Markov Decision Process Routing Games

Dan Calderone
Univ. of California, Berkeley
danjc@eecs.berkeley.edu

S. Shankar Sastry
Univ. of California, Berkeley
sastry@eecs.berkeley.edu

ABSTRACT

We explore an extension of nonatomic routing games that we call *Markov decision process routing games* where each agent chooses a transition policy between nodes in a network rather than a path from an origin node to a destination node, i.e. each agent in the population solves a Markov decision process rather than a shortest path problem. We define the appropriate version of a Wardrop equilibrium as well as a potential function for this game in the finite horizon (total reward) case. This work can be thought of as a routing-game-based formulation of continuous population stochastic games (mean-field games or anonymous sequential games). We apply our model to the problem of ridesharing drivers competing for customers.

CCS CONCEPTS

•**Networks** →*Traffic engineering algorithms;* •**Computing methodologies** →**Stochastic games;** •**Theory of computation** →*Network games;* Markov decision processes;

KEYWORDS

Stochastic games, routing games, mean-field games, anonymous sequential games, Markov decision processes

ACM Reference format:
Dan Calderone and S. Shankar Sastry. 2017. Markov Decision Process Routing Games. In *Proceedings of The 8th ACM/IEEE International Conference on Cyber-Physical Systems, Pittsburgh, PA USA, April 2017 (ICCPS 2017),* 7 pages.
DOI: http://dx.doi.org/10.1145/3055004.3055026

Classic routing games [1, 6, 16, 17, 19] are perhaps the best studied examples of continuous population potential games detailed by Sandholm[18]. The strategy choices for agents in the population of a nonatomic routing game are the various routes from their origin to their destination and each agent's goal is to find the shortest route. At the Wardrop or Nash equilibrium of the game, the population is divided up among the routes so the overall population distribution is consistent with the shortest path problem that each individual member is solving, i.e. no mass is allocated to a route with non-minimal latency. Indeed, the potential function is designed so that this is the case at its minimum.

ICCPS 2017, Pittsburgh, PA USA

978-1-4503-4965-9/17/04…$15.00
DOI: http://dx.doi.org/10.1145/3055004.3055026

It is well known that the shortest path problem can be formulated as a linear program. There are also similar linear programming formulations for solving Markov decision processes (MDP) [2, 14, 15]. This suggests that there may be a version of a continuous population potential game on a network where each agent seeks to solve an MDP as opposed to a shortest path problem. In this paper, we define such a game that we call a *Markov decision process routing game*. We define the appropriate Wardrop-type equilibrium concept and show how it can be found by minimizing a potential function in the finite horizon total reward case.

MDP routing games are a specific case of continuous population stochastic games, games where each infinitesimal agent in a population solves a Markov decision process with rewards determined by the actions of the other agents. These games were first introduced as *anonymous sequential games* by Rosenthal and Jovanovic [10]. Results have focused mostly on existence and uniqueness of equilibria [3–5, 20] and specific applications [21, 22]. Recently, stochastic population games have been studied in the mean-field game community starting with Lasry and Lions in 2006 [11–13]. Our formulation bears closest resemblance to mean-field games on graphs [7–9]. The standard mean-field model consists of a pair of coupled partial differential equations (PDEs): one backward time PDE that defines the value function or "cost-to-go" for the population of agents and one forward time PDE that defines the mass evolution of the population. As in our case, when the costs agents consider can be written as the gradient of some functional, the mean-field game is called a potential game and both PDEs can be solved by solving a single optimal control problem (similar to the finite-horizon optimization problem presented in Section 1). A significant difference between our formulation and mean-field games on graphs is that in our formulation the potential function derivative condition is defined with respect to the mass of the population taking a specific action as opposed to the mass of the population at a given node (compare Section 1 of this paper with the potential game formulations in [7, 8]).

To illustrate an application of this game, we consider the problem faced by drivers who provide ridesharing services such as Uber or Lyft drivers. Ridesharing services are fast becoming a huge part of transportation in urban areas. As members in the transportation market, ridesharing drivers have a clear incentive to try to optimize their driving strategy in order to maximize their profits. In order to employ more than just heuristics in this optimization process, drivers must consider the jobs that they take over the entire time horizon, taking into account both the fare that they will receive on an individual trip as well as how the destination of that trip will position them to take advantage of the next job. For example, a driver in the downtown area of a city could have the option of taking a lucrative job that takes them out of the city to a residential area but they should also consider how easy it will be to find a job in the residential area or if they will have to waste time and money

driving back into the city with no passenger. In order to optimize their profits, drivers should optimize over the entire available time horizon.

