Multi-User Computation Offloading in Mobile Edge Computing: A Behavioral Perspective

Ling Tang and Shibo He

Abstract

By providing cloud computing capabilities at the network edge in proximity of mobile device users, mobile edge computing offers an effective solution to help mobile devices with computation-intensive and delay-sensitive tasks. In this article, we investigate the multi-user computation offloading problem in an uncertain wireless environment. Most of the existing works assume that mobile device users are rational and make offloading decisions to maximize their expected objective utilities. However, in practice, users tend to have subjective perceptions under uncertainty, such that their behavior deviates considerably from the conventional rationality assumption. Drawing on the framework of prospect theory (PT), we formulate users' decision making of whether to offload or not as a PT-based non-cooperative game. We propose a distributed computation offloading algorithm to achieve the Nash equilibrium of the game. Numerical results assess the impact of mobile device users' behavioral biases on offloading decision making.

Introduction

In mobile cloud computing [1], mobile device users can offload their tasks to the remote cloud in order to meet the growing demand of computationally expensive applications. However, due to the long-distance wide area network (WAN) between mobile devices and the remote cloud [2], one critical challenge of mobile cloud computing is the large latency experienced by mobile device users. To address this challenge, mobile edge computing (MEC) has recently been proposed as a promising paradigm to satisfy the requirement of computation-intensive and delay-sensitive applications [3, 4]. As illustrated in Fig. 1, by deploying MEC servers at the mobile network edge, MEC can provide computing and storage resources in proximity of mobile device users. Thus, low latency can be achieved by offloading the computation to nearby MEC servers in the edge cloud.

However, due to the uncertainty of the wireless channel and the limited communication resources, the computation offloading performance in MEC hinges heavily on wireless access efficiency. On one hand, the wireless connection between the mobile device and the edge cloud is random and uncertain due to several factors (e.g., weather conditions, wall obstacles), which may result in computation offloading failure. On the other hand, if many mobile device users offload their tasks to the edge cloud simultaneously, each user would experience a long data transmission time for offloading.

To cope with the wireless access efficiency, many previous studies have investigated how to design an energy- and time-efficient computation offloading mechanism in the context of mobile cloud computing [5–10]. Most of these works assume that the mobile device users are rational and behave objectively when making the computation offloading decision. However, as observed by many empirical experiments [11], within uncertain circumstances, the decision of a human being depends on its subjective perception and may deviate from rationality significantly. Obviously, for computation offloading in an uncertain wireless environment, the subjective perception of mobile device users would play an important role in the overall decision process. Therefore, within the context of MEC, this "behavioral" effect needs to be captured when designing a computation offloading mechanism. In this article, we apply the framework of prospect theory (PT) to understand the mobile device user's realistic behavior during computation offloading in MEC, considering the channel uncertainty and competition for the limited communication resource. Specifically, we model the computation offloading decision making problem among multiple mobile device users as a non-cooperative game, where each user aims to maximize its subjective utility based on PT by choosing whether to offload or not. We then propose a distributed computation offloading algorithm to achieve the Nash equilibrium and analyze the impact of PT-related parameters on mobile device users' offloading decision making.

Overview of Mobile Edge Computing

Architecture of MEC

Mobile edge computing, which equips the radio access networks with cloud computing capabilities and hosts the application at the mobile network edge, has the potential to offer an improved user experience [12]. Figure 1 illustrates the basic architecture of MEC. We can see that there are three basic components in the architecture of MEC:

- Mobile devices such as cell phones and personal computers
- MEC servers deployed along with a wireless base station
- Wireless base stations (e.g., third/fourth/fifth generation, 3G/4G/5G, macro base stations and Wi-Fi access points) through which the mobile devices connect with the MEC servers

This work was supported in part by the National Natural Science Foundation of China (61402230), the Natural Science Foundation of Jiangsu Province (BK20140797), the Postdoctoral Foundation of Jiangsu Province (1501049B), and the Zhejiang Provincial Natural Science Foundation of China (LR16F020001).

