Emmanuel Bello-Ogunu* and Mohamed Shehab

Crowdsourcing for Context: Regarding Privacy in Beacon Encounters via Contextual Integrity

Abstract: Research shows that context is important to the privacy perceptions associated with technology. With Bluetooth Low Energy beacons, one of the latest technologies for providing proximity and indoor tracking, the current identifiers that characterize a beacon are not sufficient for ordinary users to make informed privacy decisions about the location information that could be shared. One solution would be to have standardized category and privacy labels, produced by beacon providers or an independent third-party. An alternative solution is to find an approach driven by users, for users. In this paper, we propose a novel crowdsourcing-based approach to introduce elements of context in beacon encounters. We demonstrate the effectiveness of this approach through a user study, where participants use a crowd-based mobile app designed to collect beacon category and privacy information as a scavenger hunt game. Results show that our approach was effective in helping users label beacons according to the specific context of a given beacon encounter, as well as the privacy perceptions associated with it. This labeling was done with an accuracy of 92%, and with an acceptance rate of 82% of all recommended crowd labels. Lastly, we conclusively show how crowdsourcing for context can be used towards a user-centric framework for privacy management during beacon encounters.

Keywords: Internet of things; Bluetooth low energy; beacons; location privacy; crowdsourcing

DOI 10.1515/popets-2016-0017
Received 2015-11-30; revised 2016-03-01; accepted 2016-03-02.

1 Introduction

Contextual integrity is a concept that suggests that people do not require absolute privacy but rather privacy that meets certain expectations and social norms. Research conducted by Helen Nissenbaum shows that contextual integrity is important to the privacy perceptions associated with technology; these perceptions are based on four elements: the context of a flow of information, the capacities in which those involved are acting, the type of information involved, and the principles of transmission [1]. We derive from this research an emphasis on the context-dependence of privacy concerns, and we turn to one of the latest technologies for providing proximity and indoor tracking: Bluetooth Low Energy (BLE) beacons. BLE beacons are small Bluetooth sensors that are used to provide precise location and contextual cues about users' interactions with the real world [2]. However, the current identifiers that characterize a beacon encounter are not sufficient for ordinary users to make informed privacy decisions about the location information that could be shared for each encounter. We envision a way to empower users with the means to control their privacy in beacon-enabled spaces.

For example, imagine a college student who usually visits the men's athletic wear section of the campus bookstore. This same student frequently visits the glassware section near the shot glasses. From the frequency of aisle visits and duration of stay, it may be inferred that the student is probably a male athlete but also potentially a heavy drinker or frequent partier. This results in sensitive information that reveals intimate and personal facts about the student, which he will likely be uncomfortable sharing, as it shows behavior conflicting with their role as an athlete and what is expected of them [3]. Consequently, the student should be equipped with a mechanism to manage his information privacy when it comes to beacon encounters. A better understanding of the context of each encounter, and how the related information may be shared, is required before they can effectively manage their privacy.

One solution would be to have standardized category and privacy labels associated with beacons, generated by beacon providers or an independent third-party. However, it would be difficult to ensure that all beacon providers abide by this policy. Moreover, their regard for privacy will likely differ from that of their users. Therefore, a novel approach is required to provide the necessary context, allowing users to form accurate mental

***Corresponding Author: Emmanuel Bello-Ogunu:** UNC Charlotte, E-mail: ebelloog@uncc.edu
Mohamed Shehab: UNC Charlotte, E-mail: mshehab@uncc.edu

models of beacon encounters before exercising appropriate privacy-preserving behaviors. This is where we provide a solution that does not previously exist. By designing, implementing, and deploying a beacon privacy manager that relies on crowdsourcing, we empower users who have encountered beacons to contribute their understanding of the related context for other users to leverage in order to make informed privacy preserving decisions regarding what to share with beacon enabled mobile applications.

In this paper, we propose a crowdsourcing-based approach as a plausible solution to introducing elements of context in beacon encounters, through user-provided category and privacy labels. We demonstrate the effectiveness and usability of this approach through a user study where participants use a mobile app that is designed to collect these labels from the crowd during a scavenger hunt. These labels reflect users' perceptions of the beacons they encounter based on the context of the flow of beacon-related information from them to various audiences, as well as the capacities in which participants were willing to share. Our results show that our approach was effective in helping users label beacons, according the specific context of a given beacon encounter and the privacy perceptions associated with it. This was done with an accuracy of 92%, and with an acceptance rate of 82% of all recommended crowd labels. With confidence in these results, this crowdsourcing for context will be used towards a user-centric framework for privacy management during beacon encounters.

The remainder of this paper is organized as follows: Section 2 discusses the Bluetooth Low Energy beacon technology in detail and reviews its current applications. In Section 3 we provide some background information on the concept of crowdsourcing and discuss related work in the areas of crowdsourcing and privacy. In Section 4 we introduce the study design used to demonstrate the effectiveness of a crowdsourcing-based approach, in the form of an Android app we created called "BknBkts." In Section 5 we discuss the results of the user study conducted and the implications of the results, and lastly, Section 7 discusses future work and concluding remarks.

2 BLE Beacons

Bluetooth beacons are based on Bluetooth Low Energy (BLE), or "Bluetooth Smart," which is part of the Bluetooth 4.0 specification [4]. Standard Bluetooth [5] is widely used in cars, audio equipment, mobile phones, and other technology for the purpose of transmitting large pieces of information up to approximately 100 meters. In fact, about 90% of all mobile phones sold today are Bluetooth-enabled, according to Bluetooth SIG [6]. The main focus of BLE, however, is delivering smaller amounts of data with low energy consumption through a reduced transfer rate. This minimizes the impact on a device's battery life. The BLE protocol consists of two main types of communication, namely advertising and connecting. Advertising is a one-way communication discovery mechanism, where devices that wish to be discovered transmit packets of data, up to 47 bytes, in intervals from 20 milliseconds to 10 seconds. BLE beacons only use the advertisement channel: they transmit data at regular intervals and advertise their presence, and this data is received by other devices, such as smartphones, as pictured in Figure 1. All this is done without the need to pair with other devices, as would be the case with standard Bluetooth.

