This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSIPN.2017.2682799, IEEE Transactions on Signal and Information Processing over Networks

Hidden Chinese Restaurant Game: Grand Information Extraction for Stochastic Network Learning

Chih-Yu Wang, *Member, IEEE*, Yan Chen, *Senior Member, IEEE*, K. J. Ray Liu, *Fellow, IEEE*

***Abstract*—Agents in networks often encounter circumstances requiring them to make decisions. Nevertheless, the effectiveness of the decisions may be uncertain due to the unknown system state and the uncontrollable externality. The uncertainty can be eliminated through learning from information sources, such as user-generated contents or revealed actions. Nevertheless, the user-generated contents could be untrustworthy since other agents may maliciously create misleading contents for their selfish interests. The passively-revealed actions are potentially more trustworthy and also easier to be gathered through simple observations. In this paper, we propose a new stochastic game-theoretic framework, Hidden Chinese Restaurant Game (H-CRG), to utilize the passively-revealed actions in stochastic social learning process. We propose Grand Information Extraction, a novel Bayesian belief extraction process, to extract the belief on the hidden information directly from the observed actions. We utilize the coupling relation between belief and policy to transform the original continuous belief state Markov Decision Process (MDP) into a discrete-state MDP. The optimal policy is then analyzed in both centralized and game-theoretic approaches. We demonstrate how the proposed H-CRG can be applied to the channel access problem in cognitive radio networks. We then conduct data-driven simulations using the CRAWDAD Dartmouth campus WLAN trace. The simulation results show that the equilibrium strategy derived in H-CRG provides higher expected utilities for new users and maintains a reasonable high social welfare comparing with other candidate strategies.**

I. Introduction

Agents in networks often encounter circumstances requiring them to make decisions. For examples, they decide which smartphone to buy when the last one is broken, the restaurant to have a meal when feeling hungry, or the music to listen when feeling lonely. Their choices affect their utility, or their measurement of their enjoyments on the outcome. A rational agent should seek the best decisions in order to maximize their utility given the potential influences of certain choices. Nevertheless, sometimes the influence of the choices is uncertain due to 1) unknown parameters such as the quality of a meal, and 2) the external factors such as the unfamiliar crowd within the same restaurant. A learning process is necessary for rational agents to acquire the knowledge of those uncertain parameters and external factors in order to make the right decisions.

Social learning is a learning technique which utilizes the information revealed or shared in a network to acquire certain knowledge. A typical example is choosing the smartphone on the market from dozens of choices. Customers usually do not have enough knowledge, such as the quality and usability of the smartphones, to make the right choice. One may construct the knowledge by learning from advertisements, her own experience from previous purchases, reviews or discussions shared in social networks, or some statistics on the sold quantity of each smartphone to date. All the gathered information helps the agent to construct the necessary knowledge, due to which the accuracy of the agent's decision can be enhanced. The information generated or revealed by other agents is shared through the links constructed by the social relation. Since each agent may have different social relations with others and make decisions at different time, the information one agent received may be different from others.

Chinese Restaurant Game [1], [2] is a game-theoretic framework for modeling social learning process in a system with network externality, i.e., the decisions of users affect each other. Let us consider a Chinese restaurant with multiple tables in different sizes, where customers arrive and request seats from these tables. Customers may prefer bigger space for a comfortable dining experience, or more companies for chatting. In either case, a customer's dining experience is influenced by other customers who share the same table with him, which is a kind of network externality. The goal of Chinese Restaurant Game is to understand how rational agents choose the tables wisely to enhance the dining experience, i.e. to maximize their utilities. The social learning process is introduced in sequential Chinese restaurant game [2], where the table sizes are unknown to the customers. The customers may learn the sizes with signals, either received by oneself, revealed by others, or both. In such a setting, rational customers learn the table sizes from the signals through social learning.

Dynamic Chinese Restaurant Game (D-CRG) is proposed to study how a user in a dynamic scenario acquires the knowledge to make the optimal decision through social learning [3]. In a dynamic system, a rational agent should not only consider the utility she received at a moment but also the probabilistic transitions of the system state and the influence of other agents. A Multi-dimensional Markov Decision Process (M-MDP) model [4] is proposed to integrate the social learning process into the

Chih-Yu Wang is with Research Center for Information Technology Innovation, Academia Sinica, Taiwan.

Yan Chen is with School of Electronic Engineering, University of Electronic Science and Technology of China, China.

K.J. Ray Liu is with Electrical and Computer Engineering, University of Maryland, College Park, U.S.A.

This work was partly supported by the Ministry of Science and Technology under Grant 103-2218-E-001-002-MY2, 105-2221-E-001-003-MY3, the Academia Sinica under Thematic Research Grant (to Chih-Yu Wang), the National Natural Science Foundation of China under Grant 61672137, and the Thousand Youth Talents Program of China (to Yan Chen).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSIPN.2017.2682799, IEEE Transactions on Signal and Information Processing over Networks

optimal decision process from a game-theoretic perspective. The network externality may bring performance degradation when the competition between agents is fierce, but the system still can reach social welfare optimal if the optimal pricing strategy is imposed [5].

The information revealed by agents and utilized by social learning could be user-generated contents or passively-revealed information. User-generated contents, such as public reviews, comments and ratings on certain restaurants, could be easily identified and analyzed to understand the knowledge behind the information. The knowledge behind the generated content is implicit since such information can be treated as signals generated by the system and reported by the agent, conditioning on some parameters known or unknown to the agents. A simple approach is to assume that these signals are described by some probabilistic distributions conditioning on the values of some unknown parameters [2]. Agents can then construct their belief on the unknown parameters based on their prior knowledge on the characteristics of the signals. The signaling approach is useful for simple systematic parameters and has been utilized by Chinese Restaurant Game, including the review rating on Yelp to the quality of services [6] and the sensing results to the primary user activities in cognitive radio networks [2], [3].

Nevertheless, the user-generated contents could be untrustworthy when agents have selfish interests. Agents who act rationally may choose to generate forged contents for their own benefits. Such contents can mislead other agents to have false beliefs on unknown parameters and therefore they make improper decisions. For instance, a local customer may know the best restaurants in town, but he/she may never reveal the list to others in fear that the restaurants may become too popular and she need to wait for weeks to have seats reserved. Moreover, she may choose to promote other restaurants with lower quality to lower the risk. In addition, even the providers, such as restaurant owners or movie makers, have the incentive to generate forged contents to manipulate the decisions of customers when they are in competitions with others. For instance, some restaurants will invite popular bloggers or critics to provide positive reviews or rating on the website with some discounts as rewards. This biased contents will mislead customers and therefore are not trustworthy. In cognitive radio networks, for another instance, cooperative sensing utilizes the sensing results gathered from multiple secondary users to increase the accuracy of the primary user activity detection. Nevertheless, a rational secondary user may choose to share a forged sensing results to mislead other users so she can access the unoccupied channels solely. In both cases, agents who hold the knowledge can decide to release contents that contain misleading information to confuse other agents. Additionally, the cost of forging contents is very low, since such behaviors require little effort and do not make damages to the agent's own utility. It requires a significant amount of costs and efforts, such as additional punishment or reputation system, to guarantee the trustworthiness of the contents. Given that agents may generate forged contents, agents cannot collaboratively gather the true knowledge and more likely make wrong decisions. In sum, social learning will be ineffective if we rely on user-generated contents as the sole information source.

The passively-revealed actions, such as the number of customers in certain restaurants or number of secondary users accessing certain channels, also reveal useful information but require more efforts to mine the knowledge behind. The knowledge behind such information is more explicit since they are related to not only the systematic parameters, but also the actions of some or all agents in the system. For instance, a high number of visits to a certain restaurant may suggest a high-quality service, a bad service with a short-term promotion, or the shutdown of all other restaurants. One must not only observe the explicit information but also consider the reasons behind the collaborative actions.

Nevertheless, it is easier for an agent to gather passively-revealed actions compared to the user-generated contents. For instance, one can easily observe the number of customers staying in the restaurant, but may not collect their opinions on the restaurant. For a cognitive radio network, one can detect whether a channel is accessed by another agent passively, but not the sensing result another agent holds unless an information exchange protocol and a control channel are established. Due to the ease to collect the passively-revealed actions, this type of information is commonly gathered in the real word, such as Check-ins and Likes in Facebook, Daily Visit Numbers of theme parks, Page Views of websites, Number of Access to base stations/networks, etc.

Although both user-generated contents and passively-revealed actions potentially could be altered by selfish agents according to their own benefits and interests, the passively-revealed actions are more reliable and trustworthy if we can correctly understand the logic behind their actions. In addition, it usually costs more to cheat in the revealed actions since an agent must select a sub-optimal action to reveal a forged information which adversely damages her own utility. For instance, one can only mislead other agents to choose different restaurant by actually choosing another restaurant she doesn't prefer to. Such a decision incurs a loss to her own utility, and therefore reduces her incentive for such a cheating behavior.

In light of these advantages, we propose a new stochastic game-theoretic framework to utilize the passively-revealed actions instead of user-generated contents as the main information source in social learning process. The proposed **Hidden Chinese Restaurant Game** (H-CRG) allows customers to observe the actions of other customers within a limited observation space to determine their action and belief on hidden information in a stochastic system. Two types of information may be observable: history information, which is a series of actions applied by previous customers within a limited time; and grouping information, which is the number of customers at each table. The observable information can be either one or both, depending on the network we are modeling.

Based on the stochastic state transition structure and inspired by the Bayesian learning in Chinese Restaurant Game, we propose Grand Information Extraction, a novel Bayesian belief extraction process to directly extract the belief on the hidden information from the observed actions. The proposed process is universal for any Markovian system with hidden information (state). The extracted belief is conditioned on

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSIPN.2017.2682799, IEEE Transactions on Signal and Information Processing over Networks

the policy applied by customers and may also influence their actions. The coupling relation is utilized to transform the original continuous-state formulation found in traditional Partial-Observed MDP (POMDP) into a discrete-state pseudo MDP, which greatly reduces the complexity of the problem.

The optimal policy is then analyzed in both the centralized approach and the game-theoretic approach. We propose a value-iteration solution to find the centralized policy in the centralized approach. Inspired by the centralized approach and D-CRG [3], we then propose a value-iteration solution to derive the Nash equilibrium in the H-CRG. We notice that the pure-strategy Nash equilibrium may not exist in the H-CRG. Specifically, the Information Damping phenomenon may rise in a certain H-CRG in which customers synchronously switch from one policy to another due to the information loss in the observed action. The phenomenon is similar to Information Cascade [7], which is commonly discussed in traditional social learning literature. Fortunately, the existence of naive customers, such as legacy network devices in communication networks or myopic agents in social networks, can help ensure the existence of pure-strategy Nash equilibrium.

We demonstrate how the proposed H-CRG can be applied to the channel access problem in cognitive radio networks. We then conduct data-driven simulations using the CRAWDAD Dartmouth campus WLAN trace [8], [9]. The simulation results show that the equilibrium strategy derived in H-CRG provides higher expected utilities for new customers and maintains a reasonable high social welfare compared with other candidate strategies.