Another source of complexity is that drivers are competing with each other for jobs. The reward a driver receives for a given trip depends on the fare they receive as well as the cost of fuel, the time they spend making the trip, and the time they spending waiting for a rider. If an individual area becomes crowded with drivers, they will have to wait significantly longer in order to get the job they want.

This competition naturally gives rise to a game where drivers seek to optimize their profits over some period of time by choosing a transition strategy throughout the network and their rewards depend on their own transition strategy as well as the transition strategies of the other drivers in the population. We return to this example in Section 2.

The body of the paper is organized as follows. In Section 1, we detail the MDP routing game in the finite horizon case. In Section 2, we use the model to solve a simulation of the ridesharing driver game and in Section 3, we conclude and comment on future work.

1 MARKOV DECISION PROCESS ROUTING GAMES - FINITE HORIZON CASE

We now define the *Markov decision process (MDP) routing game* in the finite-horizon total reward case. Let $\mathcal{G} = (\mathcal{N}, \mathcal{E})$ be a graph structure with a set of nodes (or states) and edges . The game is played over the course of time horizon T divided into discrete intervals $t \in \{0, 1, \dots, T\}$. Let $\mathcal{A}_j^t$ be the set of actions available at node j at time step t with associated transition probabilities denoted $P_j^{at} = (P_{ij}^{at})_{i \in \mathcal{N}}$. Each P_j^{at} satisfies

$$\sum_i P_{ij}^{at} = 1, \qquad P_{ij}^{at} \geq 0 \tag{1}$$

We will use $\mathcal{A} = \cup_{t,j} \mathcal{A}_j^t$ to refer to the set of all actions at all time steps with the understanding that $a \in \mathcal{A}$ is only available to agents at a specific node at a specific time. Let x_j^t refers to the portion of the population at node j at time t and x_j^{at} the subpopulation of x_j^t that choose action $a \in \mathcal{A}_j^t$.

$$x_j^t = \sum_{a \in \mathcal{A}_j^t} x_j^{at} \tag{2}$$

We will use $x \in \mathbb{R}_+^{|\mathcal{A}|}$ to refer to the vector of all subpopulations x_j^{at}, i.e. the whole population distribution over the state space at all times.

We can compute the total population transitioning from j to i at time t as

$$x_{ij}^t = \sum_{a \in \mathcal{A}_j^t} P_{ij}^{at} x_j^{at} \tag{3}$$

Summing over all nodes j gives us the new population in node i at the next time step

$$\sum_{a \in \mathcal{A}_i^{t+1}} x_i^{a(t+1)} = x_i^{t+1} = \sum_j x_{ij}^t = \sum_j \sum_{a \in \mathcal{A}_j^t} P_{ij}^{at} x_j^{at}. \tag{4}$$

In addition, the initial population in each node is given a priori.

$$\sum_{a \in \mathcal{A}_i^0} x_i^{a0} = x_i^0 = m_i \tag{5}$$

We define reward functions on each action, $\mathrm{R}_j^{at}(x)$, and each transition, $R_{ij}^t(x)$, at each time step that depend on the population distribution x. Note that this dependence of the reward functions on the population distribution introduces competition into the scenario.

Each individual agent in the population seeks to maximize their expected total reward over the entire time horizon which is the sum of the rewards they receive at each transition. At each transition, an agent must consider the immediate reward they expect from the action as well as the expected reward-to-go from the state they transition to. For a given population distribution x, let $v_j^t(x)$ be the expected optimal reward-to-go (or value function) from node j at time t which can be defined backwards recursively from $t = T$ as

$$v_j^T(x) = \max_{a \in \mathcal{A}_j^T} \left(\sum_i P_{ij}^{aT} R_{ij}^T(x) + \mathrm{R}_j^{aT}(x) \right) \tag{6}$$

$$v_j^{t-1}(x) = \max_{a \in \mathcal{A}_j^{t-1}} \left(\sum_i P_{ij}^{a(t-1)} \left[R_{ij}^{t-1}(x) + v_i^t(x) \right] + \mathrm{R}_j^{a(t-1)}(x) \right) \tag{7}$$

Note that $v_j^t(x)$ is dependent on x. We define the utility of taking action a from node j at time t as

$$u_j^{aT}(x) = \sum_i P_{ij}^{aT} R_{ij}^T(x) + R_j^{aT}(x) \qquad t = T \tag{8}$$

$$u_j^{at}(x) = \sum_i P_{ij}^{at} \left[R_{ij}^t(x) + v_i^{t+1}(x) \right] + R_j^{at}(x) \quad t < T \tag{9}$$

We now define the appropriate Wardrop equilibrium concept for the finite-horizon MDP routing game.

