Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence

A Tractable Leader-Follower MDP Model for Animal Disease Management

Régis Sabbadin
INRA-MIAT, Toulouse, France
Regis.Sabbadin@toulouse.inra.fr

Anne-France Viet
INRA, LUNAM Université, Oniris,
UMR1300 BioEpAR, Nantes, France
anne-france.viet@oniris-nantes.fr

Abstract

Sustainable animal disease management requires to design and implement control policies at the regional scale. However, for diseases which are not *regulated*, individual farmers are responsible for the adoption and successful application of control policies at the farm scale. Organizations (groups of farmers, health institutions...) may try to influence farmers' control actions through financial incentives, in order to ensure sustainable (from the health and economical point of views) disease management policies. Economics / Operations Research frameworks have been proposed for modeling the effect of incentives on agents. The Leader-Follower Markov Decision Processes framework is one such framework, that combines Markov Decision Processes (MDP) and stochastic games frameworks. However, since finding equilibrium policies in stochastic games is hard when the number of players is large, LF-MDP problems are intractable.

Our contribution, in this article, is to propose a tractable model of the animal disease management problem. The tractable model is obtained through a few simple modeling approximations which are acceptable when the problem is viewed from the organization side. As a result, we design a polynomial-time algorithm for animal disease management, which we evaluate on a case study inspired from the problem of controlling the spread of the Porcine Reproductive and Respiratory Syndrome (PRRS).

Content Area : Animal infectious disease management.

Introduction

The decision to control an endemic but not regulated animal disease is taken at the farmers' initiative. Each year, farmers choose to apply control actions in their own farm (biosecurity, culling, vaccination). Farmers usually decide their control action individually, without considering other farmers' decisions. However, for a disease that can be transmitted by animal purchases or direct contact, regulation decisions taken in a single or in a limited number of farms are rarely sufficient to control the disease propagation within an area, since the pathogen can spread between neighbor farms. Disease spread to one farm is not only affected by the local action applied, but also by the global infection level in the region (for example, the total number of infected herds).

The yearly revenue of a farmer is assumed to be a function of the farm status (infected or not), and of the local control action (culling and replacement costs). Farmers are then assumed to maximize their long-term profit by deciding when to apply control actions. In general, individual control actions' costs are comparable to individual costs of infection and infection is not very likely unless a significant number of farms are already infected. This is why farmers' maximization behavior often leads to not applying control actions. This, in turn, may result in new infections and money losses for the collectivity. Therefore, farmers organizations try to enforce coordination of individual decisions by using financial (for example) incentives for control actions. Regulation can then be obtained when farmers, assumed to be rationally trying to maximize their profit, balance the cost of control actions, the financial incentive and the likelihood of contamination from other "infected" farms. Despite its cost, using financial incentives is an efficient way to decrease the frequency of infections and, overall, to increase the financial and epidemiological sustainability of farm management.

The problem of endemic disease control can then be seen as a form of sequential multi-agent decision problem, in which all agents maximize their individual long-term objective. Farmers try to maximize their long-term profit by deciding when to apply control actions, while a specific distinguished agent (the Organization) has for objective to minimize the disease extension, by giving incentives to farmers. This problem can be modeled as a *Competitive Markov Decision Process* also called *Stochastic Game* (Filar and Vrieze 1996). However, our problem is more specific, due to the existence of the *Organization* agent, whose actions are interleaved with the farmers' actions (themselves acting in parallel) and do not influence directly the dynamics of the disease, but only the reward functions of the farmers. In turn, the farmers' actions and states do influence the *Organization* rewards. This specific problem is often called *Leader-Follower Markov Decision Process (LF-MDP)* (Tharakunnel and Bhattacharyya 2009) or *Dynamic Principal-Agent Problem* (Plambeck and Zenios 2000).

We propose a Leader-Follower model of the sequential decision problem of disease control. We consider that the number, n, of farms, can be large, but each one can be in only few states (typically, (S)*usceptible* or (I)*nfected*). Furthermore, farmers have only few control actions available

(typically *control* or *do nothing*). We will focus on the finite-horizon case, but the results presented in the paper can be easily extended to stationary infinite-horizon problems with discounted rewards.

