Nationwide Deployment and Operation of a Virtual Arrival Detection System in the Wild

Yi Ding[1,2], Yu Yang[3], Wenchao Jiang[4], Yunhuai Liu[5], Tian He[1,2], Desheng Zhang[3]
[1]Alibaba Group, [2]University of Minnesota, [3]Rutgers University,
[4]Singapore University of Technology and Design, [5]Peking University

ABSTRACT

We report a 30-month nationwide deployment and operation study of an indoor arrival detection system based on Bluetooth Low Energy called `VALID` in 364 Chinese cities. `VALID` is pilot-studied, deployed, and operated in the wild to infer real-time indoor arrival status of couriers, and improve their status reporting behavior based on the detection. During its full nationwide operation (2018/12-2021/01), `VALID` consists of virtual devices at 3 million shops and restaurants, where 530,859 of them are in multi-story malls and markets to infer and influence 1 million couriers' behavior, and assist the scheduling of 3.9 billion orders for 186 million customers. Although indoor arrival detection is straightforward in controlled environments, the scale of our platform makes the cost prohibitively high. In this work, we explore to use merchants' smartphones under their consent as a virtual infrastructure to design, build, deploy, and operate `VALID` from in-lab conception to nationwide operation in three phases for 30 months. We consider metrics including system evolution, reliability, utility, participation, energy, privacy, monetary benefits, along with couriers' behavior changes. We share three lessons and their implications for similar wireless sensing or communication systems with large geospatial operations.

CCS CONCEPTS

• **Networks** → **Network measurement**;

KEYWORDS

Bluetooth sensing, nationwide deployment, operational system, arrival detection

ACM Reference Format:
Yi Ding, Yu Yang, Wenchao Jiang, Yunhuai Liu, Tian He, Desheng Zhang. 2021. Nationwide Deployment and Operation of a Virtual Arrival Detection System in the Wild. In *ACM SIGCOMM 2021 Conference (SIGCOMM '21), August 23–27, 2021, Virtual Event, USA.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3452296.3472911

1 INTRODUCTION

Nowadays, instant delivery is an emerging business for Gig Economy [29], where Gig workers deliver online orders (e.g., food) within a short time (e.g., 30 minutes) from merchants (e.g., restaurants) to customers. This business proliferates with the emergence of several delivery platforms worldwide, e.g., Prime Now [7], UberEats [52], Instacart [28], and DoorDash [18] in the U.S.; Deliveroo [16] in the U.K.; Meituan [36] and Eleme (i.e., Alibaba local service company, our platform) [19] in China. For the platform, it is essential to know couriers' real-time arrival status at merchants, which is used to (1) update order status in customer's APPs for better customer experience, (2) assign new orders to the most suitable couriers, and (3) train learning models to estimate the order's preparing and delivery time for future orders [62]. Although smartphone GPS can detect outdoor arrival, inferring couriers' indoor arrival is challenging due to the lack of low-cost and reliable infrastructures at an extremely large scale.

In this paper, as one of the largest instant delivery platforms with 83 million monthly active users in China, we report an in-depth study of 30-month nationwide deployment and operation for a system called `VALID` for V̲irtuA̲L̲ arrI̲val D̲etection. `VALID` is designed to provide nationwide indoor arrival detection with practical cost/performance tradeoffs. It is deployed to infer the indoor arrival status of 1 million professional couriers at 530,859 indoor merchants in 364 Chinese cities and few of them, if any, are under our control, i.e., in the "wild". *We share one-month data of* `VALID` *for future research*[1] [1]. Admittedly, in controlled environments, e.g., labs or museums, indoor arrival detection is not technically challenging. However, it is still an open question for nationwide in-the-wild detection.

In academia, the solutions are mainly based on Wi-Fi [11, 12, 31, 32, 43], LED fixtures [33, 47, 54, 57], acoustics [39, 41, 60], RFID [3, 53], and IMU [2]. In practice, however, they are inapplicable for nationwide deployment due to monetary cost, energy consumption, or data unavailability. For example, Wi-Fi-based solutions are widely studied but inapplicable due to data unavailability (i.e., only 51% of indoor merchants on our platform are covered by Wi-Fi signals, and Wi-Fi scanning is unavailable for common APPs on iOS devices [9]), monetary cost (i.e., fingerprinting in a dynamic environment), and energy consumption (i.e., continuous Wi-Fi scanning drains smartphones' battery quickly [8]). LED-based and RFID-based methods are limited due to high deployment costs. Acoustic-based solutions are inapplicable due to data unavailability (i.e., couriers need to make phone calls frequently). IMU-based solutions are also inapplicable due environment-specific calibration and unbearable phone energy consumption with high-rate sampling.

In industry, current solutions are mainly in four categories: manual reporting [51, 58], smartphone GPS [48], cameras [22], and dedicated deployments [17]. (1) For manual reporting, it suffers from unintentional or intentional human errors [58] (Fig. 2); (2) GPS is

SIGCOMM '21, August 23–27, 2021, Virtual Event, USA

ACM ISBN 978-1-4503-8383-7/21/08…$15.00
https://doi.org/10.1145/3452296.3472911

[1] https://tianchi.aliyun.com/dataset/dataDetail?dataId=103969

inaccurate in indoor environments (e.g., multi-story malls); they are essentially limited in our setting because commodity smartphone GPS only provides reliable two-dimensional outdoor locations, but our setting is the indoor merchants in multi-story malls with multi-level basements; (3) For cameras, it is difficult for third parties (our merchants) to provide their videos (e.g., surveillance video) due to the privacy concerns; (4) For dedicated deployment, a citywide physical Bluetooth Low Energy (BLE) beacon system was introduced in [17], but the costs are prohibitively high to scale it up for nationwide deployment.

To address the limitations of the above solutions, we explore two opportunities from two perspectives, i.e., hardware and software: (i) the high penetration of low-cost smartphones (hardware), (ii) the already-installed merchant APP (software) to build a *virtual beacon* system without dedicated devices but using merchants' smartphones under their consent (See Discussion section for ethics and privacy protection). In particular, for the hardware, new models of low-cost (e.g., less than $200), fully-functional (e.g., BLE) smartphones are brought to the market every year; for the software, the percentage of merchants using merchant APP (instead of PC) for order management has been significantly increasing, e.g., from 47% in 2018/08 to 85% in 2021/01 on our platform.

However, serving as a virtual beacon is an optional but not mandatory function for merchants. Thus, one unique challenge for VALID is the incentives for merchants to participate in, i.e., benefits vs. costs of allowing the platform to use their smartphones as virtual beacons. Based on our interview with representative merchants, we found that merchants have enough incentives to participate in VALID if we can keep their participation benefits high yet the cost low. In particular, the virtual beacons can help the couriers report arrival easier and accurately, so they help the platform deliver the orders faster for a better customer experience, thus ultimately benefiting merchants themselves.

These unique opportunities and challenges of using merchants' smartphones as virtual beacons call for an elegant balance of simplicity and complexity of our VALID design for nationwide low-cost indoor arrival detection. To evaluate VALID, we comprehensively explore various metrics, including cost metrics (e.g., energy consumption and privacy risks), performance metrics (e.g., reliability and utility), and couriers' behavior changes influenced by detection results from VALID. Based on these metrics, we report VALID's deployment and operation in three phases.

- **Phase I: 1-Month Feasibility Study (2018/08-09)**. We conduct this study in a controlled environment with 20 devices to emulate couriers and merchants for reliability testing under various parameter configurations, including transmission frequency and powers in two OS.
- **Phase II: 3-Month Citywide Testing (2018/09-12)**. We embed VALID in the merchants' and couriers' APP in Shanghai under their consent to mainly compare the merchants' smartphones as virtual beacons to the physical beacons we deployed in a citywide uncontrolled environment.
- **Phase III: 26-Month Nationwide Operation (2018/12-2021/01)**. We embed VALID nationwide, and VALID has been operational as of this submission. We utilize the accounting data to conduct a post-hoc analysis to evaluate VALID in retrospect because of lacking nationwide physical beacons as ground truth. During this 30 month VALID evolution as merchants enter and leave, we evaluate VALID with metrics (e.g., reliability with different hardware and courier behavior change) that cannot obtain in Phase II.

Based on our successes and failures in the deployment and operation, we discuss three lessons learned to provide insights for other similar systems and our future work VALID+.

Lesson learned 1: Evolution in the Wild. VALID suggests that a participatory less-expensive software-based "virtual" beacon system evolves more robustly (i.e., with a gradually increasing scale) even with *high* nationwide uncertainty (i.e., 364 cities), compared to a dedicated expensive hardware-based "physical" beacon system with *low* citywide uncertainty (i.e., Shanghai), as in Fig.7 (i). However, it is essential to provide incentives for users to participate in a virtual system (even with APPs they are using) by minimizing costs and showing benefits of participation, which can be potentially addressed by system design simplicity. (Details in Sec.6.1)

Lesson learned 2: Reliability in the Wild. Although the imperfect reliability of a physical beacon system in the wild has been discussed in [17], we found that a virtual beacon system suffers *more* due to the uncontrolled factors on the senders' side (e.g., iOS's restriction on background BLE advertising, merchants' participation and mobility). Given these factors, an *asymmetric* design philosophy has the potential to improve the reliability by increasing participation, such as a simplified design for the users who require strong incentives to accommodate device diversity (e.g., merchants); whereas a more complicated design for the users who require little incentives (e.g., couriers). The in-depth understanding of virtual and physical beacons' reliability also sheds some light on the decision for future BLE-based sensing tasks in trading off the cost and reliability. (Details in Sec.6.2)

Lesson learned 3: System-Human Synergy. One of the VALID's goals was to help couriers get rid of manual reporting, which can improve their experience especially when carrying multiple orders and cannot operate on the phones timely. Although the imperfect reliability of VALID in the wild prohibits us from a complete replacement of manual reporting, we adopted and tested two complementary mechanisms based on VALID. (1) "report arrival automatically" for the couriers when arrival is detected by VALID, and (2) "notify" the couriers when they want to report arrival that is not detected by VALID, to complement human input. Similar to VALID, many systems are designed to complement human input due to unintentional errors (e.g., driving assistance [4]) or intentional manipulation (e.g., fraud detection [13]). Typically, these systems provide some suggestions to change users' behavior; in turn, the users can provide feedback to correct these systems, leading to a mutually beneficial system-human synergy to improve each other. In the operation of VALID, we found an asymmetrical synergy under a 10-month intervention with 89 thousand couriers, where the couriers improved VALID at a higher degree than VALID improved the couriers' behavior. That is, the notification from VALID indeed changes the couriers' behavior, but the feedback from the couriers are impacted by multiple factors including couriers habit and the platform policy. Therefore it is difficult to directly adopt couriers' feedback to improve VALID and design VALID+. (Details in Sec.6.5)

2 BACKGROUND AND MOTIVATIONS

Overdue Orders. Compared to other logistics (e.g., FedEx), the key feature of instant delivery is its guaranteed fast delivery time. Usually, a pre-defined deadline (e.g., 30 minutes) is shown to the customer when an order is placed. If the order is not delivered within the deadline, it is considered as an overdue order, for which the platform refunds the delivery fee or even pays back to the customer for compensation. This overdue compensation as a penalty eventually comes to the corresponding courier or merchant depending on responsibility, e.g., late merchant preparation or late courier arrival, which is typically determined by the courier's waiting time at the merchant derived from the platform accounting data.

