A practical approach to achieve private medical record linkage in light of public resources

Mehmet Kuzu,[1] Murat Kantarcioglu,[1] Elizabeth Ashley Durham,[2] Csaba Toth,[2] Bradley Malin[2,3]

► Additional appendices are published online only. To view these files please visit the journal online (http://dx.doi.org/10.1136/amiajnl-2012-000917).

[1]Department of Computer Science, University of Texas at Dallas, Dallas, Texas, USA
[2]Department of Biomedical Informatics, Vanderbilt University, Nashville, Tennessee, USA
[3]Department of Electrical Engineering and Computer Science, Vanderbilt University, Nashville, Tennessee, USA

Correspondence to
Dr Bradley Malin, 2525 West End Avenue, Suite 600, Vanderbilt University, Nashville, TN 37203, USA;
b.malin@vanderbilt.edu

Received 28 February 2012
Accepted 12 July 2012
Published Online First 30 July 2012

ABSTRACT

Objective Integration of patients' records across resources enhances analytics. To address privacy concerns, emerging strategies such as Bloom filter encodings (BFEs), enable integration while obscuring identifiers. However, recent investigations demonstrate BFEs are, in theory, vulnerable to cryptanalysis when encoded identifiers are randomly selected from a public resource. This study investigates the extent to which cryptanalysis conditions hold for (1) real patient records and (2) a countermeasure that obscures the frequencies of the identifying values in encoded datasets.

Design First, to investigate the strength of cryptanalysis for real patient records, we build BFEs from identifiers in an electronic medical record system and apply cryptanalysis using identifiers in a publicly available voter registry. Second, to investigate the countermeasure under ideal cryptanalysis conditions, we compose BFEs from the identifiers that are randomly selected from a public voter registry.

Measurement We utilize precision (ie, rate of correct re-identified encodings) and computation efficiency (ie, time to complete cryptanalysis) to assess the performance of cryptanalysis in BFEs before and after application of the countermeasure.

Results Cryptanalysis can achieve high precision when the encoded identifiers are composed of a random sample of a public resource (ie, a voter registry). However, we also find that the attack is less efficient and may not be practical for more realistic scenarios. By contrast, the proposed countermeasure made cryptanalysis impractical in terms of precision and efficiency.

Conclusions Performance of cryptanalysis against BFEs based on patient data is significantly lower than theoretical estimates. The proposed countermeasure makes BFEs resistant to known practical attacks.

INTRODUCTION

The number of scenarios where patient information is shared beyond the point at which it was collected is growing.[1 2] The motivation behind data sharing varies, with certain organizations hoping to enable transparency and facilitate scientific discovery,[3 4] while others aim to further public health through biosurveillance and pharmacovigilance programs.[5 6] These efforts are often hindered by the fact that patients' records are scattered across organizations,[7] so that, as programs expand in scale and variety, it is crucial to integrate disparate resources.[8–12] Failure to do so will lead to horizontal and vertical partitioning problems.[13] In the former, records remain duplicated across organizations, biasing statistics proportionally to the extent to which a record is replicated in the joined resource.[14 15] In the latter, records remain fragmented and an association between a clinical phenomenon at one organization (eg, distribution of a new pharmaceutical) may be disconnected from an event elsewhere (eg, adverse episode).[16 17]

For over half a century, record linkage procedures have been developed to determine when records correspond to the same individual.[18–20] Many procedures have been adapted to data derived from the health domain,[21–24] but organizations are often prohibited from sharing the features commonly utilized in record linkage (eg, personal name and residential address). Instead, organizations are encouraged to share records that are devoid of patient identifiers (ie, de-identified data).[25]

Recently, private record linkage (PRL) protocols, which obscure identifiers, have been proposed. These protocols have diverse computational costs, security guarantees, and linkage accuracy.[26] One of the latest approaches, which seeks to strike a balance between these criteria, is based on Bloom filter encodings (BFEs).[27–30] However, evidence suggests BFEs are vulnerable to cryptanalysis[31] when the encoded records are a random sample of a known resource (eg, the encoded patient names correspond to a random selection of records from the US Census Bureau's name dictionary[32]). In such a case, a constraint satisfaction mechanism can be applied as a cryptanalysis technique to crack the filters and, thereby, reconstruct encoded identifiers.

While theory suggests BFEs are vulnerable, the distribution of identifiers in such encodings in practice is unlikely to constitute a random sample of a readily available resource. For instance, the majority of patients who visit a hospital reside in the surrounding counties, but, while public resources provide identifiers on such populations (eg, county-level voter registration records[33]), they are not necessarily an exact representation of the population. Thus, in this paper, we investigate how public resources influence the extent to which real patient identifiers in BFEs can be inferred. Specifically, we use a cohort of patient records from the Vanderbilt University Medical Center (VUMC) and investigate the extent to which public voter registration records could be exploited as a representative set of identifiers. We demonstrate that the attack remains feasible in practice, but that it results in significantly lower precision in the set of predictions and longer computation time in the cryptanalysis than suggested in theory. We further show that an extension to BFEs, where multiple patient attributes (eg, personal name and

Downloaded from https://academic.oup.com/jamia/article/20/2/285/896992 by ANKARA UNIVERSITESI user on 11 December 2021

residential address) are mixed into a single encoding[34 35] may sufficiently strengthen BFEs against cryptanalysis. Specifically, this extension transforms the frequency distribution of the encodings closer to a uniform distribution, which enables the BFEs to withstand state-of-the-art attacks based on frequency analysis of the encodings.

BACKGROUND

Record linkage and privacy

Record linkage is a classification problem that predicts if a record pair belongs to a class of matches (ie, same patient) or non-matches.[i] Various record linkage frameworks have been proposed,[34–41] ranging from heuristic rules[22 42] to probabilistic models,[43–45] but the process is traditionally accomplished via identifying attributes, such as personal name or date of birth.

To respect confidentiality, a growing collection of PRL protocols have been suggested.[46–52] PRL enables similarity tests mainly through two types of models: (1) secure multi-party computation (SMC) and (2) data transformations. SMC utilizes rigorous cryptographic protocols with strong security guarantees to execute certain functions (eg, computation of Levenshtein distance), but is computationally intensive and does not scale to large data sources.[53] By contrast, transformation methods selectively reveal information to make the process more efficient. The simplest transformation maps each value to a corresponding encoding,[54–57] but this approach neglects the fact that records often contain typographical (eg, 'john' vs 'jon') or semantic (eg, maiden vs married name) errors.[8] Transformation methods are evolving and a subclass of such protocols now enable approximate matching between records to tolerate typographical errors in the records (eg, the SparseMap algorithm transforms similar strings to regions which are close in a Euclidean space to enable approximate matching[58]).