Table 1. Sample Beacon data packet

Store location		San Francisco	Paris	London
UUID		D9B9EC1F-3925-43D0-1E39D4CEA95C		
Major		1	2	3
Minor	Clothing	10	10	10
	Shoes	20	20	20
	Electronics	30	30	30

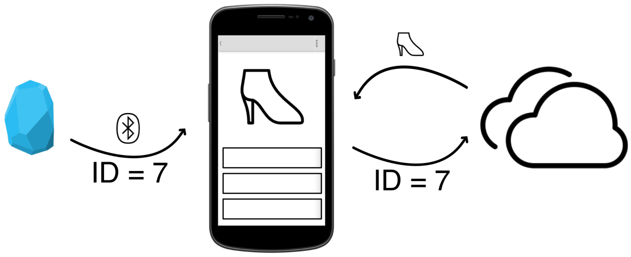


Fig. 1. A smartphone detecting a beacon using BLE technology

The format of the BLE data packet that is advertised contains a beacon ID, which is 20 bytes long and divided into three sections. Table 1 represents an example of what an advertised beacon ID would contain. There is the proximity universally unique identifier (UUID), a 16 byte identifier which is used to distinguish beacons at a specific location. For example, the beacons belonging to a store chain will all have the same proximity UUID value. The Major number (2 bytes) follows the UUID and is used to group a related set of beacons; as an example, all beacons in a specific store would be assigned the same Major number. Lastly, the Minor number (2 bytes) is used to identify individual beacons; for example, each beacon in a store will be assigned a different Minor number. In addition to the beacon ID, there is also TX power, or the Received Signal Strength Indication (RSSI), which is the strength of the beacon signal measured at 1 meter from the beacon device. The TX power and the power measurement at the receiver are used by the receiver to estimate the proximity or distance of the beacon from the smartphone or other BLE-enabled receiver. This proximity can be determined using one of two methods:

- *Beacon monitoring*, which is where the entry and exit of beacon regions is measured; this can happen while the app is running in the foreground, background, and even when the app is killed.
- *Beacon ranging*, where distances between beacons is calculated. This works only when the app is in the foreground.

2.1 Beacon Encounters

When a user is within the proximity of a broadcasting beacon, the user's smartphone receives the beacon UUID advertisement, which is subsequently sent to the server via the apps registered to listen to beacon advertisements on the smartphone. In addition, the receiver is able to provide an estimate of the distance from the beacon based on the receiver's power. Figure 1 shows the interaction between the three main components of beacon technology, namely the beacon, a user's smartphone, and an online server. In this architecture, the main computing device is the user's smartphone, which has the beacon enabled mobile app installed and is connected to the internet to send the user's beacon encounter information to a server. Without the beacon-enabled app that is configured to detect beacons, a mobile device would not be affected by nor do anything with the received BLE advertisements. The beacon architecture allows the beacon owner to know the beacons encountered by given users, the proximity to these beacons, and the temporal behavior of the users in the beacons' proximity. In addition, the provider/retailer is able to track users' movement in the beacon-equipped space by correlating the beacon encounters in the store over a period of time.

The main metrics derived from beacon encounters can be summarized as follows:

- Activity Path: How does the customer navigate in the store? The user's path in a region can be estimated by aggregating and triangulating multiple beacon encounters. This information is useful to retailers when designing product placement to ensure marketing and advertisement objectives.
- Activity Time: How long was the customer engaged in a particular activity? This information is estimated by aggregating the same beacon encounters and calculating the duration of time the user spends in the proximity of a specific beacon; this metric's accuracy is dependent on the beacon advertisement frequency and scanning rate of the smartphone.
- Visit Frequency: How often does the customer engage in a particular activity? This information is derived by computing the number of unique beacon encounters that occur per consecutive time frames. For example, it can be used to estimate the number of times the user has visited the store in a given month.
- Core Actions: Does the customer engage in actions that indicate they have adopted an idea, product, brand? For example, by performing causality analysis, it can be deduced what the user does after a beacon encounter. Did the user buy the product? Did she visit the mobile app?

Based on these metrics, a beacon provider or retailer could infer things like personality traits, gender, ethnicity and economic status based on frequency of visits to certain locations, dwell times, and purchases associated with beacon encounters.

3 Related Work

Crowdsourcing can be defined as everyday people using their spare cycles to create content, solve problems, and even do corporate R&D [7]. Rather than soliciting contributions from traditional employees or workers, crowdsourcing relies on a large number of average users, usually recruited via social networks or open calls online, to work together towards a common goal. Projects such as Wikipedia [8], Amazon Mechanical Turk [9], SETI@Home [10], and Threadless [11] are all the result of relying on the collective intelligence and input of the crowd to address a broad array of purposes. Crowdsourcing successfully demonstrates how large, loosely organized groups of people can use technology to contribute individual effort to address a larger purpose in surprisingly effective ways [12].

In the domain of security and privacy, crowdsourcing has seen similar effectiveness [13–16]. For example, Burguera et al. [17] utilized crowdsourcing to capitalize on dynamic analysis of application behavior to detect Android malware. Using their lightweight client called "Crowdroid," they were able to collect traces of applications' behavior-related data from real users, with experimental results showing that the system was able to provide a 100% detection rate for self-written malware, and 92.5% for two real malware specimens. Additionally, Lin et al. introduced a model for privacy, namely *privacy as expectations*, where crowdsourcing was employed through Amazon Mechanical Turk to capture users' expectations of what sensitive resources mobile apps tend to use, including device identifier, address book, network location, and GPS location [18]. Identifying expectation and purpose as two key factors that affect users' mental model of app privacy, Lin et al. go one step further to design and evaluate a privacy summary interface for Android apps that emphasizes an app's behaviors which do not align with the crowd's expectations. They were able to show that this interface was both more accurate and more efficient than the default Android permission interface in making users aware of the related privacy concerns. Lastly, Agarwal and Hall developed a system for iOS devices to detect an app's access to private data, such as device identifier, location data, address, and music library, and protects users by substituting anonymized data [19]. Their system used a crowdsourced recommendation approach to provide app specific privacy recommendations and was able to recommend settings for over 97.1% of the 10,000 most popular apps. Its effectiveness was also asserted through the acceptance of 67.1% of all privacy recommendations by users.

From the related work it is evident that the interaction of crowdsourcing and privacy is plausible, particularly with other tracking technologies. However, regarding beacons, the research presented here is the first of its kind. There is a need for this solution given the inherently context-driven nature of the technology. Since beacon providers decide the context of each beacon encounter [2], it is imperative that users are able to form accurate mental models of these encounters and exercise appropriate privacy-preserving behaviors when necessary. Our main objective is to build a privacy manager that leverages these mental models and equips users with policies that enforce the desired privacy-preserving behaviors based on the context of an encounter, characterized by more meaningful identifiers like the location or type of beacon, and the study we conducted is a critical first step towards this objective.