Platform Accounting Data. Our platform collects accounting data from merchants, customers, and couriers. We only focus on couriers' accounting data because they are more relevant to our detection problem. These data log the time and locations of four major order delivery statuses of a courier as shown in Table 1, including accepting an order, arrival at the merchant, departure from the merchant, and final delivery to the customer. All status data are based on couriers' manual reporting on their APPs. These data are significant to the platform because they (1) are used for the platform's new order assignment; and (2) are shown to customers in real-time to improve customers' experiences. However, we learned from couriers' feedback that manual reporting is annoying because they sometimes forget to click or cannot operate on the phone when carrying multiple orders, leading to inaccurate reported data and degraded time estimation and order assignment. Therefore, a major motivation of VALID is to ease couriers' burden so that they don't need to report "arrival" and "departure" at the merchants.

Table 1: Courier Accounting Data

Field	Value
Order/Merchant ID	O001/M001
Accepting	2018/01/10 12:00:00 & Lat./Longitude
Arrival	2018/01/10 12:10:00 & Lat./Longitude
Departure	2018/01/10 12:10:10 & Lat./Longitude
Delivery	2018/01/10 12:25:00 & Lat./Longitude

Citywide Physical Beacon System in Shanghai. We validated the accuracy of couriers' manual reports at scale at indoor merchants. Our team designed, fabricated, and deployed 12,109 physical BLE beacons in Shanghai (each merchant with one beacon as in Fig.1) under their consent with $500K budget. A physical beacon deployed at a merchant continuously advertises a unique ID tuple with the BLE protocol. A courier's smartphone receives this ID tuple in proximity (e.g., 20 meters) of the merchant. The received ID tuple is then uploaded by the courier's phone to the platform server in real-time. The server identifies the merchant matching this ID tuple and decides the courier's arrival time at this merchant. This physical beacon system has been used to obtain the real-time ground truth of the arrival time in Shanghai for our citywide testing of VALID in Phase II.

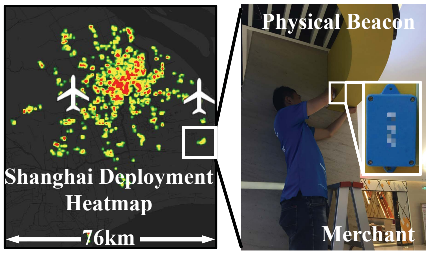

Fig. 1. Physical Beacon Deployment

Inaccurate Reporting Behavior Detected by Physical Beacons. Based on the physical beacon system in Shanghai, we performed a citywide case study of the *Reporting Behavior*. We consider a report accurate if the arrival time reported is within one minute of the actual arrival time. Fig. 2 plots the distribution of the time difference between the actual and reported arrival time of one-month orders, where the tail indicates the early reports. We found only 28.6% of the orders are accurate for their arrival time, and 19.6% of the orders are earlier for more than 10 minutes. Unfortunately, these inaccurate arrival time as courier accounting data have been used in our order assignment and showed to customers for order progress, jeopardizing both order assignment effectiveness and customer experience. Based on these findings, our deployed physical beacon system in Shanghai has been considered a successful pilot study. Motivated by this citywide system, our next stage goal is a nationwide arrival detection system. However, deploying and operating a nationwide physical beacon system nationwide would introduce unbearably high monetary and labor costs because of 530,859 indoor merchants at 364 cities the system needs to cover. It is the key motivation to design VALID as a virtual beacon system.

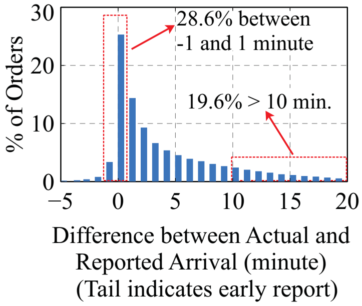

Fig. 2. Inaccurate Reporting.

3 VALID DESIGN

3.1 Requirements of Nationwide Detection

Problem Definition. Our problem is to detect the couriers' arrival time at indoor merchants. This is a simplified real-time indoor localization problem where the timing is needed for only a set of pre-defined locations (i.e., merchants). Given the enormous scale of the arrival detection, a feasible solution has two fundamental requirements as follows.

- **Requirement 1: Reliable for Indoor Coverage.** A feasible solution should cover most merchants (if not all) despite the merchant diversity (e.g., business types, Internet status, etc.). The daily detection workload is based on our business scale, i.e., 14 million daily orders from 83 million monthly active customers.
- **Requirement 2: Low Cost for Nationwide Operation.** A feasible solution should have the average total cost per merchant considerably lower than a physical beacon we deployed ($8 per unit for devices only).

3.2 Opportunity and Challenges

Smartphones as Virtual Beacons. For Requirement 1, if merchants choose to accept orders with their smartphones, they will continuously use their smartphones at their stores as long as they accept orders. Thus these merchants' smartphones will be near couriers' smartphones picking up the orders, making merchants' smartphones potentially reliable for arrival detection. For Requirement 2, we found 85% of our merchants have already been using smartphones to accept orders. As a result, we explore the opportunity of using merchants' phones as low-cost virtual beacons with reliable indoor coverage for VALID. However, the key challenge for this virtual beacon approach is the incentive, i.e., benefits vs. costs. We list two costs and two benefits for merchants to participate in VALID as follows.

- **Cost 1: Energy Consumption.** VALID's resource requirement on merchant phones (e.g., storage and communication) has to be minimized for lower energy consumption (measured by Energy Metric P_{Energy} defined in Sec. 4).
- **Cost 2: Privacy Risk.** An adversary may have devices to listen to merchants' advertising to "re-identify" them from an anonymous open dataset, so VALID has to ensure the re-identification rate (measured by Privacy Metric P_{Privacy}) is low even with large-scale adversarial devices.
- **Benefit 1: Clearer Overdue Accountability.** The platform will know whether the overdue is because of late courier arrivals or late merchant preparation, based on highly-reliable VALID (measured by Reliability Metric $P^{t\cdot n}_{\text{Reli}}$).
- **Benefit 2: Better Order Assignment.** VALID can make the new order assignments for this merchant more effective because we know which couriers are nearby (e.g., just arrived) to this merchant based on VALID. Better time estimation results (e.g., merchants' preparation time, couriers' pickup time) can also be obtained for order assignment with more accurate couriers' arrival time. This effective scheduling reduces the overdue rate (measured by Utility Metric $P^{t\cdot n}_{\text{Util}}$) and leads to monetary saving (measured by Benefit Metric B_T) by avoiding overdue.

The benefits and costs limit VALID's design space significantly, asking for a balance between the simplicity and complexity of different participants, e.g., couriers or merchants.

3.3 Key Idea

Overall Workflow. As in Fig.3, we decide to (0) ask for merchants' consents that we can use their smartphones to detect couriers when they install our merchant APP; (1) let consenting merchants' phones work as virtual beacons to advertise ID tuples continuously by the BLE 5.0 protocol when they are in the order accepting status; (2) let couriers' phones passively scan for ID tuples and then upload received ID tuples to a back-end server in real-time by Internet connection (e.g., cellular); (3) let the server with the received ID tuples check a pre-stored mapping between ID tuples and merchant IDs to detect an arrival finally. Threshold on Received Signal Strength Indicator (RSSI) (e.g, -85dB) is used to shape a moderate detectable region for each virtual beacon device. If a courier picks multiple deliveries from nearby stores and is detected by several beacons by the same time, it's reasonable to conclude the courier arrives at these stores at the same time.

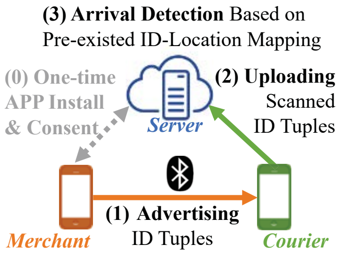

Fig. 3. Key Idea of VALID

Design Simplicity for Merchant as Sender. In practice, we embedded VALID as a software development kit (SDK) (not activate without consent) in the merchant APPs. The merchants' effort is minimized with the following mechanisms: (1) no need for configuration after initial consent, i.e., VALID is "automatic" from the merchant perspective; (2) only advertise a short message of ID tuples when the merchants are in order accepting status, which is automatically obtained from their log-in and log-off records; (3) no BLE scanning; (4) no sensor data (e.g., GPS) collecting. Note that a merchant can switch off VALID at any time in the APP, even after the initial consent. It provides flexibility but introduces potential exploits in theory. See our Discussion for details.

