BRIDGE: Administering Small Anonymous Longitudinal HCI Studies with Snowball-Type Sampling

Frode Eika Sandnes

BRIDGE: Administering Small Anonymous Longitudinal HCI Studies with Snowball-type Sampling

Frode Eika Sandnes[1,2][0000-0001-7781-748X]

[1] Oslo Metropolitan University, 0130 Oslo, Norway
[2] Kristiania University College, 0153 Oslo, Norway
frodes@oslomet.no

Abstract. When following participants across multiple sessions one needs a way to link the different session records while protecting the participants' privacy. Privacy is required by recent legislation such as the General Data Protection Regulation (GDPR). Many anonymous linking methods have been proposed, but these involve effort from the participants, involve long IDs, or require all participants to be known a priori. This study presents the BRIDGE procedure for anonymous linking of participant records with dynamically increasing samples. The procedure relies on human intervention to resolve ambiguous cases using manual recognition challenges. Simulation results show that the procedure can successfully map participant names to short and unique anonymous IDs, and that the percentage of human interventions is low. The procedure holds potential for HCI researchers who need to employ simple, flexible, and incremental sampling strategies while protecting participants' privacy.

Keywords: privacy, longitudinal, GDPR, record linking, snowball sampling

1 Introduction

Participants' anonymity is a key concern for HCI experimenters. No information should be recorded that directly identifies the participant including their name, phone number, IP-address, voice, photos, video, or information that can indirectly reveal participants' identity such as demographic or geographic patterns. Anonymity is especially essential when recruiting participants with disabilities [1, 2, 3]. Anonymity is usually not a challenge when administering experiments that can be conducted in single sessions [4]. However, the challenge arises once the experimenter needs to follow participants over time. For example, pre/post-test experiments [5] may require results from two sessions to be linked to perform pairwise analysis. Longitudinal studies [6, 7, 8] follow participants over time to observe how participants learn a particular interaction mechanism [9, 10]. To analyze the data, the session observations need to be linked.

Obviously, labelling the observations from each session with the participants name is not an option unless strict regimes are in place to protect the data. Traditionally, experimenters employed linking tables where each participant is assigned a unique running number. Observations were labelled with these IDs and the linking table was kept

separate and confidential. This allowed observations to be shared or made public while keeping the identity of the participants anonymous. Stakeholders were later able to link data from different sessions without knowing the identity of the participant. However, if the linking table is leaked, the privacy of the participants is compromised. Privacy legislation such as GDPR regulates the storage of personal information. Typically, experimenters need to apply for formal permissions to administer such tables and document that there are convincing safeguards in place and reliable procedures for disposing the tables at the end of the project. Obtaining formal permissions can be time-consuming and daunting for students and HCI researchers who are administering their first experiments. One potential consequence may be that the experimental design is altered so that it can be conducted anonymously in a single session. This is unfortunate if the research question warrants the participants to be followed over time.

Self-generated codes [11] is one approach for overcoming this challenge, where each session starts with the participants answering a questionnaire where the responses are used to generate IDs that are unique to each participant. Unfortunately, self-generated codes divert valuable time and effort away from the experiment. The other approach involves Bloom filters [12] which is a type of hash function that are robust to input errors. The participants' names are applied to a series of Bloom filters and the result is assigned the observations from the session. Bloom filter IDs are typically long (1000 bits) and may be impractical to handle manually (low usability). It has also been shown that the basic Bloom filter approach is vulnerable to systematic attacks [13].

The HIDE procedure [14] attempts to overcome the problem of long IDs, while allowing the IDs to be generated on-the-fly without collisions and intervention from the participants. One key constraint is that HIDE configuration requires all the participants to be known in advance. In practice, an experimenter may have to recruit participants incrementally. For example, snowball sampling is often employed by HCI researchers where the experimenter expands the sample of participants using the network of the already recruited participants. HIDE does not facilitate snowball sampling.

This study therefore proposes the BRIDGE procedure, which allows experiments to be administered with incremental snowball-type sampling while generating short IDs. Names that lead to collisions are resolved manually using recognition challenges. Simulations were used to measure the practical limitations of the procedure.

