I Tag, You Tag, Everybody Tags!

Hazem Ibrahim⋆, Rohail Asim⋆, Matteo VarvelloΔ, Yasir Zaki⋆
⋆ New York University Abu Dhabi, Δ Nokia Bell Labs
⋆ United Arab Emirates, Δ United States of America
hazem.ibrahim@nyu.edu,rohail.asim@nyu.edu,matteo.varvello@nokia.com,yasir.zaki@nyu.edu

ABSTRACT

Location-tags are novel devices which allow to *locate* personal belongings. This is achieved *locally*, mostly via Bluetooth communication with a paired phone, and *remotely*, by piggybacking on the location reported by eligible devices which come into proximity of such tags. Unfortunately, the security measures offered by location tag manufacturers have been insufficient in deterring stalking, *i.e.,* unsolicited tracking of *people* rather than personal *objects*. In this paper, we shed some light on the performance of both AirTags (Apple) and SmartTags (Samsung). We rely both on *controlled* experiments – with a known large distribution of location-reporting Apple and Samsung devices – as well as *in-the-wild* experiments – with no control on the number and kind of reporting devices encountered, thus emulating real-life use-cases. For the experiments in the wild, we recruited four volunteers to carry a mobile phone whose cover was equipped with both an AirTag and a SmartTag, so that ground truth tag location is obtained via the phone's GPS. Meanwhile, we continuously collect – from AirTag and SmartTag companion applications – each tag's location as reported by encountered devices in the wild. Our experiments lasted four months during which our volunteers (and location-tags) traveled 9,378 kilometers across six countries. The main finding from this experiment is that, in the wild, an AirTag's location is reported more frequently than SmartTags, enabling higher accuracy over short time periods, thus with an higher chance of stalking.

1 INTRODUCTION

Location-tags such as AirTag (Apple) and SmartTag (Samsung) enable the monitoring of the location of any object they are attached to. This is achieved *locally* by using Bluetooth Low Energy (BLE) – or using Ultra Wideband if supported – between a tag and the device it is paired with. When the location-tag is out of reach, location updates are provided *remotely* by piggybacking on any compatible iOS device, such as iPhones and iPads (for Airtag), or Samsung Galaxy devices (for SmartTag) which come into proximity of such tag. For a device to be eligible to relay a tag's location, it must support location finding, which is enabled by default on Apple devices, but must be opted in on Samsung devices.

Although the intended use-case of location-tags is for *locating* objects, there is anecdotal evidence of their misuse to *tracking* people [11, 19] (stalking). To the best of our knowledge, no scientific study has yet quantified the accuracy of location-tags, which directly correlates with their ability (or not) to act as stalking devices. Their efficacy in both locating and/or stalking depends on a few factors: 1) the technology adopted, 2) the probability of encountering an *eligible* device, e.g., a Samsung or Apple device with enabled Bluetooth, GPS location, and data connectivity. While the reach of the technology adopted can be studied in a lab, the opportunistic encountering of an eligible device requires experiments in the wild to account for realistic conditions.

The goal of this paper is to study the performance of location-tags. We tackle this problem with both controlled and in the wild experiments. We use controlled experiments to study the characteristics of tag performance with regards to update frequency in an enclosed environment with a known large distribution of device types. We use experiments in the wild to comment on the effectiveness of their opportunistic location reporting in various locations around the globe. We utilize the results of our study to comment on considerations which should be made by location-tag manufactures to deter unsolicited stalking.

For the experiments in the wild, we rely on four volunteers to carry an AirTag (Apple) and SmartTag (Samsung) while traveling to six different countries. The location-tags are mounted on the cover of an Android phone (see Figure 1)– not paired with the tags, and not an Apple or Samsung device – which is equipped with a custom application logging information like GPS location, connectivity, etc. The data collected spans 120 days, and 9,378 Kms traveled across 20 cities. For the controlled experiments, we deployed an AirTag and a SmartTag in our campus cafeteria over five days. We further collaborate with the university's IT infrastructure which provided us a count of the number of Apple and Samsung devices connected to the cafeteria's WiFi at any point in time. During both measurement campaigns, we also run *crawlers* we developed for each tag's companion app (`FindMy` and `SmartThings`) to collect fine-grained tags location history as reported by eligible devices.

Our main findings are as follows:

(F1) In union there is strength. Currently, Apple and Samsung only allow devices from their own network to report the locations of encountered tags. A more open approach where both manufactures cooperate in such a task has the potential to increase location accuracy (by 25-30%, on average). Furthermore, detaching a tag from its proprietary ecosystem enables better unknown tag detection. This more open approach would resolve issues of cross-manufacturer tracking, where, for instance, a SmartTag is being used to track an individual with an iPhone and vice versa.

(F2) Update frequency optimization is key. While AirTags provided updates to their locations more frequently in the wild, we believe that reducing the frequency of updates to the user can help reduce the effectiveness of unsolicited tracking.

Location updates should still be recorded as frequently as possible on the server, but only relayed periodically to the owner of the tag as to limit the capacity by which one could track an individual. We observe that while SmartTags relayed location updates more than twice as slowly as AirTags in the wild, both tags achieve similar accuracy rates over the duration of an hour, indicating that effective tag location can be achieved with a lower update frequency. The frequency of updates relayed to the user should be optimized by location-tag manufactures as to limit the efficacy by which one could *track* an object, while maintaining the ability to *locate* it in a timely manner.

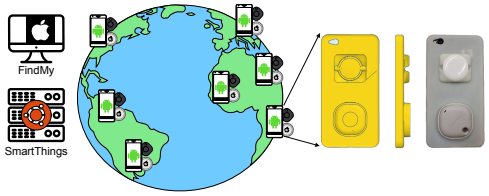


Figure 1: Visualization of our measurement platform. On the left, two data collection servers (MacOS and Ubuntu) where we run the `FindMy` and `SmartThings` crawlers. On the right, several views of our vantage point consisting of a Redmi Go equipped with two location-tags.

2 BACKGROUND AND RELATED WORK

Location-tags like AirTag (Apple), SmartTag (Samsung) and Tile (Tile) use the Bluetooth Low Energy (BLE) [18] protocol to transmit a unique identifier with a range of up to 100 meters (under ideal conditions). The "pro" model of the SmartTag (SmartTag+) and the AirTag also support Ultra Wideband [24] which further extends the range while allowing more precise device localization. Ultra Wideband is only supported by recent devices, such as iPhone models ≥ 11, and Samsung Galaxy series from the S21 onwards.