The main contributions of this paper are as follows:

1) Customers are not required to reveal their signals to other customers after their decisions. Only their actions are passively observed. The overhead in signaling exchanges therefore can be greatly reduced. In addition, the potential threat of untruthful signal reports in previous works is eliminated entirely.
2) The belief is directly extracted from the observed actions using the proposed Grand Information Extraction process instead of storing in a separate belief state. Specifically, we observe that the exact belief on the action and the action applied by rational agents in fact form a coupling relation. This motivates us to propose Grand Information Extraction to utilize the stationary probability distribution of states to calculate the probability that certain state is reached. Combining the Grand Information Extraction with the action-based M-MDP, we perform the belief update based on the policy applied by other agents and no longer need to record the belief state in the system state. The continuous belief space issue in either D-CRG and POMDP can thus be avoided. To the best of our knowledge, we are the first group to point out this relation and utilize this to design the new M-MDP model to inherently capture the belief update process without the need of separate belief state or a fixed belief update equation.
3) We provide both centralized solution and game-theoretic solutions. We analyze the H-CRG from two perspectives: centralized approach which maximizes the social welfare, and game-theoretic approach which maximizes each individual's utility. We illustrate the relation of both approaches under this framework and propose corresponding algorithms to derive the solutions.

II. Related Works

Social learning is one of the main research focuses in both economy and network science. Users in a social network may not have clear understanding on the current state of the network. They therefore may actively share their private information or passively observe the actions applied by others to improve their knowledge on the state. Most existing literatures studied how agents reach a correct consensus through social learning [7], [10]–[12]. Their studies limit to the scenario that no network externality exists, i.e., the choice of one agent will not influence the payoff received by other agents. This assumption helps them to focus on the belief formation but limits the applications. Several attempts have been made to extend the traditional social learning framework to include network externality [13], [14]. The applications of these models are still limited due to the assumptions they rely on, such as binary state space, no decision order information, and positive network externality only. Some studies on social learning in stochastic system can be found in [15], [16]. They focus on the equilibrium learning strategy concerning the stochastic characteristic of hidden state. Nevertheless, none of these consider the network externality and dynamic in agent population. To the best of our knowledge, no analysis on action-based social learning in stochastic networks with network externality has been conducted.

The proposed framework is similar with the Partially Observable Markov Decision Process (POMDP) model, which is a generalization of Markov Decision Process with hidden states. In POMDP model, a centralized user who decides the action cannot directly observe the true state of the system. Some observations, which is related to the true state, can be derived by the agent as a hint of the true underlying state. The uncertainty and the knowledge of the true state are captured by the belief, the probability distribution of the true state. The goal of the agent is to find the optimal policy to control the system in order to maximize the long-term reward. It has been shown that POMDP can be formulated as a belief MDP in which the belief is captured by additional continuous belief states. The optimal policy then can be derived using point-based methods [17], [18]. The point-based methods are based on the observations that the optimal expected value function can be formulated as a combination of the value function on a proper set of belief vectors in which the optimal action in each segment can be derived. The optimal policy can be derived then. It is non-trivial to find the feasible belief vectors for exact optimal policy. Approximated algorithms to find suboptimal but tractable belief vectors are therefore proposed [18]. Another approach is to control the POMDP system through finite-state controllers [19]. A feasible control policy can be formulated as a policy graph to describe the action to apply when certain observations is received, without the need of additional belief state [20]. The solution space can then be reduced to the policy space described by the graph. However, the size of policy graph may be intractable if we would like to guarantee the optimality of the solution. Nevertheless, the POMDP model is different from the proposed H-CRG in two aspects: the number of objective functions, and the complexity of belief update process. In traditional MDP or POMDP problems, we only have one expected value function to serve as the only objective function to maximize. In the proposed H-CRG framework, on the other hand, we are dealing with multiple objective functions, or the utilities of agents entering the system at different states and choosing different tables. As we have illustrated in previous works [4], [5], this major difference imposes a serious challenge to solve the Nash equilibrium in the proposed game since these objective functions will affect each other in a non-linear fashion. In addition, the belief update process, as we will illustrated in Section IV-B, is no longer Markovian since the updated belief depends on not only the current state, prior belief, current action, but also the action of other agents in previous and future time slots. Therefore, all existing algorithms, including point-based method and policy graph, cannot be directly applied to derive the Nash equilibrium. This motivates us to seek an alternative approach to handle this problem.

This work is different from our previous Chinese Restaurant game [1]–[3] in the following two aspects. First, customers are not required to reveal their signals to other customers after their decisions. Only their actions are passively observed. The overhead in signaling exchanges therefore can be greatly reduced. In addition, the potential threat of untruthful signal reports in previous works is eliminated entirely. Notice that this is not a trivial extension since the information contained in the action is more difficult to be extracted. We need to know the subjective intention of the agents who perform the actions, which may be influenced by the actions of other agents due to the externality in the network.

Second, the belief is directly extracted from the observed actions using the proposed Grand Information Extraction process instead of storing in a separate belief state. This belief extraction process is universal for any Markovian system with hidden information. Specifically, we observe that the exact belief on the action and the action applied by rational agents in fact form a coupling relation. This relation motivates us to propose Grand Information Extraction to utilize the stationary probability distribution of states when the policy is given to calculate the probability that certain state is reached. Then, the conditional probability of the hidden state given certain state is observed can be derived through Bayesian equations. This extraction process helps link the belief directly to the policy. Combining the Grand Information Extraction with the action-based M-MDP, we perform the Bayesian belief update based on the policy applied by other agents and no longer need to record the belief state in the system state. We then can find the Nash equilibrium in pure policy space. The continuous belief space issue in both D-CRG and POMDP can thus be avoided. To the best of our knowledge, we are the first group to point out this relation, connect all the parts in the process, and utilize this to design the new M-MDP model to inherently capture the belief update process without the need of separate belief state or a fixed belief update equation.

Finally, cooperative sensing in cognitive radio networks is an important applications of proposed framework. In cognitive radio network, secondary users are required to access the channels only if they are not accessed by primary users. Nevertheless, there is no direct communication between primary and secondary users. Secondary users should detect the activities of primary users through channel sensing in order to avoid interference. Cooperative sensing is proposed to improve the sensing accuracy by allowing secondary users share their sensing results to make decisions collaboratively [21]. Various signaling exchange schemes have been proposed including centralized or distributed mode and soft or hard collaborative decision [22]. In general, higher accuracy, such as soft decision mode in centralized sharing scheme, comes with longer latency and larger overheads due to more signal exchanges. The cheating behavior in cooperative sensing, or Byzantine attacks, gains more attentions in recent years [23]. Secondary users may intentionally report false sensing results to others in order to gain advantages in channel access. Various defense mechanisms have been proposed based on the statistic difference in false reports from other normal reports. Some additional penalties may be applied in utility-based approach to persuade rational users from launching attacks. Most proposed designs introduce further overheads in the cooperative sensing system either in implementation cost or accuracy loss. The key overheads in cooperative sensing comes from the cost of sharing sensing results among members through signal exchanges. In this paper, we propose to use H-CRG to completely eliminate the need of signal exchange process while maintaining the sensing accuracy. To best of our knowledge, we are the first one to propose an exchange-free protocol for cooperative sensing.

In the rest part of the paper, we first introduce the system model in Section III and the game structure of H-CRG in IV. Then we introduce the novel Bayesian belief extraction process called Grand Information Extraction to extract the knowledge from the observed information in Section IV-B. Based on the belief, we explain how rational agents can make use of the knowledge to make decisions and how the equilibrium is defined in Section IV-C. Two solutions based on centralized and game-theoretic approaches are provided in Section V. We then demonstrate how the proposed H-CRG can be applied to real world problems through applying H-CRG in channel access in cognitive radio networks in Section VI. Finally, we draw the conclusions in Section VII.

III. System Models

Let us consider a Chinese restaurant with M tables. We assume that the restaurant allows at most N customers to enter, where N serves as the capacity constraint of the restaurant. We consider a time-slotted system where customers may arrive at and leave the system following a Bernoulli process. That is, customers arrive at the restaurant with probability λ and leave the restaurant with probability μ. We assume that the time slot is very short so there is at most one customer arrives or departs

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSIPN.2017.2682799, IEEE Transactions on Signal and Information Processing over Networks

within a slot [1]. At each arrival, the customer requests for a seat in the restaurant by choosing one table. She may or may not know the number of customers at each table. As long as she chooses a table, she will stay seated at the chosen table until departure. A customer may not choose to enter the restaurant when 1) the restaurant is full and therefore the door is closed[2], or 2) the maximum expected utility if she enters the restaurant is negative. The latter case means that a customer may find that leaving the process without entering the restaurant could be a more valuable choice. For instance, the restaurant may offer poor-quality meals that the customer decides not to give it a try even if there are seats available. Let $x[t] \in \{0, 1, ..., M\}$ be the decision of the customer arrives at time t, where $x[t] = 0$ means that the customer chooses not to enter the restaurant or there is no customer arrives at time t. Notice that this also reflects the fact that these two events make no differences to agents who can only observe the revealed actions.

The size of a table has a positive influence on the dining experience of the customers, where we assume that a larger table is welcomed by any customer as long as the number of customers choosing the table remains the same. Some tables could be smaller or even unavailable when reserved by high-priority customers. The exact sizes of the tables are controlled by the restaurant state $\theta \in \Theta = \{1, 2, 3, ...\}$, where the size of each table x is given by $R_x(\theta)$. Nevertheless, the sizes of the tables are unknown to the customers before she actually enters the restaurant. That is, the restaurant state θ is given at the beginning of the game in advance, following a prior distribution $Pr(\theta = k)$, but unknown to the customers. Given that θ has a definite influence on the dining experience of customers, it represents the critical knowledge the customers need to acquire in the game.

A. Customers: Naive and Rational

We consider two kind of customers: naive customers and rational customers. Naive customers represent the legacy agents or devices whose the actions are predetermined and fixed without the strategic decision making capability. In cognitive radio networks, for instance, there may exist some legacy secondary devices who only have limited sensing capability without collaboration with other devices. Their channel access actions are more predictable. These naive customers may have either positive or negative impacts on the overall system performance and service quality experienced by other customers.

Rational customers, on the other hand, select the tables strategically. Their sole purpose is to maximize their expected utility. The utility of the customers is determined by two factors: the number of customers seated in the same table and the size of the table. Specifically, let $n_x[t]$ be the number of customers choosing table x at time t. The $\mathbf{n}[t] = (n_1[t], n_2[t], ..., n_M[t])$ denotes the grouping of customers at time t, i.e., the number of customers choosing each table at time t. Then, the immediate utility of a customer choosing table x at time t is $u(R_x(\theta), n_x[t])$, where $R_x(\theta)$ is the size of table x. The influence of the number of customers, that is, network externality, could be arbitrary and is captured by $\frac{\partial u(r,n)}{\partial n}$. The externality is positive when $\frac{\partial u(r,n)}{\partial n} > 0$, negative when $\frac{\partial u(r,n)}{\partial n} < 0$, and is zero when $\frac{\partial u(r,n)}{\partial n} = 0$ for all n. Finally, we assume that a larger table is welcomed by any customer, that is, $\frac{\partial u(r,n)}{\partial r} \leq 0$.

B. Observable Information

The knowledge of the unknown restaurant state can be extracted from the information collected by the customer. The basic information each customer acquires is a private signal she receives at the time of arrival. We assume that the signal $s \in \mathcal{S}$ is generated following a probability density function $f(s|\theta)$. The signal is informative, that is, there exists a non-zero correlation between restaurant state θ and signal s. Also, the signal is private, which means that other customers will not know the signal one received unless she explicitly reveals. Finally, the generated signals are independent when conditioning on the true restaurant state θ.

Despite the private signal, the information one customer may collect from others are the observed passively-revealed actions. Such information comes in two different types, **grouping information** and **history information**, as follows:

Grouping Information: The current grouping $\mathbf{n}[t]$ of customers at time t. It represents the number of customers choosing each table at the current time. This observation roughly captures the consensus among all customers in the system. Nevertheless, the decision orders of these customers are not captured in grouping information. In addition, this kind of observation could be very handy in some systems but too costly to maintain or derive. For instance, one customer may easily see the number of customers waiting to be served at each restaurant by simple counting, but she will not know the number of customers subscribing to each cellular service unless the providers explicitly announce.

