Elsevier Editorial System(tm) for International Journal of Medical Informatics
Manuscript Draft

Manuscript Number: IJMI-D-17-00036R1

Title: Anticipating Anonymity in Screening Program Databases

Article Type: Research Paper

Keywords: Anonymity; screening programs; databases; k-anonymity; constraint programming

Corresponding Author: Dr. Rafael Caballero, Ph.D.

Corresponding Author's Institution: University. Complutense de Madrid

First Author: Rafael Caballero, Ph.D.

Order of Authors: Rafael Caballero, Ph.D.; Sagar Sen, Ph.D.; Jan F Nygård, Ph.D.

Abstract: In this paper, we propose a technique for increasing anonymity in screening program databases to increase the privacy for the participants in these programs. The data generated by the invitation process (screening centre, appointment date) is often made available to researchers for medical research and for evaluation and improvement of the screening program. This information, combined with other personal quasi-identifiers such as the ZIP code, gender or age, can pose a risk of disclosing the identity of the individuals participating in the program, and eventually their test results. We present two algorithms that produce a set of screening appointments that aim to increase anonymity of the resulting dataset. The first one, based on the constraint programming paradigm, defines the optimal appointments, while the second one is a suboptimal heuristic algorithm that can be used with real size datasets. The level of anonymity is measured using the new concept of generalized k-anonymity, which allows us to show the utility of the proposal by means of experiments, both using random data and data based on screening invitations from the Norwegian Cancer Registry

Anticipating Anonymity in Screening Program Databases

Rafael Caballero[a,1,**], Sagar Sen[b,2,*], Jan F Nygård[c,3,*]

[a] *University Complutense of Madrid*
[b] *Certus SFI*
[c] *Cancer Registry of Norway (Kreftregisteret)*

Abstract

In this paper, we propose a technique for improving anonymity in screening program databases to increase the privacy for the participants in these programs. The data generated by the invitation process (screening centre, appointment date) is often made available to researchers for medical research and for evaluation and improvement of the screening program. This information, combined with other personal quasi-identifiers such as the ZIP code, gender or age, can pose a risk of disclosing the identity of the individuals participating in the program, and eventually their test results. We present two algorithms that produce a set of screening appointments that aim to increase anonymity of the resulting dataset. The first one, based on the constraint programming paradigm, defines the optimal appointments, while the second one is a suboptimal heuristic algorithm that can be used with real size datasets. The level of anonymity is measured using the new concept of generalized k-anonymity, which allows us to show the utility of the proposal by means of experiments, both using random data and data based on screening invitations from the Norwegian Cancer Registry.

Keywords: Anonymity, screening programs, constraint programming

[*] Corresponding author

[**] Principal corresponding author

[1] Rafael Caballero partially supported by the Spanish MINECO project CAVI-ART (TIN2013-44742-C4-3-R), Madrid regional project N-GREENS Software-CM (S2013/ICE-2731), UCM grant GR3/14-910502, and the NILS grant

[2] Funded by the Norwegian Research Council

[3] Funded by The Cancer Registry of Norway

1. Introduction

Privacy may be defined as *autonomy in society* [11]. Every human being has a natural desire for autonomy or independence of control of others. How to balance autonomy versus organized professional guardians of those rights is the key problem of a liberal political philosophy. Addressing this problem in the private sphere of a person's health, and in particular preventive health, sets the general context for this article.

The privacy of a person's medical history used to be locked in handwritten journals until the advent of electronic health information systems. This has made it much easier to access the medical history of the patient for the health professionals, thus greatly improving the medical advice or treatment for the patient. However, the data about peoples' health can now be easily replicated, stored, and transferred anywhere in the world making privacy hard to achieve [7]. Sharing data while maintaining people's anonymity has been a major concern in health databases where the *data already exists*.

A rudimentary approach to anonymizing a health database is to remove identifying variables such as the name of a person, social security number, and address. Despite this simple form of de-identification, the anonymity of a person in a database is often at stake due to existence of variables called *quasi-identifiers*. A specific combination of a person's birth date and the type of medical exam they underwent may reveal a very small number of people who thus could be identifiable.

Typically, medical databases emerge out of unpredictable behaviour of people going to their doctors. This results in a medical database (electronic health information system) where the primary use is to document the different diagnostics procedures and subsequent treatment. In these databases the patients need to be identifiable, and the security measures mainly consists of encryption, authentication, authorization, and logging of who have accessed the data.

Screening programs [23], present a special scenario where people are invited to come and take a screening test without having any medical symptoms. In addition to using the data from different population registries, these programs consider *appointment data* such as the date of the medical appointment, or the particular clinic where the previous screening took place. This appointment data usually become part of the quasi-identifiers, which constitutes an additional anonymity risk factor. The emerging data from these screening programs are often used to improve the program, usually by epidemiological studies, and measures to ensure anonymity should be taken into account when designing the logistics of the screening program.

For example, screening in Norway are done for three different cancer types; breast cancer, colorectal cancer and cervical cancers. These screening programs have different algorithms that initiates a logistic process by inviting people in the appropriate age groups to a screening centre or their general practitioner (GP). This selection process also involves excluding persons which previously have been diagnoses with cancer, or recently have been screened. These algorithms also specify how many times the screening test should be repeated and the interval between tests. In Norway, these logistical processes are centrally coordinated by the Cancer Registry of Norway, which also stores the results of the screening test, as well as any later diagnosis of pre-cancer or cancer. Individuals are recalled for follow-up diagnostics if the screening test turns out to be positive. In remote areas of Norway there can be very few people invited to a screening exam, and such people can be identifiable in the screening databases of the registry, specially knowing the particular date and clinic where the screening took place. The information about one person's appointment can be known e.g., by close friends, family members, or neighbours.

Therefore, the question we consider is *can we anticipate the risk of a person/patient of being re-identified in advance and modify the screening process such that the screening database is anonymized by construction?* This is the question we address in this paper. A typical example would be to invite persons with the same gender, age and medical background to the same screening centre and at the same day. There is always a trade-off between available resources and how to orchestrate the screening process to maximize anonymity of individuals which may be modelled as a *constraint optimization problem*.

The contributions of the paper to address this goal are as follows:

C1: A new definition of anonymity which extends the notion of *k-anonymity* Our approach is applied to generate screening appointments, which implies that the screening results are still not known, and anonymity measurements that rely on the distribution of this sensible information such as *l-diversity* [13], and *t-closeness* [12] cannot be applied. In contrast, the concept of *k-anonymity* is based on the quasi-ids frequency, disregarding the screening result, and hence is more suitable to our purpose. In particular, *k*-

anonymity looks for the minimum number k of repetitions of a quasi-id value. For instance, a value $k = 1$ indicates that there is a set of particular personal attributes (a quasi-id value) that occurs only once in the dataset. Thus, the anonymity of the person with these characteristics is at risk. Although useful, we have found the concept of *k-anonymity* too restrictive in practice, and we show in the paper that datasets with the same k can have in fact very different levels of anonymity. Thus, our first contribution is an extension of *k-anonymity* based on the more general concept of *generalized k-anonymity*, based on a new measurement which we call *anonymity vectors*. Anonymity vectors counts how many quasi-id values are repeated once, twice, and so on, in the dataset, thus generalizing the idea of *k-anonymity*.

C2: Defining the optimal assignment of appointments for screening We present an approach based on constraint programming [4] to model the optimal appointments from the point of view of anonymity vectors. This is not useful in practice except for very small database sizes, due to the highly combinatorial nature of the problem, but allows us to compare other, more efficient, approaches with this 'ideal' appointment assignment on small databases. It is also useful as a precise definition of the problem.

C3: A practical technique for the assignment of appointments in medical screening programs We propose an algorithm that, although not optimal, provides very good results from the point of view of anonymity vectors. The idea is to assign the same appointment data (same screening centre or same test dates) to people with similar personal characteristics. The algorithm is easy to implement and very efficient in terms of time.

C4: Experiments with real data to confirm the adequacy of the technique We conduct several experiments to check the proposed algorithm. The experiments utilize data from the Cancer Registry of Norway. Our experiments compare the anonymity vectors for the already existing appointments in the initial dataset and the anonymity that could have been obtained applying our algorithm. The results reveal a significant increase of anonymity and therefore confirm the adequacy of the algorithm to the problem of anonymity.

The rest of the paper is organized as follows. Next section presents some related work. Section 3 shows a typical screening program from the point of view of data. Then, Section 4 examines how this process affects the anonymity of the data, in particular when appointment information is added to the database. The concept of *k-anonymity* is discussed, and the extension presented in this paper, *generalized k-anonymity*, is defined. The properties of anonymity vectors, the basic notion behind *generalized k-anonymity*, are discussed in Section 5. Section 6 models the problem of finding the optimal appointment from the point of view of anonymity using constraint programming. A more practical, although suboptimal, algorithm is introduced in Section 7. Section 8 discusses the efficiency of the algorithms using experimental data. The ethical implications of our approach are discussed in Section 9. Finally, Section 10 presents the conclusions and proposed future lines of work.