Bloom filter encodings

BFEs are one of the more promising PRL strategies in terms of linkage accuracy and computational complexity.[26] Fundamentally, a Bloom filter is a bit array of length m, that is affiliated with k hash functions.[ii 59] All bits in the array are initially set to 0 and each function maps its input to an integer z, where $1 \leq z \leq m$, with a uniform probability. The bit locations in the array corresponding to the output integers of the function are set to 1.

In the BFE model, input that corresponds to a string is decomposed into its set of substrings of length n, or n-grams. Each n-gram is subject to all k functions and the BFE of an identifier corresponds to the bitwise OR of the n-gram encodings. Figure 1 depicts an example where the 2-grams of 'sam' are encoded with one hash function.

In practice, disparate data managers Alice and Bob map values from each identifying attribute from their datasets A and B to BFEs. The encodings are sent to a third-party Charlie who compares them to detect matches.[28 30] Figure 1 briefly summarizes this setting, but, due to space constraints, we refer the reader elsewhere[30] for an in-depth walkthrough of the process.

Leaked information leads to disclosure of sensitive data

Unlike PRL protocols based on SMC, those based on transformation may leave identifiers vulnerable to attack because they leak information. For instance, it was shown that frequency distributions can be leveraged to re-identify pseudonymized[iii] items in a database of transactions.[61] Additionally, it was shown that when the leaked information (eg, the relation between the original records and their transformed versions) is non-random, the attack may be formulated as a constraint satisfaction problem (CSP). A CSP is characterized by a set of variables and a set of constraints.[62] Each variable has a domain of possible values and each constraint restricts the possible assignments between variables and values. The goal is to assign all variables to values, such that the constraints are satisfied. CSPs have been successfully applied to break various encoded systems, such as simple substitution ciphers.[63]

Most relevant to our study, a CSP attack was developed for BFEs.[31] In this setting, the third party (ie, Charlie) assumes the role of the adversary, such that he attempts to compromise filters using the frequency distribution for identifying variables as derived from an available resource (eg, a public voter database). The variables of the CSP, as well as their domains, are identified through a frequency analysis on the identifiers in the available resource and the BFEs. For instance, suppose 'john' is an identifier in a public resource and its expected frequency interval is [0.1, 0.2]. In this setting, suppose BFE('john') and BFE('sam') are the only BFEs in the encoded dataset with frequencies that lie in the expected frequency interval of 'john.' As a result, 'john' is a variable of the CSP and the domain of 'john' consists of the encodings BFE('john') and BFE ('sam').

The constraints of the CSP are derived from the properties of the BFE generation mechanism (eg, the hash functions set the same bits for the same n-grams). We refer the reader to the section on 'A basic cryptanalysis protocol for Bloom filter encodings' for further details.

METHODS

Materials

We constructed datasets from three resources to gauge how BFEs are susceptible to attack.

Davidson County, Tennessee voter registration dataset

The Davidson County, Tennessee voter registration (DCVR) dataset is a public resource, made available for a moderate fee from the Davidson County electoral commission. The VUMC is located in this county and the DCVR is expected to reflect the type of resource that could be leveraged to attack data shared by the VUMC. It consists of the records of 283 173 individuals.

Vanderbilt electronic medical record dataset

The Vanderbilt electronic medical record (VEMR) dataset is a private resource that consists of patient identifiers in the VUMC electronic medical record system. We selected a subset of 239 747 records which match the demographics of the DCVR (eg, they have the same ZIP codes) and have no missing values. The details of the selection process are given in supplementary online appendix A.

For these resources, we studied two sets of identifiers: (1) the 100 most frequent residential street names and (2) the 37 five-digit ZIP codes available in the DCVR. The frequency distributions for these attributes are depicted in figure 2.

[i] In certain instances, record linkage incorporates a third class of record pairs held over for administrative review by humans.

[ii] Details for the selection of k hash functions are provided in supplementary online appendix E.

[iii] Pseudonymization is a process by which the identifying fields within a data record are replaced by artificial identifiers in a consistent manner (eg, 'sam' may be replaced with '150asd01').[60]

Downloaded from https://academic.oup.com/jamia/article/20/2/285/896992 by ANKARA UNIVERSITESI user on 11 December 2021

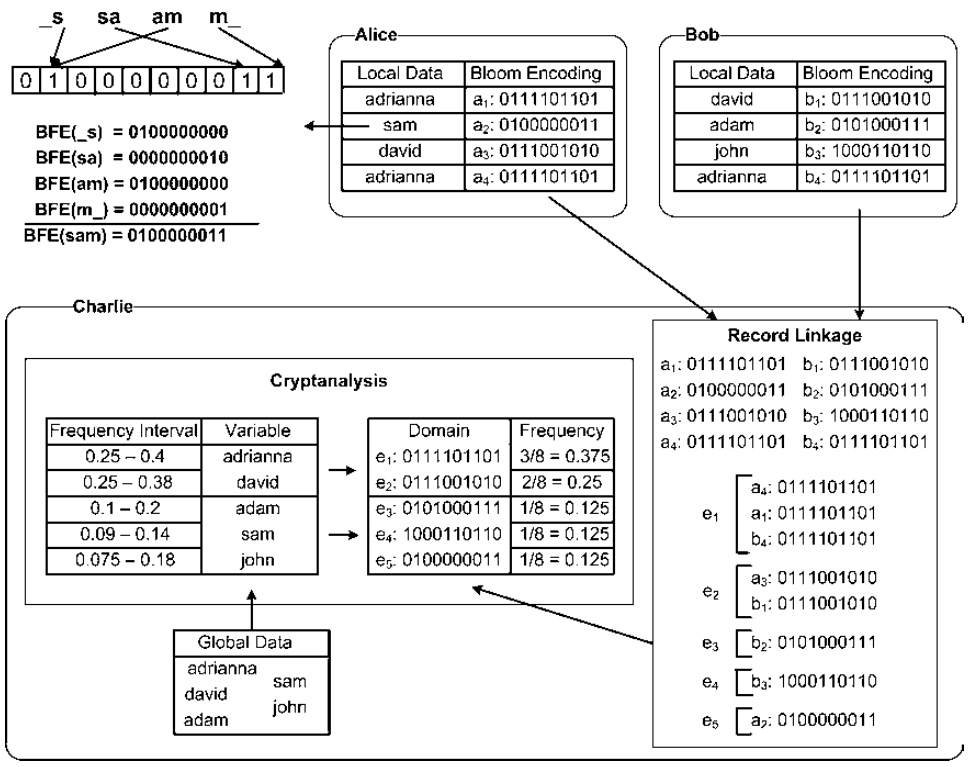


Figure 1 An example of forenames encoded into 10-bit Bloom filter encodings (BFEs) using one hash function. The names have been partitioned into 2-grams after flanking with spacers (ie, '_') to account for the ends of the strings. After encoding, data owners Alice and Bob transfer the BFEs to a third-party Charlie for linkage. This third party identifies matching records using a set-based similarity measure (eg, Dice coefficient) on the BFEs of the records. In this setting, Charlie executes the constraint satisfaction problem (CSP) attack based on a frequency analysis.

