IEEE 3rd World Forum on Internet of Things, Reston, VA (to appear)

Understanding User Privacy in Internet of Things Environments

Hosub Lee and Alfred Kobsa
University of California, Irvine
Irvine, USA
{hosubl, kobsa}@uci.edu

***Abstract*—During the past decade, user privacy has become an important issue in networked computing environments. For instance, mobile applications and devices are increasingly asking users to provide personal information, as well as monitoring users through behavioral tracking. This privacy-invasive practice is likely to increase with the proliferation of sensor devices in the upcoming era of Internet of Things (IoT). However, there has been comparatively little research so far aimed at understanding people's notion of privacy in connection with IoT. In earlier work, we unveiled five contextual parameters that characterize IoT service scenarios, and five reaction parameters that describe people's attitudes toward the scenarios. In this paper, we aim to understand *how* these contextual parameters impact people's privacy perceptions of IoT scenarios. To this end, we conducted a survey with 200 respondents on 2800 hypothetical IoT scenarios (mostly about information monitoring activities), and analyzed them using a K-modes clustering algorithm. We identified four clusters of scenarios, with clearly distinctive associated user reactions. By comparing the different clusters, we can identify contextual parameters that are associated with higher or lower acceptance of sensor tracking in IoT environments.**

***Index Terms*— IoT, privacy, online survey study, categorical data analysis, K-modes clustering.**

I. Introduction

With the widespread use of artificial intelligence technologies like machine learning, tech firms are developing new products meant to make our lives more convenient and productive. For example, Apple, Google and Microsoft are developing intelligent personal assistants, such as Apple Siri, Google Assistant and Microsoft Cortana. These products provide services tailored to each individual user by pro-actively predicting their needs. To better understand the user, service providers and device manufacturers aggressively collect personally identifiable information (e.g., location data, photos of users' faces, voice recordings, etc.) and use it as training data for their intelligent services. These industry practices will become even more powerful in future IoT environments, given that nearly all IoT devices are networked and can collectively gather personal information of users.

Without doubt, greater and more detailed volumes of personal information gathered from IoT devices can enable intelligent IoT services to better understand users and thus provide more accurately personalized services to each individual user. At the same time, however, there are still significant challenges to IoT adoption because the collection of personal information can lead to violations of privacy expectations that may harm both the user and the reputation of the firms providing these services [1]. For example, IoT devices may collect users' personal information without their permission, or may not even give any notice to them when collecting potentially sensitive information like facial photos. Hence, providing services with minimized privacy risks is very important for both protecting users' privacy expectations and keeping intelligent IoT services sustainable. In order to achieve these objectives, researchers and developers should understand how different factors influence people's privacy perceptions in an IoT environment. This understanding will enable them to better design privacy-preserving IoT systems and services.

In earlier work [2], we conducted an interview study with the goal of qualitatively assessing users' privacy perceptions regarding different IoT scenarios. We interviewed 10 participants about 9 IoT scenarios, to gather their opinions on information monitoring activities which they may encounter in their daily lives. These scenarios differed from each other in terms of five contextual parameters: place ("where"), type of collected information ("what"), agent ("who"), purpose ("reason") and frequency ("persistence") of the monitoring. We then asked participants for their thoughts on each scenario in terms of several reaction parameters: willingness to be notified ("_notification"), willingness to allow tracking ("_permission") and evaluations of comfort, risk and appropriateness of the monitoring ("_comfort", "_risk" and "_appropriateness").

Even though our previous work provided us with useful insights to extract contextual parameters that might induce privacy violations in IoT environments, it was still unclear how these parameters actually affect people's concerns about device tracking in such environments. To address this issue, we now perform a cluster analysis on online survey data, composed of IoT scenarios (i.e., contextual parameters) and user responses thereto (i.e., reaction parameters). Because all parameters have either categorical or ordinal values, we utilize K-modes, a variant of the K-means clustering algorithm. We determine that four clusters (K=4) are optimal, each of which is associated with three unique reaction parameters. We compare the identified clusters with respect to each contextual parameter, and discover some latent relationships between the given contextual information and people's privacy preferences.

