A graph theoretic linkage attack on microdata in a metric space

Martin Kroll[*]

[*]University of Duisburg-Essen, Lotharstraße 65, D-47057 Duisburg

E-mail: martin.kroll@uni-due.de

Abstract. Certain methods of analysis require the knowledge of the spatial distances between entities whose data are stored in a microdata table. For instance, such knowledge is necessary and sufficient to perform data mining tasks such as nearest neighbour searches or clustering. However, when inter-record distances are published in addition to the microdata for research purposes, the risk of identity disclosure has to be taken into consideration. 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 realistic assumptions of 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 algorithmic 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 show that publishing even approximate distances could increase the risk of identity disclosure unreasonably. We will concentrate on the perturbation of the distances; the anonymization of the vertex labels will play only a minor role in our simulations. Since the current version of our attack is not scalable, it can be launched only on datasets of sizes up to few thousands records. In the future we intend to explore possible ways of pushing further the limits of our approach.

Keywords. Anonymity, identity disclosure, linkage attack, maximum approximate common subgraph problem, microdata

1 Introduction

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 identified a contaminated water pump in London as the source of a cholera outbreak by linking the cases of mortality to their location and visualising these locations and the positions of surrounding water pumps on a map [36].

In recent years, spatial analysis techniques have become increasingly attractive in the social sciences as well [32]. However, when personal microdata containing sensitive information (e.g., gathered in a survey or health study) are published for research purposes, the anonymity of the individuals has to be guaranteed. It has been pointed out in [16] that *location is often one of the critical pieces of information for a successful re-identification attack*. Therefore, in praxis usually only microdata that contain spatial information in an aggregated form are released, which restricts the choice of applicable techniques for analysis drastically.

In particular, distance calculations that are based on aggregated data become difficult and imprecise [3], especially for entities that are spatially close to each other. Since many data mining techniques and methods in spatial analysis require accurate distance computations, it is necessary to investigate the extent to which additionally published (approximate) inter-record distances influence the risk of identity disclosure and how a possible non-acceptable increase of this risk can be prevented. Our work presented in this article provides a novel approach for tackling these questions.

Contributions of the paper

We introduce a flexible natural graph model for microdata with known inter-record distances. The search for a maximum common subgraph between two such graph models is interpreted as a novel kind of linkage attack on such microdata. We discuss the relative merits of our method in comparison to the usual linkage attacks on the basis of a small-scale example (Example 9 in Section 4).

Furthermore, in the special case of geographical distances, it is shown that, on the basis of simulated data, a non-negligible risk of identity disclosure exists if $\mathcal{N}(0,\sigma^2)$-distributed Gaussian noise is added to the input coordinates for too 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.

Note that we will focus on the effect of distance perturbations during our experimental study. The effect of quasi-identifier anonymization (via the concept of k-anonymity) will be investigated also in Section 5 but plays only a minor role.

Organisation of the paper

In Section 2, we refer to related work. The preparatory work is given in Section 3 as well as 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. In this experimental section, the scalability of our attack is discussed as well. We conclude and discuss the possible directions for future research in Section 6.

2 Related work

Statistical disclosure control and privacy preserving data mining

As already indicated in the introductory section above, the original motivation for the work presented in this article stems back to the wish to make the wide variety of distance-based methods (e.g., from spatial statistics) applicable for microdata that are published for scientific purposes. Since it is intuitively compelling that naive release of the exact distances between individuals can increase the risk of deanonymization, the question of interest, however, is how the knowledge of approximate distances might change the risk of identity disclosure, i.e. the data snooper's chance of success.

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* [22]. It is universally acknowledged that a necessary but insufficient 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., `social insurance number`) are called *(direct) identifiers*, in contrast to *quasi-identifiers*, which do not have the power to nullify an individual's anonymity on their own, a distinction which has to be ascribed to Dalenius [13].

By using 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, in [37] it was shown that based on 1990 US census data, 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* [15]: 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 [38]: 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.

Even though theoretical results on linkage attacks were recently obtained in [30], the concept of k-anonymity had already been proposed as a remedy against linkage attacks in [35]. 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, an unambiguous linkage between the identification and target file will not be possible. The graph theoretic linkage attack introduced in Section 4 contains the classical linkage attack via quasi-identifiers as a subroutine, however, it provides a way to resolve at least some of the ambiguous matches.

Several papers on *privacy preserving data mining* have already discussed privacy issues with respect to the distance-preserving transformations of microdata: However, only specific kinds of distances have been considered (e.g., ℓ_1-distance in [34] or the Euclidean (i.e. ℓ_2-) distance in [29]). Moreover, 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. Thus, in our scenario, an attack can only be based on the distances themselves and not on the knowledge of data from which they are calculated (e.g., perturbed geographical coordinates).

In contrast, the attack proposed in this paper is not limited to a special distance function but can be applied to any kind of distance function. For example, distances between participants of a health or social science survey could be either measured as geographical distances or as travelling distances. The only assumption we make is that an attacker has the ability to compute the considered distances as well.

Furthermore, a distance-preserving technique for the anonymization of binary vectors is discussed in [24]. In contrast to our approach, in that article the distance information alone is not assumed to increase the risk of identity disclosure.

Location privacy and geographical masks

There is a vast literature on the problem of identity disclosure when dealing with spatially referenced data. The opportunities and challenges with regard to spatial data in the context of social sciences are discussed in great detail in [20] and [21].

Articles [6] and [12] give illustrative examples of how naive publishing of spatially referenced data can lead to a violation of anonymity: In both cases, the respective authors were able to reconstruct many of the original addresses successfully from published low resolution maps. A currently flourishing branch of research deals with anonymization techniques for datasets containing mobility traces of individuals [18] (e.g., obtained via mobile phone tracking). This topic is usually referred to as *location privacy* [27].

In this article, however, we consider the deanonymization risk that arises from the knowledge of the (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 term of *geographical masks*. [1] and [31] provide comprehensive outlines of the existing methods.

A noteworthy method is due to Wieland et al. [40], who developed a method based on linear programming that moves each point in the dataset as little 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 geographical area as little as possible, whereas attempts predominantly focusing on the preservation of distances have not yet been discussed in the context of spatial data. It appears to be obvious that neglecting the underlying geographical area might yield more accurate results regarding distance calculations.