North Carolina voter registration dataset

The North Carolina voter registration (NCVR) dataset is a free public resource of voter registration records from the state of North Carolina.[64] We selected a subset of 1 767 510 records for which there were no missing values. We study this resource to compare our protection solution to the worst case scenario (eg, encoded identifiers are randomly selected from the available public resource) presented in earlier research.[31] For this resource, we studied four sets of identifiers: (1) 20 264 distinct forenames, (2) 30 217 distinct surnames, (3) 129 distinct ZIP codes, and (4) 131 distinct cities.

A basic cryptanalysis protocol for BFEs

We utilized the adversarial setting from the original BFE cryptanalysis study.[31] In this setting, the CSP variables and their domains were constructed under the assumption that the encoded identifiers were randomly selected from an all-encompassing global dataset. Charlie attempts to crack the encodings submitted by Alice and Bob without their assistance. To do so, he invokes a public resource as a global dataset and attempts to infer the identifiers in the BFE datasets through a CSP.

The CSP variables (ie, identifiers) and their domains are obtained from the global and encoded dataset, respectively. Normalized frequency[iv] intervals are extracted for each identifier in the global dataset through a Monte Carlo sampling technique. Specifically, multiple artificial datasets with the size of the encoded dataset are constructed by a random selection of records from the public resource. Next, frequency bounds are constructed for each identifier in the public resource using its corresponding frequencies in the artificially constructed datasets. If the lower frequency bound of an identifier is positive, it is included in the model because the corresponding identifier is expected to be in the encoded dataset. Then, each BFE is associated with the domain for each variable in which its frequency is bounded. Basically, identifiers in the public resource are used as CSP variables and the domain values of these variables consist of a subset of BFEs. Figure 1 provides an illustration of this process. Notice that the BFEs b_1 and b_2 are placed in the domain for 'adrianna' because their frequencies reside in the corresponding interval.

The CSP constraints are obtained from the properties of the n-gram encoding function, which we refer to as f. Recognize that when the function is applied to a particular n-gram, only certain bits are flipped from 0 to 1. Thus, if a BFE contains 0's in certain bit locations, Charlie concludes f does not set those locations for any of the n-grams in the corresponding identifier. This enables Charlie to simulate f on each n-gram through a belief function bf that works as follows.

To formalize the belief function, let X be a variable and v be a BFE. Initially, for an n-gram q, Charlie believes $f(q)$ may set any bit location from 0 to 1. Once a variable-to-encoding assignment $X = v$ is added to the CSP, the new belief about the encoding of X is obtained by applying the bitwise AND to $bf(q)$ and v. If v contains a 0 in location λ, then $bf(q)[\lambda] = 0$ for an updated state of the CSP. Returning to figure 1, initially bf('ad')=1111111111 because all bits are equally likely to be set when f is applied to 'ad.' Now, suppose {adrianna=0111101101} is added to the CSP. At this point, Charlie learns that f does not set the first, sixth and ninth bit locations when applied on the 2-grams of 'adrianna,' so bf('ad') is updated to 0111101101.

More generally, the belief function and properties of the BFEs are exploited to define two types of constraints. We informally describe these constraints because they influence the cryptanalysis and will be addressed in our evaluation, but we direct the reader elsewhere[31] for more rigorous and detailed definitions. An illustrative application of the constraints is provided in supplementary online appendix B.

[iv]The normalized frequency of an item is defined as its proportion among all the items in the dataset.

Downloaded from https://academic.oup.com/jamia/article/20/2/285/896992 by ANKARA UNIVERSITESI user on 11 December 2021

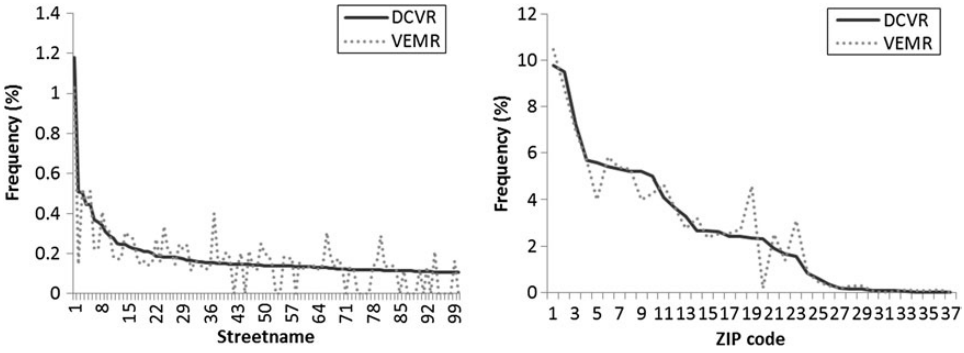


Figure 2 Frequency distribution of identifier values in the Davidson County, Tennessee voter registration (DCVR) and Vanderbilt electronic medical record (VEMR) datasets: (left) street names and (right) ZIP codes.

Belief validity constraint

The encoding of an identifier is obtained by combining the encoding of its *n*-grams with a bitwise OR operation. The belief function simulates the *n*-gram encoding function, so the encoding of any string will be formed by combining the belief about its *n*-grams in any consistent CSP state.

Minimum location constraint

The minimum number of bit locations, *w*, that will be set by an *n*-gram encoding function can be determined probabilistically. This is based on the number of hash functions applied for the encoding and the length of the Bloom filter. As such, Charlie can enforce a constraint which asserts that belief on any *n*-gram should contain at least *w* bit locations that are set to 1.

Once the CSP is formulated, a solver algorithm is applied to associate BFEs with corresponding identifiers. An example of how the solver works is provided in supplementary online appendix C.

Frequency extension to handle non-random sampling

Although public resources may have similar frequencies to the encoded patient identifiers, the latter are not necessarily a random sample. This may be due to various factors, such as the fact that voters in a certain region are not necessarily patients of a hospital in the same vicinity. And, if an encoding fails to be associated with the correct corresponding identifier in the CSP during initialization, it can lead to an incorrect belief function and an unsolvable CSP.