Design Complexity for Courier as Receiver. Couriers are our employees with obligations to join VALID, so we collect courier smartphones' sensor data when they are working (under their consent) to optimize the scanning to save energy. The sensor data are collected at low sampling rate (i.e., 10Hz), and on-device learning is used so that the data are not uploaded to the server. In particular, scanning will stop if the courier is either (1) not moving (detected by accelerometer); (2) away from (e.g., >1km) potential merchants (detected by GPS); (3) not in a delivery task. In practice, we embedded VALID as an SDK in the courier APP. Couriers can *switch off* the scanning in APP if they prefer, even with obligations.

Automatic Arrival Report and Notification on Early Manual Report. We add two functions to improve couriers' experience in the arrival reporting process, potentially influencing their reporting behavior. (1) "Automatic arrival reporting". The function will automatic report the arrival event when a courier's smartphone automatically scanned the virtual beacon in the target merchant. (2) "Early reporting warning". A notification will pop up in a courier APP if she tries to report an arrival manually before VALID detection. The notification remind the couriers that "It seems you are not arrived. Do you confirm report?" She can stop and report later by clicking the "Try Later" (i.e., VALID's detection influences her behaviors) or continue reporting it by clicking the "Confirm" (i.e., she provides feedback to correct our detection potentially). If the courier chooses "Try Later", the same notification will pop up again the next time he tries report but not detected by VALID. The significance for the second function is early reporting is invalid and likely to be penalized if a severe timeout (i.e., 1 hour late) happened. We want the courier to receive a warning during delivery, instead of getting penalized after delivery.

Alternatives and Justification. We let couriers scan and merchants advertise instead of the alternative because letting couriers scan is compatible with existing BLE beacon services in public areas like airports and museums, where couriers can also acquire their locations via BLE scanning. Meanwhile, if we can let couriers advertise and scan at the same time, it can enable fine-grained indoor localization based on couriers' frequent encounters (see our VALID+ in Discussion).

3.4 Trustworthy Advertising

Potential Weakness. The merchants' advertising message is an `ID tuple` with three parameters (`UUID`, `Major`, `Minor`): `UUID` is a 16-byte number to distinguish our beacons from the devices in other systems; both `Major` and `Minor` are 2-byte integer values where `Major` is to identify a beacon group, e.g., a mall; whereas `Minor` is to identify an individual beacon from beacons with the same `Major`. In the protocols based on BLE 5.0 (e.g., iBeacon [27]), an `ID tuple` is fixed for each device and advertised in *cleartext*. It reduces the protocol complexity but leads to merchant privacy or platform security problems under potential advertising sniffing [30].

Potential Attacks. We list two attack models. Model 1: an adversary replicate some `ID tuples` and advertise them in other locations (e.g., mall entrance), which can potentially lead to wrong detection, and thus ineffective order assignment and accounting; Model 2: an adversary deploy some mobile devices to eavesdrop merchants' `ID tuples` by war-driving [49], and build a "mapping" between the `ID tuples` and the side information like store names and locations, which can potentially lead to two problems: (a) a "free rider" problem where the "mapping" is used by unauthorized users (e.g., competing company) to detect their users; (b) a merchant privacy problem where the side information being used to attack an anonymous open dataset to re-identify the merchants (e.g., trajectory mapping with an anonymous CDR datset). [15].

Countermeasures. To address these problems, we augment our advertising with the SM3 algorithm, i.e., a public Time-based One-Time Password (TOTP) [56] algorithm to encrypt the `ID tuple`, similar to Google Authenticator. The core idea is similar with "MAC address randomization [10]" in iOS, where the device hides itself by changing its ID frequently. Specifically, the server assigns a seed ID to a merchant's phone when she logs into our platform for the first time. For every duration of K, the server (1) calculates an encrypted `ID tuple` for each phone based on its seed ID and timestamp; (2) updates the mapping of the merchant's real identity and its newly encrypted `ID tuple`; and (3) sends the encrypted `ID tuple` to the phone for advertising. We did not let the phones calculate the encrypted `ID tuple` locally because of the computation cost, the reverse-engineering risk, and the potential clock drifting on the phones. For the periodical encryption, we empirically set K as one day (Fig.6) in practice. The intuition is that we can use the non-rush hour (e.g., 2 a.m.- 5 a.m.) to complete the ID switching and reduce the potential impact on business process to minimum. A shorter period K makes the advertising safer but may cause some issues. For example, if K is one hour, the chance of encrypted `ID tuple` inconsistency (i.e., between the one sent by the merchant and the one stored at the server) will increase due to unaligned timestamps or lost connections with the server.

4 `VALID` METRICS

Cost Metric 1: Energy Consumption $P_{\textbf{Energy}}$. We use the smartphone battery decreasing ratio (i.e., battery drain) of merchants participating in `VALID` compared to non-participating merchants for energy consumption quantification [37, 40].

Cost Metric 2: Privacy $P_{\textbf{Privacy}}$. We use the re-identification ratio [15] as the privacy metric to measure a `VALID` merchant's risk of being re-identified from a set of anonymous merchants. It is calculated as the percentage of merchants re-identified correctly from all merchants in our data-driven emulation because there were no existing privacy incidents during our 30-month operation as far as we know.

Performance Metric 1: Reliability $P^{t\cdot n}_{\textbf{Reli}}$. We quantify the reliability of a virtual beacon n for a duration t as the percentage of couriers detected by n among all arrived couriers, which are obtained by either physical beacons in Phase II or the platform accounting data in Phase III. In the operation, we set t as one day when evaluating the metrics since all the platform accounting data are collected on a daily basis (detailed in Sec. 5).

Performance Metric 2: Utility $P^{t\cdot n}_{\textbf{Util}}$. We quantify the utility of a virtual beacon n for a time duration t by *delivery overdue rate reduction* of the corresponding merchant. Because overdue rate is a key metric to evaluate a platform's ability to meet deadlines, and its reduction can be easily converted to monetary returns. Since the overdue rates are influenced by many factors (e.g. dispatching, policy, etc), we calculate the overdue rate reduction as a gain between the participating and non-participating merchants as an A/B test, by assuming the other factors are the same in a close geospatial area with similar merchant types, e.g., all fast food stores within a 3 km radius. For example, if the overdue rates for merchant m and n for time period $T1$ and $T2$ are OR^m_{T1}, OR^m_{T2}, OR^n_{T1}, and OR^n_{T2}, respectively. But only n is participating in `VALID`, so n's gain as the overdue rate reduction is $[(OR^n_{\text{T1}} - OR^n_{\text{T2}}) - (OR^m_{\text{T1}} - OR^m_{\text{T2}})]$.

Performance Metric 3: Participation $P^{t\cdot n}_{\textbf{Part}}$. We measure the participation of a merchant n for a duration t by the `VALID` switch-on/off data. $P^{t\cdot n}_{\text{Part}}$ is 0 if `VALID` is switched off for n during t and 1 otherwise.

Platform Benefit Metric B_T. We quantify a merchant n's benefit until time T as

$$B^n_T = \sum_{t=1}^{T} [\ P^{t\cdot n}_{\text{Part}} \ \cdot \ F(\ O^{t\cdot n},\ P^{t\cdot n}_{\text{Reli}},\ P^{t\cdot n}_{\text{Util}},\ C^{t\cdot n}_{\text{Overdue}}\ )\ ],$$

F indicates the monetary saving from reduced overdue penalty of the orders detected by n during t. $O^{t\cdot n}$ is the number of orders during time t in the merchant with n, e.g., 100; $P^{t\cdot n}_{\text{Reli}}$ is the percentage of the orders whose couriers can be detected by n during t, e.g., 80%; $P^{t\cdot n}_{\text{Util}}$ is the absolute overdue rate reduction for orders whose couriers are detected by n during t, e.g., 20%; $C^{t\cdot n}_{\text{Overdue}}$ is the overdue penalty per order for the merchant with n during t, e.g., \$1. An example implementation of F is the product of all these terms, i.e., $O^{t\cdot n} \cdot P^{t\cdot n}_{\text{Reli}} \cdot P^{t\cdot n}_{\text{Util}} \cdot C^{t\cdot n}_{\text{Overdue}}$ (e.g., saving is $100 \cdot 80\% \cdot 20\% \cdot \$1 = \$16$). Thus, the benefits for all participating merchants, i.e., platform benefit, until T is $B_T = \sum_{n\in N_t} B^n_T$ where N_t is the participant set until t.

Behavior Intervention Metric. We measure time difference between detected and reported arrivals before and after the intervention to understand `VALID` intervention.

5 DEPLOYMENT AND OPERATION

Methodology. As in Table 2, we divide our 30-month work into 3 phases based on the ability to test `VALID`: Phase I, an in-lab feasibility

Table 2. Overview of Three Phases of VALID Nationwide Deployment and Operation

Phase		**Phase I: Feasibility Study** (2018/08-2018/09)	**Phase II: Citywide Testing** (2018/09-2018/12)	**Phase III: Nationwide Operation** (2018/12-2021/01)
Scale		10 iOS and 10 Android Phones; In-Lab	98.7 K Merchants; 33.5 K Couriers 46.4 K Orders; Shanghai	3.3 M Merchants (531K Indoor Merchants) 1M Couriers; 3.9 B Orders; 364 Cities
Metrics				
Courier Arrival Detection Reliability		91% **Factors:** Distance **Truth:** In-Lab Experiments	80.8% (Fig.4) **Factors:** Phone Diversity **Truth:** Phy. Beacon & Accounting Data	84% for Android; 38% for iOS (Fig.8, Tab.3) **Factors:** Phone Diversity; Stay Duration **Truth:** Accounting Data
Merchant	**EnergyConsum.**	3.1% per hr Battery Drain	2.6% per hr Battery Drain (Fig.5)	N/A
	Privacy	N/A	0.03% of Re-identification Risk (Fig.6)	N/A
	Utility	N/A	1% of Absolute Overdue Reduction	0.7% of Absolute Overdue Reduct. (Fig.10&11)
	Participation	N/A	81%	85% with Merchant Features (Fig.12)
Platform Benefit		N/A	42 Thousand USD in Total	7.9 Million USD in Total (Fig.7 (iii))
Behavioral Interve.		N/A	N/A	14.2% Improv. of Reporting Behavior (Fig.13)

study where we have complete ground; Phase II, a citywide study in Shanghai where we have physical beacons as the ground truth for real-time evaluation; Phase III, a nationwide operation in 364 cities in China for 25 months where we only have accounting data for post-hoc analysis.