Social network anonymization

The use of a graph model in this article might suggest a strong connection between our approach and the methods discussed in the area of *social network anonymization* [41]. However, we model the microdata with known inter-record distances using 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 from being complete and their edges are not usually weighted. For example, in [8] 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] do not seem to be sensible when investigating the risk of identity disclosure for published microdata. However, the active attack proposed in [2] is related to the one in this paper because it also makes use of graph algorithmic building blocks. It consists in the detection of a subgraph in a larger graph, whereas the attack in this paper is based on finding the common subgraphs of two different graphs.

Pattern recognition

To the best of our knowledge, this paper is the first one to make use of a graph model for a microdata file and the 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 (see [9] as a source providing an extensive outline).

Fundamental to our presentation is the article by Levi [28], which motivates to transform the problem of finding the (maximum) common subgraphs of two graphs into a (maximum) clique detection problem, and its adaption in [17] 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 [17]). The problem of finding a maximum clique in a graph is known to be NP-hard [19] and a great deal of attention has been paid to the development of techniques for solving this problem either exactly or at least approximately [5]. For the simulation study in Section 5 of this paper, we made use of the maximum clique detection algorithm introduced by Konc and Janežič in [25]. Their algorithm is based on a colouring algorithm with dynamical bound evaluation. A critical discussion of the scalability of this algorithm with regard to our attack is given in Section 5 as part of our empirical study on simulated datasets.

3 A graph model for microdata in a metric space

Preliminaries

A metric space is 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 the 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 the 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.*

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 shorthand. 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 $E' = [V']^2 \cap E$ 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 a larger clique. A clique is *maximum* if there is no other clique containing more vertices. Clearly, a maximum clique is always maximal, but generally not 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 [7].

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.

The graph model

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 labeling 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 $\mathcal{G}$ (see 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 modeling 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 from 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 were 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` (`yob`). 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.

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

Computing pairwise distances between these points yields the following distance matrix D:

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

The corresponding graph model is then given by $V = \{1, 2, 3, 4\}$, $E = [V]^2$ 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 `yob` 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	yob
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 `yob`.

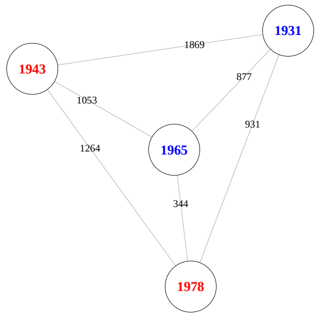

Figure 1: The graph model for the example microdata. The attribute `sex` is indicated by the colour of the vertex labels.

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, we have to at least presuppose that an appropriate external microdata file is available to the data snooper.

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

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

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

Although in some cases Assumption 4 might not be fulfilled, it is easy to find examples of when this would 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 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, any modification of the distances in the target file to be carried out by the data holder for the purpose of anonymization will have to be taken into consideration.

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 for both the target and identification file. At this step, the snooper will only consider the common quasi-identifiers of both files for the definition of the vertex labels because a comparison of records can only be based on such attributes. Hereafter, the resulting graphs will be referred to as the *target* and *identification graph*.

Hence, classical linkage attacks consist in trying to find vertices in the target graph for each vertex in the identification graph that result in matches for 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, that cannot be broken without 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 increases a snooper's chances of evaluating the credibility of potential matches. For instance, if we 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 $\lambda_V(v_2) = \lambda_W(w_2)$, we observe coincidence regarding the vertex labels. If the corresponding edge weights $\omega_E(v_1v_2)$ and $\omega_F(w_1w_2)$ are at least approximately equal (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). Conversely, a large distortion with respect to the corresponding edge weights will reduce this credibility: In this case, at least one of the considered matches should 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 to identify nearly identical substructures in both graphs which are as large as possible.

We want to emphasize that we will use the distances between the records only as a means for re-identifications and in particular to resolve ties that arise when only quasi-identifiers are considered (throughout the paper we assume that the information given by the quasi-identifiers only is not sufficient for re-identifications by the data snooper). However, we do not deal with the case where the distance information itself is sensitive and needs to be concealed (this will usually not be the case for databases arising in social or health surveys where the location information is given by the present residence or the place of birth).[1]

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 above). All of these considerations can be dealt 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:

[1] The following example might appear when tracking data of mobile phone users are published in an anonymized manner: If the location indicates the current location at some time t and it is revealed that Alice (who is married to Bob) is at the same place with David (who's married to Alice's best friend Carol), then such information may be sensitive.

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

(ii) $\omega_E(s_1 s_2) \approx \omega_F(\varphi(s_1)\varphi(s_2))$ for all distinct $s_1, s_2 \in S$.

The interpretation of $\approx$ in Definition 5 has to be made precise depending on the prevailing situation and especially on the possible perturbations of the distances caused by the data holder before publishing the microdata. This issue will be dealt with in detail in Example 9 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. For example, if numerical values such as height belong to the quasi-identifiers, one could allow some tolerance by permitting also matches between records with height 185 cm in the identification file and height values in the target file between 183 and 187 cm. In this paper, however, we do not deal with this aspect. We require exact coincidence for the labels of two vertices to be matched since 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 a clique detection problem 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

$$V_\otimes = \{(v, w) \in V \times W : \lambda_V(v) = \lambda_W(w)\} \quad \text{and}$$
$$E_\otimes = \Big\{\{(v_1, w_1), (v_2, w_2)\} : v_1 \neq v_2, w_1 \neq w_2 \text{ and } \omega_E(v_1 v_2) \approx \omega_F(w_1 w_2)\Big\}.$$

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. There is a one-to-one correspondence between the 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 $S = \{v_1, \dots, v_k\} \subseteq V$ and $T = \{w_1, \dots, w_k\} \subseteq W$, respectively. Without loss of generality, we assume $\varphi(v_i) = w_i$ for $i \in \{1, \dots, k\}$ under the corresponding subgraph isomorphism φ. Condition (i) in Definition 5 yields $(v_i, w_i) \in V_\otimes$ for $i = 1, \dots, k$. Condition (ii) in Definition 5 implies that $\omega_E(v_i v_j) \approx \omega_F(\varphi(v_i)\varphi(v_j)) = \omega_F(w_i w_j)$ and (v_i, w_i) and (v_j, w_j) are adjacent in $\mathcal{G}_\otimes$. Because i, j were chosen arbitrarily, $\mathcal{C} := \{(v_1, w_1), \dots, (v_k, w_k)\}$ forms a k-clique in $\mathcal{G}_\otimes$.