To prevent the CSP from failing, we devised a strategy to systematically expand the frequency interval. Imagine (l, h) is the low-to-high frequency interval for a patient identifier and is based on an initial prediction through the random sample assumption. To ensure a properly initialized CSP, we apply expansion factors e_l and e_h, such that the bounds become [e_l * l, e_h * h]. These expansion factors are basically multipliers for the frequency bounds of the identifiers which were derived under the random sample assumption. For instance, suppose the predicted interval for 'john' is [0.1, 0.2], e_l = 0.5, and e_h =2. Then, the expanded interval is [0.05, 0.4]. To obtain such intervals, we iteratively increment e_l and decrement e_h until the CSP solver achieves a solution.

Statistical approximation to speed up cryptanalysis

Frequency bounds were initially constructed via a Monte Carlo sampling technique.[31] However, this requires the selection of a large quantity of samples from the global source, each equal in size to the number of encoded patient records. This is a slow and computationally intensive process and suggests the attack may

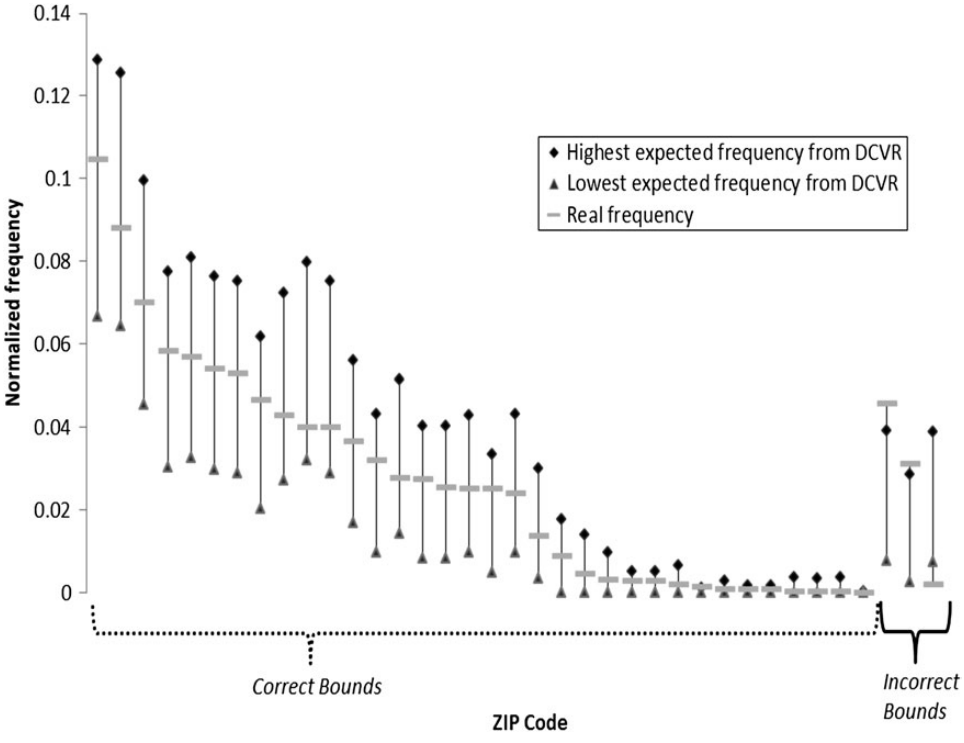


Figure 3 Frequency bounds for the ZIP codes in the Vanderbilt electronic medical record (VEMR) dataset constructed through a statistical analysis of the Davidson County, Tennessee voter registration (DCVR) dataset under the random sample assumption.

Downloaded from https://academic.oup.com/jamia/article/20/2/285/896992 by ANKARA UNIVERSITESI user on 11 December 2021

not be practical. However, we hypothesized that the CSP could be formulated in a significantly faster manner by utilizing the sampling distribution of proportion statistics to form the bounds analytically. Due to space limitations, the details of this process are deferred to supplementary online appendix D.

Mixing multiple identifying attributes into a single Bloom filter

If the frequency distribution of encoded patient records could be flattened, the CSP attack could be thwarted, even if the theoretical worst case scenario holds (ie, encoded identifiers are randomly selected from a public resource). The frequency analysis is necessary for both variable selection and domain refinement. If the lower bound of the expected proportion of a candidate variable is equal to zero, the solver cannot select it as a variable. This is because the encoding of the corresponding variable may not exist in the encoded set, which is a necessary condition for a feasible attack.

We hypothesized that combining multiple patient attributes into a single BFE[34 35] could provide such protection. To test this hypothesis, we concatenated the values of the personal attributes from the corresponding records. For illustration, a record with forename 'john' and surname 'smith' would be 'john smith' in the concatenated filter.

Performance measures

To evaluate the performance of the cryptanalysis, we adapted a precision measure. We define precision p as the ratio of correctly identified encodings to all available ⟨BFE, identifier⟩ pairs produced by the CSP solver.

To measure the computational feasibility of the cryptanalysis, we use time t required for the CSP solver to provide a solution to the problem.

Experimental setting

We assessed cryptanalysis in two settings. For both cases, BFEs were obtained by hashing 2-grams of the identifiers into a 500-bit Bloom filter. This BFE setting was suggested in an earlier study.[30]

In the first setting, we composed a real attack scenario, such that the VEMR and DCVR were applied as the sample BFE and global public datasets, respectively. We assessed the performance using ZIP codes and street names. In this setup, frequency bounds were based on an initial estimate derived from the sampling distribution of the proportion statistics (details are available in supplementary online appendix D). The bounds were further expanded with constant factors. This is because the initial bounds may not be correct for some identifiers.[v]

The second setting built on the original study[31] to assess how composite filters can mitigate vulnerability when the random sampling assumption holds true and to allow for independent validation of our experiments. In this setting, NCVR served as the global dataset and the BFE dataset was derived by randomly selecting 20 000 and 50 000 records. While the original study assessed one attribute (forename), we investigated the performance with the four attributes mentioned earlier and composite BFEs of the form (forename + surname) and (forename + surname + city + ZIP code). For this experiment, the frequency bounds were formed by an analytical analysis of proportion statistics (see supplementary online appendix D). We analyzed the domain size of the identifiers resulting from the bounds and the success of the CSP solver on the 50 most frequent identifying values.

All experiments were conducted on a Linux operating system with a two quad-core 2.5 GHz Intel Xeon processor and 30 GB memory.

RESULTS

A real attack on patient identifiers

By attacking the VEMR through the DCVR, we investigated the cracking performance and the computational efficiency of the cryptanalysis in the real world setting where the encoded identifiers are not selected randomly from a public resource.

