An Experience with a De-identifying Task to Inform About Privacy Issues

Luis Gustavo Esquivel-Quirós(✉) and E. Gabriela Barrantes

Universidad de Costa Rica, San José, Costa Rica
luis.esquivel@ucr.ac.cr, gabriela.barrantes@ecci.ucr.ac.cr

Abstract. People tend to value their privacy, but are usually unaware about the extent to which their personal information is exposed through ordinary data available online. In this paper we describe an experience in which a group of students worked to identify a group of people from partial data that had been stripped of any direct identifiers, such as name or identification number. The students were successful in the assigned task, and as an indirect consequence, there was an increase of interest in the topic of privacy. With the partial evidence collected from this case, we argue that a hands-on, exercise-solving approach could be adapted to communicate privacy issues more effectively.

Keywords: Privacy concerns · De-identification · Re-identification · Education

1 Introduction

Information is constantly flowing from most of our electronic devices. We generate all kinds of information: social, geographical, financial, and many others. If the information refers to an individual, and if it is intentionally or unintentionally published, it might expose sensitive information which the individual might wish to, and possibly has the right to, keep private. This issue has been raised everywhere, but we were interested in the situation in Costa Rica; a small country with an educated population, and relatively high information and communication technology (ICT) penetration [6]. For years, it has produced and exported software, and international software and hardware companies now have part of their operations in our country [1].

Privacy issues have developed at different paces depending on the country. Costa Rica had a slow start in legislation and general awareness about privacy. Until recently, there were no specific privacy laws, and instead privacy breaches were dealt mostly by the Constitutional Court, using *Habeas Data*, which is a constitutional court action available in some countries, used to protect the rights of an individual over its own data, and in some ways analogous to the more commonly-known *Habeas Corpus* [5].

In 2011, the first privacy-specific law was approved, the "*Ley de Protección de la Persona frente al tratamiento de sus datos personales*" [8], and in 2013 the

D. Aspinall et al. (Eds.): Privacy and Identity 2015, IFIP AICT 476, pp. 49–60, 2016.
DOI: 10.1007/978-3-319-41763-9_4

regulations for the law were published [4]. The law also created PRODHAB, a regulatory agency ("*Agencia de Protección de Datos de los Habitantes*") which started operations in 2013 [2,8].

Costa Rica has experienced scandals due to breaches in privacy, including some very prominent ones in recent years that received wide media coverage. For example, there was the case of Keylor Navas, a relatively famous national soccer player, in which it came to light that his family in Costa Rica was the subject of more than 20 frivolous consults at the hands of employees of the Organization for Judiciary Research (OIJ) [13]. The case started in October 2014, and caused a public outcry that ended up in an inquiry by a Legislative Assembly commission [23]. The commission reached its conclusion in May 2015 [15], but by November 2015 it had disappeared from the public view, without significant consequences.

However, even with the new law, the PRODHAB, the press coverage of privacy breach scandals, and the general level of education, people in Costa Rica seem mostly unaware or indifferent to privacy issues, as could be inferred from the press coverage of such cases. Each time a given person suffers a privacy breach, there is enough public outrage to keep it in the news for some time, but it is mostly forgotten soon after. We believe that this reflects a cognitive dissonance between beliefs about privacy and actual attitudes about data online. This phenomenon has been well documented in other countries, and is referred to as the "*Privacy Paradox*" [18].

A different privacy issue also present in Costa Rica has to do with the general lack of understanding of the complex interactions among different pieces of data that could link seemingly "sanitized" data to sensitive information about the individual. For example, some of the recent rulings of the Constitutional Court show that judges consider privacy important, but are not aware that the data being authorized for release could be used to put individuals in danger [12]. This so-called "re-identification" threat is also a well-documented reality, with most documented cases dealing with data published in the United States ([9,16,24]).

Given all of the above, it was important for us to test directly what the situation was in Costa Rica. More specifically, we wanted to determine how easy (or difficult) it would be for a non-expert to re-identify local individuals. Therefore, we decided to run a limited re-identification exercise with a group of educated, but non-expert participants. We also restricted it to "young" people (under 35 years of age), as there is evidence indicating that the privacy paradox is an issue in this group [14]. However, after running the task we realized that it had an unexpected outcome regardless of how easy it was for the participants to identify the local individuals, so we decided to report our findings from the re-identification task as evidence for a possible educational strategy to raise awareness about privacy issues.