Digital Object Identifier: 10.1109/MNET.2018.1700119

Ling Tang is with Nanjing University of Science and Technology; Shibo He is with Zhejiang University

0890-8044/18/$25.00 © 2018 IEEE

As the key element in the architecture of MEC, the MEC server provides computing resources and storage capacity to process tasks from mobile device users. According to [4], the MEC server platform has two parts: a hosting infrastructure and an application platform. The hosting infrastructure infers the physical hardware resources and the virtualization layer, and the application platform provides an efficient and flexible environment for hosting application.

We next present the taxonomy of MEC based on the characteristics, applications, access technologies, and key enablers, which is illustrated in Fig. 2.

Characteristics of MEC

With the idea of offering the cloud computing capabilities at the edge of mobile networks, MEC can be characterized by many advantages:

- *Proximity*: Since MEC servers are deployed along with the base station, the mobile devices within the coverage of the base station can access the nearby MEC server through one-hop wireless connection.
- *Low latency*: The round-trip time can be significantly reduced by hosting the applications at the mobile network edge.
- *High bandwidth*: Compared to a WAN, a radio access network offers higher bandwidth for users' data transmission during computation offloading.
- *Network context awareness*: The specific network context information at the mobile network edge, such as network statistics and wireless channel conditions, can be utilized to optimize the resource allocation and offer network context-aware services.
- *Location awareness*: Based on the collected low-level signal information, the MEC service provider can determine the mobile device's location and further exploit it to improve the performance of location-related tasks.

Applications of MEC

With the characteristics mentioned above, there are many applications where MEC can be exploited to enhance performance, such as augmented reality, video acceleration, the Internet of Things (IoT), dynamic content delivery, and connected vehicles. We next discuss the augmented reality application to illustrate the benefits of MEC. Augmented reality combines a person's view of a real-world object and the augmented content display [4]. Consider a visitor pointing his/her mobile device toward a painting at an art gallery. The augmented reality application will display the supplementary content of the painting when it is captured by the camera of the mobile device. Since the visitor may walk around frequently, the presented content needs to be updated quickly. If the application is hosted on the remote cloud, the visitor may receive out-of-date information. However, by bringing the application to the nearby MEC server, the latency will be dramatically reduced, and thus the visitor's experience will be improved.

Access Technologies and Deployment Scenarios

Mobile devices can access the edge cloud through a Wi-Fi access point or a cellular base station (3G/4G/5G). Depending on the access technology, there are different ways to deploy MEC servers. For the indoor scenario, such as airports and shopping malls, edge clouds can be deployed as resourceful on-premise gateways, which can be accessed through Wi-Fi and 3G/4G access points. For the outdoor scenario, powerful MEC servers can be deployed close to the macrocell base station through which users' requests are received.

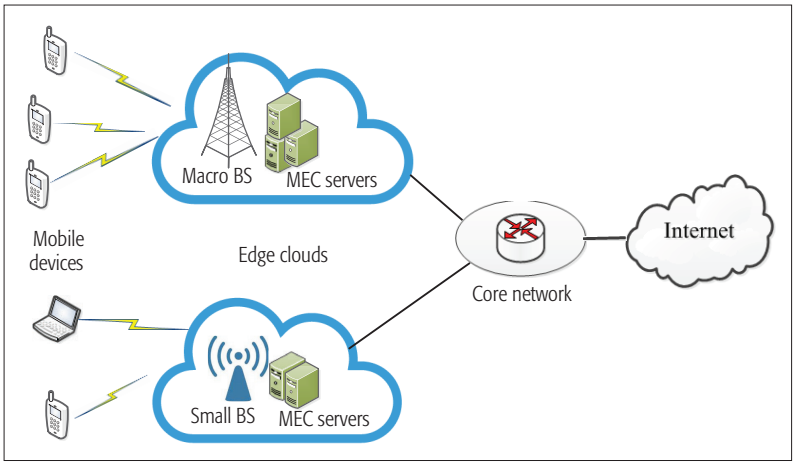

FIGURE 1. Illustration of the architecture of MEC.