History Information: The history of actions revealed by customers selecting tables at time $t - H$,$t - H + 1$,...,$t - 1$. Here we assume that a customer may observe and record the action revealed by the previous customers up to H slots before she makes decisions. The history information emphasizes the decision orders of customers and potential influences of the former actions to the later customers. History information is easier to be accessed in some networks and therefore commonly seen in the literature. Notice that when H goes to infinity, history information will become grouping information **plus** the decision order information. Nevertheless, H is usually assumed to be finite to reflect the limited observation capability one customer may have.

We assume that the only information revealed by one customer is the table she selected. That is, customers will not contribute user-generated contents such as signals, but

[1] Notice that this assumption can be easily relaxed in our model by expanding the state transition probability matrix to include multiple arrival and departure cases. We believe all methods and conclusions we made in this paper still hold with this assumption relaxed.

[2] It is also possible to offer an option for the customers to wait at outside of the restaurant until it opens again. Specifically, we may let P be the maximum number of customers waiting outside of the restaurant when the restaurant is full, and each customer enters the restaurant sequentially when some customers leave. This can be modeled as a FIFO queue and can be integrated into the framework easily by properly adjusting the state transition probabilities.

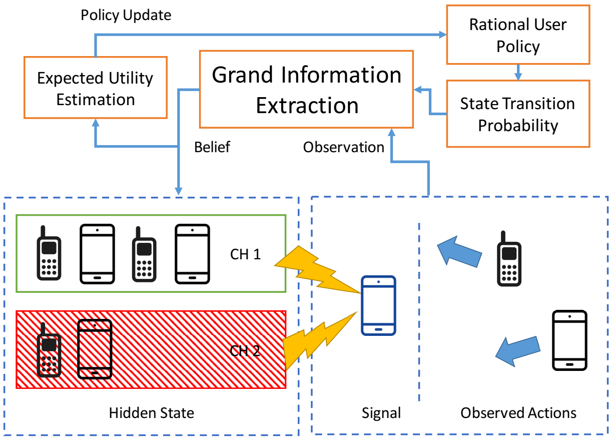


Fig. 1. Hidden Chinese Restaurant Game Framework

only reveal their actions passively. This design eliminates the scenario that a customer can untruthfully report the private signals. Formally speaking, the potential information space for a customer choosing a table at time t is as follows:

$$(\mathbf{n}[t], x[t-H], x[t-H+1], ..., x[t-1], s[t]) = (\mathbf{n}[t], \mathbf{h}[t], s[t]), \quad (1)$$

where $\mathbf{h}[t]$ denotes the history of information (actions) revealed by customers from time $t-H$ to $t-1$.

The rational customers, who are of our main interests, aim to maximize their long-term expected utility in the system, that is, to choose the table that maximizes their expected utility considering the unknown restaurant state and the potential network externality. The main challenges faced by the rational customers are two folds: 1) how to extract the information from the observed actions in order to estimate the unknown information such as restaurant state? 2) how to predict the influence of network externality in the stochastic system given there exist some naive customers?

IV. Hidden Chinese Restaurant Game

We formulate the table selection problem as a hidden Chinese Restaurant Game (H-CRG). The H-CRG is a stochastic game with an indeterministic number of players. The arrival and departure of the players (customers) are assumed to satisfy the Bernoulli distributions. A customer can choose the table to be seated or leave immediately, but she cannot change her table afterward. Her utility depends on the table she selected and the states of the system in the duration of her stay.

An illustration of H-CRG framework is shown in Fig. 1. Briefly speaking, we propose Grand Information Extraction process in the H-CRG to extract the belief on the hidden state from the observed information. The process requires inputs from the policy applied by rational customers and corresponding state transition probability. Rational customers then estimate the expected utilities with the belief extracted by the process. The updated estimation on the expected utilities will help them to make better decisions and update the rational policy accordingly. The key in H-CRG is to find the policy which maximizes the expected utility for each customer.

The system state describes the current situation of the system, including the restaurant state, grouping information, history of actions, and the generated signal. Given the current state, one may determine the current utility of a customer completely. Nevertheless, the state will transit stochastically following a state transition probability function, which we will describe later. Formally speaking, the system state of H-CRG at time t is denoted as follows:

$$\mathbf{I}[t] = \{\mathbf{n}[t], \mathbf{h}[t], s[t], \theta\}. \quad (2)$$

The information in the state $\mathbf{I}$ is differentiated into two types: observed state $\mathbf{I}^o$ and hidden state $\mathbf{I}^h$. The observed state is the information that can be readily observed by the customers when they arrive. The hidden state represents the information which can only be derived through belief estimations using observed states as inputs. Notice that both observed state and hidden state may have influences on the utility of each customer.

Whether the information is observable or hidden depends on the system we are formulating with. We assume that the restaurant state θ is always hidden, since we assume that the value is given in prior but unknown to the customers. On the other hand, the signal $s[t]$ is always observable since we assume that each customer receives at least a private signal.

For the other two pieces of information, grouping information and history information, if observable or not depends on the system settings. For instance, if we are modeling a restaurant selection problem in a food court, it is safe to assume that each customer can observe both grouping information and history information since one can easily see the number of customers waiting in lines and how these lines are formed. In this case, both grouping and history pieces of information are observable and therefore belong to the observed state. On the other hand, if we are modeling a channel selection game in a cognitive radio network, it may be impractical to assume that each secondary user can observe the grouping information since it requires a third-party base station to record and broadcast such information. Nevertheless, it could be safe to assume that each secondary user can observe the choices of other users in channel access when they arrive since it can be done by channel monitoring. In such a case, the grouping information should belong to the hidden state, while the history information belongs to the observed state.

A policy describes the table selection strategy a customer applies in H-CRG given the information she observed as inputs. As we mentioned, a customer can only observe the information in the observed state. Therefore, a policy can be defined as follows:

$$\pi(\mathbf{I}^o) \in \mathbf{A} = \{0, 1, ..., M\}, \forall \mathbf{I}^o. \quad (3)$$

Notice that $\pi(\mathbf{I}^o) = 0$ means that the customer chooses not to select any of the tables and leaves the restaurant immediately when she observes $\mathbf{I}^o$.

Recall that we have two kinds of customers, naive and rational customers. We assume that a ratio ρ of customers are rational, while others are naive customers. Naive customers follow a naive policy to determine their actions. The naive

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSIPN.2017.2682799, IEEE Transactions on Signal and Information Processing over Networks

TABLE I
NOTATIONS

Notation	Explanation
N,M	the total customer capacity and number of tables of the restaurant
$s \in S$, $\theta \in \Theta$, $f(s\|\theta)$	the signal, restaurant state, and p.d.f. of the signal
$R_x(\theta)$, $\mathbf{n}[t] = \{n_1[t], n_2[t], ..., n_M[t]\}$	the table size function and the grouping information at slot t
$\mathbf{h}[t] = \{h_1[t], h_2[t], ..., h_H[t]\}, H$	the history information and the length of history
$u(R, n)$	the immediate utility of a customer
λ, μ	the arrival and departure probability of a customer
ρ	the ratio of naive customers
$\mathbf{I}[t] = \{\mathbf{I}^o[t], \mathbf{I}^h[t]\}$	system state, observed state, and hidden state
$\pi^n(\mathbf{I}^o)$, $\pi^r(\mathbf{I}^o)$	policy followed by naive and rational customers
$\mathcal{I}^o_{\mathbf{I}^o,\pi^n,\pi^r}$	the set of system states sharing the same observed state $\mathbf{I}^o$
$W^I(\mathbf{I})$, $W(\mathbf{I}^o)$	the expected reward conditioning on the system state and observed state
$g_{\mathbf{I}\|\mathbf{I}^o,\pi^n,\pi^r}$	the belief of state $\mathbf{I}$ conditioning on observed state $\mathbf{I}^o$

policy is fixed across the whole game. We denote the naive policy as $\pi^n(\mathbf{I}^o)$.

On the other hand, the goal of a rational customer is to maximize her long-term expected utility. When a customer arrives at the system, she observes the system and receives the observed state. Then, she chooses the table providing largest long-term expected utility by considering both the network externality $\mathbf{n}$ and unknown restaurant state θ. Notice that the influence of network externality could change over time due to departure and arrival of other customers. A rational customer should not only consider the currently observed state but also predict the decisions of other customers (both rational and naive ones) in the future.

Recalling that the immediate utility of a customer choosing table x in a given time slot t is $u(R_x(\theta), n_x[t])$. The long-term expected utility of a customer arriving at time t_a is given by

$$E[U(x)|\mathbf{I}^o[t_a]] = \tag{4}$$
$$\sum_{t=t_a}^{\infty}(1-\mu)^{(t-t_a)} \sum_{\theta\in\Theta} Pr(\theta|\mathbf{I}^o[t_a]) E[u(R_x(\theta), n_x[t])|\mathbf{I}^o[t_a], \theta].$$

Notice that here a customer observes $\mathbf{I}^o$ without the knowledge of the actual state $\mathbf{I}$. It requires further efforts to estimate the corresponding hidden state, which we will introduce later.

A rational customer should maximize her expected utility, that is, choose the table as follows:

$$\pi^r(\mathbf{I}^o) = \arg \max_{x\in\{0,1,...,M\}} E[U(x)|\mathbf{I}^o], \forall \mathbf{I}^o. \tag{5}$$

The keys to estimate the expected utility are 1) extracting the hidden state from the observed state, and 2) predicting the future states conditioning on the currently observed state.

A. System State Transition

The system state transits with time. At each time slot, a new signal will be generated conditioning on the restaurant state θ. In addition, the observed actions in the history information $\mathbf{h}[t]$ will be shifted with the action observed at time $t - H$ forgotten and new action observed at time t added.

The grouping information $\mathbf{n}[t]$, which is the key to estimate the influences of network externality, changes when any of the following events occurs:

1) New Customer Arrival: When a new customer arrives, she will select the table according to either naive policy $\pi^n(\mathbf{I}^o)$ or rational policy $\pi^r(\mathbf{I}^o)$, depending on her type. It is also possible that the customer chooses not to enter the restaurant, that is, $\pi^r(\mathbf{I}^o) = 0$ or $\pi^n(\mathbf{I}^o)$, when a certain observed state $\mathbf{I}^o$ occurs. Additionally, the customer may be rejected by the restaurant when the loading reaches the maximum capacity N regardless of the table she selects. We denotes $\mathcal{I}^{e,n}$ and $\mathcal{I}^{e,r}$ as the set of system states where naive and rational customers will not enter the restaurant, respectively. Formally speaking,

$$\mathbf{I} = \{\mathbf{I}^o, \mathbf{I}^h\} \in \mathcal{I}^{e,\{n,r\}} \text{ if } \pi^{\{n,r\}}(\mathbf{I}^o) = 0 \text{ or } \sum_{j=1}^{M} n_j = N. \tag{6}$$

For any state $\mathbf{I} \notin \mathcal{I}^{e,\{n,r\}}$, there exists a set of state $\mathcal{I}^a_{\mathbf{I},\pi^{\{n,r\}}}$ where every state $\mathbf{I}' = \{\mathbf{n}', \mathbf{h}', s', \theta\}$ in the set satisfies

$$n'_{\pi^{\{n,r\}}(\mathbf{I}^o)} = n_{\pi^{\{n,r\}}(\mathbf{I}^o)} + 1, n'_j = n_j \forall j \neq \pi(\mathbf{I}^o),$$
$$\mathbf{h}' = \{h_2, ..., h_{H-1}, \pi^{\{n,r\}}(\mathbf{I}^o)\}. \tag{7}$$

In other words, when a new customer arrives and chooses a table according to the policy, the number of customers in the corresponding table plus one, and the history information $\mathbf{h}$ records this choice. Notice that there may be more than one possible transition state in the set. For any other state $\mathbf{I} \in \mathcal{I}^{e,\{n,r\}}$, the corresponding state set $\mathcal{I}^a_{\mathbf{I},\pi^{\{n,r\}}}$ is empty.

