Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/cose

Computers & Security

Decision support for releasing anonymised data

Magnus Jändel

Swedish Defence Research Agency, Sweden

ARTICLE INFO

Article history:
Received 10 February 2014
Received in revised form
3 June 2014
Accepted 6 July 2014
Available online 11 July 2014

Keywords:
Anonymisation
De-anonymisation
Privacy-preserving data mining
Privacy-preserving data publishing
Decision support

ABSTRACT

For legal and privacy reasons it is often prescribed that data bases containing sensitive personal data can be published only in anonymised form. History shows, however, that the privacy of anonymised data in many cases is easily broken by de-anonymisation attacks. This paper defines guiding principles for decisions about releasing anonymised data and provides a simple process for analysing de-anonymisation risk and for making decisions about publishing anonymised personal data. At the heart of this process is an information-theoretic de-anonymisation feasibility limit that is independent of the details of both the anonymisation procedure and the adversarial de-anonymisation algorithms. This feasibility limit relates the adversarial mutual information of the anonymised data and the attacker's background information to the number of records in the anonymised data base and the acceptable risk of privacy violations. Based on this result, we explain, discuss and exemplify the process for making decisions about releasing anonymised data.

1. Introduction

1.1. *The data anonymisation issue*

Data mining in the vast repositories of digital information that we have today and that grow at an accelerating pace can be of great value for business, research, government and law enforcement but is also a very significant threat to privacy. The wholesale collection of personal data by companies and governments combined with "big data" technologies aggravate the privacy hazard. Decision-, law- and policy makers are therefore faced with complex issues where the utility of making information available has to be balanced against legitimate legal, ethical and privacy concerns. An established solution is to publish anonymised data.

Anonymisation means that the data is processed before publishing for the purpose of hiding the true identity of the people that are described by the data. De-identification, which means that explicit identifiers such as full names and social security numbers are removed, is a necessary but typically insufficient part of the anonymisation process. Proper anonymisation means that the data is filtered such that the risk of a competent adversary identifying published information about a targeted individual is sufficiently small. De-anonymisation attacks match things that the adversary knows about the target to the published information which may enable the attacker to pinpoint sensitive information about a target person in the anonymised database. An employer could for example find out about the health status of an employee from public anonymised health care data if the salary, age, profession and zip code are included in the published health records.

Note that the data anonymisation issue is a part of a wider problem complex concerning how to avoid unwanted use of data. On many occasions, it is not only the privacy of

E-mail addresses: magnus.jaendel@foi.se, magjan@foi.se, magjan76@yahoo.se.

http://dx.doi.org/10.1016/j.cose.2014.07.001

individuals that needs to be protected. It might be important to avoid revealing statistical measures about for example ethnic groups, employee categories or groups of people with a certain medical diagnosis. Just as for anonymised personal data, it is possible to infer sensitive statistical measures by combining background knowledge with superficially uncontroversial statistical information. Inference control methods, as pioneered by Denning and Schlörer (1983), facilitate decisions about what statistical measures and marginal distributions that can be released without unduly providing opportunities for inferring sensitive information. In this paper we will, however, focus on the narrow interpretation of anonymisation as measures to protect the privacy of individuals in the context of wholesale publishing of anonymised databases thus avoiding some of the complexities of the wider problem considered by Denning and Schlörer (1983).

The differential privacy approach (Dwork, 2006, 2008) is based on the idea that the algorithm for responding to queries about the contents of a database is sufficiently safe from a privacy point of view if what an attacker can learn from the responses does not differ significantly if an extra record is added to the database. The privacy of each individual record owner is protected since the presence or absence of the personal information is guaranteed to make no significant difference in the information that the attacker gets access to. Although differential privacy formally is applicable to the release of anonymised databases, it is, however, from a practical point of view mainly useful for filtering of aggregated statistical quantities and falls hence within the wider scope of Denning and Schlörer's inference control problem. Further details on differential privacy and how it applies to our decision problem is provided in Section 1.4.

Two burgeoning branches of computer science, *Privacy-preserving data mining* and *Privacy-preserving data publishing* provide a cornucopia of methods for data anonymisation in the narrow sense considered in this paper. For recent reviews see Agrawal&Yu (2008a) and Fung et al. (2010). A tutorial discussion of the field is provided by Brynielsson et al. (2013). Important anonymisation methodologies are 1) deterministic editing of the data for the purpose of guaranteeing a certain level of privacy and 2) random distortion of the data aiming at providing a statistical measure of privacy. In both branches it is essential to optimize the data mining utility of the published data while fulfilling the privacy constraint.

An example of a deterministic editing method is the k-anonymity algorithm (Sweeney, 2002) which ensures that an adversary always finds at least k-1 database records that are indistinguishable from the target record in the k-anonymised database. The k or more records in one such *equivalence group* of a k-anonymized data set may, however, lack diversity in sensitive attributes so that attackers still may be able to learn sensitive data about the target. If all the records in an equivalence group of a k-anonymized health care database reveal the same disease, attackers will for example be able to conclude that target is among the afflicted. For the purpose of solving this problem, Machanavajjhala et al. (2007) introduce the l-diversity criterion which demands that at least l different values of a sensitive attribute should be represented in each k-anonymized equivalence group. Li and Li (2007) noted that an unusual distribution of sensitive attributes within a k-anonymised and l −diversified equivalence group still may reveal sensitive information. As a remedy, they propose the t-closeness criterion according to which the difference between the sensitive attribute distribution of each equivalence class and the corresponding overall distribution should not be larger than a threshold t.

Random distortion methods can for example entail adding random numbers to selected data attributes thus blurring individual attributes while keeping statistical averages sufficiently accurate (Aggarwal and Yu (2008b), Chen and Liu (2008)). Note that the concept of anonymisation considered here does not include methods for performing data mining in encrypted databases without revealing anything other than aggregated statistics or situations where a trusted party protects the original data and releases filtered responses to selected data mining queries.

Since this paper applies information theory to the problem of analysing the feasibility of anonymisation, we note that the information theoretic approach also has been used by Sankar et al. (2010) and Rebello-Mondedero, Forné and Domingo-Ferrer (2010) for the purpose of defining privacy and utility metrics.

In spite of the great variety and sophistication of the anonymisation algorithms, experience shows that anonymised data often is vulnerable to de-anonymisation attacks. Following Narayanan and Shmatikov (2008) we will consider two main adversarial scenarios.

I) **Single-target attack.** The assailant is interested in a specific target individual. The single-target attack has two important sub-cases:
   a. The attacker knows that the target is in the anonymised database.
   b. The attacker does not know whether the target is in the anonymised database.

II) **Large-scale attack.** The adversary wishes to assign the most likely identity to all the records in the anonymised database.

Note that the attacker's background information in the single-target case usually includes both data from public sources and information that the attacker has gained by other means such as personal contact with the target. The large-scale attack is typically automated and uses only digital background data including Internet and other public sources.

Narayanan and Shmatikov (2008) note that sparse databases are particularly vulnerable to de-anonymisation attacks. Records in sparse databases have many attributes and non-null values only for a small fraction of the attributes. Databases containing e.g. the purchasing history of customers that select from large product catalogues are typically sparse. Attackers knowing even a small subset of a target's purchases can often pinpoint the target record in a sparse database.

1.2. *History of de-anonymisation*

How serious is the threat of de-anonymisation in practice? There are several well-known cases in which the privacy of anonymised data has been broken. Sweeney (1997) show that subjects in an anonymised medical database can be identified by cross-correlating with a public voter database. Individuals were also recognized in anonymised search query data spanning more than half-million web users that America Online released for R&D purposes (Barbaro and Zeller, 2006). Two weeks after Netflix in 2006 published a hundred million anonymised movie rating records, researchers demonstrated efficient de-anonymisation algorithms (Narayanan and Shmatikov, 2008).

El Emam et al. (2011) review the history of de-anonymisation of health data and conclude that the average attack success rate in published studies is high (about 25%) but that the evidence is insufficient with considerable uncertainties. One of the published cases shows a much lower success rate (0.013%) than the average. The known cases are mainly not maliciously intended de-anonymisation attempts by researchers and journalists. Malicious attacks are, however, unlikely to come to public knowledge. Apocryphal stories tell for example that a banker broke the anonymity of published cancer data for the purpose of reviewing loan applications (Bartlett, 1993).