978-1-5090-4130-5/16/$31.00 ©2016 IEEE

II. Related Work

In emerging IoT environments, users are surrounded by various sensor devices that monitor users [3]. New types of user-centric data such as physical activities, mood or presence can be constantly and unnoticeably collected in our daily lives [4]. The collection methods are changing too. Users' personal information is collected more passively and collectively [5]. Thus, users may feel less aware and in control of personal information being collected.

Researchers have studied several factors that could influence users' privacy concerns in the IoT environment. For example, Choe et al. revealed that users are less willing to share self-appearance, intimacy behavior, cooking/eating, media use and oral expressions with in-home sensors [6]. Although the authors investigated an important contextual factor, namely location, their findings were limited to the specific place investigated, namely people's homes. Hu et al. developed a context-aware location sharing system [7]. Their assumption was that users are more likely to share location information with IoT devices in emergency situations than under normal circumstances. However, there exist no user studies or experiments to support this claim. Lederer et al. conducted a questionnaire-based study to gauge the relative importance of two contextual factors, inquirer and situation, in determining privacy preferences of users in ubiquitous computing environments [8]. They found that the identity of the information inquirer (4 possible values: spouse, employer, stranger, merchant) is a more dominant factor than situation (2 possible values: working lunch, social evening). However, there is no guarantee that this finding can also be applied to diverse IoT contexts since the situation was too coarsely defined.

Protecting user privacy is a big challenge in the adoption of IoT products and services [1]. A comprehensive understanding of user privacy in the IoT environment is required for service providers to construct privacy-preserving IoT. We find little research that experimentally investigates contextual factors influencing users' privacy expectations in an IoT environment. Our study is aimed at filling this gap.

III. In-depth Analysis of User Privacy in IoT

In this section, we discuss how we collected and analyzed the online survey data, and what we learned about people's privacy preferences in IoT environments from this analysis.

A. Data Collection

We recruited 200 participants on Amazon Mechanical Turk (MTurk), educated them about IoT and asked for their opinions on 14 IoT scenarios one by one. To assure the quality of survey responses, we restricted participants to adults who live in United States, are proficient in English and have a high worker reputation (above 95% approval ratings). 100 females and 99 males participated (one person did not disclose their gender), and the majority (57.5%) are aged 25-40.

We generated IoT scenario descriptions through random combinations of the five contextual parameters. Table I shows the possible parameter values and one sample scenario description. Since the scenarios were unique for each participant, 2800 scenarios were created in total. Like in the interview study, we enquired participants about their privacy concerns on the presented scenarios. Table II shows the possible values of the reaction parameters and sample questions. In addition, we asked participants to describe their opinions in a free text field. We then performed a cluster analysis to determine in what way these contextual parameters affect people's reactions to being tracked in IoT environments.

TABLE I. Contextual Parameters

Parameter (id)	Values	
"where" (C_1)	0. your place 1. someone else's place 2. semi-public space (e.g., restaurant) 3. public space (e.g., street)	
"what" (C_2)	1. phoneID 2. phoneID>identity 3. location 4. location>presence 5. voice 6. voice>gender 7. voice>age 8. voice>identity 9. voice>presence 10. voice>mood 11. photo 12. photo>gender	13. photo>age 14. photo>identity 15. photo>presence 16. photo>mood 17. video 18. video>gender 19. video>age 20. video>presence 21. video>mood 22. video>lookingAt 23. gaze 24. gaze>lookingAt
"who" (C_3)	1. unknown 2. colleague/fellow 3. friend 4. own device	5. business 6. employer/school 7. government
"reason" (C_4)	1. safety 2. commercial 3. social	4. convenience 5. health 6. none
"persistence" (C_5)	0. once	1. continuously
Sample Scenario Description		
A device of a *friend* (C_3=3) records your *voice to check your presence* (C_2=9). This happens *once* (C_5=0), while you are at *semi-public place* (C_1=2), for your *safety* (C_4=1).		