When trying to solve disease control problems modeled as Leader-Follower MDPs, we face several computational problems. Dynamic programming algorithms can be applied: backwards induction in the finite-horizon case, or value-iteration (for example) in the infinite-horizon case. However, these iterative algorithms imply solving many n-player non-zero-sum games at each iteration in order to build the policies of the followers (farmers). And solving games is hard, even in the simple non-zero-sum two players case. Therefore, most approaches that solve Leader-Follower games, as well as competitive Markov decision processes, rely on simulation-based *Reinforcement-Learning* approaches (Tharakunnel and Bhattacharyya 2007; 2009; Hu and Wellman 2003; Chalkiadakis and Boutillier 2003).

In this paper, our approach is different. We propose a dynamic programming approach to solve LF-MDPs in the case where the number n of followers can be large but, (i) the number of states of each follower is bounded by a small constant and (ii) the size of the joint state space of the leader is polynomial in n. These assumptions hold naturally in disease management problems, as well as in many other problems of collective sustainable management, in agriculture or ecology.

Furthermore, as our study is mainly directed towards building the Organization's incentive policy for sustainable management of animal disease, we make an additional simplifying assumption about the behavior of followers, that will help building polynomial-time solution algorithms. We will assume that followers are indistinguishable from the point of view of the leader, apart from their infection state. This means that we will assume that any two followers apply the same policy, deciding the same (possibly *mixed*) game policy, when their states are identical. We will evaluate the proposed policies, from the point of view of the followers' and leader's rewards, on a disease management case study which we will describe next.

An animal disease management case study

As an illustration, we consider a simplified version of the problem of coordination of individual decisions to limit the spread of the Porcine Reproductive and Respiratory Syndrome (PRRS) within a group of farms. PRRS is an endemic disease which impacts farm production (Nodelijk 2002). As it is a transmissible disease, choosing independently control actions within each farm may not be sufficient to limit its spread within an area. The producers Organization (leader) proposes incentives each year in order to incite followers to take control actions, thus limiting the PRRS spread within their group of farms. As nearly all animals are infected shortly after the virus introduction in a farm, we consider only two epidemiological states: S (susceptible) and I (Infected). For S farms, control actions consist of biosecurity actions reducing the probability of transmission. For I farms, depopulation can be used to change state back to S, but this action is costly. The Organization aims at limiting the total cost of the disease within the group of farms by giving incentives to producers to implement control actions, which reduces the cost of control actions for the farmer.

A Leader-Follower MDP model for the animal disease management problem

General LF-MDP model

A single leader/multiple followers finite-horizon MDP model (Tharakunnel and Bhattacharyya 2007) is a multiple time-step decision process involving several agents : one *leader* and n *followers*. At each time period $t \in \{1, \dots H\}$, the following steps occur:

- The *leader* makes an incentive decision $a^L \in A^L$, where $A^L = \{1, \dots, m\}$ is a finite number of possible incentives.
- Each follower $i \in \{1, \dots n\}$ chooses its own decision $a_i^F \in A^F = \{1, \dots, p\}$ independently, after having observed the leader's decision.
- The *global state of the system*, $\sigma \in \Sigma$, changes stochastically under the effect of the followers' actions, with transition probability $T(\sigma' | \sigma, a_1^F, \dots, a_n^F)$.
- The leader and the followers receive individual rewards, $r^L(\sigma, a^L, a_1^F, \dots, a_n^F)$ and $r_i^F(\sigma, a^L, a_i^F)$, $i = 1, \dots, n$.

The global state of the system, $\sigma \in \Sigma$, represents the *joint state* of the leader and all the followers. In the most general case, $\Sigma = S^L \times S_1^F \times \dots \times S_n^F$ is a factored state space.

As usual in sequential decision problems, we assume that agents choose their actions according to *policies*, $\delta_t^L, \{\delta_{t,i}^F\}_{i=1\dots n}$, at each time step t. In the most general case, theses policies can be *non-Markovian* and *stochastic*. However, in this paper we will focus on Markovian policies, where:

- $\delta_t^L(a_t^L | \sigma_t)$ is the probability that $a_t^L \in A^L$ is chosen by the leader at time t in the current state, given current state.
- $\delta_{t,i}^F(a_{t,i}^F | \sigma_t, a_t^L)$ is the probability that $a_{t,i}^F \in A^F$ is chosen by follower i at time t in the current state and after having observed the current action of the leader.

Strategies are *deterministic*, when the probabilities take value in $\{0, 1\}$.

