German Record Linkage Center
WORKING PAPER SERIES
NO. WP-GRLC-2014-01 | JANUARY 15, 2014

A Graph Theoretic Linkage Attack on Microdata in a Metric Space

Martin Kroll

A Graph Theoretic Linkage Attack on Microdata in a Metric Space

Martin Kroll

January 15, 2014

The knowledge of (e.g. geographic) distances between the entities whose data are stored in a microdata table makes several methods of analysis only possible. For instance, such knowledge is necessary and sufficient to perform data mining tasks such as nearest neighbour search or clustering. However, the risk of identity disclosure has to be reconsidered once more when (approximate) inter-record distances are published in addition to the microdata for research purposes. In order to tackle this problem, we introduce a flexible graph model for microdata in a metric space and propose a linkage attack based on a realistic assumption on a data snooper's background knowledge. This attack is based on the idea of finding a maximum approximate common subgraph of two vertex-labelled and edge-weighted graphs. By adapting a standard argument from graph theory to our setup this task is transformed to the maximum clique detection problem in a corresponding product graph. A toy example and experimental results on simulated data show that publishing even approximate distances only might increase the risk of identity disclosure unreasonably.

Keywords: Linkage attack, identity disclosure, maximum approximate common subgraph problem

1 Introduction

Unquestionably, enriching microdata with spatial information opens up numerous additional approaches for analysis. In the area of epidemiology, this insight goes back at least to the middle of the 19th century when John Snow detected a contamined water pump in London as the outbreak source of a cholera epidemy by linking cases of mortality to their location and visualising these locations and the positions of surrounding water pumps in a map [30]. In recent years, techniques from spatial analysis have become more and more attractive to answer questions also in the area of social sciences [26]. However, when personal microdata containing sensitive information (e.g. gathered in a survey) are to be published for research purposes, the anonymity of the individuals has to be guaranteed. Therefore, microdata are usually released containing spatial information only in aggregated form which restricts the choice of applicable techniques for analysis drastically. In particular distance calculations (especially for entities which are closely located) become difficult and imprecise on the basis of aggregated data. However, many data mining techniques and methods in spatial analysis require accurate distance computations. For this reason, it is necessary to investigate to what extent additionally published (approximate) inter-record distances influence the risk of identity disclosure and how a perhaps

non-acceptable increase of this risk can be prevented. Our work presented in this article provides a novel attempt in order to tackle these questions.

Contributions of the paper. We introduce a flexible and natural graph model for microdata with known inter-record distances. The search of a maximum common subgraph between two such graph models is interpreted as a novel kind of linkage attack on such microdata. We discuss relative merits of our method in comparison to usual linkage attacks on the basis of a small-scale example (example 10 in section 4). Further, in the special case of geographical distances, it is shown on the basis of simulated data that there exists a non-negligible risk of identity disclosure if $\mathcal{N}(0, \sigma^2)$-distributed Gaussian noise is added to the input coordinates for small values of σ. For larger values of σ (which lead to sufficiently anonymized data), however, the data become nearly useless for further analysis. These results reflect a trade-off between data utility and disclosure risk through the proposed attack.

Organisation of the paper. In section 2 we refer to related work. The preparatory section 3 provides a graph model for microdata in a metric space which forms the basis of the graph theoretic linkage attack introduced in section 4. In section 5 this attack is evaluated by means of a simulation study. We conclude and discuss possible directions for future research in section 6.

2 Related work

Statistical disclosure control and privacy preserving data mining. As indicated already in the introductory section above, the original motivation for the work presented in this article stems back to the wish of making the wide variety of distance-based methods (e.g. from spatial statistics) also applicable for microdata that are published for scientific purposes. As it is intuitively compelling that naive release of exact distances between individuals can increase the risk of deanonymization, the question of interest is how the knowledge of approximate distances only might change the risk of identity disclosure, i.e. the chance of a data snooper who attempts to identify some of the entities.
In general, the analysis of such deanonymization attacks on microdata and the development of tools for their anonymization is a central topic of *statistical disclosure control* [18]. It is universally acknowledged that a necessary but not sufficient first step during the process of anonymization consists in the removal of all attributes that can be used to identify an individual entity unambiguously (this step is usually referred to as *deidentification*). Such attributes (e.g. the `social insurance number`) are called *(direct) identifiers*, in contrast to *quasi-identifiers* that do not have the power to nullify an individual's anonymity by their own, a distinction which has to be ascribed to Dalenius [11]. By a combination of quasi-identifiers, however, it might be possible to assign an entity from the underlying population to a specific record of a published microdata file unambigously. For example, it has been shown in [31] using 1990 US census data that 87% of the population of the United States are uniquely determined by their values with respect to the quasi-identifier set {`5-digit ZIP code`, `gender`, `date of birth`}. This fact motivates a mode of attack that is commonly referred to as *linkage attack* [13]: In this scenario it is assumed that a data snooper has access to an external auxiliary microdata file (called *identification file*) containing both direct identifiers and quasi-identifiers as attributes. By making use of the quasi-identifiers the snooper attempts to identify entities by linking records from the

identification file to records from the published microdata file (termed *target file*). A real-life-example of linkage via quasi-identifiers is due to Sweeney [32]: She was able to detect the record corresponding to the governor of Massachusetts in a published health data file by linkage with a publicly obtainable voter registration list. Whereas theoretical results on linkage attacks have been obtained rather recently in [24], the concept of k-anonymity has been proposed as a remedy against linkage attacks already in [29]. The basic idea of k-anonymity is to modify the records in the released microdata such that every record coincides with at least $k-1$ other records with respect to the quasi-identifiers. For this reason, unambiguous linkage between identification and target file will not be possible for a data snooper leading to ties in the linkage process. The graph theoretic linkage attack introduced in section 4 contains the classical linkage attack via quasi-identifiers as a subroutine but will provide a way to resolve at least some of the ties using the additionally published distance information.

Several papers on *privacy preserving data mining* have already discussed privacy issues with respect to distance-preserving transformations of microdata. However, in these articles it is generally assumed that the considered distances can be directly calculated from the microdata whereas our focus is on microdata enriched with supplementary distances between the entities that cannot be calculated from the microdata itself. Moreover, in most cases only specific kinds of distances have been considered (e.g. ℓ_1-distance in [28] or euclidean (i.e. ℓ_2-) distance in [23]). In contrast, the attack proposed in this paper can be applied to any kind of distance function (albeit the special case of spatial distances motivated our research and is exclusively referred to in our examples). Further, a distance-preserving technique for anonymization of binary vectors has been discussed in [20]. In contrast to our approach, in that article the distance information is not assumed to increase the risk of identity disclosure by itself.

Location privacy and geographically masking. There is a vast literature on the problem of identity disclosure when dealing with spatially referenced data. Opportunities and challenges considering spatial data in the context of social sciences are discussed in great detail in [16] and [17]. The articles [4] and [10] give illustrative examples of how naive publishing of spatially referenced data can lead to violation of anonymity: In both cases the respective authors were able to reconstruct a large amount of original addresses successfully even from low resolution maps. A currently flourishing branch of research deals with anonymization techniques for datasets containing mobility traces of individuals (which can e.g. be obtained by mobile phone tracking). This topic is usually referred to as *location privacy* [21]. In this article, however, we consider the deanonymization risk which arises from knowledge of (approximate) distances between fixed spatial points assigned to the entities in a microdata table. Various methods for the anonymization of geographic point data (not necessarily taking additional covariates into consideration as in our case) have been discussed under the name of *geographically masking*. [1] and [25] provide comprehensive outlines of existing methods. A noteworthy method is due to Wieland et al. [33] who developed a method based on linear programming that moves each point in the dataset as less as possible under a given quantitative risk of re-identification. However, the aim of nearly all proposed anonymization techniques for spatially referenced data consists in distorting the spatial distribution with respect to the underlying geometry as less as possible, whereas attempts predominantly focusing on the preservation of distances have not been discussed yet in the context of spatial data. It appears to be obvious that neglecting the underlying geographical area might permit a higher level of accuracy regarding distance calculations.

Social network anonymization. The use of a graph model in this article might suggest a strong connection between our approach and methods discussed in the area of *social network anonymization* [34]. However, we will model microdata with known inter-record distances by a complete graph with vertex labels and edge weights which is a very specific model in contrast to the more general graph models commonly used in social network analysis. Indeed, the graphs modelling social networks are usually a long way off being complete and edges are not weighted in general. For example, in [6] the underlying graph model considers (discrete) edge labels instead of real valued weights only. Furthermore, active attacks (consisting in the addition of nodes to the published network by an intruder) as in [2] seem not to be sensible when investigating the risk of identity disclosure for published microdata. The active attack proposed in [2] is related to the one in this paper as it makes use of graph algorithmic building blocks as well. It consists in the detection of a subgraph in a larger graph whereas the attack in this paper is based on finding common subgraphs of two different graphs.

Pattern recognition. To the best of our knowledge this paper is the first one which makes use of a graph model for a microdata file and distances between its records. Finding a matching between two such graph models constitutes the basic principle of the graph theoretic linkage attack proposed in this article and is an often considered problem in the *pattern recognition* field and its various areas of application (cf. [7] as a source providing an extensive outline). Fundamental to our presentation here is the article by Levi [22] which motivates to transform the problem of finding (maximum) common subgraphs of two graphs into a (maximum) clique detection problem, and its adaption in [14] where the original approach by Levi has been relaxed in order to deal with approximate common subgraphs as well. This transformation to the maximum clique detection problem is of particular interest due to its various fields of application (e.g. biochemistry [14]). The problem of finding a maximum clique in a graph is known to be NP-hard [15] and a lot of attention has been paid to the development of techniques for solving this problem either exactly or at least approximately [3]. For the graphs considered in the experimental section 5 of this paper making use of the (maximum) clique detection algorithm provided by the R package `igraph` [9] turned out to be sufficient. Considering approximative algorithms for the (maximum) clique detection problem that work well also for larger graphs and exploring the limits of our approach in view of its scalability are topics postponed to future research.