2 Related work

Strategies for record linkage and related work are summarized in Table 1 with five key characteristics including anonymity, robustness to error, perceived trust, effort from participants and dynamic expansion. The most important characteristic is the capability to protect participants' privacy and maintain anonymity. Robustness to error has also been discussed in the literature [11, 14] as incorrect record linkage may bias the results [15]. Participants' names may for instance be incorrectly transcribed. A linkage procedure that relies on exact matching is therefore vulnerable. Participants' trust in the procedure affects participants' willingness to participate in experiments [16]. Trust and participants' experience with linking procedures have received little attention. Related

to participants' trust is also the effort required from the participants. Recruiting participants can be hard and experimenters must balance the size of the tasks against what is realistic to expect given people's general impatience. Clearly, the goal is to prevent diverting attention away from the experiment. Finally, a linkage procedure may be open or closed. An open procedure allows new participants to be added dynamically, while a closed procedure requires all participants to be known a priori.

Table 1. Characteristics of record linkage approaches.

Method	Anonymity	Robust	Trust	Effort	Flexible
Direct labelling	None	Yes	None	None	Open
Linking table	Risky	Yes	Low	None	Open
Participants remember ID	Strong	No	High	Medium	Open
Anonymous login	Strong	No	High	High	Open
Self-generated ID [17]	Strong	Yes	High	High	Open
Auto-generated ID [26]	Weak	No	Low	None	Open
Phonetic encoding [27]	Weak	Yes	High	None	Open
Ordinary hash [33]	Weak	No	Low	None	Open
Bloom filter [5]	Strong	Yes	Low	None	Open
Perfect minimal hash	Unknown	No	High	None	Closed
HIDE [14]	Strong	Yes	High	None	Closed

Clearly, labelling data sets with the identity of the participants is straightforward but does not provide anonymity nor contribute to the participants' trust [16]. Linking tables are also simple to administer and do provide anonymity if the linking table is kept private, but participants' privacy is compromised if the linking table is lost.

Another approach is to randomly generate IDs and ask participants to remember their own ID. When the participant attends a session, they must produce their unique ID. This approach is simple and provides anonymity, but there is a risk that they unintentionally or intentionally report an incorrect ID. More importantly, participants may forget their ID, or if they write it down, lose the note. Sometimes it may be possible for participants to identify themselves using some third-party login credentials (such as national authentication schemes, social media accounts, etc.). However, such schemes require that participants have access to credentials, are willing to use their credentials, remember their credentials, or are willing to create an account if they do not have one.

To reduce the participants' memory load various procedures for self-generated codes have been proposed [17-23]. A self-generated code is generated by combining the answers to a questionnaire, for example the third letter of the name of the participant's mother, the number of siblings, the month of birth, etc. Each session is typically started by generating the code using the questionnaire. Clearly, this diverts valuable effort and time away from the experiment itself. Moreover, the questions should result in consistent responses. Self-generated codes have been found to be problematic in certain situations [24], vulnerable to errors [11] and vulnerable to attacks [25].

Other methods involve generating IDs using information about the participants, typically their name, birthdate, gender, or other available information [26]. Early work employed phonetic simplifications of the participants' names using Soundex [27-30], whereby the spelling of a name is converted to brief phonetic codes. The advantage of such schemes is that they are robust to certain types of errors, i.e., vowels and double

letters are discarded, and consonants are assigned into coarse grained categories of similar sounding sounds such as *b*, *f*, *p* and *v*. Matching is possible even if a consonant is substituted for a similar sounding consonant. Although phonetic coding are lossy processes, they do not provide anonymity, unless specific mechanisms are employed such as adding fake records [31]. Phonetic methods are also known for resulting in high false positive rates [32]. Hashing is another method for generating unique IDs [33] and phonetic methods have been extended to improve anonymity [34, 35]. Although it is not possible to identify the person from a hash the identity can be discovered if searching for the ID using a list of names (phonebook attack). Hashes are therefore not anonymous. In fact, hashes can be used to confirm that a person participated in an experiment.