4 Study Design

For this study, we set out to determine whether users can come to a clear consensus through crowdsourcing regarding the specific context of a given beacon encounter, as well as the privacy perceptions associated with it. We believe users can leverage that consensus to aid them in making an informed privacy decision regarding future beacon encounters. In order to evaluate our problem statement, we conducted a between-subjects study involving the use of an Android mobile app we created called Beacon Buckets ("BknBkts"). The study took place in the on-campus bookstore, where we setup Estimote [20] beacons. Participants were instructed to find them in a scavenger hunt fashion using the mobile app, and label them based on the category of items closest to them. Figure 2 indicates the sections of the bookstore where beacons were placed, each of which represented one of the following categories: Women's Athletic Apparel, Magazines, Polo Ralph Lauren Apparel, Shot & Drinking Glasses, Clearance, Health & Beauty, Starbucks, Restrooms, and ATMs (Automated Teller Machine).

Within the app, users were programmatically assigned in round-robin fashion to one of three conditions: either they were recommended the correct category associated with each beacon, the most popular categories selected for each beacon as crowdsourced by other users from all previous sessions, or they were not provided with any recommended labeling. Table 2 summarizes these study conditions: we represented the group of participants who were provided the correct category associated with each beacon as "TopCat," while the group provided with the most popular crowdsourced categories as "CrowdCat," and those not provided with any recommended categories as "NoCat." Decisions of the participants in the other groups, TopCat and NoCat, were not taken into account when providing CrowdCat suggestions, therefore preventing any leakage of correct category labeling. Participants used the information they were given to categorize each beacon and answer a few privacy-related questions for each.

The study was conducted over a span of three weeks. Random passers-by and visitors to the bookstore served as the sample for the study. All participants were age 18

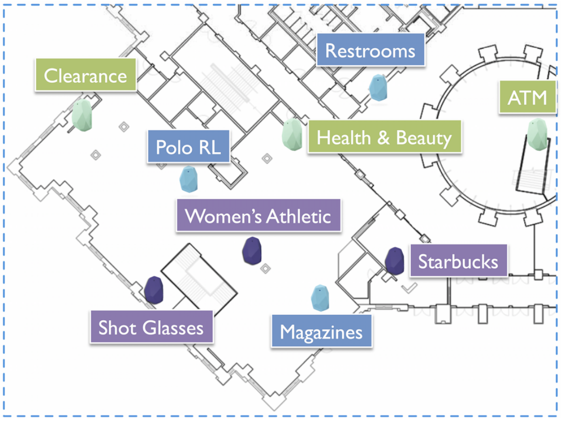


Fig. 2. Map of beacon placement in campus bookstore

Table 2. Study Conditions

Condition	Recommendation
TopCat	Correct category labels
CrowdCat	Crowdsourced category labels
NoCat	None

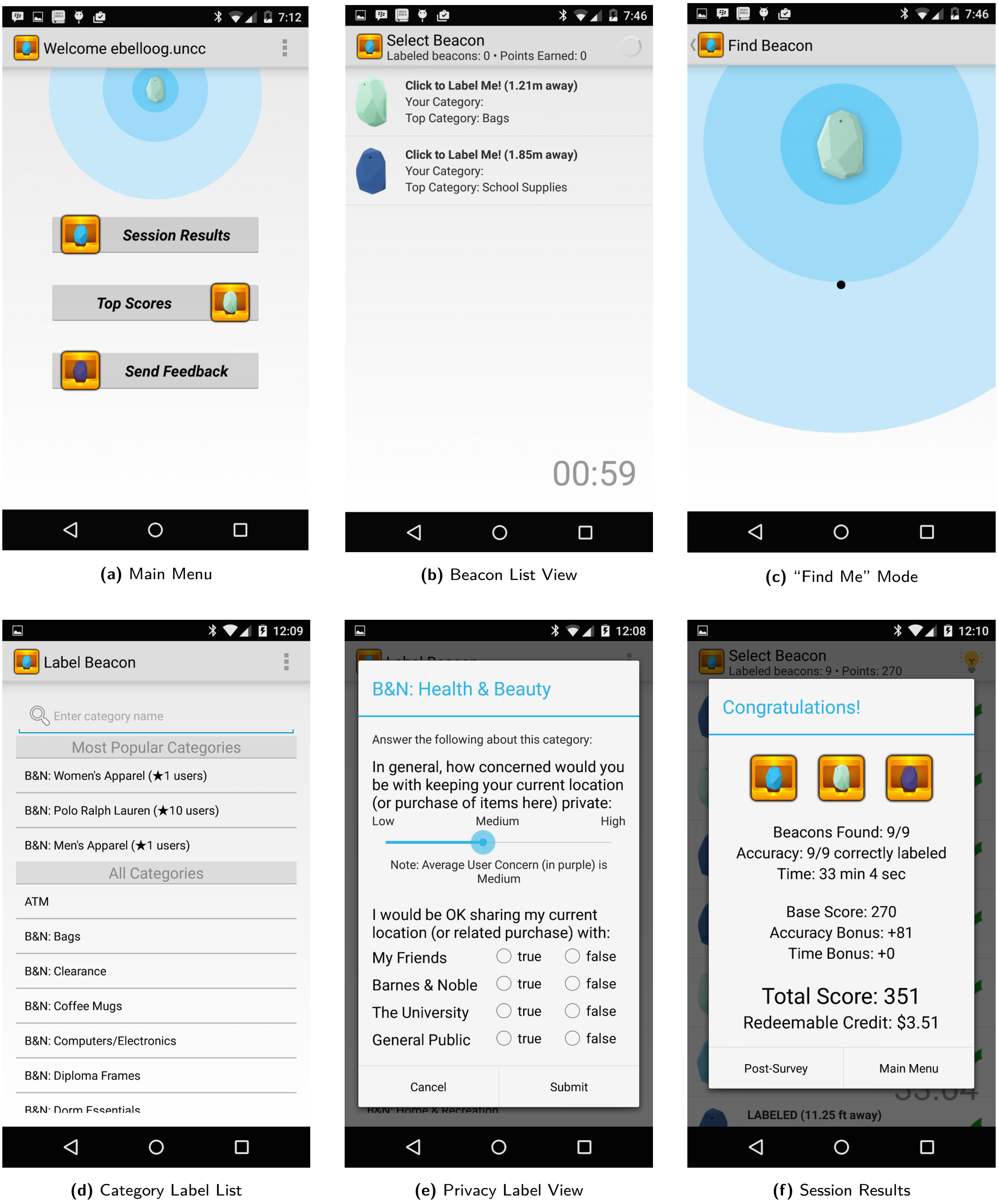
or older, and although the majority were students at the University, it was not a requirement to participate. The study was advertised as a beacon scavenger hunt game, with a reward in the form of a $5 Starbucks gift card for anyone that played. Participants with a compatible Android device installed the app on their phone; otherwise, they were given a loaner device. The scavenger hunt took on average 30 minutes to complete.

