Vision: Minimalist approaches to enforce privacy by design in surveys

ANONYMOUS, No Institute, World

Public institutions and private companies both frequently rely on user surveys for a variety of assessments (e.g. equality issues or quality of work environment). However, many such surveys struggle to garner sufficient responses, especially when they ask about sensitive subjects (such as work harassment), which also makes them exist in a legal grey area when it comes to data protection laws. One important factor in this issue is the perceived threat of deanonymisation, compounded by the frequent lack of transparency on how the data is used. The proposals seeking to address this issue often focus on complex cryptography (e.g. homomorphic encryption), without addressing the fears of non-technical users. This paper explores a radically different approach which minimises data collection on multiple fronts, partially by limiting the power of survey organisers. By design, it prevents generic attempts to deanonymise participants as the server never stores even pseudonymised information. We also try to address questions of inclusivity, once again through a minimalist approach.

CCS Concepts: • **Security and privacy** → **Social aspects of security and privacy**; **Usability in security and privacy**; • **Theory of computation** → *Theory of database privacy and security*; • **Social and professional topics** → Privacy policies.

Additional Key Words and Phrases: Privacy by design, Survey, Methodology, User experience

ACM Reference Format:
Anonymous. 2018. Vision: Minimalist approaches to enforce privacy by design in surveys. In . ACM, New York, NY, USA, 7 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Workplace surveys are seeing increased use in both public institutions and private companies, to get employee feedback on indicators ranging from the quality of the work environment to diversity or harassment issues [20]. Those surveys suffer from multiple issues, especially low and further declining response rates (unless the survey is mandatory) and the risk of self-censorship [6]. This is particularly true when there is a risk that participants could be deanonymised, which has led to a lot of work in the field of differential privacy [7].

Due to the specificities of workplace surveys and the varying and sometimes conflicting regulations, some of the questions asked can also be in a legal grey area. For example, although medical data is subject to strict confidentiality rules, some institutes approve the use of questions about discrimination which ask the reason for such discrimination[1] (with the option to choose medical reasons or disability status).

[1]Some of the surveys we were given as examples of what was done previously were operated using LimeSurvey through RENATER (the French National Research Network). The RENATER terms of use expressly forbid questions on health and sexuality [19], rendering uncertain the status of questions on harassment for cause of sexual orientation or disability — which feature in some surveys.

'XX, 2018, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/XXXXXXX.XXXXXXX

Certain questions can also create some friction depending on how they're phrased. For example, asking the participant's gender runs into multiple issues:

- If a choice is given between only two options, many queer respondents would not feel included [22].
- If a choice is given between three options such as Man/Woman/Other, this can be felt as othering by those choosing the third option [2, 13].
- As long as at least 3 options are shown, the choice to have a third option (or more) can be perceived as political (especially where such options are not socially accepted), exposing the organisers to accusations of being biased (or sometimes even to harassment) [14]. Moreover, certain jurisdictions forbid the presence of more than two options, as do most research protocols [13, 22].

The authors of this paper were tasked with designing and developing a survey system that had strong privacy guarantees for a French university[2]. This short piece is the result of reflections and exchanges we had on how to combine minimalism and privacy by design in such a context [15, 21]. It is meant as a preliminary exploration and summarisation of multiple ideas, including a few that have been proposed elsewhere but are not standard yet, and a discussion of the new issues these ideas create.

2 SETTING GOALS

A first issue we observed with some past surveys is that many of them were designed in an ad hoc fashion, adding questions that seemed interesting without necessarily looking at interactions between them — or whether any given question adds any elements that cannot be inferred from other questions. This goes against good usability practices, as more questions translate directly to higher user cost and therefore a higher dropout rate.

This ad hoc approach to survey creation falls under the paradigm and characteristic patterns of big data and data mining. The first step in the process is collecting as much data as can possibly be collected, and all this data is then sifted for meaning and correlations [11, 16]. This already becomes problematic at the data collection stage due to the above-mentioned user cost. The data analysis that follows from this data collection method also encourages bad scientific practices such as p-hacking, which may not even need to be carried out consciously [10]. With the wealth of data, the researcher is free to test all the possible correlations until they find interesting ones, without necessarily using good statistical practices (such as Holm-Bonferroni methods [1]). This is especially true when administered by people with a background in psychology and little statistical training as some bad statistical practices have been considered standard by professional organisations [5].