Considering fixed policies $\left\{\delta_t^L, \{\delta_{t,i}^F\}_{i=1\dots n}\right\}_{t=1\dots H}$, their *values* $Q^L_{\delta^L, \{\delta_i^F\}}$ and $Q^{F,i}_{\delta^L, \{\delta_i^F\}}$ to the leader and the followers are defined as follows, in every joint state and time step:

$$Q^L_{\delta^L, \{\delta_i^F\}_{i=1\dots n}}(\sigma, t) = E\left[\sum_{t'=t}^{H} r_{t'}^L \,\middle|\, \delta^L, \{\delta_i^F\}, \sigma\right], \quad (1)$$

$$Q^{F,i}_{\delta^L, \{\delta_i^F\}_{i=1\dots n}}(\sigma, t) = E\left[\sum_{t'=t}^{H} r_{t',i}^F \,\middle|\, \delta^L, \{\delta_i^F\}, \sigma\right] \quad (2)$$

Solving a LF-MDP consists in finding *equilibrium policies*, $\delta^{L*}, \{\delta_i^{F*}\}_{i=1\dots n}$, for the leader and the followers. Equilibrium policies are policies from which no one has interest to deviate.

Definition 1 (LF-MDP equilibrium policies). *Strategies $\delta^{L*}, \{\delta_i^{F*}\}_{i=1\ldots n}$ are equilibrium policies if and only if they verify, $\forall \delta^L, \{\delta_i^F\}, \sigma, t$:*

$$Q^L_{\delta^{L*},\{\delta_i^{F*}\}}(\sigma,t) \geq Q^L_{\delta^L,\{\delta_i^{F*}\}}(\sigma,t), \quad (3)$$

$$Q^{F,i}_{\delta^{L*},\{\delta_i^{F*}\}}(\sigma,t) \geq Q^{F,i}_{\delta^{L*},\delta_j^{F*},\{\delta_i^F\}_{i\neq j}}(\sigma,t), \forall j. \quad (4)$$

It has been shown (see e.g. (Filar and Vrieze 1996), (Tharakunnel and Bhattacharyya 2007)), that there exist equilibrium policies for the leader and the followers, which are Markovian. Furthermore, a deterministic equilibrium policy exists for the leader, while equilibrium policies are stochastic, in general, for the followers. Such equilibrium policies can be computed by a *backwards induction* type algorithm (or *value iteration*, for example, in the infinite horizon case), by interleaving Nash equilibrium computation steps for the followers, and classical backwards induction steps for the leader, at each time step.

However, note a few known facts about Nash equilibrium computation:

- A Nash equilibrium in a game can be *pure*, i.e. each agent chooses a fixed action, or stochastic, i.e. each agent chooses a distribution over actions.
- In a n-player non-zero sum game, there may exist several (pure and/or mixed) Nash equilibria, which cannot always be completely ordered by a dominance relation. This holds even in the simple case of a 2-player non-zero sum game.
- Even in the case of a 2-player, non-zero sum game, there exists no known polynomial time Nash equilibrium computation algorithm.
- Furthermore, the state space size Σ is exponential in the size of the description of the problem.

These facts demonstrate that the existence of an *efficient* algorithm for computing equilibrium policies in LF-MDPs is very unlikely.

In the following, we will propose a new restriction of the LF-MDP model which can model disease management problems and which will allow us, to the price of a few simplifying assumptions, to design efficient equilibrium policies computation algorithms, even for large n. This constitutes the main technical contribution of this article.

A simplified LF-MDP model of the animal disease management problem

In the animal disease management problem, the global state of infection will be described by the infection states of all farms, and we will assume a simple S-I-S model (Susceptible/Infected/Susceptible) (Hethcote 2000). That is, $\Sigma = \{S, I\}^n$.

Our main simplifying assumption will be that all the followers are considered "identical". This means that:

(i) the reward functions $r_i^F(\sigma_i, a^L, a_i^F)$ (where $\sigma_i = (s_i, c)$) of the followers are identical and are "local" to each follower, apart from the global incentive a^L and the total number of infected followers[1], c. For the leader, the reward function $r^L(c, a^L)$ is only a function the total number of infected followers, c and on the chosen incentive, a^L.

(ii) The joint transition probability has the following form:

$$T(\sigma'|\sigma, a_1^F, \ldots, a_n^F) = \prod_{i=1}^{n} p(s_i'|\sigma_i, a_i^F). \quad (5)$$

This form amounts to assuming that the infection state transition probabilities of all followers are identical, and only depend on their current state s_i, their own decision a_i^F, and c, the total number of infected followers.

Equilibrium policies in the simplified LF-MDP model

Under our simplifying assumption, we can prove the following proposition about equilibrium policies.

