Toward data-centric and trustworthy Internet of Things

Tony T. Luo

Institute for Infocomm Research, A*STAR, Singapore
E-mail: luot@i2r.a-star.edu.sg

Technical Report, November 2018

I view my recent research as a journey toward data-centric and trustworthy Internet of Things (IoT), or less formally, "small data, big data, and good data" for IoT. My future research will be steered toward intelligent IoT empowered by artificial intelligence, with a particular focus on security and privacy.

1 Recent Research

By "small data, big data", I refer to the *quantity* aspect, i.e., how to acquire a large volume of IoT data in small amounts. By "good data", I refer to the *quality* aspect, i.e., how to improve IoT data quality under different circumstances. The primary problem context for the quantity aspect is *participatory IoT*, which is a grassroots approach to IoT where sensing devices are individually owned by mobile users rather than collectively deployed by organizations as in the conventional approach. On the other hand, the problem context for the quality issue includes both participatory and convention (i.e., sensor network based) IoT as well as others (e.g., multi-agent IoT systems), as quality is a common issue in all these circumstances.

The emergence of participatory IoT was stimulated by the flourishing manufacturing industry of personal mobile devices. such as smartphones, wearables, cars, and even drones, which are equipped with an unprecedentedly rich set of sensors. By leveraging these existing versatile sensory devices, as compared to the conventional way of purchasing and deploying specialized sensors, we can benefit from several appealing advantages: remarkably lower cost, pervasive geographic coverage, rapid entry to market, and little maintenance hassle, among others. However, crowd participation is essential to this new sensing paradigm which, therefore, is also known as *mobile crowd sensing* (MCS). As such, it involves two fundamental challenges:

- *Incentive*: Why would the crowd bother to use their own devices—and inevitably incur various costs—to contribute sensor data to the data collection organization whose purpose may not be aligned with their interests?

- *Trust*: In comparison to deployed dedicated sensors, personal mobile sensing devices are much more error-prone and vulnerable to poor measurement conditions and erratic or even malicious human behaviors. Hence, how to ensure reliable and trustworthy data quality?

1.1 Incentivizing participatory IoT: Auctions and Tullock contests

The core idea to address the incentive issue is to introduce some sort of *competition* among participants. To this end, *mechanism design* [1] is a superb theoretical tool which originates from economics and describes how to design a system in which rational agents will behave in such a way that they collectively achieve a system-desired equilibrium. For this reason, the theory of mechanism design is also known as the *reverse game theory*.

I started by using the most sophisticated mechanism framework, *auction* [2], to design incentive mechanisms for participatory IoT. The objective was to motivate participants to exert their highest possible effort to contribute the most amount of data, governed by two rules that we as a designer needs to specify: an allocation rule specifies how to allocate reward to participants according to their respective *bids* (intended efforts), and a payment rule specifies how much effort each participant should actually pay. Unlike prior work, which all adopted the *winner-pay auction* (WPA) framework since it is the mainstream and a well-studied mechanism, I embarked on using a different framework called *all-pay auctions* (APA) [3, 4]. In a WPA, only winners (who outbid others and will therefore receive reward) need to pay their bids (by making actual effort to contribute sensor data). In an APA, however, all the participants need to pay their bids regardless of who win the auction. Apparently, APA does not sound intuitive, but when applied to the context of participatory IoT, it makes sense in that participants bid by making actual contributions (instead of making a "wish"). In fact, this gains APA two important advantages over WPA as identified by [3, 5]: (i) WPA requires a separate *bidding stage* while APA does not, making APA a better fit for the ad hoc nature of participatory IoT and hence easier to implement in practice; (ii) WPA suffers from a risk of *bid-nonfulfillment* which means that winners may not fulfill or only partially fulfill their bids as those are just "wishes", but APA is inherently immune to this risk.