Conversely, let $\mathcal{C}$ be a k-clique in $\mathcal{G}_\otimes$ given by vertices $(v_1, w_1), \dots, (v_k, w_k) \in V_\otimes$. We define $S = \{v_1, \dots, v_k\}$, $T = \{w_1, \dots, w_k\}$ 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 $i = 1, \dots, k$. Thus, condition (i) in Definition 5 is satisfied. The validity of the second condition follows from the fact that we have $\omega_E(v_i v_j) \approx \omega_F(w_i w_j) = \omega_F(\varphi(v_i)\varphi(v_j))$. □

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.*

We now put all the ingredients collected so far together and formulate the overall graph theoretic linkage attack.

Overview

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 brief comment on Step 4 of the attack: As already indicated in Section 2, there is a vast literature concerning the problem of maximum clique detection in graphs. A systematic comparison of the prevalent techniques to tackle this problem in the context of 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 using a small-scale example which makes use of the data summarised in Appendix A.

Example 9. Consider microdata Table 9 in Appendix A, which contains information about various important European poets. This table is anonymized by removing the direct identifier `name`, generalizing 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 anonymized Table 10 in Appendix A.

While 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 10. We assume that the snooper is in possession of the identification microdata in Table 11, i.e. the attributes `cob` and `language` serve as quasi-identifiers. By geocoding the birth locations and calculating the geographic distances, the snooper obtains the distance matrix D_2. Graph models $\mathcal{G}_1$ and $\mathcal{G}_2$ for the target and identification data can be built by using this information and are visualised in Figure 2.

Table 2 lists all the possible matches if 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 of a linkage attack where the distances are not taken 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 $\omega_E(v_1v_2) \approx \omega_F(w_1w_2) \Leftrightarrow |\omega_E(v_1v_2) - \omega_F(w_1w_2)| < 5$.[2] This definition of $\approx$ leads to the product graph shown in Figure 3.

[2] We previously mentioned that allowing such a deviation is already necessary because of errors that appear due to the fact that the data holder and snooper generally use different methods for geocoding and distance computation. This fact was also addressed 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` [23].

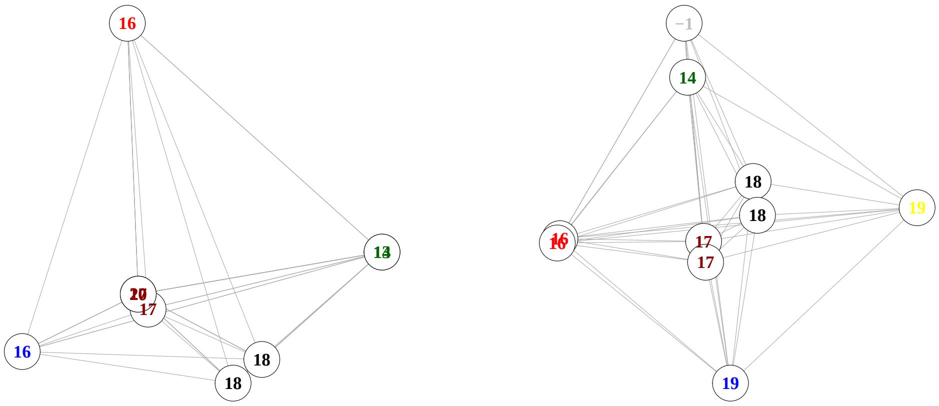

Figure 2: Graph models for target (left) and identification (right) microdata in Example 9. The layout of the graphs was chosen such that the edge lengths give an approximate indication of the distances. The attribute `language` is indicated by the vertex label colour, whereas the attribute `cob` is indicated 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 10 and 11 with respect to the quasi-identifiers `cob` and `language` only, i.e. vertex labels in the accompanying graph models.

As can be easily seen 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 would accept the potential matches in rows 1,2,3 and 5 in Table 2 as matches and reject the remaining ones.

Although Example 9 is artificial, it illustrates some of the phenomena that also appear when real-world data are taken into consideration:

- The definition of $\approx$ has to be chosen carefully. In the present example, distances between cities scattered all over the European continent are considered so that even the rather rough definition above (allowing for an absolute deviation of five kilometers)

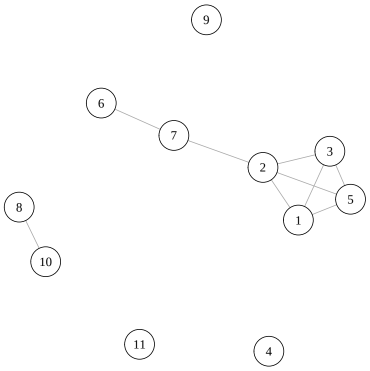

Figure 3: Product graph in Example 9.

will yield a useful result. In general, the definition of $\approx$ has to be chosen such that as many common edges as possible of the target and identification graph are detected correctly, i.e. not 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 respective first records of both tables would have already been possible unambiguously 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 credibility for this specific matching, which is now not only supported by the coincidence of the quasi-identifiers but also by the coincidence of the distances to other matches.

- However, in certain cases unambiguous matching is only possible because of the additional information about the distances. For example, record 3 of the target table could be matched with records 3 and 6 of the identification table only by taking the quasi-identifiers into consideration. This tie is resolved in our example by 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, however, they also cannot be distinguished by considering the distances to these records because the corresponding point locations (`loc`=Paris in both cases) coincide.