To design a survey that avoids these issues, one must set specific information-gathering goals and select questions with these goals in mind. An important consideration for every question is whether we are interested in receiving qualitative feedback, i.e. open-ended answers, or whether we want to compute quantitative measures from the set of answers. Moreover, in some cases statistical analysis may not be possible due to small sample size or low response rate, which renders questions that rely on returning a quantitative measurement useless to the survey's goals. One must be explicit with what each question brings to the survey, what it is meant to measure, and how this measurement will be carried out (which is a good practice to be borrowed from the social sciences). Ideally, this means that the survey will be pre-registered, with the methods and stated goals recorded before it is conducted, and ideally made available to participants for transparency [17].

[2]The system is currently being tested within the university with results to be published at the end of 2022. The system is then supposed to be released as free software.

An additional important reason to be mindful with the phrasing and inclusion of specific questions is that, even if the immediate impact of a given survey is limited, its framework may be reused in the future, especially in public institutions. This means that any badly phrased elements will be carried forth into an indeterminate number of future surveys, become standardised and have long-lasting consequences [22]. This situation can then be hard to correct, as simply removing a badly phrased question from a survey can lead to knock-on changes due (for instance) to priming.

3 PREVENTING DEANONYMISATION

One of the main risks when handling workplace surveys concerns some of the sensitive data being attributable to specific individuals, especially if the data gets leaked. This can happen if data is nominative or if it is possible to deanonymise participants.

A first way to deanonymise occurs when the survey's organisers have access to full answer sheets for each participant (or if the full data gets leaked). It can then be trivial to find the single person who answered a certain way, which increases the barrier to reporting sensitive information.

Even when people do not have access to full answer sheets, it can still be possible to deanonymise participants is to look at correlation chains. For example, let's suppose we have a single participant reporting harassment. If we have access to the average age and gender of people reporting harassment, we have initial elements which, if that person is the only one in their age-gender category, would allow us to get increasingly more information and to eventually build a profile. A first step to prevent this is to limit the use of precise data (such as age) by having wide categories (such as age brackets).

Fully preventing correlation chains is generally not possible from within the system, as doing so requires contextual information, such as the number of people from a certain demographic who occupy a specific position. It can sometimes be possible to ensure that a correlation chain is impossible no matter the context, but this requires larger survey populations — and is the context where differential privacy is often explored [7, 8].

In the case of small survey populations (i.e., around 100), we still have multiple ways to address deanonymisation using a minimalist approach.

3.1 Exchangeable codes

The easiest way to deanonymise is to obtain nominative data, such as the participant's email. This is not always stored as part of the answer sheet, but a unique identifier or password is commonly sent to users to prevent spamming and limit answers to one per person. If the password is sent by email, it creates the opportunity for organisers to directly attribute answers to known email addresses, and participants can then have legitimate privacy concerns as they cannot know whether organisers are able to track their answers.

One way to address this is to use simple passwords or passphrases — for example, two common words — and to tell users that they are free to exchange them with colleagues (although each code can only be used once). Moreover, those passwords should be checked once when the survey data is submitted (or when access is granted), but should ideally not be stored as part of the same database.

3.2 Decorrelating answers

Once the nominative data is removed, the next step to avoid deanonymisation is to avoid storing full answer sheets and never keep a database where each entry corresponds to a participant (and all their answers). Instead, each question should be stored separately. To prevent the possibility of recreating answer sheets, that means that each time data is stored for a question (that isn't a counter), it should be reordered randomly. Instead of reordering the whole array, it is thankfully

enough to only permute the last element added with another (including itself) uniformly, akin to a reversed Fisher-Yates shuffle [4].

3.3 Client-side correlations

The problem with the previous method is that if used naively, it only allows some descriptive statistics — getting the proportion of people unhappy about a particular element — and qualitative feedback. However, it does not allow the study of any between-groups differences — such as whether one gender has different work experiences. Whereas one can compute arbitrary correlations when one has full answer sheets, a minimalist decorrelated approach makes it impossible.