Another contribution we made was to introduce an *adaptive prize* into APA as its reward, which is in contrast to all the classical auctions (including APA) where either the reward is fixed or the unit reward is fixed. The adaptive prize that we proposed is a nonlinear function of winner's effort and can elicit higher effort from all the participants. Notably, this function is non-parametric in the sense that it does not assume any *a priori* form (which is inevitably artificial) with unknown parameters to be determined later (e.g., via some sort of optimization); rather, it begins with a "free form" (imagine just $f(x)$ alone) and the eventual, concrete form is purely derived as a result of optimization. The optimization process gives the adaptive prize an competitive edge of outperforming conventional fixed prizes and even the *optimal* fixed prize; in fact, we showed that it yields the maximum possible *profit*—total data contribution minus the reward payout. Later on, we generalized the problem setting to a heterogeneous one [6, 7], in which the Bayesian belief (due to incomplete information) about each other agent's sensing cost or ability is different from agent to agent, which could be a closer model of the reality. By investigating this setting, we discovered an interesting property called *strategy autonomy* where the asymmetric equilibrium collapses into a symmetric one, making the originally complex problem much more tractable in practice.

In the literature of participatory IoT as well as many other related computer science fields, auctions have been a dominant mechanism widely adopted to design incentives, since they have been extensively studied and well understood for decades. However, taking a retrospect and deliberation at that juncture, I realized that auctions are not always the best fit. The main reason is that they are *perfectly discriminating*: any agent must outbid everyone else in order to win, and as a result, the winners are always the "strongest" bidders. This makes auctions a deterministic and highly competitive mechanism which rules out the possibility for weaker bidders to win, and thus is not conducive to large participation. Excellent evidence is that auctions in reality are almost always limited to a small group of "strong" (e.g., wealthy) participants who bid for high-value goods (e.g., antiques, diamonds, and blue-fin tunas).

This observation motivated me to find an alternative mechanism that better suits the grassroots nature of participatory IoT. My exploration eventually led me to an (apparently) less known mechanism, *Tullock contests*, which are *imperfectly discriminating*: exerting higher effort only increases one's *winning odds* against others but does not guarantee winning; in other words, the winner selection is no longer deterministic but probabilistic, or informally put, "every one has a chance to win, no matter how weak you are".

This makes Tullock contests very appealing to the vast majority of grassroots users who are relatively "weak", in terms of the amount of IoT data they can contribute. Again, excellent evidence exists in reality where the most common and simplest form of Tullock contests, *lottery*, is widely practiced all over the world and the participation is often enormous, at the national scale. (This is why in the above I refer to Tullock contests as "apparently" less known.)

Hence in [8], we introduced Tullock contests as an alternative to auctions to design incentive mechanisms for participatory IoT, and we showed that it achieves superior performance in terms of several metrics such as profit and social welfare.

As a caveat, this is not to claim that auctions are no longer an option. While Tullock contests tend to foster larger participation, auctions elicit higher *per capita* contribution due to its competitiveness. As a result, there is no clear winner in terms of the *total* contribution, which is also theoretically backed up by the *revenue equivalence theorem.* Hence, the best choice of mechanism depends on the goal of each specific application [5]: auctions are more suitable for applications that favor individual quality over collective diversity, such as knowledge- or skill-intensive crowdsourcing (e.g., Kaggle) and sports contests (e.g., Olympic Games and Mr. Olympia competition), while Tullock contests are more suitable for those that favor diversity over individual quality, such as micro-task crowdsourcing (e.g., surveys and MCS with easy sensing tasks).[1]

1.2 Trustworthy IoT: Reputation systems and Machine learning

This line of research toward "good data" covers a broader range of areas including participatory and conventional IoT as well as multi-agent IoT systems, as aforementioned. As such, I also took different approaches to cater for different circumstances.