In Sects. 2 and 3 we define basic privacy and data protection concepts, respectively. In Sect. 4 we present the methodology used. Section 5 presents our results and their analysis. Section 6 summarizes our findings, and Sect. 7 proposes future directions for this work.

2 Privacy

The concept of "privacy" has no single, clear and definite meaning, and varies widely between different cultures. This concept is discussed in philosophy, politics, law, sociology, anthropology, psychology, and other areas. The difference between the public and the private have been recognized in some form in most societies, albeit there is no general agreement about the meaning, the limits, and the application of these limits (legal framework) [8,11,21].

Following Daniel Solove [22], privacy can be generically conceptualized as a barrier that people can hold to provide themselves with some relief from different sorts of social friction, which will allow them to function as a member of a society without being overwhelmed by the multiple pressures it imposes.

A definition for privacy that is relevant for information privacy was given by Alan Westin [25]: *"Privacy is the claim of individuals, groups, or institutions to determine for themselves when, how, and to what extent information about them is communicated to others."*

William Parent [17] defines privacy as the condition of not having undocumented and unauthorized personal information in possession of other parties. His position is that there is a privacy loss when others acquire personal information about an individual without his or her authorization. Personal information is defined by Parent as any information about facts that the person chooses to not reveal about him or herself. In this view, personal information becomes "documented" only if it becomes public with the authorization or actions of the individual.

The concept is in fact very dependent upon the context [22] wich makes raising awareness about issues related to it complicated, which might explain why conciousness about privacy in the general population remains an elusive goal.

3 Identification, De-identification and Re-identification

For terminology only, we will follow the decisions about data privacy definitions stated in the Internal Report 8053 of the National Institute of Standards and Technology of the United States [10]. In particular, we will **not** use the term "anonymization", and "personal information" will refer to any information from an individual, but "identifying information" would be the subset of personal information that could be used to identify that individual. The definition recognizes that personal data from which all identifying information has been removed, sometimes can be traced back to the individual.

A **de-identification** process concerns itself with stripping identifying information from personal information. The purpose is usually to decrease the risk of sensitive data about the individual being associated back to him or her. A process used to defeat de-identification is called **re-identification**. Such a process attempts to link data back to its originator. There are many possible strategies to accomplish this task [9].

4 Methodology

Given that the focus of our main exercise was to probe the ease of re-identification in our environment, the methodology focused on re-identification details. Most of the post-exercise written questions were oriented in that direction, and individual concerns about privacy were not directly referenced. For the purpose of this paper, we relied on items that provided indirect indications about the participants' concerns about privacy, and the oral responses of the participants during the interactions with the experimenters.

Subsection 4.1 describes the design of the re-identification exercise that was to be carried out by the participants in the study. Subsection 4.2 deals with the details of the exercise execution. Subsection 4.3 presents the indicators taken from the questionnaire used by the participants to qualify the exercise.

4.1 Exercise Design

The design of the re-identification exercise required choosing the following:

1. A base dataset;
2. The de-identified subsets; and
3. A group of individuals who will execute the exercise.

The choices taken are described below.

Choosing the Base Dataset. The exercise required the participants to re-identify a group of people, given some weakly de-identified data. Therefore, the first challenge was to define this group. The requirements for the base dataset were the following:

1. We had to be able to find enough publicly available information about them
2. They had to be somehow "interesting" for the participants in the exercise
3. They had to be relatively public figures so we would not infringe any local privacy laws
4. We needed a relatively large number of individuals (more than 20) to be able to choose different combinations among them

Some possibilities included local politicians, entertainment personalities, fairy tale characters, and sports figures. We ended up choosing local football (soccer) players, given the World Cup excitement, and the importance of football in Costa Rica. More specifically, we choose all the players that were called in a single selection summons, which gave us a total of 30 players.