Proposition 1 (Equilibrium policies).

(i) Equilibrium policies are identical for all followers.

(ii) Optimal policies for the leader are of the form : $\delta_t^{L*}(a^L|c)$.

(iii) Equilibrium policies for the followers are of the form : $\delta_{t,i}^{F*}(a_i^F|\sigma_i, a^L) = \delta_{t,i}^{F*}(a_i^F|s_i, c, a^L)$.

This proposition is important especially since its proof will provide us with an efficient algorithm for computing equilibrium policies. The rest of this section will be devoted to this proof.

First, note that (i) holds obviously, for simple reasons of symmetry. An important consequence of (i), and of the known fact that equilibrium policies for LF-MDP in general can be obtained through iterative Nash equilibrium computation and dynamic programming steps, is that a Nash equilibrium policy at time t for the followers can be described by a set of probability vectors $\{\pi^s\}_{s\in S^F}$, where each π^s is a probability vector over A^F. All followers which are in state $s \in S^F$ choose their action according to the same probability vector.

Furthermore, (ii) also holds for symmetry reasons, and from the fact that the leader's reward function $r^L(c, a^L)$ only depends on c and a^L, and not on any s_i in particular.

Then, we can show by backwards induction that (iii) holds for all t.

Final time step. At $t = H$, note that, by equation 2 applied to the simplified case,

$$Q^{F,i}_{\delta^{L*},\{\pi^s\}}(\sigma, H) = \sum_{a_i^F, a^L} \pi^{s_i}(a_i^F)\delta_H^{L*}(a^L|c) r^F(\sigma_i, a^L, a_i^F).$$

[1] In the experiments section, we will consider cases where followers' rewards are of the form $r^F(s_i, s_i', a^L, a_i^F)$. However, by considering *expected rewards* $\bar{r}^F(\sigma_i, a^L, a_i^F) = \sum_{s_i'} p(s_i'|\sigma_i, a_i^F) r^F(s_i, s_i', a^L, a_i^F)$ instead of immediate rewards, we do not change the global value functions, while getting back to the reward form considered here.

From the definition of equilibrium policies, we have for the leader,

$$V^L(c,H) = \max_{a^L} r^L(c,a^L) \text{ and}$$

$\delta_H^{L*}(a^L|c) = 1$ if $a^L = a^{L*} = \arg\max_a r^L(c,a)$ and 0 else.

For the followers, we have to compute a Nash equilibrium, δ^{L*} being fixed:

$$V^{F,i}(\sigma_i,H) = Nash_\pi^i \left(\sum_{a_i^F} \pi^{s_i}(a_i^F) r^F(\sigma_i, a^{L*}, a_i^F) \right).$$

In this expression, $Nash_\pi^i$ denotes the value of a mixed Nash equilibrium $\{\pi^{s_i*}\}_{s_i \in S^F}$, to the player i corresponding to state s_i, in the game defined by the set of reward functions $g_i(a_i^F) = r^F(\sigma_i, a^{L*}, a_i^F)$. But note that followers' rewards are completely independent from other follower's actions. Therefore, the computation of the $Nash_\pi^i$ equilibria amounts to $|S^F|$ independent maximizations:

$$V^{F,i}(\sigma_i,H) = \max_{a_i^F} r^F(\sigma_i, a^{L*}, a_i^F),$$

where $\delta_{H,i}^{F*}(a_i^F|\sigma_i,a^L) = 1$ if $a_i^F = a_i^{F*} = \arg\max_a r^F(\sigma_i, a^{L*}, a)$ and 0 else.

Thus, the policies have the desired form at $t = H$ and furthermore are deterministic, both for the leader and the followers.

Induction step. Now, note that mixed policies of followers are defined by sets of probability vectors $\{\pi^s\}_{s \in S^F}$ which depend on the current state σ, at time t, as well as the action a^L taken by the leader as a function of c.