Much of the literature published during the last decade has evolved around various Bloom filter approaches [12, 36]. A Bloom filter employs a series of different hash functions to the bigrams (pairs of letters) making up the participants' names. These hash functions set certain bits in a bit vector. Matches are performed by comparing these bit vectors, and degrees of similarity is related to the number of matching bits. The strength of Bloom filters is their robustness to input errors and their flexible deployment at a large scale. They are therefore used in the domain of automatic record linkage. However, Bloom filters have been identified as being vulnerable to attacks [37-40]. Bloom filters generate long IDs, typically 1000 bits, or 250 hex characters. Hash sequences of 250 characters may seem overwhelming compared to running numbers which typically comprise two or three digits. The concept of mapping a set of names to running numbers is analogous to what is achieved with minimal perfect hash functions [41, 42]. However, minimal perfect hashing requires all records to be known a priori and a hash table needs to be stored. Research has shown that the average hash table entry is short (less than 2 bits) [41, 42]. However, to the best of our knowledge, perfect minimal hash functions have not been applied to record linking. Inspired by minimal perfect hash functions the HIDE procedure generates close to running numbers [14]. Instead of running numbers, the range of numbers is expanded with the benefit of not needing a hash table. This procedure first alphabetically sorts the part of the participants' names (first, middle and family name) and then phonetically encodes the name using Soundex, making the procedure robust to many types of input errors. A salt, i.e., a random text, is added to the phonetic representation and the result is hashed. The ID is obtained by truncating the hash. This truncation step provides anonymity in that different names will yield the same ID. The initialization step of HIDE searches for a salt that provides minimum length unique IDs for a list of names.

The goal of this proposal was to reap the benefits of HIDE while facilitating dynamic inclusion of participants. This study focuses on small experiments with less than 100 participants as HCI experiments typically are small (often just 12 participants) [44].

3 The BRIDGE Procedure

The procedure proposed herein builds on the HIDE procedure [14]. First, the participant's name is sanitized for non-alphabetical symbols (hyphens, dots, etc.) and then

split into first name, middle name, and family name. These parts are sorted alphabetically (tolerant to different name orderings). Next, the name is phonetically coded using Soundex, but without the four-symbol limit. The Soundex representation is then hashed (using djb2). Finally, the resulting hash is truncated by retaining the *d* last digits. These *d* digits comprise the participant's ID. An encoding example is shown in Fig. 1.

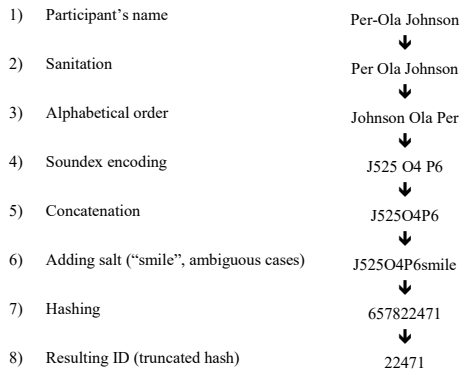

Fig. 1. BRIDGE encoding steps.

We assume that the experimenter knows if they encounter a new or a returning participant. Each time a new participant is added the resulting ID is stored in a list of IDs. If the ID already exists, the collision needs to be resolved. This is achieved by searching for a salt that results in an ID that is not already occupied. The salt is a simple text string (based on a list of common English words) that is added to the Soundex representation prior to hashing (see Fig. 1). The identified salt is then associated with the original ID. The experimenter and/or participant is asked to remember that they will be challenged with this word (the salt) in the future.

The IDs of returning participants are looked up as follows. First, the ID is encoded using the procedure above. If there is no salt associated with the resulting ID *A* there is no collision, and the correct ID has been identified as *A*. If there is a salt identified with the ID, the procedure first attempts to resolve the ambiguity automatically. If the IDs *B*, *C*, *D*, etc., obtained by encoding the participants name with the salts listed, does not match any IDs on the list we can be certain that the original ID *A* belongs to the participant. However, if one of the salts leads to a match, we cannot be certain which ID is correct, and the experimenter needs to be consulted. The experimenter is then confronted with the salts, and the experimenter and/or participant must determine if they previously were given the task of remembering any of these salts. If this participant were not asked to remember the salt, we know that the original ID *A* is correct. Otherwise, if the participant recognizes one of the salts (if there are more than one), the ID is given by resulting ID *B*, *C*, *D*, etc., resulting from applying the recognized salt.

