subject to

IC_P^i for all i

IC_A^i for all i

Now, observe that for any λ, the relaxed problem can be thought of as a hypothetical λ subsidy problem, where the principal receives a λ subsidy every time an agent is not utilized. The problem can be decoupled to an agent-by-agent problem since the remaining incentive constraints hold per agent. In particular, the principal faces the following decoupled λ subsidy problems that must be maximized agent by agent.

$$\max_{\{I_i^\tau\},\{p_i^\tau\}} \mathbb{E}\left[\sum_{\tau=0}^{\infty}\delta^\tau I_i^\tau\left[(v-c_i(s_i^\tau))+\lambda(1-I_i^\tau)-p_i^\tau\right]\right] \qquad (PP_i-\lambda)$$

subject to

$$\mathbb{E}\left[\sum_{\tau=t}^{\infty}\delta^\tau I_i^\tau\left[v-c_i(s_i^\tau)\right]-p_i^\tau\right]\geq-\gamma_i(s_i^t) \text{ for all } t$$

$$\mathbb{E}\left[\sum_{\tau=t}^{\infty}\delta^{\tau-t}p_i^\tau\right]\geq\rho_i(s_i^t) \text{ for all } t$$

Following the single-agent problem, I introduce the following definition.

Definition 7 (λ-Surplus of i-Dyad)**.** *For any relational contract* $\{I_i^t\},\{p_i^t\}$*,*

the i-dyad λ-surplus at period t is defined as:

$$\mathcal{S}^t_{\lambda,i} = \mathbb{E}\left[\sum_{\tau=t}^{\infty} \delta^\tau \left(I_i^\tau \left[v - c_i(s_i^\tau)\right] + (1 - I_i^t)\lambda\right)\right]$$

Lemma 4. *Suppose there exists a relational contract that generates a surplus $\mathcal{S}_i^t \geq \rho_i(s_i^t) - \gamma_i(s_i^t)$ for all t and λ-surplus $\mathcal{S}^t_{\lambda,i}$. Then, any pair of total payoffs $\mathcal{U}^t_{0,i}, \mathcal{U}^t_i$ such that $\mathcal{U}^t_{0,i} \geq -\gamma_i(s_i^t)$ and $\mathcal{U}^t_i \geq \rho_i(s_i^t)$ with $\mathcal{U}^t_{0,i} + \mathcal{U}^t_i = \mathcal{S}^t_i$ can be implemented in a relational contract while delivering a λ-surplus $\mathcal{S}^t_{\lambda,i}$.*

The proof of lemma 4 is identical to that of lemma 3 and hence is omitted. The only difference between the two lemmas is that there are now two ways to evaluate a contract: by the λ surplus or by the regular surplus. The regular surplus governs the incentives within the relationship, whereas the λ surplus incorporates the positive externality that nonutilization generates on the other relationships managed by the principal. Analogous to the single-agent problem, the two incentive constraints can be combined for the dynamic enforcement constraint DE_i, and the problem can be reduced to surplus maximization subject to the dynamic enforcement.

$$\max_{\{I_i^t\}} \ \mathbb{E}\left[\sum_{\tau=0}^{\infty} \delta^\tau \left(I_i^\tau \left[v - c_i(s_i^\tau)\right] + (1 - I_i^\tau)\lambda\right)\right] \qquad (SP_i - \lambda)$$

subject to

$$\mathbb{E}\left[\sum_{\tau=t}^{\infty} \delta^{\tau-t} \left(I_i^\tau \left[v - c_i(s_i^\tau)\right]\right)\right] \geq \rho_i(s_i^t) - \gamma_i(s_i^t) \text{ for all } t \qquad (DE_i)$$

Problem $SP_i - \lambda$ is similar to SP_i, but it also incorporates the benefit of relaxing the utilization constraint whenever the agent is not utilized. Despite the similarity, it is no longer possible to use Rustichini (1998) to conclude the optimality of Markovian behavior, as the objective and the constraint are now different. However, due to the strengthened assumption 3', it is possible to use Gozzi, Monte, and Tessitore (2018) to again conclude that Markovian

behavior is optimal without the need to extend the state space.

Proposition 4. *The surplus maximization problem $SP_i - \lambda$ subject to the dynamic enforcement constraint DE_i is solved optimally by a Markovian policy.*

Proof. Under assumption 3', the proposition is a straightforward application of Gozzi, Monte, and Tessitore (2018). For any given $\lambda \in \mathbb{R}$, observe that $SP_i - \lambda$ is always bounded due to the finite state space and is hence real valued. Second, the set of available actions including mixtures is $[0, 1]$, which is independent of the state at every point in history and is compact valued. Now, select any $T > 0$ and any sequence of actions that satisfy the DE_i for all t and consider the state reached at T, denoted s_i^T. Then, since the set of available actions has not changed, any ϵ optimal policy starting from state s_i^T remains feasible; hence, Gozzi, Monte, and Tessitore (2018) applies to deliver the existence of an optimal Markovian policy. □

Once the optimality of Markovian behavior is confirmed, some of the analysis from the single-agent problem carries over.

Proposition 5. *Under assumptions 1, 2 and 3', any Markovian policy is equivalent to a threshold policy identified by a threshold state $\bar{s}_i$. Only the threshold and the state immediately below are recurrent: all other states are transient.*

The proof of this proposition is analogous to that of proposition 2 and is hence omitted. In a similar fashion, active-passive payments can be altered to $q_i(s_{i,k}) = 1$ for all states in the following manner with the same optimality result.

Definition 8 (Active-Passive Payments)**.** *For any agent i and any state $s_{i,k} \in S_i$, active-passive payments are defined as*

$$p_i^a(s_{i,k}) = \rho_i(s_{i,k}) - \delta\rho_i(s_{i,k+1})$$
$$p_i^p(s_{i,k}) = \rho_i(s_{i,k}) - \delta\rho_i(s_{i,k-1})$$

With the convention that $p_i^p(s_{i,1}) = 0$.

Since the active-passive payments are identical, except with respect to deterministic movement, the incentive constraints of the agent holding with equality at every point in history also carries over.