One of my first attempts was to combine both social and economic elements to reinforce trustworthiness. Specifically, I tried to address a common limitation of game-theoretic studies, which is that all agents are self-interested, or *egoistic* as we call, to maximize their own respective utilities. My proposition was that, in reality, people are multi-faceted and can be *altruistic* at times, especially when it comes to special social circles such as their family, relatives, and close friends. Therefore, by trading off between egoism and altruism, we introduced a notion of *nepotism* and thereby proposed a new relationship called *endorsement* to connect participants into a socioeconomic network called simple endorsement web (SEW) [9]. Endorsement is a socioeconomic relationship whereby participants endorse each other either because of their special social circle (nepotism), or because of economic reasons (to tighten the possibly loose social ties). For example, Alice endorses Bob either because Bob is her boyfriend, or because Bob is a good data contributor, and this endorsement relationship results in two consequences. First, Bob will have his reputation increased due to Alice's trust, which will lead to higher reward for Bob's future contributions. Second and more importantly, Alice will receive a *share* of Bob's reward ("dividend" as we call) whenever Bob makes a good contribution. This way, Bob gets an extra and unique motivation of "working (contributing) for people whom I care about", and tends to refrain from contributing fake or malicious data because he now has both a "witness" and "beneficiary" (who may even be socially close to him). To make this idea rigorous, we formulated the economic aspect of this problem under a Stackelberg game framework, and derived the optimal contribution strategy and endorsement strategy for each participant.

I led a developer team to implement an adapted version of [9] in two software systems, FoodPriceSG (`http://foodprice.sns-i2r.org`) and imReporter (`http://imreporter.sns-i2r.org`), which are publicly available for free download at `Apple Store` and `Google Play`. Another system we developed is WiFi-Scout (`http://wifi-scout.sns-i2r.org`) which is available for free download too, and is also summarized in [10].

Based on [9], we incorporated a machine learning technique, *collaborative filtering* or more specifically *matrix factorization*, to further improve the trustworthiness of participatory IoT systems [11]. The idea was to predict, for a given sensing task, the competency of participants and the resultant sensing data quality, based on task features, participants' social attributes, and notably, their endorsement relationship with others (e.g., one who is endorsed by many doctors is likely to perform well in medical sensing tasks). This type of predictions would enable more relevant task assignment and recommendation, thereby producing more trustworthy data.

In reality, verifying data quality is often hard and entails tedious manual verification. To automate this process, we exploited the "power of crowds" [12] to a fuller extent by proposing a cross validation approach [13, 14] to not only verify but also improve IoT data quality. Specifically, the approach seeks a *validating crowd* to ratify the *contributing crowd* on the sensor data contributed by the latter,

[1] Technically speaking, auctions are a special case of Tullock contests by taking the exponent parameter in a *contest success function* to the extreme (infinity). But generally, they are considered as two different mechanism frameworks.

and uses the validation result to reshape the original sensor data into a more credible posterior belief of the ground truth. In particular, we designed a privacy-aware trust-oriented probabilistic push (PATOP^2) algorithm [13] based on the *exploration-exploitation* principle [15] and stochastic optimization, to accomplish cross validation in a timely manner with guaranteed success rate. In view of its close interaction with human, the algorithm is also robust to common security attacks such as *collusion* and *Sybil attacks*. Moreover, we also incorporated a weighted random oversampling (WRoS) technique to enable *discovery of hidden truth*, which was not possible in all prior work based on statistical methods.

Recently, the sharp rise of deep learning motivated me to leverage this powerful instrument to build a more trustworthy IoT. We set out to address a main challenge, *anomaly detection*, in the conventional IoT context of wireless sensor networks (WSN) [16]. To the best of our knowledge, we are the first to introduce *autoencoder neural networks* (ANN) into WSN to tackle this challenge. One particular innovation in [16] is that we disprove the common belief that deep learning is not suitable for WSN due to the mismatch between resource demand and supply, by constructing a WSN-applicable ANN. The core idea was to *make deep neural networks "shallow"*, by constructing an ANN of depth two only (i.e., one hidden layer) without sacrificing performance (detection accuracy in this case). We achieved this by designing a two-part algorithm that allocates computational tasks to sensors and IoT cloud based on disparate resource requirements, as well as by exploiting the *reconstructability* of ANN. Besides achieving high detection accuracy, the *unsupervised learning* feature of our proposed approach also overcomes a crucial challenge faced by most machine learning based anomaly detection methods, which is the shortage of (labeled) training data as anomalies are scarce by definition.