Definition 1.1 (Finite-Horizon Wardrop Equilibrium). We say that the population distribution x is a *finite-horizon Wardrop equilibrium* of the finite-horizon MDP routing game if at every node j at every time step t for any two actions $a, a' \in \mathcal{A}_j^t$ such that $x_j^{at} > 0$

$$u_j^{at}(x) \geq u_j^{a't}(x) \tag{10}$$

Note that this is the standard definition of Wardrop equilibria applied to the appropriate decision that agents make at each node and at each time step. Intuitively, no agent has an incentive to deviate from their chosen strategy at each transition. We now define the notion of a potential game for the finite-horizon MDP routing game.

Definition 1.2 (Finite-horizon potential game). We say the finite-horizon MDP routing game is a *potential game* if there exists a C^1 function $F : x \mapsto \mathbb{R}$ such that

$$\frac{\partial F}{\partial x_j^{at}}(x) = \sum_i R_{ij}^t(x) P_{ij}^{at} + \mathrm{R}_j^{at}(x) \tag{11}$$

for each element x_j^{at} of x.

The derivative of the potential function captures the immediate payoff of an action. If we write $F(\cdot)$ as a function of x_{ij}^t as well as x_j^{at} and satisfy

$$\frac{\partial F}{\partial x_j^{at}} = \mathrm{R}_j^{at}(x) \tag{12}$$

$$\frac{\partial F}{\partial x_{ij}^t} = R_{ij}^t(x) \tag{13}$$

then, Condition (11) is satisfied by applying the chain rule and Equation (3).

Remark 1. *In the special case where each function $R_{ij}^t(\cdot)$ is simply a function of x_{ij}^t and $\mathrm{R}_j^{at}(\cdot)$ is simply a function of x_j^{at}, we can use the potential*

$$F(x) = \sum_t \left[\sum_{ij} \int_0^{x_{ij}^t} R_{ij}^t(u)\, du + \sum_j \sum_{a \in \mathcal{A}_j} \int_0^{x_j^{at}} \mathrm{R}_j^{at}(u)\, du \right] \tag{14}$$

which bears distinct resemblance to the classic routing game potential.

Remark 2. *This definition of a potential function is a substantial deviation from mean-field games on graphs where the potential function differentiation condition is defined with respect to the mass on the nodes as opposed to the mass taking a particular action. (See [7, 8] for details.)*

We now show that a mass distribution x minimizes the potential function if and only if it is Wardrop-equilibrium for the finite-horizon game.

Theorem 1.3. *Given a potential function F for the finite-horizon MDP routing game, x satisfies the KKT first order necessary conditions for maximizing F if and only if x is a Wardrop equilibrium.*

Proof. ($\Rightarrow$) The optimization problem and corresponding Lagrangian for maximizing the potential function are given by

$$\max_{x \geq 0} \quad F(x) \tag{15a}$$

$$\text{s.t.} \quad \sum_{a \in \mathcal{A}_i^0} x_i^{a0} = m_i \tag{15b}$$

$$\sum_{a \in \mathcal{A}_i^{t+1}} x_i^{a(t+1)} = \sum_j \sum_{a \in \mathcal{A}_j^t} P_{ij}^{at} x_j^{at} \quad t < T \tag{15c}$$

and

$$L(x, \mu, \pi) = F(x) - \sum_i \pi_i^0 \left(\sum_{a \in \mathcal{A}_i^0} x_i^{a0} - m_i \right) - \sum_{t=0}^{T-1} \sum_i \pi_i^{t+1} \left(\sum_{a \in \mathcal{A}_i^{t+1}} x_i^{a(t+1)} - \sum_j \sum_{a \in \mathcal{A}_j^t} P_{ij}^{at} x_j^{at} \right) + \sum_t \sum_j \sum_{a \in \mathcal{A}_j^t} \mu_j^{at} x_j^{at} \tag{16}$$

where π are the Lagrange multipliers associated with (15b) and (15c) and μ is associated with $x \geq 0$. Computing the KKT first-order necessary conditions given that $F(x)$ is a potential function gives

$$\sum_i P_{ij}^{at} R_{ij}^t(x) + \mathrm{R}_j^{at}(x) - \pi_j^t + \sum_i \pi_i^{t+1} P_{ij}^{at} + \mu_j^{at} = 0, \tag{17a}$$

$$\sum_i P_{ij}^{aT} R_{ij}^T(x) + \mathrm{R}_j^{aT}(x) - \pi_j^T + \mu_j^{aT} = 0, \tag{17b}$$

$\mu_j^{at} \geq 0$, and $\mu_j^{at} x_j^{at} = 0$.
(In the following, we suppress the dependence of $R_{ij}^t(x)$, $\mathrm{R}_j^{at}(x)$, $u_j^{at}(x)$ and $v_j^t(x)$ on x for notational simplicity.)
Starting at $t = T$, we have that

$$\begin{aligned} \pi_j^T &= \sum_i P_{ij}^{aT} R_{ij}^T + \mathrm{R}_j^{aT} + \mu_j^{aT} \\ &= u_j^{aT} + \mu_j^{aT} \end{aligned}$$