Proposition 6. *Under assumptions 1, 2 and 3', for any agent* i *and any threshold level* $\bar{s}_i \in S_i$*, active-passive payments are optimal.*

The proof of proposition 6 follows identically to that of proposition 3, where $q_i(s_i i, k) = 1$ for all states, and is hence omitted. Once active-passive payments are identified, returns with active-passive payments are defined identically to definition 3. Here, it is important to highlight that once Markovian behavior is established, the payment scheme is identified only to satisfy the incentive constraints and is not related to the hypothetical subsidy level λ. However, the principal's restless bandit problem does incorporate the subsidy as follows:

$$[\textit{Principal's Single Restless Bandit Problem with Subsidy}]$$

$$\max_{\{I_i^t\}} \mathbb{E}\left[\sum_{\tau=0}^{\infty} \delta^\tau \left(I_i^\tau R_i^a(s_i^\tau) + (1 - I_i^\tau)\left[R_i^p(s_{i,k}) + \lambda\right]\right)\right]$$

In fact, single-arm restless bandits with λ subsidies can be combined to achieve the relaxed problem as follows:

$$\max_{\{I_i^t\}_{i\in\{1,2,\ldots,N\}}} \mathbb{E}\left[\sum_{\tau=0}^{\infty} \delta^\tau \sum_{i=1}^{N} \left(I_i^\tau R_i^a(s_i^\tau) + (1 - I_i^\tau)\left[R_i^p(s_{i,k}) + \lambda\right]\right)\right] - \lambda\frac{N-1}{1-\delta}$$

$$\text{subject to}$$

$$IC_P^i \text{ for all } i$$

Ignoring the incentive constraint of the principal, the remainder of the problem is exactly the Lagrangian relaxation that was proposed in Whittle (1988)

as the basis for the Whittle index for restless bandit problems. If the problem is indexable, that is, every single arm problem is indexable, then the Lagrangian relaxation is solved optimally by an index policy. Under assumptions 1, 2, and 3', the restless bandits in the λ subsidy problem are finite-state bidirectional restless bandits that are skip-free in at least one (both in this case) direction; therefore, according to Glazebrook, Hodge, and Kirkbride (2013), each arm is indexable without any requirement on the returns on activity or passivity. For each single agent, the reward $f_i^j(k)$ and work $g_i^j(k)$ measures are defined identically, and the index of each agent at each state is again identically defined as:

$$\lambda_i(s_{i,k}) = \frac{f_i^k(k+1) - f_i^k(k)}{g_i^k(k+1) - g_i^k(k)}$$

The λ subsidy problem provides additional insight into the definition of the index. For any agent i, observe that for a threshold policy k, starting from state k, the total returns, including the λ subsidy in the subsidy problem, are equal to $f_i^k(k)+(1-g_i^k(k))\lambda$. Keeping the same initial condition but moving the threshold to $k+1$ yields total returns equal to $f_i^k(k+1)+(1-g_i^k(k+1))\lambda$. Equating these two returns and solving for λ yields the definition of the index. That is, the index of state $s_{i,k}$ is the subsidy level that makes the principal indifferent between utilizing the agent and not utilizing the agent. With the identical indices defined, I can characterize a principal optimal contract as follows.

Theorem 2. *Under assumptions 1, 2, and 3', a principal optimal contract is as follows:*

1. *Each agent i is paid according to active-passive payments.*
2. *At each period t, agent i is utilized at state s_i^t if and only if $\lambda_i(s_i^t) \geq 0$ and $\lambda_i(s_i^t) > \lambda_j(s_j^t)$ for all t.*

Given the deterministic laws of motion, the path of play induced by the

theorem is relatively simple. The principal calculates the indices of all states of all agents. Any agent that has an index that is negative in the initial state is never utilized. If there is only one agent that has an initial state with an index that is positive, that agent is either utilized forever if the indices of all his states are positive or utilized repeatedly until his index drops below 0. From that point onward, the principal cycles every period between producing herself and utilizing the agent. If multiple agents have an initial state that is positive, then the principal starts by utilizing the agent with the highest initial index and continues to utilize him until his index drops below the initial index of another agent. At that point, the principal cycles between the two agents, switching every period. In essence, there is at most one "main" agent and potentially one "back-up" agent. Agents waiting in their initial state are never paid, the main agent is paid every period, and the back-up agent is paid the first time the main agent's state falls below his and is then paid in every other period.

4 Conclusion

4.1 A Short Discussion of the Modeling Choice

Endogenous state transitions in a contracting setup pose a significant challenge because the forward-looking constraints change along with the states. If the functions ρ_i and γ_i are constant (for example, in the case of constant hold-up threat), the optimality of Markov behavior in the surplus maximization can be achieved in a broad setting since the surplus maximization subject to dynamic enforcement problem becomes a constrained Markov decision problem with fixed constraints. For example, Feinberg (2000) can be used to capture the probabilistic resetting to the initial states in addition to the current one-step-down policy in the case of nonutilization.

After achieving the restless formulation, due to the general intractabil-

ity of restless bandits, some additional choices must be made. I chose the bidirectional setup with multiple states to achieve indexability, but partial conservation laws (Nino-Mora et al. 2001) or generalized conservation laws (Bertsimas and Niño-Mora 1996) are also promising approaches to establish the indexability of the problem. Due to the disjoint approaches, the payment structures would need to be adjusted, but an index characterization would still be possible.

Finally, the law of motion on the restlessness might appear to be limiting, but it is important to note that, in the case of two states and learning, models with Bayesian updating indeed lead to bidirectional movement. In fact, if the state spaces were reduced to be binary, Liu, Weber, and Zhao (2011) can be used to achieve indexability while allowing for more varied transition matrices. However, once again, the payment scheme would have to be adjusted.