The discussion of societal issues relating to de-anonymisation includes Ohm (2010) that provides a comprehensive review from a U.S. perspective of how privacy legislation often is based on the assumption of that anonymisation is feasible and effective. Ohm emphasizes that the ease and efficiency of de-anonymisation disrupt the intentions of lawmakers and data publishers. Rothstein (2010) elucidates problems related to using anonymisation as the legal ground for publishing health data. McGuire and Gibbs (2006) discuss the problem of protecting the privacy of genetic information given that the patient behind published DNA data readily can be identified with some background genotype data (Lin et al., 2004). Narayanan and Shmatikov (2010) discuss the societal impact of fallacious assumptions about the security provided by anonymisation and suggest that the release-and-forget approach to publishing anonymised data should be replaced by audited query interfaces.

1.3. *The decision-making problem*

The objective of this paper is to provide a simple but quantitative tool for assisting decision makers in the task of deciding whether an anonymised database can be published. The situation that we have in mind is that experts have applied anonymisation algorithms to a confidential database and produced an anonymised version of the data base. The decision makers are asked to approve the anonymised data for release either to the general public or to some other user society. We assume that the decision makers have a genuine interest in publishing the anonymised data but wish to apply a policy of caution according to the following principles.

A. Broken privacy means that an assailant correctly identifies the record in the anonymised database that is associated with a target individual.
B. The risk of privacy breach must be very small but cannot be zero.
C. All record owners have equal right to privacy.
D. Adversaries are assumed to be determined, resourceful and technically competent.
E. The de-anonymisation algorithms that attackers will use are unknown.

Principle A means that the decision maker recognizes that it is hard to define precisely what content in the anonymised database that might be sensitive for the record owners. Sensitivity depends on the personal circumstances of the target and on the relation between the target and the attacker. Most of us would not be overly worried if data about our hotel bookings is published. A stalker that knows about the possible romantic relation between the target and married person living in Portsmouth would, however, be very interested in knowing about hotel reservations in the seaside English city. Caution requires that attackers should not be allowed to identify the target's record in the database. Note that the definition of privacy breach in principle A is somewhat narrow and does not cover some situations that could be considered as privacy intrusions such as for example when the anonymised database is about persons who have filed tax returns and the attacker uncovers that the target is not included.

The decision makers understand that privacy cannot be guaranteed with absolute certainty. There is always some finite risk of privacy breach in anonymised data but this risk must, as noted in principle B, be very small. To quantify this is an important task for the decision makers.

Depending on the nature of the data, some targets may be protected by the ordinariness of their information. If the criminal record of the target is totally clean, it will be impossible to identify the specific record of the target among the large cohort of similarly blameless people. We assume, however, according to principle C that the decision maker is bound to protect the privacy of all subjects in the database including those with highly specific and unique profiles.

Even perfunctory anonymisation offers some protection against casual browsing and incompetently performed attacks. Principle D means that we have to assume that the attacker is committed and well-versed in the technology of de-anonymisation and principle E emphasizes that attackers may be more competent and creative than the technical advisors of the decision maker, as often has been the case in the historical examples reviewed in Section 1.2.

This policy focuses on a targeted attack against the privacy of a specific person. A large-scale attack that strives to identify as many record owners as possibly will usually only employ background knowledge that is easily available in computer readable form. A determined targeted attack will also use computer readable information and may furthermore include information from other types of sources such as physical surveillance, local community gossip and conversations with the target. Hence it is possible that a targeted attack rallies more information against the selected target than would have been used against the same target in a large-scale attack. Targeted attacks are therefore often more difficult to protect against compared to large-scale attacks. Depending on the situation, attackers may or may not know that the target is in the published database. Furthermore, we recognize that decision makers need simple but quantitative tools. Simple, because it is unlikely that a decision maker is a computer scientist and a privacy-preserving data publishing expert – the role of the decision maker is rather to evaluate the output of such experts; quantitative, because the decision is about probabilities. Hence, we will in Section 2 derive a formal de-anonymisation feasibility limit and in Section 3 apply this result to defining, discussing and exemplifying a decision process for releasing anonymised data.

Note that the database inference attacks considered by Denning and Schlörer (1983) have a much wider scope than the one considered here. Denning and Schlörer consider not only publishing anonymised individual records but also the release of all other kinds of derived statistical measures and partially summed tables based on the original database. This paper is furthermore limited to decisions about wholesale publishing while Denning and Schlörer in addition consider graded access controls including auditing a stream of queries and filtering the response based on the query and response history. The reason for the narrow scope of the present paper is that it, as we shall see, enables the simple quantitative result that we are looking for. Moreover, the narrow problem addressed in this paper corresponds to a common decision-making challenge since anonymisation requirements frequently refer to individuals and data-owning organisations often lack motivation, competence and resources for continuous long-term commitments such as query and response auditing.

1.4. Known de-anonymisation feasibility limits

This subsection reviews briefly results from the literature on privacy-preserving data mining and data publishing that discusses generic limitations on the feasibility of de-anonymisation in the context of the decision problem of Section 1.3.

Differential privacy offers the strongest guarantee of privacy that we have found in the literature since it is independent of the resources and knowledge of the attacker (Dwork, 2006, 2008, 2011). Quoting from Dwork (2008): "We say databases D_1 and D_2 differ in at most one element if one is a proper subset of the other and the larger database contains just one additional row. A randomized function K gives ε-differential privacy if for all data sets D_1 and D_2 differing in at most one element and all $S \subseteq \text{Range}(K)$,

$$\Pr[K(D_1)\in S] \le e^{\varepsilon} \times \Pr[K(D_2)\in S]. \tag{1}$$

The probability is taken over the coin tosses of K." Differential privacy guarantees hence that the addition of a single record to the confidential database $\mathrm{DB_S}$ will not be detectable in any analysis based on the released information $K(\mathrm{DB_S})$. In our case K is the anonymisation process and $K(\mathrm{DB_S})$ is the anonymised database.

To demonstrate that the release of an anonymised database does not satisfy ε-differential privacy for any database and any ε we set $D_1 = \mathrm{DB_S}$ in Eq. (1)while D_2 equals $\mathrm{DB_S}$ with a single target record removed. The release process K produces an anonymised version of the database in which each anonymised record is a filtered version of the corresponding record in the input database. If D_1 has n rows and m columns of attributes, D_2 will have $n-1$ rows and m columns while the output anonymised databases has the same number of rows as the input database. Select the output set S of Eq. (1) to span all possible output matrices with n rows so that $Pr[K(D_1) \in S] = 1$. However, $K(D_2) \notin S$ and hence$Pr[K(D_1) \in S] > e^{\varepsilon} \times Pr[K(D_2) \in S]$ for any positive ε thereby violating Eq. (1). Since Eq. (1) must hold for all D_1, D_2 and S, we find that the release of anonymised databases always infringes differential privacy. Dwork (2011) notes that differential privacy rules out *direct viewing of raw data*. As we have showed, direct viewing of anonymised data is also ruled out by differential privacy. It appears that if we want to release an anonymised database we have to forego the strong guarantees of differential privacy and be prepared to handle a situation where the level of privacy depends on the resources and knowledge of potential adversaries. Any weaker feasibility limits that we may consider will hence have to include assumptions about the attack scenario.

Narayanan and Shmatikov (2008) define a threshold for successful de-anonymisation based on the number of attributes m of the target that are known to an adversary and that are sufficiently similar to the corresponding attributes in the anonymised database finding that de-anonymisation is feasible if,

$$m \ge \frac{\log N - \log \varepsilon}{-\log(1-\delta)} \tag{2}$$

where N is the number of published records, ε is a measure of the maximum error in the attributes known to the adversary compared to the corresponding published attributes and δ is a measure of the precision that is required in comparing records. The attributes can be of any kind including binary, numeric, alphanumeric and tuples. We quote this equation, without restating the fairly complex complete definition of the parameters ε and δ, because it will be of interest to compare its form, but not its detailed parameters, to the main result of this paper. Note that the limit in Eq. (2) is restricted to the specific de-anonymisation algorithm employed by Narayanan and Shmatikov. The same authors find that 5–10 background attributes are needed for reliably re-identifying users in an anonymised database of 500 000 Netflix subscribers. For further analysis of the Narayanan&Shmatikov algorithm see Datta et al. (2012).