One solution is to establish beforehand a full list of all desired correlations in a way akin to pre-registration [17]. Those correlations can then be computed on the client's side and sent to the server. For example, if one wanted to correlate age and satisfaction (on any given subject), the client would send three pieces of data to the server ("age","satisfaction","age-satisfaction"). Only the pre-established correlations would then be available, and the list should be made public for transparency. Correlations between more than 2 variables can also be recorded but each additional variable makes deanonymisation easier. Even with only 2-variable correlations, care should also be taken to avoid correlation chains when designing the questions — while keeping the context in mind and how a single question could deanonymise certain persons if the sample set is small enough.

3.4 Avoiding partial results

There is one way to deanonymise participants even if the system uses the previous elements. By observing the results at multiple points in time (ideally between each participant), it becomes possible to infer the full answer sheets. A way to prevent this is to only make the results available once the survey is finished.

Of course, this raises the question of who has access to the server and the administration interface. Any person with physical access to the server has a high chance of being able to obtain the data — unless everything is secured through trusted platform modules with correct cryptography, and even this supposes resistance to side-channel attacks which is not guaranteed [12, 18].

A reasonable attacker's profile in such a context is someone — potentially a manager or an employee from human resources — trying to access data about their colleagues, hence with reasonably limited technical ability. One solution would then be to host the survey externally, or at least on a server administered by someone with no links to the participants. We can then differentiate between the person with complete server access — who has arbitrary power over the survey but no motive[3] — and the organiser. The latter should only be authorised to input the questions as well as the list of participants' emails (to access the survey), end the survey, and publish or download the results.

3.5 Post-survey correlation chains elimination and question twinning

If one is given preliminary contextual data (such as demographic information), it can become possible to perform additional correlation checks at the end of the survey. For example, let's suppose a sensitive question is correlated to a few demographic categories including gender and age, and let's suppose it is known that only three 60 year-old men work for the company and no 60-year old women do. If no women answer yes to the question but the three men do, they can be

[3]Of course, it might be possible to bribe or coerce the administrator. However, this means that the bribing party exposes themself if the administrator reveals the attempt. Keeping the administrator's identity private to most would also limit the exposure — and reduce the set of potential guilty parties if a bribe is attempted.

deanonymised (thanks to the contextual knowledge). However, uncertainty would remain if one of them did not answer (or answered differently), so the correlation by itself is not at fault.

It can be possible to automatically detect such cases (by feeding the system some contextual information initially or by guaranteeing that everyone in the population participated in the survey, giving it total demographic information). However, just removing the question from the survey is not a solution as sometimes only one answer set can be deanonymising, in which case removing the question is just as deanonymising. One option is then to analyse a priori which questions could lead to such cases and to twin them: if at least one of them is removed, then the other also is. This eliminates the risk by creating an ambiguity, although at the expense of additional data loss.

4 USABILITY AND INCLUSION

After introducing multiple ideas to prevent deanonymisation and increase participant's trust (and hence participation), it seems natural to look at a few usability elements which can also play a role in decreasing user cost (and hence dropout rates).

4.1 Self identification

For the questions that can create some friction depending on which categories are available — such as gender — a simple solution is to leave an open field for all participants (and not just for those who'd choose "other"). This allows for self-identification without making the inclusivity visible (and thus potentially avoiding some political fallout).

There remains the question of how to handle the correlations when one has an open field to correlate. Here, we must arbitrate between two different solutions, each with some drawbacks:

- We can register the detailed field with each correlation, but it can easily deanonymise certain participants (especially if some identities are rare, e.g. there is only one non-binary employee). It also makes future analyses more complex (depending on how the data is eventually clustered[4]).
- Another option is to parse the data immediately (with an extensive but non-exhaustive initial list) into a few categories, for example "Woman", "Man", "Other" and "Did not respond" (the latter two can be combined). This is somewhat more inclusive than just having an "Other" option within the survey (as participants aren't directly facing it) and facilitates the correlation analyses. It also has stronger privacy guarantees.