2. Related Work

This paper is, to the best of our knowledge, the first in considering the problem of anticipating anonymity risks in screening program databases. In contrast, existing anonymization techniques focus on anonymizing existing databases with data collected from the real-world. Instead, our approach uses constraint programming to generate appointments such that databases are *anonymous by construction*. Below, we present important related work in the area of anonymizing existing databases.

An obvious method to achieve anonymity is to release databases with data coarse-grained/generalized. The *k-anonymity* algorithm [19] is the most common approach, where data are aggregated so that there will be $k-1$ other individuals with the same attributes are found in the same equivalence class. This approach is in fact the starting point of our work, which presents a refinement of this technique which we call *generalized k-anonymity*.

In order to improve the k-level, generalization and suppression techniques [18] are usually employed. However, generalization can obscure detail in data, rendering it useless. In our case, we look for a better generalized *k-anonymity* level using neither suppression nor generalization, thus ensuring that the data quality is not decreased.

In [13], the authors identify a possible anonymity problem related to the concept of k-anonymity. It occurs when individuals within the same k-anonymity equivalence class share the same *sensitive value* on a variable. In these cases, the specific individual can be identified using other background information. ℓ-diversity adds diversity or heterogeneity to the sensitive attributes in each equivalence class with k records, avoiding the identification of those individuals who share a sensitive variable value. Another method, t-closeness [12], ensures that the distribution of values for a variable is close by a threshold t to the distribution of values in the original database. Selectively adding a random factor [1][15] has also been an approach to anonymize databases. These three disclosure risk measures have been studied an improved in many works. For instance, [10] presents a linear time approach for databases with multi-dimensional quasi-identifiers. A common characteristic of these measures is that they stress the importance of minimizing the maximum risk, that is, they focus on improving the anonymity of those individuals that can be identified more easily.

Although the approach of this paper follows this principle, our concept of generalized anonymity vectors also considers the rest of the individuals in the dataset, and thus can be considered a *global disclosure risk measure*, in the line of early works such as [20]. In this work, the authors propose the use of equivalence classes or clusters, which is related to our notion of subsets which the same anonymity level. The main difference between our proposal and [20] is that from this idea we define the anonymity vectors in order to compare the anonymity of different appointment schedules, while [20] proposes a different notion (*disclosure risk values*) with the purpose of comparing the effect of data aggregation.

3. Screening from the perspective of data

In order to better understand the goal of this paper, it is convenient to describe the phases considered in screening from the point of view of the data and the people who can access it.

3.1. Removing personal identifiers

In a first phase, the individuals that are going to be part of the program are selected from the general population. In the rest of the paper, except in the section devoted to the experimental results, we consider the small, fictional, example of selected population displayed in Table 1. The initial data contains the full name, address, ZIP code, gender and the age range, which in this case is limited to two intervals 50-54 and 55-59.

In this data subset the columns *Full Name* and *Address* are *identifiers*, since they contain information that can identify the individuals in the program. This information must be separated, and a first step is to replace it by an internal code, for instance corresponding to the number of row in the table. Now, the data subset is represented by two tables:

Table 2.(a) is kept internally, and only a few people of the screening organization have access to it, while Table 2.(b) is useful for evaluation and research after the screening has been completed. Making this data available

Full Name	Address	ZIP Code	Gender	Age
Roderick Alleyn	14 Brooks Rd.	88888	male	50-54
Martin Beck	21 Elm St.	11111	male	55-59
Nora Charles	5 Norgren St.	11111	female	55-59
Cordelia Gray	23B Martin St.	88888	female	55-59
Adam Dalgliesh	1 Linges St.	88888	male	50-54
Madelyn Hayes	811 Francis Ave.	11111	female	55-59
Jane Marple	12 Granite Lane	11111	female	55-59
Henry Merrivale	65 Baker St.	11111	male	55-59
Laura Monagan	9 Norgren St.	11111	female	55-59
Emma Page	221 Ohio St.	88888	female	55-59
Thomas Pitt	43 Creek Rd.	11111	male	55-59
Claire Rodgers	89 Lloyd Court	11111	female	55-59
Philip Trent	9 Bell St.	88888	male	50-54
Peter Wimsey	2C Ryan Ave.	88888	male	50-54
Charles Wycliffe	47 Hummel St.	11111	male	55-59

Table 1: Data sample

	Full Name	Address
1	Roderick Alleyn	14 Brooks Rd.
2	Martin Beck	21 Elm St.
3	Nora Charles	5 Norgren St.
4	Cordelia Gray	23B Martin St.
5	Adam Dalgliesh	1 Linges St.
6	Madelyn Hayes	811 Francis Av.
7	Jane Marple	12 Granite St.
8	Henry Merrivale	65 Baker St.
9	Laura Monagan	9 Norgren St.
10	Emma Page	221 Ohio St.
11	Thomas Pitt	43 Creek Rd.
12	Claire Rodgers	89 Lloyd Court
13	Philip Trent	9 Bell St.
14	Peter Wimsey	2C Ryan Ave.
15	Charles Wycliffe	47 Hummel St.

(a)

	ZIP	Gender	Age
1	88888	male	50-54
2	11111	male	55-59
3	11111	female	55-59
4	88888	female	55-59
5	88888	male	50-54
6	11111	female	55-59
7	11111	female	55-59
8	11111	male	55-59
9	11111	female	55-59
10	88888	female	55-59
11	11111	male	55-59
12	11111	female	55-59
13	88888	male	50-54
14	88888	male	50-54
15	11111	male	55-59

(b)

Table 2: Replacing identifiers by internal code

	ZIP	Gender	Age	Centre	Hour	Test
1	88888	male	50-54	A	9	✓
2	11111	male	55-59	A	9	✓
3	11111	female	55-59	A	9	✓
4	88888	female	55-59	A	9	✗
5	88888	male	50-54	A	13	✓
6	11111	female	55-59	A	13	✓
7	11111	female	55-59	B	9	✓
8	11111	male	55-59	B	9	✓
9	11111	female	55-59	B	9	✓
10	88888	female	55-59	B	9	✓
11	11111	male	55-59	B	13	✗
12	11111	female	55-59	B	13	✓
13	88888	male	50-54	C	13	✓
14	88888	male	50-54	C	13	✗
15	11111	male	55-59	B	13	✓

Table 3: Screening appointments

to researchers, is important because it can lead to important improvement of the screening program, as well as discovery of new relations, such as prevalence of the disease by age, gender or ZIP code.

3.2. Including appointment data and test results

The selected population is then assigned to an appointment for a medical test/examination. The appointment process usually involves several places (hospitals, screening centres) and for each place a number of slots of time, each one with its own *capacity* that is the maximum amount of people that can be attended in that slot.

Table 3 contains an example of possible appointments for our running example. We assume that there are three medical centres involved in the screening, A, B and C. Centre A admits 4 people at 9h, and two people at 13h. Centre B admits 4 people at 9h, and 3 people at 13h. Finally, C only admits two people at 13h. The dates of these appointments are the same, and have been omitted from the table.

Finally, after the screening test has been finished, a last piece of information is added to the table, namely the result of the screening test, in our example just an indication if the test is positive or negative (column Test).

It is important to remark that this column, the test result, is not considered in our setting as we focus on the phase of appointment assignments, which occurs before the screening test takes place and the test results are available.

4. Anonymity in the screening process

In this section, we describe how the screening stages described in the previous section can affect anonymity.

4.1. The role of quasi-identifiers

We have seen that a first, obvious, measure to increase privacy is to remove the personal identifiers, as shown in Table 2. However, the information obtained after removing the identifiers is not safe from the point of view of anonymity. In [22] is shown how in the 1990 U.S. Census, more than the 87% (216 million of 248 million) of the population in the United States can be identified from its 5-digit ZIP, gender and date of birth. For this reason, we say that the columns *ZIP*, *Gender* and *Age* form a *quasi-identifier* (quasi-id, in short). In the rest of the paper we assume tables where the personal identifiers have been removed, and that there is a set of attributes forming a quasi-id.

If we consider a table containing N individuals, the quasi-id attributes can take Q different values $v_1, \dots, v_Q$, with $Q \leq N$. In the rest of the paper we denote by q_i , $1 \leq i \leq Q$ to the number of rows in the table such that the quasi-id takes the value v_i. Obviously, $q_1 + \cdots + q_Q = N$.

In 2002, L. Sweeney proposed a simple method for measuring the degree of anonymity in a data set given a quasi-id [19]. The measurement is known as *k-anonymity*, and consists in determining the number of repetitions of each quasi-id value, and determining the minimum of all these values. This number is denoted by k, and the data set is said to verify a level k of anonymity. The next Definition formalizes this concept and introduces some notations employed in the rest of the paper.

Definition 1. k-anonymity

- *Let T be a table with N rows ($N \geq 0$) with a well-determined quasi-id.*
- *Let Q be the number of different values that takes the quasi-id in T, and let v_1, $\dots$, v_Q be these values.*

	ZIP	Gender	Age	Test
1	88888	male	50-54	✓
2	11111	male	55-59	✓
3	11111	female	55-59	✓
4	88888	female	55-59	✗
5	88888	male	50-54	✓
6	11111	female	55-59	✓
7	11111	female	55-59	✓
8	11111	male	55-59	✓
9	11111	female	55-59	✓
10	88888	female	55-59	✓
11	11111	male	55-59	✗
12	11111	female	55-59	✓
13	88888	male	50-54	✓
14	88888	male	50-54	✗
15	11111	male	55-59	✓

Table 4: Quasi-ids in the data set of table 2.(b)

- *Let q_i, $1 \leq i \leq Q$ be the frequency of v_i.*

Then, we define $k = min\{q_i \mid 1 \leq i \leq Q\}$, and say that T verifies k-anonymity with respect to the quasi-identifier q.