Table 2 shows an example ID table with two salts (three collisions). Imagine we want to find the ID of "Per-Ola Johnson". We first computed the ID which is 22471. The table shows that there are no salts listed for this ID, and we have a unique match. Next, imagine we want to find the ID of "Lena Hansson". We first computed the ID

99175. This ID is associated with two salts. We then compute the IDs of "Lena Hansson" with the two salts and get 82023 and 61955, respectively. As none of these IDs are listed, we have identified the unique ID of "Lena Hansson" as being 99175.

Next, imagine we look up the ID of "Gunnar Green" in another session and find that the corresponding ID 99175 has two salts. We find that when applying the salt "sand" we get the valid ID 26294. We therefore are unable to automatically determine if the participants ID is 26294 or 99175. However, the participant confirms that she was told to remember "sand", and we therefore know that the correct ID is 26294.

Table 2. ID lookup example.

ID	Salts
22471	
26294	
99175	elevator, sand
32512	

4 Method

To assess the proposed procedure a simulation was conducted as this allowed many cases to be evaluated. The lists of names used in [14] derived from [44] were chosen as base populations of which random samples could be drawn. A few errors in the lists were removed and the resulting list comprised 103,472 unique names.

Sample sizes were varied from 5 to 95 in increments of 5 participants, and the ID lengths were varied from 1 to 7 digits. Clearly, the simulations were not performed for sample sizes above 10 with one digit as it is not possible to uniquely map more than 10 participants using one decimal digit. For each configuration, the simulation was repeated 10,000 times, each time with a random draw of participants. The goal of the simulation was to determine the success rate of the BRIDGE procedure, the manual intervention rate, and how many ambiguous cases that could occur at once. The simulations were written in Java.

5 Results

The simulation results show the BRIDGE procedure was able to successfully resolve all IDs (100% success rate). This simulation assumed that the experimenter and participant responded correctly to the challenges.

Fig. 2 shows the probability of having to manually intervene with BRIDGE under different conditions. A log-linear plot was chosen to emphasize the small probabilities. The probability of having to manually intervene increased linearly with the sample size. Moreover, the rate of increase was strongly related to the length of the ID. With IDs comprising 1 or 2 digits the probability of manually having to resolve ambiguities was relatively high (more than 20% with more than 20 participants). With IDs comprising 3 digits, the probability of less than 14% with 95 participants, and less than 1% with 20

participants or fewer. With 4 digits the probability of manual intervention did not exceed 0.4%, and with 5 digits this probability did not exceed 0.05%. The reductions in the probability for manual intervention is marginal with 6 and 7 digits.

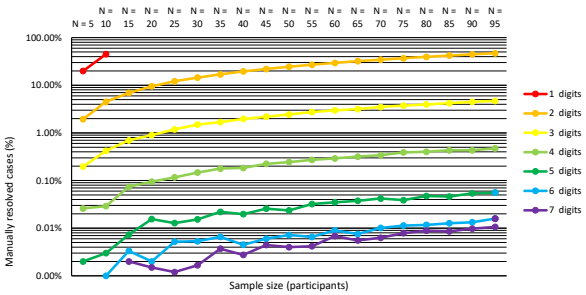

Fig. 2. Log-linear plot of manual resolution percentages with BRIDGE for different ID lengths (10,000 trials, zero values for 6- and 7-digit lines are not shown in the log-plot).

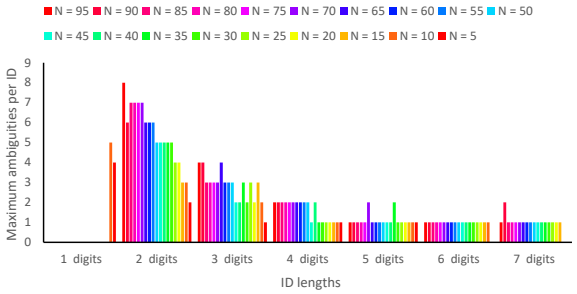

Fig. 3. Maximum number of ambiguities per ID as a function of ID length and sample size (10,000 trials).