Key Enablers of MEC

The implementation of MEC depends on several technological advancements:

- *Cloud technology*: Due to the advancement of cloud computing technology, multiple applications can be run simultaneously on an MEC server in an efficient way.
- *Network technology*: Improvements in network technology can provide a higher data rate and network context awareness, which would enhance users' experience.
- *Smarter mobile device*: Due to the pervasiveness of smarter mobile devices, a peer-to-peer mobile cloud within the edge network can be formed by multiple mobile devices within a vicinity. With device-to-device communications, energy and network efficiency can be improved.

Computation Offloading

As one of the main use cases of MEC, computation offloading can save the mobile device's computation energy consumption at the cost of additional transmission energy consumption. From the perspective of a mobile device user, it needs to decide whether it is beneficial to offload the computation to the MEC server. Different approaches have been proposed under various scenarios, such as vehicular networks, the single-user case [7], and the multi-user competition case [5, 6, 9, 10]. In this article, we consider the multi-user computation offloading decision problem from a behavioral perspective.

System Model

We consider a network consisting of a set of $\mathcal{N} = \{1, 2, ..., N\}$ mobile device users collocated within the coverage of an edge cloud. The mobile device users connect with the MEC servers in the edge cloud through a wireless base station. Each user in $\mathcal{N}$ needs to complete a task that can-

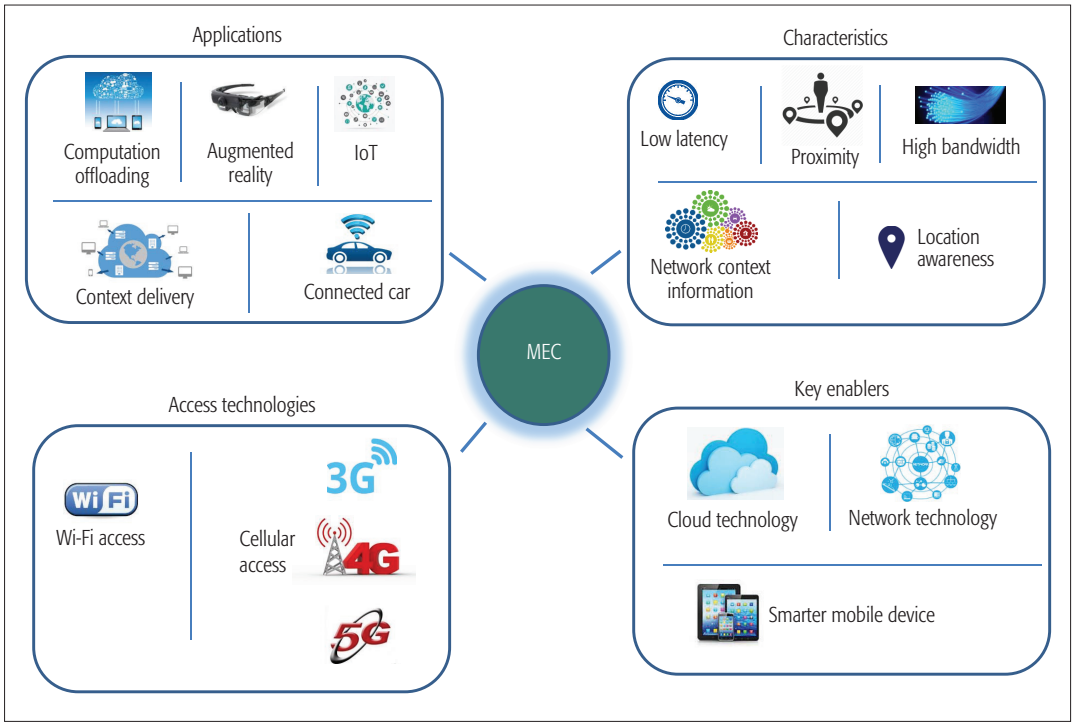

FIGURE 2. Taxonomy of MEC.