Aggarwal (2008) explains lucidly why it is difficult to anonymise high-dimensional data noting that both the k-anonymity and the randomization approach must degrade the utility of the data substantially in all cases in which we do not know in advance what background knowledge an adversary may have access to. This is called "*the curse of dimensionality*" in the privacy-preserving data mining literature (note that this phrase has different meanings in different fields of research). Aggarwal's conceptual analysis is strengthened by detailed mathematical investigations of the high-dimensional behaviour of several specific models of k-anonymity and randomization. The conclusions of Aggarwal can be summarized as "*… privacy preservation by anonymisation becomes impractical in very high-dimensional cases, since it leads to an unacceptable level of information loss*" and "*The results seem to suggest that the curse of dimensionality may be a fundamental one from the point of view of privacy and cannot be easily solved using more effective algorithms and techniques*". Aggarwal notes, however, that it may be possible to exploit some special benign structures even in high-dimensional data sets.

2. Generic de-anonymisation feasibility limit

Section 2.1 defines a data and adversarial model for which an information-theoretic de-anonymisation feasibility limit is derived in Section 2.2. In Section 2.3 we argue that this limit is generally applicable as a worst-case estimate of de-anonymisation feasibility.

2.1. *Attack procedure and data model*

A database owner wishes to publish an anonymised version DB_A of a secret database DB_S. Both DB_A and DB_S contain N records each of which is associated with a unique individual – the record owner. The anonymised database is produced by applying an anonymisation algorithm to DB_S thereby removing any explicit identifiers and furthermore processing the data for the purpose of protecting the privacy of the record owners against adversarial attacks. Many different anonymisation algorithms are known (see the brief review in Section 1.1) but the conclusions of this section are agnostic with respect to the type of anonymisation algorithm that is employed for producing DB_A.

An adversary knows the real-world identity of a target person and has some further background information Y_t about the target. In addition to this target-specific background information, the attacker has general knowledge KW about the world that may include information in any form including digital data repositories, human memory and various anonymised databases including in particular DB_A which is assumed to be published in a context where it is accessible for the attacker. The attacker may or may not know whether the target is included in DB_A but the target is in fact a record owner in DB_A; in Section 1.3 we defined broken privacy as meaning that the attacker correctly identifies the anonymised record of the target.

Some information may overlap between the adversary's specific background information Y_t and the target record X_t in DB_A. The *pseudo-identifier* Q_t of the target t represents the maximum amount of information about the target that is shared between X_t and Y_t, given the adversary's general knowledge KW. The pseudo-identifier may help the attacker to pinpoint the target record in DB_A as discussed in the following.

As a running example we will consider a database consisting of anonymised responses to an employee satisfaction survey. Each entry consists of some facts about the employee followed by the employee's opinions about the company and its management. The facts include the number of children, age group in five year intervals and years of employment. Imagine that the disgruntled employee Alice in the survey has entered that she has four children, age in the 50–55 interval and three years of employment followed by some rather scathing comments about the competence and credibility of the management. Her boss, Bob, reads Alice's diatribe and feels a strong urge to identify and punish the disloyal underling. Bob has records on the age and employment date of everyone in the company and knows by hearsay if employees have children or not. Bob suspects that the somewhat defiant Alice may be the writer of the infuriating comments and compiles hence the pseudo-identifier of Alice which amounts to: *have children, age in the 50-55 interval and three years of employment*. Note that this neither is the exact information in the anonymised database nor is the precise information held by Bob, but rather is the overlapping information relating to Alice of these sources.

Since we are interested in worst-case situations with maximally competent and resourceful attackers, we shall assume that the adversary, for any target, is able to extract the pseudo-identifier from X_t and Y_t. Hence we assume the existence and availability of methods G and F such that,

$$Q_t = G(X_t) = F(Y_t). \tag{3}$$

These functions will, if applied to target record data and adversarial target background data respectively, by definition give the same output, namely the pseudo-identifier. If there is no overlapping information in X_t and Y_t, both functions will return a null value. Pseudo-identifiers are in the privacy-preserving data mining literature often defined as the target record attributes that are known to the attacker. This definition is subsumed by our more general concept which also includes situations in which the attacker's knowledge is uncertain and perhaps in an unstructured format such as free text or human memory.

When Bob, in our running example, compiled Alice's pseudo-identifier, it required not just mechanical matching of records in different databases. The precise age had to be compared to an age interval. The years of employment had to be computed from the employment date. Furthermore, Bob had to recall what Alice had told him about her offspring. The process for computing the pseudo-identifier required hence both semantic matching of records in different formats and comparing digital data to the attacker's memories. Once Bob has figured out how to do this for Alice, it can be formulated as a generic process that can be applied to any target. In our abstract model, this process corresponds to applying the functions G and F to the target data.

The assailant wishes to identify the DB_A record that belongs to the target individual and applies therefore the following procedure.

2.1.1. *Attack procedure*

1) Compile background information Y_t about the target.
2) Extract the pseudo-identifier $Q_t = F(Y_t)$ of the target.
3) Search DB_A for candidate records X_i such that$Q_t = G(X_i)$.
4) For each candidate record estimate the probability that it is the target record.

Steps 1–4 are in the following called *the attack procedure*. If the target record is in DB_A, it will be among the set of candidate records that are found. Assume now that the adversary knows that the target is in the anonymised database. If only one candidate is found in step 3, privacy is broken in the worst possible way since the attacker with certainty has identified the target's record in DB_A. If k candidate records are found and no other relevant statistical information is available, the attacker will be able to identify the target with probability $1/k$.

The risk analysis is more complex if the attacker can use generic background knowledge for reasoning about identification probabilities. Another complication is that the adversary may not know whether the target is in DB_A. These issues will be discussed in Section 2.3.

In our running example, Bob already has Alice's pseudo-identifier and can now apply step 3 of the attack procedure in which he computes the pseudo-identifier of all records in the anonymised survey. Any record that matches Alice's pseudo-identifier *have children, age in the 50–55 interval and three years of employment* is added to the list of candidate records. Eventually, it turns out that there is only one candidate so step 4 is easy. Bob initiates an impromptu performance review of Alice.

2.1.2. *Data model*

For the purpose of deriving a generic risk analysis process for de-anonymisation attacks, we need an abstract model of the data that is used in the attack procedure. Hence, we will now describe a model for how to generate the records of DB_A and the corresponding adversarial information. We are not claiming that such probabilistic generative models always exist or are available for decision makers. The data model is just a scaffold that in the following subsection will be used for deriving the feasibility limit. In Section 2.3 we discuss the applicability of the feasibility limit and application examples are discussed in Section 3.

In the generative data model, a record i of DB_A is represented by a random variable X_i and the corresponding adversarial background data by a random variable Y_i. Variables X_i and Y_i are correlated according to the model in Fig. 1 where V_i, Q_i and W_i are mutually independent random variables and the pseudo-identifier Q_i represents the information that is common to X_i and Y_i .

According to the data model of Fig. 1, X_i is computable from V_i and Q_i by some deterministic function,

$$X_i = K(V_i, Q_i). \tag{4}$$

Similarly Y_i can be computed from W_i and Q_i by some other deterministic function

$$Y_i = M(W_i, Q_i). \tag{5}$$

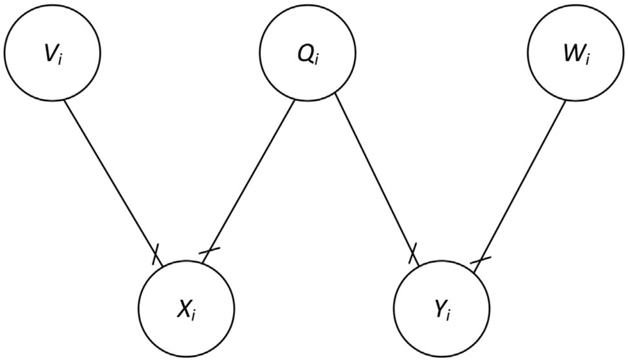

Fig. 1 – Generative model for the random variables X_i and Y_i. The connectors indicate deterministic dependencies to the random variables V_i, Q_i, and W_i according to Eqs. (4) and (5).