For any two actions $a_1, a_2 \in \mathcal{A}_j^T$ such that $x_j^{a_1 T} > 0$, we have that $\mu_j^{a_1 T} = 0$ and $\mu_j^{a_2 T} \geq 0$ and thus

$$\pi_j^T = u_j^{a_1 T} \geq u_j^{a_2 T}$$

Thus, we have that Condition (10) is satisfied at $t = T$. We also have that

$$\pi_j^T \geq \max_{a \in \mathcal{A}_j^T} u_j^{aT} = v_j^T \tag{18}$$

We would have equality except for the possibility that $x_j^T = 0$ and thus $x_j^{aT} = 0$ for all $a \in \mathcal{A}_j^T$. In this case, $\mu_j^{aT} > 0$ for all a and π_j^T could be shifted up by an arbitrary amount. In the case where $x_j^T > 0$ however, there must exist $a_1 \in \mathcal{A}_j^T$ such that $x_j^{a_1 T} > 0$. It follows that $\mu_j^{a_1 T} = 0$ and

$$\pi_j^T = u_j^{a_1 T} = \max_{a \in \mathcal{A}_j^T} u_j^{aT} = v_j^T. \tag{19}$$

Thus, we have that π_j^T is an upper bound on the optimal reward to go from node j at time T with equality achieved whenever $x_j^T > 0$.

Moving on to $t = T - 1$ by Equation (17a), we have that

$$\begin{aligned} \pi_j^{T-1} &= \sum_i P_{ij}^{a(T-1)} \left[R_{ij}^{T-1} + \pi_i^T \right] + \mathrm{R}_j^{a(T-1)} + \mu_j^{a(T-1)} \\ &\geq \sum_i P_{ij}^{a(T-1)} \left[R_{ij}^{T-1} + v_i^T \right] + \mathrm{R}_j^{a(T-1)} + \mu_j^{a(T-1)} \\ &\geq u_j^{a(T-1)} + \mu_j^{a(T-1)} \end{aligned} \tag{20}$$

for all $a \in \mathcal{A}_j^{T-1}$

If $x_j^{T-1} > 0$, for any two actions $a_1, a_2 \in \mathcal{A}_j^{T-1}$ such that $x_j^{a_1(T-1)} > 0$, we have that $\mu_j^{a_1(T-1)} = 0$ and $\mu_j^{a_2(T-1)} \geq 0$. For any i such that $P_{ij}^{a_1(T-1)} > 0$, $x_i^T > 0$ and thus $\pi_i^T = v_i^T$ by (19). It follows that

$$\pi_j^{T-1} = u_j^{a_1(T-1)} \tag{21}$$

Thus we have that

$$\pi_j^{T-1} = u_j^{a_1(T-1)} \geq u_j^{a_2(T-1)} \tag{22}$$

which is Condition (10) at time $t = T - 1$. In addition we have that

$$\pi_j^{T-1} \geq \max_{a \in \mathcal{A}_j^{T-1}} u_j^{a(T-1)} = v_j^{T-1} \tag{23}$$

with equality achieved whenever $x_j^{T-1} > 0$. The result follows by induction.

($\Leftarrow$) Conversely suppose x satisfies Condition (10). The primal feasibility conditions are satisfied a priori. We need to construct dual variables π_j^t and μ_j^{at} that satisfy dual feasibility, complementary slackness, and the gradient condition at each time step.

Starting from $t = T$ for each j, define

$$\pi_j^T = \max_{a \in \mathcal{A}_j^T} \sum_i P_{ij}^{aT} R_{ij}^T + \mathrm{R}_j^{aT} \tag{24}$$

$$= \max_{a \in \mathcal{A}_j^T} u_j^{aT} = v_j^T \tag{25}$$

By Condition (10), we have that $\pi_j^T = u_j^{a_1 T}$ for all $a_1 \in \mathcal{A}_j^T$ such that $x_j^{a_1 T} > 0$ and $\pi_j^T \geq u_j^{aT}$ if $x_j^{a_2 T} = 0$; thus, setting

$$\mu_j^{aT} = \pi_j^T - u_j^{aT}(x) \tag{26}$$

$$= \pi_j^T - \sum_i P_{ij}^{aT} R_{ij}^T(x) - \mathrm{R}_j^{aT}(x) \tag{27}$$

satisfies dual feasibility, complementary slackness, and the gradient constraint for $t = T$.