2) Existing Customer Departure: When an existing customer in the restaurant leaves, the number of customers at the table decreases by one. Since no new customer enters the restaurant at this moment, the history information records zero (no observed action) at this moment. Let $\mathcal{I}^d_{\mathbf{I}}$ be the set of transition states from state $\mathbf{I}$ that one customer departs from the restaurant. For every state $\mathbf{I}' \in \mathcal{I}^d_{\mathbf{I}}$, we have

$$\exists d, n'_d = n_d - 1, n'_j = n_j \forall j \neq d, \mathbf{h}' = \{h_2, ..., h_{H-1}, 0\}. \tag{8}$$

3) No Change: When both previous events are not occurred, the grouping information remains unaltered. In such a case, only history information and signal changes at next time slot. We have $\mathcal{I}^u_{\mathbf{I}}$ be the set of transition states from state $\mathbf{I}$ that no customer arrives at or departs from the restaurant. For every state $\mathbf{I}' \in \mathcal{I}^u_{\mathbf{I}}$, we have

$$\mathbf{n}' = \mathbf{n}, \mathbf{h}' = \{h_2, ..., h_{H-1}, 0\}. \tag{9}$$

Given all the probability distributions we defined in Section III and the discussions above, we can derive the state transition probability in (10).

B. Grand Information Extraction

In order to estimate the expected utility provided by certain tables, it is necessary to estimate the hidden state, which is unobservable by the customer, conditioning on the observed state. Specifically, the **belief** on the hidden state, i.e, the probability distribution of the hidden state, should be derived. This problem is similar to the belief update in Partial Observed Markov Decision Process (POMDP) except that the belief in POMDP is usually assumed to be an input of the policy while the belief update process is known and given. Traditionally, the optimal policy is derived through transforming the POMDP into a belief MDP with a continuous state space in belief state. The optimal policy then can be derived using value-iteration or policy-iteration algorithms on a finite set of value function where the expected value are formulated as a linear combination of the value function of a proper set of belief vectors. Nevertheless, the main disadvantage is that the exponential increase in the size of belief vector set makes this approach computational intractability. Approximated algorithms are more preferred for practicability.

We proposed a novel belief estimation method, Grand Information Extraction, to extract the distribution of hidden state directly from the observed state without the needs of belief update process. The basic idea is extending the Bayesian belief method in CRG from the signal domain to system state domain. We utilize the stationary probability distribution of the system states, which can be derived from the state transition probability, to directly derive the belief on the restaurant state θ. Conditioning on θ, we then can estimate the belief on the hidden state accordingly. The main advantage of this process is the possibility to formulate the system purely with discrete state spaces without the needs of belief update process and belief vector set. The problem can then be reduced to discrete-space pseudo MDP problem.

We first formally defined the belief in H-CRG as follows:

$$g_{\mathbf{I}|\mathbf{I}^o} = Pr(\mathbf{I}|\mathbf{I}^o). \tag{11}$$

When the restaurant state $\theta = k$ and policy applied by the customers are given, the state transition probability can be derived directly through (10). Then, the stationary state distribution of H-CRG $\left[Pr(\mathbf{I}|\theta = k, \pi^n, \pi^r)\right]$ is given by:

$$\begin{aligned}&\left[Pr(\mathbf{I}|\theta = k, \pi^n, \pi^r)\right] \\ &= \left[Pr(\mathbf{I}'|\mathbf{I}, \theta = k, \pi^n, \pi^r)\right]\left[Pr(\mathbf{I}|\theta = k, \pi^n, \pi^r)\right].\end{aligned} \tag{12}$$

The stationary state distribution of system states $\mathbf{I}$ can be derived through finding the normalized eigenvector of the transition matrix with eigenvalue as 1.

Lemma 1. *The stationary state distribution* $\left[Pr(\mathbf{I}|\theta = k, \pi^n, \pi^r)\right]$ *is unique.*

Proof. It has been known that the sufficient condition to have a unique stationary state distribution in a Markov system is to have exactly one closed communication class in the state transition. It can be easily seen from (10) that all states have the positive probability to transit to the zero states with no customer in the restaurants, no actions observed in the history, and arbitrary signals when all customers depart from the restaurant. This means all states will be linked to the zero state and it is impossible to have two closed communication class in H-CRG. Therefore, the stationary state distribution, conditioning on the restaurant state $\theta = k$, is unique. □

The uniqueness of the stationary state distribution guarantees that all rational customers will reach a consensus on the belief as long as they have the same observations and knowledge on the state transition.

The stationary state probability $Pr(\mathbf{I}|\theta = k, \pi^n, \pi^r)$ represents the probability that a customer will encounter a certain state $\mathbf{I}$ when he arrives the system at any time, if the restaurant state θ is actually k. The Bayesian belief rule then can be applied to derive the probability of the restaurant θ as follows. Specifically, when the stationary probability distribution $Pr(\mathbf{I}|\theta, \pi^n, \pi^r)$ is derived for all $\theta \in \Theta$, we then can derive the probability of the restaurant state θ as follows:

$$Pr(\theta = k|\mathbf{I}, \pi^n, \pi^r) = \frac{Pr(\mathbf{I}|\theta = k, \pi^n, \pi^r)}{\sum_{k' \in \Theta} Pr(\mathbf{I}|k', \pi^n, \pi^r)}. \tag{13}$$

Nevertheless, the above probability is conditioning on the system state $\mathbf{I}$, while the customer in fact can only observe the observed state $\mathbf{I}^o$. It requires further efforts to derive the actual belief of the customer on the hidden state. Let $\mathcal{I}^o_{\mathbf{I}^o}$ be the set containing all the states sharing the same observed state $\mathbf{I}^o$. The probability that one may observe certain observed state $\mathbf{I}^o$ conditioning on $\theta = k$ is the sum of the stationary state probability of all states in $\mathcal{I}^o_{\mathbf{I}^o}$, which is as follows:

$$Pr(\mathbf{I}^o|\theta = k, \pi^n, \pi^r) = \sum_{\mathbf{I} \in \mathcal{I}^o_{\mathbf{I}^o}} Pr(\mathbf{I}|\theta = k, \pi^n, \pi^r). \tag{14}$$

Then again, we can estimate the probability of the restaurant state θ conditioning on the observed state $\mathbf{I}^o$ following the Bayesian rule:

$$\begin{aligned}&Pr(\theta = k|\mathbf{I}^o, \pi^n, \pi^r) = \\ &\frac{Pr(\mathbf{I}^o|\theta = k, \pi^n, \pi^r)Pr(\theta = k)}{\sum_{k' \in \Theta} Pr(\mathbf{I}^o|\theta = k', \pi^n, \pi^r)Pr(\theta = k')}.\end{aligned} \tag{15}$$

The above belief is sufficient for the case that only the restaurant state θ is unobservable. Nevertheless, it is also possible that the grouping information or history information is not observable as we discussed in Section III. In such a case, we still need to estimate the hidden information in the hidden state. Recall the stationary state distribution $\left[Pr(\mathbf{I}|\theta = k, \pi^n, \pi^r)\right]$ we already derived. Conditioning on a given $\theta = k$, we can derive the probability of the actual system state $\mathbf{I}$ conditioning on the observed state $\mathbf{I}^o$ as follows:

$$Pr(\mathbf{I}|\mathbf{I}^o, \theta = k, \pi^n, \pi^r) = \frac{Pr(\mathbf{I}|\theta = k, \pi^n, \pi^r)}{\sum_{\mathbf{I}' \in \mathcal{I}^o_{\mathbf{I}^o}} Pr(\mathbf{I}'|\theta = k, \pi^n, \pi^r)}. \tag{16}$$

$$Pr(\mathbf{I}[t+1]|\mathbf{I}[t],\pi^n,\pi^r) = \tag{10}$$
$$\begin{cases} \rho\lambda f(s[t+1]|\theta), & \mathbf{I}[t+1]\in\mathcal{I}^a_{\mathbf{I}[t],\pi^r};\\ (1-\rho)\lambda f(s[t+1]|\theta), & \mathbf{I}[t+1]\in\mathcal{I}^a_{\mathbf{I}[t],\pi^n};\\ (n_j[t])\mu f(s[t+1]|\theta), & \mathbf{I}[t+1]\in\mathcal{I}^d_{\mathbf{I}[t]}, n_j[t+1]=n_j[t]-1;\\ (1-\mu\sum_{j=1}^M n_j-\lambda)f(s[t+1]|\theta), & \mathbf{I}[t+1]\in\mathcal{I}^u_{\mathbf{I}[t]}, \mathcal{I}^a_{\mathbf{I}[t],\pi^r}\neq\emptyset, \mathcal{I}^a_{\mathbf{I}[t],\pi^n}\neq\emptyset;\\ (1-\mu\sum_{j=1}^M n_j-\rho\lambda)f(s[t+1]|\theta), & \mathbf{I}[t+1]\in\mathcal{I}^u_{\mathbf{I}[t]}, \mathcal{I}^a_{\mathbf{I}[t],\pi^r}\neq\emptyset, \mathcal{I}^a_{\mathbf{I}[t],\pi^n}=\emptyset;\\ |-|(1-\mu\sum_{j=1}^M n_j-(1-\rho)\lambda)f(s[t+1]|\theta), & \mathbf{I}[t+1]\in\mathcal{I}^u_{\mathbf{I}[t]}, \mathcal{I}^a_{\mathbf{I}[t],\pi^r}=\emptyset, \mathcal{I}^a_{\mathbf{I}[t],\pi^n}\neq\emptyset;\\ (1-\mu\sum_{j=1}^M n_j)f(s[t+1]|\theta), & \mathbf{I}[t+1]\in\mathcal{I}^u_{\mathbf{I}[t]}, \mathcal{I}^a_{\mathbf{I}[t],\pi^r}=\mathcal{I}^a_{\mathbf{I}[t],\pi^n}=\emptyset;\\ 0, & \text{else.} \end{cases}$$

Combining (15) with (16), we can derive the belief on the system state $\mathbf{I}$ conditioning on the observed state $\mathbf{I}^o$:

$$\begin{aligned} &g_{\mathbf{I}|\mathbf{I}^o,\pi^n,\pi^r} = Pr(\mathbf{I}|\mathbf{I}^o) \\ &= \sum_{k\in\Theta} Pr(\mathbf{I}|\mathbf{I}^o,\theta=k,\pi^n,\pi^r)Pr(\theta=k|\mathbf{I}^o,\pi^n,\pi^r). \end{aligned} \tag{17}$$

It should be noted that the belief from the proposed Grand Information Extraction process on the state $\mathbf{I}$ is conditioned on not only the observed state $\mathbf{I}^o$ but also the policies applied by rational and naive customers. The accuracy of the estimated belief depends on how informative the observed actions are, which is determined by the policies.

C. Equilibrium Conditions

In H-CRG, rational customers will seek to maximize their long-term expected utility. Nevertheless, the expected utility is determined by not only the current state of the system but also the transition of the state in the future.