All equations that are indexed with a single index i are in the following tacitly assumed to hold for all$1 \le i \le N$. Values of random variables X_i and Y_i are, according to the data model, generated by first drawing values of V_i, Q_i, and W_i. The values of X_i and Y_i are then computed according to Eqs. (4) and (5) respectively. According to the attack procedure above, the pseudo-identifier Q_i is retrieved from X_i and Y_i by applying functions G and F respectively. Note that variables in Fig. 1 that belong to different record owners may be correlated; e.g. X_i may be correlated with X_j for $i \neq j$. Such inter-record dependencies could be caused by real-world relations, for example family connections, or by anonymisation processing, for example k-anonymity algorithms. Our data model does not describe intra-record correlations in detail since this is not required for deriving the feasibility limit. Section 2.3 discusses and exemplifies how the feasibility limit applies to situations with intra-record dependencies. Note that Sankar et al. (2010) use as similar data modelling approach for the purpose of information theoretic analysis of application independent utility and privacy metrics although their model assumes independent records.

In our running example, X_i is the employee's response in the survey and Y_i is what the manager knows about the employee. Clearly, it is almost always impossible to find a probabilistic generative model for such information sources. Nevertheless, we shall find that the feasibility limit, which in the next section is derived from the data model, often is useful for analysing real-world problems.

2.2. *Feasibility limit*

The variable Q_i of the data model in Section 2.1 captures all correlations between X_i and Y_i and is hence the optimal pseudo-identifier from an attacker's point of view. We can express this by,

$$\begin{aligned} P(X_i|Y_i) &= P(X_i|Q_i) \\ P(Y_i|X_i) &= P(Y_i|Q_i), \end{aligned} \tag{6}$$

where the symbol P indicates a conditional probability distribution. In information theoretic representation, the dependency relations of or model can be summarized as

$$I(X_i, Y_i) = H(Q_i), \tag{7}$$

where $I(X_i,Y_i)$ is the mutual information of X_i and Y_i whereas $H(Q_i)$ is the Shannon entropy of Q_i. It is straightforward to show that Eq. (7) follows from the definition of mutual information applied to our generative data model. The entropy of Q_i is measured in information bits and can be thought of as the length of the shortest possible binary code that describes Q_i. Such minimal length codes can be produced by applying optimal compression algorithms to data.

Eq. (7) can be specialized to a situation where all probability distributions and information entropies are conditioned on the generic background knowledge KW of the attacker,

$$I(X_i, Y_i|KW) = H(Q_i|KW). \tag{8}$$

Both the *adversarial mutual information* on the left side and the entropy on the right side of Eq. (8) are hence conditioned on all the knowledge of the attacker that is not specifically attributable to the target and thus included in Y_i. The proof is again readily found from the definition of the mutual information and our data model. In the compression analogy, $H(Q_i|KW)$ is understood as the bit length of the optimally compressed code that is achieved by applying the knowledge KW in the compression process.

In computing the Shannon entropy of a pseudo-identifier such as *parenthood, age interval and years of employment* the outcome will clearly depend on what statistical information that is available for estimating the underlying probability distributions. Bob can use company data about the actual cohort of employees while an attacker without access to company records would have to do with for example national census data and some guesswork. In Eq. (8) we make this dependence on the attacker's knowledge explicit by conditioning the information-theoretic quantities on KW.

The information-theoretic considerations that are summarized in Eq. (8) will now be used for analysing the feasibility of de-anonymisation attacks. Consider an adversary that extracts the pseudo-identifier Q_t according to the attack procedure of Section 2.1 and then uses all available generic background knowledge to compute an ideally compressed code for Q_t. The length n of this code is according to Eq. (8)

$$n = I(X_t, Y_t|KW). \tag{9}$$

Suppose that the attacker knows that the target belongs to some encompassing group of people $\hat{S}$ with $\hat{N}$ members and that the population S of DB_A is included in $\hat{S}$ so that $S \subseteq \hat{S}$ and $N \leq \hat{N}$. Astute attackers will select the smallest group $\hat{S}$ for which they have good background knowledge. What this means will become clear in Section 3 where examples are provided. If the adversary knows that the target is in the published database, $S = \hat{S}$ and$\hat{N} = N$. Otherwise, some larger encompassing group has to be selected. It is always possible to find an encompassing group since the attacker at worst could make $\hat{S}$ equal to the entire population of the world.

The target's anonymity is broken if step 3 in the attack process finds the target's record as the sole candidate and the attacker in step 4 reasonably can conclude that the targets pseudo-identifier with high probability is unique among all members of $\hat{S}$. The anonymised database may not span all of $\hat{S}$ so the attacker may not have access to the actual values of Q_i for all members of$\hat{S}$. The adversary can, however, apply generic knowledge to reason about the statistical behaviour of pseudo-identifiers in the encompassing group. If the target is known to be in the DB_A, the attacker has in step 3 of the attack procedure compiled a table of Q_i for all of $\hat{S}$ and can in step 4 use this to identify the target.

Suppose that, in our running example, for privacy reasons only 75% of the anonymised responses are provided to the management. The employees whose responses are included correspond to the set S above. Bob is aware of Alice's brazen defiance and wants to find her response in the anonymised data if it should happen to be included. Bob can still perform steps 1–3 of the attack procedure in Section 2.1 as before. In Step 4, Bob implicitly selects $\hat{S}$ to be all employees in the company by arguing that if Alice's pseudo-identifier is unique in$\hat{S}$ and matches a single response in the anonymised survey, that response must be hers. Note that Bob might not need access to precise personal details about all employees. Aggregated workforce statistics could suffice for concluding that there is just one single parent in the 50–55 age span with three years of employment.

What is the probability that Q_t in fact is unique within the encompassing group? This is easy to analyse using the random-looking bit strings that are the ideally compressed representation of the pseudo-identifiers. We find that the probability p that Q_t with entropy n is unique among the $\hat{N}$ pseudo-identifier samples of $\hat{S}$ is given by,

$$p = \left(1 - 2^{-n}\right)^{\hat{N}-1}. \tag{10}$$

This is the probability that none of the other $\hat{N} - 1$ pseudo-identifiers have the same bit string representation as Q_t. From Eq. (10) and the assumption $2^{-n} << 1$, we extract the entropy n_p that corresponds to a given *de-anonymisation probability* p,

$$n_p \approx \log_2\left(\hat{N}\right) - \log_2\left(-\ln\left(p\right)\right). \tag{11}$$

Combining Eqs. (9) and (11) we now define an information-theoretic limit for de-anonymisation feasibility. The probability that an optimal de-anonymisation attack against target t is successful is less than p only if the adversarial mutual information $I(X_t, Y_t|KW)$ is less than the critical entropy n_p of Eq. (11) according to,

$$I\left(X_t, Y_t\middle|KW\right) < n_p \approx \log_2\left(\hat{N}\right) - \log_2\left(-\ln\left(p\right)\right). \tag{12}$$

Eq. (12) provides hence an upper limit to the amount of information that may overlap between the target record X_t and the adversarial background knowledge Y_t given that the de-anonymisation risk is as most p. The quantity that is limited by Eq. (12) is the adversarial mutual information which incorporates both the specific knowledge about the target and any generic background information that the attacker may possess.

2.3. *Relevance and applicability of the feasibility limit*

To get a foretaste of the impact of Eq. (12), consider for example that the anonymised database includes the world population ($7.2 \cdot 10^9$ records) and that the decision maker accepts a de-anonymisation risk of 50%. The maximum overlapping information is, according to Eq. (12), 33 bits of information. This corresponds to approximately 33 characters of text in English (red font indicates the first 33 characters of this sentence). Real databases contain typically fewer subjects than the world population and acceptable de-anonymisation risks are normally much less than 50%. Tolerable values of the adversarial mutual information $I(X_t, Y_t|KW)$ are therefore in all practical cases less than 33 bits. Given the vast and increasing archives of personal information, this means that proper anonymisation in many cases will have a significant impact on the utility of data mining and that publishing useful data while respecting the privacy of the record owners often will be impossible.

Note that this result is independent of the technical format of the data. In the following, database means any kind of information repository and record means all the information about a specific individual in the information repository, whatever the technical format. The adversarial background information could likewise be in any conceivable format including natural language and pieces of information stored only in the attacker's mind. The advantage of applying the concept of adversarial mutual information rather than counting overlapping database attributes as in (Narayanan&Shmatikov, 2008) is that all such disparate information sources are encompassed in the former more abstract concept.