Moving on to $t = T - 1$, since π_j^T is the optimal reward-to-go from j at time T, we have that

$$u_j^{a(T-1)}(x) = \sum_i P_{ij}^{a(T-1)} \left[R_{ij}^{T-1} + \pi_i^T \right] + \mathrm{R}_j^{a(T-1)} \tag{28}$$

Let

$$\pi_j^{T-1} = \max_{a \in \mathcal{A}_j^{T-1}} \sum_i P_{ij}^{a(T-1)} \left[R_{ij}^{T-1} + \pi_i^T \right] + \mathrm{R}_j^{a(T-1)} \tag{29}$$

$$= \max_{a \in \mathcal{A}_j^{T-1}} u_j^{a(T-1)} = v_j^{T-1} \tag{30}$$

By Condition (10), $\pi_j^{T-1} = u_j^{a_1(T-1)}$ if $x_j^{a_1(T-1)} > 0$ and $\pi_j^{T-1} \geq u_j^{a_2(T-1)}$ if $x_j^{a_2(T-1)} = 0$; thus, setting

$$\mu_j^{a(T-1)} = \pi_j^{T-1} - u_j^{a(T-1)} \tag{31}$$

satisfies dual feasibility and complementary slackness and by substituting Equation (28) into Equation (31) gives the gradient condition. □

Remark 3. *In the finite-horizon case if the transitions are fully deterministic, the problem could be framed as a classic routing game by making T copies of the state space, connecting the proper nodes (where transitions are allowed) between each time step, and then enumerating all possible paths through this new network. It should be noted that this is not possible however when the actions are probabilistic. It might be the case for a specific sequence of transitions that a realization of the first t transitions make the $t+1$ transition impossible even if it would have been possible for another realization. Indeed, this might be true for* all *possible sequences of transitions available to an agent.*

Rate	Velocity	Fuel Price	Fuel Eff
\$6 /mi	8 mph	\$2.5/gal	20 mi/gal

Table 1: Common values for reward function calculations

2 RIDESHARING GAME SIMULATION

To illustrate the model, we run a simulation of the ridesharing driver game as it might be played in downtown San Francisco on a weekend night. We abstract the city as a set of neighborhoods (nodes) that ridesharing drivers travel between. We assume the graph is fully connected and that all transitions are fully deterministic. x_{ij}^t is the population of drivers transitioning from neighborhood j to neighborhood i at time t. The reward functions that drivers consider are influenced by multiple factors including the fare they receive, their fuel costs, the time they spend traveling, and the time they spend waiting for customers. We use linear transition rewards of the form

$$R_{ij}^t(x_{ij}^t) = M_{ij}^t - (C_{ij}^t)_{\text{travel}} - (C_{ij}^t)_{\text{wait}} x_{ij}^t \tag{32}$$

Since the transitions are deterministic, there is no need to define separate rewards on the actions ($\mathrm{R}_j^{at}(x) = 0$). The monetary reward of a trip M_{ij}^t has the form

$$M_{ij}^t = k \cdot \underbrace{\left(\text{Rate}\right)}_{\text{\$/mi}} \cdot \underbrace{\left(\text{Dist}\right)}_{\text{mi}} \tag{33}$$

where k is the surge pricing multiplication factor. The travel cost of the trip consists of travel time plus fuel costs.

$$(C_{ij}^t)_{\text{travel}} = \tau \cdot \underbrace{\left(\text{Dist}\right)}_{\text{mi}} \cdot \underbrace{\left(\text{Vel}\right)^{-1}}_{\text{hr/mi}} + \underbrace{\left(\frac{\text{Fuel}}{\text{Price}}\right)}_{\text{\$/gal}} \cdot \underbrace{\left(\frac{\text{Fuel}}{\text{Eff}}\right)^{-1}}_{\text{gal/mi}} \cdot \underbrace{\left(\text{Dist}\right)}_{\text{mi}} \tag{34}$$

where τ is a time-money tradeoff parameter which we calculate by multiplying the ride rate (\$/mi) times the average distance between neighborhoods times the length of one time interval (20 min), assuming one trip per time interval. The last portion of the reward is the cost of waiting for jobs that depends on the other ridesharing drivers attempting to make the same transition. The coefficient $(C_{ij}^t)_{\text{wait}}$ has units of \$ / driver and is defined as

$$(C_{ij}^t)_{\text{wait}} = \tau \cdot \underbrace{\left(\frac{1}{\frac{\text{Customer}}{\text{Demand Rate}}} \right)}_{\text{hr/rides}} \tag{35}$$

The values that are not specifically edge dependent are listed in Table 1.

We simulate the activity of ridesharing drivers in San Francisco over the course of a weekend evening from 7 pm to 1 am with every time step representing 20 min. A population of 20 drivers starts at each node.