4.2 Concluding Comments

The restless specification of the dynamic contracting framework enables investigating problems where utilization choices have direct impacts on the utilized part capturing a wide variety of phenomena, from learning by doing, organizational forgetting and entrenchments. Such phenomena occurs frequently in outsourcing, especially in contract manufacturing where the entire product is outsourced as opposed to just parts. Despite the cost advantages and broad usage contract manufacturing relationships usually suffer from problems tied to utilization which could be captured by restless formulation explored here. In many contract manufacturing agreements parties soon find themselves immersed in a "melodrama replete with promiscuity, infidelity, and betrayal" (Arrunada and Vázquez 2006). In some cases a contract manufacturer(agent) is in a prime position to compete or even overtake the client. "Adding insult to injury, if the client had not given its business to the traitorous contract manufacturer, the CM's knowledge might

have remained sufficiently meager to prevent it from entering its patron's market".(Arrunada and Vázquez 2006). Indeed, Intel, Cisco Systems and Alcatel firms juggle their production in order to curb the learning and efficiency of the CM, a problem that could be readily captured by an increasing pair of "break-off" functions $(\rho(\cdot), \gamma(\cdot))$ in the setup explored. Alternatively in different industries CM's must be able to meet a client's needs for flexible scheduling and capacity (Langer 2015). However, as McCoy (2003) notes, when a client approaches a contractor she may discover that he is entrenched with little flex capacity. The relationships between a client and a potential contractor become necessarily intertwined despite their bilateral nature as the client might just want to cycle through CMs to avoid such entrenchments. Contractors manage these diverse relationships by trying to keep their facilities running at $70 - 80\%$ capacity and they meet extra demands by working overtime (Tully 1994). Thus, a contractor may have to over-utilize his assets, which increases his costs. Again a problem that can be readily captured by the restless setup via an increasing production cost function $(c(\cdot))$.

Problems tied to utilization is present in many other settings beyond the contract manufacturing example as well, and the restless setup explored here provides a simple, tractable framework to study many economic phenomena where a principal might have to juggle different agents. The index solution here provides an optimal and intuitive method of "juggling", with the familiar interpretation as a marginal increase, where the margin is on the policy space. The restless bandits literature has some limitations due to the tractability of the bandit problem, but the dynamic contracting setting not only captures a lot of economic phenomena naturally but also is more promising than the bandit problem. On the one hand payments needs to be pinned down for each different incentive friction in addition to the scheduling problem which might seem like a complication, but the payments being a choice allows for construction of bandits, that might help alleviate the intractability issues by a suitable choice, while simultaneously capturing the relevant phenomena.

References

Abreu, D., D. Pearce, and E. Stacchetti (1990): "Toward a Theory of Discounted Repeated Games with Imperfect Monitoring," *Econometrica*, 58(5), 1041–1063.

Andrews, I., and D. Barron (2013): "The allocation of future business: Dynamic relational contracts with multiple agents," *American Economic Review*.

Arrunada, B., and X. H. Vázquez (2006): "When your contract manufacturer becomes your competitor," *Harvard business review*, 84(9), 135.

Baker, G., R. Gibbons, and K. J. Murphy (2002): "Relational Contracts and the Theory of the Firm," *Quarterly Journal of economics*, pp. 39–84.

Bergemann, D., and J. Välimäki (1996): "Learning and strategic pricing," *Econometrica: Journal of the Econometric Society*, pp. 1125–1149.

Bertsimas, D., and J. Niño-Mora (1996): "Conservation laws, extended polymatroids and multiarmed bandit problems; a polyhedral approach to indexable systems," *Mathematics of Operations Research*, 21(2), 257–306.

Blackwell, D. (1965): "Discounted Dynamic Programming," *The Annals of Mathematical Statistics*, 36(1), pp. 226–235.

Board, S. (2011): "Relational contracts and the value of loyalty," *The American Economic Review*, pp. 3349–3367.

Bolton, P., and C. Harris (1999): "Strategic experimentation," *Econometrica*, 67(2), 349–374.

Cisternas, G., and N. Figueroa (2015): "Sequential procurement auctions and their effect on investment decisions," *The RAND Journal of Economics*, 46(4), 824–843.

Feinberg, E. A. (2000): "Constrained discounted Markov decision processes and Hamiltonian cycles," *Mathematics of Operations Research*, 25(1), 130–140.

Fryer, R., and P. Harms (2017): "Two-armed restless bandits with imperfect information: Stochastic control and indexability," *Mathematics of Operations Research*.

Gittins, J., K. Glazebrook, and R. Weber (2011): *Multi-armed bandit allocation indices*. John Wiley & Sons.

Gittins, J. C. (1979): "Bandit processes and dynamic allocation indices," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 148–177.

Glazebrook, K., D. Hodge, and C. Kirkbride (2013): "Monotone policies and indexability for bidirectional restless bandits," *Advances in Applied Probability*, 45(1), 51–85.

Glazebrook, K., J. Nino-Mora, and P. Ansell (2002): "Index policies for a class of discounted restless bandits," *Advances in Applied Probability*, pp. 754–774.

Glazebrook, K., D. Ruiz-Hernandez, and C. Kirkbride (2006): "Some indexable families of restless bandit problems," *Advances in Applied Probability*, pp. 643–672.

Gozzi, F., R. Monte, and M. E. Tessitore (2018): "On the dynamic programming approach to incentive constraint problems," in *Control Systems and Mathematical Methods in Economics*, pp. 81–96. Springer.

Halac, M. (2012): "Relational contracts and the value of relationships," *American Economic Review*, 102(2), 750–79.

Jacko, P. (2011): "Optimal index rules for single resource allocation to stochastic dynamic competitors," in *Proceedings of the 5th International ICST Conference on Performance Evaluation Methodologies and Tools*, pp. 425–433. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering).

Keller, G., S. Rady, and M. Cripps (2005): "Strategic experimentation with exponential bandits," *Econometrica*, 73(1), 39–68.

Klein, N., and S. Rady (2011): "Negatively Correlated Bandits," *The Review of Economic Studies*.

Kwon, S. (2016): "Relational contracts in a persistent environment," *Economic Theory*, 61(1), 183–205.

Langer, E. S. (2015): "CMOs Facing Significant Capacity Constraints," *Contract Pharma*.

Levin, J. (2002): "Multilateral contracting and the employment relationship," *Quarterly Journal of economics*, pp. 1075–1103.

——— (2003): "Relational incentive contracts," *The American Economic Review*, 93(3), 835–857.

Ligon, E., J. P. Thomas, and T. Worrall (2002): "Informal insurance arrangements with limited commitment: Theory and evidence from village economies," *The Review of Economic Studies*, 69(1), 209–244.