4.2 Cookies

Due to the survey not saving full user sheets, any problem in the database could corrupt the data in a way that cannot be handled by simply simulating the inputs on the server side. More importantly, if a correlation that was meant to be measured failed, it is impossible to get it back from the available data — unlike with the other structure where the organisers can choose what to analyse a posteriori.

One of the solutions is to ask the participants to retake the whole survey, but that has a high user cost and compounds with the dropout risk. Another option is to store all the participants' data on the client's side (as a full sheet in a cookie). In case of a server-side issue, it then becomes possible to update the code then ask users to go back to the survey page and resubmit their original data — plus eventual new correlations computed on the client's side.

This does create some security and privacy risk depending on the exact context. It can also have a small usability cost as the cookie storage requires compliance with various regulations

[4]If the data is not meant to be clustered at all, then one can question whether it should be correlated at all.

such as GDPR[5] [23]. The cookie information being available in cleartext on the client machine is a privacy risk, which can be mitigated by encrypting the cookie data. This can be done either in an asymmetric way — in which case the server is asked to decrypt the data upon a second login — or using symmetric encryption without storing the key on the client's machine except during the session. If there is a risk of a participant stealing another participant's cookies, the password should be different for all users (and could be partially based on the user's password).

4.3 Data modification and deletion

Another issue with avoiding user sheets is that it is not directly possible to remove or modify one user's data. However, if one is using cookies as above, then an option becomes available — with an additional security risk if anyone has full access to the server, although as stated above this renders most points moot. In addition to the decorrelated user data, the server can store a hash[6] of each sheet (including the password). Then if a user tries to login back into the system, they can go into a special modification mode where the client keeps the old cookie with the old answers, and send one message to the server with the initial list of answers and the corresponding hash (which the server checks before deleting each answer from the corresponding database) and a the corrected list of answers. The list of hashes should in any case not be public and should be deleted when the final results are computed to prevent future bruteforce attempts.

5 CONCLUSION

Many of the ideas above were implemented and we expect some user feedback from the test that is due to end by autumn 2022. Some proposals — such as question twinning — were left out because they seemed too complex for the task at hand.

We also found multiple questions, social and technical, for which we have no good answers and which could be investigated in the future. Here are the main ones:

- As transparency is good for trust and we advocate for publishing in advance the question and correlation list. However, the question remains of whether to automatically make public (to the participants) the results of the study once it is completed, and the corresponding trade-offs deserve an analysis.
- Our model assumes that an attacker does not have direct access to the server (as it is generally beyond the technical purview of the survey organisers to prevent such attacks). For situations where higher security is required, this vulnerability needs to be addressed. A simple method is to have a double system where two machines in different locations are in continuous contact with each other and the Internet. If contact is broken at any point or if someone tries to access one of the machines physically, a public alert is sent (for example by email, on Twitter or a blockchain). This is already a better system but it is extremely prone to false alarms (and to denial-of-service attacks). Using systems such as proactive secret sharing [9] as well as TPMs, could a distributed encrypted system avoid this issue while being able to recover when one machine fails?

[5]One option is to have at the end of the survey a checkbox with the option to store the data locally if the user wants to, in which case a GDPR cookie warning wouldn't be necessary.

[6]It would be preferable to use a costly hash to resist bruteforce (e.g., Argon2 [3]), as it is computed rarely.

REFERENCES

[1] Hervé Abdi. 2010. *Holm's Sequential Bonferroni Procedure.* SAGE Publications.

[2] Greta Bauer. 2012. Making Sure Everyone Counts: Considerations for Inclusion, Identification and Analysis of Transgender and Transsexual Participants in Health Surveys. https://open.library.ubc.ca/cIRcle/collections/facultyresearchandpublications/52383/items/1.0132676

[3] Alex Biryukov, Daniel Dinu, and Dmitry Khovratovich. 2016. Argon2: new generation of memory-hard functions for password hashing and other applications. In *IEEE European Symposium on Security and Privacy – EuroS&P*. IEEE, 292–302.

[4] Paul E. Black. 2019. Fisher-Yates shuffle. *Dictionary of Algorithms and Data Structures* 19 (2019).