The various neighborhoods (modeled as nodes, and divided loosely into downtown and residential neighborhoods) are shown in Figure 1 and are listed in Table 2. We assume that throughout the night there are at least a few customers (10 customers/hr) who want to travel between any two nodes. During the first few hours,

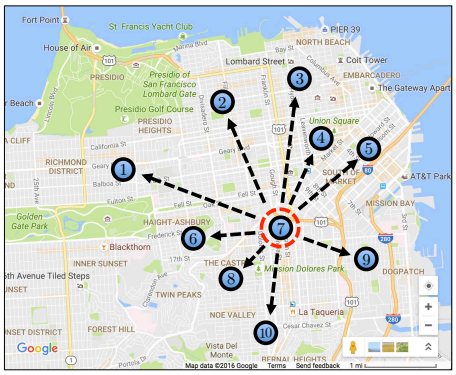

Figure 1: Neighborhoods in San Francisco

#	Neighborhood	Type
1	Richmond	Resident
2	Presidio	Resident
3	North Beach	Downtown
4	Union Square	Downtown
5	S. of Market	Downtown
6	Haight-Ashbury	Resident
7	Mission	Downtown
8	Castro	Resident
9	Dogpatch	Resident
10	Noe Valley	Resident

Table 2: Neighborhood Types

Rates $\left(\frac{\text{rides}}{\text{hr}}\right)$	Resident to Downtown	Downtown to Downtown	Downtown to Resident	Resident to Resident
7 pm – 9 pm	300	100	10	20
9 pm – 11 pm	100	200	100	20
11 pm – 1 am	10	50	300	20

Table 3: Customer demand rates (rides/ hr)

most customers are traveling from residential neighborhoods to downtown neighborhoods. As the evening progresses, more customers are looking for rides among downtown neighborhoods, and then towards the end of the evening, most customers are looking to travel back to residential neighborhoods. The demand for rides between each of the different types of neighborhoods is detailed in Table 3. We also add a surge pricing factor of 2 between the downtown nodes from 9-11pm and a surge pricing factor of 3 from downtown to residential nodes from 11pm-1am. We note that all the values in this simulation could be chosen much more accurately given google maps data and driver demand data. We solve the game by optimizing the potential function given in (14).

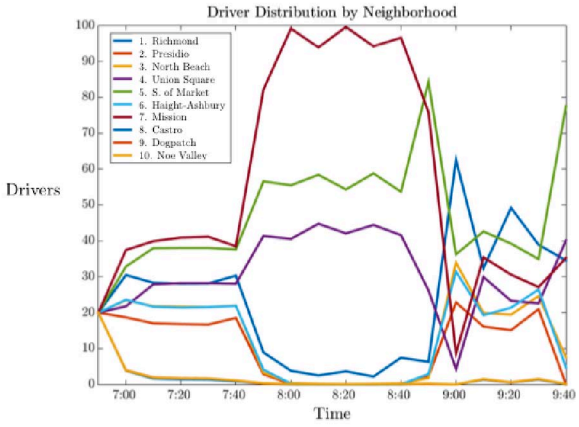

Figure 2: Population of drivers in each neighborhood at each time.

In Figure 2, we show the population of drivers in each neighborhood over the course of the evening. Low numbers of drivers in certain neighborhoods could indicate a need to adjust the fares in those neighborhoods to maintain service for all customers throughout the evening.

Given the population distribution at equilibrium, there are many optimal routes that drivers starting from each node can take over the course of the evening. In Figures 3 and 4, we show the running reward and cumulative average reward for several optimal routes as well as several random routes starting from Node 1. Note that for the optimal routes, the instantaneous running reward that drivers experience might go down or even be negative at one time step in order to set up for a large reward in the future. We note also that while the cumulative averages for each optimal route vary separately over time, they all become equal at the final time step. This has to be true at equilibrium for any two optimal routes starting from the same node. Optimal routes starting at different nodes could have different total rewards. As expected, random routes achieve significantly less total reward over the time horizon. Two of the optimal traces are shown in Figures 5 and 6.

Finally, we consider the decision that drivers make at an individual node at a specific time step. In Figures 7 and 8, we show decision criteria that drivers face at node 7 (the Mission, Figure 1) at time steps $t = 9$ and $t = 17$. This decision criteria includes both the immediate reward for a specific transition and the expected reward-to-go. Notice that population mass is only distributed among transition choices that achieve the maximum expected reward.