Liu, K., R. Weber, and Q. Zhao (2011): "Indexability and whittle index for restless bandit problems involving reset processes," in *2011 50th IEEE Conference on Decision and Control and European Control Conference*, pp. 7690–7696. IEEE.

Liu, K., and Q. Zhao (2010): "Indexability of restless bandit problems and optimality of whittle index for dynamic multichannel access," *IEEE Transactions on Information Theory*, 56(11), 5547–5567.

Malcomson, J. M. (2016): "Relational incentive contracts with persistent private information," *Econometrica*, 84(1), 317–346.

Malcomson, J. M., et al. (2010): *Relational incentive contracts*. Department of Economics, University of Oxford.

Marcet, A., and R. Marimon (2011): "Recursive contracts," .

McCoy, M. (2003): "Serving emerging pharma," *Chemical & engineering news*, 81(16), 21–33.

Nino-Mora, J. (2002): "Dynamic allocation indices for restless projects and queueing admission control: a polyhedral approach," *Mathematical programming*, 93(3), 361–413.

Niño-Mora, J. (2006): "Restless bandit marginal productivity indices, diminishing returns, and optimal control of make-to-order/make-to-stock M/G/1 queues," *Mathematics of Operations Research*, 31(1), 50–84.

——— (2007): "Dynamic priority allocation via restless bandit marginal productivity indices," *Top*, 15(2), 161–198.

Nino-Mora, J., et al. (2001): "Restless bandits, partial conservation laws and indexability," *Advances in Applied Probability*, 33(1), 76–98.

Niño-Mora, J., and S. S. Villar (2011): "Sensor scheduling for hunting elusive hiding targets via Whittle's restless bandit index policy," in *International Conference on NETwork Games, Control and Optimization (NetGCooP 2011)*, pp. 1–8. IEEE.

Papadimitriou, C. H., and J. N. Tsitsiklis (1999): "The complexity of optimal queuing network control," *Mathematics of Operations Research*, 24(2), 293–305.

Pavoni, N., C. Sleet, and M. Messner (2018): "The dual approach to recursive optimization: theory and examples," *Econometrica*, 86(1), 133–172.

Puterman, M. L. (2014): *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons.

Rosenberg, D., E. Solan, and N. Vieille (2007): "Social Learning in One-Arm Bandit Problems," *Econometrica*, 75(6), 1591–1611.

Rustichini, A. (1998): "Dynamic programming solution of incentive constrained problems," *Journal of Economic Theory*, 78(2), 329–354.

Serfozo, R. (2009): *Basics of applied stochastic processes*. Springer Science & Business Media.

Strulovici, B. (2010): "Learning while voting: Determinants of collective experimentation," *Econometrica*, 78(3), 933–971.

Thomas, J., and T. Worrall (1988): "Self-enforcing wage contracts," *The Review of Economic Studies*, 55(4), 541–554.

Tully, S. (1994): "You'll never guess who really makes," *Fortune*, 130(7), 124–128.

Whittle, P. (1988): "Restless bandits: Activity allocation in a changing world," *Journal of applied probability*, pp. 287–298.

Yang, H. (2013): "Nonstationary relational contracts with adverse selection," *International Economic Review*, 54(2), 525–547.

5 Appendix - For Online Publication

In most calculations, it is necessary to use a common version (see Serfozo (2009) pp 399-400) of Wald's identity for discounted partial sums with stopping times. For convenience, I include the identity here as well.

Identity 1 (Wald's Identity for Discounted Sums)**.** *Suppose that $X_1, X_2, \ldots$ are i.i.d. with mean $\bar{x}$. Let $\delta \in (0,1)$ and τ be a stopping time for $X_1, X_2, \ldots$ with $E(\tau) < \infty$ and $E(\delta^\tau)$ exists. Then,*

$$E(\sum_{t=0}^{\tau} \delta^t X_t) = \frac{\bar{x}(1 - \delta E(\delta^\tau))}{1-\delta}$$

5.1 Proof of Lemma 1

Proof. Note that due to assumption 4 in any principal optimal contract no agent ever breaks off, that is $d_i^t = 1$ for all t and for all i. Thus in any principal optimal contract the continuation payoff for any agent i at any period t is given by

$$v_i^t = \mathbb{E}\left[\sum_{\tau=t}^{\infty} \delta^{\tau-t} p_i^\tau\right]$$

Observe that if the agent ever breaks off at period t he receives a payoff $\rho_i(s_i^t)$. Thus if $\mathbb{E}\left[\sum_{\tau=t}^{\infty} \delta^{\tau-t} p_i^\tau\right] \geq \rho_i(s_i^t)$ for all t then the agent never has an incentive to break of from the principal. On the other hand if there exists a history h^τ and a period τ where the inequality does not hold, then in that period the agent can gain by breaking off. Since the agent does not have any additional choices the condition is necessary and sufficient. □

5.2 Proof of Lemma 2

Recall that a strategy of the principal is a mapping $\sigma_0(h^t)$ that maps a history to a payment vector $\{p_i^t\}_{i\in\{1,2,\ldots,N\}} \in \mathbb{R}^N$ and utilization vector $\{I_i^t\}_{i\in\{0,1,2,\ldots,N\}} \in \{0,1\}^{N+1}$ with the restriction that $\sum_{i=0}^{N} I_i^t = 1$ and a public signal $y_t \in Y$. Since the public signal sent is payoff irrelevant, on the equilibrium path it is without loss to assume that the principal sends $y_t = (p^t, s^t)$ every period. I will denote a deviation at history h^t from an equilibrium strategy σ_0 a *deviation against* i if σ_0 recommends I_i^t and p_i^t after h^t and the principal deviates by either offering $\tilde{I}_i^t \neq I_i^t$ or $\tilde{p}_i^t \neq p_i^t$. I will assume that any deviation that does not involve a deviation against i for some i is ignored both by the agents and the principal. Observe that the agent's minmax action is to break off from the principal. Given the bilateral nature of the relationships it is without loss to assume that any deviation against i is punished by agent i by immediate rejection and breaking off, that is $d_i^t = 0$. Similarly all agents who are not deviated against will accept the offer. Suppose σ is a principal optimal relational by assumption 4, on the equilibrium path no agent ever breaks off, thus, under σ the principal's payoff starting from period t following a history h^t is given by

$$v_0^t(\sigma) = \mathbb{E}\left[\sum_{\tau=t}^{\infty} \delta^{\tau-t}\left(\sum_{k=1}^{N} I_k^\tau \left[v - c_k(s_k^\tau)\right] - \sum_{k=1}^{N} p_k^\tau\right)\right]$$