We also applied deep learning to large-scale multi-agent IoT systems where nodes are more autonomous. In such systems, agents interact more frequently with each other to perform transactions or consume services. However, such transactions break down when there is lack of trust among agents, thus making trust management critical. To this end, we proposed DELTA [17], a deep learning based trust assessment mechanism, to allow each agent to compute an accurate trustworthiness score for other agents of its interest, even if it has never interacted with those other agents. DELTA achieves high predictive accuracy by (i) converting traditional second-hand evidence into our proposed *conjectured evidence* which incorporates context information, and (ii) aggregating the conjectured evidence with the agent's first-hand evidence using a deep neural network (with carefully chosen activation and loss functions as well as its depth and width). Besides accuracy, DELTA is also robust to malicious and fake second-hand evidence even when such evidence dominates; for example, it is resistant to the *51-percent attack* under common attack schemes such as ballot-stuffing and bad-mouthing.

1.3 Other Research Work

Some of my other research not elaborated herein includes, to name a few, data privacy [18, 19], device-to-device networks [20], market-based approach to data quality analysis [21], and mobile edge computing [22].

2 Future research

My future research will be steered toward intelligent IoT empowered by artificial intelligence (AI), with a particular focus on security and privacy.

The first thread of research I plan to do is intelligent manufacturing in *industrial IoT*. I intend to apply statistical and machine learning methods to production lines and make real-time predictions about performance, reliability and safety, and thereby decide whether the production lines are conformant to corresponding industry standards and how likely they will become non-conformant in a future period. I will be analyzing both historical and streamed data, including time series data as well. Some interesting research issues and the main challenges include the stringent security and privacy requirement imposed by manufacturers, highly complex and heterogeneous factory environments, as well as the evolving industry standards.

The second line of research I plan to do is AI-empowered IoT security. Security used to be the "last mile" to build, or "something good to have", for many systems. However, my past six years' industry experience tells me that such a mindset and practices simply result in fragile systems that no longer sustain. For large-scale and open IoT systems, the consequence of security breach is much severer and hence such systems are more likely to be the target of malicious attackers. Given the high complexity and dynamics of IoT systems, I am interested to take an AI-based approach and focus on using different learning techniques to discover and capture the elusive patterns veiled under various and constantly changing phenomena. One possible starting point for me is security games [23] which is one of the key efficacious approaches in the general area of security. However, the idealistic game-theoretical assumptions thereof can restrict practical application, and the new challenges brought forth by the heterogeneous IoT environments also call for new innovations. Therefore, I would inject machine learning techniques into security games with idealistic assumptions altered, to cope with the new challenges.

My third path of research is about IoT privacy. IoT requires a plethora of data, for example via sharing, in order to offer compelling services and produce effective results. This puts both personal and organizational privacy at stake, and several recent serious incidences about privacy leaks have corroborated this and widely aroused public awareness and concern. Besides policy control such as GPDR in the Europe, technological measures must also be put in place. One of the major technical issues is the longstanding trade-off between privacy and utility, and this becomes even more challenging in the context of IoT, because the openness of IoT inherently calls for massive collaboration and data sharing among millions of heterogeneous IoT nodes. One interesting project toward addressing the IoT privacy tension is IoT databox [24], which takes a hardware approach to build a physical device that acts as a gateway to an individual's or a household's IoT data. This could be one of my starting points, yet I will explore possibilities of minimizing or even eliminating the necessity of having a hardware component since it may present a barrier to practical adoption.

References

[1] L. Hurwicz and S. Reiter, *Designing Economic Mechanisms.* Cambridge University Press, 2006.

[2] V. Krishna, *Auction theory*, 2nd ed. New York: Academic Press, 2009.

[3] T. Luo, S. K. Das, H.-P. Tan, and L. Xia, "Incentive mechanism design for crowdsourcing: An all-pay auction approach," *ACM Transactions on Intelligent Systems and Technology*, vol. 7, no. 3, pp. 35:1–26, April 2016.

[4] T. Luo, H.-P. Tan, and L. Xia, "Profit-maximizing incentive for participatory sensing," in *IEEE INFOCOM*, April 2014, pp. 127–135.