- Finally, the attack has reduced the number of matches from eleven in Table 2 to four. These matches indeed correspond to the actual overlap of the 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 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

The data for the simulation study were generated as follows: In the first step, addresses from the German telephone book were sampled at random. Subsequently, geographic latitudes and longitudes based on the World Geodetic System 1984 were assigned to these addresses using the `geocode` command from the R package `ggmap` [23]. Finally, the 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 containing (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 derived during the German census 2011.[3]

We generated data where both the size N_1 of the target and N_2 of the identification file were equal to 500. The overlap N_{common} of common records was chosen equal to 50. The target and the identification file are visualised in Figure 4. Note that the classification with respect to age (eleven age intervals) is rather rough; this guaranteed the existence of many duplicates with respect to the quasi-identifiers in our test microdata, which would result in ties when performing a classical linkage attack. Indeed, the order of the resulting product graph was equal to $|V_{\otimes}| = 15517$.

In order to study the scalability of our attack we also considered a slightly larger data set with $N_1 = N_2 = 2000$ and $N_{\text{common}} = 200$. For significantly larger files we were not able to perform our attack using the exact maximum clique algorithm of [25] any longer which is in coincidence with the fact that the problem of finding a maximum clique of a graph is NP-hard. More information including runtime results can be found below in the subsection *Scalability of the attack*. In addition to the test for scalability we used this second pair of files in order to test the effect of k-anonymization (for more details, see the next paragraph).

Perturbation

A standard technique for the anonymization of spatial point data consists in the addition of random noise to their coordinates (see Section 3.2 in [1]). In this section, we consider the performance of the proposed graph theoretical linkage attack under this anonymization technique. To be more precise, $\mathcal{N}(0, \sigma^2)$-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.

In addition to the anonymization of distances we also investigated the anonymization of the vertex labels systematically by means of the second data set. To be more precise, we considered k-anonymization which was simply achieved by suppressing combinations of the key variables `sex` and `age` that appeared less than the given predefined threshold k. We considered values of k between 1 and 10.

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

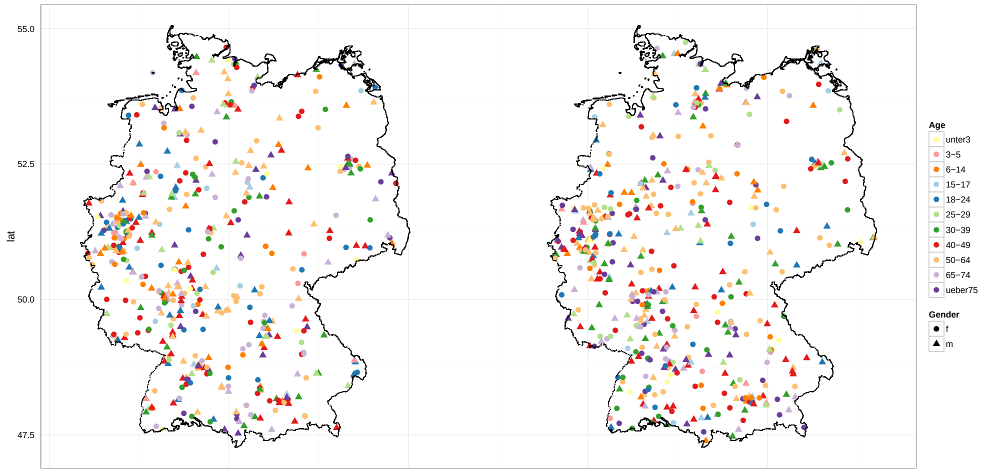

Figure 4: One of the data sets for the simulation study: Both the target (left) and identification file (right) contain 500 records of which 50 are common. Quasi-identifiers (`age`, `gender`) were sampled according to the actual distribution of these attributes due to the demographic statistics derived during the German census in 2011. Without the additionally released approximate distances the target file can be regarded as sufficiently anonymized: a classical linkage attack leads to a set of 15517 potential matches between the target and identification file.

Fine-tuning of the attack

A suitable definition for the relation $\approx$ has to be found for the generation of the product graph in the graph theoretical linkage attack. Following Kerckhoffs' principle [33] (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 the distances are calculated and, furthermore, that the standard deviation σ is known to him (the latter assumption is in conformance with [1], who emphasise that *all useful spatial analyses of masked data require some knowledge about the characteristics of the mask used*).

Under the assumption of a Euclidean distance function, the effect of random perturbation of the input coordinates on the squared distances can be studied theoretically, an approach which has been considered in [24]. Such a rigorous mathematical analysis appears to be more difficult in the case of geographical distances, i.e. distances on the sphere. 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 of the deviation of the 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 $d-d'$ is greater than the empirical $\frac{1-\alpha}{2}$-quantile and smaller than the $\frac{1+\alpha}{2}$-quantile

					quantile			
		0.05	0.1	0.25	0.5	0.75	0.9	0.95
	0.005	-1.1088	-0.8558	-0.4261	0.0226	0.4496	0.9256	1.2192
	0.010	-2.3909	-1.7191	-0.8750	0.0798	0.9733	1.8218	2.2642
	0.015	-3.3063	-2.4633	-1.2288	0.0810	1.4714	2.7492	3.4624
	0.020	-4.6132	-3.6261	-2.0787	-0.0615	1.7278	3.4013	4.3512
	0.025	-5.6147	-4.2924	-2.2826	-0.1592	2.2177	4.1254	5.3089
σ	0.030	-6.4952	-4.9210	-2.5024	0.1763	2.9190	5.2730	6.6511
	0.035	-8.3848	-6.2351	-3.0673	-0.0665	3.0206	6.0428	7.9411
	0.040	-9.1530	-6.7866	-3.7315	-0.1884	3.4698	7.0085	8.7236
	0.045	-11.0830	-8.2160	-4.1860	-0.0680	3.6953	7.6620	10.2638
	0.050	-11.4906	-8.9544	-4.7386	-0.0057	4.5955	8.6728	11.4998

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

for the current value of σ. In this case, the distances from the identification file take on the role of d and the distances from the target file the one of d'. If $\omega_E(v_1, v_2) = d$ and $\omega_F(w_1, w_2) = d'$ an edge will be inserted in the product graph between (v_1, w_1) and (v_2, w_2) if and only if $d - d' \in [q_{\frac{1-\alpha}{2}}, q_{\frac{1+\alpha}{2}}]$. The threshold parameter α chosen by the snooper is supposed to guarantee that a common edge of the target and identification graph is detected by the snooper with probability approximately equal to α. Its effect will also be considered within this section. In our scenario the difference between Euclidean and geographical distances should be so small that it can be neglected. However, other kinds of distances such as travelling distance (time needed to travel from one location to the other) could be used and even be more senseful in some applications. Thus, we decided to demonstrate the study of the effect of perturbation by means of a simulation study which can also be performed in this more general case under the validity of Assumption 4.