Post-Hoc Analysis. The final "order delivery" reporting in the accounting data in Table. 1 gave us the offline ground truth *in retrospect* for the reliability evaluation nationwide. After an order was delivered, the courier will manually update it as "delivered" on the courier APP. With this reported final order delivery time, we know a courier must have arrived at the merchant some time ago to pick up this order. Otherwise, the courier could not deliver the order to the customer. Therefore, we can find a false negative detect result in retrospect, e.g., a courier arrived at a merchant but never detected by VALID. The reported order delivery time is typically accurate because inaccurate reporting may cause customer complaints, making the couriers rarely do it. Although cases exist that couriers negotiate with customers, and report earlier, our analysis is not affected since we only use the timestamp of the reported acceptance and delivery as the time window to extract beacon data. Even the couriers report earlier than true delivery at the customers, the report are almost certainly after the true arrival event at the merchants. Very few – if not no – couriers would report delivery even before he really arrives at the merchants.

5.1 Phase I (2018/08-09): Feasibility Study

We use 5 iOS and 5 Android phones as senders and other 10 phones as receivers. We test different advertise frequencies and powers by obtaining the results for average RSSI [59] and percentages of advertise messages scanned at five distances, i.e., 5m, 15m, 20m, 25m, 50m. When the APP is active, iOS phones perform better as senders where the advertising signal is stable within 15m with 91% reliability but degrades dramatically beyond 25m. iOS has a restriction on these fine-grained advertise configuration; Android has four advertising powers (i.e., HIGH, MEDIUM, LOW, ULTRA_LOW) where we set HIGH, and three advertise frequencies (i.e., LOW_POWER, BALANCED, LOW_LATENCY) where we selected BALANCED. For energy, continuous advertising only cost 3.1% additional battery on average. We omit the results since they are consistent with existing BLE works [24, 35, 46].

5.2 Phase II (2018/09-12): Citywide Testing

In 2018/09, the VALID modules were embedded in the merchants' and couriers' APP in Shanghai. As merchants update the APP gradually, the number of participating merchants increased from 23 on 2018/09/07 to 98,787 (81%) on 2018/12/07, We select the ones with ground truth (i.e., with physical beacons deployed) for testing. Given the parameters in Phase I, our goal is to validate if the merchant's phones, as virtual beacons in a citywide uncontrolled environment, have the similar reliability as the physical beacons deployed. We also study some metrics that cannot be obtained in Phase I, e.g., Participation, Privacy, and Benefits.

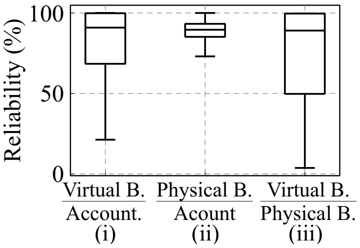

Fig. 4. Reliab. in Three Settings

Reliability. To evaluate the reliability of VALID and compare it with physical beacons, we use accounting data (Table 1) as ground truth. Fig. 4 (i) and (ii) show the percentage of arrival events detected by VALID and physical beacons among all the arrival events. The error bar in Fig.4 indicates the variation of different beacon devices in different days. The average reliability is 80.8% for virtual beacons and 86.3% for physical beacons. The result is reasonable because virtual beacons are also impacted by merchants' mobility and merchants' smartphone hardware and software, as well as the common factors for both virtual and physical beacons (e.g., beacon positions, hardware and software of couriers smartphones). We also calculate the reliability of virtual beacons using physical beacons as ground truth (Fig. 4 (iii), 74.8% on average). It suggests that a gap of reliability between virtual and physical beacons, and a hybrid deployment is promising in improving the overall reliability. Note that a long tail is observed for virtual beacons, possibly due to the impact of different sender-receiver combinations as shown in Table 3.

Energy Consumption. Fig.5 shows the smartphones' energy consumption of participating merchants measured by battery decreasing (2.6%) is very similar to the merchants not participating in both iOS and Android phones. The error bar in Fig.5 indicates the variation of different smartphone devices.

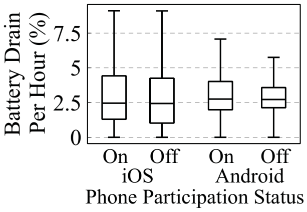

Fig. 5. Energy Consum.

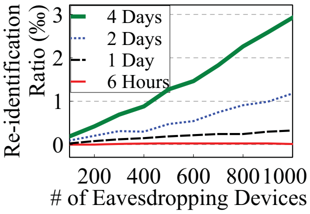

Fig. 6. Privacy Risk

Privacy. Location privacy is quantified with some metrics in [15, 45]. One of the privacy risks is that users' real identities can be derived from their traces by linking with profile data from other sources (mobility data or medical data). Here we use the *re-identification ratio* to measure a user's risk of being identified due to BLE advertising. We assume that an anonymous mobility data of merchants are obtained by the attackers and a bunch of mobile devices are used to eavesdrop. The attackers' goal is to re-identify the real identity of each merchant by linking the trace they eavesdropped and the mobility data they had. Specifically, the re-identification ratio is defined as

$$\text{Re-identification ratio} = \frac{\text{\# of merchants identified correctly}}{\text{\# of all merchants}} \quad (1)$$

Fig.6 shows the re-identification ratio increases as an adversary utilizes more eavesdropping devices based on a data-driven emulation of Model 2 in Sec.3.4. The emulation is conducted as follows. We use one day of merchants' location data in Shanghai – collected for risk control purpose – as a supposedly-leaked anonymous platform dataset owned by the attackers. 1,000 couriers are viewed as attackers to eavesdrop on merchants' advertises to collect side information (e.g., merchants' locations and shop name) to attack this supposedly-leaked data. The attackers' goal is to first match the merchants' traces in the supposedly-leaked anonymous dataset and the merchants' traces collected by the attackers from eavesdropping, and then identify the merchants's identity in supposedly-leaked anonymous dataset. We found that with the ID update cycle set as 1 day (our default setting), a merchant is uniquely re-identified in a supposedly-leaked anonymous platform accounting data set with a set of 73.8K merchants in Shanghai is smaller than 0.03% even under a powerful adversarial attack model. Even we use 4 days as ID update cycles, the risk ratio is still below 0.3%.

Utility. For all participating merchants, their utility (i.e., overdue rate reduction) improves by 25% on average (from 5% to 4%). *The platform benefits are 42 thousand USD* (out of 211 thousand USD of total overdue compensation during the same time in Shanghai).

Detailed analysis of a citywide physical beacon system can be found in [17]. The new findings in the nationwide virtual beacon system is described and discussed in the next section.

5.3 Phase III (18/12-21/01): National Operation

After testing in Shanghai, we started our nationwide deployment in 2018/12, and VALID has been operating until this submission. As VALID evolves as merchants enter and leave, we have been utilizing the accounting data to conduct daily post-hoc analysis to monitor the operation of VALID. We show the results in the next section.

6 NATIONWIDE OPERATION RESULTS

6.1 VALID Evolution

Overview. In Fig.7, we show a panorama of three phases from 2018/08 to 2021/01. In Fig.7 (i), given a day t, we show both the number of merchant phones as virtual beacons N_t with "participating" status in VALID and the number of delivery orders O_t whose couriers are detected by VALID, so every order is associated with an arrival detection by VALID. We also show the active physical beacons we deployed in Shanghai in 2018/01 to compare their evolution with that of virtual beacons. In Fig.7 (ii), we visualize VALID' evolution at 4 key time points. (a) 2018/12: 2nd week of Phase III where VALID has not been uniformly deployed due to batched merchant APP update; (b) 2019/01: 1st month where VALID is fully operated on a few metro hubs; (c) 2020/01: 13th month where VALID is evolved to all major cities and reaches a very high percentage of cities our platform served (336/367); (d) 2021/01: the latest scale as of this submission.

Scales of Virtual Beacons and Arrival Detection. In Fig.7 (i), we found our Phase II testing in Shanghai leads to some fluctuations starting 2018/10 on virtual devices numbers and resultant detection because we tested the operation of VALID in Shanghai by opening and closing scanning functions in some regions to ensure it has no impact on couriers APP's main function (i.e., accepting orders). From the start of Phase III, the scale of VALID has been increasing gradually as more merchants jointed than left every day, partially because of the low cost of being virtual beacons. For these left merchants, we found the main reasons are either merchants switch to other platforms or merchants closed permanently. Based on our data, the merchants' turnover rate is high, i.e., 76.5% of new merchants in 2018 were closed or changed to another store within one year of the opening. We found in the operation stage of Phase III, the number of arrivals detected (i.e., orders) is around 10 times of the number of VALID's virtual devices, which implies each virtual device detects 10 arrivals on average every day. This ratio remains similar throughout Phase III except the mid-February, during which the detection decreases sharply due to the Spring Festival, i.e., the biggest holiday in China. We observed some sharp decreases and recoveries around 2019/02, but the recoveries in 2020 took much longer due to COVID-19, and the corresponding recovery impact on the benefits is in Fig.7 (iii).