We collected the following publicly-available information about the players:

- Local ID number (*Cédula de identidad*)
- Complete name
- Birth date
- Marital status
- Number of children
- Province were registered to vote
- Current club

Choosing the De-identified Subsets. After collecting the base dataset, we had to define the subsets of de-identified data that would be given to study participants to re-identify. We defined two subsets for the exercise, to test if restricting the amount of de-identified information would make a difference in the re-identification efficacy. The first subset was comprised of data from three queries on the base dataset, and no further data was provided during the exercise. The definition of the queries follows:

- Birth year, age, place of birth, province where registered to vote, for all 30 players
- Marital status and number of children, but only for married players
- Number of children and current club, but only if the current club was local.

The second de-identified subset was potentially different for each participant. No data was given outright, but a participant could request from the experimenters up to three queries from the base dataset, with the following rules:

- In the first query, a participant could request all the data in a non-direct ID column (for example, column "number of children", but not "Local ID number"). The student could specify the column by name, but could only pick one, two or three columns.
- In the second query, a participant could request at most two columns, but only 50 % of the data for that column was given (data for 15 randomly chosen players).
- In the third query, the participant could request only one attribute, and again, only 50 % of the data would be returned.

Choosing the Participants. We needed as participants a group of young people who were not very familiar with either computer security or privacy issues. We had the generous cooperation of the students and the professor in the FS408 Thermodynamics class, at the end of 2013. This is a second-year course in the curriculum of the Bachelor Degree in Physics at the Universidad de Costa Rica.

None of the students in the group had previous, specialized experience in security, data mining, or protection of data privacy, and they were all Physics majors. There were 49 participating students. They were naturally divided into two attendance modalities: 34 in-class and 15 remote students. We used this natural division to test two different types of query.

The participants were mostly male given that the student population for the Physics major at the Universidad de Costa Rica presents a marked gender imbalance, as it can be seen in Table 1. However, both the group and each of the subgroups were very homogeneous in age. The youngest student was 19, and the oldest 29 at the time. Table 2 summarizes the age distribution of the participants in the exercise.

Table 1. Summary of gender distribution of participants.

Population	Male	Female	Total population
In-person	26	8	34
Remote	14	1	15
Whole group	40	9	49

Table 2. Summary of age distribution of participants.

Population	Average	Stand. Dev	Median	Mode
In-person	21,21	1,49	21	20
Remote	23,07	3,13	22	21
Whole group	21,78	2,27	21	20

4.2 Execution

The full class had to attend a brief presentation on general privacy issues. This presentation included basic concepts on privacy protection, and legal consequences of information published online, showing local and international examples ([3,19]). We showed them examples of potential dangers such as identity theft, and misuse of sensible information. After the talk, all students had to participate in a group activity where they re-identified characters in a fairy tale, with a subsequent discussion of methods and results. During both activities, they were allowed to ask questions, make suggestions and offer opinions.

The next step was to explain the rules of the identification exercise, but before that, an explanation was given about the possible uses of the data to be collected, and the possibility of opting-out. Those students who were willing to participate had to complete and sign informed consent forms.

There was a small reward associated to the completion of the exercise. If the student demonstrated serious effort towards the re-identification, and answered a question about the entropy of the data, the completed exercise counted as an extra quiz. The actual re-identification success did not count towards the grade.

Afterwards, all students were given their respective instruments, which included the de-identified datasets, and a questionnaire to be completed after the exercise. They were given a week to complete it. In-class students were given the fixed dataset and remote students had to generate their own using the rules explained in Sect. 4. The goal was to identify (obtain the names) of the largest number of players. As mentioned in Sect. 4, the reason that different datasets were used for each group is that originally the exercise was meant to compile information about the ease of re-identification in Costa Rica.

The answers to the exercise were collected a week after it was handed out. The experimenters had a brief discussion session with the students at that point. The students in the remote modality had the responsibility to be present if there

was an evaluation activity, so all of them attended the class in the first week, but most of them were not present for the discussion on the second week.

4.3 Questionnaire and Analysis Strategy

The questionnaire that the students had to complete after completing the re-identification exercise, contained nine items, divided into three closed, and six open-text questions. The questions were oriented to determine the difficulty of the re-identification task. Because the written instrument was not explicitly designed to measure levels of concern about privacy, we used the answers to the following items as proxies for concern:

- **Perception about the exercise:** This was a single-choice question. It stated: *"In which category would you assign this exercise?"*. It had four possibilities: "challenging", "interesting", "normal" and "boring".
- **Guess about the correct number of re-identifications:** The answer to this question required a number from 0 to 30. The question was: "How many people did you re-identify?".