The VEMR and DCVR datasets covered the same set of 37 ZIP codes. Yet, as figure 3 illustrates, the initial estimates on the normalized frequency bounds were incorrect for three (or 8.1%) of the ZIP codes. Execution of the CSP attack on such a system failed to yield a correct solution due to the propagation of belief errors. However, we were able to successively expand the frequency intervals to achieve a satisfiable CSP that correctly assigned 95% of all ZIP codes to their corresponding BFEs. This was accomplished when the lower and upper bound expansion factors were set to e_l=0.2 and e_h=1.3, respectively.

Despite the attack's success, it should be recognized that the expansion process led to significant growth in the domains of the ZIP codes, as shown in figure 4. Once the expansion factors were correctly defined, the CSP solver required approximately 1 day to complete the task. However, an adversary needs to search for an appropriate e_l and e_h. In the experiments, it was found that each trial in the search resulted in the CSP solver halting without a solution, but required approximately 2 days per run.

Regarding street names, we observed that the frequencies of many names in the two resources were similar (as shown in figure 2), but that certain names in the DCVR were not in the top 100 VEMR names. A CSP attack based on these data is guaranteed to fail because, to achieve a satisfiable CSP, the domains for all the names in the global dataset must be correctly defined. In other words, the domain of each name should contain its correct encoding. For instance, the domain of name 'john' should contain BFE('john'). However, this does not imply the system is sufficiently protected. Consider, an adversary might conduct a smaller scale attack on a smaller set of names. To address this case, we composed a CSP with the most frequent names in the DCVR that were in the list of most frequent names in the VEMR. This CSP consisted of a set of 42 names from which we were able to construct a satisfiable CSP formulation once the expansion factors were set to e_l=0.6 and e_h=1.2, respectively. In this setting, the solver was unable to discover a solution after almost a week of continuous runtime. Therefore, we reduced the number of names in the CSP to the point at which the solver completed the task in a practical time period. This occurred at the 20 most frequent names, where a mapping was returned in a few seconds. In this solution, four of the 20 mappings (or 20%) were correct.

These results illustrate that attacks on BFEs based on real patient records are feasible, although challenging.

Mitigation of the worst case attack

The results of the previous section indicate the importance of an efficient frequency analysis on the cryptanalysis. Although it is difficult to assert well-defined assumptions on the frequency distribution of the encoded dataset, encodings should withstand worst case scenarios (eg, when encoded identifiers are randomly

[v]The bounds were constructed under the assumption that encoded records were randomly selected from the public source, which was not necessarily true for real datasets.

Downloaded from https://academic.oup.com/jamia/article/20/2/285/896992 by ANKARA UNIVERSITESI user on 11 December 2021

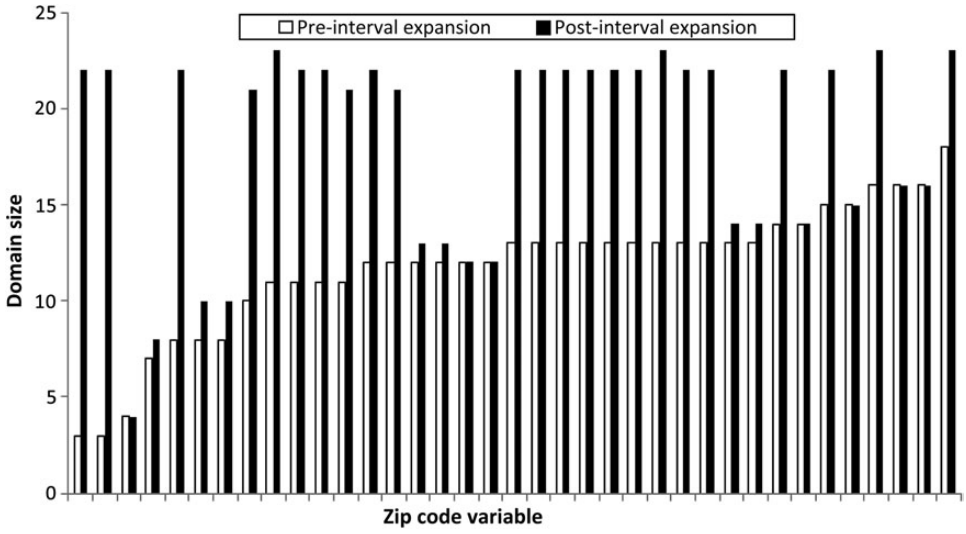


Figure 4 Expansion of lower and upper frequency bounds with constants $e_l = 0.2$ and $e_h = 1.3$ enables construction of correct bounds for all ZIP codes in the Vanderbilt electronic medical record (VEMR) dataset with the cost of considerable increase in the domains of the constraint satisfaction problem (CSP) variables.

selected from a public resource). In the theoretical study,[31] such a scenario was constructed by drawing encoded identifiers as a random sample of the global dataset (ie, the NCVR dataset).

Our investigations reveal that the CSP for this setting has variables with relatively small domains due to an efficient frequency analysis. Figure 5 depicts the domain sizes for the four NCVR identifying attributes. For all attributes, the domains of the first few variables were sufficiently small to help prune large parts of the solver's search space, leading to efficient assignments of the variables. The results given in table 1 demonstrate the efficiency of the CSP solver in this setting, where it can be seen that the precision of the attack ranged from 0.94 to 1.0 and the runtime ranged from several seconds to an hour.

To protect BFEs, we found that the frequency distribution can be pushed toward a uniform distribution by combining multiple attributes of a patient into a common BFE. To illustrate the effect, we executed the CSP on random samples of 20 000 and 50 000 NCVR records for the (firstname + surname) and (firstname + surname + city + ZIP code) attributes. When the number of records was 20 000, we were unable to instantiate the CSP for an attack. This was because the lower bound of the expected proportion for all candidate variables was equal to zero.