3 A graph model for microdata in a metric space

Preliminaries. A metric space is defined as a pair (X, d) where X is a set and d is a (distance) function $d : X \times X \to \mathbb{R}$ satisfying the following three conditions: (i) $d(x, x) = 0$ and $d(x, y) > 0$ whenever $x \neq y$, (ii) $d(x, y) = d(y, x)$ and (iii) $d(x, y) \leq d(x, z) + d(z, y)$. We assume that a deduplicated microdata table T at hand contains information with respect to an attribute set $\mathcal{A} := \{A_1, \ldots, A_m\}$ about $N_T \in \mathbb{N}$ entities from an underlying population. The fact that distances between the entities of T are known can be modelled, in mathematical terms, by means of a function $\tau : [N_T] := \{1, \ldots, N_T\} \to (X, d)$, $i \mapsto \tau(i)$ which maps the ith record/entity of T to a point $\tau(i)$ in a metric space X such that the distance between records i and j of T is equal to $d_{ij} := d(\tau(i), \tau(j))$. The distances between all the entities can then be stored in the $N \times N$ distance matrix $D = (d_{ij})$. Such a pair (T, D) is hereafter referred to as *microdata in a metric space*. Note that we did not state any assumptions on the function τ such as injectivity or surjectivity. It is easy to see

that $[N_T]$ itself becomes a metric space by the pullback of d by τ if and only if τ is injective (cf. page 81 in [12]). In general, $[N_T]$ becomes a pseudometric space only. From our point of view, this flexibility regarding τ is intended as the records of a microdata table often only form a pseudometric instead of a metric space which is illustrated by the following example: Consider microdata about individuals which have been gathered in a scientific survey. If two respondents share a common residence the geographical distance between these respondents will be equal to zero and thus the set of respondents with accompanying distances between them forms a pseudometric space only. Thus, the distance matrix D is not assumed to be a proper distance matrix, i.e. zeroes outside the diagonal are permitted.

Some terms from graph theory. Given a set S we denote the set of its two-element subsets by $[S]^2$. A *(simple undirected) graph* $\mathcal{G} = (V, E)$ consists of a set V (whose elements are termed *vertices*) and a set $E \subseteq [V]^2$ of *edges*. The cardinality $|V|$ of V is called the order of $\mathcal{G}$. Two distinct vertices v and w of V are *adjacent* if $\{v, w\} \in E$. The existence of an edge between v and w will sometimes be denoted by $vw \in E$ as a short hand. A graph is called *complete* if any two of its vertices are adjacent. A graph $\mathcal{G}' = (V', E')$ with $V' \subseteq V$ and $E' \subseteq [V']^2 \cap E$ is a *subgraph* of $\mathcal{G} = (V, E)$. If even $E' = [V']^2 \cap E$ holds, the graph $\mathcal{G}'$ is called an *induced subgraph* of $\mathcal{G}$ or we say that the subset V' of vertices induces $\mathcal{G}'$ in $\mathcal{G}$ which is denoted by $\mathcal{G}' = \mathcal{G}[V']$. A subset of the vertex set V is a *clique* if the subgraph induced by these vertices is complete. A clique containing k elements is termed a *k-clique*. A clique is *maximal* if it is not contained in any properly larger clique. A clique is *maximum* if there is no other clique containing more vertices. Clearly, a maximum clique is always maximal but not generally vice versa.

The notion of a vertex-labelled and edge-weighted graph is of fundamental importance to the graph model for microdata in a metric space introduced below. This notion is just a special case of the more general notion of an *attributed graph* which is frequently used in the pattern recognition community [5].

Definition 1. Let $\mathcal{L}_V$ be a set of vertex labels. A *vertex-labelled* and *edge-weighted graph* is a four-tuple $\mathcal{G} = (V, E, \lambda, \omega)$, where V is the vertex set, $E \subseteq [V]^2$ the edge set, $\lambda : V \to \mathcal{L}_V$ the vertex-labelling function and $\omega : E \to \mathbb{R}$ a weight function which assigns real numbers to the edges.

Graph model for microdata in a metric space. Let (T, D) be microdata in a metric space and N_T the number of records in T as above. An associated vertex-labelled and edge-weighted graph $\mathcal{G} = \mathcal{G}(T, D) = (V, E, \lambda, \omega)$ can be defined as follows: Set $V = \{1, \ldots, N_T\}$, $E = [V]^2$ and define $\omega_E : E \to \mathbb{R}$ via $\omega_E(ij) = d_{ij} := d(\tau(i), \tau(j))$; the labelling function $\lambda_V : V \to \mathcal{L}_V$ assigns a certain part of the information stored in T for a record to the corresponding vertex of the graph $\mathcal{G}$ (cf. example 2 below). Note that the simple undirected graph $\mathcal{G}_{\text{simple}} := (V, E)$ obtained from $\mathcal{G}$ by forgetting vertex labels and edge weights is the complete graph K_{N_T} with N_T vertices. This graph theoretical structure appears adequate for modelling microdata in a metric space: Loops, i.e. edges linking a vertex with itself, are not necessary because $d_{ii} = 0$ for any vertex $i \in V$ and undirected edges are sufficient for reflecting the distance by the corresponding edge weights due to the symmetry $d_{ij} = d_{ji}$ of the distance matrix $D = (d_{ij})$. Obviously, it would be easy to widen this model, e.g. by introducing directed edges, if this would be necessary for a specific application.

Example 2. Consider the imaginary microdata provided by table 1 containing personal microdata with respect to the attributes `name`, `sex`, `birth location` and `year of birth`.

The function τ maps each individual to the geographic coordinates (longitude λ and latitude θ in degrees) of the correspoding birth location with respect to the World Geographic System WGS 84, i.e.

$$\begin{aligned}
\tau(1) &= (-0.1198244, 51.51121) \quad \text{(Alice was born in London)} \\
\tau(2) &= (2.3522219, 48.85661) \quad \text{(Bob was born in Paris)} \\
\tau(3) &= (-3.7037902, 40.41678) \quad \text{(Eve was born in Madrid)} \\
\tau(4) &= (13.4049540, 52.52001) \quad \text{(Walter was born in Berlin)}
\end{aligned}$$

Assuming a spherical shape with radius $R = 6371$ km for the earth and converting degrees to radians the geographical distance d between two locations (λ_1, θ_1), (λ_2, θ_2) can be calculated as $d = R \cdot \phi$ where

$$\cos\phi = \sin\theta_1 \sin\theta_2 + \cos\theta_1 \cos\theta_2 \cos(\lambda_1 - \lambda_2).$$

Using this formula leads to the following distance matrix D.

$$D = (d_{ij}) = \begin{pmatrix} 0 & 343.6 & 1264.0 & 930.9 \\ 343.6 & 0 & 1052.9 & 877.5 \\ 1264.0 & 1052.9 & 0 & 1869.1 \\ 930.9 & 877.5 & 1869.1 & 0 \end{pmatrix}.$$

Then, for the corresponding graph model we have $V = \{1, 2, 3, 4\}$, $E = [V]^2$ and the edge weights are defined via $\omega(ij) = d_{ij} = d_{ji}$. We define the vertex-labelling function by assigning the information regarding the attributes `sex` and `year of birth` to each vertex, i.e. formally we have $\lambda_V : V \to \text{dom}(\texttt{sex}) \times \text{dom}(\texttt{yob})$.
The resulting vertex-labelled and edge-weighted graph can be visualised as in figure 1.

name	sex	birth location	year of birth
Alice	f	London	1978
Bob	m	Paris	1965
Eve	f	Madrid	1943
Walter	m	Berlin	1931

Table 1: Example microdata table. The table contains the attributes `name`, `sex`, `birth location` and `year of birth`.

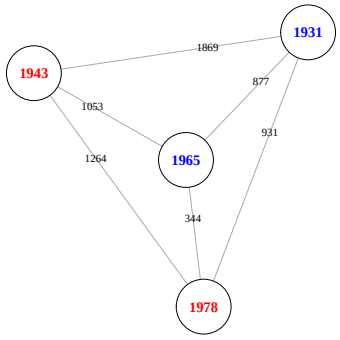

Figure 1: A graph model for the example microdata. The attribute `sex` is indicated by the colour of the vertex label.

4 A graph theoretic linkage attack

Prerequisites for the attack. In order to make any kind of linkage attack with the objective of identity disclosure at least possible we have to presuppose the availability of an appropiate external microdata file to the data snooper.

Assumption 3. *The snooper is in possession of an identification file containing direct identifiers.*

Under this assumption classical linkage attacks which are based on comparisons considering quasi-identifiers of identification and target file can be conducted. As already mentioned in section 2 such attacks yield an important mode of attack aiming for identity disclosure in the literature on deanonymization of microdata. In order to perform a linkage attack that goes beyond these ordinary ones by also taking the information given by the pairwise distances between the records into consideration, we have to widen the setup by a second assumption.