Since the visitor may walk around frequently, the presented content needs to be updated quickly. If the application is hosted on the remote cloud, the visitor may receive out-of-date information. However, by bringing the application to the nearby MEC server, the latency will be dramatically reduced, and thus the visitor's experience will be improved.

not be further partitioned. There are two options for each user: execute its task locally on its own mobile device or offload it to the MEC server.

Let $a_n \in \{0, 1\}$denote user n's decision, with $a_n = 1$ denoting that user n chooses to offload the computations and $a_n = 0$ denoting that user n chooses local computing. To obtain insightful results, we adopt a commonly used assumption in mobile cloud computing: that the locations of mobile device users are unchanged during a computation offloading period [5–10, 13].

Communication model: For the wireless communication resources, a round-robin policy is adopted to schedule the data transmission among multiple offloading users [5, 6]. As the number of offloading users increases, the expected throughput of each user will decrease. If user n chooses to offload and is scheduled for transmission, there are two possible outcomes: a successful computation offloading with probability $p_n \in [0, 1]$ and a computation offloading failure with probability $1 - p_n$. Note that probability p_n is used to model the random wireless channel condition between user n and the base station. If computation offloading failure occurs, user n will switch to local computing and incur an additional penalty, d_n.

Computation model: If user n executes its task locally, the time and energy consumption are not dependent on others' choices. If user n offloads the computation to the MEC server, the transmission time and energy increase as more users choose to offload. In this case, the total energy consumption is mainly contributed by the transmission energy, and the total time delay is the sum of transmission time and the task processing time, which is determined by its subscribed computing resources on the MEC server.

For user n, let c_n^l denote the total overhead of local computing in terms of energy consumption and processing time. Let $c_n^e(\boldsymbol{a})$ denote the total overhead of remote computing on the MEC server, where $\boldsymbol{a} = (a_1, ..., a_n)$ is the decision profile of all the users. Based on the communication and computation models, $c_n^e(\boldsymbol{a})$ can be written as a linear function with respect to the total number of offloading users [6]. Due to the randomness of the wireless channel, if user n chooses to offload its computation, it will occur overhead $c_n^e(\boldsymbol{a})$ with probability p_n and overhead $c_n^l + d_n$ with probability $1 - p_n$.

We define the utility of user n as the overhead reduction with respect to local computing. Therefore, if user n chooses local computing, it obtains zero utility; if user n chooses computation offloading, it achieves utility $c_n^l - c_n^e(\boldsymbol{a})$ with probability p_n and negative utility $-d_n$ with probability $1 - p_n$.

PT-Based Computation Offloading Game

The Framework of Prospect Theory

When faced with uncertainty, it is observed that people may not behave in a rational manner established by the expected utility theory (EUT) model [14]. To account for the realistic irrational behavior, PT provides a viable model for the scenario with uncertainty [15]. Next, we introduce two key notions in PT.

Weighting Effect: When making a decision in real life, people are inclined to use their subjective probability to weigh the values of possible outcomes instead of the objective probability p. In particular, people usually underweigh high

probability events and overweigh low probability events. The weighting effect can be illustrated by a probability weighting function $w(p)$. The commonly used one is known as the Prelec function [15], which is characterized by a probability distortion parameter α. This parameter indicates how the objective probability p is distorted by an individual's subjective evaluation. A smaller α corresponds to a larger probability distortion and deviation from rationality.

Framing Effect: In PT, people change the ways to compute their utilities. The objective utility is usually framed by each individual's subjective perception. Subjective value function $v(\cdot)$ is commonly adopted to capture this framing effect. One typical subjective value function is the Tversky value function [15], where the framing effect is controlled by two parameters: risk aversion parameter β and loss penalty parameter γ. A smaller β indicates a larger impact of framing effect, and a larger γ means more sensitivity to loss.

Table 1 shows a comparison between EUT and PT. Obviously, if the method of computing utility is changed, the whole decision making process will be different from the traditional thought established by EUT.