[5] T. Luo, S. S. Kanhere, J. Huang, S. K. Das, and F. Wu, "Sustainable incentives for mobile crowdsensing: Auctions, lotteries, and trust and reputation systems," *IEEE Communications Magazine*, vol. 55, no. 3, pp. 68–74, March 2017.

[6] T. Luo, S. S. Kanhere, S. K. Das, and H.-P. Tan, "Incentive mechanism design for heterogeneous crowdsourcing using all-pay contests," *IEEE Transactions on Mobile Computing*, vol. 15, no. 9, pp. 2234–2246, September 2016.

[7] T. Luo, S. S. Kanhere, S. K. Das, and H.-P. Tan, "Optimal prizes for all-pay contests in heterogeneous crowdsourcing," in *IEEE International Conference on Mobile Ad Hoc and Sensor Systems (MASS)*, 2014, pp. 136–144.

[8] T. Luo, S. S. Kanhere, H.-P. Tan, F. Wu, and H. Wu, "Crowdsourcing with Tullock contests: A new perspective," in *IEEE INFOCOM*, April 2015, pp. 2515–2523.

[9] T. Luo, S. S. Kanhere, and H.-P. Tan, "SEW-ing a simple endorsement web to incentivize trustworthy participatory sensing," in *IEEE International Conference on Sensing, Communication, and Networking (SECON)*, July 2014, pp. 636–644.

[10] F.-J. Wu and T. Luo, "WiFiScout: A crowdsensing WiFi advisory system with gamification-based incentive," in *IEEE International Conference on Mobile Ad-hoc and Sensor Systems (MASS)*, 2014.

[11] C. Wu, T. Luo, F. Wu, and G. Chen, "EndorTrust: An endorsement-based reputation system for trustworthy and heterogeneous crowdsourcing," in *IEEE Globecom*, December 2015.

[12] P. Michelucci and J. L. Dickinson, "The power of crowds," *Science*, vol. 351, no. 6268, pp. 32–33, 2015.

[13] T. Luo, J. Huang, S. S. Kanhere, J. Zhang, and S. K. Das, "Improving IoT data quality in mobile crowdsensing: A cross validation approach," *IEEE Internet of Things Journal*, under review.

[14] T. Luo and L. Zeynalvand, "Reshaping mobile crowd sensing using cross validation to improve data credibility," in *IEEE Globecom*, December 2017.

[15] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction.* Cambridge: MIT Press, 1998.

[16] T. Luo and S. G. Nagarajan, "Distributed anomaly detection using autoencoder neural networks in WSN for IoT," in *IEEE International Conference on Communications (ICC)*, May 2018.

[17] L. Zeynalvand, T. Luo, and J. Zhang, "DELTA: Deep learning based trust assessment using conjectured evidence in multi-agent Internet of Things," in *International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, 2019, under review.

[18] F.-J. Wu and T. Luo, "CrowdPrivacy: Publish more useful data for less privacy exposure in crowdsourced location-based services," *ACM Transactions on Privacy and Security*, under review.

[19] F.-J. Wu, M. R. Brust, Y.-A. Chen, and T. Luo, "The privacy exposure problem in mobile location-based services," in *IEEE Globecom*, December 2016.

[20] Y. Han, T. Luo, H. Wu, and D. Li, "Competition-based participant recruitment for delay-sensitive crowdsourcing applications in D2D networks," *IEEE Transactions on Mobile Computing*, vol. 15, no. 12, pp. 2987–2999, December 2016.

[21] C. K. Tham and T. Luo, "Quality of contributed service and market equilibrium for participatory sensing," *IEEE Transactions on Mobile Computing*, vol. 14, no. 4, pp. 829–842, April 2015.

[22] M.-V. Ngo, H.-T. Hoang, T. Luo, and T. Q. Quek, "Orchestrated container migration and handover in edge computing for mobile users," in *IEEE ICC*, 2019, under review.

[23] M. Tambe, *Security and Game Theory: Algorithms, Deployed Systems, Lessons Learned.* New York, NY, USA: Cambridge University Press, 2011.

[24] http://iotdatabox.com, The University of Nottingham, accessed November 2018.