In this paper we generally assume that record owners are individual persons. Depending on the context record owners could, however, be for example families, companies, military units or any other kind of group or organisation. Even though the anonymised database in such cases would not be about individuals it would still have records compiling anonymised information about the record owner, the objective of the attacker would be to identify the record of the target and Eq. (12) would then still indicate the feasibility of a successful attack.

What if the assailant doesn't know if the target is in the anonymised database? Some anonymisation methods apply for example a strategy of row-suppression in which records that are deemed to be particularly vulnerable are deleted from the database. In such cases, attackers must be uncertain about whether the target can be found the published database. Introducing the encompassing group $\widehat{S}$ in the derivation of Eq. (12) means that we can handle such situations as explained in the following example. Consider a scenario in which DB_A is a medical database of U.S. army patients and the adversary knows that the target is female and a U.S. citizen but does not know whether or not the target is included in DB_A. In this case, the attacker may select $\widehat{S}$ to be the set of female U.S. citizens. The adversary may know some useful statistical properties $\widehat{K}$ relating to $\widehat{S}$ such as for example the age distribution or blood group distribution. The adversarial generic knowledge KW consists in this example of the combination of DB_A and$\widehat{K}$. Furthermore, assume that the attacker in step 3 of the attack procedure finds one single record that matches the pseudo-identifier Q_t. Suppose that the attacker in step 4 of the attack procedure can apply statistical reasoning based on KW to show that the pseudo-identifier with high probability would be unique in a hypothetical anonymised database that is of the same type as DB_A but spans the entire encompassing group$\widehat{S}$. If so, the attacker can conclude that the identified record with high probability is the target. Using the encompassing set $\widehat{S}$ in Eq. (12) is critical for capturing this kind of reasoning.

The ensuing part of this subsection will discuss various technical issues relating to the applicability of the feasibility limit. Readers that are mainly interested in how to apply the feasibility limit can move on to Section 3.

Adding noise to W_i, Q_i, V_i, X_i or Y_i in Fig. 1 represents a variety of situations where the data is unreliable or corrupted including for example that the attacker has partially uncertain background information. Noise injection is, however, easily subsumed in the model of Fig. 1 by redefining W_i, Q_i, V_i, X_i and Y_i as needed. Introducing noise by adding new nodes in Fig. 1 can also be incorporated by collapsing sub-networks and making suitable re-definitions of the variables in Fig 1. This is best illustrated in the information diagram of Fig. 2.

The attack procedure of Section 2.1 presupposes that the attacker has functions $Q_i = G(X_i) = F(Y_i)$ that reliably recover the pseudo-identifier. It is perhaps more realistic to assume that the adversary only has access to functions that return approximations of Q_i. The attacker would when have to apply some similarity metric as for example in (Narayanan&Shmatikov, 2008). Knowing the analytic capabilities of the attacker, it would be possible to construct a tighter de-anonymisation feasibility limit than in Eq. (12) as exemplified in Eq. (2). Following condition E of Section 1.3 we assume, however, that decision makers lack such detailed insights in the de-anonymisation tools of the assailants. The feasibility limit of Eq. (12) is based on the worst-case assumption that adversaries can utilise the full potential of the background information.

By using the adversarial mutual information $I(X_t, Y_t|KW)$ in Eq. (12) where KW includes knowledge of DB_A, we handle all kinds of dependency relations between different records in the anonymised database. We can for example conclude that the adversarial mutual information is less or equal to $log_2(N)$ if the attacker knows that the target is included in DB_A. It is always possible to construct a $log_2(N)$ long code for Q_t by computing a table of Q_i for all records and use a pointer to the table entry Q_t as a code for Q_t thus demonstrating that $I(X_t, Y_t|DB_A) \leq \log_2(N)$ with equality if all Q_i table entries are different. If there are correlations within DB_A to the effect that some of the pseudo-identifiers are identical, we know that it is possible to construct compressed codes for Q_t that are shorter than $log_2(N)$.

Consider for example the k-anonymity algorithm where typically a subset of the DB_S attributes are identified as possible pseudo-identifiers and the data is processed so that the same pseudo-identifier pattern is shared by at least k records. Universal k-anonymity means that all attributes are assumed to be included in the pseudo-identifier. If universal k-anonymity is applied, assailants would invariable find that

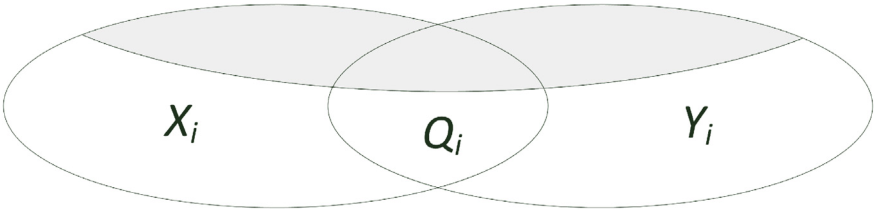

Fig. 2 – Information diagram showing the anonymised record X_i, the adversarial background information Y_i and the pseudo-identifier information Q_i. The shaded parts are noise that may be added or subtracted at will to any of the variables without changing the results of Section 2.2.

any pseudo-identifier matches at least k different records. By computing a table of X_i for all records and use a pointer to this table to code for both X_i and Q_i, it can be shown that both the entropy of X_i and the adversarial mutual information in this case are less or equal to $log_2(N/k)$, again assuming that the attacker knows that the target is in DB_A. Since the adversarial mutual information is the information overlap between X_i and Y_i, it can never be larger than the entropy of X_i itself and any further specific or generic background information of the attacker can therefore not make the adversarial mutual information larger than $log_2(N/k)$. Universal k-anonymity is, however, very restrictive with respect to the amount of information that can be published. The maximum information content of a universally k-anonymised record is equal to $log_2(N/k)$ corresponding for example to 29.42 bits for 10-anonymity and 7.2 billion records. In order to preserve the data mining utility of the published data it is therefore common to apply k-anonymity only to a smaller set of the attributes for which it is assumed that attackers have background knowledge. In this case, the adversarial mutual information may hence be larger than $log_2(N/k)$ if assailants have background information not foreseen by the designers of the anonymisation process.

The feasibility limit in Eq. (12) is consistent with other feasibility limits in the literature (see Section 1.4). The de-anonymisation threshold of Narayanan and Shmatikov (2008) (see Eq. (2)) is derived for a specific de-anonymisation algorithm but shows the same characteristic linear relationship between the logarithm of the number of records and a measure of the knowledge of the attacker. This measure is the maximum number of attributes known by the assailant and the adversarial mutual information in Eq. (2) and Eq. (12) respectively. The qualitative discussion of the curse of dimensionality (Aggarwal, 2008) presages the impact of Eq. (12). Aggarwal's in depth discussion of specific examples of k-anonymity and randomization anonymisation is consistent with and supports the more abstract and generic limit proposed in the present paper. The main difference is that the feasibility limit proposed here is intended for decision support applications rather than for analysis of specific anonymisation and de-anonymisation algorithms.

Smith (2009) shows that Shannon entropy and mutual information may not be ideal for estimating the risk of information leakage in settings where an adversary tries to guess a secret H based on the output L of a program $f(H) = L$. The problem is that measures based on the Shannon entropy may severely underestimate the risk if the probability weight of H conditioned on L is very unevenly distributed so that a first guess has a high probability of being on the spot. Smith proposes alternative risk measures based on the concept of vulnerability which is the worst-case probability that the adversary correctly guesses H in one try. Our feasibility limit in Eq. (12) is based on the Shannon entropy so it is interesting to relate our result to Smith's critique of using Shannon entropy for estimating information leakage risks. In our case, the secret to be guessed is the identity of the target based on knowledge of the anonymised database, adversarial background data and other background knowledge. This is not exactly the problem type analysed by Smith but the lesson that measures based on Shannon entropy may underestimate the risk of guessing outliers in skewed probability distributions should be taken seriously.