Reorganizing the sums leads to:

$$v_0^t(\sigma) = \mathbb{E}\left[\sum_{k=1}^{N}\left(\sum_{\tau=t}^{\infty} \delta^{\tau-t} I_k^\tau \left[v - c_k(s_k^\tau)\right] - \sum_{k=1}^{N} p_k^\tau\right)\right]$$

Once the principal deviates against an agent i, then agent i will leave the game and his states will transition downwards until they reach the initial state. Now consider the following *replication of* i by the principal. Let τ

be the period where agent i breaks off from the principal. And let $\hat{p}_i^\tau$ be the payment that the principal would have paid on path, and let $\hat{s}_i^\tau$ be the random variable that denotes state of agent i at the end of period τ on path, conditional on I_i^τ. At the end of period τ let s_j^τ and p_j^τ respectively denote the states and payments made to agents $j \neq i$. Then in period τ the principal sends $y_\tau = (\hat{s}_i^\tau, \hat{p}_i^\tau, \{s_j^\tau\}_{j\neq i}, \{p_j^\tau\}_{j\neq i})$. That is the principal sends what the public history would be (including randomization by the principal conditional on whether agent i is employed or not) as a public signal. In period $\tau + 1$ the principal makes offers to all agents $j \neq i$ as if the history $h_{\tau-1}$ was followed by (y_τ, y_τ). In this manner in period $\tau + 1$ for all agents $j \neq i$, $I_j^{\tau+1}$ and $p_j^{\tau+1}$ is the same both on path and after a deviation against i in period τ. For now assume the agents accept these offers. If in period $\tau + 1$ conditional on the history $(h_{\tau-1}, (y_\tau, y_\tau))$, $I_i^{\tau+1} = 1$ then the principal utilizes herself $I_0^{\tau+1} = 1$ if one of the agents were to be utilized then that agent is utilized. Again conditional on $\hat{s}_i^\tau$'s realization and $I_i^{\tau+1}$, the principal can randomize according to P_i^a or P_i^p respectively to generate the random variable $\hat{s}_i^{\tau+1}$. At the end of period $\tau + 1$ then the principal would then send a signal $y_{\tau+1}$ where $y_\tau = (\hat{s}_i^{\tau+1}, p_i^{\tau+1}, \{s_j^\tau + 1\}_{j\neq i}, \{p_j^\tau + 1\}_{j\neq i})$ Clearly for all periods $\tau + k$ for $k \geq 1$ the principal can continue randomizing to replicate what the on path public history would be as if agent i has not broken off and send it as a signal $y_{\tau+k}$. If the principal uses this replication strategy then all agents $j \neq i$ have no incentive to reject the offers at any period $\tau + k$ since they are the same both on the equilibrium path as well as after a deviation against i. In particular in such a strategy y_t will follow the on equilibrium path of (p^t, s^t) for all t even after a deviation against i. Finally since σ was principal optimal, keeping the rest of the schedule $I_j^{\tau+k}$ and $p_j^{\tau+k}$ has to be optimal. But then the after deviation optimal payoff for some $t > \tau$ denoted

by $v_0^t(\sigma|dev-i)$ is as follows:

$$v_0^t(\sigma|dev-i) = \mathbb{E}\left[\sum_{k\neq i}\left(\sum_{r=t}^{\infty}\delta^{r-t}I_k^\tau\left[v-c_k(s_k^r)\right]-\sum_{k\neq i}p_k^r\right)\right]$$

At the point of deviation, the principal also loses $\gamma_i(s_i^\tau)$. Thus if the principal is to not deviate against i in period τ the following must hold:

$$\begin{aligned}&\mathbb{E}\left[\sum_{k=1}^{N}\left(\sum_{r=\tau}^{\infty}\delta^{r-\tau}I_k^r\left[v-c_k(s_k^r)\right]-\sum_{k=1}^{N}p_k^r\right)\right]\\&\geq-\gamma_i(s_i^\tau)+\mathbb{E}\left[\sum_{k\neq i}\left(\sum_{r=\tau}^{\infty}\delta^{r-\tau}I_k^r\left[v-c_k(s_k^r)\right]-\sum_{k\neq i}p_k^r\right)\right]\end{aligned}$$

Notice that the expectation operator on both sides have the same law due to the replication of i strategy. Simplifying the above yields:

$$\mathbb{E}\left[\sum_{r=\tau}^{\infty}\delta^{r-\tau}I_i^r\left[v-c_i(s_i^r)\right]-p_i^r\right]\geq-\gamma_i(s_i^\tau)$$

Observe that since $\sum_{i=0}^{N}I_i^t=1$ which the principal can replicate multiple agents in a similar fashion to above as well, just by replacing the agent that was deviated against by her own utilization in the relevant period. Thus in order for the principal to not deviate against any agent in any period we need IC_0^i holding. Finally observe that the special form of the public signal is completely unnecessary as any on path history can in principle be mapped to a public randomization device that takes values in $[0,1]$, which is standard in the literature.

5.3 Proof of Proposition 2

Observation 1. *Any pure Markov policy will map states into utilization decisions. Let π be any pure Markov policy, and let S_i^π denote its active set*

such that $I_i^t = 1 \Leftrightarrow s_i^t \in S_i^\pi$.