Fig. 3 shows the maximum number of ambiguities observed for each configuration. Note that 1 ambiguity means there are two items that need to be resolved, with 2 ambiguities there is a set of 3 ambiguous items, etc. The plot shows that the maximum number of ambiguities is related to the length of the IDs. With a length of 1 digit the maximum number of ambiguities is 4 and 5, and with 2 digits the maximum number of ambiguities ranges from 8 to 2, where the maximum number of ambiguities is higher for larger samples. With 3 digits the max ranged from 2 to 4, while with 4 or more digits the max never exceeded 2. With 5 digits there were mostly a max of one ambiguity.

6 Discussion

The simulation results demonstrate that the procedure is successfully capable of establishing a unique mapping between participants and the IDs, and the short ID codes ensure that this mapping is anonymous. The procedure assumes that the experimenter and or participants can resolve the manual challenges. These challenges rely on recognition

which is less demanding than recall. It is much easier to recognize a word you have been told to remember than to recall the same word.

However, the simulations show that the procedure in more than 99% of cases can be administered without the need for challenges if the IDs comprise 5 digits. IDs with 5 digits are within the classic 7+/-2 short-term memory limit [45]. The extent of manually resolving challenges is therefore moderate. IDs with 5 digits will provide sufficient protection against most phonebook attacks, given the phonebook is large, since several names will lead to the same IDs making it impossible to confirm with certainty that a given person was a participant in an experiment (high k-anonymity) [46].

If the experimenter or participant are unable to respond to the challenge, it may still be possible for the experimenter to manually piece the parts together by using a combination of timestamps in the records and the experimenter's memory. Failing that, statistical testing procedures should be sufficiently robust to give correct conclusions with two incorrectly linked data points, provided the sample size is sufficiently large. Alternatively, the experimenter may choose to discard such observations from the analysis.

7 Conclusion

The BRIDGE procedure for incrementally generating anonymous IDs was presented. The procedure requires the experimenter to resolve ambiguous cases using manual word recognition challenges. The procedure generates a challenge (salt) that the experimenter and/or the participant recognize or not. The phonetic encoding means that the procedure is robust to several types of input errors. Simulation results show that IDs with 5 digits results in less than a 1% chance of manual intervention, while providing a high level of anonymity. Simulations demonstrated that the procedure can be used with up to 95 participants, but the procedure is likely to successfully handle larger sample sizes. An implementation of the procedure is available as a browser-based tool (https://www.cs.oslomet.no/~frodes/BRIDGE/). The source is also available at (https://github.com/frode-sandnes/BRIDGE). Future work involves developing a scheme for resolving collisions automatically.

References

1. Berget, G., Mulvey, F., Sandnes, F. E.: Is visual content in textual search interfaces beneficial to dyslexic users?. International Journal of Human-Computer Studies **92**, 17-29 (2016).
2. dos Santos, A. D. P., Medola, F. O., Cinelli, M. J., Ramirez, A. R. G., Sandnes, F. E.: Are electronic white canes better than traditional canes? A comparative study with blind and blindfolded participants. Universal Access in the Information Society **20**, 93-103 (2021).
3. Sankhi, P., Sandnes, F. E.: A glimpse into smartphone screen reader use among blind teenagers in rural Nepal. Disability and Rehabilitation: Assistive Technology, (2020).
4. Aschim, T. B., Gjerstad, J. L., Lien, L. V., Tahsin, R., Sandnes, F. E.: Are split tablet keyboards better? A study of soft keyboard layout and hand posture. In: IFIP Conference on Human-Computer Interaction, pp. 647-655. Springer, Cham (2019).