TABLE II. Reaction Parameters

Parameter (id)	Values
"_notification" (R_1)	1. notify me always 2. notify me just this time 3. don't notify me
"_permission" (R_2)	1. don't allow always 2. don't allow just this time 3. allow just this time 4. allow always
"_comfort" (R_3)	1. very uncomfortable/risky/inappropriate 2. uncomfortable/risky/inappropriate 3. somewhat uncomfortable/risky/inappropriate 4. neutral 5. somewhat comfortable/safe/appropriate 6. comfortable/safe/appropriate 7. very comfortable/safe/appropriate
"_risk" (R_4)	
"_appropriateness" (R_5)	
Sample Questions	
• If this situation happens, would you want to be *notified* (R_1) about it? • If this situation happens, would you want to *allow* (R_2) it? • How *appropriate* (R_5) do you consider this situation to be?	

B. Cluster Analysis

1) K-modes clustering

The K-means clustering algorithm is the most popular data mining technique to partition observations into K clusters. Each observation is assigned to the cluster with the nearest mean, which itself serves as a representative value of the cluster. However, the applicability of K-means is restricted to continuous numeric values. A variant of K-means, the K-modes clustering algorithm, aims to utilize the K-means paradigm for clustering categorical data without the need for data conversion. The K-modes clustering algorithm makes the following extensions to K-means: (1) replacing cluster means with modes, (2) using the simple matching dissimilarity function in place of the Euclidean distance function to compute the distance between categorical objects and (3) updating modes with the most frequent categorical attributes in each iteration of the clustering [9, 10]. To be specific, it divides categorical objects into K groups such that the distance from objects to the assigned cluster modes is minimized. Default simple-matching distance is used to determine the dissimilarity of two objects. It is computed by counting the number of mismatches in all variables. This distance is weighted by the frequencies of the categories (modes) in data. We used *klaR* [11], an R implementation of the K-modes clustering algorithm, on our Amazon MTurk survey data to find cluster modes and assign each data point to the corresponding cluster based on its dissimilarity function through an iterative clustering process.

2) Selecting the number of clusters

Determining the correct number of clusters is an important step in unsupervised data clustering like K-modes. As a-priori knowledge of the appropriate value of K does not exist for our data set, we heuristically searched for the optimal K by applying the Elbow method [12]. First, we computed the sum of errors (SE) of the K-modes clustering with a limit of 50 iterations, while increasing K from 2 to 10. The SE is defined as the sum of the distance between each member of the cluster and the cluster's centroid (mode):

$$SE_K = \sum_{i=1}^{K} \sum_{x \in c_i} dist(x, c_i)$$

where x is a data point belonging to the ith cluster and c_i is the mode of the ith cluster. Then, we calculated the difference values between SE_K and SE_{K-1}, and found that the largest decrease in errors occurs when we increase K from 3 to 4 (Table III). Therefore, we chose 4 as the appropriate number of clusters, and used it as a parameter (*modes*) for initializing the K-modes clustering algorithm. The algorithm then randomly choses 4 rows from the data set as the initial modes, and updates the modes through iterative clustering. As we did not configure the maximum number of iterations allowed (*iter.max*), the algorithm continued until the clustering error was minimized.