4.1 User Study Flow

Participants began the user study by registering their profile in the BknBkts app, and proceeded through a brief tutorial with accompanying screenshots depicting how to use the app. Figure 3 represents screenshots of the app, similar to what users would view during the tutorial. After registering, users were brought to the main menu, as seen in Figure 3a, where they would enable Bluetooth on the device and start the session. From here, users could view a dynamic list of BLE beacons within range of the device, shown in Figure 3b, sorted in order of distance from the user. From this view, the user also had the option to access hints for where beacons might be placed around the bookstore. Selecting a beacon from the list provided two options: "Find Me," which displayed the radar view shown in 3c to help users determine their proximity to an unseen beacon, and "Label Me," which displayed a list of categories, as shown in Figure 3d, from which to choose and associate a beacon. Users were advised to visually locate the beacon in question before labeling it, to minimize incorrect labeling. The app actually required users to be within approximately 10 feet for the beacon to show up in the list of available beacons to label, and it was designed this way to limit the interference from other beacons when one was within view.

When choosing from a list of categories for a given beacon, participants in the "TopCat" group were presented a subsection for an assigned "Top Category." These participants were made aware through verbal instruction that the one category at the top represented a label as if provided by the bookstore itself. The "CrowdCat" participants were shown the three "Most Popular" categories, which were crowdsourced from other participants' responses. These users were informed that the "Most Popular" category labels were the most popular choices for a beacon that were crowdsourced among participants across all previous sessions. The first CrowdCat user actually received no recommendation. All subsequent CrowdCat users received the top recommendations at the time (in particular, the second user is recommended whatever the first user chose, and so on), which may include ties. The "NoCat" group did not see anything other than the standard list of all categories.

Figure 4 represents the view that participants were shown based on their condition. While on this view, the app kept track of the amount of time taken to select a category, as well as whether the selected category was a recommended one, if applicable. The last step required users to provide a privacy label for a beacon, which involved a few questions, as shown in 3e, regarding the "sensitivity" based on the selected category. In this scenario, a "sensitive" beacon was described as "worth keeping my presence here, or purchase of related items here, as private." Therefore for each beacon, users indicated the level of concern associated with the privacy/sensitivity of the beacon using a slider on a scale from 0 to 100. Here, those in the CrowdCat group would see "Average User Concern" level represented on the slider. This was done to again provide crowdsourced feedback on what other users had prescribed as a sensitivity level, even though there were no a priori "correct" privacy labels. Participants also indicated which circles of people they would be willing to share this location with, including friends, the bookstore, the university, and general public, using radio buttons. These last few questions that pertained to the privacy label represented other levels of context to consider in location information sharing for each beacon found.



Select Beacon
Labeled beacons: 0 • Points Earned: 0

Click to Label Me! (1.21m away)
Your Category:
Top Category: Bags

Click to Label Me! (1.85m away)
Your Category:
Top Category: School Supplies

00:59

Label Beacon

Enter category name

Most Popular Categories

- B&N: Women's Apparel (★1 users)
- B&N: Polo Ralph Lauren (★10 users)
- B&N: Men's Apparel (★1 users)

All Categories

- ATM
- B&N: Bags
- B&N: Clearance
- B&N: Coffee Mugs
- B&N: Computers/Electronics
- B&N: Diploma Frames
- B&N: Dorm Essentials

B&N: Health & Beauty

Answer the following about this category:

In general, how concerned would you be with keeping your current location (or purchase of items here) private:

Low — Medium — High

Note: Average User Concern (in purple) is Medium

I would be OK sharing my current location (or related purchase) with:

My Friends	○ true	○ false
Barnes & Noble	○ true	○ false
The University	○ true	○ false
General Public	○ true	○ false

Cancel | Submit

B&N: Home & Recreation

Select Beacon
Labeled beacons: 9 • Points: 270

Congratulations!

Beacons Found: 9/9
Accuracy: 9/9 correctly labeled
Time: 33 min 4 sec

Base Score: 270
Accuracy Bonus: +81
Time Bonus: +0

Total Score: 351
Redeemable Credit: $3.51

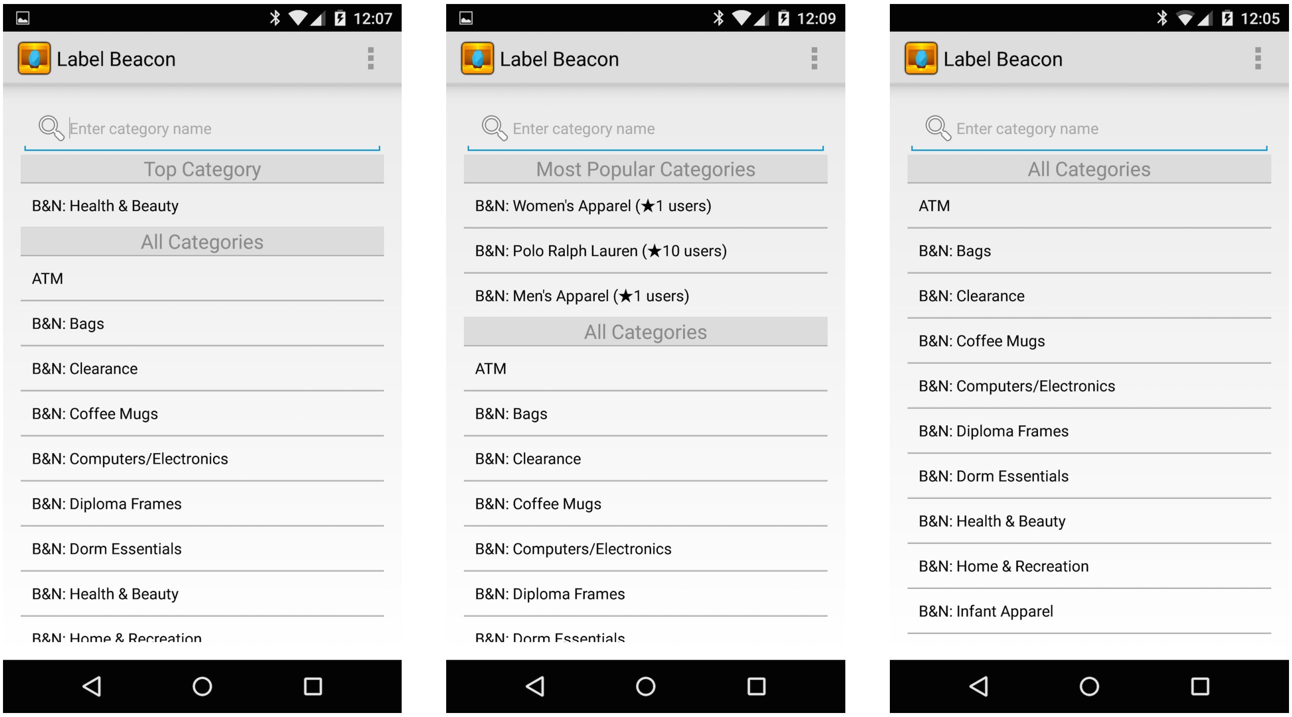
Post-Survey | Main Menu

LABELED (11.25 ft away)