In addition to *local* tracking, e.g., when a tag is in Bluetooth range of its associated device, location-tags allow *remote* tracking. This is achieved by allowing eligible devices – iOS devices for AirTags, Samsung devices for SmartTags, any device on which the Tile application is installed for Tiles – to report the location of tags encountered in the wild. Whenever an eligible device comes in the proximity of a location-tag, *i.e.,* receives a Bluetooth beacon advertisement, it updates the tag's location in the cloud using its own GPS coordinates as an approximation. The owner of the tag can check its location at any time using the tag's companion application. This process is private, without leaking any information about either the tag's owner or the device which has reported its last location.

Apple and Samsung have implemented measures to deter malicious and unsolicited tracking in their tag ecosystems, yet these measures have been insufficient. The main issue is that each vendor only alerts a user if an unpaired tag from the same vendor has been in their vicinity for an extended period of time. This means that, for example, an AirTag can easily be used to stalk Samsung users and vice-versa. To address this concern, Apple released the application "Tracker Detect" [6] which allows Android users to manually scan for nearby AirTags. Heinrich et al. [17] improved this design by automatically alerting users if they encounter the same AirTag in three separate locations within a 24 hour period. Similarly, Briggs et al. [10] extends this design to generic tags, not just AirTags. These applications are however only partially effective due to the fact that location-tags rely on MAC address randomization, *i.e.,* they eventually appear as a new tag to a third-party application. Last but not least, Mayberry et al. [20] demonstrate the ability to build a custom location-tag which mimics an AirTag's functionality and has the capacity to be tracked in Apple's `FindMy` network while not triggering item safety alerts, circumventing Apple's system for detecting malicious AirTags. This is done by utilizing three different approaches, one of which involves bit flipping to mark the custom location-tag as "lost", since lost AirTags do not trigger alerts regarding unsolicited tracking.

To the best of our knowledge, no previous paper has investigated the efficacy (*i.e.,* accuracy and responsiveness) of location-tags in real-world scenarios, which is the main contribution of this paper. Instead, Givehchian et al. [14] have investigated the privacy of devices using the BLE protocol, such as location-tags. While cryptographic anonymity protects the devices at the application layer, it is unclear whether the signal at the physical layer can uniquely identify such devices. The paper investigates multiple factors which can impact fingerprinting accuracy. They set up a passive listener of BLE beacons and evaluated how much each of these factors limits accurate fingerprinting in a large-scale field study, where they find that physical-layer identification is viable although sometimes unreliable.

3 METHODOLOGY

Figure 1 visualizes the methodology we have devised to study the two most popular location-tags in the market: AirTag (Apple) and SmartTag (Samsung). The methodology is generic and can be applied to other tags, which is part of our future work. The figure shows two *data collection servers* (iOS and Ubuntu) whose goal is to continuously monitor the data available at the companion application of each location-tag. The figure also shows multiple *vantage points*, each consisting of a mobile device mounting the two location-tags via a custom cover. In the remainder of this section, we detail the measurement methodology.

3.1 Location-tags Pairing

Apple AirTag: This tag must be paired and registered via Bluetooth with an iOS or iPadOS device above version 14.5, *i.e.,* no MacOS. Once the tag is linked to the Apple ID of the device it is registered with, it is then displayed in the `FindMy` app across all devices that have signed in with that Apple ID (including MacOS devices). From the `FindMy` app, a user can see a map with the tag at its most recent reported location.

Samsung SmartTag: This tag can only be paired and registered via Bluetooth with a Samsung Galaxy device running Android 8.0 or above. The tag is linked to the Samsung account of the device it is registered with, and it is displayed as a linked device in the Samsung `SmartThings` app. This app, currently only available on Android, contains some basic features to interact with SmartTags, such as view their location on a map. The `SmartThings` app is not only dedicated to SmartTags; it also displays and controls Internet of Things (IoT) Samsung devices linked with the same account.

3.2 Tag Data Collection

Neither Samsung nor Apple offer public APIs to access tag's location data (<timestamp, GPS location>) as maintained by each tag's companion app: `FindMy` (Apple) and `SmartThings` (Samsung). In addition, `FindMy` does not support location history, and `SmartThings` only provides some low resolution location history for up to 6 days. Accordingly, we developed "crawlers" for both apps which constantly monitor – one crawl per minute – for location changes and can thus build fine-grained tags location history.

`FindMy` Crawler: The `FindMy` application is available for most Apple devices, e.g., Macbook, iPhone, and iPad. For ease of instrumentation, we write our crawler for MacOS. Note that MacOS version 11 or above is needed, since `FindMy` on older MacOS versions does not support AirTags. In `FindMy`, users can find the last reported coordinates of any AirTag paired with their account as follows. First, by clicking on the targeted tag from the list of devices in `FindMy` and selecting the option to open the location in Apple Maps. Once Apple Maps is launched, a pin is placed on the map with the latest reported location of the tag. With a right-click on the pin, the user is given the option to "copy coordinates".

We wrote a `FindMy` crawler in Python using the `pyautogui` [4] library to automate the above operations, and store in a file the last reported coordinates of each available AirTag. Along with a tag's coordinates, we also store a timestamp which approximates when the coordinates were reported. This is computed using the crawling epoch time and the time at which a tag was last seen which is reported by `FindMy` as "X minutes ago", thus adding a potential error of up to one minute. Given this "last seen" time cannot be extracted from the `FindMy` app, we resort to Optical Character Recognition (OCR) [22] to convert a screenshot of its value into usable text.

`SmartThings` Crawler: The `SmartThings` application is only available for Android devices. In the application, users can retrieve the coordinates of a SmartTag as follows. First, they select a tag from the list of tags associated with their account; then they click "view location" which opens Google Maps with a pin showing the location of the tag. At this point, the tag's coordinates are available in the search bar and can be copied. We automate `SmartThings` via the Android Debugging Bridge (ADB [1]), a rich Android protocol which allows to automate app operations like launching, scrolling, and GUI interaction. Specifically, we connect an Android device, previously paired with one or more SmartTags, to a Linux machine via USB. ADB is then used to launch `SmartThings` and iterate over the tags. Once a tag's coordinates are available in the Google Maps' search bar, they are copied and logged to a file along with a current timestamp and the time at which the tag location was updated last. As per `FindMy`, this "last seen" time is shown as "X minutes ago" and can only be extracted via OCR [22].