It is, in principle possible to replace the Shannon mutual information in Eq. (12) with some alternative measure of mutual information such as the one based on Smith's vulnerability concept. In doing so, it is crucial that the mutual information measure is equal to the bit-length of a code that could serve as an index in a table of pseudo-identifiers (see Eq. (9)). This is indeed the case both for Shannon entropy and for Smith's alternative min-entropy; the latter is related to the shortest possible code length. In the decision process of Section 3.2 we apply, however, the Shannon mutual information for the following reasons. Firstly, we expect a competently performed anonymisation process to even out sharp variations in the relevant probability distributions. An anonymised employee satisfaction survey will not include the education category “Ph.D.” if there is just one Ph.D. in the work force but will rather use a broader category encompassing everyone with a university degree. Secondly, we want to offer a decision process that is useful for decision makers with scant computer science preparation. Employing widely known concepts such as the Shannon entropy makes the process easier to explain and accept. Thirdly, the conclusions of the Shannon-based decision process, appears, as we shall see, to be very restrictive. It is in fact hard to find realistic examples of situations where data can be both useful and properly anonymised. Using the Smith measure would make our conclusions even more restrictive but perhaps less believable since the result could be discarded as an artefact of using a little-known advanced information measure. Fourthly, we would like to use the decision process even in situations where detailed probability distributions are unavailable. Since the Shannon entropy is an average quantity rather than an extreme value, it is often easier to make back-of-an-envelope estimates of Shannon entropy. However, while sticking to using the Shannon entropy in Eq. (12), we at least partially take Smith's important lesson into account by including a step in the decision process of Section 3.2 for reasoning about outlier risks.

3. Decision-making

This section describes and exemplifies a process for making decisions about the release of anonymised data where release means dissemination either to the general public or to a more limited audience in which there may be adversarial elements. Section 3.1 relates the feasibility limit in Eq. (12) to the process requirements in Section 1.3. The decision making process is defined in Section 3.2 and two application examples are provided in Section 3.3.

3.1. *The feasibility limit as a basis for decisions*

Section 1.3 defines five principles A–E for guiding decisions about publishing anonymised data. Is the feasibility limit of Eq. (12) what a decision maker needs to make quantitative judgements in the spirit of these principles? The feasibility limit is based on the same privacy definition as in principle A i.e. it is concerned about attacks where the anonymised record of an individual target is identified. Equation (12) requires the decision maker to define the acceptable probability of broken privacy (principle B). Any of the records could be the target of the attack which means that none of the record owners are discriminated (principle C). The feasibility limit is furthermore based on pessimistic assumptions about the skills and perseverance of the adversary who is assumed to be able to optimally utilize the available information (principle D). We are according to principle E) sceptical about any prescience that we may have today about what kind of de-anonymisation algorithms that may be employed in the future and therefore Eq. (12) includes only a fundamental information-theoretic constraint on the efficiency of de-anonymisation.

Is the feasibility limit too complex for executive decision making? Computing the right-hand part of Eq. (12) can be done on a scientific pocket calculator. The difficult part is estimating the left-hand adversarial mutual information $I(X_t, Y_t|KW)$. This quantity is the essence of what the decision maker must know about the target and the adversary in order to make an informed risk estimate. It is, however, difficult to compute because decision makers typically lack the required information. In most practical situations, decision makers will not have access to generative data models, detailed statistical distributions or quantitative insights into the knowledge of the attacker. Detailed computations of the adversarial mutual information are therefore in many cases out of the question. However, many decision problems turn out to be quite straightforward even based on rudimentary estimates of the adversarial mutual information. This will be demonstrated in Section 3.3. The essential advantage of Eq. (11) and the concept of adversarial mutual information is that the decision maker has a clear definition of what to estimate. Using an information-theoretic measure in Eq. (12) makes the analysis independent of both the anonymisation and the de-anonymisation algorithms. This removes the need for algorithm expert support in decision making. The central analytical task is to reason about what kind of adversaries to worry about and what background information that they may have access to.

3.2. *Decision-making process*

We recommend the following process for making decisions about the release of an anonymised database.

3.2.1. *Decision-making process*

1) Find out precisely what information that is proposed for publishing. Ignore the anonymisation process and what information it has removed. Focus on the real-world meaning of the anonymised information. What can you learn about the individuals described by the data?
2) Generate an attack scenario by specifying the objectives of the adversary, the likelihood of the scenario, the consequences of a successful attack, and what the adversary knows about the target. To save work, start with the worst-case scenario.
3) Select the largest acceptable de-anonymisation probability p for the attack scenario. If the adversary makes a determined effort to identify the individual behind a given anonymised record, the probability for success is required to be less than p. In selecting p, consider the likelihood of the scenario and the consequences of de-anonymisation. If $p = 0$ is the only acceptable choice, do not publish the anonymised data since even pure guessing gives an adversary a small chance of pinpointing the target.
4) Determine if the assailant knows whether the target is in the anonymised database.
   a. If yes, count the number of records in the anonymised database N, set $\hat{N} = N$ and use Eq. (12) for computing the maximum adversarial mutual information n_p.
   b. If no, identify a larger population which the adversary knows encompasses both the target and the population of the anonymised database. Find the size $\hat{N}$ of this encompassing group and use Eq. (12) for computing n_p.
5) Estimate the adversarial mutual information $I(X_t, Y_t|KW)$. This may seem to be a daunting task but back-of-an-envelope estimates will often suffice. If the decision makers need technical support for this step it is recommended that expertise at arms-length from the designers of the anonymisation process is enrolled.
6) Compare the adversarial mutual information to the threshold n_p. If $I(X_t, Y_t|KW) \geq n_p$ terminate this process and do not release the anonymised data, else proceed to step 7.
7) Consider if the target may be an outlier in a statistical sense and hence could be more vulnerable than average record holders. If the make of automobiles owned by the record owner is included in the anonymised data, Toyota drivers would be safely anonymous but Koeningsegg owners might be at risk. If the scenario is about stalkers re-identifying wealthy celebrities it would be prudent to refuse releasing the data based on such scenario-critical outliers. If an unacceptable threat against outliers has been found, do not release the anonymised data, else proceed to step 8.
8) Repeat steps 2–7 for as many attack scenarios as deemed necessary for achieving the required level of security. What sufficient security means is a critical judgement call. The decision makers with the final responsibility for releasing the data must therefore be involved in this process.
9) If none of the attack scenarios prohibits release of the data according to steps 6 or 7, consider publishing the anonymised data.

Note that coming up with the relevant attack scenarios is a critical part of the process which requires domain knowledge, experience, insight and creativity. The thinking that goes into the attack scenarios is often much more crucial than the precision of computing the adversarial mutual information. A lot of time and resources can be saved by considering worst-case attack scenarios first and start with quick-and-dirty estimates of the adversarial mutual information. The process may well terminate after crudely analysing a first worst-case scenario which means that further scenario analysis or more precise adversarial mutual information computations are superfluous.

Note that identifying the worst-case scenario is based on intuition on what would be the most threatening combination of scenario likelihood, consequences of broken privacy and adversarial knowledge. What we consider to be unwanted consequences are also not objectively definable but depends on cultural, political and legal factors. In some cultural settings we may for example not consider it a privacy risk if parents learn sensitive information about their children even if the target may suffer considerable because of the leaked information. The person with the maximal background knowledge is also not necessarily the worst adversary. The spouse of the target may for example know all the data in the target's anonymised data base record but this means also that he or she learns nothing new by identifying the target's record other than that the target is included in the database. The latter may or may not be a privacy risk; it would for example not be a privacy risk if the released database lists the home owners of the county while it might be a considerable intrusion on the target's privacy if the database is about the users of an infidelity dating site. In the special case where publishing the anonymised data reveals only information that was publicly available before the release, decision makers could argue that there is no privacy risk since adversaries can learn nothing new from the published data. In such cases it will not be possible to come up with a credible attack scenario. Consider for example a database containing the age, gender and political affiliation of all the political candidates in a general election. Since all attributes of the any possible target already are open information and it also is a public fact that the persons in the list are candidates, decision makers could reasonably conclude that there is no conceivable attack scenario related to the release of the database.

It is crucial to document each step of the decision process. This documentation may have legal ramifications if the data is published. Anonymisation experts will find the documentation very useful if the data is not released and a more thorough anonymisation process is requested. Having good documentation is also convenient if the decision process need to be repeated for releases of new versions of the data or of the same data to different users.