For instance, consider the Table 4 with $N = 15$ rows. It shows in different colors the $Q = 4$ quasi-identifier values in our example:

- $v_1 = \boxed{88888}\boxed{\text{male}}\boxed{\text{50-54}}$, repeated 4 times (thus, $q_1 = 4$).
- $v_2 = \boxed{11111}\boxed{\text{male}}\boxed{\text{55-59}}$, repeated 4 times ($q_2 = 4$).
- $v_3 = \boxed{11111}\boxed{\text{female}}\boxed{\text{55-59}}$, repeated 5 times ($q_3 = 5$).
- $v_4 = \boxed{88888}\boxed{\text{female}}\boxed{\text{55-59}}$, repeated twice ($q_4 = 2$).

Thus, the minimum frequency of the quasi-identifier values is 2,

$$k\text{= min } \{ \ 4, 4, 5, 2 \ \} = 2$$

q	T
A	✓
B	✓
C	✓
D	✗
D	✓
D	✓
D	✗
D	✗

(a)

q	T
A	✓
B	✗
C	✓
C	✓
C	✗
C	✓
C	✓
C	✗

(b)

q	T
A	✓
B	✗
B	✓
C	✗
C	✓
C	✓
C	✓
C	✗

(c)

q	T
A	✓
B	✗
B	✓
B	✓
C	✓
C	✓
C	✓
C	✗

(d)

	1	2	3	4	5	6
(a)	3	0	0	0	1	0
(b)	2	0	0	0	0	1
(c)	1	1	0	0	1	0
(d)	1	0	1	1	0	0

Global anonymity vectors

Table 5: Same k-anonymity but different generalized k-anonymity

and we say this table verifies 2-anonymity.

For instance, suppose that the attacker knows a female with age between 55 and 59 and living in a place with ZIP code 88888, and has access to the screening results. With this information is still not possible to guess the result of the test for this person, because there is another individual with the same quasi-id.

As a side note, observe that in order to preserve the anonymity we need to ensure that the test results for the two people are different, or in general that the set of tests associated to every quasi-id value (represented here by the column *Test*) follow the same distribution. This is related to the concepts of ℓ-diversity [13] and t-closeness [12]. However, in this work we consider the problem of k-anonymity in relation to appointments in screening. The two points of view are reconcilable. While k-anonymity tries to anticipate anonymity problems *before* the test results are obtained, a general screening procedure will also include measurements for increasing anonymity *after* the test results have been incorporated. Hence, in the rest of the paper we assume that the test results have a 'good' distribution, and focus on the anonymity problems coming from quasi-ids and appointment data.

Next, we introduce a novel refinement of the concept of k-anonymity employed in our screening appointment proposal.

4.2. Generalized k-anonymity

Consider now the examples (a)-(d) of Table 5. In all the cases the quasi-id is composed of just one column, *q*, while the test result corresponds to column T. The four tables verify 1-anonymity, since the minimum number of repetitions of quasi-id values is 1: in (a) the quasi-id value C is repeated just once, in (b) this happens with quasi-id values A and B, in table (c) with quasi-id value A, and also with A in table (d). Is then fair to say that the four tables are equally anonymous? The answer is *no*. Table 5.(a) includes 3 people whose test results can be disclosed if someone knows their quasi-id (quasi-ids A, B and C), Table 5.(b) contains two rows in this situation (A and B), while Tables 5.(c) and 5.(d) only have one quasi-id with one repetition (A). This means that Table 5.(a) is the least anonymous, followed by Table 5.(b).

In order to compare Tables 5.(c) and 5.(d), we need to check the next level of anonymity risk. Table 5.(d) contains no element with two elements, while Table 5.(c) has one quasi-identifier with two repetitions (B), and thus Table 5.(c) is less anonymous than Table 5.(d).

In general, we propose to establish an order that allows us to compare the anonymity of different tables. First, we need to introduce the new concept of anonymity vector:

Definition 2. Anonymity vector

Let T, N, Q, v_i and q_i be as defined in Definition 1. Then, we define:

- $k(j)$*, The number of different quasi-ids values with* j *repetitions in* T*:*

$$k(j) = |\,\{v_i \mid q_i = j,\, 1 \leq\, i \leq\, Q\}\,|$$

- *The* anonymity vector *of* T*,* $avector(T)$*:*

$$avector(T) = (k(1), k(2), \dots)$$

Thus, anonymity vector counts how many quasi-ids are repeated once, twice, and so on. In principle it is defined as an infinite vector, but it becomes constantly zero after some finite position p, $p \leq N$, and thus it allows a finite representation.

The right-hand side of Table 5 shows the first 6 positions of the *avector*s for the four tables at the left-hand side (the value at positions beyond 6 are zero in all the cases). For instance, the *avector* of Table 5.(a) is (3,0,0,0,1) indicating that it contains three quasi-ids with one repetition and one identifier with five repetitions. The *avector*s can be be employed for comparing the anonymity of two tables:

Definition 3. Generalized k-anonymity

Let T_1, T_2 be two tables with the same size N, each one with a well-defined quasi-id. Then, we say that:

1. *T_1 has the same generalized k-anonymity as T_2, and write $anon_k(T_1) = anon_k(T_2)$ when $avector(T_1) = avector(T_2)$.*
2. *T_1 has better generalized k-anonymity than T_2, and write $anon_k(T_1) > anon_k(T_2)$, when $avector(T_1) <_{LEX} avector(T_2)$, with $<_{LEX}$ the lexicographical order.*

The lexicographical order is the usual order for comparing numbers and strings. In the case of anonymity vectors, $v_1 <_{LEX} v_2$ means that there is a position j such that $v_1[j] < v_2[j]$ and $v_1[i] = v_2[i]$ for every $i < j$, that is the two vectors are equal in the first $j-1$ components, and the j-th position of v_1 is less than the same position in v_2 [6].

For instance, in the examples of Table 5 we have:

$$anon_k(5.(\mathrm{d})) > anon_k(5.(\mathrm{c})) > anon_k(5.(\mathrm{b})) > anon_k(5.(\mathrm{a}))$$

because, from the right-hand side of Table 5:

$$\underbrace{(1,0,1,1)}_{avector(5.(\mathrm{d}))} <_{LEX} \underbrace{(1,1,0,0,1)}_{avector(5.(\mathrm{c}))} <_{LEX} \underbrace{(2,0,0,0,0,1)}_{avector(5.(\mathrm{b}))} <_{LEX} \underbrace{(3,0,0,0,1)}_{avector(5.(\mathrm{a}))}$$

4.3. Influence of appointments in anonymity

The columns defining a screening appointment can be considered part of the quasi-identifier, although their values for a particular individual are not as easy to discover as the gender, ZIP code or age. The appointment can be either found by chance, by meeting someone known at the medical center, or in the case of famous people if they are followed by journalists or photographers.

Thus, this information can create a security risk when combined when the rest of the quasi-identifiers and the screening result, which can be obtained either because they are published as part of the results, made available to some researches, or simply because this information is available to different people during the development of the screening.

Table 6 shows the quasi-identifier values after extending the quasi-identifier in Table 3 with the appointment data. The anonymity vector of this table is then (11,2), meaning 1-anonymity for 11 non-repeated quasi-identifier values

	ZIP	Gender	Age	Centre	Hour	Test
1	88888	male	50-54	A	9	✓
2	11111	male	55-59	A	9	✓
3	11111	female	55-59	A	9	✓
4	88888	female	55-59	A	9	✗
5	88888	male	50-54	A	13	✓
6	11111	female	55-59	A	13	✓
7	11111	female	55-59	B	9	✓
8	11111	male	55-59	B	9	✓
9	11111	female	55-59	B	9	✓
10	88888	female	55-59	B	9	✓
11	11111	male	55-59	B	13	✗
12	11111	female	55-59	B	13	✓
13	88888	male	50-54	C	13	✓
14	88888	male	50-54	C	13	✗
15	11111	male	55-59	B	13	✓

Table 6: Screening appointments

(depicted as white rows in the Table). In contrast, the anonymity vector before including the appointments was (0,1,0,2,1) (see Table 4) which corresponds to a level of 2-anonymity.

In this small example, it is possible to use a simple program to show that there are 14504 possible appointment schedules satisfying the resources maximum capacity. Of these, 14446 (more than the 99 percent) correspond to $k = 1$, while 58 possible appointment schedules verify k-anonymity with $k \geq 1$. However, in this small case it would be easy to find one of these 58 2-anonymity appointment schedules. We could even find the optimal assignment from the point of view of generalized k-anonymity in a few seconds, simply trying all the combinations. However, in a bigger, real example, involving several hundreds or even thousands of individuals the exponential explosion in the number of assignments makes the problem unsolvable using a simple combinatorial search.

Before presenting our solution to this problem, we need to study anonymity vectors in depth.