Evolution of Benefits. In Fig.7 (iii), we show the platform benefits (defined in Sec. 4), i.e., the money saved due to overdue reduction. We show two cumulative platform benefits based on the empirical value of utility $P_{\text{Util}}^{t \cdot n}$ (i.e., overdue rate reduction after a virtual

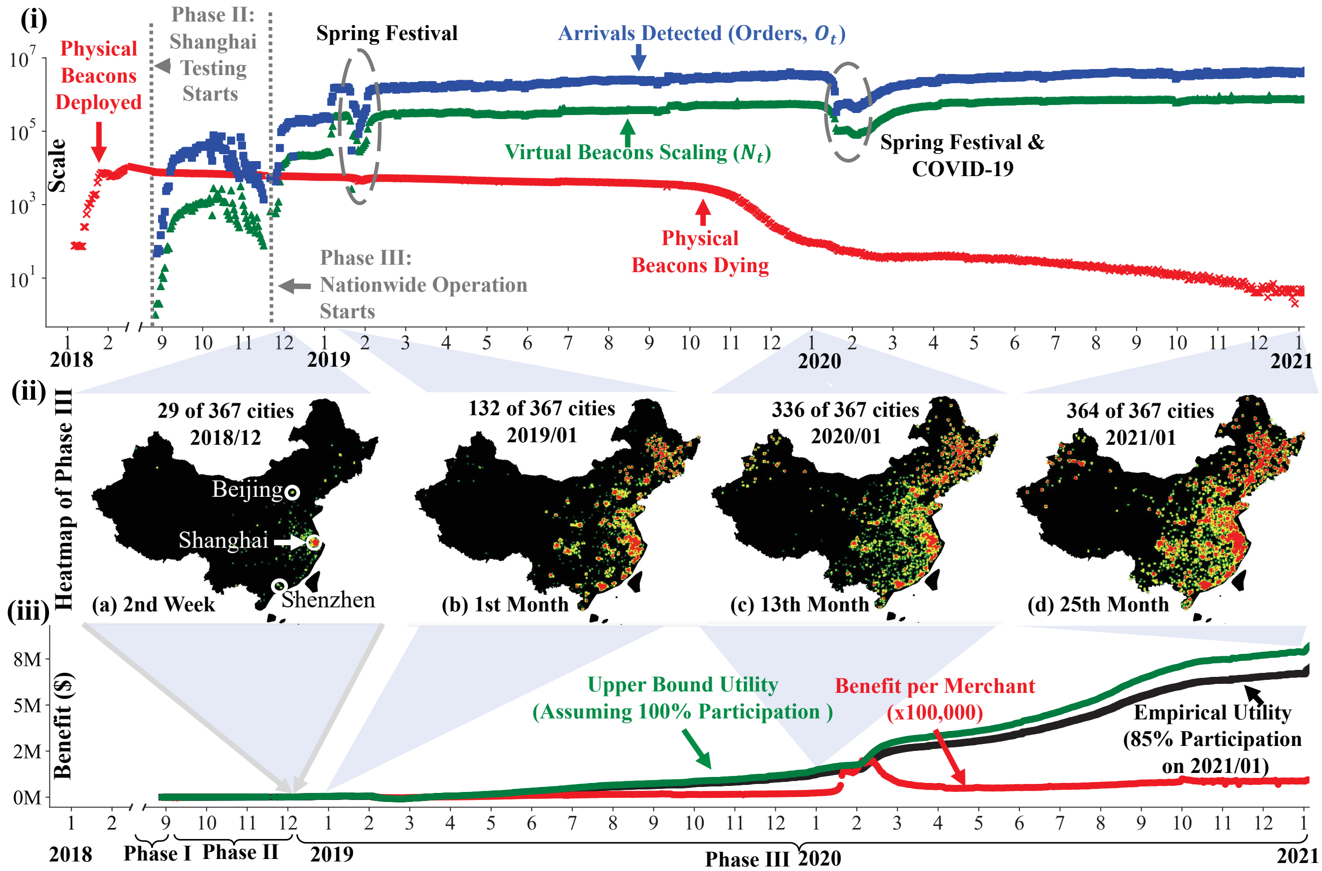

Fig. 7. (i) VALID timeline including Phase I Feasibility Study, Phase II Citywide Testing in Shanghai, and Phase III Nationwide Operation; (ii) VALID nationwide virtual beacon heatmaps for 4 key months and the corresponding time expanded to (i) and (iii); (iii) Benefits with upper-bound and empirical utility and Per-Merchant Benefit.

beacon n participates), along with its upper bound $\overline{P}^{t \cdot n}_{\text{Util}}$ (i.e., we assume all merchants participate), respectively. We also show the average individual benefits for each merchant as the red line. As of 2021/01, the empirical cumulative benefits are $7.9 million, which is impressive compared to the $65 million of estimated overdue compensation in the whole country in 2020. The number is close to its upper bound due to the high participation rate (85%). The platform benefit jumped significantly in 2020/02 due to the merchants' reopening after the strike of COVID-19, which leads to a lower average benefit.

Detailed Lesson Learned 1: **System Evolution in the Wild.** Comparing two systems as in Fig.7 (i), we found that the scale of our virtual VALID has been gradually increasing nationwide; whereas the scale of our physical beacon system (even with a much smaller scale compared to VALID) has been constantly decreasing due to beacon dying for various factors (e.g., vandalism, battery). We have to retire the physical beacon system starting 2019/11. This fundamental observation suggests a participatory less-expensive "virtual" beacon system evolves more robustly (i.e., with a gradually increasing scale) even in a highly uncertain nationwide environment (i.e., 364 cities), compared to a dedicated more-expensive "physical" beacon system in a less uncertain citywide environment (i.e., Shanghai as the most advanced city of China). However, it is essential to provide incentives for merchants to participate in a virtual system by minimizing the participation costs and showing the participation benefits, which can be potentially addressed by the system simplicity and benefit quantification as in Fig.7 (iii). To further increase the platform benefit, our on-going work for the next generation of VALID, i.e., VALID+, enables courier phones to advertise as well to work as "mobile virtual beacons". This is because the merchant phones can only work as "stationary virtual beacons" because they can only detect couriers at merchant locations, making current VALID lacking of beacon mobility. But in VALID+ with courier advertising, we can further increase the system scale and benefit since VALID+ can detect couriers by each others based on their encounters outside the premise of indoor merchants (see Discussion for details).

6.2 Performance Metric 1: Reliability $P^{t \cdot n}_{\textbf{Reli}}$

Impact of Stay Duration on $P^{t \cdot n}_{\textbf{Reli}}$. The stay duration is the time difference between a courier arrives at and departs from a merchant, which is obtained by our accounting data as shown in Table 1. Even

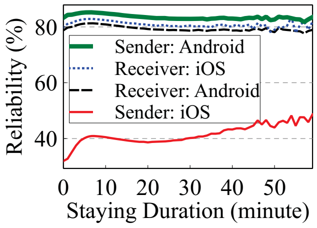

Fig. 8. Stay Duration Impact

there were inaccurate report data due to human errors (Fig.2), our results can provide statistically significant insights based on 3.9 billion orders during more than 25 months. The stay duration varies due to multiple factors but the main factor is waiting for orders. Fig. 8 shows the reliability in four settings with different OS in merchants' (i.e., senders) and couriers' (i.e., receivers) smartphones. We found the reliability is very low (i.e., 38%) with iOS merchant phones compared to Android (i.e., 84%), because of a recent iOS update on permission management that an APP cannot advertise in the background. However, in other three settings, the reliability is around 80%. In particular, we found that within 7-minute stay, the longer that a courier stays, the higher the reliability; after 7 minutes, the reliability reduces gradually because a longer time is not helpful for them to be in proximity.

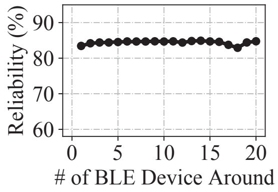

Fig. 9. Density Impact

Impact of BLE Device Density. One potential impact factor on reliability is the interference from nearby BLE devices. We evaluate this impact by finding the BLE advertisement sent by different numbers of merchant BLE devices but scanned by the same couriers' phone at the same time. The results in Fig.9 shows that no obvious impact is observed even there are around 20 different BLE devices (i.e., merchant smartphones) advertising nearby. That is, BLE-based detection is robust to interference from other BLE devices at a relative high density.

Table 3. Impacts on Reliability

Merchant as Sender (Shares) / Courier as Receiver (Shares)	Apple(14%)	HUAWEI(15%)	Xiaomi(6%)	vivo(15%)	Samsung(1%)	Others(49%)
Apple(29%)	37%	80%	89%	77%	79%	86%
HUAWEI(20%)	38%	78%	87%	76%	79%	85%
Xiaomi(2%)	38%	72%	77%	75%	70%	81%
vivo(33%)	37%	77%	79%	76%	77%	84%
Samsung(1%)	39%	81%	83%	78%	82%	87%
Others(15%)	37%	78%	80%	77%	79%	85%

Impact of Smartphones Diversity on $P^{t \cdot n}_{\textbf{Reli}}$. Since the data show our nationwide 1 million couriers use phones with 258 brands and 5,251 models, our goal is to have most merchant and courier phones (if not all) to be compatible at both software (i.e., OS) and hardware (i.e., phone models) levels. It is impossible for us to know if they are using unsupported smartphones or to force them to use specific models or OS. We show five major brands from merchants and couriers in Table 3. We find that when Apple phones were used by merchants to advertise, the reliability is significantly lower because of iOS restricting background advertise. This will be a problem since there is no clue that Apple will change this mechanism in the future. The problem cannot be resolved completely but can be alleviated by deploying physical beacons and VALID+ (Details in Section 7.3). In contrast, we found Xiaomi performs the best as senders and Samsung performs the best as receivers.

Other Impact Factors on Reliability. The position of the merchants' smartphones also impacts the reliability. Although we do not have the ground truth data of merchants' smartphones position (e.g., on the reception desk or in the kitchen), our in-field study in deploying the physical beacons shows that the reliability is significantly impacted by the smartphones' locations [17]. For example, if there are barriers (e.g., wall) between the merchant's and the courier's smartphones, most BLE signals will be blocked thus the reliability drops sharply. We also study the impact of the smartphone battery and find that the reliability is NOT impacted by the battery level.