(a) Main Menu (b) Beacon List View (c) "Find Me" Mode

(d) Category Label List (e) Privacy Label View (f) Session Results

Fig. 3. Screenshots of "BknBkts" Android app



(a) TopCat **(b)** CrowdCat **(c)** NoCat

Fig. 4. Category list views presented to users based on their study condition

Users would repeat this process for each beacon they encountered as they explored the bookstore. Should they determine that they had inaccurately labeled any beacon, or they wanted to change their responses to the sensitivity-related questions, they simply had to click on the labeled beacon to redo the process. Since the app was designed as a Game With A Purpose [21], where every beacon labeled earned users a number of points contributing to a base score, this motivated users to provide truthful labels and gamified the labeling process. Additionally, users earned bonus points for labeling accuracy, as well as for completion time. The total possible score that could be earned was 500 points, which translated to the $5 gift card reward for participation. Although the app was structured this way to further incentivize participants to do their best in labeling, in the end all participants were equally paid a $5 gift card. Once all nine beacons had been found and labeled, or if the participant chose to quit early, the app would trigger the conclusion of the session and present the user with a report of the score earned, as seen in 3f. From the main menu, users were able to see a leaderboard of the top ten scorers and compare their scores with those of other participants. The purpose of the leaderboard was simply to contribute to making the app more like a game, and therefore incentivize users to complete the study. Participants were not able to redo the study in an effort to improve their score, and while it is possible that participants could have encouraged their friends to take part in the study and coach them on the correct answers, this was not something for which we controlled.

4.2 Analysis Procedure

In this study, our main focus for this analysis was the effectiveness of crowdsourcing as a method to incorporate user input and introduce levels of context regarding beacon encounters. This effectiveness is represented by the accuracy (correctness) and efficiency (time or user burden) of beacon category labeling. We also report on the percentage of recommended category labels that are selected by participants assigned to the respective conditions. Furthermore, we analyze the privacy labels that participants contribute, in order to test for significant differences between the privacy labels based on the beacons themselves.

5 Descriptive Results

We analyzed the app usage from a total of 90 participants. The demographic breakdown of these participants can be seen in Table 3. Males made up 65.6% of participants and females were 34.4%. Regarding age, 62.2% were between 18-24, while 32.2% were ages 25-34, and the remaining 5.6% fell between 35-54. Lastly, the main levels of education that users completed included Some college, with 43.3% of participants, Bachelor's degree at 22.2% of participants, and Graduate degree at 26.7%.

Table 3. Participants' demographic summary

Gender	%	Age	%	Education	%
Male	65.6	18-24	62.2	Some college	43.3
Female	34.4	25-34	32.2	Graduate	26.7
		35-64	5.6	Bachelor	22.2
				High school	4.4
				Associate	3.3

5.1 Accuracy of Crowdsourcing

Of the 90 participants that were recruited, 30 were grouped into each condition, and the mean accuracy for each group is represented in Table 4. The TopCat group showed the highest accuracy of all the participants, at 94.074%, while the CrowdCat group had an accuracy of 92.592%, and the NoCat group was the least accurate in labeling, at 86.667%. This shows that without any recommended labels, participants are not as accurate, and it would appear that crowdsourcing labels provides an accuracy that is close to that of exact category recommendations. Furthermore, we note that the TopCat group arrived at 94% accuracy instead of 100%, even though they were informed that the one category at the top of their list represented a label generated by the bookstore itself, and therefore was the most correct one. This was for a number of reasons. First, participants were still free to choose whatever label they believed was correct, and in some cases, they did chose an incorrect label. Additionally, some beacon categories were more prone to erroneous labeling, due to their placement; for example, not all beacons were eye-level, and some were placed in between sections. Still, for those in the TopCat group, who were given the correct answer, we anticipated that they would not be as prone to incorrect labeling with these. Lastly, a few participants were not able to find and label all the beacons during their session; in the TopCat group, however, out of total possible 270 labels, 269 were provided, meaning only one label was missing from a single participant in this group. In the CrowdCat group, two labels were missing, and one label in the NoCat group. Hence, the major contributing factor that led to the 6% error in this condition was incorrect labeling.

Table 4. Mean Accuracy Percentage Per Condition

Condition	Mean (μ)	St. Dev	N
TopCat	94.074	9.103	30
CrowdCat	92.592	9.823	30
NoCat	86.667	10.275	30

In order to compare the effect of condition on the mean accuracy for the three groups, we used a one-way ANOVA test. This revealed a significant effect of condition on accuracy, with F(2,87) = 4.853, p = .010. Table 5 reflects the results of the post hoc comparisons performed, using the Tukey HSD test, with statistically significant results yielding a p-value less than .05. We found that there was no significant difference between the accuracy for the TopCat group and CrowdCat group, but there was between CrowdCat and NoCat (p = .05), as well as between TopCat and NoCat (p = .011). Logically we would expect the TopCat group to perform better in labeling beacons than the NoCat group, which was not given any labeling recommendation. However, these results also confirm that crowdsourcing is more effective in accuracy than no recommendation at all, and that the crowd can be relied on to provide categorizations that are comparable in accuracy with the exact categories.

Table 5. Mean Accuracy Post Hoc Comparisons

Condition A	Condition B	μ_A-μ_B	p-value
TopCat	CrowdCat	1.481	.827
CrowdCat	NoCat	5.925	.050
TopCat	NoCat	7.407	.011

5.2 Time Efficiency of Crowdsourcing

In addition to accuracy, the other aspect of effectiveness that we sought to prove regarding our beacon labeling crowdsourcing approach is efficiency. This was represented by the average amount of time users took to complete a labeling task for a beacon. Table 6 captures the mean time for each of the three groups, where N is the number of beacon labels captured that were attributed to that group. The TopCat group, as expected, had the fastest mean time, of 6.662 seconds, while the CrowdCat group had a mean label time of 8.349 seconds. The NoCat group had a mean label time of 11.280 seconds, meaning those who were not recommended any label had the slowest labeling time.