Model	Assumption	Weighting probability	Optimization function
Expected utility theory	Rational behavior under uncertainty	Objective probability	Objective expected utility
Prospect theory	Realistic irrational behavior under uncertainty	Weighted subjective probability	Framed utility relative to a reference point

TABLE 1. Comparison between expected utility theory and prospect theory.

PT-Based Computation Offloading Game

Due to the random wireless communication channel, each user obtains a probabilistic utility if it chooses to offload. Moreover, with limited wireless communication resources, the computation offloading decisions among the mobile device users are coupled. When the number of offloading users increases, the wireless communication resources for each offloading user will decrease, resulting in longer transmission time and higher energy consumption. To capture both the uncertain channel condition and coupling among multiple users' decisions, we formulate the computation offloading decision problem among multiple subjective users as a non-cooperative game and apply the framework of PT to provide a user-centric view of computation offloading decision.

Based on the weighting and framing effects in PT, if user n chooses to offload, its subjective utility $u_n^{PT}(\boldsymbol{a})$ under PT is the summation of $w(p_n)v(c_n^l - c_n^e(\boldsymbol{a}))$ and $w(1-p_n)v(-d_n)$. If user n chooses local computing, there is no uncertainty involved, and its subjective utility is the same as its objective utility. Without loss of generality, we use the Prelec function and Tversky value function as the probability distortion function and utility value function in our model, respectively. Based on the subjective utility, we formulate a non-cooperative computation offloading game Γ defined as follows:

- *The players* are the set of mobile device users $\mathcal{N}$.
- *The strategy* of each player is to choose whether to offload its computation or not.
- *The payoff* is the subjective utility u_n^{PT}, $\forall n \in \mathcal{N}$.

In this game, each user in $\mathcal{N}$ individually chooses whether to offload or not to maximize its payoff.

Equilibrium Analysis

Nash equilibrium is a widely adopted solution in non-cooperative game. At the Nash equilibrium, all the players adopt the best response strategy, and no player has the incentive to change its strategy if others' strategies remain fixed. For the computation offloading game, we show that the best response of user n is a threshold-based strategy given as follows: If the total number of offloading users is smaller than its threshold t_n, it is better to choose to offload the computation to the MEC server; otherwise, it is better to compute its task locally. The threshold t_n of user n can be computed by letting $u_n^{PT}(\boldsymbol{a})$ equal zero.

Note that the threshold t_n can indicate the probability of user n choosing to offload. With a higher threshold t_n, user n has higher probability of choosing to offload its computation. From the expression of t_n, we find that if the probability distortion parameter α decreases, the probability of user n choosing to offload decreases for the case with a larger successful computation offloading probability p_n, and increases for the case with a smaller p_n. With a smaller α, the decision of each user deviates from the EUT more severely. Even if its channel is in good condition with higher probability p_n, the user may choose local computing due to overweighing the small probability of offloading failure. Moreover, the probability of choosing to offload decreases as the loss penalty parameter λ increases and the risk aversion parameter β decreases. Therefore, we find that there would be less offloading users due to the behavioral impact.

Since each user's best response depends on its threshold, we have the following observation: for a given user, if it is better off choosing to offload, all the other users with higher thresholds are also better off choosing to offload; if it is better off choosing local computing, all the other users with smaller thresholds are also better off choosing local computing. Based on this observation, we next propose an efficient algorithm to compute the Nash equilibrium of the computation offloading game Γ.

Distributed Algorithm for Computation Offloading

Before making a computation offloading decision, each user sends the value of threshold t_n ($\forall\ n \in \mathcal{N}$) to the MEC server. Based on the received thresholds, the MEC server orders the set $\mathcal{N}$ of mobile users such that $t_1 \le t_2 \le \ldots \le t_n$, and assigns each user a unique ID (e.g., 1, 2, ..., N) based on the order. After being notified of their IDs, the users begin to perform distributed computation offloading decision. Specifically, each user sets its initial decision to be local computing. Given other users' strategies, user 1 is the first to make its best response strategy and broadcasts its updated strategy. Then user 2 begins to update its best response strategy based on user 1's updated strategy and the remaining non-updated users' strategies. Based on the strategy updating order from 1 to n, the updating process continues until user n has updated its strategy. At the end of the

By making decisions based on the subjective utilities, more users choose local computing, although computation offloading offers a better performance in expectation. Thus, the subjective perception in human decision making plays an important role in practical computation offloading mechanism design.