Detailed Lesson Learned 2: **Reliability in the Wild.** We found that a well-calibrated virtual beacon system for a simple sensing task (i.e., arrival) has low reliability in uncertain environments, even though it has high reliability in controlled in-lab environments. It is mainly affected by many real-world factors including sensing subject status (e.g., stay duration for couriers in Fig.8) and phone OS & hardware combination and permission (e.g., 258 brands and 5,251 models in our platform, and new iOS permission updates). While some impact factors are out of system designer's control, an asymmetric design philosophy has the potential to address the device diversity issues (e.g., 258 phone brands and 5,251 models in our platform), for example, design simplicity for the users who need strong incentives to accommodate diversity (e.g., merchants); whereas design complexity for the users who need little incentives to accommodate diversity (e.g., couriers). Moreover, based on the in-depth understanding of virtual beacons' reliability in the wild, one can achieve a hybrid system based on the tradeoff between physical beacons (high cost, high reliability) and virtual beacons (low cost, low reliability). For example, for high-end merchants requiring more tight delivery time constraints, we can deploy the physical beacons; whereas for normal merchants where arrival detection is only used for data collection for time estimation, we can deploy the virtual beacons. Importantly, based on APP usage data, we found the chance of courier APPs going to background is much lower than that of merchants because couriers have to actively engage with their APPs to report order status, especially when they are near merchants. Thus, in our VALID+, we let couriers to advertise and merchants to receive to increase the reliability.

6.3 Performance Metric 2: Utility $P^{t \cdot n}_{\textbf{Util}}$

We use the overdue rate reduction to measure the utility $P^{t \cdot n}_{\text{Util}}$ of a virtual beacon n during time t. We study two factors impacting the utility as follows.

Impact of Demand/Supply Ratios on $P^{t \cdot n}_{\textbf{Util}}$. The reason for overdue orders is high demand and low supply in some regions of cities, e.g., downtown. We compute the demand/supply ratios in five different cities based on the order number (i.e., demand) and the courier number (i.e., supply) in their 5 km proximity, which is the delivery range limitation. As shown in Fig. 10, we found a higher ratio leads to a higher utility in general, shown as the absolute overdue rate reduction. These areas typically have higher overdue rates compared to others, so VALID works more effectively, which leads to better scheduling and thus higher overdue rate reduction. The error bar in Fig.10 shows the variation in different days. The variance may be due other impact factors like weather and holiday, whereas the positive mean value of utility proves the effectiveness of the system. Note that in our platform, an 0.7% nationwide reduction in the absolute overdue rate is significant because of 14 million daily orders.

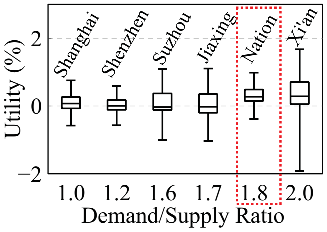

Fig. 10. Demand/Supply Impact

Impact of Different Building Floors on $P^{t\cdot n}_{\text{Util}}$. We found that a virtual beacon device's utility is not proportional to the number of interacting users at its deployed location, but proportional to the uncertainty of user behaviors (e.g., higher floors and lower basements). In general, the ground floor has the lowest overdue rate; whereas higher floors have higher overdue rate due to uncertain indoor mobility to across different floors. So if VALID can detect a courier at higher floor merchants, it has the potential to reduce overdue rates of these merchants for a higher utility. As in Fig.11, where the error bar indicates the variation for different merchants, the utility is higher for VALID in higher floors or lower basements because the variance of the courier mobility is proportional to the indoor travel distance. This is because the higher the merchant floor, the longer the distance from the merchant to the building entrance, the more variance of the couriers' arrival time to the merchant. It leads to a higher utility for VALID at these merchants for time estimation and order assignment. Specifically, couriers tend to report arrival once they enter the merchants' building, which leads to inaccurate arrival reports, especially for basement and high-floor merchants. The inaccurate arrival reports then result in wrong data for the estimation module and introduce wrong dispatching decisions for those merchants [62]. Moreover, GPS-based arrival detection cannot detect this inaccurate report since the couriers and the merchants are close enough in the horizontal dimension. These findings have the potential to provide practical design guidance for both our VALID+ (e.g., deciding where to let couriers to advertise indoor) and other sensing systems in the wild.

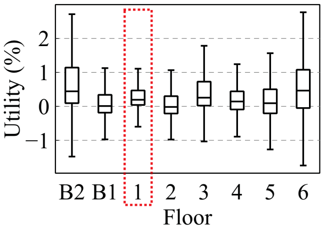

Fig. 11. Impact of Floor

6.4 Performance Metric 3: Participation $P^{t\cdot n}_{\text{Part}}$

Impact of Merchant Experience on $P^{t\cdot n}_{\text{Part}}$. The participation rate (85% on average) is affected by the merchants' choice because they can open and close the advertising at will. Fig.12 shows experience impact measured by the time they open on our platform, where the error bar indicates the variation for different merchants. Although we expect different participation between newer and older merchants, there is no obvious correlation between merchant experience and participation.

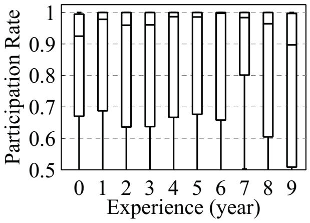

Fig. 12. Experience Impact

6.5 System-Human Synergy

In this part, we discuss the asymmetrical synergy between system and human users (i.e., couriers) including (i) intervention from the system's notification to the couriers' behavior and (ii) the feedback from couriers' behavior to the system.

Behavior Change due to System Intervention. We observe behavior changes in couriers' manual reporting after we add the "early report warning" notification. Fig.13 shows the time difference between detected (by VALID) and reported arrivals before and after the notification was added to their APPs based on the nationwide detection results of VALID. To show the behavior evolving process, we show the results of 2 weeks, 1, 3, 6 and 10 months after the notification was added. We found that the longer the notification was added, the higher percentage of orders with the difference closer to 0, indicating the couriers have been improving their reporting behavior. In particular, the percentage of reporting with errors smaller than 30 seconds increases from 36.1% to 49.5% after three-month nationwide intervention, while increasing subtly to 50.3% after 10 months. It indicates the marginal effect of intervention decreases after longer time.

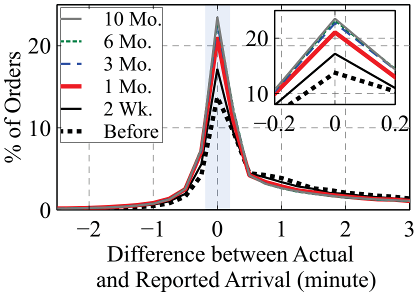

Fig. 13. Reporting Behavior Change

Behavior as Feedback to the System. In a retrospective analysis with physical beacons and couriers' GPS traces, we investigated whether our VALID-based notification shown to a courier for their potential early reporting behavior was correct or not. Note that an incorrect notification is a result from multiple factors including VALID's reliability (e.g.,

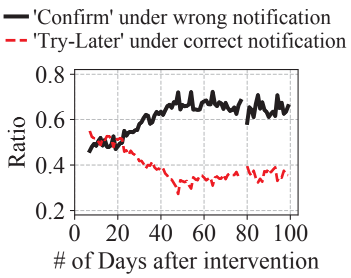

Fig. 14. Behavior as Feedback

merchants' and couriers' smartphone hardware and software), couriers' mobility (e.g., picking up at the door but not entering), and delivery process management (e.g., whether entering building count as arrival). We have been working together other teams to improve the notification quality. At this point, we are interested to see given our intervention, how would the couriers respond, i.e., stop reporting (by clicking 'Try-Later') or continue reporting (by clicking 'Confirm'). As shown in Fig. 14, we collect 3-month notification data in a city under consent after we add the notification function, and calculate two ratios:

(1) Ratio of 'Confirm' clicks when the notification is wrong (i.e., a courier's behavior improves VALID where a courier arrived and reported but VALID did not detect due to limited reliability, 'Confirm'-ratio);
(2) Ratio of 'Try Later' clicks when the notification is correct (i.e., VALID improves a courier's behavior where a courier did not arrived yet but want to report and yet stopped when seeing our warning, 'Try-Later'-ratio).

Ideally, we want both the ratios to be high so that we can trust the feedback. As shown in Fig.14, both ratio is around 0.5 in the first month, which may be caused by random trial clicks since couriers are not familiar with the mechanism; after the first month, 'Confirm'-ratio increases as we expected, but 'Try-later'-ratio decreases, which means even under correct notification, couriers still tend to continue report even he is not arrived yet. A possible reason for this phenomenon is that in current business process, the system tends to trust human report and has not penalized incorrect 'Confirm' or 'Try-Later' clicks. The reason of not instituting penalties is that trustworthy approaches do NOT exist ubiquitously to arbitrate between VALID and couriers on accurate arrival time. A false negative will cost the courier much time to apply for human reconsideration, which also leads to potential cost for the platform. Given the facts, the couriers tend to 'Confirm' to save time (no notification will pop up after 'Confirm') and pretend to have been arrived. This phenomenon suggested that the users improved VALID at a higher degree than VALID improved the users (i.e., an asymmetrical synergy). However, we are still working to filter out useful feedback (e.g. clustered 'Confirm' feedback at specific merchants) to improve the VALID (e.g., tune the parameters in 'automatic arrival report' and adjust 'notification mechanism' accordingly).

Detailed Lesson Learned 3: **Asymmetrical System Human Synergy.** Similarly to VALID, many systems (e.g., AI systems) have been deployed to replace or complement the human input due to unintentional human errors (e.g., driving assistance based on Machine-learning [4]) or intentional human manipulation (e.g., fraud detection [13]). Typically, these systems can provide some suggestions to their users to potentially change their behavior, and, in turn, their users can provide feedback to potentially correct these systems in real time. This interaction enables a bidirectional system-human synergy, i.e., a mutually beneficial partnership to improve each other. We found that under our nationwide intervention based on VALID's detection results, this system-human synergy is actually asymmetrical where VALID benefits more from users' behavior, but only a small percentage (i.e., 14.2%) of users improves their behavior (e.g., stopped the early reporting). It suggested that if we use courier feedback as "labels" to design machine learning models (e.g., our VALID+ introduced in Discussion), the positive courier feedback (where couriers were correct and VALID was incorrect) can serve as more accurate "labels" compared to the negative feedback (where they were incorrect and VALID was correct).