TABLE III. CLUSTERING ERRORS

No. of Clusters (K)	**Sum of Errors (SE_K)**	**Error Difference ($SE_{K-1} - SE_K$)**
2	15765	
3	15075	-690
4	**14170**	**-905**
5	13655	-515
6	13129	-526
7	12917	-212
8	12562	-355
9	12329	-233
10	12311	-18

3) Interpretation of clusters

Table IV shows the modes that the clustering algorithm generated as the centroids of the four clusters. Interestingly, the clusters differ from each other primarily in the contextual parameters "what" (C_2) and "who" (C_3): each mode has a unique categorical value for these parameters. This means that "what" and "who" define clusters more than the remaining contextual parameters "where" (C_1), "reason" (C_4) and "persistence" (C_5). In addition, each mode has identical and unique values for the reaction parameters "_comfort" (R_3), "_risk" ($R_{4,}$ reverse-coded) and "_appropriateness" (R_5). These three parameters indicate respondents' attitudes about a scenario on a scale of 1 to 7. For instance, R_3=1, R_4=1 and R_5=1 indicate that the scenario is perceived as "very uncomfortable", "very risky" and "very inappropriate", respectively (see Table II). In contrast, the remaining reaction parameters "_notification (R_1)" and "_permission (R_2)" do not have unique values per cluster.

Since the reaction parameters R_3, R_4 and R_5 pertaining to each mode are unique, we can characterize each cluster along these parameters. We will label scenarios that belong to the cluster CL_1 as "acceptable" to the study participants because its mode has the second highest value for R_3, R_4 and R_5 (namely 6 on a 7-item scale). We label CL_2 scenarios as "somewhat unacceptable" (since the value of its reaction parameters (3) falls slightly below the scale average), CL_3 scenarios as "unacceptable" and CL_4 scenarios as "very unacceptable". Only 12.6% of the scenario descriptions that we presented to participants fall into the "acceptable" cluster, while 40.8% fall into the "very unacceptable" cluster.

TABLE IV. MODES OF CLUSTERS

Mode	**Contextual Parameters**					**Reaction Parameters**					**Cluster**	**Label**	**Number of Instances**	**Color Code**
	C_1	C_2	C_3	C_4	C_5	R_1	R_2	R_3	R_4	R_5				
M_1	0	**8**	**4**	6	0	3	4	**6**	**6**	**6**	CL_1	Acceptable	352/2800	Green
M_2	2	**9**	**3**	1	0	1	1	**3**	**3**	**3**	CL_2	Somewhat unacceptable	466/2800	Yellow
M_3	3	**22**	**5**	6	1	1	1	**2**	**2**	**2**	CL_3	Unacceptable	840/2800	Red
M_4	0	**24**	**1**	6	0	1	1	**1**	**1**	**1**	CL_4	Very unacceptable	1142/2800	Black

4) *Verification of clustering results*

To verify the distinctiveness of the clusters, we first conducted three Welch's t-tests on the R_3 parameter between the following pairs of clusters: (CL_1, CL_2), (CL_2, CL_3) and (CL_3, CL_4). The reason for using Welch's t-test is that all clusters have different variances in the R_3 parameter. The tests confirm that the difference in the means of the R_3 parameter between each of these clusters is statistically significant (p-values < 0.016, Bonferroni-corrected for three comparisons). Then, we also performed Welch's t-tests on the R_4 and R_5 parameters, and reached the same conclusion. Thus, we find that the clusters are sufficiently distinct from each other in terms of user reactions to the scenario descriptions pertaining to each cluster.

Next, we visually inspect clusters so that we can confirm our cluster labeling is reasonable. In doing this, we utilized the reaction parameter R_5 since *appropriateness* is a crucial element for assessing potential privacy risks in a given context. This concept has been proposed by Helen Nissenbaum's Contextual Integrity theory [13], which provides a systematic way of determining when and why people perceive certain usage and disclosure of personal information as appropriate, or as a privacy violation.