3.3 Vantage Point

A vantage point consists of an Android device (Xiaomi Redmi Go)[1], an AirTag and a SmartTag; both tags are mounted on a custom cover for the mobile device which we designed and 3D printed (see Figure 1). The tags are paired with testing Samsung and Apple accounts we have created, using the procedure described in Section 3.1. Note that the Android device used for a vantage point is not capable of reporting the location of neither the AirTag nor the SmartTag, thus not impacting the accuracy of the experiments to be conducted.

The Android device is equipped with a mobile application we developed which collects GPS data, if available. The application records pairs of <timestamp, GPS location> with a 5-second frequency which are buffered on the phone for up to five minutes. Note that only GPS variations are recorded, thus avoiding redundant data. After five minutes, the buffered data is POSTed to a server in our lab, if a data connection is available. Otherwise, the data is kept in the buffer until a connection becomes eventually available. The <timestamp, GPS location> pairs are used as the ground truth of where the tags were located at a given point in time. This allows us, at any point in time, to evaluate the accuracy of a tag's location as shown by its companion app, *i.e.,* as reported by eligible devices opportunistically encountered by location-tags.

4 DATA COLLECTION

This section describes two independent experiments (controlled and in-the-wild) we have conducted using the previous methodology. It further details the crawling infrastructure we used, along with a few challenges we have experienced.

Controlled Experiment – We deployed an AirTag and a SmartTag within a busy cafeteria at a university over the course of five days. The cafeteria serves roughly 1,000 students, faculty, and staff, and operates between 7:30 A.M. and 10 P.M. everyday, with peak hours during lunch (12 P.M. to 3 P.M.) and dinner (6 P.M. to 9 P.M.). Meanwhile, we ran our crawlers and collaborated with the university's IT infrastructure team to monitor the number of distinct Apple and Samsung devices connected to the WiFi access point in the cafeteria. This is achieved by investigating the destinations of the traffic generated by each connected device, which was done by the university's IT team. The rationale is that a clear distinction arises between Samsung and Apple devices since they rely on disjoint and proprietary data-centers, *i.e.,* no

[1]It is a low-end Android devices mounting a 1.4 GHz Quad-core and a 1 GB RAM.

Country	# of cities	# of Samsung pings	# of Apple pings	Distance travelled (km)	Walking (km)	Jogging (km)	Transit (km)	Time Spent (Days)
United States	2	145	4821	906.92	14.28	21.56	871.07	29.9
Italy	10	1361	4520	3395.13	157.05	68.26	3169.83	28.16
UAE	2	1442	9572	3679.69	144.95	150.92	3383.82	52.33
Pakistan	1	129	454	193.69	12.59	15.94	165.16	1.97
Switzerland	1	331	489	91.11	13.76	15.58	61.77	3.06
Germany	4	187	1225	1112.33	45.9	45.17	1021.26	4.96
Total	20	3595	21081	9378.87	388.53	317.43	8672.91	120.38

Table 1: High level statistics from the data-set collected in the wild.

third-parties, to run their services. In the past, it was possible to reliably infer a device's vendor from its MAC address's Organizational Unique Identifier (OUI). As modern mobile phones utilize MAC address randomization [9], this technique is no longer viable. The information collected was aggregated into a count of the number of Apple and Samsung devices at different time periods, and thus completely anonymized, and returned to us for analysis.

One limitation of this experiment is that we might miss devices which do not connect to the cafeteria WiFi. While we cannot quantify this limitation, we believe most phones rely on the cafeteria WiFi due to poor mobile coverage in the area. Another limitation is that we approximate the number of Apple and Samsung devices connected to WiFi to the number of *eligible* devices for reporting a tag location. This can be an overestimate for Samsung devices whose users are required to *opt-in* to enable this behavior. This is less of a concern for Apple devices which instead require their users to *opt-out* of this behavior.

In The Wild Experiment – We deployed our vantage points via four study participants between March and August 2022. In total, the tags were carried over 9,378 Kms across six countries (Germany, Italy, Pakistan, Switzerland, UAE, and USA, New Jersey) and 20 cities. Table 1 provides detailed information about the data-set we collected. Study participants were instructed to carry the vantage point as much as possible, and only interact with the phone to charge it, connect it to a WiFi network, or insert a SIM card with a mobile data plan. To avoid biasing results in favor of either tag, participants ensured the location reporting option was disabled on any personal Samsung or Apple device they owned that meet the requirements to interact with the tags. Other family members were not required to do so.

We can do this, but it may change a lot of our results/figures. This will take some time to run, I will work on it and update the figures when I can. Intuitively, decreasing the radius to 200 meters makes more sense, but i'm getting weird results so far, so I'm investigating why, which is taking some time. We also filtered any data which was recorded within a kilometer radius of our participants' home locations, as to not bias our data in the event of a neighbor or family member's phone repeatedly reporting a tag's location. Home locations are assumed as our participants homes, hotels, or any other area in which they slept overnight. Overall, this filter accounted for 74% of all data collected.

Ethics –The underlying intention of our research is to assess the performance of location-tagds in the wild. Given that we recruited participants to carry custom prepare mobile devices, we obtained an institutional review board (IRB) approval (HRPP-2021-185) to conduct these studies. In addition, one of the authors has completed the required research ethics and compliance training, and was CITI certified [2]. Participants were also provided with a consent form to read, and sign, acknowledging their willingness to participate. They were given the opportunity to ask questions about the study and what was being collected. We do not collect any identifiable, sensitive, or personal information about the participants. The only foreseeable concern that might put our users' privacy at risk is the collection of the phones' GPS data, which in principle can reveal the participants movements. We did inform the participants about this concern and we obtained their written consent that they approve this collection. As such, we believe that this experiment is deemed to be of low-risk.

5 RESULTS

This section analyzes the location-tag data-sets collected in the wild and in our university cafeteria. We first overview a common methodology we have devised to analyze location-tag data-sets, and then dive into the actual analysis.