6.6 Correlation between Different Metrics

Due to the space limit, we briefly report our main findings. For the same group of VALID beacons, when the reliability is low (e.g., $< 50\%$ for Apple phones as senders), its correlation with both utility and participation is high, which (i) weakens the utility due to limited data gathered for order assignment, and (ii) lowers the participation because the overall benefits are low. However, when their reliability is high, their participation is more impacted by their utility.

7 DISCUSSIONS

7.1 Limitations of VALID

Limited Baselines. Many baselines could be implemented to further validate VALID's performance. A possible baseline is a reversed asymmetric design where we could deploy a system that is complex on the merchant side to show that it scales or not (our hypothesis is that it does not). Another baseline could be a nationwide physical beacons or Wi-Fi. However, we argue that different from evaluations in a controlled setting, VALID was mainly evaluated and operated in the wild. Given the constraints we have, i.e., reliable indoor coverage and low-cost nationwide deployment, other baselines are almost impossible to be implemented without significant investment. Further, some drawbacks of the above baselines have been discussed in-depth [17, 50] based on costs and incentives, which makes them hard to succeed in a practice setting. Finally, due to privacy concerns, we did not collect merchant smartphone sensor data, so our design to optimize advertising is also limited, which makes some baselines requiring detailed merchant data impractical.

Exploit by Merchants. In VALID, due to our design minimalism on the merchant end, we did not add sophisticated security design. So it is possible that some merchants will exploit VALID by switching off advertising when they were late to prepare the order but the couriers are awaiting in store, and then starting advertising when their orders are ready to pick up. In this case, VALID would detect the courier just arrived and thus the overdue responsibility is on the courier side. However, in practice, we only have very limited complaints from couriers related to such an exploit because the couriers still report manually and sometimes take pictures with timestamp as evidence of their on-time arrival. Moreover, we found that 93% of merchants don't switch VALID states (i.e., turn on or turn off) during the whole day; 99% of merchants switch VALID states ≤ 2 times; 99.9% of merchants switch VALID states ≤ 4 times; only 0.01% of merchants switch VALID states ≥ 10 times. This indicates that potential exploits by merchants is NOT widely observed. The correlate between merchants' behavior and the late deliveries will be studied in future works.

7.2 Ethics, Privacy, IRB, and Data Release

The collection of the couriers' and merchants' data are under their agreement as a part of the *Privacy Policy and User Agreement* for couriers' APP (Item 1.2.2 in [20]) and merchants' APP (Item 1.2.2 in

[21]). In the agreements, couriers and merchants are notified that the data are being collected and their data will be used to support and improve products and services (including BLE data as location). Any type of raw data (GPS trace, BLE, etc.) are deleted from the database completely after a preset life-cycle (i.e., 3 months for the current policy), we only keep the statistical data (e.g., Fig. 7) for the whole 3-year process. Our intervention notification was exempted from IRB because it has no greater than minimal risk. We have been working on aggregate and anonymous data and did not focus on individuals. The couriers and merchants ID are an anonymous keys to join different data sets. As a result, our results cannot be used to trace back to individuals. We did not utilize any personal information, e.g., age and gender. We will release one month of our data collected in VALID for the research community to validate our results and conduct further research. We will follow the data format of a previously released data set from the "aBeacon" platform [6] to protect privacy during the data release.

7.3 VALID+: Next Generation of VALID

Built upon VALID, we initiated a new system called VALID+ to retain its strengths and address its limitation of stationary beacons. We adopt a similar three-phase approach for VALID+, and now we are in Phase I feasibility study. In VALID+, under the couriers' consent, we let couriers' smartphones advertise as well by working as a mobile virtual beacon in addition to merchant phones, which are also mobile phones but they are mostly stationary in the merchant store. If we have courier phones as mobile virtual beacons, we can infer couriers' indoor locations based on the massive courier encounter events in a crowdsourcing manner [44]. Compared to the encounter events detected by VALID at known locations (i.e., merchants' locations), the encounter events detected by VALID+ are at unknown locations since both the senders and the receivers are couriers and are in movement. However, this information can be viewed as the "sample" locations when couriers are traveling among indoor merchants. Based on the data we have, in the rush hour (11am) within a mall area, 79 couriers move around 37 merchants, making 389 courier-merchants interactions and 2,534 courier-courier encounter events. The deployment and operation insights we obtained from VALID have been guiding our development of VALID+ based on machine learning with label data (i.e., courier feedback data to our early reporting reminder) collected from our nationwide behavior intervention in VALID. In particular, based on these label data and our accounting data, we will train high-performance learning models to continuously predict working couriers' status, e.g., indoor locations.

7.4 Generalization

Although VALID is designed for couriers' arrival detection in on-demand food delivery, the core technique, i.e., using smartphone as virtual beacons, can be generalized in both academic and industrial community. For academic research, as a potential proximity detection method, the technique as well as the lessons learned can be applied to interaction tracking [14, 42], range-free localization [25, 58], and context sensing [26, 38]. For industrial applications, AirTag [55] is a sample application, where the tag's location information is relayed by the nearby iPhone in crowdsourcing mechanism.

8 RELATED WORK

Table 4: Operational BLE Device Systems

Nation	Deployment Site	Application	Scale
Iceland	Eldh. museum [35]	Localization	54 devices
U.S.	Beale Street[46]	Presence detection	100 devices
U.K.	Gatwick airport [24]	Localization	2,000 devices
India	Railway station [23]	Presence detection	2,000 devices
Brazil	Tom Jobim airport [5]	Localization	3,000 devices
China	Shanghai [17]	Presence detection	12,000 devices

Operational BLE System. As in Table 4, a few BLE systems are operated in public sites such as airports or museums for presence detection or indoor localization. The largest BLE system we found is the aBeacon system in Shanghai with 12,000 devices. However, all of them require physical device deployment, which is not scalable to nationwide operation.

Indoor Localization. Indoor localization is widely studied for presence detection, indoor navigation, etc, and are mainly in two categories. *(1) Using existing infrastructures.* Wi-Fi devices are the most common indoor infrastructure, and their various signal characteristics are used to achieve indoor localization such as RSSI [59], Time-of-Flight [11], and Angle-of-Arrival [31]. However, their applications to nationwide detection are limited due to configuration costs. *(2) Deploying new infrastructures.* Camera [34] and LED [61] based approaches are recently introduced but are only validated on a small scale because of the data feed restriction and extra costs. Radio-frequency identification (RFID) based approaches [53] are usually used for short-distance proximity measurements (i.e., within 1 meter) that are not reliable for arrival detection, because the distance between couriers and RFID sources cannot be controlled.

9 CONCLUSION

We introduce VALID, an indoor arrival detection system, from its in-lab conception to its nationwide operation in 364 Chinese cities in 30 months. VALID is built upon both opportunities and challenges of using smartphones as beacons for arriving detection, considering a system design balance between simplicity and complexity. We quantify VALID's performance by seven metrics in our nationwide operation in the wild. We identify three lessons regarding system scale evolution, reliability, and system-human synergy. We envision these lessons have implications for other systems requiring long-term operations with large geospatial coverage. We will share one-month data of VALID with the research community to reproduce and build upon our work.

ACKNOWLEDGMENTS

We greatly thank our shepherd, Dr. Kun Tan, and the anonymous reviewers for their constructive comments. We also thank the engineers who helped build and maintain the VALID system over the past several years, which made the observations and lessons learned possible. Prof. Yunhuai Liu is partially supported by the National Key R&D Program of China 2018YFB2100300, 2018YFB0803400, and National Natural Science Foundation of China (NSFC) 61925202, 61772046.

REFERENCES

[1] Alibaba Group, Local Service BU. 2021. VALID Data-set Link. https://tianchi.aliyun.com/dataset/dataDetail?dataId=103969. (2021).

[2] Heba Abdelnasser, Reham Mohamed, Ahmed Elgohary, Moustafa Farid Alzantot, He Wang, Souvik Sen, Romit Roy Choudhury, and Moustafa Youssef. 2015. SemanticSLAM: Using environment landmarks for unsupervised indoor localization. *IEEE Transactions on Mobile Computing* 15, 7 (2015), 1770–1782.

[3] Fadel Adib, Zachary Kabelac, and Dina Katabi. 2015. Multi-person localization via RF body reflections. In *NSDI'15 Proceedings of the 12th USENIX Conference on Networked Systems Design and Implementation.* 279–292.

[4] Abdul Rafey Aftab. 2019. Multimodal Driver Interaction with Gesture, Gaze and Speech. In *2019 International Conference on Multimodal Interaction.* 487–492.

[5] Airport Benchmarking. 2016. Airlines and Airports Are Beaconizing. Webpage. (2016).

[6] Alibaba Group. 2021. aBeacon Data-set. Webpage. (2021).

[7] Amazon. 2021. Amzon Prime Now. Webpage. (2021).

[8] Ganesh Ananthanarayanan and Ion Stoica. 2009. Blue-Fi: enhancing Wi-Fi performance using bluetooth signals. In *Proceedings of the 7th international conference on Mobile systems, applications, and services.* 249–262.

[9] Apple. 2021. Apple Developer Forums. Webpage. (2021).

[10] Apple Inc. 2021. Wi-Fi privacy. Webpage. (2021).

[11] Roshan Ayyalasomayajula, Aditya Arun, Chenfeng Wu, Sanatan Sharma, Abhishek Rajkumar Sethi, Deepak Vasisht, and Dinesh Bharadia. 2020. Deep learning based wireless localization for indoor navigation. In *Proceedings of the 26th Annual International Conference on Mobile Computing and Networking.* 1–14.

[12] Paramvir Bahl and Venkata N Padmanabhan. 2000. RADAR: An in-building RF-based user location and tracking system. In *INFOCOM*, Vol. 2. Ieee.