Initialization:
Each user n chooses the initial decision to be $a_n = 0$, $\forall\, n \in \mathcal{N}$.

ID assignment:
The MEC server sorts the users' thresholds t_ns in a decreasing order and assigns each user with a unique ID based on the order.

Computation offloading decision update:
for users with ID from 1 to N **do**
  if the number of offloading users is smaller than its threshold t_n
    update $a_n = 1$.
  else
    stick to the initial decision $a_n = 0$.
  end if
  broadcast its updated decision to others.
end for

ALGORITHM 1. Algorithm for computing the Nash equilibrium.

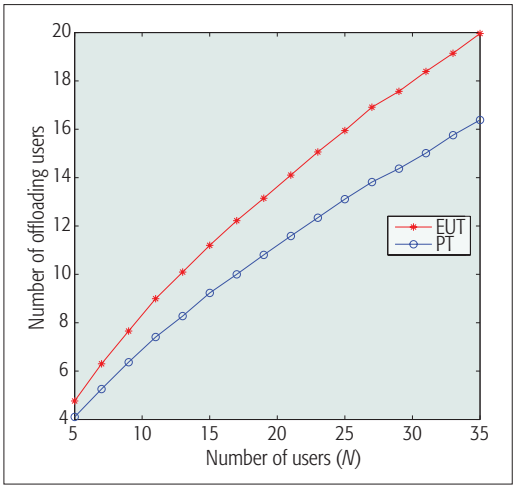

FIGURE 3. Average number of offloading users with different n.

strategy updating process, no user has the incentive to change its strategy, and Nash equilibrium is achieved. The algorithm is summarized in Algorithm 1.

Algorithm 1 has a low computational complexity of $\mathcal{O}(N \log N)$, which is contributed by the process of threshold ordering and ID assignment. Note that Algorithm 1 has n iterations, and the user with ID n updates its strategy at iteration n. When the last user with ID n finishes its updating, Nash equilibrium is achieved and Algorithm 1 terminates, since no user has the incentive to deviate after n iterations.

Numerical Results

In this section, we present numerical results to illustrate the impact of subjective parameters (α, β, and γ) on the users' offloading decisions. We consider the setting that an edge cloud is centered at a 100 m × 100 m square area, and the n mobile device users are randomly located over the area. In the simulation, the task of each user is characterized by the input data size 420 kB and the number of needed instructions 1000×10^6, and the computation resource is characterized by the number of CPU cycles per second. For each user, the local computation resource is uniformly distributed between 100 and 200, and the subscribed computation resource at the MEC server is uniformly distributed between 800 and 1000. Each user uses the same transmit power with 1.5 W, and the local computing power is uniformly distributed between 0.5 and 1 W.

Figure 3 shows the average number of offloading users with different n under both the EUT and PT models. Here we set $\alpha = 0.5$, $\beta = 0.8$, and $\gamma = 3$. The successful computation offloading probability p_n is uniformly generated in the interval [0.7, 1]. We can see that the number of offloading users is smaller for PT than EUT. With subjective perception, PT users overweight the possibility of computation offloading failure and are more willing to choose local computing.

Impact of weighting effect: Figure 4a illustrates the impact of probability distortion parameter α on a computation offloading decision. With larger successful computation offloading probability p_n, the number of offloading users increases under both the EUT and PT models. In Fig. 4a, we observe that the number of offloading users increases as the probability distortion parameter α increases under PT. For the case with a larger α, each user evaluates the successful computation offloading probability in a less subjective way and is more apt to offload their computations. For the case with a smaller α, corresponding to a larger deviation from rationality, users are more unwilling to offload their computations.