5. Kaushik, H. M., Eika, E., Sandnes, F. E.: Towards Universal Accessibility on the Web: Do Grammar Checking Tools Improve Text Readability?. In: International Conference on Human-Computer Interaction, pp. 272-288. Springer, Cham (2020).
6. Vissers, J., De Bot, L., Zaman, B.: MemoLine: evaluating long-term UX with children. In: Proc. 12th Int. Conf. Interaction Design and Children, pp. 285-288. ACM (2013).
7. Jain, J., Boyce, S.: Case study: longitudinal comparative analysis for analyzing user behavior. In CHI'12 Extended Abstracts, pp. 793-800. ACM (2012).
8. Karapanos, E., Zimmerman, J., Forlizzi, J., Martens, J. B.: User experience over time: an initial framework. In: Proc. SIGCHI CHI'09 conferenc, pp. 729-738. ACM (2009).
9. Ye, L., Sandnes, F. E., MacKenzie, I. S.: QB-Gest: qwerty bimanual gestural input for eyes-free smartphone text input. In: International Conference on Human-Computer Interaction, pp. 223-242. Springer, Cham (2020).
10. Sandnes, F. E.: Can spatial mnemonics accelerate the learning of text input chords?. In: Proc. working conference on Advanced visual interfaces, pp. 245-249. ACM (2006).
11. Schnell, R., Bachteler, T., Reiher, J.: Improving the use of self-generated identification codes. Evaluation Review **34**(5), 391-418 (2010).
12. Schnell, R., Bachteler, T., Reiher, J.: Privacy-preserving record linkage using Bloom filters. BMC medical informatics and decision making 9(1), (2009).
13. Christen, P., Schnell, R., Vatsalan, D., Ranbaduge, T.: Efficient cryptanalysis of bloom filters for privacy-preserving record linkage. In: Pacific-Asia Conference on Knowledge Discovery and Data Mining, pp. 628-640. Springer, Cham (2017).
14. Sandnes, F.E.: HIDE: Short IDs for Robust and Anonymous Linking of Users Across Multiple Sessions in Small HCI Experiments. In: CHI '21 Conference on Human Factors in Computing Systems Extended Abstracts Proceedings. ACM (2021).
15. Harron, K., Dibben, C., Boyd, J., Hjern, A., Azimaee, M., Barreto, M. L., Goldstein, H.: Challenges in administrative data linkage for research. Big data & society **4**(2), (2017).
16. Audrey, S., Brown, L., Campbell, R., Boyd, A., Macleod, J.: Young people's views about consenting to data linkage: findings from the PEARL qualitative study. BMC medical research methodology **16**(1), (2016). DOI: https://doi.org/10.1186/s12874-016-0132-4
17. Yurek, L. A., Vasey, J., Sullivan Havens, D.: The use of self-generated identification codes in longitudinal research. Evaluation review **32**(5), 435-452 (2008).
18. Damrosch, S. P.: Ensuring anonymity by use of subject-generated identification codes. Research in nursing & health **9**(1), 61-63 (1986). DOI: https://doi.org/10.1002/nur.4770090110
19. DiIorio, C., Soet, J. E., Van Marter, D., Woodring, T. M., Dudley, W. N.: An evaluation of a self-generated identification code. Research in nursing & health **23**(2), 167-174 (2000).
20. Grube, J. W., Morgan, M., Kearney, K. A.: Using self-generated identification codes to match questionnaires in panel studies of adolescent substance use. Addictive behaviors **14**(2), 159-171 (1989). DOI: https://doi.org/10.1016/0306-4603(89)90044-0
21. Kearney, K. A., Hopkins, R. H., Mauss, A. L., Weisheit, R. A.: Self-generated identification codes for anonymous collection of longitudinal questionnaire data. Public Opinion Quarterly **48**(1B), 370-378 (1984). DOI: https://doi.org/10.1093/poq/48.1B.370
22. Vacek, J., Vonkova, H., Gabrhelík, R.: A successful strategy for linking anonymous data from students' and parents' questionnaires using self-generated identification codes. Prevention Science **18**(4), 450-458 (2017). DOI:https://doi.org/10.1007/s11121-017-0772-6
23. Lippe, M., Johnson, B., Carter, P.: Protecting student anonymity in research using a subject-generated identification code. Journal of Professional Nursing **35**(2), 120-123(2019).
24. Galanti, M. R., Siliquini, R., Cuomo, L., Melero, J. C., Panella, M., Faggiano, F.: Testing anonymous link procedures for follow-up of adolescents in a school-based trial: the EU-DAP pilot study. Preventive medicine **44**(2), 174-177 (2007).