Table 6. Mean Label Time Per Condition, in Seconds

Condition	Mean (μ)	St. Dev	N
TopCat	**6.662**	**13.618**	**269**
CrowdCat	**8.349**	**18.440**	**268**
NoCat	**11.280**	**14.546**	**269**

Since the time measurements are not an unbounded normal variable, we performed a logarithmic transformation on the time, and then performed a one-way ANOVA to compare the effect of the study condition on time. This revealed a significant effect of condition on time, with F(2,87) = 9.535, $p < .001$. Table 7 reflects the post hoc comparisons done using the Tukey HSD test. The results were similar to the findings related to labeling accuracy: we found that there was not a significant difference between the time efficiency for the TopCat group and CrowdCat group, but there was a significant one between CrowdCat and NoCat (p = .024), as well as between TopCat and NoCat ($p < .001$). Taken together, this means that the crowd can again be relied upon to quickly determine beacon labels with an efficiency that is comparable to users who are given the correct category. Another way to consider this outcome is that the user burden for crowdsourcing is sufficiently acceptable compared to being recommended the category.

Table 7. Mean Log Time Post-Hoc Comparisons

Condition A	Condition B	μ_A-μ_B	p-value
TopCat	**CrowdCat**	**-.1211**	**.223**
CrowdCat	**NoCat**	**-.1978**	**.024**
TopCat	**NoCat**	**-.3190**	**.000**

With confidence in the accuracy and efficiency of the crowdsourcing approach established, we examine the acceptance of the crowdsourcing approach: of the 268 recommended labels generated for the CrowdCat group, 82% of the recommendations were taken by those participants. What we mean here is that 82% of participants in this group accepted some label recommendation that was provided by the crowd. This was not necessarily the response that was selected the most by the crowd. As Figure 4b indicates, labels recommended via the crowd had an accompanying count of users that chose this label. Compare the CrowdCat group's acceptance rate to the 94% acceptance of recommendations by users in the TopCat group. Again, one might expect the acceptance rate of the TopCat to be 100%, but as mentioned earlier, participants in this group did not always chose the top category, for various reasons. Participants were still able to make a selection beyond the recommendation, by choosing from the entire list. We observed that 3% of TopCat users noted the recommendation, but still searched through the whole list for a correct label, and though they may have ultimately selected the correct label, it was done so by choosing it from the main list, instead of choosing the recommended label at the top of the list.

5.3 With Privacy, Context is King

Looking at the privacy labels generated by users, we recall that these reflected the perceived sensitivity of a beacon based on its assigned category label, as well as users' willingness to share their location with various circles of people. By averaging these responses, we get an aggregate view of the privacy labels, which is represented in Table 8. In this table, the Sensitivity column represents a rating between 0 and 100, where 0-32 is be considered Low Sensitivity and therefore of minimal privacy concern, 33-65 as Medium, and 66-100 as High and therefore of highest privacy concern. Additionally, the percentages under the "Share with" columns in the Table represent the amount of users who were willing to share their presence at that particular beacon location with the corresponding social circle: Friends, the University, the Bookstore, or the General Public.

Based on Table 8, we see that the Sensitivity is highest for the ATM beacon, followed by the Restrooms beacon, which had an average rating of 87.84 and 71.88 respectively. Furthermore, the percentage of users willing to share their location here was lowest across the various social circles for these two beacon categories as compared to the others, as we expected. On the other end of the spectrum is the Starbucks beacon, which had a mean Sensitivity rating of 46.10. Even with this rating of

Table 8. Averages for Privacy Label Responses per Beacon Category: Sensitivity is on a scale from 0-100, and Sharing represents percentage of users willing to share; A plus (+) represents significantly different from ATM, and an asterisk (*) represents significantly different from Starbucks.

Beacon	Sensitivity	Share w/Friends	Share w/Bookstore	Share w/University	Share w/Public
ATM	**87.84**	**44%***	**21%***	**30%***	**13%***
Restrooms	**72.00**	**53%***	**31%***	**35%***	**24%***
Health & Beauty	**64.37**	**73%$^{+*}$**	**68%$^{+}$**	**47%***	**33%$^{+*}$**
Shot Glasses	**63.57**	**70%$^{+*}$**	**68%$^{+}$**	**46%***	**36%$^{+*}$**
Ralph Lauren	**49.73**	**83%$^{+}$**	**87%$^{+*}$**	**71%$^{+}$**	**50%$^{+}$**
Clearance	**48.99**	**76%$^{+*}$**	**83%$^{+*}$**	**66%$^{+}$**	**51%$^{+}$**
Women's Athletic	**47.13**	**75%**	**70%**	**57%**	**38%**
Magazines	**46.23**	**84%$^{+}$**	**86%$^{+*}$**	**70%$^{+}$**	**55%$^{+}$**
Starbucks	**46.10**	**91%$^{+}$**	**62%$^{+}$**	**71%$^{+}$**	**58%$^{+}$**

Medium Sensitivity, it was the lowest of all the beacons, which was again as we expected. Similarly, the "Share with" percentages were the highest across most of the social circles for this beacon, with a reported 91.11% of participants willing to share their presence here with Friends, the highest amount of all the beacons.

For the remaining beacon categories, we observed a variation of sensitivity in the privacy labels reported. We note that none of them were regarded as Low Sensitivity beacons. Shot Glasses and Health & Beauty were the two categories with the next highest Sensitivity ratings, at 63.57 and 64.37 respectively. This would place them on the higher end of the Medium Sensitivity range, bordering High Sensitivity. For the Shot Glasses beacon, we believed it was perceived this way because of the association with alcohol and drinking; any indication of visiting this section of the bookstore too often might suggest the user engages in frequent drinking, which tends to have negative connotations. Concerning the Health & Beauty category, we found that the Bookstore sold products here that included condoms, feminine care, and other personal hygiene items. It makes sense that users would consider this a sensitive beacon, as it is unlikely that they would be comfortable sharing their presence or purchase of these kinds of products with many others.

Lastly, the beacons with categories Clearance, Magazines, Polo Ralph Lauren, and Women's Athletic all had a Medium Sensitivity rating that was similar to that of the Starbucks beacon, ranging between 46.23 and 49.73. Their respective percentages for users' willingness to share location information with different social circles were also comparable as well. This also makes sense, as three of the four categories were clothing-related, and furthermore, given the type of clothes and magazines sold at the campus bookstore, most of the related items nor a known proximity to these items would likely pose a privacy threat to visitors.

For these beacons that were observed, the ATM beacon was used as a "ground truth" for what we hypothesized could be regarded the most sensitive beacon, because we believed users would not feel comfortable sharing every time they visited an ATM, likely to withdraw money. It could be further hypothesized that this comes from a sense of physical safety/security when carrying money on themselves. On the other hand, the Starbucks beacon was used as the benchmark for the least sensitive beacon, given how public of a location Starbucks is generally considered. The remaining beacons could then be compared against these ground truths to determine what kind of privacy concern users associated with a beacon based on its category when sharing their encounter, as well as potential audiences of that information (Friends, Bookstore, University, and Public). Using the established ground truth beacons, our null hypothesis was that probability of a "Yes" response in willingness to share was the same for all categories of beacons, and our alternative hypothesis is that the probability that participants responded with a "Yes" in willingness to share was different depending on beacon, where they are more likely to respond "No" for the ATM beacon and more likely to respond "Yes" for the Starbucks beacon, in each of the different audiences.