5.1 Methodology

We analyze the performance of AirTag and SmartTag both independently and *combined*. The combined performance emulates a scenario where Apple and Samsung devices can report the location of each other tags, functionally detaching the two tags from their proprietary ecosystems. This is achieved by assuming a combined ecosystem, where Samsung and Apple devices are allowed to report the locations of both AirTags and SmartTags. To assess the performance of each tag, we rely mainly on two metrics which we detail in the following: *accuracy* and *responsiveness*.

Assessing Tag Accuracy – At a high level, assessing the accuracy of a tag consists of comparing its reported location, at a given time, with the location of its associated vantage

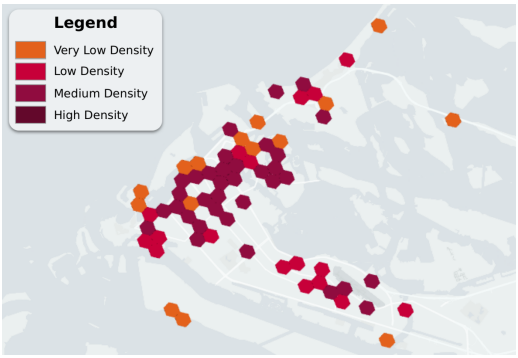


Figure 2: Visualization of hexagons – assuming Uber's H3 index and a resolution of 8 – visited by one of our study participants over XX days in Abu Dhabi. The color of each hexagons relates to an estimate of population density provided by the Kontur data-set. – FIXME

point. We proceed as follows. We group all locations reported within the same X-minutes interval into the same "bucket". For each X-minutes "bucket", we calculate the distance between the location reported by the vantage point and the locations crawled from each companion app of both AirTag and SmartTag. If the distance between the vantage point's location and a tag's location is below a (radius) threshold we count a "hit", otherwise we count a "miss". We then compute a tag's *accuracy* as the percentage of hits in a given scenario, e.g., when considering a given time of the day or radius.

Intuitively, several factors impact a tag's accuracy. First and foremost, the tag's location is approximated by the GPS location of the reporting device. Given Bluetooth has a 100 meter range, this can cause an error of up to 100 meters. Don't think so since we don't know what device is reporting, it could be the same device reporting twice in a given time frame, or different devices. . Another source of error is due to potential movement of both the tag and the reporting device: as these devices move, the time needed to extract and report the GPS location can introduce some error. For example, we sample GPS locations every 5 seconds; when moving on a high speed train (300 Kmh) this can introduce an error of 400 meters. For these reasons, when studying tags accuracy we introduce the following range of radii to consider whether a location reported was a hit or a miss: 10m, 500m, and 5km. We chose 10 meters as a challenging radius, given it is much smaller than the maximum Bluetooth range. Achieving high accuracy with such a small radius is useful to reduce the stress of searching an object in a larger space, but it can be dangerous from a stalking perspective. Next, we choose [...] Good idea, will work on this. We consider radii greater than 100 meters, which is the maximum range of BLE, due to the fact that devices which may relay a tag location could also be on the move and/or do not relay their locations instantly.

All GPS locations collected are grouped into hexagons in accordance with Uber's Hexagonal Hierarchical Spatial Index (H3 index) [7]; grouping GPS locations allows to aggregate location data over a small area to average out anomalies. The H3 index models the globe as an icosahedron, and creates 12 pentagons centered on each of its vertices joined by 110 hexagons. Each cell is then recursively filled with a number of hexagons which depends on the desired resolution. The total number of cells at a given resolution r is given by $c = 2\,120 * 7^r$. For instance, at a resolution of zero, the earth is covered by 110 hexagons, whereas at a resolution of eight, the number of hexagons increases to 691,776,110 [5]. Naturally, as the number of cells which cover the surface of the earth increases, each cells occupies less area overall. At a resolution of eight, an individual hexagon has an average area of 0.737 km^2.

In our analysis what were doing is that we are measuring things with raw data, but grouping by hexagon. For example, lets say in a city we spend a lot of time (let's say 20 percent of the time) in one area and we get really bad results, then grouping by hexagons will mean that it only accounts for 1 hexagon of the many hexagons in the city, rather than 20 percent of the data. , we use a resolution of 8 as it coincides with the resolution used in the Kontur Hexagon Population density data set [3], which reports population densities within H3 hexagons inferred from satellite images of building density. Figure 2 shows an example of the hexagons visited by one of our study participants in Abu Dhabi (UAE). We consider an hexagon *visited* if our participant (and tag) spent at least 5 minutes within it, thus ignoring hexagons which a user has only visited briefly, e.g., while driving on the highway. We color code each hexagon using the Kontur data-set for population density, from very low density (orange) to high density (dark read). need to update figure because we moved to a 3 bucket model (low, medium, high) instead of 4 buckets (very low, low, medium, high).

Assessing Tag Responsiveness – Having accurate tag locations is important, but for these tags to be a viable way of locating misplaced items, their locations also need to be reported in a timely manner. If a tag's location is updated frequently, then the owner will have less area to backtrack as (s)he realizes that the "tagged" object was lost. At the same time, a high update frequency is also an enabler of stalking or unsolicited tracking of a person. We calculate tag responsiveness as the time difference between the timestamp of the first hit – *i.e.,* when the distance between the vantage point's location and a tag's location is below a radius — and the first time that the vantage point reported such location.

5.2 Controlled

We start with the analysis of controlled experiments which were performed over five days in our campus cafeteria (see Section 4). Specifically, we analyze each tag's *update rate*, computed as the number of location updates reported by eligible devices every hour. This analysis allows to shed some light on the behavior of each tag ecosystem, which will then allow us to explain our observations in the wild.