5. Anonymity vectors

The main goal of this section is to define a suitable notion of distance between anonymity vectors. We start introducing some easy properties and definitions.

5.1. Properties

Let v be an anonymity vector for a population of size N. Then:

P_1 The k-anonymity of the population represented by v corresponds to the position of the leftmost non-zero component of v. That is, $(5, 2, 2, 0, 0)$ and $(5, 2, 2)$ represent the same anonymity vector. In general, we call *size* of an anonymity vector v, denoted as $|v|$ to the rightmost position in v with a non-zero value. For instance, if $v = (5, 2, 2, 0, 0)$, then $|v| = 3$.

P_2 Let v_1, v_2 two anonymity vectors for two tables T_1, T_2 of size N, such that T_1 verifies k_1-anonymity and T_2 k_2-anonymity. Then, $v_1{<}v_2$ implies $k_1 \geq k_2$.

P_3 The following equality holds:

$$\sum_{i=1}^{|v|} v[i] \times i = N$$

For instance, the anonymity vector $v = (5, 2, 2, 0, 0)$ corresponds to a population of $5 * 1 + 2 * 2 + 2 * 3 = 15$ (meaning that 5 quasi-id values have a frequency of 1, two quasi-id values have a frequency of 2, and another two quasi-id values occur three times each one).

Properties P_1, P_2 indicate that generalized k-anonymity refines the concept of k-anonymity and that it is a conservative extension, respectively. Property P_3 observes that anonymity vectors contain implicitly the population size.

5.2. Partitions

An interesting and relevant question is: *how many anonymity vectors exist for a population of size N?*. To solve this question we must consider the problem of *partitions* [3]. In number theory and combinatorics, a partition of a positive integer N is a way of writing N as a sum of positive integers, where different orders of the summands are not considered to be distinct. For instance, the number $N = 4$ has 5 partitions. The interesting point for us, is that each partition corresponds to an anonymity vector for the same N.

Partition of 4	Anonymity vector
4	(0,0,0,1)
1+ 3	(1,0,1)
2+2	(0,2)
1+1+2	(2,1)
1+1+1+1	(4)

In fact, it is easy to prove that there is a bijective correspondence between partitions and anonymity vectors, and thus that the number of anonymity vectors for a given population size N is the number of partitions of N, denoted in number theory as $p(N)$. The value $p(N)$ has no known explicit expression, but it can be obtained defining the following auxiliary recursive expression (easy to prove using induction on N, M):

$$p\prime(N, M) = \begin{cases} 0 & \text{if } N < 0 \\ 1 & \text{if } N = 0 \text{ or } M = 1 \\ p\prime(N - M, M) \; + \; p\prime(N, M - 1) & \text{otherwise} \end{cases}$$

$p\prime(N, M)$ represents the number of partitions of N with summands less than or equal to M. Thus $p(N) = p\prime(N, N)$. The idea behind $p\prime(N, M)$ is that if $N < 0$ there are no partitions, if $N = 0$ only the partition 0 is allowed (by convenience 0 can be used in the partition of $N = 0$). Also, if $M = 1$ there is just one partition, represented by $\underbrace{1 + \cdots + 1}_{N}$, or analogously by the anonymity vector (N). Using techniques of memoization [17] to avoid repeating recursive calls, the recursive function can be applied to medium-size population (thousands of individuals). For instance, for $N = 5000$ the function yields an integer which can be approximated as $p(N) \sim 1.698 \times 10^{74}$ after a few seconds.[4] If we wish to approximate the value of $p(N)$

[4] Time: 11 seconds. Processor: 4x Intel(R) Core(TM) i7-6560U CPU at 2.20GHz, 16 GB memory, operating system Ubuntu 16.04 LTS. Program in Java using the *BigInteger* package, Java Virtual machine argument -Xss8M. Exact value 169820168825442121851975101689306431361757683049829233322203824652329144349.

for larger values of N, we can use the asymptotic formula proposed by the mathematicians G. H. Hardy and S. Ramanujan found in 1918 [9]:

$$p(N) \sim \frac{1}{4N\sqrt{3}} \exp\left(\pi\sqrt{\frac{2N}{3}}\right) \text{ as } N \to \infty$$

For values of $N \geq 5000$ the percent error of this formula is below 0.3%.

5.3. Measuring anonymity improvement

In order to compare different anonymity options for screening programs we use the concept of *lexicographic index*, which is the position that an anonymity vector v in the lexicographical order. We represent this number as $index(v)$. We assume that for every N,

$$index(\,(N)\,) = 0 \qquad \text{and} \qquad index(\,(\underbrace{0, \ldots, 0}_{N-1}, 1)\,) = p(N)$$

where p the function defined in the previous subsection, (N) is the anonymity vector of a set of N individuals with 1-anonymity, that is the worst possible anonymity, and $(\underbrace{0, \ldots, 0}_{N-1}, 1)$ represents a set of individuals whose quasi-id values are all different, which corresponds to the best possible level of anonymity.

Let b be the anonymity vector corresponding to some population with some well-defined quasi-id, and let v_1, v_2 be two anonymity vectors obtained after extending this quasi-id with the appointment information, and after generating the appointments with two different techniques. Then, *anonymity improvement* of v_1 with respect to v_2 is the value:

$$\frac{index(v_1) - index(v_2)}{index(b)}$$

The idea behind this definition is to compare v_1 and v_2 with respect to the starting point from the point of anonymity, which is the base anonymity we already have before introducing appointments.

Observe that after extending a quasi-id including new attributes the anonymity can only decrease, and thus $0 \leq index(v_1) \leq index(b)$, $0 \leq index(v_2) \leq index(b)$, and thus the improvement is a value between -1 and 1, where negative values indicate that v_1 is in fact decreasing the anonymity with respect to v_2. By convenience we the improvement as 0 when $index(b) =$

0, because in this case $index(v_1) = index(v_2) = 0$ and no improvement is possible.

6. Anonymity as an optimization problem

Our goal is to obtain a set of appointments that minimizes the loss of anonymity when considering the appointment information as part of the table quasi-id. In this section, we do this applying *constraint programming* [14], a programming paradigm that *models* problems by defining variables with some initial domain (possible set of values), and then defines constraints that these variables must hold. A *solver*, that is, a program specialized in finding values for the variables that satisfy all their constraints, is then employed in order to find the possible solutions. When finding one (or more) solutions satisfying the constraints is the goal, we talk about a constraint satisfaction problem. When look for the solution that minimizes/maximizes some function we talk about an optimization problem.

In this paper we use the system MiniZinc [16], a system defining a clear modelling language and many off-the-shelf solvers. In particular, we use here the *finite domain solver*, employed when the initial domain of the variables contains a finite set of values, and define the problem of obtaining the best appointment as an optimization problem.

In order to define our MiniZinc model, we start with a population table similar to Table 2.(b). From this table is very easy to obtain a table of quasi-id values with their frequencies, represented by Table 7.(a). We also assume the existence of a table of *resources* which contains the possible medical centres, the hours or the different available slots, and the capacity (number of people that can be admitted) in each slot. Table 7.(b) contains the resources of our running example. From now on we represent each quasi-id and each resource by its position in Tables 7.(a) and 7.(b), respectively.

These two tables correspond to the initial parameters of our model. In particular, we call:

- Q: Number of quasi-id values, that is number of rows in Table 7.(a).
- f: Frequency of each quasi-id. It corresponds to column f of Table 7.(a). We use the notation $f[i]$ to refer to the frequency of the i-th quasi-id in the table, $1 \leq i \leq Q$.
- R: number of resources, that is number of rows in Table 7.(b).

	ZIP	Gender	Age	f
1	88888	male	50-54	4
2	11111	male	55-59	4
3	11111	female	55-59	5
4	88888	female	55-59	2

(a)

	Centre	Hour	Capacity
1	A	9	4
2	A	13	2
3	B	9	4
4	B	13	3
5	C	13	2

(b)

Table 7: Quasi-id & Resources

- c: capacity of each resource. It corresponds to column *capacity* of Table 7.(b). We use the notation $c[j]$ to refer to the capacity of the j-th resource in the table, $1 \leq j \leq R$.

With these parameters, we define a two-dimensional array a of $Q \times R$ integer decision variables, that represent the optimal appointment schedule for our problem. The decision variables are variable whose values will be computed by the constraint solver, satisfying the requirements of the model. In this array, $a[i, j]$ represents the number of people with quasi-id i ($1 \leq i \leq Q$) that are assigned to resource j ($1 \leq j \leq R$).

In order to obtain the appointment with the best generalized k-anonymity we need to find the appointment with the smaller anonymity vector (see Definition 3). This means that we must choose values a_{ij} as high as possible in order to increase the anonymity. Algorithm 1 shows a technique for obtaining the optimal anonymity vector iteratively.

The algorithm starts by computing in the parameter p the total population, obtained as the sum of the frequencies of all the identifiers. The assignment $l \leftarrow 1$ indicates that the next iteration of the loop will compute the first component of the anonymity vector v.