Assumption 4. *The snooper is able to calculate the distances between the entities in the identification file at least approximately.*

Although in some cases assumption 4 might not be fulfilled, it is easy to find examples when this will indeed be the case. For instance, when the target file containing survey data is enriched by the geographic distances between the respondents' residences, we assume that the snooper can geocode the addresses of the individuals in the identification file and calculate the corresponding distance matrix. In this example, there will be at least some dependence on the methods used for geocoding and distance calculation, a fact which has to be considered in the creation of an attack mode. Analogously, an intended modification of the distances in the target file by the dataholder for the purpose of anonymization has to be thought about later.

Approximate common subgraphs. Due to assumptions 3 and 4 a data snooper can create a vertex-labelled and edge-weighted graph as defined in section 3 both for target and identification file. At this step, the snooper will only consider common quasi-identifiers of both files for the definition of vertex labels because comparisons of records can only be based on such attributes. Hereafter, the resulting graphs will be referred to as the *target* and the *identification graph*. Hence classical linkage attacks consist in trying to find for each vertex in the identification graph vertices in the target graph that result in matches regarding the accompanying vertex labels. In the parlance of graph theory this approach is equivalent to the search for *common subgraphs* of order 1, a notion which will be made precise below. This course of action will usually (e.g. if the target file satisfies k-anonymity for some $k > 1$) lead to ties which cannot be broken without any extra information. However, due to the additional information given by the edge weights in the graph model, the snooper is able to search for complete common subgraphs of order > 1 which forms the essence of our attack. It is intuitively apparent that taking edge weights into consideration augments a snooper's chance to evaluate the credibility of potential matches. For instance, consider vertices v_1, v_2 in the target graph $\mathcal{G}_1 = (V, E, \lambda_V, \omega_E)$ and w_1, w_2 in the identification graph $\mathcal{G}_2 = (W, F, \lambda_W, \omega_F)$ such that $\lambda_V(v_1) = \lambda_W(w_1)$ and $\lambda_V(v_2) = \lambda_W(w_2)$, i.e. we observe coincidence regarding the vertex labels. Then, if the corresponding edge weights $\omega_E(v_1v_2)$ and $\omega_F(w_1w_2)$ are at least approximately equal (which will be denoted by $\omega_E(v_1v_2) \approx \omega_F(w_1w_2)$), this fact will augment the credibility of the two matches (v_1, w_1) and (v_2, w_2). Vice versa, a large distortion with respect to the corresponding edge weights will reduce this credibility: In this case at least one of the considered matches has supposed to be false. These considerations can easily be generalised to more than two matches and all accompanying edge weights. The more potential matches preserve all the accompanying edge weights, the more the credibility of all these potential matches will increase. This motivates the snooper's intention to detect nearly identical substructures in both graphs which are as large as possible. As indicated above it seems convenient to allow some deviation with respect to the edge weights in this context due to deviations which cannot be circumvented by a snooper (as mentioned in the

special case of geographic distances already above). All these thoughts can be dealed with rigorously using the notion of an *approximate common subgraph* of two vertex-labelled and edge-weighted graphs. This notion is made precise by means of the following definition.

Definition 5. Let $\mathcal{G}_1 = (V, E, \lambda_V, \omega_E)$ and $\mathcal{G}_2 = (W, F, \lambda_W, \omega_F)$ be two vertex-labelled and edge-weighted graphs in the sense of definition 1. An *approximate common subgraph* of $\mathcal{G}_1$ and $\mathcal{G}_2$ is given by subsets $S \subseteq V$, $T \subseteq W$ and a bijection $\varphi : S \to T$ such that the following two statements are true:

(i) $\lambda_V(s) = \lambda_W(\varphi(s))$ for all $s \in S$.

(ii) For all $s_1, s_2 \in S$ we have either
   (a) $s_1 s_2 \in E$, $\varphi(s_1)\varphi(s_2) \in F$ and $\omega_E(s_1 s_2) \approx \omega_F(\varphi(s_1)\varphi(s_2))$, or
   (b) $s_1 s_2 \notin E$ and $\varphi(s_1)\varphi(s_2) \notin F$.

Discussion. Condition (ii) of definition 5 guarantees that vertices $v_1, v_2 \in V$ can only be mapped to vertices $w_1, w_2 \in W$ if either both pairs of vertices are adjacent or non-adjacent (the non-adjacency yields even a sufficient condition for making such a mapping possible). Because we consider complete graphs in this article exclusively, only requirement (a) in condition (ii) has to be checked as requirement (b) will never be fulfilled. Furthermore, the interpretation of $\approx$ in definition 5 has to be made precise depending on the prevailing situation and particularly on possible perturbations of the distances conducted by the dataholder before publishing the microdata. This issue will be dealt with in detail in example 10 in this section and the simulation study in section 5. It would have certainly been possible to allow some amount of deviation regarding the vertex labels as well by introducing a similarity measure on the set of vertex labels. In this paper, however, we will not deal with this aspect and require exact coincidence for the labels of two vertices to be matched as we are primarily interested in the effect of how publishing (perturbed) distances influences the risk of identity disclosure.

The product graph. In order to tackle the problem of finding approximate common subgraphs of two vertex-labelled and edge-weighted graphs $\mathcal{G}_1$ and $\mathcal{G}_2$, we transform this problem to the problem of clique detection in an appropriately defined simple undirected graph $\mathcal{G}_\otimes$, the product graph of $\mathcal{G}_1$ and $\mathcal{G}_2$.

Definition 6. Let $\mathcal{G}_1 = (V, E, \lambda_V, \omega_E)$ and $\mathcal{G}_2 = (W, F, \lambda_W, \omega_F)$ be two vertex-labelled and edge-weighted graphs as in definition 1. The *product graph* $\mathcal{G}_\otimes = (V_\otimes, E_\otimes)$ of $\mathcal{G}_1$ and $\mathcal{G}_2$ is a simple undirected graph defined through

$$V_\otimes = \{(v, w) \in V \times W : \lambda_V(v) = \lambda_W(w)\} \quad \text{and}$$

$$E_\otimes = \Big\{\{(v_1, w_1), (v_2, w_2)\} : v_1 \neq v_2, w_1 \neq w_2 \text{ and either (a) } v_1 v_2 \in E, w_1 w_2 \in F \text{ and } \omega_E(v_1 v_2) \approx \omega_F(w_1 w_2), \text{ or (b) } v_1 v_2 \notin E \text{ and } w_1 w_2 \notin F\Big\}.$$

The announced transformation of the maximum approximate common subgraph problem into the maximum clique problem is achieved via the following theorem.

Theorem 7. *Consider the setup of definition 6. Then, there exists a one-to-one correspondence between approximate common subgraphs of order k and k-cliques of $\mathcal{G}_\otimes$.*

Proof. Let an approximate common subgraph of $\mathcal{G}_1$ and $\mathcal{G}_2$ of order k be given by the vertex sets $S = \{v_1, \ldots, v_k\} \subseteq V$ and $T = \{w_1, \ldots, w_k\} \subseteq W$, respectively. Without loss of generality we assume $\varphi(v_i) = w_i$ for $i \in \{1, \ldots, k\}$ under the corresponding subgraph isomorphism φ. Condition (i) in definition 5 yields $(v_i, w_i) \in V_\otimes$ for $i = 1, \ldots, k$. Moreover, for distinct $i, j \in \{1, \ldots, k\}$ we have $v_i v_j \in E \Leftrightarrow w_i w_j \in F$ and $\omega_E(v_i v_j) \approx \omega_F(\varphi(v_i)\varphi(v_j)) = \omega_F(w_i w_j)$ if $v_i v_j \in E$ due to condition (ii) in definition 5, which implies that (v_i, w_i) and (v_j, w_j) are adjacent in $\mathcal{G}_\otimes$. Because i, j were chosen arbitrarily, $\mathcal{C} := \{(v_1, w_1), \ldots, (v_k, w_k)\}$ forms a k-clique in $\mathcal{G}_\otimes$.

Conversely, let $\mathcal{C}$ be a k-clique in $\mathcal{G}_\otimes$ given by vertices $(v_1, w_1), \ldots, (v_k, w_k) \in V_\otimes$. We define $S = \{v_1, \ldots, v_k\}$, $T = \{w_1, \ldots, w_k\}$ and $\varphi : S \to T$ via $\varphi(v_i) = w_i$. Then φ is a bijection and we obtain $\lambda_V(v_i) = \lambda_W(w_i) = \lambda_W(\varphi(v_i))$ for $i = 1, \ldots, k$. Thus condition (i) in definition 5 is satisfied. The validity of the second condition follows from the fact that either $v_i v_j \notin E$ and $w_i w_j \notin F$ or $v_i v_j \in E$ and $w_i w_j \in F$ and that we have $\omega_E(v_i v_j) \approx \omega_F(w_i w_j) = \omega_F(\varphi(v_i)\varphi(v_j))$ in the latter case. □

Corollary 8. *The problem of finding a maximum approximate common subgraph of two vertex-labelled and edge-weighted graphs is equivalent to the problem of detecting a maximum clique in the associated product graph.*

Discussion. Before putting the ingredients discussed so far together we state some remarks regarding theorem 7 which is folklore within the pattern recognition community. For the first time the correspondence between common subgraphs of two graphs and cliques in the corresponding product graph was considered in [22] in case of exact isomorphisms between simple graphs including vertex labels. This approach has become a standard tool for tackling graph matching problems in various fields of application since then [7]. The definition of the product graph recently presented in [14] is equivalent to our one here. However, in that paper the authors relax the concept of a clique (which appears to be too restrictive for their application) to the less restrictive concept of a γ-quasi-clique and propose an algorithm for quasi-clique detection based on local-clique-merging.
Analogous to the remark after definition 5 only requirement (a) in the construction of the edge set $E_\otimes$ is relevant for our application as we deal with complete graphs only. The condition that $v_1 \neq v_2$ and $w_1 \neq w_2$ in the construction of $E_\otimes$ guarantees that cliques in the product graph correspond to one-to-one matches between vertices of the two graphs to be matched. For complete graphs without loops this condition is implicitly part of the requirement that $\omega_E(v_1 v_2) \approx \omega_F(w_1 w_2)$. If we had permitted loops in our graph model and assigned the weight 0 to each of these loops requiring $v_1 \neq v_2$ and $w_1 \neq w_2$ would have been necessary in order to keep the statement of theorem 7 true which is illustrated by the following example.