In order to prove that these differences in privacy labels supported the existence of context as an influencer of privacy perceptions in our study, we relied on statistical analysis to determine significant results. To do this, we used the Cochran Q test for k related samples. This provides a reliable way to test whether multiple matched sets of frequencies differ significantly among themselves. In this design, our "k" related samples are the beacons, and they are considered matched because each participant provides a response on willingness to share for each beacon, and for each audience. This test is particularly useful when the data are categorical or nominal, which is the case with our beacon data; the response for willingness to share is dichotomized as "Yes" or "No." With this test, the number of individual responses in each of the matched response sets (N=59) was less than the total 90 participants, because the test analyzes only the complete matched sets, those being where a response was provided for all beacons. Not every participant provided 9 beacon labels, and not every participant that did so was able to correctly identify the 9 beacons in question; hence, the test dropped the Women's Athletic Apparel beacon, the beacon that was missed the most (either incorrectly labeled or not labeled at all), and considered "k" to be the 8 remaining beacons, thereby causing N to be 59.

Looking at participants' willingness to share beacon location information among the Friends audience first, we found that the Cochran's Q test resulted in Q = 76.245, with a p-value less than .001. With an alpha of .05, we rejected our null hypothesis in favor of the alternative hypothesis. Similarly, making the same comparisons within the Bookstore audience, we generated Q = 129.284 with a p-value less than .001. For the University audience, the test resulted in Q = 87.062, p-value less than .001. Lastly, for the Public audience, we generated Q= 78.821, with p less than .001. Consequently, on the basis of these data, we conclude that the probability that participants are willing to share their presence at a beacon differs significantly between the beacons, for each audience.

We conducted multiple post hoc pairwise tests using the McNemar test to determine where the difference was in willingness to share for different beacons for each audience. Cochran's Q test is an extension of McNemar, which is a nonparametric test specifically for two-sample cases, making the latter an appropriate test for post hoc comparisons. For the purpose of simplicity, instead of conducting McNemar on every pairing of beacons, we only used our ground truth beacons of ATM ("most sensitive") and Starbucks ("least sensitive"), and we conducted the McNemar test between each of these beacons respectively and the remaining seven beacons. This resulted in 14 pairwise comparisons under each audience, seven with the ATM beacon and seven with the Starbucks beacon. Our alpha was .0035 (.05/14), which was Bonferroni corrected, to counteract the problem of multiple testing. The pairings that resulted in significant differences are represented in Table 8, where a plus (+) represents significantly different from ATM, and an asterisk (*) represents significantly different from Starbucks. Through examination of crosstabs for each comparison, we confirmed that participants were more likely to respond "No" in willingness to share for the ATM beacon than any beacon from which it was significantly different, and "Yes" for the Starbucks beacon versus any others.

With these results, we confidently assert that context does impact the privacy perceptions of users in their willingness to share beacon information.

6 Discussion

When analyzing the value of adopting a crowdsourcing approach to introducing context in beacon encounters, we recall that no significant difference was found between the acceptance rates of recommendations in the TopCat group and the CrowdCat group. This shows that users sufficiently relied on the provided recommendations in both groups. While CrowdCat performed at an equivalent level as the TopCat group in accuracy, efficiency, and acceptance rate, it is still of greater value implementing a crowdsourcing system, for two main reasons. First, it is difficult to compel every retailer or beacon provider to go through the extra work of providing these category labels for the beacons they deploy, particularly if it does not also benefit them. Secondly, with this crowdsourcing method, we put the control of labeling in the hands of the users, and we demonstrate that both the category and privacy labels which they provide can be trusted as if provided by the retailers themselves.

In actual practice, users of the crowdsourcing feature, when it is ultimately incorporated into the larger beacon privacy manager, would not be required to label every beacon they encounter, only those that were not labeled by a minimum number of people. Otherwise, when a user sets a policy, whether for a specific category of beacon, privacy level of beacon, or beacons in a particular sublocation, then other beacons that fall under the set policy's rules would automatically be handled appropriately, even if this user had not labeled that beacon himself. Similarly, users would not likely get paid to label beacons in practice, but incentives could be provided for setting policies that share more information with beacons during encounters. This is one way to convince users to set policies that are not completely restrictive, since neither beacon providers nor users benefit when nothing is shared with these beacons.

6.1 Limitations and Extensions

As the first work of its in kind in the domain of BLE beacons, it is important to identify both strengths and weaknesses of this crowdsourcing approach, so that it can continually be refined and improved. One clear strength is that users did not have a difficult time in expressing privacy labels for beacons. Yet, as [18] admitted in a similar approach, one limitation is the realization that users may often weigh utility over privacy and security when making decisions. An extension to this research could include explicitly investigating the role of utility to maintain a more honest level of context. Security of a crowdsourcing approach is an angle that was outside the scope of this study, but is certainly worth researching in a future iteration, such as when exploring how to best incorporate this approach into the beacon privacy manager we are developing.

Furthermore, while using the concept of contextual integrity as inspiration for this study, admittedly this only applies in a limited scope. When considering the privacy labels for beacons, the study examines primarily the "context of flow of information" aspect of contextual integrity, specifically the per-beacon and per-audience context of a flow of information, and the capacities in which the participating were acting in choosing to share. An extension to the study could additionally focus on the other components of contextual integrity, such as type of information involved. For example, we could expand the concept of beacon information involved to include the different beacon metrics, such as beacon encounter frequency, duration, or travel path. Another logical extension is to explore the principles of transmission. Here, we could investigate on what basis the beacon information could be collected and shared, be it legal requirement, in exchange for an incentive, or the promise of anonymity. This could lead to more comprehensive results from the crowd, based on a more accurate representation of context. Concerning the user study design, another limitation was the presence of "Average User Concern" in the user interface for those in the CrowdCat group. This may have somewhat artificially inflated the findings of the privacy label analysis, and more extensive experiments are required to make truly conclusive results.

In this study there is also concern for confounding variables within our study design. For example, participants were instructed about where the labels came from; TopCat users were told that the recommended label could be considered as if "provided by the bookstore itself" may have made TopCat participants question the point of labeling. Additionally, it would appear that the study confounds the number of recommended labels provided with the type of labels, however we defend this design decision with the justification that the number of labels is actually an integral part of the different conditions. There is only one correct category for each beacon, and so the TopCat should only see one recommended label. For the CrowdCat group, on the other hand, there may be a tie in recommended labels, particularly in that initial period where there are not sufficient contributions for any label to pull ahead as the top label. So it is important to reveal more than one option in this condition. Furthermore, users are still able to come to a reasonable decision regarding the top choice among the crowdsourced labels, given that these labels were marked with a count of the number of users who had selected that label before them. Yet this count in itself could be considered a confounding variable.