The condition in the *while* loop checks if the current value in v covers the totality of the population. If this is the case, we can ensure that rest of the anonymity vector components are zero, and the algorithm stops. Otherwise, the next iteration determines the value of the l-th position. This is done looking for the minimum number of quasi-ids with l repetitions. This ensures that, by construction, the vector is minimal with respect to the lexicographic order, and in consequence that the appointment is optimal with respect to generalized k-anonymity.

```
Algorithm 1 Appointment with optimal generalized k-anonymity

Input: $Q$, $f$, $R$, $c$, as defined above.
Output: An array $a$ with optimal generalized k-anonymity, and $v$, its associated anonymity vector.

Define:
    $p \leftarrow \sum_{i=1}^{Q} f[i]$    % total population
    $l \leftarrow 1$    % next level of the anonymity vector to be computed
    $v \leftarrow [\,]$    % anonymity vector, initially with 0 components

While $\sum_{n=1}^{l-1} v[n] \times n < p$ do % compute $l$-th component of $v$
    Define and solve the following constraint optimization problem:
        Variables:
        - $a$: bidimensional array containing $Q \times R$ decision variables
          with domain the non-negative integers
        - $k$: integer decision variable containing the next component
          of the anonymity vector

        Constraints:
          $(C_1)$ For $i = 1 \ldots Q : \sum_{j=1}^{R} a[i,j] = f[i]$

          $(C_2)$ For $j = 1 \ldots R : \sum_{i=1}^{Q} a[i,j] \leq c[j]$

          $(C_3)$ For $n = 1 \ldots l-1$ :
                $|\ \{\, a[i,j] \mid i = 1 \ldots Q,\ j = 1 \ldots R,\ a[i,j] = n \,\}\ | \ = \ v[n]$

          $(C_4)$ $k \ = \ |\ \{\, a[i,j] \ \mid \ i = 1 \ldots Q,\ j = 1 \ldots R,\ a[i,j] = l \,\}\ |$

        Problem Goal: minimize k;

    Increase the size of $v$, add a new component at the right (position l)
    $v[l] \leftarrow k$, with $k$ the value obtained by the constraint solver
    $l \leftarrow l + 1$
end While
return  $a$ as the best appointment schedule, with $v$ as anonymity vector
```

The optimization problem starts defining two variables, k that will contain the new element of the vector, and a, that contains the variables that will constitute the optimal appointment. It is worth noticing that the array a is new in each iteration, that is, if there is a previous array a it is removed and a new array is defined.

The optimization problem is defined by four constraints. The constraints C_1, C_2, define valid appointment schedules. C_1 indicates that each row i of a must provide appointments for all people with the quasi-id i. That is, everybody is appointed. C_2 ensures that the appointment does not exceed the capacity of any resources. C_3 and C_4 ensure that the appointment is not only valid but also optimal. C_3 ensures that the new solution keeps the already computed values for $v[n]$ with $1 \leq n \leq l-1$. C_4 indicates that k is the new value for v counting the number of repetitions of value l. The goal of the optimization problem is to minimize the new value for v. Finally, the new value is stored in v.

After the last iteration of the loop, a contains the optimal appointment schedule. In our example, the algorithm needs four iterations displayed in Table 8. Each iteration determines one position of the final generalized anonymity vector. This is indicated in the first column, which corresponds to variable *level*. The second column displays the anonymity matrix a after the iteration. The third column indicates the part of the anonymity vector that is already determined after the iteration. Finally, the last column shows the complete anonymity vector associated to the array a. In the last iteration the values of *anom* and v must be the same, since *anom* contains now the whole anonymity vector.

In order to relate a and with the real appointment schedule, observe for instance that in the last iteration, the first row of a indicates that the 4 individuals with quasi-id 1 are appointed to the third resource, which corresponds to medical centre B at 9h. Table 9 shows this appointment including the quasi-id and resource data.

The new anonymity vector $(0, 2, 1, 2)$ is more anonymous than the vector $(11, 2)$ of the random appointment in Table 6. Observe that it even keeps the 2-anonymity that the table had before the appointments, whose anonymity vector was $(0, 1, 0, 2, 1)$.

It is worth observing that the optimal algorithm presented here creates an appointment schedule that maximizes the worst anonymity level, that is, that increases the anonymity of the individuals which are at risk at the cost of a possible decrease in the anonymity of others. Informally, we might

level	a	anom	v
1	$\begin{pmatrix} 0 & 0 & 2 & 0 & 2 \\ 2 & 2 & 0 & 0 & 0 \\ 0 & 0 & 2 & 3 & 0 \\ 2 & 0 & 0 & 0 & 0 \end{pmatrix}$	(0)	$(0, 6, 1, 0)$
2	$\begin{pmatrix} 0 & 0 & 4 & 0 & 0 \\ 4 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 3 & 2 \\ 0 & 2 & 0 & 0 & 0 \end{pmatrix}$	$(0, 2)$	$(0, 2, 1, 2)$
3	$\begin{pmatrix} 4 & 0 & 0 & 0 & 0 \\ 0 & 0 & 4 & 0 & 0 \\ 0 & 0 & 0 & 3 & 2 \\ 0 & 2 & 0 & 0 & 0 \end{pmatrix}$	$(0, 2, 1)$	$(0, 2, 1, 2)$
4	$\begin{pmatrix} 0 & 0 & 4 & 0 & 0 \\ 4 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 3 & 2 \\ 0 & 2 & 0 & 0 & 0 \end{pmatrix}$	$(0, 2, 1, 2)$	$(0, 2, 1, 2)$

Table 8: Iterations of the algorithm 1 for the running example

	ZIP	Gender	Age	Centre	Hour	Test
1	88888	male	50-54	B	9	✓
2	11111	male	55-59	A	9	✓
3	11111	female	55-59	B	13	✓
4	88888	female	55-59	A	13	✗
5	88888	male	50-54	B	9	✓
6	11111	female	55-59	B	13	✓
7	11111	female	55-59	B	13	✓
8	11111	male	55-59	A	9	✓
9	11111	female	55-59	C	13	✓
10	88888	female	55-59	A	13	✓
11	11111	male	55-59	A	9	✗
12	11111	female	55-59	C	13	✓
13	88888	male	50-54	B	9	✓
14	88888	male	50-54	B	9	✗
15	11111	male	55-59	A	9	✓

Table 9: Optimal screening appointment

say that in our algorithm individuals with good anonymity level "sacrifice" part of their anonymity in order to increase the anonymity of those whose anonymity is at risk.

For instance, imagine a small dataset with eight individuals, and a quasi-id which already contains the appointment data. Suppose that the quasi-id takes two values, a and b. Suppose also, that the random appointment obtains the quasi-id values distribution, a, b, b, b, b, b, b, b, which corresponds to the anonymity vector (1, 0, 0, 0, 0, 0, 1). This means that the individual with quasi-id value 'a' is in a risky situation with respect to anonymity, with only just one repetition. Instead, provided that there are enough resources available, our algorithm finds the optimal solution a, a, a, a, b, b, b, b, which corresponds to the anonymity vector (0, 0, 0, 2). The result is really good, since we have increased the k-level from 1 to 4, and the quasi-id value a is now repeated four times. However, this increase is achieved by decreasing the anonymity level of the seven individuals with quasi-id value b in the first assignment schedule, which has passed from seven repetitions to just four.

7. A heuristic solution

The constraint programming model of Section 6 defines and obtains the optimal appointment with respect to the generalized k-anonymity. However, we will see in Section 8 that this solution is only applicable to very small datasets due to its high inefficiency in terms of time. This is not surprising, because anonymity vectors refine the concept of k-anonymity, and optimizing the k-anonymity of a dataset is already a NP-hard problem [2].

Thus, we propose a simple heuristic algorithm that, although does not guarantee optimality, provides good results in general.

The basic idea is sketched in Algorithm 2. The inputs of this algorithm are the same as those of the algorithm 1. First, the algorithm declares some variables, such as the returned values v (anonymity vector) and a (appointments array). It also declares a vector o which contains the quasi-id value frequencies already stored in f, but ordered in descending order. In each step of the main *for* loop, the algorithm generates appointments for all the people whose quasi-id value corresponds to the frequency $o[i]$.

The value $o[i]$ is copied to the auxiliary variable q inside of the *for* loop. The inner *while* loop iterates until $q = 0$, ensuring that all the people with the associated quasi-id have been appointed to some resource when the loop exits.

Algorithm 2 Heuristic Appointments

Input: Q, f, r, c, as defined in section 6, that is: Q total number of quasi-id values, f the frequency of each value, r the number of resources and $c[i]$ the capacity of each resource.