Proof. Note that the initial state is $s_{i,1}$, and consider any pure Markov policy π, identified with its active set S_i^π. Let $s_{i,\underline{x}} = \max\{s_i \in S_i : s_i \notin S_i^\pi\}$. Then, by definition under policy π, for all t $s_i^t \in \{s_{i,1}, \dots s_{i,\underline{x}}\}$. Moreover, for all t, $I_i^t = 1 \Leftrightarrow s_i^t < s_{i,\underline{x}}$. However, observe that due to the bidirectional law of motion, once state $s_{i,\underline{x}}$ is reached, the agent becomes passive and hence returns to state $s_{i,\underline{x}-1}$. By definition $s_{i,\underline{x}-1} \in S_i^\pi$, the agent becomes active again and continues to alternate between the two states from that point onward. □

5.4 Proof of Proposition 3

Proof. First, observe that with any threshold policy for agent i, only the threshold level $s_{i,\underline{x}}$ and the state immediately before $s_{i,\underline{x}-1}$ are recurrent. Any other state $s_{i,\underline{x}-k}$ for $k > 1$ will be transient, and any state $s_{i,\underline{x}+l}$ for $l > 0$ will never be reached. Below, I first show that the incentive constraints of the agent hold with equality in the recurrent states; then, I show that the incentive constraints also hold with equality in the transient states. Hence, the principal leaves no slack for the agent.

Lemma 5. *When* $p_i^p(s_{i,k}) = \rho_i(s_{i,k}) - \delta\rho_i(s_{i,k-1})$ *for all* $k \in \{1, 2, \dots, N_i\}$*, the incentive constraints of the agent hold with equality in the recurrent states.*

Proof of Lemma 5. Let $T_i(x, y, z)$ denote the expected discounted time agent i spends in state $s_{i,x}$ under the threshold policy with threshold $s_{i,y}$ starting from initial state $s_{i,z}$. Consider an arbitrary threshold k; then, the incentive

conditions holding with equality implies

$$\begin{aligned} p_i^a(s_{i,k-1})T_i(k-1,k,k-1) + p_i^p(s_{i,k})T_i(k,k,k-1) &= \rho_i(k-1) \\ &\quad \text{(IC at } s_{i,k-1}) \\ p_i^a(s_{i,k-1})T_i(k-1,k,k) + p_i^p(s_{i,k})T_i(k,k,k) &= \rho_i(k) \\ &\quad \text{(IC at } s_{i,k}) \end{aligned}$$

Observe that the left-hand side (lhs) of first line corresponds to the expected discounted value of all future payments to the agent starting from state $s_{i,k-1}$ under the k threshold policy and the right-hand side (rhs) is the benefit of breaking off. Similarly, the lhs of the second line corresponds to the expected discounted value of all future payments to the agent starting from state $s_{i,k}$ under the k threshold policy, and the rhs is the benefit of breaking off. Rearranging the second equation, we obtain

$$p_i^a(s_{i,k-1})T_i(k-1,k,k) = \rho_i(k) - p_i^p(s_{i,k})T_i(k,k,k)$$

Now, by the bidirectional law of motion, observe that after spending a single period in state k, the agent returns to $k-1$ before cycling again, which implies

$$T_i(k-1,k,k) = \delta T_i(k-1,k,k-1)$$

Plugging in the equality from the second line and using the relationship $T_i(k-1,k,k) = \delta T_i(k-1,k,k-1)$, the system reduces to:

$$\rho_i(k) - p_i^p(s_{i,k})T_i(k,k,k) + \delta(p_i^p(s_{i,k})T_i(k,k,k-1)) = \delta\rho_i(k-1)$$

with $p_i^a(\cdot)$ being free. Again, since the agent returns to $k-1$ after spending a single period in state k before returning to state $k-1$, we also have the

following relation

$$T_i(k,k,k) = 1 + \delta T_i(k,k,k-1)$$

Plugging in the second relation pins down $p_i^p(s_{i,k}) = \rho_i(s_{i,k}) - \delta\rho_i(s_{i,k-1})$ for the incentive conditions to hold at the recurrent states, regardless of $p_i^a(s_{i,k})$, for any threshold k. □

Lemma 6. *When $p_i^a(s_{i,k}) = (1-\delta(1-q_i))\rho_i(s_{i,k}) - \delta q_i \rho_i(s_{i,k+1})$ for all $k \in \{1,2,\ldots,N_i\}$, the incentive constraints of the agent hold with equality at the transient states.*

Proof of Lemma 6. From lemma 5, for any threshold k, we know that at the recurrent states, the incentive constraints hold with equality with no restrictions on the active payments. Let $\tau_{k-1} = \inf_{t\geq 0}\{t : s_i^t = s_{i,k-1}, s_i^0 = s_{i,k-2} \text{ and } I_i^s = 1 \forall s\}$; that is, τ_{k-1} is the time to reach state $k-1$ starting from state $k-2$ by utilizing the agent in every period. Given the law of motion, this process is equivalent to repeated Bernoulli trials with odds $q_i(s_{i,k-2})$ until the first success. By definition, we have

$$\mathbb{E}(\delta^{\tau_{k-1}}) = \frac{q_i(s_{i,k-2})\delta}{1-\delta(1-q_i(s_{i,k-2}))}$$

Observe that under the threshold policy with active-passive payments, once the agent reaches state $k-1$, the expected discounted payment from the first time state $k-1$ is reached is equal to $\rho_i(s_{i,k-1})$. Then, the expected discounted payment starting from state $k-2$ under the threshold k can be written as

$$\mathbb{E}\left[\sum_{t=0}^{\tau_{k-1}-1} \delta^t p_i^a(s_{i,k-2}) + \delta^{\tau_{k-1}}\rho_i(s_{i,k-1})\right]$$

Now, using identity 1 and the identity above, we can calculate the expectation

in closed form and equate it to the incentive constraint at state $k-2$ to obtain

$$\frac{p_i^a(s_{i,k-2})}{1-\delta(1-q_i(s_{i,k-2})))}+\frac{q_i(s_{i,k-2})\delta\rho_i(s_ik-1)}{1-\delta(1-q_i(s_{i,k-2}))}=\rho_i(s_{i,k-2})$$

After minor algebra and shifting of the indices, we can pin down the active payment as $p_i^a(s_{i,k})=(1-\delta(1-q_i(s_{i,k})))\rho_i(s_{i,k})-\delta q_i(s_{i,k})\rho_i(s_{i,k+1})$, which ensures that the incentive conditions hold at all states in which the agent is active, both in the transient and recurrent states. □