[5] Denny Borsboom. 2006. The attack of the psychometricians. *Psychometrika* 71, 3 (2006), 425.

[6] Don A. Dillman. 2020. *Towards Survey Response Rate Theories That No Longer Pass Each Other Like Strangers in the Night.* Springer International Publishing, Cham, 15–44. https://doi.org/10.1007/978-3-030-47256-6_2

[7] Cynthia Dwork. 2008. Differential privacy: A survey of results. In *International conference on theory and applications of models of computation*. Springer, 1–19.

[8] Úlfar Erlingsson, Vitaly Feldman, Ilya Mironov, Ananth Raghunathan, Kunal Talwar, and Abhradeep Thakurta. 2019. Amplification by shuffling: From local to central differential privacy via anonymity. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms*. SIAM, 2468–2479.

[9] Yair Frankel, Peter Gemmell, Philip D MacKenzie, and Moti Yung. 1997. Optimal-resilience proactive public-key cryptosystems. In *Proceedings 38th Annual Symposium on Foundations of Computer Science*. IEEE, 384–393.

[10] Andrew Gelman and Eric Loken. 2013. The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time. *Department of Statistics, Columbia University* 348 (2013).

[11] Nitza Geri and Yariv Geri. 2011. The Information Age Measurement Paradox: Collecting Too Much Data. *Informing Sci. Int. J. an Emerg. Transdiscipl.* 14 (2011), 47–59.

[12] Dan Goodin. 2021. Trusted platform module security defeated in 30 minutes, no soldering required. Ars Technica. http://web.archive.org/web/20220523092016/https://arstechnica.com/gadgets/2021/08/how-to-go-from-stolen-pc-to-network-intrusion-in-30-minutes/

[13] Haute Autorité de Santé. 2020. *Sexe, genre et santé.* Technical Report. Haute Autorité de Santé. https://www.has-sante.fr/upload/docs/application/pdf/2020-12/rapport_analyse_prospective_2020.pdf

[14] S. Jaroszewski, D. Lottridge, O. L. Haimson, and K. Quehl. 2018. "Genderfluid" or "Attack Helicopter": Responsible HCI Practice with Non-Binary Gender Variation in Online Communities. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/3173574.3173881

[15] Marc Langheinrich. 2001. Privacy by design—principles of privacy-aware ubiquitous systems. In *International conference on ubiquitous computing*. Springer, 273–291.

[16] Alec Levenson and Alexis Fink. 2017. Human capital analytics: too much data and analysis, not enough models and business insights. *Journal of Organizational Effectiveness: People and Performance* (2017).

[17] Jennifer M. Logg and Charles A. Dorison. 2021. Pre-registration: Weighing costs and benefits for researchers. *Organizational Behavior and Human Decision Processes* 167 (2021), 18–27. https://doi.org/10.1016/j.obhdp.2021.05.006

[18] Daniel Moghimi, Berk Sunar, Thomas Eisenbarth, and Nadia Heninger. 2020. {TPM-FAIL}:{TPM} meets Timing and Lattice Attacks. In *29th USENIX Security Symposium (USENIX Security 20)*. 2057–2073.

[19] RENATER. 2020. Conditions d'utilisation du service d'enquête. http://web.archive.org/web/20201203200601/https://services.renater.fr/groupware/enquetes/conditions

[20] Paul M. Sanchez. 2007. The employee survey: More than asking questions. *Journal of Business Strategy* (2007).

[21] Sarah Spiekermann. 2012. The challenges of privacy by design. *Commun. ACM* 55, 7 (2012), 38–40.

[22] Mathieu Trachman, Tania Lejbowicz, and Katharine Throssell. 2018. Putting LGBT and non-binary people in boxes. Statistical categorization and criticism of gender and sexuality assignations in a study on violence. *Revue française de sociologie* 59, 4 (2018), 677–705.

[23] Razieh Nokhbeh Zaeem and K Suzanne Barber. 2020. The effect of the GDPR on privacy policies: Recent progress and future promise. *ACM Transactions on Management Information Systems (TMIS)* 12, 1 (2020), 1–20.