25. McGloin, J., Holcomb, S., Main, D. S.: Matching anonymous pre-posttests using subject-generated information. Evaluation Review **20**(6), 724-736 (1996).
26. Thoben, W., Appelrath, H. J., Sauer, S.: Record linkage of anonymous data by control numbers. In: From Data to Knowledge, pp. 412-419. Springer, Berlin, Heidelberg (1996).
27. Friedman, C., Sideli, R.: Tolerating spelling errors during patient validation. Computers and Biomedical Research **25**(5), (1992). DOI: https://doi.org/10.1016/0010-4809(92)90005-U
28. Mortimer, J. Y., Salathiel, J. A.: 'Soundex' codes of surnames provide confidentiality and accuracy in a national HIV database. Communicable disease report. CDR review **5**(12), R183-6 (1995).
29. Rogers, H. J., Willett, P.: Searching for historical word forms in text databases using spelling-correction methods: Reverse error and phonetic coding methods. Journal of Documentation **47**(4), 333-353 (1991). DOI:https://doi.org/10.1108/eb026883
30. Holmes, D., McCabe, M. C.: Improving precision and recall for soundex retrieval. In: Proc. Int. Conf. Information Technology: Coding and Computing, pp. 22-26. IEEE (2002).
31. Karakasidis, A., Verykios, V. S., Christen, P.: Fake injection strategies for private phonetic matching. In: Data Privacy Management and Autonomous Spontaneus Security, pp. 9-24. Springer, Berlin, Heidelberg (2011). DOI:https://doi.org/10.1007/978-3-642-28879-1_2
32. Camps, R., Daudé, J.: Improving the efficacy of approximate searching by personal-name. In: Natural language processing and information systems. Bonn, Germany (2003).
33. Johnson, S. B., Whitney, G., McAuliffe, M., Wang, H., et al.: Using global unique identifiers to link autism collections. J. Am. Med. Inform. Assoc. **17**(6), 689-695. (2010).
34. Bouzelat, H., Quantin, C., Dusserre, L.: Extraction and anonymity protocol of medical file. In: Proceedings of the AMIA Annual Fall Symposium, pp. 323-327. AMIA (1996).
35. Quantin, C., Binquet, C., Allaert, F. A., Cornet, B., Pattisina, R., Leteuff, G., Ferdynus, C., Gouyon, J. B.: Decision analysis for the assessment of a record linkage procedure. Methods of Information in Medicine **44**(1), 72-79 (2005).
36. Benhamiche, A. M., Faivre, J.: Automatic Record Hash Coding and Linkage for Epidemiological. Meth Inform Med **37**, 271-278 (1998).
37. Durham, E. A., Kantarcioglu, M., Xue, Y., Toth, C., Kuzu, M., Malin, B.: Composite bloom filters for secure record linkage. IEEE T Knowl Data En **26**(12), 2956-2968 (2013)
38. Kroll, M., Steinmetzer, S.: Automated cryptanalysis of bloom filter encryptions of health records. German Record Linkage Center, Working Papers, No. WP-GRLC-2014-05 (2014).
39. Randall, S. M., Ferrante, A. M., Boyd, J. H., Bauer, J. K., Semmens, J. B.: Privacy-preserving record linkage on large real world datasets. J. Biomed. Inform **50**, 205-212 (2014).
40. Niedermeyer, F., Steinmetzer, S., Kroll, M., Schnell, R.: Cryptanalysis of basic bloom filters used for privacy preserving record linkage. German Record Linkage Center, Working Paper Series, No. WP-GRLC-2014-04 (2014).
41. Cichelli, R. J.: Minimal perfect hash functions made simple. Communications of the ACM **23**(1), 17-19 (1980). DOI:https://doi.org/10.1145/358808.358813
42. Sager, T. J.: A polynomial time generator for minimal perfect hash functions. Communications of the ACM **28**(5), 523-532 (1985). DOI:https://doi.org/10.1145/3532.3538
43. Caine, K.: Local standards for sample size at CHI. In: Proceedings of the 2016 CHI conference on human factors in computing systems, pp. 981-992. ACM (2016).
44. Ioannidis, J. P., Baas, J., Klavans, R., Boyack, K. W.: A standardized citation metrics author database annotated for scientific field. PLoS biology **17**(8), e3000384 (2019).
45. Miller, G. A.: The magical number seven, plus or minus two: Some limits on our capacity for processing information. Psychological review **63**(2), 81-97 (1956).
46. Sweeney, L.: k-anonymity: A model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems **10**(05), 557-570 (2002).