The value function of any follower i can be written (using hypothesis at time $t+1$ for $V_{t+1}^{F,i}$) :

$$\begin{aligned} V^{F,i}(\sigma,a^L,t) &= Nash_\pi^i \left(\sum_{\{a_j^F\}_j} \prod_j \pi^{s_j}(a_j^F) \times \right. \\ &\quad \Big[r^F(\sigma_i,a^L,a_i^F) + \sum_{\sigma'} T(\sigma'|\sigma,\{a_j^F\}) \\ &\quad \left. V^{F,i}(\sigma', \delta_{t+1}^{L*}(c'), t+1) \Big] \right) \end{aligned}$$

By Bayes rule:

$$T(\sigma'|\sigma,\{a_j^F\}) = p(s_i'|\sigma_i,a_i^F) p(c'|\sigma_i,s_i',\{a_j^F\}_{j\neq i}). \quad (6)$$

Where c' is the number of infected followers at time step $t+1$, without considering follower i. The transition functions $p(s_j'|\sigma_j,a_j^F)$ are inputs of the problem, and $p(c'|\sigma_i,s_i',\{a_j^F\}_{j\neq i})$ can be computed easily from these inputs, albeit tediously.

Now, by replacing $T(\sigma'|\sigma,\{a_j^F\})$ with the right hand side expression of equation 6 and by assuming that (iii) holds, we notice that the value function for follower i has form:

$$V^{F,i}(\sigma_i,a^L,t) = Nash_\pi^i \left(E_{\{\pi^{s_j}\}} \Big[g_i(\sigma_i,a^L,t,\{a_i^F\}) \Big] \right),$$

$$\text{where } g_i(\sigma_i,a^L,t,\{a_i^F\}) = r^F(\sigma_i,a^L,a_i^F) +$$

$$\sum_{\sigma_i'} p(s_i'|\sigma_i,a_i^F) p(c'|\sigma_i,s_i',\{a_j^F\}_{j\neq i}) V^{F,i}(\sigma', \delta_{t+1}^{L*}(c'), t+1)$$

and $\delta_{t,i}^{F*}(a_i^F|\sigma_i,a^L) = \pi^{s_i*}(a_i^F)$, the corresponding probability in the Nash equilibrium mixed policy for s_j has the desired form (iii).

Note that the non-cooperative game between the followers may not have a unique (mixed or pure) equilibrium. However, we do not dwell here on the problem of the choice of the Nash equilibrium to return, which is an important but difficult topic in game theory. In the following Section we will come back to this point in the specific S-I-S case, in which finding followers' equilibrium policies amounts to solving a *bi-matrix game*. We will see in the Experiments Section that, in the S-I-S case we consider, the returned equilibria are, most of the time, pure and unique.

However, let us go on with the induction step, by considering now the leader's case. Considering equation 1, assuming that $\forall t' > t$ equilibrium policies of the form of the induction hypothesis are followed by the leader and followers, as well as at time t for the followers, the Q-function for the leader takes form, at time t:

$$\begin{aligned} Q^L(\sigma,a^L,t) &= r_t^L(c,a^L) \\ &+ \sum_{\sigma'} T(\sigma'|\sigma,a^L,\{\delta_{t,i}^{F*}\}) V^{L*}(c',t+1) \\ &= r_t^L(c,a^L) \\ &+ \sum_{c'} T(c'|\sigma,a^L,\{\delta_{t,i}^{F*}\}) V^{L*}(c',t+1) \end{aligned}$$

Once again, for reasons of symmetries, $T(c'|\sigma,a^L,\{\delta_{t,i}^{F*}\})$ can be rewritten in the form $T(c'|c,a^L,\{\delta_{t,i}^{F*}\})$, and Q^L only depends on c, a^L and t:

$$Q^L(c,a^L,t) = r_t^L(c,a^L) + \sum_{c'} T(c'|c,a^L,\{\delta_{t,i}^{F*}\}) V^{L*}(c',t+1). \quad (7)$$

Thus, $V^L(c,t) = \max_{a^L} Q^L(c,a^L,t) = Q^L(c,a^{L*},t)$ and $\delta_t^{L*} = \arg\max_{a^L} Q^L(c,a^L,t)$ has the desired form.

We now have all the necessary elements for designing a backwards induction dynamic programming algorithm for solving LF-MDP problems, by interleaving *followers' Nash equilibrium* computation steps and polynomial time *leader optimal strategies* computation steps, in our simplified case.

However, before going to the experiments, let us add a few words about the Nash equilibrium computation step, in the S-I-S case.

Bi-matrix games and followers' equilibrium policies computation in the S-I-S case

In the S-I-S case, two functions, $g_S(s_i = S, c, a^L, t, a_S^F, a_I^F)$ and $g_I(s_i = I, c, a^L, t, a_S^F, a_I^F)$, which can be represented by matrices $G_S(a_S^F, a_I^F)$, $G_I(a_S^F, a_I^F)$ for any (c, a^L), represent the value to followers respectively in states S and I of applying respectively actions a_S^F and a_I^F, when the global

number of I followers is c and the leader's incentive is a^L, at time t.