5 Results and Analysis

We will start by presenting the efficacy of the actual re-identification performed by the students (Subsect. 5.1). Next we show the perceptions about the exercise itself (Subsect. 5.2) which is the closest to a direct indicator of a change in concern about privacy. We then analyze the self-reported perception of accuracy (Subsect. 5.3) as an indicator of learning. Attitudes during the in-class interactions are described in Subsect. 5.4. Finally, in Subsect. 5.5 there is a brief discussion of the results.

5.1 Success in Re-identification

The students were quite successful finding who the players were. All participants completed the exercise, and turned it in on time. As explained in Sect. 4, being able to correctly identify the players was not being graded in itself. Even so, a large percentage of students (90 % for the in-class and 93.3 % for the remote groups) identified correctly more than 96 % (29 or 30) of the de-identified players. The student who identified the smallest number of players, "only" managed to identify 16 of them, although he claimed to know next to nothing about football.

In fact, students were so effective in the re-identification, that they *corrected* one player's birth date, on which we inadvertently had introduced a mistake when collecting the information.

Figure 1 presents the normalized frequency histograms for the number of correctly re-identified players.

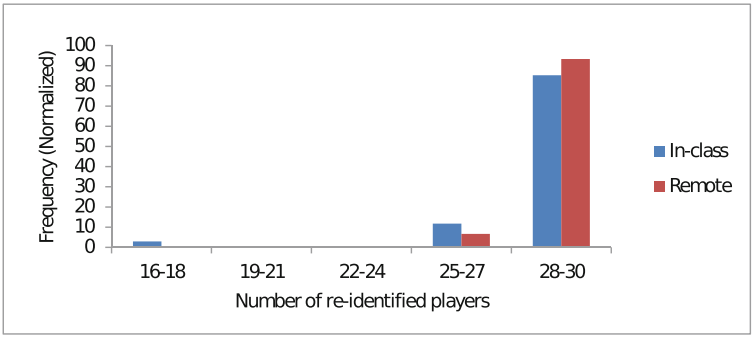

Fig. 1. Frequency (normalized) of correctly re-identified players. (Color figure online)

5.2 Perceptions About the Exercise

The ratings given by the students to the exercise are summarized in Table 3. Given the unconcerned attitudes revealed during the first week activities (see Subsect. 5.4), the change to consider the exercise as mostly "interesting" (67.7 and 73.3 % for in-class and remote students, respectively), constitute a clear indicator of a growing concern. Note that the exercise itself belongs to a discipline different from the group's major, so it is unlikely that the rating reveals a disciplinary issue.

Table 3. Exercise ratings

Rating	In-class	Remote
Challenging	1 (2.9 %)	0 (0 %)
Interesting	23 (67.7 %)	11 (73.3 %)
Normal	6 (17.6 %)	3 (20 %)
Boring	4 (11.8 %)	1 (6.7 %)

5.3 Efficacy Perception

When students had to complete the questionnaire, they still did not know if their re-identifications were correct. We asked them to guess how many players they had identified correctly. Table 4 compares their self-evaluated accuracy with the reality of how many players they had identified correctly. It turns out that their guesses are quite close to reality, which indicates that they had really understood the underlying processes. This result, combined with the fact that 48 out of 49 students identified 25 or more of the 30 players with no previous knowledge about re-identification techniques, shows that this is not a difficult process with common tools and data available today online.

Table 4. Self-evaluation of success in re-identification

Population	Real		Guessed	
	Average	Stand. Dev.	Average	Stand. Dev.
In-person	28,82	2,64	29,29	2,39
Remote	29,4	1,06	28,6	5,15
Whole group	29	2,28	29,08	3,43

5.4 Attitude Change: The In-class Interaction

We describe the subjective assessment of the authors about students' attitudes before and after the exercise.