Figure 3 shows the update rate as a function of the surrounding devices which can potentially report a tag's location, or *eligible devices*. The figure shows, for each hour of the day, the average (over 5 days) tag's update rate and device count,

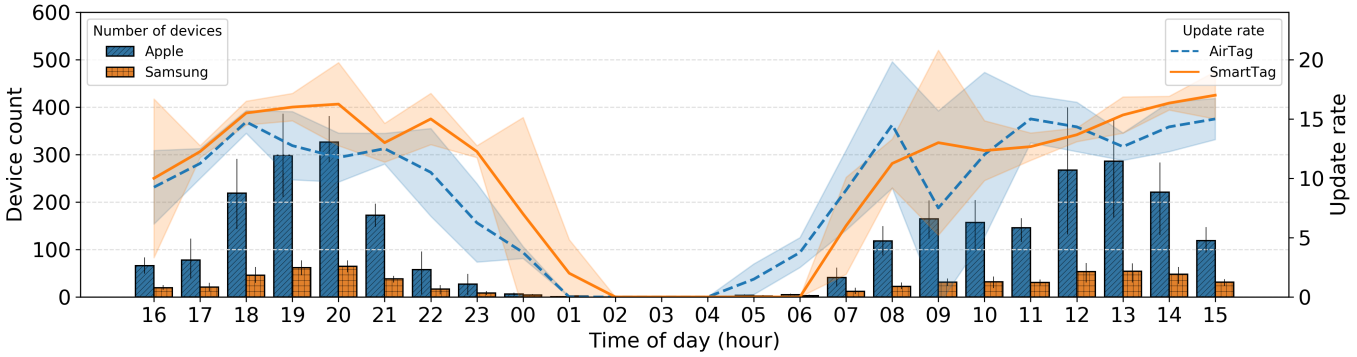


Figure 3: Update rates of AirTag and SmartTag at different times of day in a busy university cafeteria.

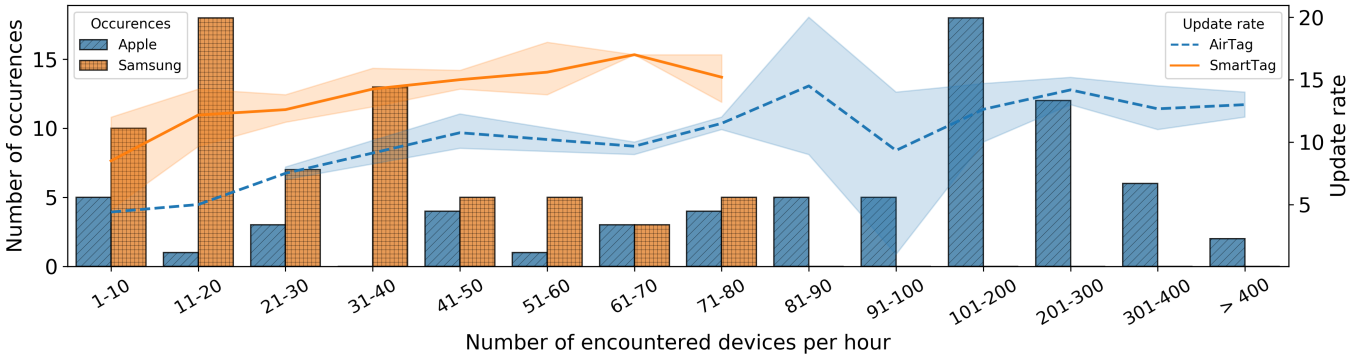


Figure 4: Update rates of AirTag and SmartTag with varying numbers of eligible devices.

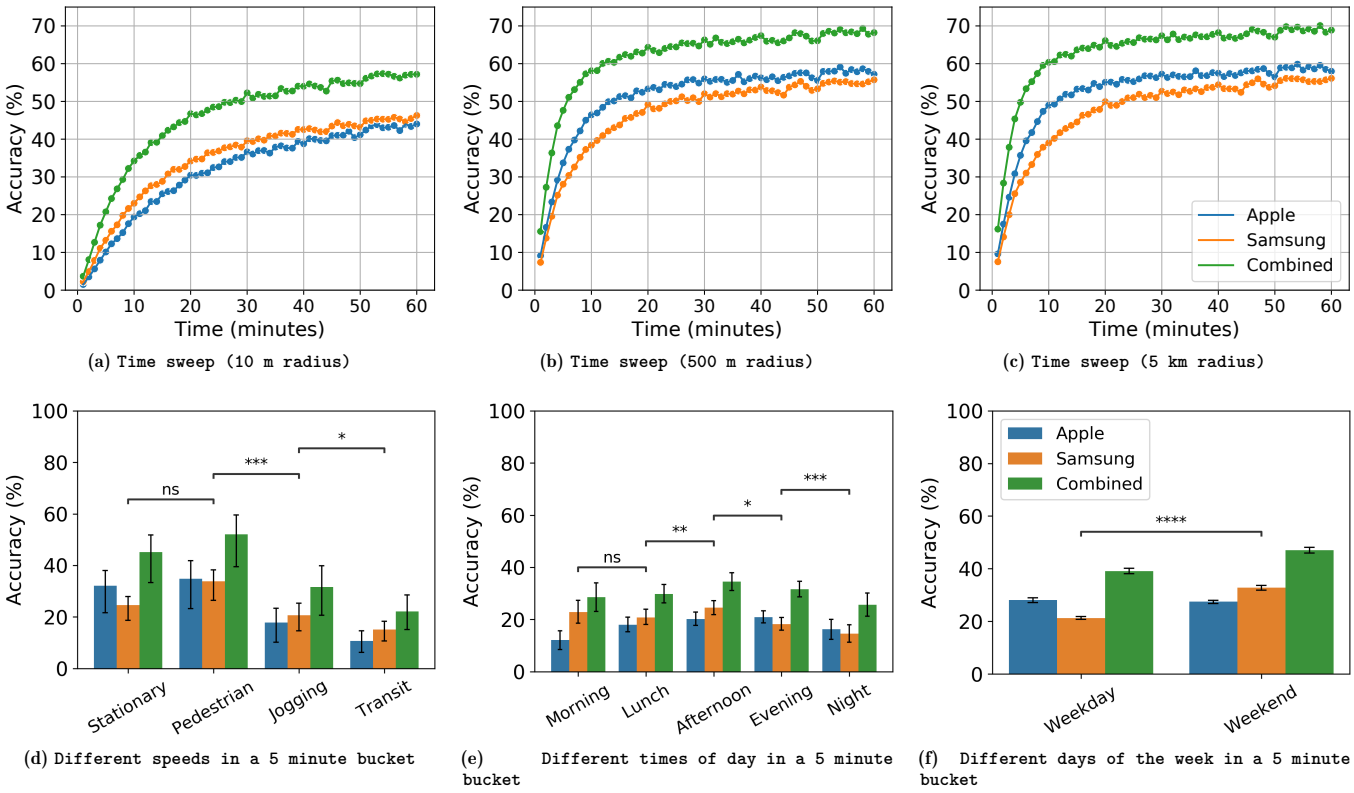
i.e., the number of Apple and Samsung devices present in the cafeteria. The shaded areas and errorbars in the figure report the standard deviation of each metric. The figure shows an overall similar update rate between tags, peaking at roughly 15 updates per hour during lunch and dinner, and dipping to zero as the cafeteria closes over night. However, the figure also shows that there were far more Apple than Samsung devices, up to 6 times more devices during peak hours, e.g., 320 Apple devices versus only 50 Samsung devices at 8pm.