Lastly, one of the more critical limitations to acknowledge is the lack of a proper bootstrap or seeding mechanism, in order to address the "cold start" problem that is typical of early-stage recommender and crowdsourcing systems. This study did not address the negative implications, such as potential of information cascading, where users may observe incorrect beacons labels provided from the crowd, and despite their own inclinations, follow the same labeling. This would lead to a waterfall of incorrect labels, thereby weakening the accuracy of the crowdsourcing approach. Fortunately we did not observe any evidence of this in our study results, but we recognize that this is something for which we would have to account and correct in a future application of this approach. For example, we could implement a heuristic where recommendations are not made for beacons until at least 3 or 4 labels in agreement have been provided. Alternatively, we could collect some manual input, or use another set of heuristics to generate likely labels. In the end, the purpose of this study was to show what was possible with crowdsourcing for context and privacy, and going forward, we have the confidence to implement a more stable and robust crowdsourcing approach.

7 Conclusion

The long term vision of this research is to design a beacon privacy management framework that leverages crowdsourcing to incorporate context into privacy policy configuration for beacons. In this way, we can empower users to manage their own location information privacy during beacon encounters. We intend to achieve this framework by defining beacon privacy policy creation at the mobile architecture level, in such a way that both the beacon provider and the user can benefit from the data gleaned from location-based profiling, without putting users at risk of privacy invasion. Through the results of the study addressed in this paper, we have demonstrated the feasibility of crowdsourcing as a critical component of this vision. Regarding future work, research is already underway to build this user-centric, context-based system that incorporates the core components of the crowdsourcing application used here. Ultimately, we can even use the crowdsourced considerations to extend the beacon privacy framework, moving from a pure management system to a recommender system, suggesting privacy policy configurations based on similar beacon categories and configurations, as well as users' general privacy concern profile.

8 Acknowledgments

This research was supported by a Google Research Award, as well as a GAANN Fellowship (U.S. Department of Education grant P200A130088). We thank all the participants in our research study, as well as the reviewers of this paper. In particular, we thank Dr. Osa Omokaro, Dr. Scott Heggen, Stephen MacNeil, and Jonathan Lee, who provided critical comments on various iterations of this manuscript. Additionally, we thank Dr. David Wilson and Sanika Joshi for providing insight and expertise that greatly assisted the related research.

References

[1] Helen Nissenbaum. Privacy as contextual integrity. *Washington law review*, 79(1), 2004.

[2] Galen Grumen. What you need to know about using bluetooth beacons, 2014. http://www.infoworld.com/article/2608498/mobile-apps/what-you-need-to-know-about-using-bluetooth-beacons.html.

[3] Cathy Goodwin. A conceptualization of motives to seek privacy for nondeviant consumption. *Journal of Consumer Psychology*, 1(3):261–284, 1992.

[4] Bluetooth SIG. Bluetooth smart technology: Powering the internet of things, 2015. http://www.bluetooth.com/Pages/Bluetooth-Smart.aspx.

[5] Bluetooth SIG. Welcome to bluetooth technology 101, 2015. http://www.bluetooth.com/Pages/Fast-Facts.aspx.

[6] Bluetooth SIG. Are ibeacons cookies for the physical web?, 2015. http://www.bluetooth.com/Pages/Consumer-Electronics-Market.aspx.

[7] Jeff Howe. The rise of crowdsourcing. *Wired magazine*, 14(6):1–4, 2006.

[8] Jakob Voss. Measuring wikipedia. 2005.

[9] Michael Buhrmester, Tracy Kwang, and Samuel D Gosling. Amazon's mechanical turk a new source of inexpensive, yet high-quality, data? *Perspectives on psychological science*, 6(1):3–5, 2011.

[10] David P Anderson, Jeff Cobb, Eric Korpela, Matt Lebofsky, and Dan Werthimer. Seti@ home: an experiment in public-resource computing. *Communications of the ACM*, 45(11):56–61, 2002.

[11] Daren C Brabham. Moving the crowd at threadless: Motivations for participation in a crowdsourcing application. *Information, Communication & Society*, 13(8):1122–1145, 2010.

[12] Thomas W Malone, Robert Laubacher, and Chrysanthos Dellarocas. The collective intelligence genome. *IEEE Engineering Management Review*, 38(3):38, 2010.

[13] Jialiu Lin Bin Liu Norman Sadeh and Jason I Hong. Modeling users' mobile app privacy preferences: Restoring usability in a sea of permission settings. In *Symposium on Usable Privacy and Security (SOUPS)*, 2014.

[14] Bin Liu, Jialiu Lin, and Norman Sadeh. Reconciling mobile app privacy and usability on smartphones: could user privacy profiles help? In *Proceedings of the 23rd international conference on World wide web*, pages 201–212. ACM, 2014.

[15] Jialiu Lin, Guang Xiang, Jason I. Hong, and Norman Sadeh. Modeling people's place naming preferences in location sharing. In *Proceedings of the 12th ACM International Conference on Ubiquitous Computing*, UbiComp '10, pages 75–84, New York, NY, USA, 2010. ACM.

[16] Norman Sadeh, Alessandro Acquisti, Travis D Breaux, Lorrie Faith Cranor, Aleecia M McDonalda, Joel R Reidenbergb, Noah A Smith, Fei Liu, N Cameron Russellb, Florian Schaub, et al. The usable privacy policy project. Technical report, Tech. report CMU-ISR-13-119, Carnegie Mellon University, 2013.

[17] Iker Burguera, Urko Zurutuza, and Simin Nadjm-Tehrani. Crowdroid: behavior-based malware detection system for android. In *Proceedings of the 1st ACM workshop on Security and privacy in smartphones and mobile devices*, pages 15–26. ACM, 2011.

[18] Jialiu Lin, Shahriyar Amini, Jason I Hong, Norman Sadeh, Janne Lindqvist, and Joy Zhang. Expectation and purpose: understanding users' mental models of mobile app privacy through crowdsourcing. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, pages 501–510. ACM, 2012.

[19] Yuvraj Agarwal and Malcolm Hall. Protectmyprivacy: detecting and mitigating privacy leaks on ios devices using crowdsourcing. In *Proceeding of the 11th annual international conference on Mobile systems, applications, and services*, pages 97–110. ACM, 2013.

[20] Estimote. Estimote beacons, 2015. http://estimote.com/.

[21] Luis Von Ahn. Games with a purpose. *Computer*, 39(6):92–94, 2006.