3.3. *Application examples*

In this subsection we provide two examples on how to apply the decision-making process. The same steps as in Section 3.2 are followed in the examples. For brevity we consider just one attack scenario in each example although a real-life process could iterate over several scenarios.

3.3.1. *Anonymised demographic data*

The first application example considers including demographic data in anonymised databases.

1) Each record in the anonymised database consists of gender, zip code, year, month and day of birth as well as sensitive data on individual use of certain healthcare services. For brevity we will in this example not spell out the details of the healthcare-related data.
2) The attack scenario is about a prospective employer suspecting that a job candidate has health problems and trying to identify the record of the target for the purpose of finding out about the candidate's health status. The employer knows the gender, zip code, year, month and day of birth of the target but has no information related to healthcare.
3) The decision maker accepts a 5% risk of de-anonymisation in this attack scenario.
4) The database comprises all U.S. citizens registered in the year 2000 census and the attacker knows that the target belongs to that group. This means that process step 4.a can be applied. There are 281 million records in the database. Inserting $\widehat{N} = 281\,000\,000$ and $p = 0.05$ in Eq. (12) reveals that the maximum adversarial mutual information is $n_p \approx 26.5$.
5) A rough estimate of the adversarial mutual information assumes that gender, zip code and date of birth are mutually independent with uniform statistical distributions and proceeds to estimate the number of bits that is required to specify the components of the pseudo-identifier as follows: gender (1 bit); zip code (15.0 bits); year (6.3 bits); month (3.6 bits) and day (4.9 bits) where we assume 33233 U.S. zip codes and a uniform age distribution over a life span of 80 years. The total estimated adversarial mutual information is 30.8 bits.
6) Since the adversarial mutual information is larger than the threshold n_p (30.8 > 26.5), the privacy risk is unacceptable and the decision must be to not release the anonymised data. The decision makers realize that a more sophisticated statistical analysis may produce a different estimate of the adversarial mutual information but they find it unlikely that their decision would be swayed even if some advanced statistical model pushes the estimate closer to the threshold.
7) The target is not likely to be an outlier with respect to any of the attributes gender, zip code, year, month and day of birth. Gender and the time of birth are known to be evenly distributed for the age group that the target belongs to. None of the very few people who have their own zip codes are likely to apply for jobs in the firm.

The last step of outlier analysis is strictly not necessary as steps 1–6 indicate that the data should not be released. It is, however, included here to exemplify how this step can be performed. A different attack scenario may for example be about con artists targeting ageing people. In that case it might be relevant to note that a person aged 116 may be identifiable based on age alone.

The decision is based on using, in some cases, unrealistic uniform distributions. A careful statistical analysis by Sweeney (2000) based on the year 1990 U.S. census finds that 87% of the U.S. population is identifiable based on gender, zip code and date of birth. Similarly Golle (2006) finds that 63.3% of the U.S. population is identifiable based on gender, zip code and date of birth using data from the year 2000 U.S. census. This corroborates the simplified process in the example. Inserting $N = 281.000.000$ as well as the crudely estimated adversarial mutual information of step 5 as the value of n_p in Eq. (12) and solving for p leads to $p = 86.0\%$ meaning that the decision maker must accept a de-anonymisation probability of at least 86.0% in order to make the decision to publish the anonymised data set. This demonstrates that the simplified estimate in step 6 as expected somewhat exaggerates the probability for a successful attack compared to Golle's result. This makes, however, little difference to the release decision.

Clearly, we could use the result of the decision process as input to another round of anonymisation processing. After having the first attempt rejected according to the example above, anonymisation engineers could filter out the month and day of birth leaving only gender, zip code and year of birth in the demographic part of the anonymised data. Re-running the process according to steps 1–7 above, decision makers now estimate the adversarial mutual information to 22.3 bits (gender (1 bit); zip code (15.0 bits); year (6.3 bits)). While pondering further attack scenarios, it is found that possible adversaries would have no knowledge on the medical status of the targets and that they cannot have more relevant demographic knowledge since the first attack scenario already assumed maximum knowledge. As the estimated adversarial mutual information in the only relevant attack scenario is less than the threshold value ($22.3 < 26.5$) and no worse attack scenario is conceivable the decision should be to release the anonymised data.

3.3.2. *Anonymised air travel data*

The second application example considers releasing anonymised data on air travel. Decision makers could apply the process of Section 3.2 as follows.

1) The data about each individual in the anonymised database consists of a list of airline flights that the subject has travelled on or have tickets for. The information about each flight includes departure airport, destination airport, airline and flight code.
2) The villain of the attack scenario is a stalker who is tracking a woman living in a mid-sized European city. If the stalker finds out about the target's travel plans he may attempt to pursue and possibly attack the target physically. It is assumed that the stalker has access to the released data and knows about two outbound flights that the target has written about in her blog.
3) The decision maker accepts a 1% risk of de-anonymisation in this attack scenario.
4) The attacker is not sure that the target is in the anonymised database so process step 4.b is applied. The attacker knows, however, that the roughly one billion record owners are a subset of the world population. Hence $\widehat{N} = 7.2 \cdot 10^9$ and $p = 0.01$ are plugged into Eq. (12) with the result that the maximum adversarial mutual information is $n_p \approx 30.5$.
5) A rough and ready estimate of the adversarial mutual information may be performed as follows. The anonymised data spans a ten year period during which the target takes many trips starting out from the home airport. The total number of aeroplane departures from the home airport over ten years is 10^6 (loosely based on Arlanda airport statistics). Pinpointing any outbound flight requires therefore $Log_2(10^6) \approx 19.9$ bits of information. Based on the stalker's knowledge of two outbound flights we conclude that the adversarial mutual information is 39.8 bits.
6) Since the estimated adversarial mutual information (39.8 bits) exceeds the upper limit of 30.5 bits we conclude that the anonymised data cannot be published.
7) The target is unlikely to be an outlier with respect to departure airport, destination airport, airline and flight code since she is an ordinary passenger on a flight from a mid-sized European city.

The assailant's lack of knowledge about whether or not the target is in the database has, in this case, little effect on the conclusion. Using process step 4.a rather than 4.b gives an upper limit of 27.7 bits rather than 30.5 bits which is of no consequence for the decision.

Note that the adversarial mutual information in this example is estimated by comparing the information content of free text in a blog with attributes of the anonymised database. This is fundamentally different from counting the number of overlapping attributes as in the Narayanan&Shmatikov limit (Eq. (2)). The concept of adversarial mutual information is more versatile since it encompasses using uncertain information and comparing syntactically incongruent data. Consider for example a scenario as above where the target blogged about a visit to the Copenhagen Tivoli Gardens in 26 May 2014; the adversary had a 25% chance of observing the target entering the airport train in the morning of 26 May 2014 combined with the public knowledge that there is just one morning flight to Copenhagen from the relevant airport. None of these three pieces of information corresponds directly to an attribute in the anonymised database but can be combined to a 25% risk of the attacker correctly identifying a flight taken by the target. This would correspond to 17.9 bits added to the adversarial mutual information where the 2 bit penalty compared to the 19.9 bits per flight estimated above is due to the uncertainty of the information.

3.3.3. *Discussion of the application examples*

The decision-making process in Section 3.2 is based on the precept that privacy breach means identification of the target record in the anonymised database. This may not be the only risk that decision makers need to take into account. Consider for example a stalker, according to the second example, with access to information about just one outbound flight. The stalker may not realize that a uniquely identified record may not belong to the target and could therefore assault the wrong person, perhaps with grave consequences. Although it is impossible to guard against incompetent attackers making incorrect inferences the lesson to be learnt from this example is that decision makers should look out for scenario-specific risks that cannot be captured by the decision process provided in this paper.

False positives could in particular have grave consequences in government, intelligence and military de-anonymisation operations where we, however, can assume competent reasoning about identification probabilities. If the anonymised data has been released according to the process of Section 3.2, the parameter p that is selected in process step 3 will be an upper limit of the probability of correct de-anonymisation in the attack scenario that are fleshed out in process step 2. This means that the probability for that de-anonymisation algorithm points to the wrong individual as the most likely target is larger than 1-p. The competent adversary will, however, be able to reason correctly about the probability of false positives. The possibility of a ruthless adversary in a high-stake conflict indiscriminately targeting individuals that most likely are false positives for the off chance of striking the intended target should if applicable be considered in the risk estimate of process step 3.