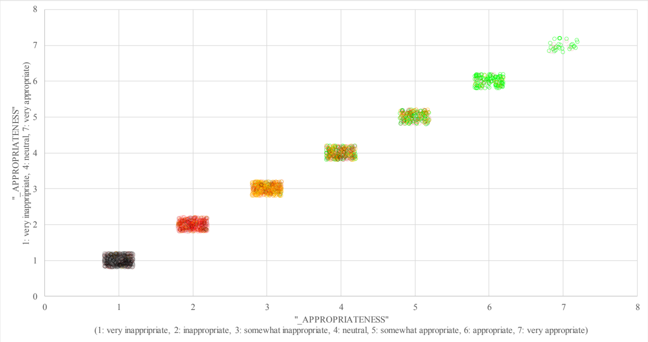

Fig. 1. Visualized clustering results in terms of "_appropriateness"

We first assign colors to clusters: green for CL_1, yellow for CL_2, red for CL_3 and black for CL_4. We then project all scenario descriptions from all clusters onto a 2-dimensional space, using their R_5 reaction parameter as both their x and y values and their cluster color as their surface code. We add a small amount of random noise to the coordinates of each data point to make them visible. Figure 1 shows that situation descriptions that respondents deemed very inappropriate (R_5=1) mostly became clustered into CL_4 (black). In contrast, situation descriptions that respondents deemed appropriate or very appropriate (R_5=6, 7) became clustered into CL_1 (green).

C. Comparison of Clusters based on Contextual Parameters

In this section, we compare the clusters in terms of the five contextual parameter values pertaining to their modes, to understand how contextual information influences people's reactions toward their privacy in IoT environments.

1) *where*

Regarding the "where" parameter (see Fig. 2), participants consider monitoring that occurs at personal places like their homes as very unacceptable ("where"=0, see CL_4; p < .0001, Cohen's d = 0.479[1]). This is probably because people do not exercise self-control in such private spaces, and therefore do not want to be monitored. In addition, many participants also have privacy concerns if the monitoring is performed in a public space ("where"=3, see CL_3; p < .0001, d = 0.4921). A participant commented:

> "*Serious invasion of privacy (yes, even in a public place). If the data is stored, a profile could be created as to what I am doing or where I am going.*" [P46]

As for semi-public spaces such as restaurants, participants feel that monitoring is somewhat unacceptable ("where"=2, see CL_2; p < .0001, d = 0.6109). Interestingly, personal place is a dominant factor for making scenarios acceptable ("where"=0, see CL_1). Therefore, we need to further investigate other contextual factors like the "what" and "who" parameters to fully understand this cluster.

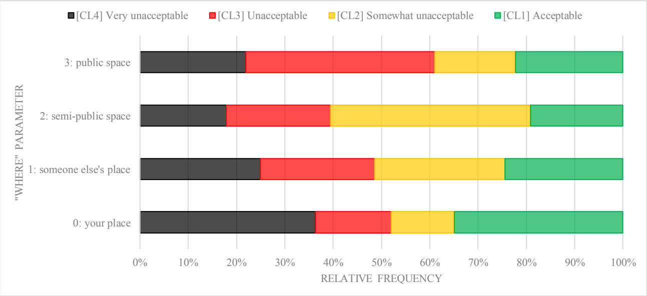

Fig. 2. Relative distribution of "where" parameter per cluster

2) *what*

With regard to the "what" parameter (see Fig. 3), participants do not accept situations in which someone is monitoring them to figure out what they are looking at ("what"=23, 24, see CL_4; p = 0.0001, d = 0.3041). Participants also considered photo-taking and/or video monitoring unacceptable for various purposes ("what"=12, 16, 22, see CL_3; p < .0001, d = 0.319).

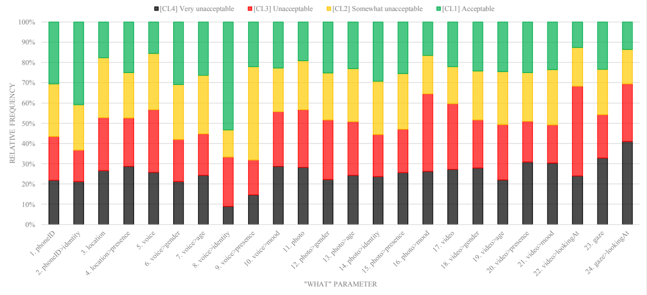

Fig. 3. Relative distribution of "what" parameter per cluster

[1] In the Social Sciences, effect sizes less than 0.3 are commonly regarded as small, effect sizes between 0.3 and 0.6 as medium, and effect sizes larger than 0.6 as large [14].