Example 9. Consider $\mathcal{G}_1 = (V, E, \lambda_V, \omega_E)$ given by $V = \{1\}$. Then, necessarily $E = \emptyset$ and ω_E is redundant. We will consider the product graph $\mathcal{G}_\otimes$ of $\mathcal{G}_1$ with $\mathcal{G}_2 = (W, F, \lambda_W, \omega_F)$ where $W = \{2, 3\}$, $F = \{\{2, 3\}\}$ and $\omega_F(\{2, 3\}) = \frac{\varepsilon}{2}$ for some $\varepsilon > 0$. Moreover, we assume that $\lambda_V(1) = \lambda_W(2) = \lambda_W(3)$. Thus, the vertex set $V_\otimes$ of the product graph is given by $V_\otimes = \{(1, 2), (1, 3)\}$. Let us define that two edge weights are approximately the same if the absolute value of their difference is less than ε. Then, due to our definition of the edge set $E_\otimes$ the two vertices from $V_\otimes$ are not adjacent as they coincide in their first component. Without the condition $v_1 \neq v_2$ and $w_1 \neq w_2$ and allowing loops of zero weight in the graph model, however, these two vertices would have been adjacent. Obviously, in this case the resulting clique $\mathcal{C} = V_\otimes$ would not be related to an approximate common subgraph of $\mathcal{G}_1$ and $\mathcal{G}_2$.

Overview. Let us now formulate the overall graph theoretic linkage attack.

Graph Theoretic Linkage Attack on Microdata in a Metric Space

INPUT Target data (T_1, D_1), identification data (T_2, D_2)
OUTPUT List of matches between records from T_1 and T_2

1. Build target graph $\mathcal{G}_1$ from (T_1, D_1).
2. Build identification graph $\mathcal{G}_2$ from (T_2, D_2) (possible under assumptions 3 and 4).
3. Build product graph $\mathcal{G}_\otimes$ (requires reasonable definition of $\approx$).
4. Find a maximum clique $\mathcal{C}_{\max}$ in $\mathcal{G}_\otimes$ (using some maximum clique detection algorithm).
5. Extract matches from $\mathcal{C}_{\max}$.

Let us make a short comment on step 4 of the attack: As indicated already in section 2 there is a vast literature concerning the problem of maximum clique detection in graphs. A systematic comparison of prevalent techniques to tackle this problem regarding our application goes beyond the scope of this paper and is postponed to future research.
To conclude this section we illustrate the process of the proposed graph theoretic linkage attack by a small-scale example which makes use of the data placed together in appendix A.

Example 10. Consider the microdata table 8 in appendix A which contains information about some important european poets. This table is anonymized by removing the direct identifier `name`, generalising the attribute `yob` (year of birth) to `cob` (century of birth) and removing the information about the birth location (`loc`). The attribute `language` remains unchanged. This yields the anonymized table 9 in appendix A. Whereas the spatial information `loc` has been deleted from this table, the distance matrix D_1 (see appendix A) containing the geographic distances between the birth locations is meant to be published in addition to table 9. We assume that the snooper is in possession of the identification microdata in table 10, i.e. the attributes `cob` and `language` serve as quasi-identifiers. By geocoding the birth locations and calculation of the geographic distances the snooper obtains the distance matrix D_2. The graph models $\mathcal{G}_1$ and $\mathcal{G}_2$ for target and identification data can be build using this information and are visualised in figure 2. Table 2 lists all possible matches when the snooper takes only the vertex labels into consideration. These eleven matches form the vertex set of the product graph as well. Note that this set would already constitute the final outcome when performing a linkage attack not taking the distances into consideration.
For the construction of the product graph we allow an absolute deviation of five kilometers with respect to the edge weights, i.e. we define $\omega_E(v_1v_2) \approx \omega_F(w_1w_2) \Leftrightarrow |\omega_E(v_1v_2) - \omega_F(w_1w_2)| < 5$.[1] This definition of $\approx$ leads to the product graph shown in figure 3.
As can be seen easily from figure 3 the product graph contains a unique maximum clique $\mathcal{C} := \{1, 2, 3, 5\}$. Therefore a snooper following the protocol of the graph theoretic linkage attack will accept the potential matches in rows 1,2,3 and 5 in table 2 as matches and reject the remaining ones.

Although example 10 is artificial it illustrates some phenomena that will also appear when one considers real-world data:

[1] We mentioned previously that allowing such a deviation is already necessary because of errors appearing due to the fact that dataholder and snooper will in general use different methods for geocoding and distance computation. This fact was also addressed to in this example by geocoding the birth locations of the target microdata via Wikipedia and the birth locations of the identification file by means of the command `geocode` provided by the R package `ggmap`.

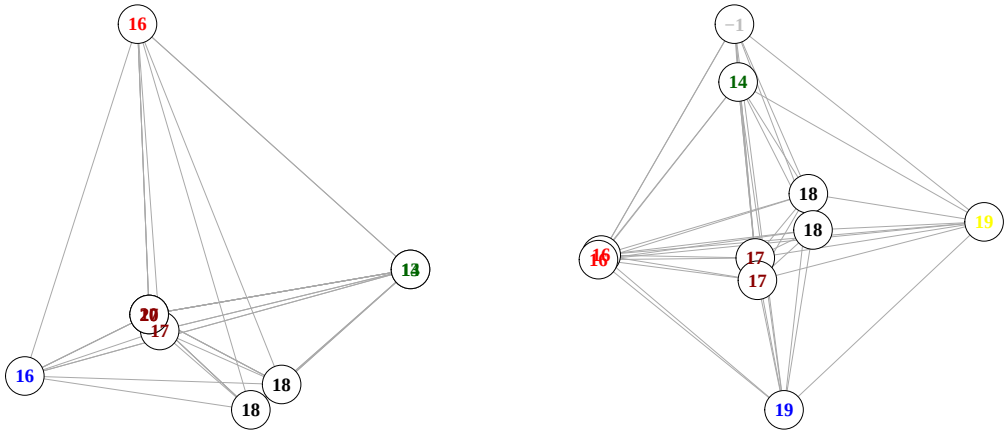

Figure 2: Graph models for target (left) and identification (right) microdata in example 10. Layout of graphs was chosen such that edge lengths indicate the pseudodistances approximately. The attribute `language` is indicated by the vertex label colour, whereas the attribute `cob` by the vertex label itself.

vertex of product graph	rownumber target file	rownumber identification file
1	1	1
2	2	2
3	3	3
4	6	3
5	4	4
6	7	4
7	3	6
8	6	6
9	4	7
10	7	7
11	2	9

Table 2: Possible matches between tables 9 and 10 with respect to the quasi-identifiers `cob` and `language` only, i.e. vertex labels in the accompanying graph models.

- The definition of $\approx$ has to be chosen reasonably. In the present example distances between cities scattered over the whole European continent are considered so that even the rather coarse definition above (allowing an absolute deviation of 5 kilometers) will yield a useful result. In general, the definition of $\approx$ has to be chosen such that as much as possible common edges of target and identification graph are detected correctly without classifying too many edges as approximately the same that are actually different. The definition of $\approx$ will be studied in greater detail as part of the simulation study in section 5.

- A successful match of the respectively first records of both tables would have been possible unambiguously already without the additional distance information because both records are unique in their tables with respect to the corresponding quasi-identifiers. Nevertheless, using the additional distance information increases the

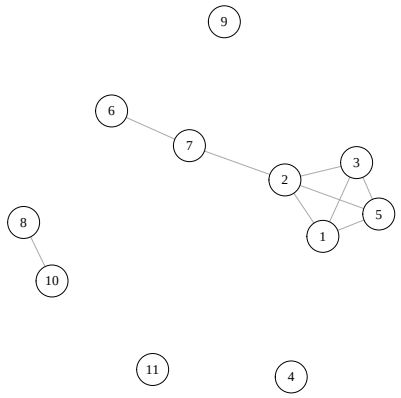

Figure 3: Product graph in example 10.

credibility regarding this specific matching which is now not only supported through the coincidence of the quasi-identifiers but also through coincidence of distances to other matches.

- However, in certain cases unambiguous matching will only be possible due to the additional information about distances. For example, record 3 of the target table could be matched with records 3 and 6 of the identification table when only taking quasi-identifiers into consideration. This tie is resolved in our example through the extra information given by the edge weights.

- Evidently, in practise there will be ties in the data that cannot be resolved by our method either. In our example the records 9 and 10 of the target file do not differ according to their quasi-identifiers, but cannot be distinguished considering distances to these records as well because the corresponding point locations (`loc`=Paris in both cases) coincide.

- Last, but not least, the attack has reduced the number of matches from eleven in table 2 to four. These matches indeed correspond to the actual overlap of target and identification file.