The state transition is determined by the choice of customers at different states. The grouping $\mathbf{n}$, for instance, is determined by the policy applied by not only the naive customers but also other rational customers in the future. Additionally, the belief of a customer on the hidden state relies on the observed state, which contains the actions of other previous customers. As long as the actions applied by other customers change, the state transition changes, and so is the expected utility experienced by the customer. In sum, the belief on the hidden state and the state transition depends on the choices of all customers, while each customer's choice depends on their belief on the states. The complex interactions between customers is therefore captured by the proposed H-CRG.

We now analyze the pure-strategy Nash equilibrium of H-CRG. Nash equilibrium is a popular solution concept in game theory which is used to predict the outcome of a game. Let $E[U(x_{\mathbf{I}^o},\mathbf{x}_{-\mathbf{I}^o})]$ be the expected utility of a customer observing $\mathbf{I}^o$, where $x_{\mathbf{I}^o}$ is her choice and $\mathbf{x}_{-\mathbf{I}^o}$ are the choices of other customers at other states. The pure-strategy Nash equilibrium of H-CRG is defined as follows:

Definition 1 (Nash Equilibrium)**.** *The Nash equilibrium, or pure-strategy Nash equilibrium, in the proposed H-CRG is a policy π^* where for all $\mathbf{I}^o$,*

$$E[U(x^*_{\mathbf{I}^o},\mathbf{x}^*_{-\mathbf{I}^o})] \geq E[U(x,\mathbf{x}^*_{-\mathbf{I}^o})], \forall x\in\{0,1,2,...,M\}, \tag{18}$$

where $x^*_{\mathbf{I}^o}=\pi^*(\mathbf{I}^o)$, $\mathbf{x}^*_{-\mathbf{I}^o}=\{\pi^*(\mathbf{I}^{'o})|\mathbf{I}^{'o}\neq\mathbf{I}^o\}$.

The expected utility in (18) can be analyzed by modeling H-CRG as a Multi-Dimensional Markov Decision Process (M-MDP) [4]. Let the system state $\mathbf{I}$ be the state and the $\pi^r(\mathbf{I}^o)$ be the policy in M-MDP, we define the immediate reward as

$$R(\mathbf{I},x) = R(\mathbf{n},\mathbf{h},s,\theta,x) = u(R_x(\theta),n_x). \tag{19}$$

The expected reward of a customer choosing table x at state $\mathbf{I}$ in the system can be denoted as $W^I(\mathbf{I},x)$ and derived through Bellman equation. When the game reaches stationary states, the expected reward for a customer to stay at a table x is equal to the immediate reward she receive at the current state plus the expected reward she will receive in the future if she keeps staying in the restaurant. Therefore we have:

$$\begin{aligned} &W^I(\mathbf{I},x,\pi^r) = R(\mathbf{I},x) \\ &+(1-\mu)\sum_{\mathbf{I}'} Pr(\mathbf{I}'|\mathbf{I},\pi^n,\pi^r,x)W^I(\mathbf{I}',x,\pi^r), \forall\mathbf{I},x. \end{aligned} \tag{20}$$

Nevertheless, the state transition probability $Pr(\mathbf{I}'|\mathbf{I},\pi^n,\pi^r,x)$ we denoted here is different from (10) since it is conditioned on the fact that this customer does not depart at next time slot. Specifically, when $x>0$, the number of customers who may depart from table x will be $n_x[t]-1$ instead of $n_x[t]$. The transition probability therefore is given by (21).

The Bellman equations in (20) describe the inherent expected rewards if one has the full knowledge of the system state. Nevertheless, the inherent expected reward is unknown to the customers since they only have the knowledge of observed state $\mathbf{I}^o$. It requires further efforts to estimate the expected utility conditioning on the observed state $\mathbf{I}^o$. The idea is to utilize the belief we extracted through Grand Information Extraction to estimate the expected immediate reward and corresponding state transition probability. Let $W(\mathbf{I}^o,x)$ be the expected utility of a customer at table x if she observes $\mathbf{I}^o$. Recalling that $\mathcal{I}^o_{\mathbf{I}^o}$ is the set of states sharing the same observed state $\mathbf{I}^o$ and $g_{\mathbf{I}|\mathbf{I}^o}$ is the distribution of the states in the set conditioning on the observed state $\mathbf{I}^o$, we have

$$W(\mathbf{I}^o,x) = \sum_{\mathbf{I}\in\mathcal{I}^o_{\mathbf{I}^o}} g_{\mathbf{I}|\mathbf{I}^o,\pi^n,\pi^r} W^I(\mathbf{I},x), \forall\mathbf{I}^o, x\in\{1,2,...,M\}. \tag{23}$$

Additionally, let $W(\mathbf{I}^o,0)$ be the utility if the customer chooses to leave the system immediately. This represents the

$$
\begin{aligned}
&Pr(\mathbf{I}[t+1]|\mathbf{I}[t],\pi^n,\pi^r,x) = \\
&\begin{cases}
\rho\lambda f(s[t+1]|\theta), & \mathbf{I}[t+1]\in\mathcal{I}^a_{\mathbf{I}[t],\pi^r}; \\
(1-\rho)\lambda f(s[t+1]|\theta), & \mathbf{I}[t+1]\in\mathcal{I}^a_{\mathbf{I}[t],\pi^n}; \\
(n_j[t])\mu f(s[t+1]|\theta), & \mathbf{I}[t+1]\in\mathcal{I}^d_{\mathbf{I}[t]}, n_j[t+1]=n_j[t]-1, j\neq x; \\
(n_x[t]-1)\mu f(s[t+1]|\theta), & \mathbf{I}[t+1]\in\mathcal{I}^d_{\mathbf{I}[t]}, n_x[t+1]=n_x[t]-1; \\
(1-\mu(\sum_{j=1}^M n_j-1)-\lambda)f(s[t+1]|\theta), & \mathbf{I}[t+1]\in\mathcal{I}^u_{\mathbf{I}[t]}, \mathcal{I}^a_{\mathbf{I}[t],\pi^r}\neq\emptyset, \mathcal{I}^a_{\mathbf{I}[t],\pi^n}\neq\emptyset; \\
(1-\mu(\sum_{j=1}^M n_j-1)-\rho\lambda)f(s[t+1]|\theta), & \mathbf{I}[t+1]\in\mathcal{I}^u_{\mathbf{I}[t]}, \mathcal{I}^a_{\mathbf{I}[t],\pi^r}\neq\emptyset, \mathcal{I}^a_{\mathbf{I}[t],\pi^n}=\emptyset; \\
(1-\mu(\sum_{j=1}^M n_j-1)-(1-\rho)\lambda)f(s[t+1]|\theta), & \mathbf{I}[t+1]\in\mathcal{I}^u_{\mathbf{I}[t]}, \mathcal{I}^a_{\mathbf{I}[t],\pi^r}=\emptyset, \mathcal{I}^a_{\mathbf{I}[t],\pi^n}\neq\emptyset; \\
(1-\mu(\sum_{j=1}^M n_j-1)f(s[t+1]|\theta), & \mathbf{I}[t+1]\in\mathcal{I}^u_{\mathbf{I}[t]}, \mathcal{I}^a_{\mathbf{I}[t],\pi^r}=\mathcal{I}^a_{\mathbf{I}[t],\pi^n}=\emptyset; \\
0, & \text{else.}
\end{cases}
\end{aligned}
\tag{21}
$$

$$
\begin{aligned}
&Pr(\mathbf{I}^{'o}|\mathbf{I}^o,\pi^r,x) = \\
&\begin{cases}
\sum_{\theta\in\Theta} Pr(\theta=k|\mathbf{I}^o,\pi^n,\pi^r)f(s'|\theta), & \mathbf{h}'=\{h_2,h_3,...,x\},\ n'_x=n_x+1,\ n'_j=n_j,\forall j,x>0,; \\
\sum_{\theta\in\Theta} Pr(\theta=k|\mathbf{I}^o,\pi^n,\pi^r)f(s'|\theta), & \mathbf{h}'=\{h_2,h_3,...,0\},\ \mathbf{n}'=\mathbf{n},\ x=0 \text{ or } \sum_{j=1}^M n_j=N; \\
0, & \text{else.}
\end{cases}
\end{aligned}
\tag{22}
$$

lower bound of the expected utility a customer will enter the restaurant. In this paper, we let$W(\mathbf{I}^o,0)=0,\forall\mathbf{I}^o$ without losing generality.

Rational customers seek to maximize their long-term expected utility given the action they applied by other customers in previous and future states. The actions applied by other customers are captured by the rational policy π^r. Nevertheless, its own action will also influence the system state transition. Let $Pr(\mathbf{I}^{'o}|\mathbf{I}^o,\pi^r,x)$ be the probability that the observed state transits from $\mathbf{I}^o$ to $\mathbf{I}^{'o}$ if she selects table x. The optimal action for the customer is to choose the table that maximizes the expected utility. We thus have

$$\pi^r(\mathbf{I}^o) = \arg\max_{x\in\{0,1,2,...,M\}}\sum_{\mathbf{I}^{'o}} Pr(\mathbf{I}^{'o}|\mathbf{I}^o,\pi^r,x)W(\mathbf{I}^{'o},x). \tag{24}$$

The exact transition probability from $\mathbf{I}^o$ to $\mathbf{I}^{'o}$ depends on the form of the observed state. The transition in grouping information and history information is unique since it only involves exactly one customer joining a table. Clearly, when $x>0$, we have $n'_x=n_x+1$ and $\mathbf{h}'=\{h_2,h_3,...,x\}$ in the transition state $\mathbf{I}$ and the corresponding $\mathbf{I}^o$. On the other hand, when the customer chooses not to enter the restaurant, there is no change in the grouping. Therefore, we have $n'_x=n_x$ and $\mathbf{h}'=\{h_2,h_3,...,0\}$ when $x=0$.

Nevertheless, the newly generated signal s' conditions on the restaurant state θ, which is in the hidden state $\mathbf{I}^h$. We therefore need to estimate the new signal s' based on the belief on the restaurant state. Concluding from above, we have the transition probability of observed states in (22) based on the observed state $\mathbf{I}^o$ and the belief from Grand Information Extraction $g_{\mathbf{I}|\mathbf{I}^o,\pi^n,\pi}$.

The coupling relation between the long-term expected utility $W(\mathbf{I}^o)$ and the rational policy $\pi^r(\mathbf{I}^o)$ captures the influences of any customer's action on the expected utility. Then, according to the Nash equilibrium of H-CRG we defined in Definition 1, we have the equilibrium conditions of H-CRG as follows:

Theorem 1. *The Nash equilibrium of H-CRG is* $\pi^*(\mathbf{I}^o)$ *if*

$$W^{I*}(\mathbf{I},x,\pi^*) = R(\mathbf{I},x) \tag{25}$$
$$+(1-\mu)\sum_{\mathbf{I}'} Pr(\mathbf{I}'|\mathbf{I},\pi^n,\pi^*,x)W^{I*}(\mathbf{I}',x,\pi^*),$$

$$W^*(\mathbf{I}^o,x) = \sum_{\mathbf{I}\in\mathcal{I}^o_{\mathbf{I}^o}} g_{\mathbf{I}|\mathbf{I}^o,\pi^n,\pi^*}W^I(\mathbf{I},x), \tag{26}$$

$$\pi^*(\mathbf{I}^o) = \arg\max_x\sum_{\mathbf{I}^{'o}} Pr(\mathbf{I}^{'o}|\mathbf{I}^o,\pi^*,x)W^*(\mathbf{I}^{'o},x), \tag{27}$$

for all $\mathbf{I}$, $\mathbf{I}^o$, $x\in\{1,2,...,M\}$.