□

5.5 Proof of Theorem 1

Recall that the principal's problem was the following:

$$[Principal's\ \ Problem]$$

$$\max_{\{I_i^t\},\{p_i^t\}}\mathbb{E}\left[\sum_{\tau=0}^{\infty}\delta^\tau\left(\sum_{k=1}^{N}I_k^\tau\left[v-c_k(s_k^\tau)\right]-\sum_{k=1}^{N}p_k^\tau\right)\right]$$

subject to

$$\sum_{l=0}^{N}I_k^t=1 \text{ for all } t$$

$$IC_0^i \text{ for all } i$$

$$IC_i \text{ for all } i$$

According to lemma 3, the problem above is equal to:

$$\max_{\{I_i^t\}} \mathbb{E}\left[\sum_{\tau=0}^{\infty} \delta^\tau \left(I_i^\tau \left[v - c_i(s_i^\tau)\right]\right)\right] \tag{SP_i}$$

subject to

$$\mathbb{E}\left[\sum_{\tau=0}^{\infty} \delta^\tau \left(I_i^\tau \left[v - c_i(s_i^\tau)\right]\right)\right] \geq \rho_i(s_i^t) - \gamma_i(s_i^t) \tag{DE_i}$$

Furthermore, according to proposition 1, there is a Markovian solution to SP_i. Due to proposition 2, any Markov policy is equivalent to a threshold policy, and proposition 3 for active-passive payments guarantee that the incentive condition of the agent is satisfied exactly at every reachable history with any threshold policy, regardless of the threshold. Hence, assuming active-passive payments and a threshold policy without loss, the principal's problem reduces to:

$$[\textit{Principal's Single Restless Bandit Problem}]$$
$$\max_{\{I_i^t\}} \mathbb{E}\left[\sum_{\tau=0}^{\infty} \delta^\tau \left(I_i^\tau R_i^a(s_i^\tau) + (1 - I_i^\tau) R_i^p(s_{i,k})\right)\right]$$

subject to her own incentive constraint. However, the objective identified here is a finite-state single-arm restless bandit with one side skip-free; hence, it is indexable, and the scheduling policy is optimally solved by the Whittle index due to Glazebrook, Hodge, and Kirkbride (2013). Furthermore, the positive indices imply that the expected discounted payoff of the principal is increasing as long as the index is positive by increasing the threshold marginally; hence, the principal's profits never fall below 0, which implies the principal's incentive constraint is not binding at any history under the index policy.

5.6 Proof of Theorem 2

Recall that the principal's problem was the following:

$$[Principal's\ \ Problem]$$

$$\max_{\{I_i^t\},\{p_i^t\}} \mathbb{E}\left[\sum_{\tau=0}^{\infty}\delta^\tau\left(\sum_{k=1}^{N} I_k^\tau\left[v - c_k(s_k^\tau)\right] - \sum_{k=1}^{N} p_k^\tau\right)\right]$$

subject to

$$\sum_{k=1}^{N}(1 - I_k^t) \geq N - 1 \text{ for all } t$$

$$IC_0^i \text{ for all } i$$

$$IC_i \text{ for all } i$$

Let PP denote the solution to the principal's problem. The principal's relaxed problem was introduced by relaxing the utilization constraint

$$[Principal's\ \ Relaxed\ \ Problem]$$

$$\max_{\{I_i^t\},\{p_i^t\}} \mathbb{E}\left[\sum_{\tau=0}^{\infty}\delta^\tau\left(\sum_{k=1}^{N} I_k^\tau\left[v - c_k(s_k^\tau)\right] - \sum_{k=1}^{N} p_k^\tau\right)\right]$$

subject to

$$\mathbb{E}\left[\sum_{\tau=0}^{\infty}\delta^\tau\sum_{k=1}^{N}(1 - I_k^\tau)\right] \geq \frac{N-1}{1-\delta}$$

$$IC_0^i \text{ for all } i$$

$$IC_i \text{ for all } i$$

Let PPR denote the solution to the relaxed problem. By definition, we know that $PPR \geq PP$.

Observe that the relaxed problem can be decoupled by introducing a Lagrange multiplier to the utilization constraint, which leads to the problems

$PP_i - \lambda$, which have to be solved optimally for a single λ.

$$\max_{\{I_i^\tau\},\{p_i^\tau\}} \mathbb{E}\left[\sum_{\tau=0}^{\infty} \delta^\tau I_i^\tau \left[(v - c_i(s_i^\tau)) + \lambda(1 - I_i^\tau) - p_i^\tau\right]\right] \qquad (PP_i - \lambda)$$

subject to

$$\mathbb{E}\left[\sum_{\tau=t}^{\infty} \delta^\tau I_i^\tau \left[v - c_i(s_i^\tau)\right] - p_i^\tau\right] \geq -\gamma_i(s_i^t) \text{ for all } t$$

$$\mathbb{E}\left[\sum_{\tau=t}^{\infty} \delta^{\tau-t} p_i^\tau\right] \geq \rho_i(s_i^t) \text{ for all } t$$

Due to lemma 4, $PP_i - \lambda$ problems are equivalent to λ-surplus maximization problems, subject to the surplus satisfying the dynamic enforcement constraints DE_i. Now, observe that due to proposition 4, each λ-surplus maximization subject to DE_i is solved by a Markov policy, and due to proposition 5, each Markov policy is a threshold policy. Therefore, when solving $PP_i - \lambda$, we can restrict our attention to threshold policies. However, due to proposition 6, we know that the principal cannot do any better than active-passive payments for any threshold, so the $PP_i - \lambda$ problem can be reduced to the following:

$$\max_{\{I_i^\tau\},\{p_i^\tau\}} \mathbb{E}\left[\sum_{\tau=0}^{\infty} \delta^\tau I_i^\tau \left[(v - c_i(s_i^\tau)) + \lambda(1 - I_i^\tau) - p_i^\tau\right]\right] \qquad (PP_i - \lambda)$$

subject to

$$\mathbb{E}\left[\sum_{\tau=t}^{\infty} \delta^\tau I_i^\tau \left[v - c_i(s_i^\tau)\right] - p_i^\tau\right] \geq -\gamma_i(s_i^t) \text{ for all } t$$

$$p_i^t = p_i^a(s_{i,k}) \text{ if } I_i^t = 1$$

$$p_i^t = p_i^p(s_{i,k}) \text{ if } I_i^t = 0$$

Plugging in the closed forms of the active-passive payments to obtain the returns from activity-passivity reduces the $PP_i - \lambda$ problem to the restless

bandit problem with λ subsidy, subject to the principal's incentive constraint.