Our toy example has shown that publishing inter-record distances might increase the risk of identity disclosure for microdata files. We will confirm this result in the following section by investigating the effect of random noise addition to the input coordinates which is a standard technique for the anonymization of spatial point data.

5 Experimental Results

Data. For the simulation study data were generated as follows: In a first step addresses from the German telephone book were sampled at random. Then, geographic latitude and longitude with respect to the World Geodetic System 1984 were assigned to these addresses using the `geocode` command from the R package `ggmap` [19]. Finally, geographic distances between the addresses were calculated to obtain the corresponding distance matrix. We

randomly assigned the points of the resulting metric spaces to example microdata which contained (besides an ID) attributes concerning gender and age which served as quasi-identifiers in our experiments. The attribute values were sampled in accordance with the actual distribution of these attributes due to the demographic statistics derived during the German zensus 2011.[2] Note that the classification with respect to age (eleven age intervalls) is rather coarse which guaranteed the existence of duplicates with respect to the quasi-identifiers in our test microdata and would lead to many ties when performing a classical linkage attack. We generated data for different values of common records N_{common} of target and identication file and different sizes N_1 and N_2 of the metric spaces.

Perturbation technique. A simple technique for the anonymization of spatial point data consists in the addition of random noise to their coordinates (see, e.g., section 3.2 in [1]). In this section, we consider the performance of the proposed graph theoretical linkage attack in dependence on this anonymization technique. To be more precisely, $\mathcal{N}(0, \sigma^2)$-distributed Gaussian noise was added to the input coordinates of the target file before the distance matrix was calculated. Different instances of the standard deviation σ were considered.

Fine-tuning of the attack. A suitable definition for the approximative relation $\approx$ has to be found for the generation of the product graph in the graph theoretical linkage attack. Following Kerckhoffs' principle [27] (which implies that the security of a cryptosystem/anonymization technique must not depend on the concealment of the algorithm in use), we assume that the data snooper knows that Gaussian noise is added to the geographic coordinates before distances are calculated, and furthermore that the standard deviation σ is known to him (the latter assumption is in conformance with [1] where it is emphasized that *all useful spatial analyses of masked data require some knowledge about the characteristics of the mask used*). Under the assumption of an euclidean distance function the effect of random perturbation of the input coordinates on the squared distances could be studied theoretically, an approach which has been considered in [20], for example. Such a rigorous mathematical analysis appears to be more difficult in the case of geographical distances. For this reason, we assume that the snooper performs a little simulation study by which she/he investigates the effect of perturbation by Gaussian noise to the calculation of distances. To imitate this course of action, we sampled 1000 pairs of points from the area of the Federal Republic of Germany for each considered value of σ and compared the distances before and after addition of Gaussian noise. Several sample quantiles and the sample variance (the latter will only be used for the evaluation of our experiments) of the deviation of distances (which is defined as $d - d'$ where d denotes the original distance and d' the distance after perturbation) have been gathered and are recorded in table 3. We use the empirical quantiles to define the interpretation of $\approx$: For a threshold parameter $\alpha \in (0, 1)$ we define that two edge weights satisfy the relation $\approx$, if the corresponding deviation is greater than the empirical $\frac{1-\alpha}{2}$-quantile and smaller than the $\frac{1+\alpha}{2}$-quantile for the current value of σ. Here, the distances from the identification file take on the role of d and the distances from the target file the one of d'. The threshold parameter α chosen by the snooper is supposed to guarantee that a common edge of target and identification graph is detected by the snooper with probability equal to α. Its effect will also be considered within this section.

[2]These demographic statistics can be downloaded from `https://ergebnisse.zensus2011.de/auswertungsdb/download?pdf=00&tableId=BEV_1_1_1&locale=DE`.

σ	0.05	0.1	0.25	0.5	0.75	0.9	0.95	sample variance
0.0	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
0.01	-2.1185	-1.6547	-0.8877	-0.0041	0.8742	1.6732	2.3365	1.7613
0.02	-4.4577	-3.5257	-1.9054	-0.0889	1.7744	3.2325	4.0726	7.0433
0.03	-6.7860	-5.2663	-2.7144	0.0031	2.6000	4.6025	6.1992	16.2465
0.04	-9.3577	-6.5848	-3.4838	-0.0493	3.6967	7.0553	8.8865	30.2513
0.05	-11.5771	-8.9097	-4.6609	0.0990	4.1113	8.5682	11.3509	47.8643
0.06	-11.9136	-9.9088	-5.3549	-0.4764	5.1601	10.6607	13.7124	64.3454
0.07	-15.1725	-11.6241	-6.7557	-0.5914	5.4345	11.2392	15.4473	84.8152
0.08	-17.0787	-13.0403	-7.4898	-0.5243	7.0405	13.6658	17.9175	117.1461
0.09	-20.1859	-15.3711	-7.8677	-0.2037	8.2208	15.2992	20.1929	146.4311
0.10	-25.8312	-18.8113	-9.4552	-0.9983	7.5855	15.6075	21.4478	194.2151

Table 3: Sample quantiles and variance of the considered distance deviation $d - d'$ for different values of σ.

Implementation. All experiments reported here and the accompanying visualisations were performed in R using the `igraph` package [9]. In particular, we made use of the (maximum) clique detection algorithm provided by this package which turned out to be sufficient for the (relatively small) filesizes considered. When the maximum clique was not uniquely determined, the set theoretic union of all maximum cliques was taken as the result of the matching process.

Evaluation of the attack. The matches and non-matches between target and identification file gathered by our graph theoretical linkage attack were classified as true positives (successful deanonymization), false positives (failed deanonymization), false negatives (records belonging to the same entity have been missed) and true negatives (records have been correctly classified as belonging to distinct entities). The quality measures considered are based on the number of true positives (**TP**), false positives (**FP**) and false negatives (**FN**). More precisely, we consider

$$\text{prec} = \frac{\mathbf{TP}}{\mathbf{TP} + \mathbf{FP}} \quad \textit{(precision)}, \quad \text{and}$$
$$\text{rec} = \frac{\mathbf{TP}}{\mathbf{TP} + \mathbf{FN}} \quad \textit{(recall)},$$

which are two standard measures in the evaluation of data linkage processes [8].

Simulation design and results. In a first experiment the sizes N_1 and N_2 of target and identification file were kept fixed and equal to 100. We varied the size N_{common} of the overlap between both files as well as the noise parameter σ and the threshold α. For each parameter setup the simulation was repeated $n = 100$ times. The mean of precision and recall over all iterations for some parameter setups can be found in tables 4 and 5. Visualisations of the results belonging to this first experiment can be found in figures 4, 5 and 6. In a second experiment also a second value ($N_1, N_2 = 300$) for the sizes of target and identification file was considered. In this second experiment the simulation was repeated 50 times for each parameter combination. The respective averages of precision and recall are recorded in tables 6 and 7, see figures 7 and 8 for visualisations.

N_{common}	$\alpha \mid \sigma$	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	0.10
10	0.5	0.9013	0.8593	0.7508	0.6494	0.5621	0.4704	0.3772	0.2815	0.2368	0.1899
	0.9	0.9120	0.9076	0.9049	0.8903	0.8631	0.8185	0.7637	0.6805	0.6177	0.5295
25	0.5	1.0000	0.9954	0.9814	0.9718	0.9568	0.9522	0.9375	0.9303	0.9246	0.9089
	0.9	1.0000	0.9908	0.9745	0.9588	0.9456	0.9337	0.9216	0.9068	0.8943	0.8858
50	0.5	1.0000	0.9992	0.9973	0.9866	0.9766	0.9660	0.9563	0.9470	0.9422	0.9258
	0.9	1.0000	0.9984	0.9927	0.9765	0.9622	0.9514	0.9417	0.9322	0.9204	0.9113
100	0.5	1.0000	1.0000	1.0000	0.9997	0.9980	0.9970	0.9936	0.9909	0.9877	0.9814
	0.9	1.0000	1.0000	1.0000	0.9992	0.9976	0.9948	0.9915	0.9853	0.9816	0.9768

Table 4: **Average precision** in dependence on simulation parameters σ, α and N_{common} over $n = 100$ repetitions. Number of records of target and identification file was equal to 100.

N_{common}	$\alpha \mid \sigma$	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	0.10
10	0.5	0.6100	0.6210	0.5780	0.5760	0.5510	0.4950	0.4100	0.4110	0.3910	0.3060
	0.9	0.8870	0.8850	0.8920	0.8940	0.8980	0.8710	0.8430	0.8330	0.7910	0.7500
25	0.5	0.4464	0.4724	0.4540	0.4496	0.4472	0.4468	0.4560	0.4680	0.4708	0.4492
	0.9	0.8000	0.7996	0.8112	0.8184	0.8044	0.8020	0.8020	0.7884	0.8016	0.7916
50	0.5	0.3626	0.3806	0.3508	0.3796	0.3784	0.3686	0.3708	0.3488	0.3512	0.3466
	0.9	0.7404	0.7414	0.7442	0.7726	0.7578	0.7498	0.7476	0.7474	0.7538	0.7268
100	0.5	0.2984	0.2947	0.2757	0.3053	0.3011	0.2966	0.2888	0.3007	0.2920	0.2738
	0.9	0.6867	0.6918	0.6966	0.7186	0.7136	0.7048	0.7050	0.6977	0.7106	0.6884

Table 5: **Average recall** in dependence on simulation parameters σ, α and N_{common} over $n = 100$ repetitions. Number of records of target and identification file was equal 100.