Participants' reactions toward voice monitoring were generally positive compared to the above cases. For instance, many participants are likely to allow their voice to be monitored for gender identification and location awareness ("what"=6, 9, see CL_2; $p = 0.0006$, $d = 0.2713$). They are also very open to giving their personally identifiable information such as unique phone id or voiceprint ("what"=2, 8, see CL_1; $p < .0001$, $d = 0.6237$), presumably due to convenience or habituation. For instance, participant P19 deemed voice-based authentication and personalization as acceptable:

> "*Maybe a voice recording could be used in place of credit/debit cards for transaction purposes. The system analyzes the recording and knows what you want from the business and prepares your order or services.*" [P19]

3) who

In our previous interview study, some interviewees stated that "who" is an important parameter affecting their privacy preferences regarding IoT services. The results of our present study clarify its impact (see Fig. 4). If the monitoring entity is unknown to online survey participants, their responses on the given scenarios are very conservative ("who"=1, see CL_4; $p < .0001$, $d = 0.7268$). They also do not trust the government ("who"=7, see CL_4; $p < .0001$, $d = 0.2603$). Participants also have privacy concerns if a nearby business tracks their personal information ("who"=5, see CL_3; $p < .0001$, $d = 0.5845$). This may be because they doubt that the company safeguards their information. People feel safe if the monitoring is performed by either their friends ("who"=3, see CL_2; $p < .0001$, $d = 0.6305$) or own devices such as their smartphone ("who"=4, see CL_1; $p < .0001$, $d = 0.9989$).

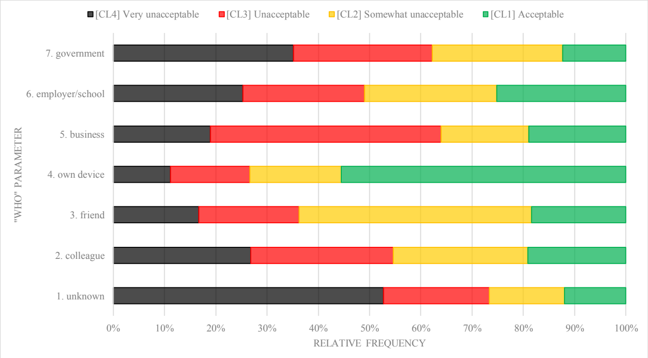

Fig. 4. Relative distribution of "who" parameter per cluster

4) reason

We specified the purpose of each IoT scenario using the "reason" parameter. This parameter has six possible values, namely "safety", "commercial", "social", "convenience", "health" and "not specified" (see Fig. 5). The absence of a purpose causes the greatest number of unacceptable scenarios ("reason"=6, see CL_3 and CL_4; $p < .0001$, $d = 0.3221$). On the other hand, a considerable amount of scenarios was still considered acceptable ("reason"=6, see CL_1), even though the purpose of monitoring was not indicated. This suggests that participants have a tendency to base a privacy decision mainly on concrete contextual factors like the "who" and "what" parameters, as we explained in the previous sections. In fact, several participants reacted to purposeless scenarios by imagining the possible purposes by themselves:

- "*It would be able to find criminals and catch criminal behavior on tape so the benefit could possibly improve public safety.*" [P10]
- "*If I was on a trip to get to my friend's house, they could see how far I am from them without having to call or text me.*" [P105]
- "*If I have a health related accident or injury then the person watching can come assist me immediately.*" [P112]