Output: An array a with optimal generalized k-anonymity

Variables and initializations:

```
p        ← $\sum_{i=1}^{Q} f[i]$                    % total population
o        ← sort(f)                                   % f sorted in descending order
a        ← $[[0,\ldots,0],\ldots,\ [0,\ldots,0]]$    % array of dim. QxR
v        ← $[0,\ldots,0]$                            % vector of dim. p
q,i,j,h  ← 0                                         % auxiliary variables
For each $i \in 1 \ldots Q$ :
  q = o[i]
  % find appointments for people with the $i-th$ quasi-value

  while o[i] > 0
    % Find a resource
    If exists at least one $h$ such that c[h] = q then
        Let $j$ be any of the indices $h$ such that c[h] = q
        a[i,j] = q
        c[j] = q = 0
    elseif exists some $h$ such that c[h] > q then
        Let $j$ be the index of the maximum c[j]
        a[i,j] = q
        c[j] = c[j] - q
        q = 0
    else % there is no resource with capacity enough for $q$
        Let $j$ be the index of the maximum c[j] less than or equal to q/2
        a[i,j] = c[j]
        q = q - c[j]
        c[j] = 0
    end if
end for
for i ∈ $1 \ldots Q$, j ∈ $1 \ldots R$ % obtain v from a
  v[a[i,j]] = v[a[i,j]] + 1
end for
return  $a$ as the best appointment, with $v$ as anonymity vector
```

Inside of the *while* loop three cases are distinguished using and *if . . . elseif . . . else* statement:

1. At least one resource h fits perfectly q, that is $c[h] = q$. In this case we pick up any of the resources with this property, called j in the code, appoint all the people to this resource ($a[i, j] = q$), and indicate that the resource is empty ($c[j] = 0$) and that all the remaining people for this quasi-id have been appointed ($q = 0$).
2. There is no resource whose capacity is exactly q, but some of them have a capacity greater than q ($c[h] > q$ in the code). As in the previous case, we will pick up one of them and appoint all the people to this resource, which will end the *while* loop. However, in this case the particular index j must be carefully chosen.
   Consider for instance that $q = 3$ and the resources that cover q have capacities $c_1 = 4$, $c_2 = 6$, and $c_3 = 7$. Any of then is a good choice for the current quasi-id iteration, but if we choose c_1 with capacity 4 we left c_1 with $4 - 3 = 1$, which means that if this resource is needed in the future it will result in 1-anonymity. Therefore the best choice is the maximum capacity, that is c_3, and the code of this case chooses the maximum value $c[j]$.
3. The last case (the *else* branch), is employed when no resource can cover the number of people in q. In this case more *while* iterations are needed before reaching $q = 0$, and any selected resource will change its capacity to 0. However, it is difficult to choose the best resource in this step because:
   (a) On one hand, choosing the greatest resource available seems good in order to end with the appointments in q as soon as possible, but it gives raise to a decrease in the final anonymity level. For instance, imagine that $q = 4$ and $c_1 = 1$, $c_2 = 2$, and $c_3 = 3$. If we choose the resource with the maximum capacity, c_3, after this iteration we will have $q = 1$, $c_1 = 1$, $c_2 = 2$, $c_3 = 0$. In any case, $q = 1$ implies that we are enforcing again 1-anonymity, which is precisely what we want to avoid.
   (b) On the other hand, choosing the smallest resource available improves the perspectives for the next iteration, but can anticipate the problem to the current iteration. Suppose now $q = 4$, $c_1 = 1$, $c_2 = 2$, $c_3 = 3$. In this case the resource with the smallest capacity is $c_1 = 1$, but choosing means ensuring 1-anonymity in the current loop iteration.

After some experiments, we have found that finding the resource capacity closest to value $q/2$ gives the best results.

For instance, consider again our running example. In this case $Q = 4$, and thus the main loop iterates four times, each one adding a new row to the appointment a. The final appointment schedule a is

$$\begin{pmatrix} 4 & 0 & 1 & 0 & 0 \\ 0 & 0 & 3 & 1 & 0 \\ 0 & 2 & 0 & 2 & 0 \\ 0 & 0 & 0 & 0 & 2 \end{pmatrix}$$

which corresponds to the anonymity vector $(2, 3, 1, 1)$. As expected, this vector is greater (less anonymous) than the optimal vector $(0, 2, 1, 2)$ found in Section 6. However, it corresponds to the rank 339 out of 14504 possible appointment schedules, and thus belongs to best two percent of the possible appointment schedules.

Although our approach is simple, and many improvements could be introduced, the experiments of Section 8 show that in practice it produces very good results.

8. Experimental results

In this section, we check experimentally the benefits in terms of anonymity improvement of the algorithms presented in Section 6 and 7.

8.1. Random Appointments

In this first set of experiments we compare three appointment techniques:

1. Random appointments, where each individual is assigned to some resource randomly, only taking into account the capacity of each resource.
2. Constraint programming. The approach of Section 6, which produces optimal appointments from the point of view of anonymity.
3. Heuristic approach. The simple heuristic system presented in Section 7.

Table 10 shows the results of our experiments. In the Table, each improvement value corresponds to the average of 10000 experiments if $N \leq 250$, and to the average of 1000 experiments for $N > 250$. Each experiment applies the appointment techniques to a random population generated with the following parameters:

N	Heuristic / random	C.P. / Heuristic
20	24%	1%
25	20%	2%
30	36%	5%
35	42%	12%
40	54%	14%
45	50%	14%
50	62%	10%
100	82%	_
250	93%	_
500	92%	_
1000	92%	_

Table 10: Comparing random appointments, the heuristic algorithm and the optimal result

1. Number of quasi-id values: $Q = N/100 + 5$. This formula has been chosen because it approximates quite well the real data of several possible screening scenarios analyzed. Small changes in this formula have no effect in the results of the experiments. However, it is worth mentioning that other possible proportions between N and Q can yield very different results. For instance, if we choose $Q = N/10 + 5$ and $N = 1000$, then the improvement of the heuristic approach over the random assignment drops from 92% (Table 10) to just about the 35%. This happens because in this case there is a large number of quasi-id values ($Q = 1000/10 + 5 = 105$). Thus, the anonymity base is already very low, and the appointment algorithm has little space for improvement.
2. The quasi-id value frequencies are chosen from a random distribution around the mean N/Q with a standard deviation of 20%, again realistic from our experiments with real data.
3. In the case of the resources (the appointment slots), we assume that in average a resource has a capacity of 10 people, allowing a standard deviation of 20%.

The first column indicates different population sizes. The second column represents the anonymity improvement of the heuristic algorithm with respect to the pure random assignment. Finally, the third column shows the improvement of the optimal solution, obtained by the constraint programming model, with respect the heuristic approach. Observe that this third column is not included for population sizes over 50, because the constraint model takes too much time in finding a solution (over one hour in a standard computer for N=55). However, the optimal solution is useful for measuring the efficiency of the heuristic approach. For instance, for $N = 45$ the heuristic algorithm introduces an improvement of a 50% with respect to the random appointments, while the optimal solution yields another 14% of improvement.

For larger values, the table shows how the improvement of anonymity of the heuristic approach with respect to the random appointments increases reaching values over 90%. It is worth to comment that the heuristic algorithm can deal with values of N over one million people in less than one second in a standard computer. However, the table only considers the maximum size $N = 1000$ because the function *index* is quite inefficient, and the measurement of the improvement becomes unfeasible for greater values.

8.2. A real case: the Cancer Registry of Norway

Our second set of experiments is based on data from of a hypothetical Norwegian Cancer Screening program, using Norwegian population data to form a realistic population to be screened. Although this dataset is hypothetical, it is based on the demographics of the Norwegian cervical cancer screening program. We have divided Norway into 5 regions, and within each region there are associated screening centres, between 3 and 13 within each region. We have only assigned 1 day each month at each centre to perform screening, the first screening invitation date corresponds to January of 1992 and the last to March of 2015. This amounts to a total of 282 different screening dates.

In our simulation, we consider each of the 282 different screening dates a different screening program. Moreover, since the set of screening centres depend on the region we consider the table the union of five datasets, one for each region, which amounts to a total $282 \times 5 = 1410$ experiments. In practice, this means that there were 1410 times that invitations needed to be issued assigning participants to screening centres. To achieve this, based on the Norwegian population and the capacity of the centres, we created a table of approx 2,4 million screening invitations, randomly assigning persons to invitations. We then compared the anonymity of this dataset with the anonymity that wE have obtained by using the heuristic algorithm of Section 5. The dataset schema consists of the following relevant columns:

region	Appointments	Dates	centres	Min.(%)	Max.(%)	Avg.(%)
1	1370100	282	13	65.4	96.9	82.6
2	348487	282	4	1.7	65.3	34.4
3	303226	282	3	35.8	92.2	58.8
4	244990	282	3	9.5	89.3	53.7
9	132395	282	3	0.0	21.4	3.6

Table 11: Improvement of anonymity in the simulation with the data from the Cancer Registry of Norway

- *reg.* This column determines the region within Norway.
- *year.* The birth year of the patient. The minimum value is 1904 and the maximum value 1996.
- *screening_date.* The screening invitation date.
- *center_nr.* The number of centres where the screening took place.

For each region and screening date, the basic anonymity vector b corresponds to the single attribute quasi-id *(year)*. The extended quasi-id $(year, center\_nr)$ produces the anonymity vector v_1, which represents the anonymity vector after randomly adding the appointed screening centre. This vector is obtained from the data already included in the table. Finally, our heuristic algorithm generates a new assignment of screening centres and thus a new anonymity vector v_2. The anonymity improvement of the heuristic algorithm with respect to the original data is obtained as $\frac{index(v_1)-index(v_2)}{index(b)}$ following the ideas of Subsection 5.3.