Implementation

All the experiments reported here were performed using R and the exact maximum clique detection algorithm proposed in [25].[4] Thus the algorithm guarantees that a maximum clique is indeed found and not only a large clique which might be the case for approximative algorithms only. All the accompanying visualisations were created in R.

Evaluation of the attack

The matches and non-matches between the target and identification file gathered by the proposed 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

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

[4] We adapted the C++ implementation of this algorithm, which is available from `http://www.sicmm.org/~konc/maxclique/`, for our purposes.

which are two standard measures in the evaluation of data linkage processes [10].

Simulation design and results

In our experiments, we varied the noise parameter σ as well as the threshold parameter α. For each parameter setup, the simulation was repeated $n = 100$ times for the first experiment ($N_1 = N_2 = 500$) and $n = 50$ times for the second experiment ($N_1 = N_2 = 2000$). The mean of precision and recall over all iterations for the chosen parameter setups can be found in Tables 4– 7. Visualisations of some of these results can be found in Figures 5 and 6. In addition, typical outcomes of the graph theoretic linkage attack are visualized in Figure 7.

						σ					
		0.005	0.010	0.015	0.020	0.025	0.030	0.035	0.040	0.045	0.050
	0.1	0.7800	0.7215	0.3415	0.2752	0.3232	0.2193	0.1549	0.1571	0.1136	0.1584
	0.2	0.9717	0.8929	0.8769	0.8229	0.7569	0.6121	0.5657	0.5094	0.4780	0.4785
	0.3	0.9831	0.9513	0.9047	0.8502	0.8255	0.7728	0.6862	0.6567	0.6037	0.5975
	0.4	0.9829	0.9558	0.9037	0.8700	0.8358	0.7766	0.7411	0.6651	0.6428	0.6410
α	0.5	0.9808	0.9458	0.9133	0.8721	0.8374	0.7834	0.7505	0.6832	0.6526	0.6274
	0.6	0.9830	0.9436	0.9102	0.8725	0.8315	0.7780	0.7430	0.6974	0.6604	0.6229
	0.7	0.9803	0.9405	0.9087	0.8675	0.8255	0.7707	0.7434	0.6948	0.6556	0.6086
	0.8	0.9795	0.9373	0.9008	0.8539	0.8027	0.7666	0.7248	0.6894	0.6484	0.6065
	0.9	0.9764	0.9304	0.8884	0.8351	0.7954	0.7513	0.7017	0.6605	0.6159	0.5876

Table 4: **Average precision** in dependence on the parameters σ and α over $n = 100$ repetitions while $N_1 = N_2 = 500$ and $N_{\text{common}} = 50$.

						σ					
		0.005	0.010	0.015	0.020	0.025	0.030	0.035	0.040	0.045	0.050
	0.1	0.0664	0.0612	0.0302	0.0284	0.0326	0.0228	0.0168	0.0174	0.0130	0.0196
	0.2	0.1166	0.1034	0.1126	0.1144	0.1120	0.0870	0.0864	0.0770	0.0714	0.0818
	0.3	0.1586	0.1552	0.1556	0.1592	0.1642	0.1416	0.1398	0.1354	0.1292	0.1390
	0.4	0.2084	0.2204	0.2112	0.2164	0.2162	0.1934	0.1966	0.1894	0.1828	0.2026
α	0.5	0.2774	0.2874	0.2754	0.2880	0.2722	0.2628	0.2526	0.2518	0.2404	0.2564
	0.6	0.3622	0.3714	0.3308	0.3582	0.3364	0.3322	0.3166	0.3146	0.3084	0.3132
	0.7	0.4402	0.4490	0.4250	0.4408	0.4198	0.4080	0.3976	0.3992	0.3754	0.3936
	0.8	0.5638	0.5538	0.5420	0.5408	0.5096	0.5110	0.5204	0.5068	0.5104	0.4966
	0.9	0.7202	0.7146	0.6960	0.6790	0.6568	0.6568	0.6838	0.6554	0.6658	0.6468

Table 5: **Average recall** in dependence on the parameters σ and α over $n = 100$ repetitions while $N_1 = N_2 = 500$ and $N_{\text{common}} = 50$.

						σ					
		0.005	0.010	0.015	0.020	0.025	0.030	0.035	0.040	0.045	0.050
	0.1	0.99	0.99	0.97	0.95	0.93	0.93	0.88	0.91	0.80	0.84
α	0.5	1.00	0.99	0.98	0.96	0.95	0.92	0.91	0.89	0.87	0.85
	0.9	1.00	0.99	0.97	0.96	0.93	0.91	0.88	0.86	0.83	0.81

Table 6: **Average precision** in dependence on the parameters σ and α over $n = 50$ repetitions while $N_1 = N_2 = 2000$ and $N_{\text{common}} = 200$.

Utility of the Perturbed Data

The simulations show that, in principle, a sufficient level of anonymity can be achieved (even for our small datasets and the ideal preconditions for the attacker, i.e. target and

		σ									
		0.005	0.010	0.015	0.020	0.025	0.030	0.035	0.040	0.045	0.050
	0.1	0.04	0.04	0.03	0.04	0.04	0.03	0.03	0.04	0.03	0.04
α	0.5	0.19	0.21	0.20	0.21	0.20	0.20	0.19	0.19	0.18	0.20
	0.9	0.65	0.65	0.63	0.62	0.61	0.60	0.64	0.61	0.64	0.62

Table 7: **Average recall** in dependence on the parameters σ and α over $n = 50$ repetitions while $N_1 = N_2 = 2000$ and $N_{\text{common}} = 200$.