In this case, a *pure equilibrium policy* for the followers is a pair of actions $(\bar{a}_S^F, \bar{a}_I^F)$ which satisfies, for all pairs of followers' actions $(a_S^F, a_I^F) \in A^F \times A^F$:

$G_S(\bar{a}_S^F, \bar{a}_I^F) \geq G_S(a_S^F, \bar{a}_I^F)$ and $G_I(\bar{a}_S^F, \bar{a}_I^F) \geq G_I(\bar{a}_S^F, a_I^F)$.

This problem is called *Bi-matrix Game* (Lemke and Howsion 1964). It does not always admit a pure policy, however, if a pure policy (which may not be unique) exists, it can be found in time polynomial in $|A^F|$. When no pure policies exist, mixed policies (corresponding to stochastic policies $\delta_{t,i}^{F*}(\cdot|s_i, c, a^L)$) can be found by classical algorithms. However, in this case, mixed policies are more complex to compute. (Chen, Deng, and Teng 2009) have shown that this two-player problem is PPAD-complete, where PPAD is an intermediate complexity class between P et NP (provided that these classes be distinct). In the mean time, (Daskalakis, Goldberg, and Papadimitriou 2009) have shown that the problem is also PPAD-complete for a number of players (hence of health status) greater than or equal to 3. Yet, (Lipton, Markakis, and Mehta 2004) have shown that computing ε-Nash optimal mixed policies was *quasi polynomial*.Furthermore, the computed policies have at most k non-zero probability actions.

Taking these considerations into account, we have implemented an algorithm that searches exhaustively for pure equilibria and then, if none or more than 1 equilibria are found, mixed policies are looked for using a *quadratic programming* modeling of a bi-matrix game (Mangasarian and Stone 1964) and the SCILAB built-in solver QUAPRO.

Experiments on the case study

Scenarios We consider a group of n herds (followers). The model is defined by:

- $p(s_i' = I|s_i = S, c, a^F) = \frac{\beta(a^F)*c}{n} + out(a_F)$ with c the number of infected herds, a^F the action applied, β the disease transmission rate and out the external risk (both β and out depend on a^F, as shown in Table 1).
- $p(s_i' = S|s_i = I, a^F) = P_{I\to S}(a^F)$ only depends on the action undertaken in the I herds (Table 1).
- The rewards of the followers are the negated sum of the cost (incentive taken into account) of the control action plus a cost associated to the herd's resulting state.

$$r^F(s_i', a^L, a^F) = -\Big((1 - a^L * percent) * control(a^F) + cost(s_i')\Big).$$

- The reward of the leader is the negated sum of: (i) an incentive cost if $a^L = 1 : c_{incent}$, (ii) the expected control cost of infected herds:

$$C_I(c, a^L) = c * \delta^{F,i*}(a_i^F = control|s_i = I, c, a^L) * percent * control(s_i = I),$$

(iii) the expected control cost of non-infected herds:

$$C_{NI}(c, a^L) = c * \delta^{F,i*}(a_i^F = control|s_i = S, c, a^L) * percent * control(s_i = S),$$

and (iv) penalties for infected herds: $C_p(c) = c * pen(I)$.

$$r^L_{\{\delta^{F,i*}\}}(c, a^L) = -\Big(c_{incent} + c_I(c, a^L) + c_{NI}(c, a^L) + C_p\Big).$$

Note that the rewards have a more complex form than the ones we adopted earlier. However, the backwards induction algorithm we have defined can still be applied with slight modifications.

We computed the equilibrium policies for parameters in Table 1 for an horizon of 50 time-steps, without discount. We then simulated the dynamics under the equilibrium policies to evaluate the leader's use of incentive, the associated cost and the impact on the prevalence of the disease (I/n). As our model is stochastic, we ran 10,000 replications with an initial state of $n - 2$ infected herds.