Table 11 shows the results of the experiments for the five regions. The first column indicates the region identifier and the second column is the number of table rows that correspond to this region. Column *Dates* includes the number of different screening date values, and column *centres* the number of screening centres for this region. The last three columns, *Min.*, *Max.* and *Avg.* indicate the minimum, maximum and average anonymity improvement. On average, the anonymity improves in all the cases, ranging from a huge improvement of 82.6% in the first region to a small but still positive improvement in the last region.

The application that obtains the data of the experiments is available at `https://github.com/RafaelCaballero/anonymity-experiments`

9. Ethical considerations

The appointment generation can have different ethical implications that must be taken into account.

E1. A particular screening centre cannot be convenient for the patient In our experiments, we have divided the population into regions, allowing only the appointment to screening centres in the same region. This is easy to implement in practice, either splitting the population in advance, or adding distance constraints to the algorithms.

However, in some cases there is just one screening centre that can be chosen without affecting the patients comfort. Even in this case, our proposal can make sense in there are enough people to use several days or time slots. In this cases we would consider as a resource the date and time, and use the heuristic algorithm presented in Section 7.

A different issue arises when the patient cannot attend to the appointment. Maybe someone forgets the appointment, or the location provided is difficult to access (for instance for people with limited mobility). If the associated quasi-id has been split into two or more slots, the application/staff in charge of this service should offer first the alternatives in the same area which already have people with the same identifier. If this is not possible, then there is nothing that can be done, and the person's data and results will be surely anonymized *after* obtaining the screening results (not discussed in this paper) since it will verify 1-anonymity.

E2. Too many people with the same quasi-id meet each other in the screening centre

Imagine someone living in a foreign country who receives an appointment for some screening exam. She attends to the screening centre, date and time indicated in the appointment, and finds that all the people waiting there for the screening exam have her own country of origin. It is only logical that this person thinks that the screening is only aimed to people from this country.

This situation can arise if the quasi-id includes attributes such as race or country of origin. Perception of discrimination in health screening is a hot issue [5, 21] which must be avoided.

Fortunately, it is easy to introduce additional constraints that ensure that this situation will not arise. For instance, we can add an additional constraint to the model of Algorithm 1 (Section 6) to ensure that each appointment includes no more than 15 percent of people from different countries of origin from the country where the screening exam is taking place, and that among this 15 percent there are at least three different nationalities.

Adding these constraints can result in a final decrease of the generalized anonymity, but they are necessary to avoid a possible perception of discrimination.

10. Conclusion

The publication of medical data requires a delicate trade-off between sharing information for preventive health measures and research on one side, and preserving individual privacy on the other side. In the case of screening programs, appointment data can be an additional risk factor, because knowing this information can contribute to deidentification among people with same individual characteristics (age, sex, ZIP code, etc.).

In this paper, we propose to have the anonymity in mind when preparing the screening program appointments, and select appointments that include several people with the same individual characteristics in the same resource (screening centre and date, for instance). Our proposal is independent of other usual measurements such as generalization, suppression or micro-aggregation [8], and our experiments show that it can contribute to keep the anonymity from the beginning of the program, not only with respect to the final publication of results, but also considering the anonymity of the individuals during the process, that is with respect to both administrative and clinic staff involved in the screening program.

We propose two algorithms. The first one is presented as a constraint programming model. Although quite inefficient in practice in terms of time, this first model defines the optimal appointments with respect to our notion of anonymity, and allows us to compare other algorithms with respect to the optimal result, at least for experiments involving a small population size.

The second algorithm is a heuristic approach that, although only provides suboptimal appointments with respect to anonymity, performs very well in practice, both in terms of efficiency and in terms of the anonymity level reached. The algorithm is very easy to implement and, although not discussed here for the sake of simplicity, allows introducing several considerations that can be needed in practical cases, such as including only clinics that are within a given distance of each individual address.

An additional contribution of the paper is the introduction of anonymity vectors as a refinement of the well-known k-anonymity measurement. While k-anonymity only considers the worst possibility of a privacy breach, our proposal considers *how many* individuals are in this situation, repeating the operation with the rest of the individuals. The final result is a vector presenting the different levels of anonymity risk, each one with the number of people involved. Our technique shows that datasets with the same k-anonymity can have in practice very different levels of privacy.

Acknowledgements.. We would like to thank Mari Nygard and Giske Ursin for preparing the dataset used in the experiments, and to Richard J. Cabana for his contribution to the algorithm implementation.

References

[1] C. C. Aggarwal and S. Y. Philip. A general survey of privacy-preserving data mining models and algorithms. In *Privacy-preserving data mining*, pages 11–52. Springer, 2008.

[2] G. Aggarwal, T. Feder, K. Kenthapadi, R. Motwani, R. Panigrahy, D. Thomas, and A. Zhu. Anonymizing tables. In *Database Theory - ICDT 2005, 10th International Conference, Edinburgh, UK, January 5-7, 2005, Proceedings*, pages 246–258, 2005.

[3] G. E. Andrews. *The theory of partitions*. Encyclopedia of mathematics and its applications. Cambridge University Press, 1984.

[4] K. Apt. *Principles of Constraint Programming*. Cambridge University Press, New York, NY, USA, 2003.

[5] L. L. Black, R. Johnson, and L. VanHoose. The relationship between perceived racism/discrimination and health among black american women: a review of the literature from 2003 to 2013. *Journal of Racial and Ethnic Health Disparities*, 2(1):11–20, 2015.

[6] R. Fagerberg. String sorting. In M. Y. Kao, editor, *Encyclopedia of Algorithms*, pages 2117–2121. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2016.

[7] L. Frank. When an entire country is a cohort. *Science*, 287(5462):2398–2399, 2000.

[8] B. C. M. Fung, K. Wang, R. Chen, and P. S. Yu. Privacy-preserving data publishing: A survey of recent developments. *ACM Comput. Surv.*, 42(4):14:1–14:53, June 2010.

[9] S. R. G. H. Hardy. Asymptotic formulæ in combinatory analysis. *Proceedings of the London Mathematical Society*, 2(XVII):75–115, 1918.

[10] G. Ghinita, P. Karras, P. Kalnis, and N. Mamoulis. A framework for efficient data anonymization under privacy and accuracy constraints. *ACM Transactions on Database Systems (TODS)*, 34(2):9, 2009.

[11] J. Hirshleifer. Privacy: Its origin, function, and future. *The Journal of Legal Studies*, 9(4):649–664, 1980.

[12] N. Li, T. Li, and S. Venkatasubramanian. t-closeness: Privacy beyond k-anonymity and l-diversity. In *Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference on*, pages 106–115. IEEE, 2007.

[13] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam. ℓ-diversity: Privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 1(1), Mar. 2007.

[14] K. Marriott and P. J. Stuckey. *Programming with Constraints. An Introduction.* The MIT Press, 1998.

[15] I. Neamatullah, M. M. Douglass, L.-w. H. Lehman, A. Reisner, M. Villarroel, W. J. Long, P. Szolovits, G. B. Moody, R. G. Mark, and G. D. Clifford. Automated de-identification of free-text medical records. *BMC Medical Informatics and Decision Making*, 8(1):32, 2008.

[16] N. Nethercote, P. J. Stuckey, R. Becket, S. Brand, G. J. Duck, and G. Tack. MiniZinc: Towards a standard CP modelling language. In C. Bessière, editor, *CP 2007*, volume 4741 of *LNCS*, pages 529–543. Springer, 2007.

[17] P. Norvig. Techniques for automatic memoization with applications to context-free parsing. *Comput. Linguist.*, 17(1):91–98, Mar. 1991.

[18] P. Samarati. Protecting respondents identities in microdata release. *IEEE transactions on Knowledge and Data Engineering*, 13(6):1010–1027, 2001.

[19] L. Sweeney. K-anonymity: A model for protecting privacy. *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.*, 10(5):557–570, Oct. 2002.

[20] T. M. Truta, F. Fotouhi, and D. Barth-Jones. Global disclosure risk measures and k-anonymity property for microdata. In *International Conference on Theory and Applications of Mathematics and Informatics ICTAMI 2005*.

[21] C. Valdovinos, F. J. Penedo, C. R. Isasi, M. Jung, R. C. Kaplan, R. Espinoza Giacinto, P. Gonzalez, V. L. Malcarne, K. Perreira, H. Salgado, M. A. Simon, L. M. Wruck, and H. A. Greenlee. Perceived discrimination and cancer screening behaviors in us hispanics: the hispanic community health study/study of latinos sociocultural ancillary study. *Cancer Causes & Control*, 27(1):27–37, 2016.

[22] L. Willenborg and T. de Waal. *Statistical Disclosure Control in Practice*. LNCIS Series. Springer, 1996.

[23] J. Wilson and G. Jungner. *The Principles and Practice of Screening for Disease*. World Health Organization, 1966.

Author Contributions

Regarding our paper "*Anticipating Anonymity in Screening Program Databases*" submitted to IJMI.

We claim that all authors should have made substantial contributions to all of the following:

(1) the conception and design of the study, or acquisition of data, or analysis and interpretation of data,

(2) drafting the article or revising it critically for important intellectual content,

(3) final approval of the version to be submitted.

Signed by all authors as follows:

R. Caballero

15 January 2017

Rafael Caballero

Sagar Sen

16 January 2017

Sagar Sen

January 16th, 2017

Jan F Nygård

***Conflict of Interest Form**

AUTHOR DECLARATION

Regarding our paper "*Anticipating Anonymity in Screening Program Databases*" submitted to IJMI.