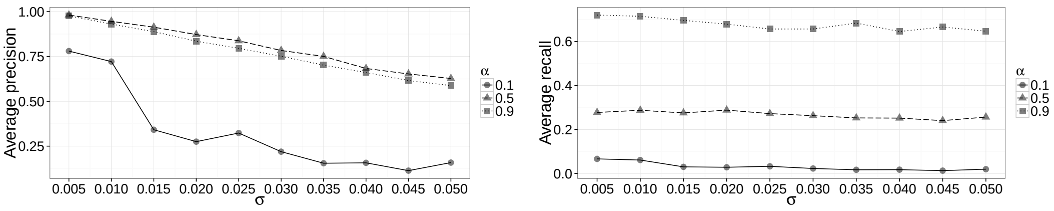

Figure 5: Dependence of average precision (left) and recall (right) on the standard deviation σ for different values of α (see Tables 4 and 5).

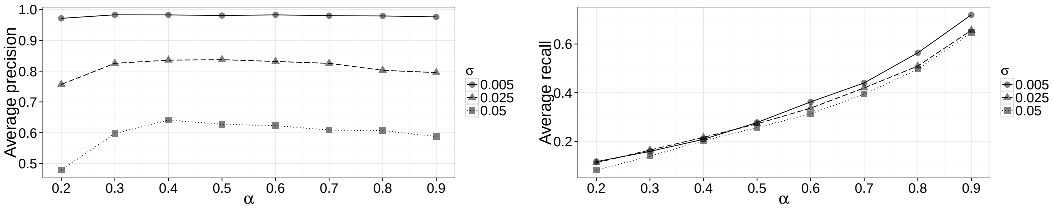

Figure 6: Dependence of average precision (left) and recall (right) on the threshold parameter α for different values of σ (see Tables 4 and 5).

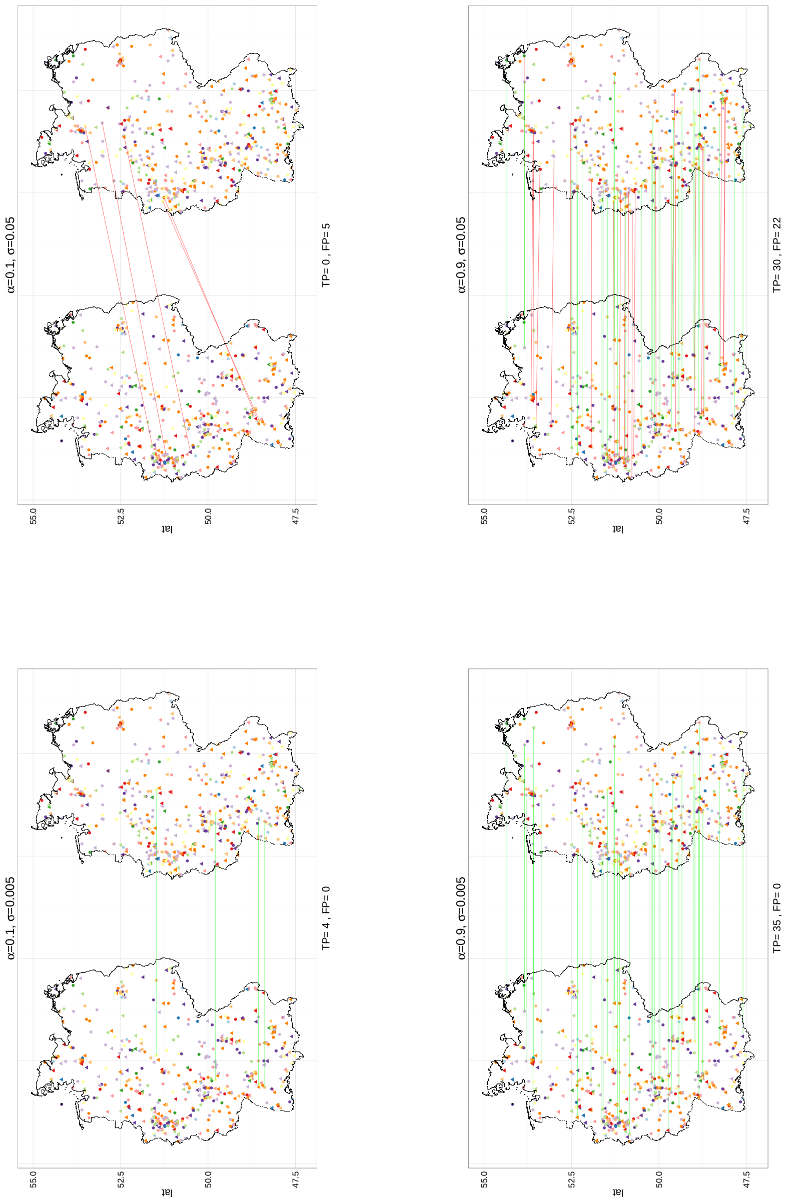

Figure 7: Typical results of the graph theoretic linkage attack for different combinations of the noise parameter σ (chosen by the data holder of the target file) and the threshold parameter α (chosen by the data snooper). Line segments between the target and identification file indicate matches made by the data snooper (the green lines indicate true, the red ones false positives). Larger values of α lead to more matches and a increase in recall, larger values of σ to a decrease in precision.

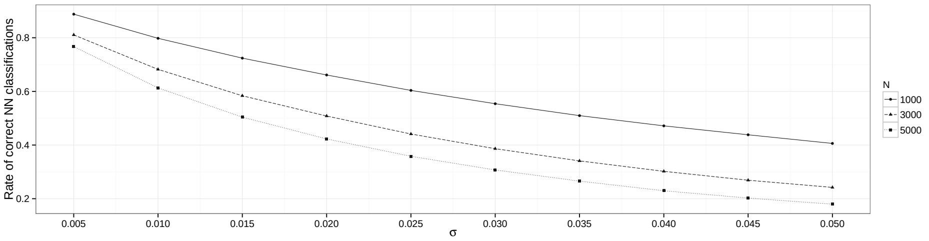

Figure 8: Effect of anonymization: For three sets of points from Germany of different size N the effect of the anonymization given by the threshold σ (on the x-axis) was evaluated as the portion of correct nearest neighbour classifications (on the y-axis) after Gaussian noise was added to the geographical coordinates before computing the distance matrix.

identification file are small and there exists a certain overlap between both files) by the addition of random noise to the input coordinates before computing the distance matrix. However, this anonymity is not free, which is illustrated by means of Figure 8, where the portion of correct nearest neighbour classifications is plotted against the anonymization parameter σ. In some cases the anonymization threshold σ would have to be chosen large enough to guarantee at least some degree of anonymity that useful analyses based on the distances would become difficult. For this reason, the development of distance modification techniques that guarantee a certain degree of anonymity, and make it possible to also conduct useful analyses on the anonymized data, will be an important aspect of future research.