Parameter	Notations	Value
Transmission rate (action)	β	$\{0.8, 0.4\}$
External risk (action)	out	$\{0.05, 0.025\}$
Transition I to S (action)	$P_{I\to S}$	$\{0.05, 1\}$
Individual cost $\{S, I\}$	$cost$	$\{0, 30\}$
Leader penalties $\{S, I\}$	pen	$\{0, 15\}$
Control cost $\{S, I\}$	$control$	$\{15, 150\}$
Incentive cost	c_{incent}	$\{0, 1\}$
Proportional incentive cost	$percent$	$0.1\|0.25\|0.4$

Table 1: Parameters values

Results First, note that followers policies were always deterministic with the parameters' values we used in the experiments. Figure 1 shows an apparently paradoxical fact. The Organization proposes incentives more often when they represent a larger percentage of the followers' control action costs. This is especially true when the proportion of infected herds is high. It can be explained by looking at the followers' policy (Figure 3). Followers in state S tend to control more when incentives are increased, since control actions are less expensive in this case. Still, note that in state S when we consider an incentive of 40% of the control cost, followers control even without incentive. Varying the number of herds has no effect on the leader and followers' policies nor on the proportion of infected herds (results not shown). Figure 2 (left) shows that increased incentives lead to fewer infected herds, which is rather natural. Figure 2 (right) even shows that, for a sufficiently long horizon, considering an incentive of 40% of control cost leads to less cumulated expected cost for the leader, since the followers are more reactive to incentives.

A remarkable point of the experiments (which could be expected) is that equilibrium policies are more or less stationary, except for the last (and initial) few steps. This allows a simple *rule-based* representation of policies, which is particularly convenient for decision-makers in animal disease management. And, even though simple, these policies improve on the ones currently used, in which incentives are decided unconditionally, without considering the prevalence of the disease.

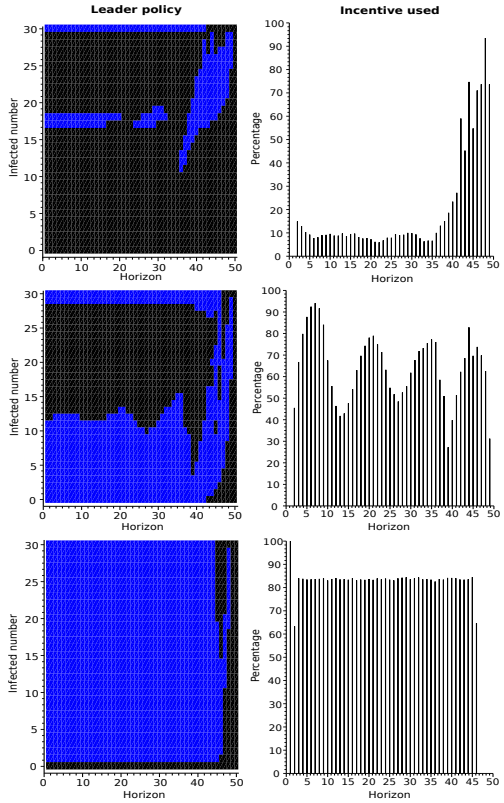

Figure 1: Deterministic policy for the leader (blue areas represent situations when the incentive is retained) and use of the incentive, when incentives represent 10% (top), 25% (middle) and 40% (bottom) of followers' control costs.

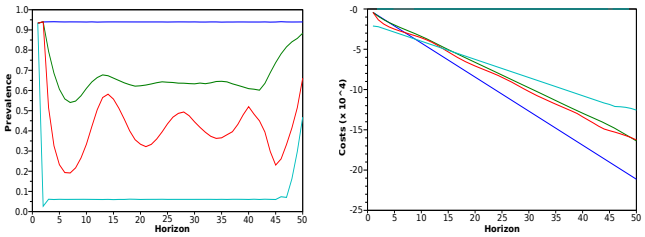

Figure 2: Evolution over time of the mean proportion of infected herds (left) and of the mean total leader cost (right) when incentives represent 0% (blue), 10% (green), 25% (red) and 40% (light blue) of followers' control costs.

Concluding remarks, further work

In this article we have proposed a LF-MDP model for animal disease management in the context of a set of herds and a global organization. Our main technical contribution is an adaptation of the LF-MDP model which allows to efficiently deal with a large (realistic) number of herds, in the case of S-I-S disease models. In the framework of sustainable disease management, our approach offers (i) a way to compute *adaptive* (with respect to the number of infected herds) incentive policies, which is not usual in the domain and (ii) a framework for analyzing leaders incentive policies and their impact on follower's policies. Our work also differs from existing work, such as (Tharakunnel and Bhattacharyya 2007; 2009), in that exact dynamic programming can be applied, when the latter use a reinforcement learning approach.