Impact of framing effect: Figure 4b illustrates the impact of loss penalty parameter γ and risk aversion parameter β on a computation offloading decision. As γ increases, the number of offloading users decreases, since the perceived loss caused by offloading failure is larger and users are less willing to choose computation offloading. We also observe that a smaller β results in a smaller number of offloading users, due to the fact that users are more risk averse in the computation offloading gain. Although computation offloading offers better performance, users are more willing to choose local computing.

From Figs. 3 and 4, we are able to see that by considering the weighting and framing effects in the PT model, the results of computation offloading can significantly change compared to the usually adopted EUT model. By making decisions based on the subjective utilities, more users choose local computing, although computation offloading offers better performance in expectation. Thus, the subjective perception in human decision making plays an important role in practical computation offloading mechanism design.

Conclusion

In this article, we study the multi-user computation offloading problem in MEC from a behavioral perspective. We apply the framework of prospect theory to model mobile device users' realistic behavior when making the computation offloading decision. We cast the users' decision making

of whether to offload or not as a PT-based non-cooperative game and propose a distributed computation offloading algorithm to achieve the Nash equilibrium. Numerical results analyze the impact of behavioral biases on users' decision making, and show that the number of offloading users becomes smaller under the PT model than the classical EUT model. This article could motivate a deeper investigation of the role of subjective human perception in future work. For instance, for the MEC service provider, how to design an optimal price policy considering the irrational and subjective user behavior remains a challenging problem.

References

[1] S. Abolfazli *et al.*, "Cloud-based Augmentation for Mobile Devices: Motivation, Taxonomies, and Open Challenges," *IEEE Commun. Surveys & Tutorials*, vol. 16, no. 1, Jan. 2014, pp. 337–68.

[2] Z. M. Fadlullah *et al.*, "State-of-the-Art Deep Learning: Evolving Machine Intelligence Toward Tomorrow's Intelligent Network Traffic Control Systems," *IEEE Commun. Surveys & Tutorials*, May 2017., DOI: 10.1109/COMST.2017.2707140.

[3] J. Ren *et al.*, "Serving at the Edge: A Scalable IoT Architecture Based on Transparent Computing," *IEEE Network*, vol. 31, no. 5, Aug. 2017, pp. 96–105.

[4] Y. Hu *et al.*, "Mobile Edge Computing: A Key Technology toward 5G", ETSI White Paper, Sept. 2015.

[5] X. Chen *et al.*, "Efficient Multi-User Computation Offloading for Mobile-Edge Cloud Computing," *IEEE Trans. Net.*, vol. 24, no. 5, Oct. 2016, pp. 2795–2808.

[6] L. Tang, X. Chen, and S. He, "When Social Network Meets Mobile Cloud: A Social Group Utility Approach for Optimizing Computation Offloading in Cloudlet," *IEEE Access*, vol. 4, Sept. 2016, pp. 5868–79.

[7] Y. Li, M. L, and Y. Zheng, "Adaptive Multi-Resource Allocation for Cloudlet-Based Mobile Cloud Computing System," *IEEE Trans. Mobile Computing*, vol. 15, no. 10, Oct. 2016, pp. 2398–2410.

[8] M. V. Barbera *et al.*, "To Offload or Not to Offload? The Bandwidth and Energy Costs of Mobile Cloud Computing," *IEEE INFOCOM*, July. 2013.

[9] Y. Mao *et al.*, "Power-Delay Tradeoff in Multi-User Mobile-Edge Computing Systems," *IEEE GLOBECOM*, Dec. 2016.

[10] S. Sardellitti *et al.*, "Joint Optimization of Radio and Computational Resources for Multicell Mobile-Edge Computing," *IEEE Trans. Signal and Info. Processing over Networks*, vol. 1, no. 2, June 2015, pp. 89–104.

[11] N. Barberis, M. Huang and T. Santos, "Prospect Theory and Asset Prices," *Quarterly J. Economics*, vol. 116, no. 1, Feb. 2001, pp. 1–53.