Proof. Given a policy π^*, the expected utility of any rational customer applying action x at state $\mathbf{I}^o$ is given by (26). In addition, given (27), we have

$$
\begin{aligned}
E[U(\pi^*(\mathbf{I}^o),\pi^*_{-\mathbf{I}^o}(\mathbf{I}^o))] &= W^*(\mathbf{I}^o,\pi^*(\mathbf{I}^o) \\
&\geq W^*(\mathbf{I}^o,x) = E[U(x,\pi^*_{-\mathbf{I}^o}(\mathbf{I}^o))]
\end{aligned}
$$

Therefore, the policy π^* is a Nash equilibrium according to Definition 1. □

V. Solutions

A. Centralized Policy

We first analyze the socially-optimal policy for the proposed H-CRG. The socially-optimal policy is the solution that maximizes the expected social-welfare of the whole system. This solution serves as the performance bound provided by centralized-control solutions. We define the social welfare as the average total utility of all customers in the restaurant,

$$SW = \lim_{T\to\infty}\sum_{t=0}^{T}\frac{\sum_{j=1}^M n_jU(R_j(\theta),n_j)|_{\mathbf{I}[t]}}{T}. \tag{28}$$

The expected total utility given the observed state $\mathbf{I}^o$, which we defined as $W^s(\mathbf{I}^o)$, can be given by the Bellman equation

$$\begin{aligned} W^s(\mathbf{I}^o) &= \mu' \sum_{\mathbf{I}'^o} Pr(\mathbf{I}'^o|\mathbf{I}^o, \pi^r) W^s(\mathbf{I}'^o) \\ &+ \sum_{\mathbf{I} \in \mathcal{I}^o_{\mathbf{I}^o}} g_{\mathbf{I}|\mathbf{I}^o,\pi^n,\pi^r} \sum_{j=1}^{M} n_j U(R_j(\theta), n_j)|_{\mathbf{I}}. \end{aligned} \tag{29}$$

Notice that $Pr(\mathbf{I}'^o|\mathbf{I}^o, \pi^r)$ can be derived by reducing (10) to the observed state domain.

This form resembles a Markov Decision Process (MDP) except that the immediate reward is not only related to the current action but also the actions in other states due to the Grand Information Extraction. In such a case, it is very challenging to find the socially-optimal solution. Instead, we seek to find the centralized policy which maximizes the current expected social-welfare at each instance, or the expected social-welfare given the currently applied policy. Notice that this is also a common objective in traditional MDP problems.

The centralized policy is given by

$$\pi^s(\mathbf{I}^o) = \arg \max_{x \in \{0,1,2,\ldots,M\}} \sum_{\mathbf{I}'^o} Pr(\mathbf{I}'^o|\mathbf{I}^o, x) W^s(\mathbf{I}'^o) \tag{30}$$

We then propose to use value-iteration algorithm to find the centralized policy. The proposed algorithm is different from the traditional value-iteration algorithm in MDP. When a policy is updated, not only the expected social welfare but also the immediate social welfare is updated. The immediate reward social welfare is updated by the Grand Information Extraction in order to derive the correct belief on the hidden states under the new policy. The algorithm is shown in Algorithm 1.

Algorithm 1 Value Iteration for centralized Solution

```
Initialize $\pi^{s,0}, W^{s,0}, l = 0$
while $1$ do
    $l \leftarrow l + 1$
    $Pr(\mathbf{I}'^o|\mathbf{I}^o, \pi^n, \pi^r) \leftarrow$ (22)
    for all $\mathbf{I}^o$ do
        $\pi^{s,l} \leftarrow$ (30)
        $W^{s,l} \leftarrow$ (29)
    end for
    $W^{s,d} \leftarrow W^{s,l+1} - W^{s,l}$
    if $\max W^{s,d} - \min W^{s,d} < \epsilon$ then
        Break
    end if
end while
Output $\pi^{r,l}$ and $W^{s,l}$
```

The centralized policy may provides superior performance from the service operator (i.e. restaurant owner)'s perspective. Nevertheless, it doesn't consider the rationality of rational customers. In some cases, the centralized policy requires some rational customers choose the tables which are beneficial to the system but sub-optimal for their own utility. In such a case, These customers may refuse to follow the centralized policy if no extra incentive mechanism is introduced [5].

B. Nash Equilibrium

When it turns to the original H-CRG setup where rational customers choose to maximize their own long-term expected utilities, it is more challenging to derive the final outcome, that is, the Nash equilibrium, due to the competitions among customers. Inspired by the value-iteration algorithm for centralized solution, we propose a value-iteration algorithm to finding the Nash equilibrium in the proposed H-CRG. The algorithm is shown in Algorithm 2.

Algorithm 2 Value-Iteration Algorithm for Nash Equilibrium

```
Initialize $\pi^r, W, W^I$;
while $1$ do
    $g_{\mathbf{I}|\mathbf{I}^o,\pi^n,\pi^r} \leftarrow (17)$
    for all $\mathbf{I}^o$ do
        $\pi^{r'} \leftarrow (27)$;
        $W^{I'} \leftarrow (25)$;
        $W' \leftarrow (26)$;
    end for
    $W^d \leftarrow W' - W$
    if $\max W^d - \min W^d < \epsilon$ then
        Break
    else
        $W \leftarrow W'$, $W^I \leftarrow W^{I'}$, $\pi^r \leftarrow \pi^{r'}$
    end if
end while
Output $\pi^r$, $W$, and $W^I$
```

Lemma 2. *The output of Algorithm 2, if converged, is the Nash equilibrium of the H-CRG when $epsilon \rightarrow 0$.*

Proof. It can be easily seem that when Algorithm 2 converges with $\epsilon = 0$, all conditions in Theorem 1, that is, (25), (26), and (27) are satisfied. Therefore, the output policy π^r is the Nash equilibrium of the proposed H-CRG. □

Unfortunately, we find that the pure-strategy Nash equilibrium may not exist when the belief on the hidden state is highly influenced by the choices applied by other customers observed in the history information. In some cases, the rational policy applied by rational customers will be damping from one to another in the stochastic system, which we called **Information Damping**. Specifically, the optimal choice of a rational customer depends on her belief on the restaurant state. The belief is conditioned on both her received signal and observed actions, while the information contained in the later term depends on whether other customers make decisions following their signals or not. When all rational customers follow their own signal, the information contained in previous actions could be stronger than her own signal when the length of observed history is long enough. When the information is strong enough to overcome the signal, the customers may choose to follow the actions while ignoring their own signals. This decision, on the other hand, reduces the information contained in the observed actions. In some cases, the information in the action is reduced to a degree that the received signal becomes more informative than the observed actions. Then, customers switch back to follow the observed signals instead of the observed actions. We called this Information Damping as customers may alternatively choose to follow the signals or the action of others. This phenomenon is similar to Information Cascade in traditional social learning problems, where the information contained in the observed action will be constrained due to the information diffusion structure [24], [25]. Nevertheless, the Information Damping we discussed here further points out that the information contained in the actions could be **reduced** in a stochastic system. Information damping is one of the key factor that leads to the nonexistence of pure-strategy Nash equilibrium, which may influence the stability of the system.

The root cause of information damping is the loss of information in the observed action due to rational choices of customers. Nevertheless, it turns out that a way to avoid this is to consider not only the rational customers but also naive customers in the system. The naive customers, who follow a predetermined policy to select the table given the observed state, is commonly seen in most systems. In wireless networks, for instance, the naive customers could be the legacy devices which follow the existing protocols without strategic thinking. In social networks, on the other hand, the naive customers could be the agents who are naive with short-term memory, which is commonly observed in the literature [26], [27]. The naive policy π^n followed by the naive customers are predictable, not influenced by the rational customers, and potentially can be informative if the received signal influences the output of the policy π^n.

The action of these naive customers will be treated as an external information source to reveal the hidden state in the proposed Grand Information Extraction process. Nevertheless, it can be difficult to distinguish the action of naive customers from rational ones. In the proposed H-CRG, as we illustrated in Section III, we assume that these actions are indistinguishable. The naive policy π^n applied by these naive customers, on the other hand, is known by the rational customers. When the ratio of naive customers increases, the observed action potentially will be more informative and predictable. The proposed Grand Information Extraction process will automatically extract the information conditioning on the naive policy and the customer type ratio. Due to the complexity of the process, it is still an open problem to derive the lower bound of customer type ratio to guarantee the existence of pure-strategy Nash equilibrium.

VI. Application: Channel Access in Cognitive Radio Networks

In this section we introduce an important application of H-CRG: channel access in cognitive radio networks. We describe the problem first and then illustrate the corresponding H-CRG model. We then evaluate the performance of H-CRG through data-driven simulations [8].

We consider a cognitive radio network with some primary users and secondary users who share the channels. The primary users have the higher priority to access the channels. That is, secondary users are not allowed to access the channel as long as the primary users already occupied it. In some cases, the secondary users may need to pay a penalty if they accidentally interfere the primary user transmissions. Secondary users therefore are required to detect the activity of primary users through channel sensing before the actual transmission. Nevertheless, the channel sensing is imperfect, especially when the protocol of primary users is unknown. Either miss detection or false alarm may damage the service quality experienced by both primary and secondary users. Cooperative sensing is a popular approach to enhance the detection accuracy through aggregating the sensing results from all nodes. The aggregated results could lead to a better consensus on the channel states and therefore better decisions on channel access.

Here we propose a new approach for cooperative sensing inspired by the stochastic social learning techniques in H-CRG. The secondary users now detect not only the activity of primary users but also the access attempts of other secondary users. Specifically, a secondary user will first wait for few slots and detect the access attempts of other secondary users in the channels. Then, it will detect the activity of primary user through traditional channel sensing. The secondary user may learn the sensing results of other secondary customers from the collected channel access pattern. The advantage of this approach is that no control channel or signaling exchanges between secondary users are required, which makes it more practical for networks with limited channel resources. Nevertheless, such an advantage comes with a cost: each secondary user must wait for a period of time before access the channel. This introduces an extra delay and therefore reduces the average throughput. When the accuracy of sensing result is high, the cost in extra delay may cancel out the benefit retrieved in the proposed learning process.

We now formulate the system model. Let us consider M channels, while one of them is currently occupied by the primary users. The secondary users sense the channels and determine the channel occupied by the primary user. We assume that the sensing is imperfect, that is, with a probability of $p < 1$ that the occupied channel will be detected correctly. The access probability of the secondary users per slot is given by $p_a < 1$. Notice that when multiple users access the channel at the same time, the transmissions will collide with each others and failed. Therefore, given that the channel is unoccupied by the primary users and k secondary users select the same channel, the expected access opportunity a secondary user may get is

$$E[u(1,k)] = p_a(1-p_a)^{k-1}. \tag{31}$$

On the other hand, when the channel is occupied by the primary user, the secondary user who attempts to access the same channel will need to pay a penalty:

$$E[u(0,k)] = C < 0. \tag{32}$$

When a secondary user arrives, she selects a channel for accessing. We assume that she will wait for H slots and detect the access pattern of other customers. At the final slot, she will also sense the activity of primary users in each channel. We called these slots as sensing slots. Then, she will choose the channel to access with until her departure. We call these

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSIPN.2017.2682799, IEEE Transactions on Signal and Information Processing over Networks

following slots are accessing slots. We assume that there are two kinds of secondary users: legacy and strategic users. The legacy secondary users will access the channel following her own sensing results. This represents the legacy devices without cooperative sensing capability. Strategic secondary users, on the other hand, will utilize all the observed information to select the channel giving him the largest expected utility.