In the first session, after our talk, most students agreed that they cared about their privacy, but considered that it was not at risk because of the information publicly available about them. Some even offered details on people they knew whose identity was stolen, but the general feeling was that it was something that happened to other people, and that they were safe. Even after running the mini-exercise of re-identifying characters in a fairy tale, their concern did not increase, as they claimed that it was "too easy".

A week later, after completing the exercise, the authors met with the students to pick up the answers and have a group discussion about the results. The mood of the group was more guarded, and some of the students expressed apprehension about their data online, and some of them explained that they did not realize that it was "so easy" to find a trove of information about people.

5.5 Discussion

The indicators used suggest that there may be an attitude change about privacy after executing the re-identification exercise, which was engineered to be relatively easy (it was weakly de-identified). Interestingly enough, it made no difference whether students received a pre-made group of de-identified data, or could devise queries by themselves. We thus believe that the most valuable part of the experience was the exercise itself, because it was hands-on [20].

The talk itself, which warned about what the students later discovered through the exercise, had no measurable effect, so it was surprising when some of the previously sceptic students expressed concern about privacy in the after-exercise discussion. Although we are just describing a single case, our results are consistent with constructivist literature in education [7].

We believe that the most valuable insight is that the task described, being very easy, but definitely practical, could be adapted to raise the privacy awareness of almost any person by executing a re-identification exercise.

We cannot say at this time whether any attitude changes will be maintained, which would be interesting to investigate in the future.

6 Conclusions

We presented a case where a group of students showed an attitude change about privacy issues after completing a hands-on re-identification exercise.

Almost all students (48 out of 49) were able to re-identify 25 players or more, even though they were not trained for the process, confirming the assumption that it is not very complex.

During the first session, students were not particularly interested in the exercise. After the students finished the re-identification process, some of them expressed more concern for their privacy in the post-exercise session. From the answers in the questionnaire, we know they were interested, which is in itself an attitude change. We argue that the interest about the exercise reflects an internal process that led to the greater concern expressed by some in the second session, which potentially suggests an attitude change about privacy.

The ease of resolution, plus the increased interest, points us in the direction that similar exercises could be used to raise awareness about privacy in the general population.

7 Future Work

We intend to test the results of the case presented with a larger, more diverse sample. However, the delivery described (talk, synthetic re-identification exercise, re-identification exercise with real individuals), is relatively expensive. We expect to eventually develop automated, online versions of similar exercises, which would allow us to reach a larger public.

Acknowledgments. This work was done for research project 834-B4-150 at Universidad de Costa Rica (UCR), the Research Center in Information and Communication Technologies (CITIC) and the Department of Computer Science (ECCI). Funding was also received from the Ministry of Science and Technology of Costa Rica (MICITT) and the National Council for Scientific and Technological research (CONICIT). Special thanks to Professor Hugo Solis, his FS0408 students, and the Physics Department at UCR.

References

1. Costa Rica líder en Latinoamérica en exportación de software — Cámara de Tecnologías de Información y Comunicación de Costa Rica (CAMTIC). http://www.camtic.org/hagamos-clic/costa-rica-lider-en-latinoamerica-en-exportacion-de-software/
2. PRODHAB — Agencia de Protección de Datos de los Habitantes. http://www.prodhab.go.cr/. Accessed 30 Nov 2015
3. Twitter sued over Hardy tweet. http://www.smh.com.au/technology/technology-news/twitter-sued-over-hardy-tweet-20120216-1tbxz.html