In the case of 50 000 NCVR (firstname + surname) records, we were able to set up the CSP for 12 variables. However, as shown in figure 6, the domains for these variables were too large to mount a successful attack. The average domain size was

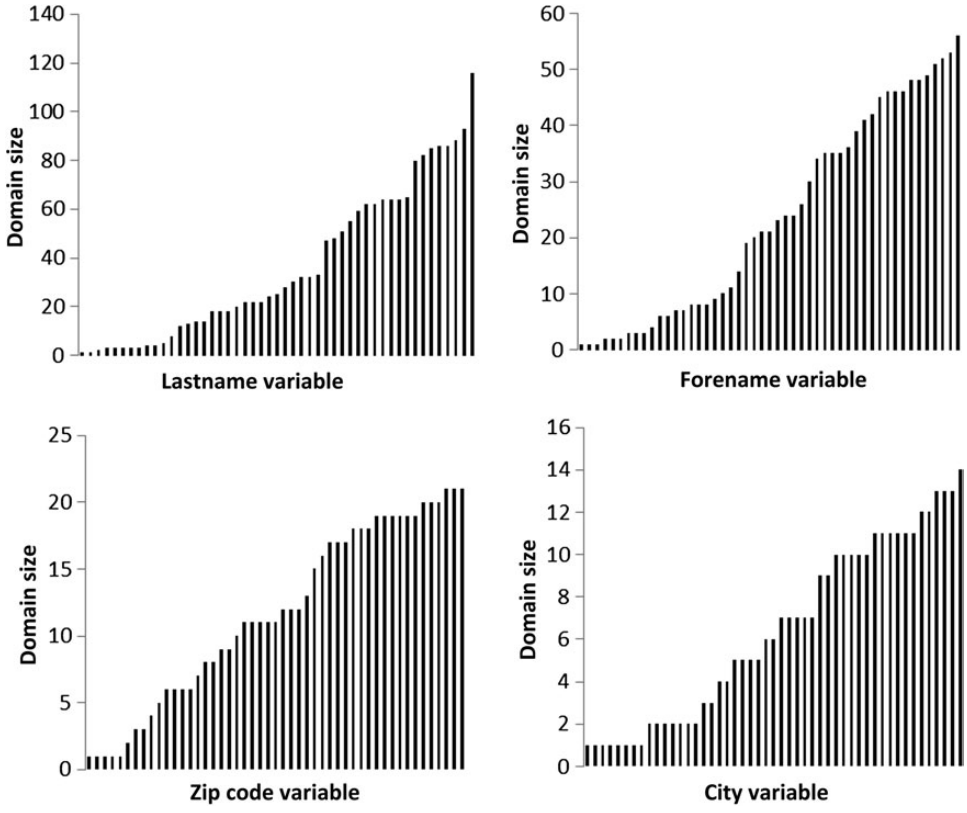


Figure 5 Frequency analysis of the North Carolina voter registration (NCVR) dataset under the random sample assumption enabled construction of constraint satisfaction problem (CSP) variables with relatively small domains.

Downloaded from https://academic.oup.com/jamia/article/20/2/285/896992 by ANKARA UNIVERSITESI user on 11 December 2021

Table 1 Performance of the constraint satisfaction problem (CSP) cryptanalysis with the North Carolina voter registration (NCVR) dataset when identifying attributes are drawn according to a random sampling assumption

	Identifying attribute			
	Forename	Surname	Five-digit ZIP code	City
Precision	1.00	0.94	0.94	1.00
Time to complete attack (seconds)	10.19	11.95	4792.21	7.2

approximately 28 000. As a result, the assignments proposed by the CSP were almost random. Only one identifier was correctly assigned to its encoding. By the time the BFE was extended to incorporate (firstname + surname + city + ZIP code), the CSP was unable to crack any identifiers.

DISCUSSION

BFEs are a potentially useful technique for enabling PRL in a practical way. We recognize that cryptographic techniques based on SMC[53 65] provide more rigorous security guarantees, but, at the moment, are not sufficiently scalable for the integration of large data sources. According to a recent survey, integration of two datasets of 1000 strings with a well-known SMC technique requires more than 2 years on a commodity server.[26] By contrast, BFEs enable such integration in only a few seconds.

However, care must be taken in the application of BFEs. Our experiments prove public resources, such as voter registration lists, can be leveraged to crack BFEs based on real patient identifiers. Although the speed and precision of state-of-the-art cryptanalysis is worse than previously posited,[31] BFEs can still be compromised using the computational resources available in a commodity server. Fortunately, we have shown that combining multiple patient identifying attributes into a common BFE offers a protection mechanism. This simple, yet effective, technique ensures the CSP model is incorrectly formulated or practically intractable to solve. We illustrated this approach precludes the attack even if the theoretical worst case assumptions (eg, encoded identifiers are randomly selected from a public resource) hold true.

Despite the merits of this work, we wish to highlight certain limitations of our study. First, while we have shown that the attack is feasible, we assumed an adversary had appropriate knowledge of the population from which the BFEs were generated. However, in practice, it may be difficult for an adversary to acquire such knowledge. Supplementary online appendix A provides further details about how such knowledge would influence the CSP attack with the VEMR dataset. Second, our cryptanalysis is based on a specific technique. Although it is the best known attack to date, it may not expose all vulnerabilities of the encodings. Third, the empirical analysis is specific to the datasets and personal attributes we studied. The frequency distributions of the attributes may differ across datasets and attributes. Finally, the proposed BFE protection mechanism was only evaluated from the perspective of its resistance against the attack with the goal of constructing a more secure BFE. It will be necessary to extend BFEs in a more principled approach by evaluation of the impact of the attack on record linkage accuracy.

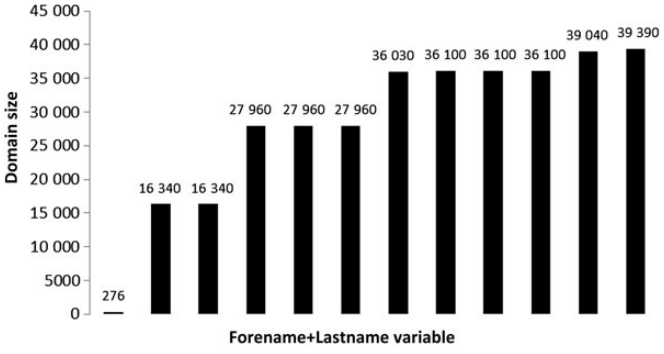


Figure 6 Flattening of the encoding frequency distribution led to significant increase in the number of possible encodings (domain) that may be assigned to a string in the global dataset (variable).

CONCLUSION

BFE is a promising approach for private medical record linkage, but this investigation confirms that BFEs may leave identifiers vulnerable to compromise via public resources. Nonetheless, our work further illustrates there are limitations to frequency-based attacks and when patient identifiers are not a proper random sample of a resource available to an attacker (eg, voter list), cryptanalysis is less likely to succeed. Finally, we showed that combining multiple patient identifying attributes into a single BFE offers a practical protection mechanism against cryptanalysis.

Acknowledgments The authors thank Dr Dario Giuse for providing the patient data used in this study.

Funding This research was funded by grants R01LM009989 from the National Library of Medicine and CCF-0424422, CNS-0845803, and CNS-0964350 from the National Science Foundation.

Competing interests None.