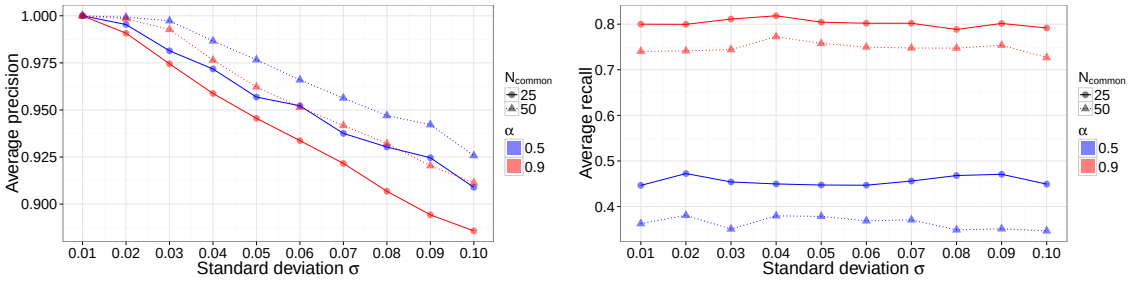

Figure 4: Dependence of average precision and recall on the standard deviation σ for different values of N_{common} and α in the first experiment (cf. tables 4 and 5).

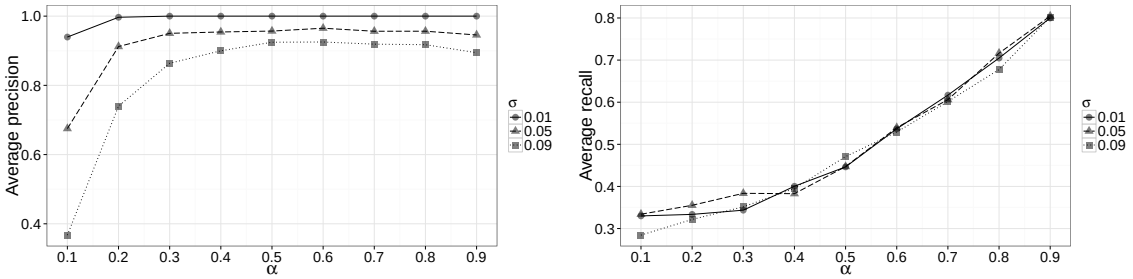

Figure 5: Dependence of average precision and recall on the threshold parameter α for different values of σ in the first experiment (cf. tables 4 and 5). $N_1 = N_2 = 100$. $N_{\text{common}} = 50$.

	N_1	100		300	
N_{common}	σ | N_2	100	300	100	300
10	0.01	0.9188	0.7971	0.8281	0.5062
	0.02	0.8873	0.6765	0.6441	0.3602
	0.03	0.7808	0.4443	0.5532	0.2740
	0.04	0.6743	0.3684	0.4447	0.2738
	0.05	0.5913	0.1980	0.4144	0.2328
	0.06	0.4973	0.1598	0.3359	0.1911
	0.07	0.4285	0.1710	0.2974	0.1818
	0.08	0.3130	0.1202	0.2523	0.1426
	0.09	0.2664	0.0917	0.2363	0.1163
	0.10	0.2107	0.0735	0.2191	0.1086
25	0.01	1.0000	0.9980	0.9721	0.9370
	0.02	0.9947	0.9737	0.9375	0.8611
	0.03	0.9724	0.9291	0.9093	0.7668
	0.04	0.9673	0.8704	0.8812	0.6827
	0.05	0.9515	0.8029	0.8479	0.5923
	0.06	0.9544	0.7343	0.8036	0.5638
	0.07	0.9390	0.6903	0.7238	0.4952
	0.08	0.9368	0.6508	0.6641	0.4425
	0.09	0.9269	0.6391	0.5925	0.3832
	0.10	0.9165	0.6091	0.5801	0.3123
50	0.01	1.0000	0.9753	0.9838	0.9801
	0.02	0.9985	0.9643	0.9772	0.9406
	0.03	0.9980	0.9564	0.9571	0.8936
	0.04	0.9858	0.9380	0.9397	0.8356
	0.05	0.9754	0.9335	0.9290	0.7736
	0.06	0.9645	0.9178	0.9116	0.7082
	0.07	0.9582	0.9010	0.8827	0.6759
	0.08	0.9496	0.8895	0.8665	0.6385
	0.09	0.9434	0.8666	0.8366	0.6015
	0.10	0.9281	0.8368	0.8280	0.5658
100	0.01	1.0000	1.0000	1.0000	0.9922
	0.02	1.0000	1.0000	0.9985	0.9839
	0.03	1.0000	0.9986	0.9934	0.9634
	0.04	0.9994	0.9957	0.9817	0.9398
	0.05	0.9979	0.9913	0.9754	0.9100
	0.06	0.9965	0.9881	0.9691	0.8837
	0.07	0.9927	0.9845	0.9618	0.8535
	0.08	0.9877	0.9760	0.9557	0.8168
	0.09	0.9835	0.9741	0.9454	0.7944
	0.10	0.9745	0.9601	0.9411	0.7694

Table 6: **Average precision** in dependence on simulation parameters N_1, N_2, N_{common} and σ over $n = 50$ repetitions. The threshold parameter α was kept fixed and equal to 0.5.

	N_1	100		300	
N_{common}	σ | N_2	100	300	100	300
10	0.01	0.8900	0.8820	0.8660	0.8160
	0.02	0.8900	0.8840	0.8600	0.8200
	0.03	0.9000	0.8580	0.8300	0.7760
	0.04	0.8960	0.8340	0.8220	0.7540
	0.05	0.9120	0.7840	0.7960	0.7200
	0.06	0.8800	0.7380	0.8040	0.6660
	0.07	0.8560	0.6540	0.7680	0.6840
	0.08	0.8500	0.6320	0.7680	0.6760
	0.09	0.8600	0.5240	0.7860	0.6680
	0.10	0.7840	0.4800	0.7420	0.6540
25	0.01	0.8000	0.7952	0.7864	0.7952
	0.02	0.8016	0.7920	0.7920	0.7784
	0.03	0.8096	0.7920	0.7872	0.7824
	0.04	0.8088	0.8088	0.8056	0.7624
	0.05	0.8088	0.7936	0.7912	0.7448
	0.06	0.7992	0.7896	0.7808	0.7232
	0.07	0.8008	0.7832	0.7840	0.7112
	0.08	0.7984	0.7600	0.7568	0.6640
	0.09	0.7960	0.7752	0.7536	0.6656
	0.10	0.7904	0.7512	0.7152	0.6656
50	0.01	0.7384	0.7272	0.7580	0.7324
	0.02	0.7360	0.7348	0.7512	0.7304
	0.03	0.7448	0.7368	0.7572	0.7320
	0.04	0.7660	0.7668	0.7752	0.7488
	0.05	0.7592	0.7552	0.7676	0.7364
	0.06	0.7424	0.7444	0.7508	0.7192
	0.07	0.7532	0.7456	0.7444	0.7116
	0.08	0.7464	0.7348	0.7360	0.6928
	0.09	0.7512	0.7480	0.7540	0.7100
	0.10	0.7264	0.7180	0.7204	0.6712
100	0.01	0.6826	0.6798	0.6904	0.6872
	0.02	0.6874	0.6842	0.6836	0.6858
	0.03	0.6908	0.7034	0.6946	0.6948
	0.04	0.7170	0.7274	0.7188	0.7144
	0.05	0.7174	0.7190	0.7100	0.6956
	0.06	0.7010	0.7012	0.6938	0.6876
	0.07	0.7010	0.7164	0.6922	0.6890
	0.08	0.6972	0.6944	0.6910	0.6766
	0.09	0.7088	0.7092	0.7026	0.6944
	0.10	0.6870	0.6824	0.6784	0.6646

Table 7: **Average recall** in dependence on simulation parameters N_1, N_2, N_{common} and σ over $n = 50$ repetitions. The threshold parameter α was kept fixed and equal to 0.5.

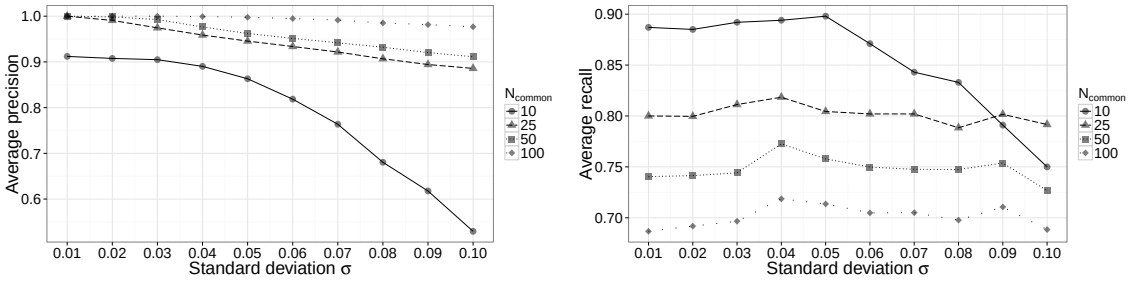

Figure 6: Dependence of average precision and recall on the standard deviation σ for different values of N_{common} in the first experiment (cf. tables 4 and 5). $\alpha = 0.9$.