$$\max_{\{I_i^t\}} \mathbb{E}\left[\sum_{\tau=0}^{\infty} \delta^\tau \left(I_i^\tau R_i^a(s_i^\tau) + (1 - I_i^\tau)\left[R_i^p(s_{i,k}) + \lambda\right]\right)\right]$$

subject to

$$\mathbb{E}\left[\sum_{\tau=t}^{\infty} \delta^\tau I_i^\tau R_i^a(s_i^\tau) + (1 - I_i^\tau) R_i^p(s_{i,k})\right] \geq -\gamma_i(s_i^t) \text{ for all } t$$

Now, ignoring the constraint, the objective in this problem is a one-armed, bidirectional, skip-free bandit with λ subsidy that, due to Glazebrook, Hodge, and Kirkbride (2013), is indexable, and the optimal policy is the Whittle index policy that sets $I_i^t = 1$ whenever $\lambda_i(s_i^t) > \lambda$, $I_i^t = 0$ whenever $\lambda_i(s_i^t) < \lambda$ and $I_i^t \in [0, 1]$ whenever $\lambda_i(s_i^t) = \lambda$, with $\lambda_i(s_i^t)$ defined as in definition 6. Moreover, whenever the indices are positive, the payoff from that state onward under the index policy is positive and hence greater than $-\gamma_i(\cdot)$.

Recombining all the individual problems yields

$$\max_{\{I_i^t\}_{i\in\{1,2,\ldots,N\}}} \mathbb{E}\left[\sum_{\tau=0}^{\infty} \delta^\tau \sum_{i=1}^{N} \left(I_i^\tau R_i^a(s_i^\tau) + (1 - I_i^\tau)\left[R_i^p(s_{i,k}) + \lambda\right]\right)\right] - \lambda \frac{N-1}{1-\delta}$$

subject to

$$IC_P^i \text{ for all } i$$

where the optimum has the value PPR.

Now, consider each of the individual single-arm problems and the collection of all indices $\{\{\lambda_i(s_{i,k})\}_{\{s_{i,k}\in S_i\}}\}_{i\in\{1,2,\ldots,N\}}$. Let $\bar{i}$ be the agent show has the maximum initial index across all agents that are positive; that is, $\bar{i}$ is the agent $i \in \{1, 2, \ldots, N\}$ such that $\lambda_i(s_{i,1}) \geq \lambda_j(s_{j,1})$ for $j \neq i$ and $\lambda_i(s_{i,1}) \geq 0$. If no such agent exists, the solution to all individual problems is to produce in-house all the time. Similarly, let $\underline{i}$ be the agent who has

the maximum initial index across agents other than $\bar{i}$; that is, $\underline{i}$ is the agent $i \in \{1, 2, \ldots, N\} \setminus \{\bar{i}\}$ such that $\lambda_i(s_{i,1}) \geq \lambda_j(s_{j,1})$ for $j \in \{1, 2, \ldots, N\} \setminus \{\bar{i}\}$. Now, consider the principal's relaxed problem with active-passive payments.

$$
\begin{aligned}
&[\textit{Principal's Relaxed Problem}] \\
&\max_{\{I_i^t\}} \mathbb{E}\left[\sum_{\tau=0}^{\infty} \delta^\tau \left(\sum_{k=1}^{N} \left[I_k^\tau \left[v - c_k(s_k^\tau) - p_k^a(s_k^\tau)\right] + (1 - I_k^\tau) p_k^p(s_k^\tau) + \lambda\right]\right)\right] - \lambda \frac{N-1}{1-\delta} \\
&\quad \text{subject to} \\
&\quad IC_0^i \text{ for all } i
\end{aligned}
$$

Setting $\lambda = \max\{\lambda_{\underline{i}}(s_{\underline{i},1}), 0\}$ results in only agent $\bar{i}$ or $\underline{i}$ ever being active. There are three possible cases:

Case 1 If $\lambda_{\underline{i}}(s_{\underline{i},1}) < 0$, then only agent $\bar{i}$ is ever utilized whenever the index of the state of $\bar{i}$ is positive, which achieves optimality in the relaxed problem and is also feasible in the restricted problem since the utilization constraint is not binding, resulting in $PPR = PP$ with the index policy.

Case 2 If $\lambda_{\underline{i}}(s_{\underline{i},1}) < \min_{s_{\bar{i},k} \in S_{\bar{i}}} \lambda_{\bar{i}}(s_{\bar{i},k})$, then again only agent $\bar{i}$ is ever utilized whenever the index of the state of $\bar{i}$ is positive, which achieves optimality in the relaxed problem and is also feasible in the restricted problem since the utilization constraint is not binding, resulting in $PPR = PP$ with the index policy.

Case 3 If $\lambda_{\underline{i}}(s_{\underline{i},1}) > 0$ and $\lambda_{\underline{i}}(s_{\underline{i},1}) > \min_{s_{\bar{i},k} \in S_{\bar{i}}} \lambda_{\bar{i}}(s_{\bar{i},k})$, then with the chosen λ, the principal is indifferent between utilizing agent $\underline{i}$ at every period and thus utilizes agent $\bar{i}$ whenever the index of the state of $\bar{i}$ is greater than $\lambda_{\underline{i}}(s_{\underline{i},1})$. In this case, consider the calendar-based "randomization" strategy that utilizes $\underline{i}$ whenever the index of the state of $\bar{i}$ is less than $\lambda_{\underline{i}}(s_{\underline{i},1})$.[2] This policy is an optimal policy in the relaxed problem and is feasible in

[2]The strategy is deterministic and dependent on calendar time. Randomization refers only to how ties are broken when the index is equal to λ.

the restricted problem, resulting in $PPR = PP$. Additionally, this policy overlaps with the index policy suggested in theorem 2.