To further understand the previous result, Figure 4 shows the update rate as a function of the number of eligible devices present in the cafeteria, e.g., up to 10 and between 10 and 20. The figure also shows how likely it was to find N eligible devices, Samsung or Apple, in term of "number of occurrences" over the five days. As expected from Figure 3, it is more likely to find few Samsung devices, e.g., less than 20, whereas it is more likely to find lots of Apple devices, e.g., between 100 and 300. The key result of this analysis is that, while both AirTags and SmartTags converge to a similar maximum update rate (15-20 updates per hour), they do so in a very different way. Samsung implements an *aggressive* update strategy, which quickly converges to the maximum update rate. In contrast, Apple implements a *conservative* strategy, e.g., half the update rate than Samsung when less than 20 devices are present.

the problem is that the IT department were unwilling to provide any information outside of a count of the number of devices during a given hour. That's all the information we have, so I don't think we can do further analysis for point 2. As for point 1, we can do something similar to the time sweep if you wish.

5.3 In-The-Wild

Tags Accuracy and Responsiveness – cool idea, will work on this We begin our analysis by investigating each tag's accuracy as a function of its responsiveness within a given radius. Figure 5 summarizes this analysis as we consider a radius of 10 meters (5a), 500 meters (5b) and 5 km (5c); note that "combined" refers to a unified ecosystem where Apple and Samsung devices can report the location of each other tags. Each point in the figure is computed as the ratio of all hits and the sum of all hits and misses, across space (location where each tag was deployed) and time (when a specific tag was deployed). Intuitively, Figure 5 (a,b,c) shows that relaxing the responsiveness, *i.e.,* allowing more time to locate a tag within a radius, improves tag accuracy, e.g., the combined tag's accuracy for larger radii (500 meters and 5 km) grows from 10% to 70% as the responsiveness grows from one to 60 minutes. The figure also shows that the accuracy of the combined tags offers a 15% accuracy improvement, on average, over the accuracy of each individual tag.



(a) Time sweep (10 m radius) (b) Time sweep (500 m radius) (c) Time sweep (5 km radius)

(d) Different speeds in a 5 minute bucket (e) Different times of day in a 5 minute bucket (f) Different days of the week in a 5 minute bucket

Figure 5: Evaluation of AirTag, Smartag and "combined" accuracy in the wild.

The previous observations also apply to a small radius (10 meters, see Figure 5a) although with a few important differences. First, one minute is too fast to locate a tag within such a small radius, e.g., an accuracy of 2% versus 8-10% at larger radii. Second, as we relax the responsiveness, the tag's accuracy increases much slower than what is observed for larger radii, e.g., 20-25% versus 40-47% assuming a responsiveness of 10 minutes. This happens because, as both tags and reporting users might move, it is more challenging to correctly report the right location with such small radius and high responsiveness. Finally, the maximum accuracy caps at 60%, when considering both tags combined, or 10% less than what observed for larger radii. lets revist once we know which radii were going to be interested in. Given the slow responsiveness allowed, this reflects errors introduced by approximating a tag's location with the reporting device location, which can be up to 100 meters over Bluetooth. Overall, this analysis shows that very accurate (10 meters radius) localization is rare, and only possible if considering a combined ecosystem or tolerating slow responsiveness. This is an important result when thinking about stalking, since it implies that a victim can rarely be tracked in real time, but almost half of their movement can be back-traced with very high accuracy after just a one hour delay.

Finally, if we focus on each tag independently, Figure 5a shows that SmartTag (orange lines) slightly outperforms AirTag (blue lines) at a radius of 10 meters. At larger radii, this trend reverses with AirTag achieving higher accuracy, although the gap between the two curves reduces as we allow more time to locate a tag. This result likely depends on Samsung's aggressive strategy (see Figure 4), which allows higher accuracy in more challenging scenarios, e.g., small radius. Both should be using BLE which hypothetically is the same distance. SmartTags advertise that they use Bluetooth 5.0, which has a range of 120 meters, but AirTags do not specify the type of Bluetooth connectivity, although some say its also 5.0. We had attempted to test this out in the parking lot, but it was pretty unreliable even at short distances.

Mobility and Time of the Day – We continue our analysis by exploring the effect of different mobility and temporal characteristics on the accuracy of each tag. For this analysis, we assume a responsiveness of 5 minutes, and report the accuracy across different radii. . to be updated once we know the radii we are interested in. We also compute the statistical significance between different mobility and temporal groups by running t-tests across the average value of the different groups. In Figures 5d-f, statistical significant tests are denoted using the following symbols: ns denotes a $p > 0.05$, * denotes

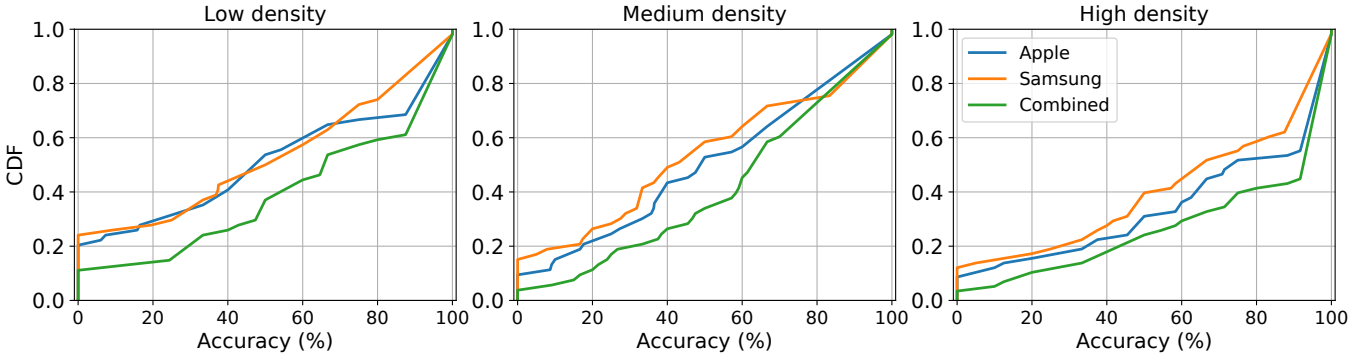


Figure 6: CDF of the accuracy in 1 hour interval at different population densities.