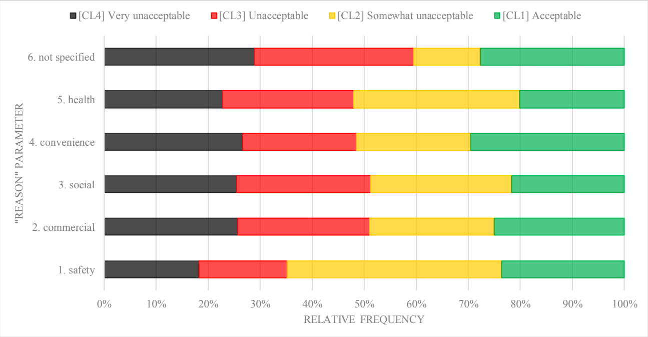

Fig. 5. Relative distribution of "reason" parameter per cluster

Other than "not specified", "convenience" is the most significant purpose that participants found acceptable ("reason"=4, see CL_1). Additionally, safety is a reasonable justification for participants to generally accept the situation ("reason"=1, see CL_2).

5) persistence

We assumed that participants will have strong privacy concerns if information monitoring happens continuously rather than just once. However, our analysis results are inconclusive: no clear tendency towards the one or the other can be seen in Fig. 6.

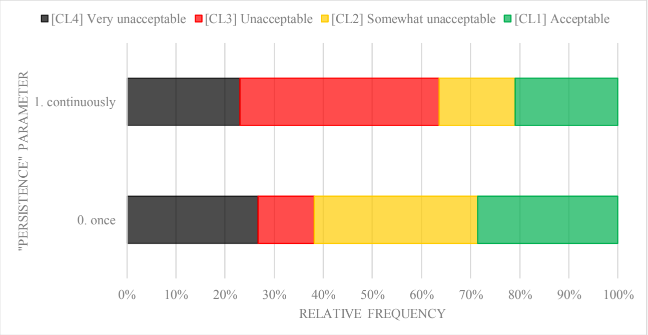

Fig. 6. Relative distribution of "persistence" parameter per cluster

IV. Limitations and Future Work

Our analysis shows how each contextual factor impacts people's privacy preferences in IoT environments. Yet, this study still has some limitations that need to be considered.

First, we notice that some contextual parameters were defined with coarse granularity. For example, "someone else's place" of the "where" parameter might be interpreted differently by different participants because the meaning of "someone else" is broad. Furthermore, given that participants took this online survey at a location that has no association with the IoT scenarios described in the survey, there could have been a decreased sense of realism to the scenarios. Privacy research has repeatedly shown that people's stated attitudes with regard to privacy often differ from their actual behaviors in a concrete situation [15-17]. For these reasons, out-of-context attitudinal studies like our online survey must be viewed with some caution and be verified.

We plan to tackle these limitations by developing a location-based survey system on a mobile/wearable device which presents scenarios that are specifically related to participants' current locations. The aim of this system is to simulate user experience in a *real* IoT environment as closely as possible. To that end, our research team will collaboratively create realistic scenarios that are mapped to real locations (e.g., through embedded GPS information), and feed the scenarios into the survey system. Participants will then be asked to walk around while carrying the device. As participants move towards a certain location, the device will display an IoT scenario description related to this location. Then, participants will answer questions about their privacy preferences for the presented scenario. We will use the same questionnaires from our online study to investigate whether any discrepancies in the responses of the participants exist between the online survey and the location-based survey.

V. Conclusion

In this paper, we showed that IoT scenarios can be grouped into four clusters in terms of their potential privacy risks. By comparing these clusters according to the five contextual parameters, we also extracted some contextual factors that cause users' privacy concerns in IoT. To verify whether our findings are also applicable to broader IoT contexts, we plan to design and develop a location-based survey system. One of our research hypotheses is that a location-based survey system would be more suitable for collecting genuine responses from users than a traditional survey system, since it situates users in IoT scenarios that are comparatively more realistic than clicking through a survey in a web browser.