Ethics approval The Behavior and Social Sciences section of the Vanderbilt Institutional Review Board approved this study.

Provenance and peer review Not commissioned; externally peer reviewed.

REFERENCES

1. **National Committee on Vital and Health Statistics.** Enhanced protections for uses of health data: a stewardship framework for "secondary uses" of electronically collected and transmitted health data. *Report to the Secretary of the U.S. Department of Health and Human Services*. Hyattsville, MD: US Department of Health and Human Services, 2007.
2. **Safran C,** Bloomrosen M, Hammond WE, *et al*; Expert panel. Toward a national framework for the secondary use of health data: an American Medical Informatics Association white paper. *J Am Med Inform Assoc* 2007;**14**:1–9.
3. **National Institutes of Health.** *Final Statement on Sharing Research Data*. NOT-OD-03–032. Bethesda, MD: National Institutes of Health, 2003.
4. **National Institutes of Health.** *Policy for Sharing of Data Obtained in NIH Supported or Conducted Genome-Wide Association Studies (GWAS)*. NOT-OD-07–088. Bethesda, MD: National Institutes of Health, 2007.
5. **Behrman R,** Benner JS, Brown JS, *et al*. Developing the Sentinel System—a national resource for evidence development. *N Eng J Med* 2011;**364**:498–9.
6. **Platt R,** Wilson M, Chan KA, *et al*. The new Sentinel Network—improving the evidence of medical-product safety. *N Engl J Med* 2009;**361**:645–7.
7. **Malin B,** Sweeney L. How (not) to protect genomic data privacy in a distributed network: using trail re-identification to evaluate and design anonymity protection systems. *J Biomed Inform* 2004;**37**:179–92.
8. **Hernandez M,** Stolfo S. Real-world data is dirty: data cleansing and the merge/purge problem. *Data Min Knowl Discov* 1998;**2**:9–37.
9. **Holman CD,** Bass AJ, Rouse I, *et al*. A decade of data linkage in Western Australia: strategic design, applications and benefits of the WA data linkage system. *Aust Health Rev* 2008;**32**:766–77.
10. **Jaro MA.** Probabilistic linkage of large public health data files. *Stat Med* 1995;**14**:491–8.
11. **Kelman CW,** Bass AJ, Holman CD. Research use of linked health data—a best practice protocol. *Aust N Z J Public Health* 2002;**26**:251–5.
12. **Méray N,** Reitsma JB, Ravelli AC, *et al*. Probabilistic record linkage is a valid and transparent tool to combine databases without a patient identification number. *J Clin Epidemiol* 2007;**60**:883–91.
13. **Vaidya J,** Clifton C, Zhu Y. *Advances in Information Security: Privacy Preserving Data Mining*. New York, NY: Springer, 2006.

Downloaded from https://academic.oup.com/jamia/article/20/2/285/896992 by ANKARA UNIVERSITESI user on 11 December 2021