4. Reglamento a la Ley de protección de la persona frente al tratamiento de sus datos personales (2013). https://www.tse.go.cr/pdf/normativa/reglamentoley-proteccionpersona.pdf
5. Avendaño, A.: Protección de datos en el gobierno digital. In: Ciberseguridad en Costa Rica, pp. 349–356. Programa Sociedad de la Información y el Conocimiento, PROSIC. Universidad de Costa Rica, San José, October 2010. http://www.prosic.ucr.ac.cr/sites/default/files/documentos/ciberseguridad_en_costa_rica.pdf
6. Bolaños, R.: Acceso y uso de las tic en la administració pública, empresas y hogares. In: Informe Anual 2014. Programa Sociedad de la Información y el Conocimiento, PROSIC, pp. 123–156. Universidad de Costa Rica, San José (2014). http://www.prosic.ucr.ac.cr/sites/default/files/documentos/cap4_3.pdf
7. Bransford, J.D., Schwartz, D.L.: Rethinking transfer: a simple proposal with multiple implications. In: Review of Research in Education, pp. 61–100. American Educatonal Research Association, Washington, DC (1999)
8. Rica, C., Legislativa, A.: Ley de protección de la persona frente al tratamiento de sus datos personales (2011). http://www.tse.go.cr/pdf/normativa/leydeproteccion-delapersona.pdf
9. El Emam, K., Jonker, E., Arbuckle, L., Malin, B.: A systematic review of re-identification attacks on health data. PloS one **6**(12), e28071 (2011). http://dx.plos.org/10.1371/journal.pone.0028071
10. Garfinkel, S.L.: NISTIR 8053. de-identification of personal information. Technical report, National Institute of Standards and Technology, Gaithersburg, MD, USA, october 2015
11. Gopalan, R., Antón, A., Doyle, J.: UCON LEGAL. In: Proceedings of the 2nd ACM SIGHIT Symposium on International Health Informatics - IHI 2012, p. 227, No. 111. ACM, New York, January 2012. http://dl.acm.org/citation.cfm?id=2110363.2110391
12. Herrera, M.: Sala IV: Salario de los empleados del Estado es un dato público (2014). http://www.nacion.com/nacional/sala-iv/Sala-IV-Salario-empleados-publico_0_1404459700.html
13. MARCA.com: Keylor Navas spied on by the Costa Rican police. Marca (2014). http://www.marca.com/en/2014/10/29/en/football/real_madrid/1414591431.html
14. Marwick, A.E., Murgia-Diaz, D., Palfrey, J.G.: Youth, Privacy and Reputation (Literature Review) (2010). http://papers.ssrn.com/abstract=1588163
15. Mata, E.: Congreso señala débil protección de datos en el OIJ (2015). http://www.nacion.com/nacional/politica/Congreso-achaca-debil-proteccion-OIJ_0_1487051331.html
16. Narayanan, A., Shmatikov, V.: Robust De-anonymization of Large Sparse Datasets. In: 2008 IEEE Symposium on Security and Privacy (sp 2008), pp. 111–125. IEEE, May 2008. http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4531148
17. Parent, W.A.: Ethical issues in the use of computers. In: Privacy, Morality, and the Law, pp. 201–215. Wadsworth Publ. Co., Belmont (1985)
18. Preibusch, S.: Guide to measuring privacy concern: review of survey and observational instruments. Int. J. Hum. Comput. Stud. **71**(12), 1133–1143 (2013). http://www.sciencedirect.com/science/article/pii/S1071581913001183
19. Rhodes, M.G., Somvichian, W., Wong, K.C.: Google-Motion-to-Dismiss-061313. Technical report. Attorneys for Defendant GOOGLE INC. (2013)

20. Schneider, B., Wallace, J., Blikstein, P., Pea, R.: Preparing for future learning with a tangible user interface: the case of neuroscience. IEEE Trans. Learn. Technol. **6**(2), 117–129 (2013). http://dx.doi.org/10.1109/TLT.2013.15
21. Shklovski, I., Vertesi, J.: "un-googling" publications: the ethics and problems of anonymization. In: CHI 2013 Extended Abstracts on Human Factors in Computing Systems, pp. 2169–2178. CHI EA 2013. ACM, New York (2013). http://doi.acm.org/10.1145/2468356.2468737
22. Solove, D.J.: A taxonomy of privacy. Univ. Pennsylvania Law Rev. **154**(3), 477–560 (2006). http://www.jstor.org/stable/40041279, https://www.law.upenn.edu/journals/lawreview/articles/volume154/issue3/Solove154U.Pa.L.Rev.477(2006).pdf
23. Soto, J.: Diputados crearn comisin para investigar espionaje en el OIJ contra Keylor Navas. La Nación (2014). http://www.crhoy.com/diputados-crearan-comision-para-investigar-espionaje-en-el-oij-contra-keylor-navas/
24. Sweeney, L.: Uniqueness of Simple Demographics in the U.S. Population, LIDAP-WP4 (2000)
25. Westin, A.: Privacy and Freedom. Bodley Head, London (1970)