Acknowledgment

This research has been funded by grants from Intel and the NSF (SES-1423629). The human subjects research conducted for this paper is covered under IRB protocol #2014-1600 at the University of California, Irvine. The authors would like to thank Bart Knijnenburg for his support in designing and administering the Amazon MTurk survey.

References

[1] J. Gubbi, R. Buyya, S. Marusic, and M. Palaniswami, "Internet of Things (IoT): A vision, architectural elements, and future directions," Future Gener. Comput. Syst., vol. 29, no. 7, pp. 1645–1660, Sep. 2013.

[2] R. Chow, S. Egelman, R. Kannavara, H. Lee, S. Misra, and E. Wang, "HCI in Business: A collaboration with academia in IoT privacy," HCI in Business, F. F.-H. Nah and C.-H. Tan, Eds. Springer International Publishing, 2015, pp. 679–687.

[3] A. C. Sarma and J. Girão, "Identities in the Future Internet of Things," Wirel. Pers. Commun., vol. 49, no. 3, pp. 353–363, Mar. 2009.

[4] J. Virkki and L. Chen, "Personal perspectives: Individual privacy in the IoT," Adv. Internet of Things, vol. 3, no. 2, pp. 21–26, Apr. 2013.

[5] C. M. Medaglia and A. Serbanati, "An overview of privacy and security issues in the Internet of Things," The Internet of Things, D. Giusto, A. Iera, G. Morabito, and L. Atzori, Eds. Springer New York, 2010, pp. 389–395.

[6] E. K. Choe, S. Consolvo, J. Jung, B. Harrison, and J. A. Kientz, "Living in a glass house: a survey of private moments in the home," in Proc. 13th Int. Conf. Ubiquitous Computing (UbiComp), Beijing, China, 2011, pp. 41–44.

[7] C. Hu, J. Zhang, and Q. Wen, "An identity-based personal location system with protected privacy in IoT," in Broadband Network and Multimedia Technology (IC-BNMT), 2011 4th IEEE International Conference on, Shenzhen, China, 2011, pp. 192–195.

[8] S. Lederer, J. Mankoff, and A. K. Dey, "Who wants to know what when? privacy preference determinants in ubiquitous computing," in Proc. CHI '03 Extended Abstracts on Human Factors in Computing Systems, Lauderdale, FL, USA, 2003, pp. 724–725.

[9] Z. Huang, "A fast clustering algorithm to cluster very large categorical data sets in data mining," in Research Issues on Data Mining and Knowledge Discovery, Tucson, AZ, USA, 1997, pp. 1–8.

[10] Z. Huang, "Extensions to the k-Means algorithm for clustering large data sets with categorical values," Data Min. Knowl. Discov., vol. 2, no. 3, pp. 283–304, Sep. 1998.

[11] C. Neumann, kmodes {klaR}: K-Modes Clustering. CRAN repository.

[12] T. S. Madhulatha, "An overview on clustering methods," *arXiv preprint arXiv:1205.1117*, May 2012.

[13] H. Nissenbaum, "Privacy as contextual integrity," Wash. Law Rev., vol. 79, pp. 119–157, 2004.

[14] J. Cohen, Statistical power analysis for the behavioral sciences (revised ed.). New York: Academic Press, 1977.

[15] A. Acquisti and J. Grossklags, "Privacy attitudes and privacy behavior," Economics of Information Security, vol. 12, L. Camp and S. Lewis, Eds. Springer US, 2004, pp. 165–178.

[16] C. Jensen, C. Potts, and C. Jensen, "Privacy practices of Internet users: self-reports versus observed behavior," Int. J. Hum.-Comput. Stud., vol. 63, no. 1–2, pp. 203–227, July 2005.

[17] K. Connelly, A. Khalil, and Y. Liu, "Do I do what I say?: observed versus stated privacy preferences," Human-Computer Interaction – INTERACT 2007, vol. 4662, C. Baranauskas, P. Palanque, J. Abascal, and S. Barbosa, Eds. Springer Berlin / Heidelberg, 2007, pp. 620–623.