$0.01 < p < 0.05$, ** denotes $0.001 < p < 0.01$, *** denotes $0.0001 < p < 0.001$, and **** denotes $p < 0.0001$.

Figure 5d shows the accuracy as we vary how fast a tag is moving (as per our ground truth). We find that while walking at a pedestrian speed (< 6.0 km/h), the accuracy is maximized for both tags and even when combined. The rationale behind this finding is that walking represents a good equilibrium between number of devices the tag may be exposed to, e.g., higher than when being stationary, and the length of the time window for the BLE signal to be picked up by a eligible device. As the speed increases, e.g., when jogging (speed comprised between 6.0 and 12.0 km/h) or in transit ($\geq$ 12.0 km/h), the accuracy deteriorates due to the little time allowed for Bluetooh communication. will revisit once new analysis is done, same for next two comments.

Figure 5e shows average tag's accuracy –95% confidence intervals reported as error-bars – during different times of the day. The figure shows no significant differences between morning (6 A.M. to 10 A.M.) and lunch hours (10 A.M. and 2 P.M.), but a statistically significant increase in accuracy in the afternoon (2 P.M. to 6 P.M.). A decrease in accuracy is seen in the evening (6 P.M. to 10 P.M.) followed by a further decrease at night (10 P.M. to 2 A.M.). We next explore potential impact of weekdays and weekends on the accuracy. Figure 5f shows significant tag's accuracy increase

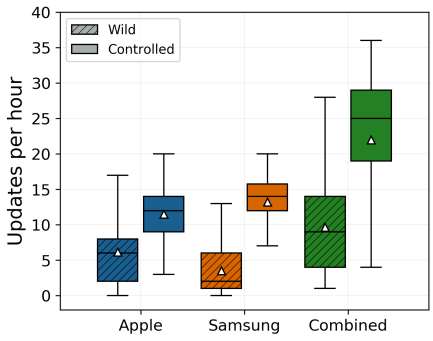


Figure 7: Update rates of the location tags in the different experiments.

on weekends as compared to weekdays, likely due to greater outdoor activity by the general public in the locations visited.

Population Density – Intuitively, the accuracy of a location-tag depends on the number and type of devices in their vicinity. In the wild, this information is unavailable. We thus resort to explore the effect of *population density* using the Kontur data set [3], the rationale being that tag's accuracy increases im more densely populated areas. We threshold the different population density buckets as the 33rd, 66th, 100th percentiles of the population densities of all hexagons visited in our study. As such, we designate hexagons which hold a $population < 600$ (33rd percentile) as "low density", those with a $600 \leq population < 1,750$ (66th percentile) as "medium denisty" and those with $population \geq 1,750$ as "high density".

Figure 6 shows the Cumulative Distribution Function (CDF) of the accuracy at different population densities visited during our in the wild experiments. The figure shows that we received zero Apple or Samsung hits in roughly 20% of low density hexagons visited. The CDF shows that the median accuracy for both AirTag and SmartTag fall around 45%, whereas the combined accuracy is roughly 65%. In medium density hexagons, the likelihood of receiving zero AirTag or SmartTag hits drops below 20%. We do not see an improvement in the median accuracy for both tags in medium density hexagons, with both tags achieving a accuracy of roughly 40 to 45%, with the combined accuracy also remaining at roughly 65%. However, in high density hexagons, we see an improvement in both AirTag and SmartTag accuracy, with the median accuracy reaching 70% and 65% respectively. On average, we see the least combined improvement in high density hexagons (28% improvement), since the likelihood of encountering both an eligible Apple and Samsung device in the same area increases with a higher population density.

Update Rate – To conclude, we report on a tag's update rate. Figure 7 shows boxplots of the update rate for each tag (as well as combined), distinguishing between controlled and in the wild experiments. In the wild, the locations of Samsung SmartTags were reported about three times an

hour, at the median. This update rate is lower than the minimum update rate measured in our controlled experiments, observed when less than 10 Samsung devices were concurrently present in the cafeteria. This implies that a very small number of Samsung devices was encountered "long enough" to provide a tag location update. Note that the controlled experiments purposely represents the least challenging scenario for a location-tag since: i) the tag is not moving, ii) people spend several minutes at the cafeteria, choosing, ordering, and eating, or enough to at least send out one location update. The median update rate measured for AirTags, six updates per hour, suggest that instead at least 10-20 eligible Apple devices are normally encountered in the wild.

6 CONCLUSION AND DISCUSSION

I think we should meet and discuss the takeaways from the paper to guide writing the discussion and the intro. Still working on getting the analysis for a smaller home location radius, will add those figures to the bottom of the paper when it's done.

This paper has provided a novel methodology to measure and analyze the efficacy (accuracy and responsiveness) of location-tags, novel devices which rely on Bluetooth communication to localize lost objects they have been attached to. The key novelty of location-tags – such as the popular AirTags (Apple) and SmartTags (Samsung) – is that they rely on opportunistic encountering of Apple or Samsung devices to anonymously report their last seen location to their owners. Our key idea is to deploy vantage points consisting of a mobile phone, used to report ground truth location data, to which we mounted, via a cover, both an AirTag and a SmartTag. We then used *crawlers* we developed to monitor the tag location as reported by Apple and Samsung devices encountered in the wild. We recruited four to carry our vantage point while travelling; overall, this experiment in the wild covered 9,378 Kms across six countries. The analysis of the data collected shows that location-tags achieve very good accuracy in a number of real-life scenarios, with varying speeds, population densities, times of day, and days of the week. When given a sufficient amount of time to come across an eligible device to relay its location, both AirTags and SmartTags perform similarly, reporting an accurate location roughly 60% of the time. Furthermore, we show that combining the network of users for both tags would allow to increases such accuracy by up to 30%.