14. **Zingmond DS,** Ye Z, Ettner SL, *et al*. Linking hospital discharge and death records—accuracy and sources of bias. *J Clin Epidemiol* 2004;**57**:21—9.
15. **Abowd J,** Vilhuber L. The sensitivity of economic statistics to coding errors in personal identifiers. *J Econom* 2005;**23**:133—52.
16. **van Herk-Sukel MP,** van de Poll-Franse LV, Lemmens VE, *et al*. New opportunities for drug outcomes research in cancer patients: the linkage of the Eindhoven Cancer Registry and the PHARMO record linkage system. *Eur J Cancer* 2010;**46**:395—404.
17. **Wang HE,** Cook LJ, Chang CC, *et al*. Outcomes after out-of-hospital endotracheal intubation errors. *Resuscitation* 2009;**80**:50—5.
18. **Fellegi IP,** Sunter AB. A theory for record linkage. *J Am Stat Assoc* 1969;**64**:1183—210.
19. **Herzog T,** Scheuren F, Winkler W. *Data Quality and Record Linkage Techniques*. New York, NY: Springer, 2007.
20. **Newcombe HB,** Kennedy JM, Axford SJ, *et al*. Automatic linkage of vital records. *Science* 1959;**130**:954—9.
21. **Duvall SL,** Fraser AM, Rowe K, *et al*. Evaluation of record linkage between a large healthcare provider and the Utah Population Database. *J Am Med Inform Assoc* 2012;**19**:e54—9.
22. **Grannis S,** Overhage J, McDonald CJ. Analysis of identifier performance using a deterministic linkage algorithm. *Proc AMIA Symp* 2002:305—9.
23. **Newman TB,** Brown AN. Use of commercial record linkage software and vital statistics to identify patient deaths. *J Am Med Inform Assoc* 1997;**4**:233—37.
24. **Roos L,** Walld R, Wajda A, *et al*. Record linkage strategies, outpatient procedures, and administrative data. *Med Care* 1996;**34**:570—82.
25. **US Department of Health and Human Services, Office for Civil Rights.** *Standards for Privacy of Individually Identifiable Health Information, Final Rule*. Washington, DC, USA: Federal Register, 2003;45 CFR: Pt 164.
26. **Durham E,** Xue Y, Kantarcioglu M, *et al*. Quantifying the correctness, computational complexity, and security of privacy-preserving string comparators for record linkage. *Information Fusion* 2012;**13**:245—59.
27. **Bachteler T,** Schnell R, Reiher J. An empirical comparison of approaches to approximate string matching in private record linkage. *Proceedings of Statistics Canada Symposium*. Ottawa, Canada: Statistics Canada. 2010:08C—1.
28. **Durham E,** Xue Y, Kantarcioglu M, *et al*. Private medical record linkage with approximate matching. *AMIA Annu Symp Proc* 2010:182—6.
29. **Karakasidis A,** Verykios V. Secure blocking + secure matching = secure record linkage. *J Comput Sci Eng* 2011;**5**:223—35.
30. **Schnell R,** Bachteler T, Reiher J. Privacy-preserving record linkage using Bloom filters. *BMC Med Inform Decis Mak* 2009;**9**:41.
31. **Kuzu M,** Kantarcioglu M, Durham E, *et al*. A constraint satisfaction cryptanalysis of Bloom filters in private record linkage. *Proc Privacy Enhancing Technologies Symp* 2011:226—45.
32. **Population Division, US Census Bureau.** *Frequently Occurring First and Surnames from the 1990 Census*. http://www.census.gov/genealogy/names/ (accessed 25 Jan 2011).
33. **Benitez K,** Malin B. Evaluating re-identification risks with respect to the HIPAA Privacy Rule. *J Am Med Inform Assoc* 2010;**17**:169—77.
34. **Schnell R,** Bachteler T, Reiher J. *A Novel Error-Tolerant Anonymous Linking Code*. Working Paper Series No. WP-GRLC-2011-02. Nürnberg, Germany: German Record Linkage Center, November 2011.
35. **Durham E.** *A Framework Accurate, Efficient Private Record Linkage*. Ph.D. Thesis, Nashville, TN: Vanderbilt University, 2012. http://etd.library.vanderbilt.edu/available/etd-03262012-144837/unrestricted/dissertation.pdf (accessed 30 Apr 2012).
36. **Bell G,** Sethi A. Matching records in a national medical patient index. *Comm ACM* 2001;**44**:83—8.
37. **Campbell K,** Deck D, Krupski A. Record linkage software in the public domain: a comparison of Link Plus, The Link King, and a 'basic' deterministic algorithm. *Health Inform J* 2008;**14**:5—15.
38. **Christen P,** Churches T, Hegland M. Febrl—a parallel open source data linkage system. *Proc Pacific-Asia Symposium on Knowledge-Discovery and Data Mining*, Sydney, Australia. 2004:638—47.
39. **Elmagarmid A,** Ipeirotis P, Verykios V. Duplicate record detection: a survey. *EEE Trans Knowl Data Eng* 2007;**19**:1—16.
40. **Fair M.** Generalized record linkage software—Statistics Canada's record linkage software. *Austrian Journal of Statistics* 2004;**33**:37—53.
41. **Winkler W.** Matching and record linkage. In: Co X B, Binder D, Chinnappa A, *et al*, eds. *Business Survey Methods*. Wiley, 1993:355—84.
42. **Langley J,** Botha J. Use of record linkage techniques to maintain the Leicestershire Diabetes Register. *Comput Methods Programs Biomed* 1994;**41**:287—95.
43. **Duval S,** Kerber R, Thomas A. Extending the Fellegi—Sunter probabilistic record linkage method for approximate field comparators. *J Biomed Inform* 2010;**43**:24—30.
44. **Grannis S,** Overhage J, Hui S, *et al*. Analysis of a probabilistic record linkage technique without human review. *AMIA Annu Symp Proc* 2003:259—63.
45. **Winkler W.** Using the EM algorithm for weight computation in the Fellegi-Sunter model of record linkage. *Proc Section on Survey Research Methods, American Statistical Association*. Alexandria, VA: American Statistical Association, 1998:667—71.
46. **Al-Lawati A,** Lee D, McDaniel P. Blocking-aware private record linkage. *Proc International Workshop on Information Quality, Baltimore, MD*. New York, NY: ACM Press, 2005:59—68.
47. **Churches T,** Christen P. Some methods for blindfolded record linkage. *BMC Med Inform Decis Mak* 2004;**4**:9.
48. **Inan A,** Kantarcioglu M, Bertino E, *et al*. A hybrid approach to private record linkage. *Proc IEEE International Conference on Data Engineering, Cancun, Mexico*. Washington, DC: IEEE Computer Society, 2008:496—505.
49. **Inan A,** Kantarcioglu M, Ghinita G, *et al*. Private record matching using differential privacy. *Proc International Conference on Extending Data Base Technology, Lausanne, Switzerland*. New York, NY: ACM Press, 2010:123—34.
50. **Yakout M,** Atallah M, Elmagarmid A. Efficient private record linkage. *Proc IEEE International Conference on Data Engineering, Shanghai, China*. Washington, DC: IEEE Computer Society, 2009:1283—6.
51. **Agrawal R,** Evfimievski A, Srikant R. Information sharing across private databases. *Proc ACM SIGMOD, San Diego, CA*. New York, NY: ACM Press, 2009:86—97.
52. **O'Keefe CM,** Yung M, Gu L, *et al*. Privacy-preserving data linkage protocols. *Proc ACM Workshop on Privacy in the Electronic Society, Washington, DC*. New York, NY: ACM Press, 2004:94—102.
53. **Atallah M,** Kerschbaum F, Du W. Secure and private sequence comparisons. *Proc ACM Workshop on Privacy in the Electronic Society, Raleigh, NC*. New York, NY: ACM Press, 2003:39—44.
54. **Fischbacher CM,** Bhopal R, Povey C, *et al*. Record linked retrospective cohort study of 4.6 million people exploring ethnic variations in disease: myocardial infarction in South Asians. *BMC Public Health* 2007;**7**:142.
55. **Malin B,** Airoldi E. Confidentiality preserving audits of electronic medical record access. *Stud Health Technol Inform* 2007;**129**:320—4.
56. **Quantin C,** Bouzelat H, Allaert FA, *et al*. How to ensure data security of an epidemiological follow-up: quality assessment of an anonymous record linkage procedure. *Int J Med Inform* 1998;**49**:117—22.
57. **Weber SC,** Lowe H, Das A, *et al*. A simple heuristic for blindfolded record linkage. *J Am Med Inform Assoc* 2012;**19**:e157—61.
58. **Scannapieco M,** Figotin I, Bertino E, *et al*. Privacy preserving schema and data matching. *Proc ACM SIGMOD Conference on Management of Data, Beijing, China*. New York, NY: ACM Press, 2007:653—64.
59. **Bloom B.** Space/time trade-offs in hash coding with allowable errors. *Comm ACM* 1970;**13**:422—6.
60. **Pfitzmann A,** Hansen M. Anonymity, unobservability, and pseudonymity—a proposal for terminology. *Proc Designing Privacy Enhancing Technologies Workshop, Berkeley, CA*. New York, NY: Springer-Verlag, 2001:1—9.
61. **Lakshmanan L,** Ng R, Ramesh G. On disclosure risk analysis of anonymized itemsets in the presence of prior knowledge. *ACM Trans Knowl Discov Data* 2008;**2**:13.
62. **Russell S,** Norvig P. *Artificial Intelligence: a Modern Approach*. 2nd edn. Upper Saddle River, NJ: Prentice Hall, 2003.
63. **Lucks M.** A constraint satisfaction algorithm for the automated decryption of simple substitution ciphers. *Proc Annual International Cryptology Conference, Santa Barbara, CA*. New York, NY: Springer-Verlag, 1990:132—44.
64. *North Carolina Voter Registration Database*. ftp://www.app.sboe.state.nc.us/data (accessed 5 Dec 2011).
65. **Feigenbaum J,** Ishai Y, Nissim K, *et al*. Secure multiparty computation of approximations. *ACM Trans Algorithm* 2006;**2**:435—72.

Downloaded from https://academic.oup.com/jamia/article/20/2/285/896992 by ANKARA UNIVERSITESI user on 11 December 2021