Note that we in both of the examples used uniform probability distributions for estimating the adversarial mutual information. The uniform distribution gives a worst-case limit of the adversarial mutual information and is hence concordant with the principle of caution. If estimates based on uniform distributions suggest that the data cannot be released, decision makers have the option to postpone the decision and let experts make a more careful estimate using well-founded statistical distributions.

4. Conclusions

We have described and exemplified a process for making decisions about the release of anonymised information. At the core of this process is an information-theoretic limit on the adversarial mutual information which is independent of both anonymisation and de-anonymisation algorithms. Decision makers using this process need a basic understanding of probability and some training in the concept of adversarial mutual information but will not need to handle complex algorithms. The number of attack scenarios to be considered and how much analytic effort that is reasonable to spend depend on the privacy and security requirements of the application. The higher the cost of a privacy breach the more effort should go into the decision process.

The frequent failures of anonymisation as described in Section 1.2 is not hard to understand in the light of the feasibility limit that we have discovered in this paper. In Section 2.3 we demonstrate that 33 bits of adversarial mutual information normally is more than sufficient for successful de-anonymisation. Given the rich sources of digital information that is available about most of us, it is not surprising that it in many historical cases has been possible to muster at least 33 bits of personal background information that overlaps with the published data and thus re-identify record owners. Using the decision process provided in Section 3.2 it should be possible for future decision makers to avoid publishing too weakly anonymised information.

One could argue that the process recommended here is an effective instrument for disputing the release of an anonymised database but that it appears to far less suitable for finding reasons in favour of releasing the data. This apparent prejudice is, however, an unavoidable property of the decision problem rather than a bias in the decision-making process. Releasing an anonymised database is, as discussed in Section 1.4 not an operation characterized by differential privacy and it is therefore not possible to provide strong privacy guarantees independent of the knowledge of adversaries. This means that we must reason about attack scenarios and that one single attack scenario can be sufficient reason for not releasing the data while only the combined analysis of all relevant attack scenarios can form an adequate foundation for publishing the data.

The decision-making process provides documentation of the attack scenarios that were considered and how the risks where analysed. This documentation is valuable in periodic security reviews, audits and legal processes. Having considered all relevant attack scenarios gives the decision makers confidence in their decisions and makes the decisions transparent and defendable. However, making the decision process auditable and transparent may also open up new avenues for criticism and litigation since the documentation exposes the unavoidable compromises and risk-benefit balancing that is inherent in any decisions about publishing anonymized sensitive data. Understanding the brittleness of data anonymisation and how to make informed decisions about data publishing are useful also in political deliberations and in law making related to privacy of personal information. The concept of adversarial mutual information and how it relates to privacy risks could be helpful also in this context.

Acknowledgements

Fredrik Johansson has reviewed multiple versions of this manuscript and provided penetrating questions and much helpful discussion. We are grateful for profound insights provided by three anonymous reviewers. This work has been funded by the European FP7 FIDELITY project (SEC-2011-284862) and by the R&D programme of the Swedish Armed Forces.

REFERENCES

Aggarwal C. Privacy and the dimensional curse. In: Aggarwal CC, Yu PS, editors. Privacy-preserving data mining: models and algorithms. Springer; 2008. pp. 433–60.

Aggarwal C, Yu P, editors. Privacy-preserving data mining – Models and algorithms, vol. 34. Springer; 2008.

Aggarwal C, Yu P. A survey of randomization methods for privacy-preserving data mining. In: Aggarwal CC, Yu PS, editors. Privacy-preserving data mining: models and algorithms. Springer; 2008. pp. 137–56.

Barbaro M, Zeller T. A face is exposed for AOL. New York Times; 2006 [Searcher No. 4417749].

Bartlett E. RMS need to safeguard computerized patient records to protect hospitals. Hospital Risk Manag 1993;15(9):129–32.

Brynielsson J, Johansson F, Jändel M. Privacy-preserving data mining – a literature review'(FOI-R-3633-SE) [Technical report]. Swedish Defence Reserach Agency; 2013.

Chen K, Liu L. A survey of multiplicative perturbation for privacy-preserving data mining. In: Aggarwal CC, Yu PS, editors. Privacy-preserving data mining: models and algorithms. Springer; 2008. pp. 157–81.

Datta A, Sharma D, Sinha A. Provable de-anonymization of large datasets with sparse dimensions. In: Proc. ETAPS Conference on Principles of Security and Trust; 2012.

Denning DE, Schlörer J. Inference controls for statistical databases. IEEE Comput 1983;16(7):69–82.

Dwork C. Differential privacy. In: Proc. of the 33rd international colloquium on automata, languages and programming. Springer; 2006. pp. 1–12.

Dwork C. Differential privacy: a survey of results. In: Proc.. of the 5th international conference on theory and applications of models of computation. Springer; 2008. pp. 1–19.

Dwork C. The promise of differential privacy: a tutorial on algorithmic techniques. In: IEEE 52nd annual symposium on foundations of computer science. IEEE; 2011. pp. 1–2.

El Emam K, Jonker E, Arbuckle L, Malin B. A systematic review of re-identification attacks on health data. PLoS ONE 2011;6(12):e28071.

Fung BCM, Wang K, Chen R, Yu PS. Privacy-preserving data publishing: a survey of recent developments. ACM Comput Surv 2010;42(4):14:1–14:53.

Golle P. Revisiting the uniqueness of simple demographics in the US population. In: Proceedings of the 5th ACM workshop on privacy in electronic society. New York, NY, USA: ACM; 2006. pp. 77–80.

Li N, Li T. t-closeness: privacy beyond k-anonymity and *l*-diversity. In: Proceedings of IEEE international conference on data engineering. IEEE; 2007. pp. 106–15.

Lin Z, Owen A, Altman RB. Genomic research and human subject privacy. Science 2004;305:183.

Machanavajjhala A, Kifer D, Gehrke J, Venkitasubramaniam M. *l*–diversity: privacy beyond k-anonymity. ACM Trans Knowl Discov Knowl Discov 2007;1(1) [article 3].

McGuire AL, Gibbs R. No longer de-identified. Science 2006;312:370–1.

Narayanan A, Shmatikov V. Privacy and security: myths and fallacies of "Personally Identifable information". Commun ACM 2010;53:24–6.

Narayanan A, Shmatikov V. Robust de-anonymization of large sparse datasets. In: Proceedings of the 2008 IEEE Symposium on Security and Privacy; 2008. pp. 111–25.

Ohm P. Broken promises of privacy: responding to the surprising failure of anonymization. UCLA Law Rev 2010;57:1701–77.

Rebollo-Monedero D, Forné J, Domingo-Ferrer J. From t-closeness-like privacy to postrandomization via information theory. IEEE Trans Knowl data Eng 2010;22(11):1623–36.

Rothstein M. Is deidentification sufficient to protect health privacy in research? Am J Bioeth 2010;10(9):3–11.

Sankar L, Raj Rajagopalan S, Poor HV. A theory of utility and privacy of data sources. In: Proceedings of the 2010 IEEE symposium information theory; 2010. pp. 2642–6.

Smith G. On the foundations of quantitative information flow. In: Proceedings of the 12th International Conference on Foundations of Software Science and Computational Structures; 2009. pp. 288–302.

Sweeney L. k-anonymity: a model for protecting privacy. Int J Uncertain Fuzziness Knowledge-based Syst 2002;10:557–70.

Sweeney L. Simple demographics often identify people uniquely [Technical report]. Pittsburgh: Carnegie Mellon University; 2000.

Sweeney L. Weaving technology and policy together to maintain confidentiality. J Law, Med Ethics 1997;35:2–3.

Magnus Jändel is a deputy research director at the Swedish Defence Research Agency (FOI) and associate professor in theoretical physics at the Royal Institute of Technology in Stockholm. He obtained a PhD in theoretical physics 1985, was Senior Fellow at the theory division of CERN, research manager in image and video coding at Ericsson Research and founded start-up IT company Terraplay Systems. His research interests focus on artificial intelligence including computational creativity, situation analysis and reasoning, support vector machines and brain-inspired intelligent systems. He is the author or co-author of more than 42 peer-reviewed papers and 14 multi-national patents.