3 CONCLUSION

We have presented a new kind of routing game that models agents solving a Markov decision process rather than a shortest path problem. This formulation provides clear connections between traditional routing games and continuous population stochastic games. In this paper, we present the finite-horizon total reward case and used our model to explore the decisions of ride-sharing drivers competing for customers. Future work includes exploring

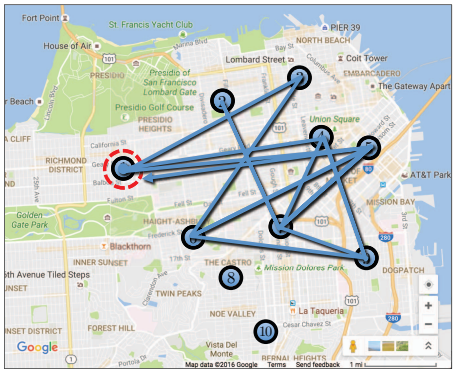

Figure 5: Trace 1

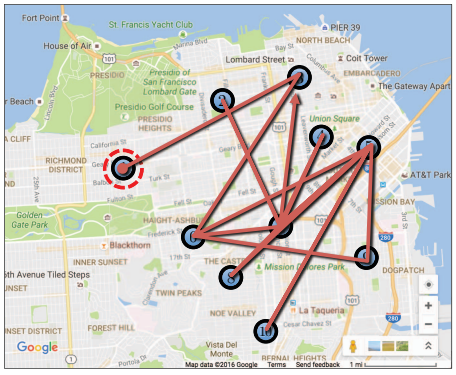

Figure 6: Trace 2

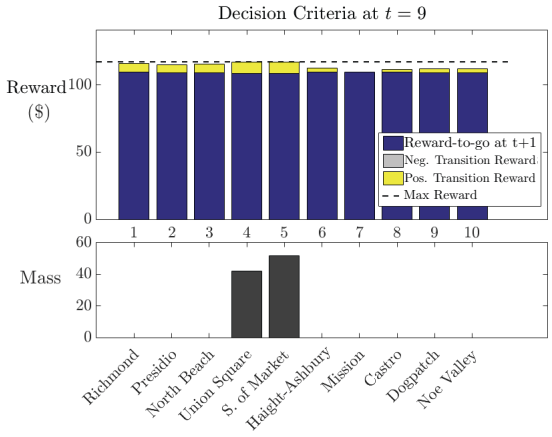

Figure 7: Decision criteria at node 7 (the Mission) at $t = 9$ showing the immediate reward and expected reward-to-go for the transitions to other neighborhoods and the portion of mass that chooses each transition.

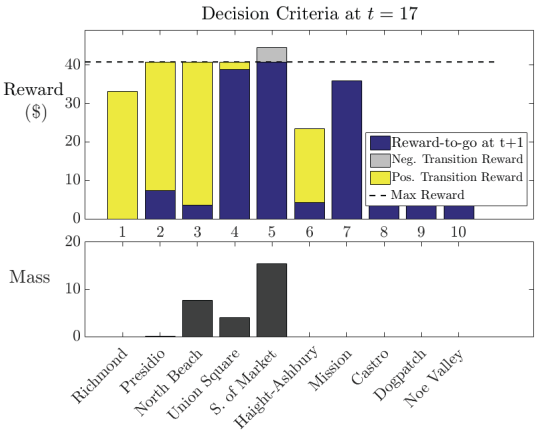

Figure 8: Decision criteria at node 7 (the Mission) at $t = 17$ showing the immediate reward and expected reward-to-go for the transitions to the other neighborhoods and the portion of mass that chooses each transition.

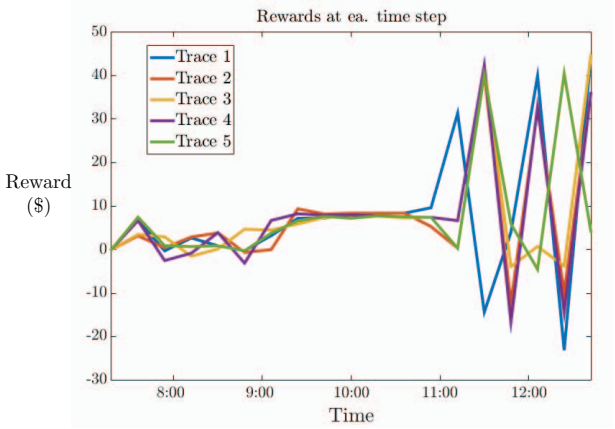

Figure 3: Running reward for various routes.

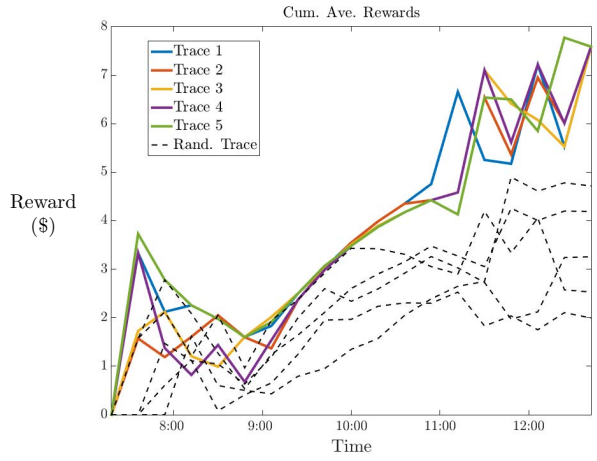

Figure 4: Cumulative average reward for various routes.

the infinite-horizon case (both the average expected and discounted reward cases) and examining traditional routing game concepts (such as price of anarchy and Braess's paradox) in the MDP routing game framework.