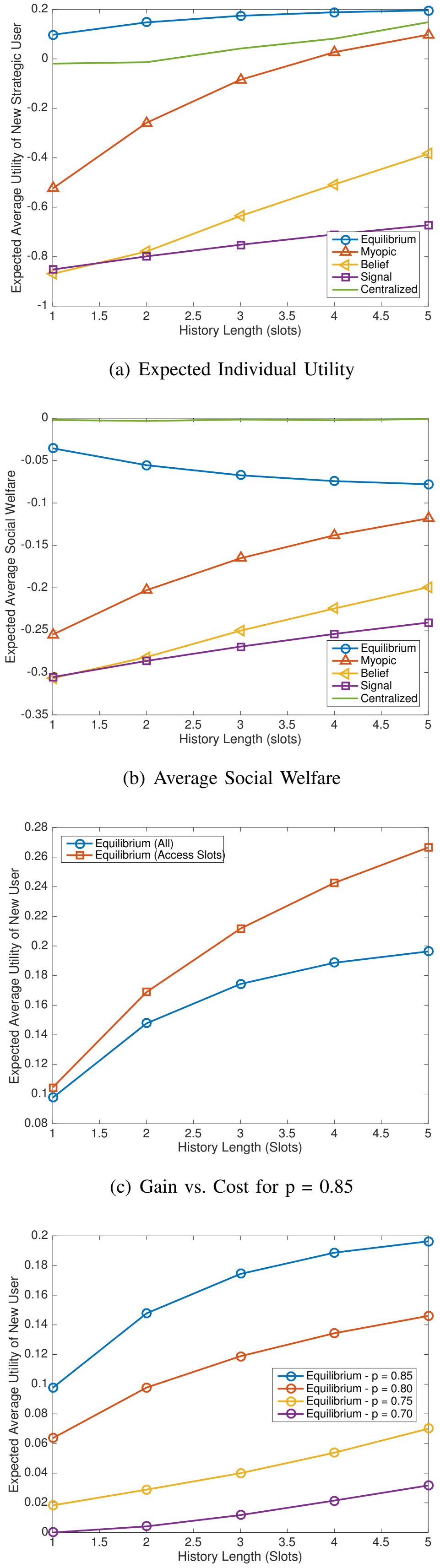
The long-term expected utility of a secondary user is the average expected successful access attempt before he leaves the system, including both sensing slots and accessing slots. Notice that all secondary users will wait in the sensing slots before access the channel, which means that each secondary users have zero access opportunities for the first H slots. The larger the H, the smaller portion of accessing slots in the duration of a secondary user's stay.

We can formulate the channel access problem as a H-CRG if we treat M channels as tables, the sensing results as the signal s, and the utility as the access opportunity minus the penalty. In addition, the restaurant state $\theta = \{1, 2, ..., M\}$ denotes the channel occupied by primary users. We may derive both the centralized policy and Nash equilibrium policy for the strategic secondary users. Notice that in this system a secondary user may observe the action applied by previous users but not the current grouping. That is, the observed state $\mathbf{I}^o = \{\mathbf{h}, s\}$. Therefore, the rational customers not only need to derive the belief on the primary user occupation but also estimate the number of secondary users choosing the same channel.

A. Simulation Results

We evaluate the performance of proposed policies through data-driven simulations with user process and network models using time-invariant parameters extracted from the real dataset. Specifically, the proposed H-CRG framework requires several components, such as the user arrival process, departure process, and utility functions. We extract the required parameters for each component from the dataset and determine the appropriate settings for the problem. In the simulation, the parameters of user process and network models are extracted from the CRAWDAD Dartmouth campus WLAN trace [8], [9]. The arrival and departure of secondary users follow the distribution extracted from the trace in the dataset. For the utility function, we consider the slotted ALOHA access mechanism and assume that secondary users focus on the successful access attempts, with a penalty imposed by the primary user if an interference occurred. The sensing accuracy, on the other hand, depends on the sensing techniques the system applied and here we left it as an adjustable parameter in the simulations. It can be replaced with the corresponding accuracy when certain sensing technique is chosen.

We simulate a cognitive radio network with 2 channels and maximum 8 secondary users. The length of a slot is 5 minutes. The primary users may choose either channel to occupy with equal probability. The arrival and departure of secondary users follow the distribution extracted from [8], that is, the arrival rate of naive and strategic secondary users are 0.2106 and 0.1479 per slot, respectively. The departure rate per secondary user is 0.0715 per slot [9]. For each secondary user in the



(a) Expected Individual Utility

(b) Average Social Welfare

(c) Gain vs. Cost for p = 0.85

(d) Gain Gained from History Length

Fig. 2. Influence of History Length

network, the access probability per slot is 0.7. The penalty for interfering primary user is -0.7 per slot.

We compare the performance of the derived policy from H-CRG with four other policies: signal, belief, myopic, and centralized. The signal policy is the same as the policy applied by legacy devices, where the users always access the channel following their own signal. This demonstrates the performance bound if the secondary users do not cooperate at all. The belief policy represents the strategy to follow the belief on the primary user occupation extracted by the Grand Information Extraction by (15), but to ignore the estimation on the number of users. This shows the performance upper bound if secondary users cooperate with each other in the sensing but ignore the effect of network externality. The myopic policy represents the strategy that both the primary user occupation and number of secondary users are estimated by Grand Information Extraction, but the payoff in the future slots is ignored. This shows the performance bound if secondary users consider both the sensing results and network externality, but ignore the influences of other secondary users in the future. Finally, the centralized policy is the policy we derived from Algorithm 1. This represents the policy that there exists a centralized-control node to determine the channel access policy for secondary users. It serves as the performance upper bound of social welfare if a centralized control mechanism is applied.

We evaluate the performance of different policies with two metrics: expected long-term individual utilities for new strategic users and average social welfare per slot.

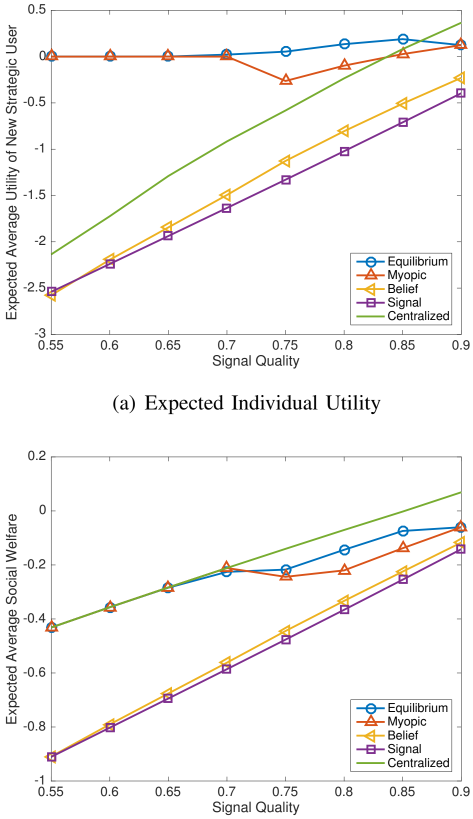
We first evaluate the influence of history length to the performance of all policies. We let the signal quality $p = 0.85$ and then simulate with different history length from 1 to 5. This simulation helps us understand whether the increased history information helps improve the accuracy of the belief on the hidden state and the utilities of the secondary users. The simulation results are shown in Fig. 2. A clear trend shows that the increase in history length benefits the expected individual utility of the user. This is due to the increase in the observed state space, which contains more information to be extracted by Grand Information Extraction. We observe that the proposed equilibrium policy from H-CRG provides highest individual expected utility among all the policies. In addition, H-CRG is the only policy which guarantees positive expected utilities for new users. Interestingly, the increases in H also brings a positive effect on the expected utility of signal policy. This is due to the fact that the signal policy receives a negative expected utility and the accuracy is not affected by the history length. Given that the expected utility is negative in accessing slots, a reduction in the portion of accessing slots will bring a positive impact on the expected utility under the signal policy.

For the social welfare, on the other hand, we observe that the trend is inverse for the equilibrium policy for H-CRG. This is due to the fact that the increase in the history length suggests better understanding on the hidden state, which also leads to a fiercer competitions among the secondary users. This may have a negative impact on the social welfare. Nevertheless, the proposed equilibrium policy is still closest to the centralized policy. We also observe that H-CRG outperforms centralized policy in expected utility of new secondary users but not in social welfare. The reason is that some users may be sacrificed in order to achieve better social welfare in centralized policy. For instance, some secondary users may be forbidden to enter the system in centralized policy in order to protect other existing users from higher collision rates in channel access, even if these new users may receive positive utilities if they enter. Such protections lead to a higher social welfare but may impale the utility of new customers. For H-CRG, on the other hand, new secondary users will access the channel which maximize their own utilities, regardless whether this attempt will damage the social welfare. Therefore, the expected utility of new customer should be higher with this policy, in exchange of a lower social welfare.

Next, we discuss the impact of sensing slots on the average utility. We first illustrate the average utility of secondary users in the whole duration and access slots in Fig. 2(c). The differences between these two lines are the overhead due to the extra delay in sensing slots. We observe that the overhead introduced by the sensing slots grows significantly with the increase of history length. On the other hand, the increase of average utility in access slots diminishes with the increase of history length. This suggests that a proper history length exists in order to balance the accuracy and the extra overhead brought by the sensing slots. The results also suggest that the resulting performance is concave in history length, therefore the optimal history length can be easily found. We also illustrate the average utility of secondary users under different signal quality in Fig. 2(d). We observe that the increase of utility brought by larger history lengths are more significant when the signal quality is low. This is due to the fact that when the signal quality is low, the increase of accuracy due to the extended history is beneficial enough to compensate the overhead introduced by the extra delay. The results suggest that the sensing slots should be expanded when the signal quality is low, and vice versa.

We then evaluate the influence of signal quality (channel sensing accuracy) to the performance of all policies. We let the history length $H = 4$ and then simulate with different signal quality from 0.9 to 0.45. The results are shown in Fig. 3. We observe that the increase in signal quality benefits the expected individual utility of the user. This is due to the increased signal quality and therefore the information contained within. Nevertheless, we observe that the proposed equilibrium policy from H-CRG still provides highest individual expected utility among all the policies except when the signal quality is high. For the case that the signal quality is high, the centralized policy provides a higher expected utility. Nevertheless, the centralized policy is sub-optimal for some users at certain states, therefore is unstable and cannot be implemented without additional incentive mechanisms.

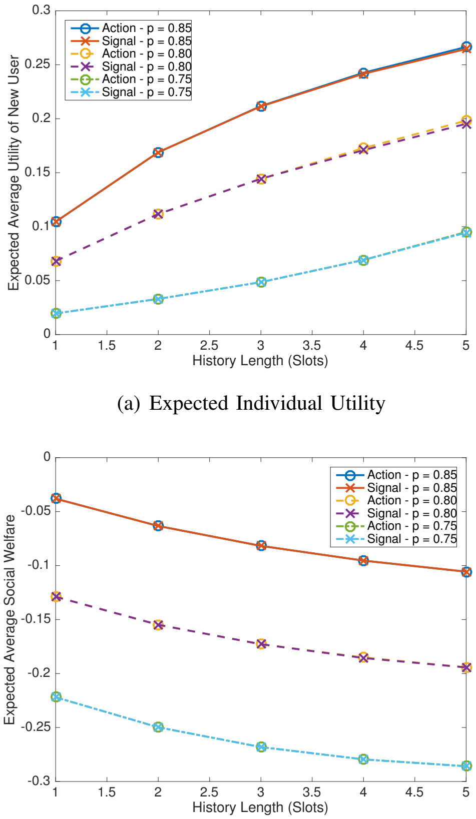
For the social welfare, on the other hand, we observe that equilibrium, myopic, and centralized policies provide same social welfare when the signal quality is low, but the social welfare degrades for both equilibrium and myopic ones when the signal quality is high. This is still the influence of fiercer competition when the system state is more accurately identified by better signal. This may have a negative impact on the social welfare. Nevertheless, the performance of the



(a) Expected Individual Utility

(b) Average Social Welfare

Fig. 3. Influence of Signal Quality



(a) Expected Individual Utility

(b) Average Social Welfare

Fig. 4. Action vs. Signal-based Model

proposed equilibrium policy still is closest to the centralized policy in most cases.