Effect of k-anonymization of the vertex labels

As announced above, we studied the effect of vertex k-anonymization via suppression on our attack. For this purpose, we considered a target and an identification file each containing $N = 2000$ records. Again `age` and `sex` were chosen as quasi-identifiers but in contrast to the first experiment the age was not given by an age interval (e.g., 25-29) but by one integer number only. Then, k-anonymized versions for values of k between 1 and 10 were generated by suppression of `age`/`sex`-combinations that appeared less than k times. Then the attack was performed where only records from the target file were used for which the quasi-identifier information was not suppressed (another option would have been to regard these records as potential instances of every record in the identification file which, however, had lead to an essential enlargement of the vertex set of the product graph, see the next paragraph on the scalability of our attack). The results of the vertex anonymization experiment are summarized in Figure 9. A discussion of these results is included in the subsection *Discussion* below.

Scalability of the Attack

As already mentioned above we were not able to perform our attack when the size of target and identification file exceeded $N_1 = N_2 = 2000$ significantly. Since the runtime and applicability of the algorithm from [25] depends on the size of the product graph only, we generated random graphs of different size and connectedness and evaluated the algorithm from [25] on these graphs. To be more precise, we generated random graphs with $n \in \{10000, 20000, 30000\}$. Each edge was assumed to be existent randomly with probability $p \in \{0.001, 0.01, 0.1\}$ and existence of different edges was assumed to be independent.

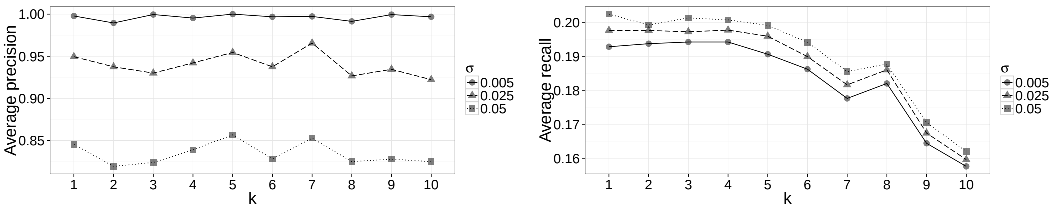

Figure 9: Dependence of average precision (left) and recall (right) on the anonymity parameter k. Note that for small values of σ the precision is close to 1 even for $k = 10$ whereas he precision would be less than 0.1 if matching could only be based on the quasi-identifier values. Experiments were performed using the datasets where both target and identification file contained 2000 records, respectively.

For this purpuse, we used the command `erdos.renyi.game(n, p)` from the R package `igraph` [11]. Then the exact maximum clique algorithm from [25] was applied to these graphs. The obtained runtimes are summarized in Table 8. We stress that the algorithm was not able to treat significantly larger graphs on a commercially available personal computer. Thus, our attack is limited to the case when the resulting product graph is of limited size which will be the case when both target and identification file are themselves of limited size (both containing roughly 2000 records). The maximal product graph we were able to deal with contained 30000 nodes and 44994803 edges. In that case the maximum clique algorithm took 3078 seconds. Another realistic attack scenario would be to investigate the case that the target file is very large and the identification file of moderate size. However, also in this case only limited sizes for the identification file will be possible such that we renounced a further investigation in this direction. Thus, our attack might not be applicable when considering datasets originating from surveys of large populations but when the population is limited, for example when data about individuals with a rare illness or data about infrastructure facilities are to be published.[5] Hence a further aspect of future research could be to investigate how the size of the product graph can be further reduced in a senseful manner for our attack.

Discussion

The main effect of the threshold parameter α concerns the recall. The probability of detecting a common edge of the target and identification graph is approximately equal to α. For this reason, higher values of α lead to a higher recall (see Table 5 and Figures 5 and 6).

Simultaneously, the effect of α on the precision appears to be twofold: On the one hand, for increasing α a larger portion of the overlap between the identification and target file can be successfully detected by the snooper, which makes false positives less likely (leading to a larger precision). On the other hand, for too high values of α also the chance for non-common edges of the target and identification graph (but which coincide with respect to the vertex labels of their endpoints) to be classified as common edges increases leading to a slight decrease in precision. The latter phenomenon, together with the increase in recall for

[5] In the paper [26] by the author together with Rainer Schnell we study the effect of our attack on another anonymization method on a dataset of 847 hospitals in England which might be a realistic dataset, say, for example, in a national health survey.

n	p	E	maximum clique size	time (in s)
10000	0.001	50318	3	0.160858
	0.01	499765	4	0.273826
	0.1	5002813	8	22.0877
20000	0.001	199909	3	0.631664
	0.01	1999927	4	1.45158
	0.1	19996877	8	468.74
30000	0.001	449574	3	1.4361
	0.01	4502415	5	3.89454
	0.1	44994803	8	3077.83

Table 8: Scalability of the exact maximum clique algorithm: Random graphs with n vertices and edge probability p yielding random graphs with E edges were generated and the maximum clique algorithm from [25] was applied. For significantly larger graphs the application of the algorithm was not feasible any more.

increasing α mentioned above, would reflect a trade-off between precision and recall, which is a well-known phenomenon in data linkage [10]. Thus, combining these two thoughts, for increasing α the precision should rapidly increase initially and then slightly decrease when α becomes too large. This expectation is confirmed by our experiments (see Table 4 and Figures 5 and 6), although the decrease in precision when α becomes too large is not significant for the considered values of σ.