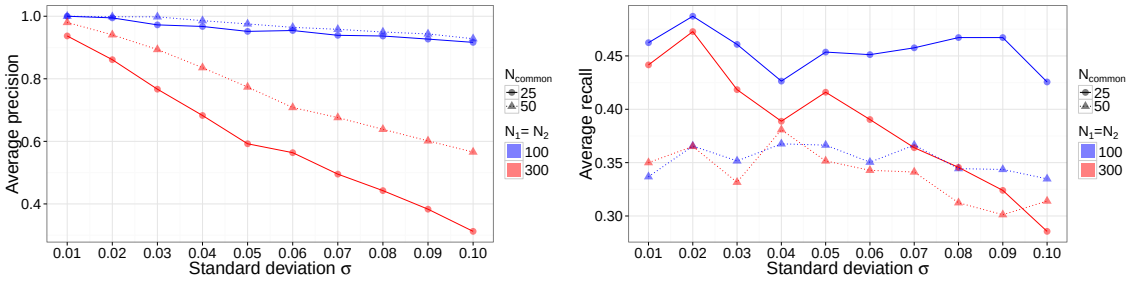

Figure 7: Dependence of average precision and recall on the standard deviation σ for different values of N_{common} and $N_1 = N_2$ in the second experiment (cf. tables 6 and 7). $\alpha = 0.5$.

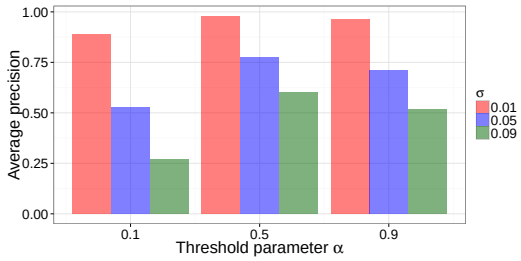

Figure 8: Dependence of average precision on the threshold parameter α for different values of σ in the second experiment. $N_1 = N_2 = 300$. $N_{\text{common}} = 100$.

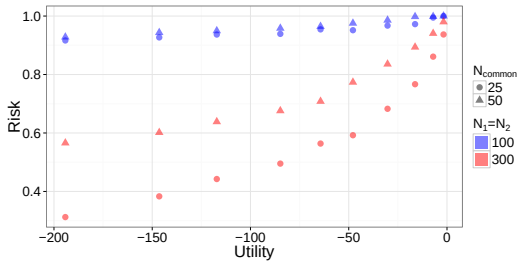

Figure 9: R-U confidentiality map for some parameter setups in the second experiment. The risk is measured by the average precision whereas the utility by the negative of the sample variance in table 3. $\alpha = 0.5$.

Discussion. The main effect of the threshold parameter α concerns the recall which becomes evident by rethinking that α is designed so that the probability of detecting a common edge of target and identification graph is approximately equal to α. For this reason, higher values of α lead to a higher recall (cf. table 5 and figures 4 and 5). Simultaneously, the effect of α on the precision appears to be twofold: On the one hand, for too high values of α the precision should decrease, because the chance for non-common edges of target and identification graph (but which coincide with respect to the vertex labels of their endpoints) to be classified as common edges increases. This would reflect a trade-off between precision and recall which is a well-known phenomenon in data linkage in general [8]. However, this trade-off is only of limited validity as, on the other hand,

for increasing α also a larger portion of the overlap between identification and target file can be successfully detected by the snooper which makes false positives (leading to less precision) less likely. Thus, combining these two thoughts, for increasing α the precision should initially also increase, but later on decrease when α is getting too large. This expectation is confirmed by our experiments (cf. table 4 and figures 5 and 8) although the decrease of precision for α getting too large can only be anticipated which might amongst others be explained by the small file sizes considered. For filesizes $N_1 = N_2 = 100$ a rapid growth of the precision is observable when α increases from 0.1 to 0.4 and the precision does not vary severely for α between 0.4 the and 0.8. For α getting larger than 0.8 only a slight decrease of the precision is recorded (cf. figure 5). This effect can be observed more clearly when considering larger filesizes $N_1 = N_2 = 300$ in the second experiment (cf. figure 8).
From the definition of $\approx$ (see the paragraph *Fine-tuning of the attack* above) it should be supposed that the recall is more or less independent of σ as the probability of correctly detecting an edge should be nearly α (which is independent of σ). This non-dependence is impressively confirmed by the performed simulations and illustrated in figures 4 and 5. However, σ strongly influences the precision (for larger values of σ the precision evidently decreases): The data snooper has to accept false positives (resulting in less precision) if she/he wants to achieve a certain predetermined recall. For small values of σ the sizes of target and identification file have only a small effect on the recall (cf. table 7 and figure 7). This phenomenon can be explained analogously to the non-existing dependence of σ on the recall: The filesize has no influence on the chance of a common edge of target and identification file to be matched. As the file sizes of both target and identification file increase, the precision decreases (cf. table 6) because there will appear more instances of actually distinct edges that are matched which makes false positives more likely. However, if the snooper would reduce the number of records of the identification file (e.g. by subsampling) to overcome this effect he would simultaneously reduce the number of common entities of target and identification graph which would also lead to a decrease of precision (cf. table 6 and figure 7).
Altogether the simulations show that, in principle, a sufficient level of anonymity can be achieved by addition of random noise to the input coordinates before computing the distance matrix. However, this anonymity is not for free which is illustrated by means of the risk-utility (R-U) confidentiality map in figure 9 (which is for its part based on figure 7) where the risk of identity disclosure (measured by the average precision of the linkage attack; cf. also the discussion below) is plotted against the utility (measured as the negative of the sample variance of the distance deviation $d - d'$ for the current value of σ as recorded in table 3). In the datasets considered in the simulation study σ would have to be chosen so large to guarantee at least some amount of anonymity that a useful analysis based on the distances would not be possible any more. For this reason the development of distance modification techniques that guarantee anonymity but also make useful analysis on the anonymized data possible will be an important aspect of future research.
Note that in our specific example the snooper would primarily attempt to achieve high precision: In the case of geographic distances a point is uniquely determined by the exact distances to three other points. If the snooper could deanonymize at least three entities, exploiting this fact would be a good starting point to identify even more individuals. For arbitrary metric spaces such a relationship does not hold in general albeit the successful deanonymization of some entities will alleviate a snooper's work also in this more general case. Obviously, for other techniques used for distance modification than perturbation of the input coordinates a snooper will have to modify the graph theoretical linkage at-

tack, especially the definition of $\approx$. However, due to Kerckhoffs' principle it has to be assumed that the snooper knows at least the distance modification technique used by the holder of the target file and exploits this knowledge in the precise construction of the attack. For instance, if noise is not added to the input coordinates before computing the distance matrix but rather on the distance matrix itself (a technique discussed e.g. in [20]), the attack has to be slightly adapted. In this case, when defining the relation $\approx$ the quantiles of the noise distribution can be used directly making the empirical study on distance deviations originating from perturbation of the input coordinates needless. Moreover, in this specific case it might be reasonable to further modify the attack by relaxing the (relatively strong) notion of a maximum clique to the less restrictive notion of a maximum quasi-clique, a relaxation which has been successfully applied in [14] for the purpose of protein classification. In a similar way, our attack can be adapted to many other anonymization techniques and thus provides a useful tool for the analysis of methods for distance-preserving anonymization.

6 Conclusion

In this article we have introduced a novel graph theoretic linkage attack on microdata with additionally published (approximate) inter-record distances. In the special case of spatial distances, we have demonstrated – by means of simulated data – that the release of distances increases the risk of identity disclosure unreasonably even if geographical coordinates have been perturbed by random Gaussian noise before distances are calculated. Furthermore we showed that augmenting the standard deviation of the added random noise gradually will lead to a sufficient level of anonymity, but also make the perturbed distances useless for further analysis.
Our approach motivates attractive questions for future research: First of all, the development and analysis of anonymization techniques for microdata in a metric space that distort the distances as less as possible (particularly with regard to the applicability of data mining techniques) is important. Further, in order to empirically evaluate such techniques with test data of larger sizes, a further development of the proposed attack appears necessary. To be more precise, the choice of an appropriate approximative algorithm for detecting sufficiently large cliques in step 4 of the attack will be part of future research since the exact maximum clique detection is the computationally most expensive step of the proposed attack

Acknowledgements

This research has been supported by a grant from the German Research Foundation (DFG) to Rainer Schnell. I am grateful to Rainer Schnell for valuable suggestions that helped to improve the presentation of the paper.

References

[1] M.P. Armstrong, G. Rushton and D.L. Zimmerman (1999) Geographically masking health data to preserve confidentiality. *Statistics in Medicine* 18:497–525.

[2] L. Backstrom, C. Dwork and J. Kleinberg (2007) Wherefore art thou r3579x?: Anonymized social networks, hidden patterns, and structural steganography. *ACM Proceedings of the 16th International Conference on World Wide Web*, 181–190.

[3] I.M. Bomze, M. Budinich, P.M. Pardalos and M. Pelillo (1999) The maximum clique problem. In: D.-Z. Du and P. Pardalos (eds.): *Handbook of combinatorial optimization*, 1–74.

[4] J.S. Brownstein, C.A. Cassa, I.S. Kohane and K.D. Mandl (2006) An unsupervised classification method for inferring original case locations from low-resolution disease maps. *International Journal of Health Geographics* 5:56.

[5] H. Bunke and K. Riesen (2012) Towards the unification of structural and statistical pattern recognition. *Pattern Recognition Letters* 33:811–825.

[6] S. Chester, B.M. Kapron, G. Srivastava and S. Venkatesh (2013) Complexity of social network anonymization. *Social Network Analysis and Mining* 3:151–166.