Finally, we compare the performance of the proposed framework with different information sources. Specifically, we replace the actions revealed in the observed state in H-CRG with signals each user receives when she arrives the system. In other words, users reveal their signals to other users when they access the channel in the revised model. We define the original H-CRG as action-based model while the revised one as signal-based model. The revised model represents the social learning system using user-generated signals as information source. Notice that all proposed algorithms still apply to the revised model. The expected utility of new customers and social welfare under both models are shown in Fig. 4. We observe that the performance loss from signal-based to action-based model, if any, is negligible. The result showed that the revealed actions already contain enough information for rational customers to learn the hidden information and make proper decisions.

VII. Conclusions

We propose a new stochastic game-theoretic framework, Hidden Chinese Restaurant Game, to utilize the passively-revealed actions instead of user-generated contents as the main information source in social learning process. Based on the stochastic state transition structure and inspired by the Bayesian learning in Chinese Restaurant Game, the propose Grand Information Extraction can extract the belief on the hidden information directly from the observed actions. The proposed belief extraction process is universal for any Markovian system. The coupling relation between the belief and the policy is further utilized to transform the original continuous-state formulation in traditional Partial-Observed MDP (POMDP) into a discrete-state MDP. The optimal policy is then analyzed in both the centralized approach and the game-theoretic approach. We notice that the pure-strategy Nash equilibrium may not exist in the H-CRG. Specifically, the Information Damping phenomenon may rise in a certain H-CRG in which customers synchronously switch from one policy to another due to the information loss in the observed action. Fortunately, the existence of naive customers can help ensure the existence of pure-strategy Nash equilibrium. Their actions can be treated as signals to stabilize the belief of agents on the observed actions. We evaluate the performance of H-CRG through simulations by applying the framework to the channel access problem in cognitive radio networks. We conduct data-driven simulations using the CRAWDAD Dartmouth campus WLAN trace. The simulation results showed that the equilibrium strategy derived in H-CRG provides higher expected utilities for new users and maintains a reasonable high social welfare comparing with other candidate strategies.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSIPN.2017.2682799, IEEE Transactions on Signal and Information Processing over Networks

References

[1] C.-Y. Wang, Y. Chen, and K. J. R. Liu, “Chinese restaurant game,” *IEEE Signal Processing Letters*, vol. 19, no. 12, pp. 898 –901, Dec. 2012.

[2] C. Y. Wang, Y. Chen, and K. J. R. Liu, “Sequential chinese restaurant game,” *IEEE Transactions on Signal Processing*, vol. 61, no. 3, pp. 571–584, Feb 2013.

[3] C. Jiang, Y. Chen, Y.-H. Yang, C.-Y. Wang, and K. J. R. Liu, “Dynamic chinese restaurant game: Theory and application to cognitive radio networks,” *IEEE Transactions on Wireless Communications*, vol. 13, no. 4, pp. 1960–1973, April 2014.

[4] Y.-H. Yang, Y. Chen, C. Jiang, C.-Y. Wang, and K. J. R. Liu, “Wireless access network selection game with negative network externality,” *Wireless Communications, IEEE Transactions on*, vol. 12, no. 10, pp. 5048–5060, October 2013.

[5] C. Y. Wang, Y. Chen, H. Y. Wei, and K. J. R. Liu, “Scalable video multicasting: A stochastic game approach with optimal pricing,” *IEEE Transactions on Wireless Communications*, vol. 14, no. 5, pp. 2353–2367, May 2015.

[6] Y. Chen, C. Jiang, C. Y. Wang, Y. Gao, and K. J. R. Liu, “Decision learning : Data analytic learning with strategic decision making,” *IEEE Signal Processing Magazine*, vol. 33, no. 1, pp. 37–56, Jan 2016.

[7] V. Bala and S. Goyal, “Learning from neighbours,” *The Review of Economic Studies*, vol. 65, no. 3, p. 595, 1998.

[8] D. Kotz, T. Henderson, I. Abyzov, and J. Yeo, “CRAWDAD dataset dartmouth/campus (v. 2009-09-09),” Downloaded from http://crawdad.org/dartmouth/campus/20090909, Sep. 2009.

[9] Y. H. Yang, Y. Chen, C. Jiang, and K. J. R. Liu, “Wireless network association game with data-driven statistical modeling,” *IEEE Transactions on Wireless Communications*, vol. 15, no. 1, pp. 512–524, Jan 2016.

[10] B. Golub and M. O. Jackson, “Naive Learning in Social Networks and the Wisdom of Crowds,” *American Economic Journal: Microeconomics*, vol. 2, no. 1, pp. 112–149, 2010.

[11] D. Acemoglu, M. Dahleh, I. Lobel, and A. Ozdaglar, “Bayesian Learning in Social Networks,” *LIDS report 2780, Review of Economic Studies*, Jan. 2011.

[12] D. Acemoglu and A. Ozdaglar, “Opinion dynamics and learning in social networks,” *Dynamic Games and Applications*, vol. 1, pp. 3–49, 2011.

[13] G. Angeletos, C. Hellwig, and A. Pavan, “Dynamic global games of regime change: Learning, multiplicity, and the timing of attacks,” *Econometrica*, vol. 75, no. 3, pp. 711–756, 2007.

[14] J. Costain, “A herding perspective on global games and multiplicity,” *The BE Journal of Theoretical Economics*, vol. 7, no. 1, p. 22, 2007.

[15] V. Krishnamurthy, “Quickest detection pomdps with social learning: Interaction of local and global decision makers,” *IEEE Transactions on Information Theory*, vol. 58, no. 8, pp. 5563–5587, Aug 2012.

[16] J. A. Bohren, “Stochastic Games in Continuous Time: Persistent Actions in Long-Run Relationships, Second Version,” *SSRN Electronic Journal*, 2014.

[17] H. Zhang, “Partially Observable Markov Decision Processes: A Geometric Technique and Analysis,” *Operations Research*, vol. 58, no. 1, pp. 214–228, Jul. 2009.

[18] G. Shani, J. Pineau, and R. Kaplow, “A survey of point-based POMDP solvers,” *Autonomous Agents and Multi-Agent Systems*, vol. 27, no. 1, pp. 1–51, 2013.

[19] N. Meuleau, L. Peshkin, K.-E. Kim, and L. P. Kaelbling, “Learning finite-state controllers for partially observable environments,” in *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*, ser. UAI’99, 1999, pp. 427–436.

[20] N. Meuleau, K.-E. Kim, L. P. Kaelbling, and A. R. Cassandra, “Solving pomdps by searching the space of finite policies,” in *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*, ser. UAI’99, 1999, pp. 417–426.

[21] T. Yucek and H. Arslan, “A survey of spectrum sensing algorithms for cognitive radio applications,” *IEEE Communications Surveys & Tutorials*, vol. 11, no. 1, pp. 116–130, 2009.

[22] A. Ali and W. Hamouda, “Advances on Spectrum Sensing for Cognitive Radio Networks: Theory and Applications,” *IEEE Communications Surveys & Tutorials*, vol. PP, no. 99, pp. 1–1, 2016.

[23] L. Zhang, G. Ding, Q. Wu, Y. Zou, Z. Han, and J. Wang, “Byzantine Attack and Defense in Cognitive Radio Networks: A Survey,” *IEEE Communications Surveys & Tutorials*, vol. 17, no. 3, pp. 1342–1363, 2015.

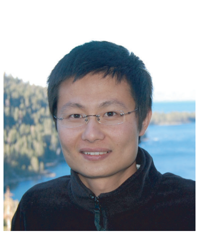
[24] F. Chierichetti, J. Kleinberg, and A. Panconesi, “How to schedule a cascade in an arbitrary graph,” in *Proceedings of the 13th ACM Conference on Electronic Commerce (EC)*, Jun. 2012.

[25] M. Hajiaghayi, H. Mahini, and D. Malec, “The polarizing effect of network influences,” in *Proceedings of the fifteenth ACM conference on Economics and computation (EC)*, Jun. 2014, pp. 131–148.

[26] D. Fudenberg and A. Peysakhovich, “Recency, records and recaps: learning and non-equilibrium behavior in a simple decision problem,” in *Proceedings of the fifteenth ACM conference on Economics and computation (EC)*, Jun. 2014, pp. 971–986.

[27] S. Zhang and A. J. Yu, “Forgetful Bayes and myopic planning: Human learning and decision-making in a bandit setting,” in *Advances in Neural Information Processing Systems (NIPS)*, 2013, pp. 2607–2615.

Chih-Yu Wang (M’13) received the B.S. and Ph.D. degrees in electrical engineering and communication engineering from National Taiwan University (NTU), Taipei, Taiwan, in 2007 and 2013, respectively. He has been a visiting student in University of Maryland, College Park in 2011. He is currently an Assistant Research Fellow with the Research Center for Information Technology Innovation, Academia Sinica, Taipei, Taiwan. His research interests include game theory, wireless communications, social networks, and data science.



Yan Chen (SM’14) received the bachelor’s degree from the University of Science and Technology of China in 2004, the M.Phil. degree from the Hong Kong University of Science and Technology in 2007, and the Ph.D. degree from the University of Maryland, College Park, MD, USA, in 2011. He was with Origin Wireless Inc. as a Founding Principal Technologist. Since Sept. 2015, he has been a full Professor with the University of Electronic Science and Technology of China. His research interests include multimedia, signal processing, game theory, and wireless communications.

He was the recipient of multiple honors and awards, including the Best Student Paper Award at the IEEE ICASSP in 2016, the best paper award at the IEEE GLOBECOM in 2013, the Future Faculty Fellowship and Distinguished Dissertation Fellowship Honorable Mention from the Department of Electrical and Computer Engineering in 2010 and 2011, the Finalist of the Dean’s Doctoral Research Award from the A. James Clark School of Engineering, the University of Maryland in 2011, and the Chinese Government Award for outstanding students abroad in 2010.

K. J . Ray Liu (F’03) was named a Distinguished Scholar-Teacher of University of Maryland, College Park, in 2007, where he is Christine Kim Eminent Professor of Information Technology. He leads the Maryland Signals and Information Group conducting research encompassing broad areas of information and communications technology with recent focus on smart radios for smart life.

Dr. Liu was a recipient of the 2016 IEEE Leon K. Kirchmayer Technical Field Award on graduate teaching and mentoring, IEEE Signal Processing Society 2014 Society Award, and IEEE Signal Processing Society 2009 Technical Achievement Award. Recognized by Thomson Reuters as a Highly Cited Researcher, he is a Fellow of IEEE and AAAS.

Dr. Liu is a member of IEEE Board of Director. He was President of IEEE Signal Processing Society, where he has served as Vice President Publications and Board of Governor. He has also served as the Editor-in-Chief of IEEE Signal Processing Magazine.

He also received teaching and research recognitions from University of Maryland including university-level Invention of the Year Award; and college-level Poole and Kent Senior Faculty Teaching Award, Outstanding Faculty Research Award, and Outstanding Faculty Service Award, all from A. James Clark School of Engineering.