We wish to draw the attention of the Editor to the following facts which may be considered as potential conflicts of interest and to significant financial contributions to this work.

We wish to confirm that there are no known conflicts of interest associated with this publication and there has been no significant financial support for this work that could have influenced its outcome.

We confirm that the manuscript has been read and approved by all namedauthors and that there are no other persons who satisfied the criteria for authorship but are not listed.

We further confirm that the order of authors listed in the manuscript has been approved by all of us.

We confirm that we have given due consideration to the protection of intellectual property associated with this work and that there are no impediments to publication, including the timing of publication, with respect to intellectual property. In so doing we confirm that we have followed the regulations of our institutions concerning intellectual property.

We understand that the Corresponding Author is the sole contact for the Editorial process (including Editorial Manager and direct communications with the office). He/she is responsible for communicating with the other authors about progress, submissions of revisions and final approval of proofs. We confirm that we have provided a current, correct email address which is accessible by the Corresponding Author and which has been configured to accept email from rafacr@ucm.es

Signed by all authors as follows:

R. Caballero
15 January 2017
Rafael Caballero

Sagar Sen
16 January, 2017
Sagar Sen

January 16th, 2017
Jan F Nygård

Anticipating Anonymity Risks in Screening Program Databases

Summary points

- Medical screening programs are a common practice and it is interesting to make available the anonymized results for future research
- The invitation data (screening centre, appointment date) combined with other personal quasi-identifiers can pose a risk of disclosing the identity of the individuals
- We propose to consider anonymity during the generation of the invitations to the screening program. Individuals with similar characteristics will share screening centres and appointment dates in order to difficult the identification even knowing the individual characteristics and the invitation data
- Two algorithms that produce a set of screening appointments that aim to increase anonymity of the resulting dataset are presented
- The level of anonymity is measured using the new concept of generalized k-anonymity

Anticipating Anonymity Risks in Screening Program Databases

Highlights

- A technique for increasing anonymity in screening program databases is proposed.
- The invitation data (screening centre, appointment date) combined with other personal quasi-identifiers can pose a risk of disclosing the identity of the individuals
- Two algorithms that produce a set of screening appointments that aim to increase anonymity of the resulting dataset are presented
- The level of anonymity is measured using the new concept of generalized k-anonymity

Changes on the paper “Anticipating Anonymity in Screening Program Databases” by Rafael Caballero, Sagar Sen, and Jan F Nygrd

Dear IJMI Editors,

This submission is a revised version of the submission *IJMI-D-17-00036*. Thanks to the reviewer comments, which we believe that have greatly improved the quality of the paper. In the following, the reviewer comments will be colored in blue to be distinguished from ours.

Best regards,

Rafael, Sagar, and Jan

Related work. Please add a section to discuss related work. Explain if there exists or not any other approach to anonymization using constraint programming. Is there anything similar to the concept of general kanonymity and anonymity vectors? Is the problem you are solving similar to any other applications, where anonymization is builtin the creation process of the dataset if yes, how is the anonymization incorporated? In particular, please see the paper:
”Global Disclosure Risk Measures and KAnonymity Property for Microdata”, Truta, Fotouhi, BarthJones, ICTAMI 2005.
In this work, there is a description of microaggregation for masking microdata; the concept of equivalence classes used in that presentation seems to bear some resemblance to your proposed anonymity vectors.

We have created a new section of related work as suggested. We didn’t know about the Truta, Fotouhi and Barth-Jones paper, thank you. With respect to this work now the paper says

> *. . . our concept of generalized anonymity vectors also considers the rest of the individuals in the dataset, and thus can be considered a* global disclosure risk measure*, in the line of early works such as [20]. In this work, the authors propose the use of equivalence classes or clusters, which is related to our notion of subsets which the same anonymity level. The main difference between our proposal and [20] is that from this idea we define the anonymity vectors in order to compare the anonymity of different appointment schedules, while [20] proposes a different notion (*disclosure risk values*) with the purpose of comparing the effect of data aggregation.*

Describe the properties of anonymity vectors before presenting the algorithms. I think this will help the reader understand the algorithms easier. For example, to understand the working of the two algorithms, I had to figure out the fact that each new lower a_{ij} value in the appointment matrix induces a new anonymity level of the dataset. This can indeed be deduced from Property P1 in fact, I think it should be explained that algorithm 1 tries to pick a_{ij}s that are as high as possible, to avoid the kanonymity level of the solution be low. So, if you delay presenting the anonymity vectors properties until after the algorithms, the reader will have figured them out by that time.

Modified as suggested. Now the properties of anonymity vectors are in Section 5, preceding the presentation of the two algorithms (Sections 6 and 7).

With respect to the comment about the election of the a_{ij}, we have introduced the following paragraph (page 19) when explaining Algorithm 1:

In order to obtain the appointment with the best generalized k-anonymity we need to find the appointment with the smaller anonymity vector (see Definition 3). This means that we must choose values a_{ij} as high as possible in order to increase the anonymity. Algorithm 1 shows a technique for obtaining the optimal anonymity vector iteratively.

In section 7, when presenting the datasets:

- Last paragraph on page 25 states that Q was chosen to be N/100 + 5, which for N = 1000 means there were 15 different QI values.

- Then, on the next page, same paragraph, you state that ”for N/10 with N = 1000, that is 100 different quasiid values, the improvement of the heuristic approach over the random assignment drops from 92% to just about the 35%”

Please clarify how is Q set (N/100, or N/10)? Or did you use multiple settings for Q?

The text is confusing, we are sorry. We have rephrased it and now says:

Number of quasi-id values: $Q = N/100 + 5$. This formula has been chosen because it approximates quite well the real data of several possible screening scenarios analyzed. Small changes in this formula have no effect in the results of the experiments. However, it is worth mentioning that other possible proportions between N and Q can yield very different results. For instance, if we choose $Q = N/10 + 5$ and $N = 1000$, then the improvement of the heuristic approach over the random assignment drops from 92% (Table 10) to just about the 35%. This happens because in this case there is a large number of quasi-id values ($Q = 1000/10 + 5 = 105$). Thus, the anonymity base is already very low, and the appointment algorithm has little space for improvement.

That is, Table 10 uses $Q = N/100 + 5$. However we advise that changing this formula has a significant impact in the Table results. We hope now it is better explained.

Another suggestion for increasing the clarity of the presentation: for algorithm 1, show how the appointment matrix evolves in four steps, for the working example that you are using in the paper. Currently you only present the final matrix obtained at the completion of the algorithm.

Done. Now Table 8 shows the four iterations, including the matrix a, the partial anonymity vector *anom* and the anonymity vector associated to matrix a.

Also show the appointment solution that algorithm 2 creates for the same working example.

Done. At the end of Section 7, now we show the appointment matrix a and its associated anonymity vector. We have also included a brief discussion about the anonymity of this solution compared to the optimal solution of the previous section and also with respect the total number of possible appointment schedules.

Typos and other corrections suggested in the edited (attached) pdf.

Thank you, we have corrected the typos and addressed the suggestion in the edited pdf. Here we comment the main suggestions (those enclosed in a red square). The page numbers correspond to the file "IJMI-D-17-00036_Marked.pdf".

Page 5, text: *The properties of anonymity vectors, the support of generalized k-anonymity, are discussed in Section 6.*
Reviewer's comment: what do you mean by support?

The sentence is confusing. Rephrased as: *The properties of anonymity vectors, the basic notion behind* generalized k-anonymity*, are discussed in Section 5*

Page 13, text: *In this small example, it would be easy to find an alternative appointment assignment which yields k=2, although of the all the possible assignments more than the 99.94% correspond to a level $k = 1$.*
Reviewer's comment: proof?

We have no proof. The data have been obtained using a program (in fact the model of algorithm 1, transformed from an optimization problem to a satisfaction problem to find all the possible anonymity vectors). We have change the text to clarify this point:

In this small example, it is possible to use a simple program to show that there are 14504 possible appointment schedules satisfying the resources maximum capacity. Of these, 14446 (more than the 99 percent) correspond to $k = 1$, while 58 possible appointment schedules verify k-anonymity with $k \geq 1$.

Page 23, text: *After some experiments, we have found that finding the resource capacity closest to value q/2 gives the best results.*
Reviewer's comment: this contradicts the choice of j stated on the else branch in algorithm 2

There was an error in the algorithm. Fixed in the first sentence of the *else* branch of Algorithm 2.

Page 23, text: *Recursive formula for p'*
Reviewer's comment: reference for formula?

We have deduced the formula ourselves and proved it by induction. The Wikipedia includes another but different formula. The first version of the formula were posted by us in this message.

Page 23, formula defining index
Reviewer's comment: explain (N) reflects a set on N 1−anonymous = de−identifiable individuals.

Done.

Page 24, Text:Table 9 shows the results of our experiments
Reviewer's comment: Describe dataset before reporting results .

Done, now the experiment conditions is explained before describing the table columns.

LaTeX Source Files

Click here to download LaTeX Source Files: ijmi.tex

Figure

Click here to download high resolution image

Figure

Click here to download high resolution image