Immediate extensions to our work would be to consider more realistic disease propagation models, which are generally used in disease management (S-I-R, S-E-I-R models, with potentially different levels of excretion). Our approach would naturally extend to these problems, the main difficulty being to have k-players equilibrium policies to compute. Other natural extensions would be to consider (i) non-homogeneous disease transmission rates (in this case, recent *graph-based* MDP approaches could be applied (Sabbadin, Peyrard, and Forsell 2012)) and (ii) partial state observability. Considering the latter problem would require to consider problems of the form LF-POMDP.

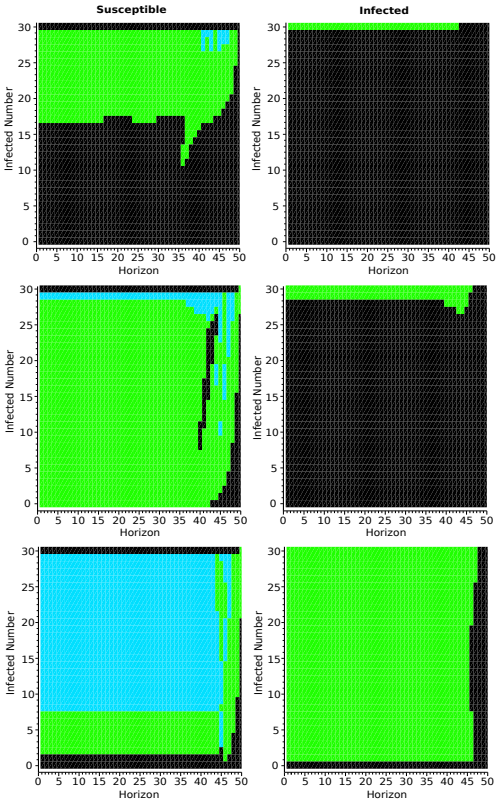

Figure 3: Deterministic policy for followers in states S (left) and I (right) when incentives represent 10% (top), 25% (middle) and 40% (bottom) of control costs: "no action" even when incentive (black), "action" only when incentive (green) and "action" even when no incentive (light blue).

Acknowledgments

This work was supported by the National French Research Agency projects ANR-2010-BINF-07 (MIHMES) and ANR-2010-BLAN-0215-04 (LARDONS).

References

Chalkiadakis, G., and Boutillier, C. 2003. Coordination in multiagent reinforcement learning: A bayesian approach. In *Proc. of AAMAS'03*.

Chen, X.; Deng, X.; and Teng, S.-H. 2009. Settling the complexity of computing two-player nash equilibria. *Journal of the ACM* 56(3).

Daskalakis, C.; Goldberg, P.; and Papadimitriou, C. 2009. The complexity of computing a nash equilibrium. *SIAM Journal on Computing* 39(3):195–259.

Filar, J., and Vrieze, K. 1996. *Competitive Markov Decision Processes*. Springer.

Hethcote, H. W. 2000. The mathematics of infectious diseases. *SIAM review* 42:599–653.

Hu, J., and Wellman, M. P. 2003. Nash q-learning for general-sum stochastic games. *Journal of Machine-Learning Research* 4:1039–1069.

Lemke, C., and Howsion, J. T. 1964. Equilibrium points of bimatrix games. *J. Soc. Ind. Appl. Math.* 12(2):413–423.

Lipton, R.; Markakis, E.; and Mehta, A. 2004. Playing large games using simple strategies. In *Proceedings of the 4th ACM conference on Electronic Commerce*, 36–41.

Mangasarian, O., and Stone, H. 1964. Two-person non-zero sum games and quadratic programming. *Journal of Mathematical Analysis and applications* 9:348–355.

Nodelijk, G. 2002. Porcine reproductive and respiratory syndrome (prrs) with special reference to clinical aspects and diagnosis: A review. *Veterinary Quarterly* 24:95–100.

Plambeck, E. L., and Zenios, S. A. 2000. Performance-based incentives in a dynamic principal-agent model. *Manufacturing and Service Operations Management* 2(3):240–263.

Sabbadin, R.; Peyrard, N.; and Forsell, N. 2012. A framework and a mean-field algorithm for the local control of spatial processes. *International Journal of Approximate Reasoning* 53(1):66–86.

Tharakunnel, K., and Bhattacharyya, S. 2007. Leader-follower semi-markov decision problems: theoretical framework and approximate solution. In *IEEE international conference on Approximate Dynamic Programming and Reinforcement Learning (ADPRL)*, 111–118.

Tharakunnel, K., and Bhattacharyya, S. 2009. Single-leader-multiple-follower games with boundedly rational agents. *Journal of Economic Dynamics and Control* 33:1593–1603.