We also uncovered key insights in the location-tag update strategies adopted by both vendors, finding that Samsung utilizes a more aggressive and frequent update strategy. Our results show that SmartTags send beacon advertisements more frequently than AirTags, due to SmartTags having fewer eligible devices which can relay their location. Despite Samsung having a larger market share than Apple globally, only a subset of Samsung devices may relay a SmartTag's location if the necessary permissions are provided and the technology needed is supported. While it has been shown in literature that AirTag beacon advertisements are irregular with regards to the intervals between consecutive advertisements [10], our results indicate that, on average, these advertisements are less frequent than those of SmartTags. This comes without a loss in accuracy due to more eligible phones being available to AirTags. Such a strategy by Apple also helps maximize an AirTag's battery life, which has an advertised battery life of over a year, as compared to SmartTag's battery life of 300 days.

The recent emergence of widely-used location-tags has come with a number of concerns. Current considerations made by location-tag manufactures, such as updating privacy warnings [8], have been insufficient in completely deterring unsolicited tracking of people (stalking) [11–13, 15, 16, 19, 21, 23]. First, since each vendor is blind with respect to each other, e.g., Apple fail in accurately reporting stalking via SmartThings. Second, since people not owning a smartphone, e.g., almost 50 millions people in the US only, cannot be notified, Therefore, we believe that tag detection, while an important step in deterring tracking, is not sufficient alone and must be supplemented with further anti-tracking approaches. One such approach is to limit the frequency of updates to the end-user with regards to the location of their tag. Location-tags may still frequently advertise their existence by sending beacon advertisements to nearby devices to share their approximate locations on their behalf, but information relayed from the server to the end user should be optimized to limit the capacity by which one could unsolicitedly track an individual. Our in the wild results have shown that in the case of SmartTags, despite the fact that they had much lower update rates, they did manage to have a similar—if not higher—accuracy in comparison to Apple's AirTags over a longer period of time such as a one hour interval, which is enough for locating lost items. This could be further optimized by segmenting use-cases into "lost" and "found" modes. In the "found" mode, the server would relay a tag's location infrequently, reducing the network load needed, as well as the capacity to track the object/person. However, in the "lost" mode, the frequency of updates to be displayed to the user can be increased, but further security measures could be placed, such as a constant sound played from the tag, and more frequent beacon advertisements which indicate that the tag is lost. We believe that taking such measures would allow these tags to be used for their intended purpose while significantly limiting the ability of any malicious party to misuse them for tracking other individuals.

REFERENCES

[1] Android debug bridge. https://developer.android.com/studio/command-line/adb.
[2] Citi program: Research, ethics, and compliance training. https://about.citiprogram.org/.
[3] Kontur population: Global population density for 400m h3 hexagons. https://data.humdata.org/dataset/kontur-population-dataset.
[4] Pyautogui's documentation. https://pyautogui.readthedocs.io/en/latest/.
[5] Tables of cell statistics across resolutions. https://h3geo.org/docs/core-library/restable/.
[6] Tracker detect - apple. https://play.google.com/store/apps/details?id=com.apple.trackerdetect.

[7] H3: Uber's hexagonal hierarchical spatial index. https://www.uber.com/en-AE/blog/h3/, Jun 2018.
[8] An update on airtag and unwanted tracking. https://www.apple.com/newsroom/2022/02/an-update-on-airtag-and-unwanted-tracking/, journal=Apple Newsroom, Aug 2022.
[9] Apple. Wi-fi privacy - mac address randomization. https://support.apple.com/guide/security/wi-fi-privacy-secb9cb3140c/web.
[10] J. Briggs and C. Geeng. Ble-doubt: Smartphone-based detection of malicious bluetooth trackers. In , pages 208–214. IEEE, 2022.
[11] J. C. J. Dyer. Apple airtags - 'a perfect tool for stalking'. https://www.bbc.com/news/technology-60004257, Jan 2022.
[12] L. Eadicicco. Airtags are linked to stalking, and apple can't solve this problem alone.
[13] G. A. Fowler. Review | am i being tracked? anti-stalking tech from apple, tile falls short., May 2022.
[14] H. Givehchian, N. Bhaskar, E. R. Herrera, H. R. L. Soto, C. Dameff, D. Bharadia, and A. Schulman. Evaluating physical-layer ble location tracking attacks on mobile devices.
[15] D. Grossman. Father sends warning on tracking device after someone put airtag in son's back pocket. https://www.click2houston.com/news/local/2022/10/08/father-sends-warning-on-tracking-device-after-someone-put-air-tag-in-sons-back-pocket/, Oct 2022.
[16] I. A. Hamilton. Thieves are using apple airtags to track luxury vehicles to owners' driveways before stealing them, police say. https://www.businessinsider.com/apple-airtags-thieves-track-steal-luxury-vehicles-police-2021-12.
[17] A. Heinrich, N. Bittner, and M. Hollick. Airguard-protecting android users from stalking attacks by apple find my devices. In , pages 26–38, 2022.
[18] R. Heydon and N. Hunn. Bluetooth low energy. *CSR Presentation, Bluetooth SIG https://www. bluetooth. org/DocMan/handlers/DownloadDoc. ashx*, 2012.
[19] M. Levitt. Airtags are being used to track people and cars. here's what is being done about it. https://www.npr.org/2022/02/18/1080944193/apple-airtags-theft-stalking-privacy-tech, Feb 2022.
[20] T. Mayberry, E. Fenske, D. Brown, J. Martin, C. Fossaceca, E. C. Rye, S. Teplov, and L. Foppe. Who tracks the trackers? circumventing apple's anti-tracking alerts in the find my network. In , pages 181–186, 2021.
[21] A. Moore. 'i didn't want it anywhere near me': How the apple airtag became a gift to stalkers, Sep 2022.
[22] S. Mori, H. Nishida, and H. Yamada. *Optical character recognition*. John Wiley & Sons, Inc., 1999.
[23] M. Peterson. Sports illustrated swimsuit model says she was tracked for hours with airtag, Jan 2022.
[24] Z. Sahinoglu, S. Gezici, and I. Guvenc. Ultra-wideband positioning systems. *Cambridge, New York*, 2008.