[13] Alex Beutel, Leman Akoglu, and Christos Faloutsos. 2015. Graph-based user behavior modeling: from prediction to fraud detection. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining.* 2309–2310.

[14] Andreas Biri, Neal Jackson, Lothar Thiele, Pat Pannuto, and Prabal Dutta. 2020. SociTrack: infrastructure-free interaction tracking through mobile sensor networks. In *Proceedings of the 26th Annual International Conference on Mobile Computing and Networking.* 1–14.

[15] Yves-Alexandre De Montjoye, César A Hidalgo, Michel Verleysen, and Vincent D Blondel. 2013. Unique in the crowd: The privacy bounds of human mobility. *Scientific reports* 3 (2013).

[16] Deliveroo. 2021. Deliveroo. Webpage. (2021).

[17] Yi Ding, Ling Liu, Yu Yang, Yunhuai Liu, Desheng Zhang, and Tian He. 2021. From Conception to Retirement: a Lifetime Story of a 3-Year-Old Wireless Beacon System in the Wild. In *18th USENIX Symposium on Networked Systems Design and Implementation (NSDI 21).* USENIX Association, Boston, MA. https://www.usenix.org/conference/nsdi21/presentation/ding

[18] DoorDash. 2021. DoorDash. Webpage. (2021).

[19] Eleme. 2021. Eleme. Webpage. (2021).

[20] Eleme. 2021. Privacy Policy and User Agreement for Couriers' APP. Webpage. (2021).

[21] Eleme. 2021. Privacy Policy and User Agreement for Merchants' APP. Webpage. (2021).

[22] Dumitru Erhan, Christian Szegedy, Alexander Toshev, and Dragomir Anguelov. 2014. Scalable object detection using deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition.* 2147–2154.

[23] Fablian Technologies. 2018. Eddystone beacon installation at Indian Railway stations by Google. Webpage. (2018).

[24] Future Travel Experience. 2017. Gatwick's beacon installation provides partners with blue dot navigation and augmented reality wayfinding. . (2017).

[25] Fei Gu, Jianwei Niu, and Lingjie Duan. 2017. WAIPO: A fusion-based collaborative indoor localization system on smartphones. *IEEE/ACM Transactions on Networking* 25, 4 (2017), 2267–2280.

[26] Shaohan Hu, Lu Su, Hengchang Liu, Hongyan Wang, and Tarek F Abdelzaher. 2015. Smartroad: Smartphone-based crowd sensing for traffic regulator detection and identification. *ACM Transactions on Sensor Networks (TOSN)* 11, 4 (2015), 1–27.

[27] Apple Inc. 2021. iBeacon. Webpage. (2021).

[28] InstaCart. 2021. InstaCart. Webpage. (2021).

[29] Jason T Jacques and Per Ola Kristensson. 2019. Crowdworker economics in the Gig Economy. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems.* 1–10.

[30] Markus Jakobsson and Sid Stamm. 2007. Web camouflage: Protecting your clients from browser-sniffing attacks. *IEEE Security & Privacy* 5, 6 (2007), 16–24.

[31] Manikanta Kotaru, Kiran Joshi, Dinesh Bharadia, and Sachin Katti. 2015. Spotfi: Decimeter level localization using wifi. In *Proceedings of the 2015 ACM Conference on Special Interest Group on Data Communication (SIGCOMM).* 269–282.

[32] Swarun Kumar, Stephanie Gil, Dina Katabi, and Daniela Rus. 2014. Accurate indoor localization with zero start-up cost. In *ACM MobiCom.* 483–494.

[33] Ye-Sheng Kuo, Pat Pannuto, Ko-Jen Hsiao, and Prabal Dutta. 2014. Luxapose: Indoor positioning with mobile phones and visible light. In *ACM MobiCom.* 447–458.

[34] Mingkuan Li, Ning Liu, Qun Niu, Chang Liu, S-H Gary Chan, and Chengying Gao. 2018. SweepLoc: Automatic video-based indoor localization by camera sweeping. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 3 (2018), 1–25.

[35] Locatify. 2021. Eldheimar museum, Intuitively designed beacon based museum audio guide. (2021). https://locatify.com/blog/eldheimar-museum/

[36] Meituan. 2021. Meituan. Webpage. (2021).

[37] Yan Michalevsky, Aaron Schulman, Gunaa Arumugam Veerapandian, Dan Boneh, and Gabi Nakibly. 2015. Powerspy: Location tracking using mobile device power analysis. In *24th {USENIX} Security Symposium ({USENIX} Security 15).* 785–800.

[38] Akshay Uttama Nambi, Ishit Mehta, Anurag Ghosh, Vijay Lingam, and Venkata N Padmanabhan. 2019. ALT: towards automating driver license testing using smartphones. In *ACM SenSys.* 29–42.

[39] Rajalakshmi Nandakumar, Krishna Kant Chintalapudi, and Venkata N Padmanabhan. 2012. Centaur: Locating devices in an office environment. In *ACM MobiCom.* 281–292.

[40] Abhinav Pathak, Y Charlie Hu, and Ming Zhang. 2012. Where is the energy spent inside my app? Fine Grained Energy Accounting on Smartphones with Eprof. In *Proceedings of the 7th ACM european conference on Computer Systems.* 29–42.

[41] Chunyi Peng, Guobin Shen, Yongguang Zhang, Yanlin Li, and Kun Tan. 2007. Beepbeep: a high accuracy acoustic ranging system using cots mobile devices. In *ACM SenSys.* 1–14.

[42] Timothy J Pierson, Travis Peters, Ronald Peterson, and David Kotz. 2019. Proximity detection with single-antenna IoT devices. In *The 25th Annual International Conference on Mobile Computing and Networking.* 1–15.

[43] Kun Qian, Chenshu Wu, Yi Zhang, Guidong Zhang, Zheng Yang, and Yunhao Liu. 2018. Widar2. 0: Passive human tracking with a single wi-fi link. In *ACM MobiSys.* 350–361.

[44] Sanae Rosen, Sung-ju Lee, Jeongkeun Lee, Paul Congdon, Z Morley Mao, and Ken Burden. 2014. MCNet: Crowdsourcing wireless performance measurements through the eyes of mobile devices. *IEEE Communications Magazine* 52, 10 (2014), 86–91.

[45] Reza Shokri, George Theodorakopoulos, Jean-Yves Le Boudec, and Jean-Pierre Hubaux. 2011. Quantifying location privacy. In *2011 IEEE symposium on security and privacy.* IEEE, 247–262.

[46] THINKPROXI. 2017. THINKPROXI ANNOUNCES FAMOUS BEALE STREET IMPLEMENTED BEACON TECHNOLOGY. Webpage. (2017).

[47] Zhao Tian, Yu-Lin Wei, Wei-Nin Chang, Xi Xiong, Changxi Zheng, Hsin-Mu Tsai, Kate Ching-Ju Lin, and Xia Zhou. 2018. Augmenting Indoor Inertial Tracking with Polarized Light. In *ACM MobiSys.* 362–375.

[48] Hien To, Gabriel Ghinita, Liyue Fan, and Cyrus Shahabi. 2017. Differentially private location protection for worker datasets in spatial crowdsourcing. *IEEE Transactions on Mobile Computing* 16, 4 (2017), 934–949.

[49] Arvin Wen Tsui Tsui, Wei-Cheng Lin, Wei-Ju Chen, Polly Huang, and Hao-Hua Chu. 2010. Accuracy performance analysis between war driving and war walking in metropolitan Wi-Fi localization. *IEEE Transactions on Mobile Computing* 9, 11 (2010), 1551–1562.

[50] Daniel Turner, Stefan Savage, and Alex C Snoeren. 2011. On the empirical performance of self-calibrating wifi location systems. In *2011 IEEE 36th Conference on Local Computer Networks.* IEEE, 76–84.

[51] Uber. 2021. Delivering using the Uber Eats app. . (2021).

[52] Uber. 2021. Uber Eat. Webpage. (2021).

[53] Ju Wang, Liqiong Chang, Omid Abari, and Srinivasan Keshav. 2019. Are RFID Sensing Systems Ready for the Real World?. In *ACM MobiSys.* 366–377.

[54] Yu-Lin Wei, Chang-Jung Huang, Hsin-Mu Tsai, and Kate Ching-Ju Lin. 2017. Celli: Indoor positioning using polarized sweeping light beams. In *ACM MobiSys.* 136–147.

[55] Wikipedia. 2021. AirTag. Webpage. (2021).

[56] Wikipedia. 2021. Time-based One-Time Password. Webpage. (2021).

[57] Bo Xie, Guang Tan, and Tian He. 2015. Spinlight: A high accuracy and robust light positioning system for indoor applications. In *ACM SenSys.* 211–223.

[58] Yu Yang, Yi Ding, Dengpan Yuan, Guang Wang, Xiaoyang Xie, Yunhuai Liu, Tian He, and Desheng Zhang. 2020. TransLoc: transparent indoor localization with uncertain human participation for instant delivery. In *ACM MobiCom.* 1–14.

[59] Zheng Yang, Zimu Zhou, and Yunhao Liu. 2013. From RSSI to CSI: Indoor localization via channel response. *ACM Computing Surveys (CSUR)* 46, 2 (2013), 1–32.

[60] Sangki Yun, Yi-Chao Chen, Huihuang Zheng, Lili Qiu, and Wenguang Mao. 2017. Strata: Fine-grained acoustic-based device-free tracking. In *ACM MobiSys.* 15–28.

[61] Chi Zhang and Xinyu Zhang. 2017. Pulsar: Towards ubiquitous visible light localization. In *ACM MobiCom.* 208–221.

[62] Lin Zhu, Wei Yu, Kairong Zhou, Xing Wang, Wenxing Feng, Pengyu Wang, Ning Chen, and Pei Lee. 2020. Order Fulfillment Cycle Time Estimation for On-Demand Food Delivery. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining.* 2571–2580.