[7] D. Conte, P. Foggia, C. Sansone and M. Vento (2004) Thirty years of graph matching in pattern recognition. *International Journal of Pattern Recognition and Artificial Intelligence* 18:265–298.

[8] P. Christen and K. Goiser (2007) Quality and Complexity Measures for Data Linkage and Deduplication. In: F. Guillet and H.J. Hamilton (eds.): *Quality Measures in Data Mining*, 127–151.

[9] G. Csardi and T. Nepusz (2006) The igraph software package for complex network research. *InterJournal, Complex Systems* 1695(5). `http://igraph.sf.net`.

[10] A.J. Curtis, J.W. Mills and M. Leitner (2006) Spatial confidentiality and GIS: Re-engineering mortality locations from published maps about hurricane Katrina. *International Journal of Health Geographics* 5:44.

[11] T. Dalenius (1986) Finding a needle in a haystack – or identifying anonymous census records. *Journal of Official Statistics* 2(3):329–336.

[12] M.M. Deza and E. Deza (2009) Encyclopedia of distances. *Springer, New York*.

[13] G.T. Duncan, M. Elliot and J.-J. Salazar-González (2011) Statistical confidentiality: Principles and practise. *Springer, New York*.

[14] T. Fober, G. Klebe and E. Hüllermeier (2013) Local Clique Merging: An extension of the maximum common subgraph measure with applications in structural bioinformatics. In: B. Lausen, D. van den Poel and A. Utsch (eds.): *Algorithms from and for Nature and Life*, 279–286.

[15] M.R. Garey and D.S. Johnson (1979) Computers and intractability: A guide to the theory of NP-completeness. *Freeman, New York*.

[16] M.P. Gutmann and P. C. Stern (2007) Putting people on the map: Protecting confidentiality with linked social-spatial data. *National Academies Press, Washington, D.C.*

[17] M.P. Gutmann, K. Witkowski, C. Colyer, J.M. O'Rourke and J. McNally (2008) Providing spatial data for secondary analysis: Issues and current practises relating to confidentiality. *Population research and policy review* 27:639–665

[18] A. Hundepool, J. Domingo-Ferrer, L. Franconi, S. Giessing, E. Schulte Nordholt, K. Spicer and P.-P. de Wolf (2012) Statistical Disclosure Control. *Wiley series in survey methodology.*

[19] D. Kahle and H. Wickham (2013) ggmap: A package for spatial visualization with Google Maps and OpenStreetMap. R package version 2.3. `http://CRAN.R-project.org/package=ggmap`.

[20] K. Kenthapadi, A. Korolova, I. Mironov and N. Mishra (2013) Privacy via the Johnson-Lindenstrauss transform. *Journal of Privacy and Confidentiality* 5(1):39–71.

[21] J. Krumm (2009) A survey of computational location privacy. *Personal and Ubiquitous Computing* 13(6):391–399.

[22] G. Levi (1973) A note on the derivation of maximal common subgraphs of two directed or undirected graphs. *Calcolo* 9(4):341–352.

[23] K. Liu, C. Giannella and H. Kargupta (2006) An attacker's view of distance preserving maps for privacy preserving data mining. In: J. Fürnkranz, T. Scheffer and M. Spiliopoulou (eds.): *Knowledge Discovery in Databases: 10th European Conference on Principles and Practice of Knowledge Discovery in Databases*, 297–308.

[24] M. Merener (2012) Theoretical results on de-anonymization via linkage attacks. *Transactions on Data Privacy* 5(2):377–402.

[25] C. M. O'Keefe (2012) Confidentialising maps of mixed point and diffuse spatial data. In: J. Domingo-Ferrer and I. Tinnirello (eds.): *Privacy in statistical databases*, 226–240.

[26] R.N. Parker and E.K. Asencio (2008) GIS and spatial analysis for the social sciences: Coding, mapping, and modeling *Routledge, New York.*

[27] F.A.P. Petitcolas (2011) Kerckhoffs' principle. In: H.C.A. van Tilborg and S. Jajodie (eds.): *Encyclopedia of cryptography and security*, 675.

[28] S. Rane, W. Sun and A. Vetro (2010) Privacy-preserving approximation of L1 distance for multimedia applications. *IEEE International Conference on Multimedia and Expo (ICME)*, 492–497.

[29] P. Samarati and L. Sweeney (1998) Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. *Technical report, SRI International.*

[30] J. Snow (1855) On the Mode of Communication of Cholera. *John Churchill.*

[31] L. Sweeney (2000) Uniqueness of simple demographics in the US population. *Technical report.*

[32] L. Sweeney (2002) k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 10:557–570.

[33] S.C. Wieland, C.A. Cassa, K.D. Mandl and B. Berger (2008) Revealing the spatial distribution of a disease while preserving privacy. *Proceedings of the National Academy of Sciences* 105:17608–17613.

[34] E. Zheleva and L. Getoor (2011) Privacy in social networks: A survey. In: C.C. Aggarwal (ed.): *Social Network Data Analytics*, 277–306.

A Example Dataset: European Poets

	name	yob	language	loc
1	Giovanni Boccaccio	1313	italian	Firenze
2	Miguel de Cervantes	1547	spanish	Alcala de Henares
3	Johann Wolfgang Goethe	1749	german	Frankfurt am Main
4	Moliere	1622	french	Paris
5	Dante Alighieri	1265	italian	Firenze
6	Friedrich Schiller	1759	german	Marbach am Neckar
7	Jean-Baptiste Racine	1637	french	La Ferte-Milon
8	William Shakespeare	1564	english	Stratford-upon-Avon
9	Simone de Beauvoir	1908	french	Paris
10	Jean-Paul Sartre	1905	french	Paris

Table 8: Microdata containing information about famous european poets. Attribute `yob` contains the year of birth, `loc` the birth location of the poets.

	cob	language
1	14	italian
2	16	spanish
3	18	german
4	17	french
5	13	italian
6	18	german
7	17	french
8	16	english
9	20	french
10	20	french

Table 9: Anonymized version of table 8 obtained by removing the direct identifier `name`, generalizing year of birth (`yob`) to century of birth (`cob`) and removing the birth location (`loc`).

The distances between birth locations `loc` are stored in the distance matrix D_1.

$$D_1 = \begin{pmatrix} 0 & 1261 & 729 & 886 & 0 & 593 & 864 & 1341 & 886 & 886 \\ 1261 & 0 & 1424 & 1034 & 1261 & 1369 & 1093 & 1307 & 1034 & 1034 \\ 729 & 1424 & 0 & 479 & 729 & 137 & 414 & 762 & 479 & 479 \\ 886 & 1034 & 479 & 0 & 886 & 507 & 67 & 469 & 0 & 0 \\ 0 & 1261 & 729 & 886 & 0 & 593 & 864 & 1341 & 886 & 886 \\ 593 & 1369 & 137 & 507 & 593 & 0 & 449 & 856 & 507 & 507 \\ 864 & 1093 & 414 & 67 & 864 & 449 & 0 & 478 & 67 & 67 \\ 1341 & 1307 & 762 & 469 & 1341 & 856 & 478 & 0 & 469 & 469 \\ 886 & 1034 & 479 & 0 & 886 & 507 & 67 & 469 & 0 & 0 \\ 886 & 1034 & 479 & 0 & 886 & 507 & 67 & 469 & 0 & 0 \end{pmatrix}$$

	name	cob	language	loc
1	Giovanni Boccaccio	14	italian	Firenze
2	Miguel de Cervantes	16	spanish	Alcala de Henares
3	Johann Wolfgang Goethe	18	german	Frankfurt am Main
4	Moliere	17	french	Paris
5	James Joyce	19	english	Dublin
6	Heinrich Heine	18	german	Duesseldorf
7	Pierre Corneille	17	french	Rouen
8	Publius Ovidius Naso	-1	latin	Sulmona
9	Lope de Vega	16	spanish	Madrid
10	August Strindberg	19	swedish	Stockholm

Table 10: Identification microdata table which is used by the data snooper in example 10.

Geocoding of the locations from table 10 using the R package `ggmap` and calculation of the mutual distances via the command `spDists` from the package `sp` yields the distance matrix D_2.

$$D_2 = \begin{pmatrix} 0 & 1260 & 731 & 887 & 1666 & 894 & 999 & 291 & 1290 & 1791 \\ 1260 & 0 & 1423 & 1033 & 1446 & 1427 & 1055 & 1457 & 30 & 2574 \\ 731 & 1423 & 0 & 479 & 1091 & 183 & 551 & 983 & 1447 & 1188 \\ 887 & 1033 & 479 & 0 & 782 & 412 & 112 & 1177 & 1052 & 1546 \\ 1666 & 1446 & 1091 & 782 & 0 & 919 & 671 & 1956 & 1450 & 1633 \\ 894 & 1427 & 183 & 412 & 919 & 0 & 450 & 1156 & 1448 & 1149 \\ 999 & 1055 & 551 & 112 & 671 & 450 & 0 & 1290 & 1071 & 1548 \\ 291 & 1457 & 983 & 1177 & 1956 & 1156 & 1290 & 0 & 1487 & 1942 \\ 1290 & 30 & 1447 & 1052 & 1450 & 1448 & 1071 & 1487 & 0 & 2595 \\ 1791 & 2574 & 1188 & 1546 & 1633 & 1149 & 1548 & 1942 & 2595 & 0 \end{pmatrix}$$

IMPRINT

Publisher

German Record-Linkage Center
Regensburger Str. 104
D-90478 Nuremberg

Template layout

Christine Weidmann