ACKNOWLEDGMENTS

This work is supported by NSF FORCES (Foundations of Resilient Cyber-physical Systems) CNS-1239166, AFOSR MURI CHASE FA9550-1 0-1-0567, and ONR N00014-09-1-0230.

REFERENCES

[1] Martin Beckmann, CB McGuire, and Christopher B Winsten. 1956. *Studies in the Economics of Transportation.* Technical Report.

[2] David Bello and German Riano. 2006. Linear programming solvers for markov decision processes. In *Systems and Information Engineering Design Symposium.* 90–95.

[3] James Bergin and Dan Bernhardt. 1992. Anonymous sequential games with aggregate uncertainty. *Journal of Mathematical Economics* 21, 6 (1992), 543–562.

[4] James Bergin and Dan Bernhardt. 1995. Anonymous sequential games: existence and characterization of equilibria. *Economic Theory* 5, 3 (1995), 461–489.

[5] James Bergin, Dan Bernhardt, and others. 1991. *Anonymous sequential games with general state space.* Technical Report.

[6] Stella C Dafermos and Frederick T Sparrow. 1969. The traffic assignment problem for a general network. *Journal of Research of the National Bureau of Standards, Series B* 73, 2 (1969), 91–118.

[7] Diogo A Gomes, Joana Mohr, and Rafael Rigao Souza. 2010. Discrete time, finite state space mean field games. *Journal de mathématiques pures et appliquées* 93, 3 (2010), 308–328.

[8] Olivier Guéant. 2011. From infinity to one: The reduction of some mean field games to a global control problem. *arXiv preprint arXiv:1110.3441* (2011).

[9] Olivier Guéant. 2015. Existence and uniqueness result for mean field games with congestion effect on graphs. *Applied Mathematics & Optimization* 72, 2 (2015), 291–303.

[10] Boyan Jovanovic and Robert W Rosenthal. 1988. Anonymous sequential games. *Journal of Mathematical Economics* 17, 1 (1988), 77–87.

[11] Jean-Michel Lasry and Pierre-Louis Lions. 2006. Jeux à champ moyen. i–le cas stationnaire. *Comptes Rendus Mathématique* 343, 9 (2006), 619–625.

[12] Jean-Michel Lasry and Pierre-Louis Lions. 2006. Jeux à champ moyen. II–Horizon fini et contrôle optimal. *Comptes Rendus Mathématique* 343, 10 (2006), 679–684.

[13] Jean-Michel Lasry and Pierre-Louis Lions. 2007. Mean field games. *Japanese Journal of Mathematics* 2, 1 (2007), 229–260.

[14] Alan S Manne. 1960. Linear programming and sequential decisions. *Management Science* 6, 3 (1960), 259–267.

[15] John L Nazareth and Ram B Kulkarni. 1986. Linear programming formulations of Markov decision processes. *Operations research letters* 5, 1 (1986), 13–16.

[16] Michael Patriksson. 2015. *The traffic assignment problem: models and methods.* Courier Dover Publications.

[17] Tim Roughgarden. 2005. *Selfish routing and the price of anarchy.* Vol. 174. MIT press Cambridge.

[18] William H Sandholm. 2001. Potential games with continuous player sets. *Journal of Economic Theory* 97, 1 (2001), 81–108.

[19] John Glen Wardrop. 1952. ROAD PAPER. SOME THEORETICAL ASPECTS OF ROAD TRAFFIC RESEARCH.. In *ICE Proceedings: Engineering Divisions*, Vol. 1. Thomas Telford, 325–362.

[20] Piotr Wiecek and Eitan Altman. 2015. Stationary anonymous sequential games with undiscounted rewards. *Journal of optimization theory and applications* 166, 2 (2015), 686–710.

[21] Piotr Wiecek, Eitan Altman, and Yezekael Hayel. 2009. An Anonymous Sequential Game Approach for Battery State Dependent Power Control. In *International Conference on Network Control and Optimization.* Springer, 121–136.

[22] Piotr Wiecek, Eitan Altman, and Yezekael Hayel. 2011. Stochastic state dependent population games in wireless communication. *IEEE Trans. Automat. Control* 56, 3 (2011), 492–505.