[12] Y. Zhang *et al.*, "A Survey on Emerging Computing Paradigms for Big Data," *Chinese J. Electronics*, vol. 26, no. 1, Jan. 2017, pp. 1–12.

[13] L. Tang *et al.*, "Double-Sided Bidding Mechanism for Resource Sharing in Mobile Cloud," *IEEE Trans. Vehic. Tech.*, vol. 66, no. 2, Feb. 2017, pp. 1798–1809.

[14] D. Schmeidler, "Subjective Probability and Expected Utility Without Additivity," *Econometrica*, vol. 57, no. 3, May 1989, pp. 571–87.

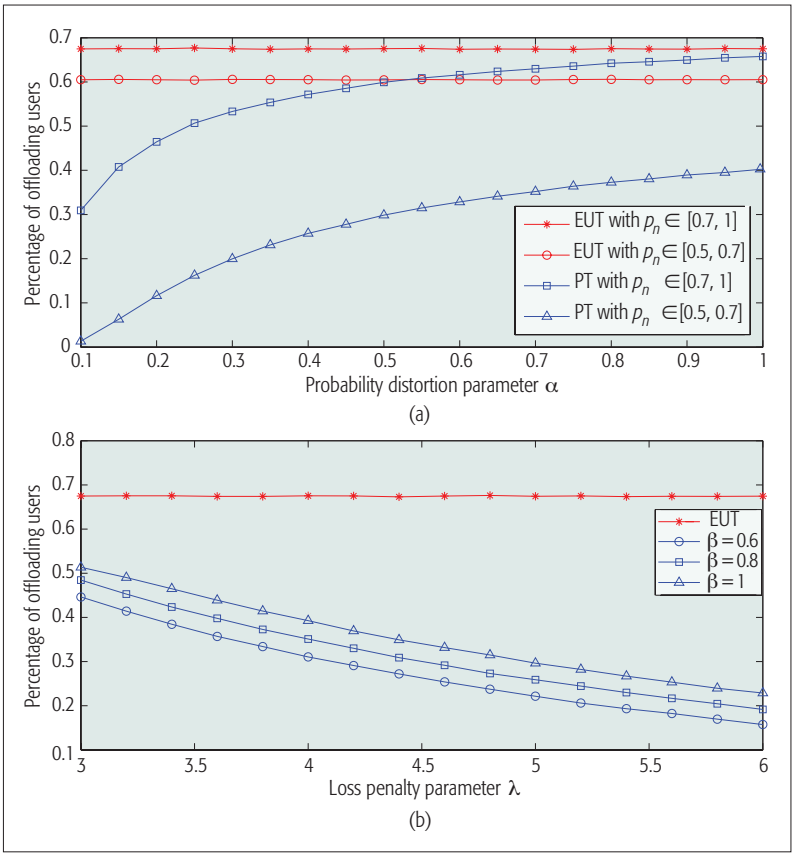

FIGURE 4. Impact of weighting and framing effects: a) impact of weighting effect; b) impact of training effect.

[15] D. Kahmeman and A. Tversky, "Prospect Theory: An Analysis of Decision Under Risk," *Econometrica*, vol. 47, no. 2, Mar. 1979, pp. 263–91.

Biographies

Ling Tang (ling.tang@njust.edu.cn) is currently an associate professor with the School of Computer Science and Engineering at Nanjing University of Science and Technology. She received her Ph.D. degree in electronic science and engineering from Southeast University, Nanjing, China, in 2013. Her current research interests include mobile cloud computing, cooperative communications, and physical layer security.

Shibo He [M'13] (s18he@iipc.zju.edu.cn) is currently a professor at Zhejiang University, Hangzhou, China. He was an associate research scientist from March 2014 to May 2014, and a post-doctoral scholar from May 2012 to February 2014, with Arizona State University, Tempe. He received his Ph.D. degree in control science and engineering from Zhejiang University in 2012. His research interests include wireless sensor networks, crowdsensing, and big data analysis.