From the definition of $\approx$ (see the paragraph *Fine-tuning of the attack* above), it is supposed that the recall does not change significantly in dependence on σ because the probability of correctly detecting an edge should be nearly α (which is independent of σ). This non-dependence is clearly confirmed by the performed simulations and illustrated in Figures 5 and 6. 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.

Note that in our specific example, the snooper would primarily attempt to achieve a 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 successfully, 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 would also alleviate a snooper's work in this more general case.

Regarding the effect of k-anonymization on our attack, for the considered values of $k \in \{1, \ldots, 10\}$ no significant loss in precision and a slight loss concerning the recall was observed, see Figure 9. In particular, when only taking quasi-identifiers (without inter-record-distances) into consideration one would expect a precision which is bounded from above by $\frac{1}{k}$ whereas in our case the precision is close to one for small values of σ even for a value of k equal to 10. Following our strategy and not taking records with suppressed quasi-identifiers into account for the generation of the product graph, a further increase of k would lead to a product graph with only few vertices and thus lead to a further decrease of the recall. An alternative would certainly be to match suppressed quasi-identifiers in the target file with all quasi-identifier values in the identification file during the generation of the product graph. This would, however, lead to a product graph with so many vertices that the attack could not be performed due to its limited scalability. In the extreme case – when all quasi-identifiers would have been suppressed – the problem would be equivalent to the problem of finding the largest approximate common subgraph of target and identification file (with no information about the quasi identifiers at all). In this case the product graph would have 250000 vertices in the case $N_1 = N_2 = 500$ and thus not be feasible with our current implementation.

Obviously, for distance modification techniques other than perturbation of the input coordinates, a snooper will have to modify the graph theoretical linkage attack, especially the definition of $\approx$. However, due to Kerckhoffs' principle, it has to be assumed that the snooper at least knows 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 to the distance matrix itself (a technique discussed in [24]), 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, thereby making the empirical study on distance deviations originating from perturbation of the input coordinates unnecessary. 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 [17] 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 and flexible 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. The main message of the article is that when microdata are enriched by (noisy) spatial inter-record-distances the risk of de-identification increases but the risk seems to be controllable. Only for target and identification files of moderate size we were able to perform the suggested attack completely. Thus, the fact that the graph matching and the maximum clique problem are NP-hard problems seems to provide a barrier against the attacker's wish to link as much information between target and identification file as possible (this might be seen in coincidence with a well-known phenomenon from classical cryptography where the hardness of the problem of factorizing product of two large primes into its factors provides a barrier on which the applicability of many cryptographic methods is based). However, for the case that the size of the data sets is limited (e.g., when already the underlying populations are not very large), the release of distances might increase the risk of identity disclosure unreasonably even if geographical coordinates have been perturbed by random Gaussian noise before the distances are calculated. Furthermore, we showed that an increase of the standard deviation of the added random noise will gradually lead to a sufficient level of anonymity also in this case, but also make the perturbed distances useless for further analysis (see Figure 8). However, the practicability of the proposed attack to real world scenarios seems to be limited especially when only small samples from a large population are published and sufficient anonymity with respect to the quasi-identifiers is given. In spite of this, the development and analysis of anonymization techniques for microdata in a metric space that allows for a certain degree of anonymity but distort the distances as little as possible (particularly with regard to the applicability of data mining techniques) will be an important aspect of future research.

Acknowledgements

This research has been supported by the research grant SCHN 586/19-1 from the German Research Foundation (DFG) awarded to the head of the Research Methodology Group, Rainer Schnell. I am grateful to Rainer Schnell for valuable suggestions that helped to improve this paper.

I also thank two anonymous referees for their constructive comments.

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] K.M.M. Beyer, A.F. Saftlas, A.B. Wallis, C. Peek-Asa and G. Rushton (2011) A probabilistic sampling method (PSM) for estimating geographic distance to health services when only the region of residence is known. *International Journal of Health Geographics* 10:4.

[4] R.S. Bivand, E. Pebesma, V. Gomez-Rubio (2013) Applied spatial data analysis with R, Second edition. *Springer, New York.*

[5] 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*, Springer 1–74.

[6] 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.

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

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

[9] 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.

[10] 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*, Springer, Berlin 127–151.

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

[12] 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.

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

[14] M.M. Deza and E. Deza (2009) Encyclopedia of distances. *Springer, Heidelberg.*

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

[16] K. El Emam and L. Arbuckle (2013) Anonymizing health data. *O'Reilly.*

[17] 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*, Springer, Cham 279–286.

[18] S. Gambs, M.-O. Killijian and M. Núñez del Prado Cortez (2011) Show Me How You Move and I Will Tell You Who You Are. *Transactions on Data Privacy* 4:103–126.

[19] M.R. Garey and D.S. Johnson (1979) Computers and intractability: A guide to the theory of NP-completeness. *W. H. Freeman.*

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

[21] 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.

[22] 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*, John Wiley & Sons, Ltd.

[23] 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`.

[24] 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.

[25] J. Konc and D. Janežič (2007) An improved branch and bound algorithm for the maximum clique problem. *MATCH Commun. Math. Comput. Chem.* 58:569–590.

[26] M. Kroll and R. Schnell (2015) . Anonymization of geo-referenced health data via Lipschitz embedding. *Forthcoming.*

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

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

[29] 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*, Springer, Berlin 297–308.

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

[31] 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*, Springer, Berlin 226–240.

[32] R.N. Parker and E.K. Asencio (2008) GIS and spatial analysis for the social sciences: Coding, mapping, and modeling. *Taylor & Francis.*

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

[34] 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.

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

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

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

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

[39] M. Templ, A. Kowarik and B. Meindl (2014) sdcMicro: Statistical Disclosure Control methods for anonymization of microdata and risk estimation. *R package version 4.4.0.* `http://CRAN.R-project.org/package=sdcMicro`.

[40] 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.

[41] E. Zheleva and L. Getoor (2011) Privacy in social networks: A survey. In: C.C. Aggarwal (ed.): *Social Network Data Analytics*, Springer, New York 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 9: Microdata containing information about famous European poets. The attribute `yob` contains the year of birth, and `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 10: The anonymized version of Table 9 is obtained by removing the direct identifier `name`, generalising the 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 11: Identification microdata table used by the data snooper in Example 9.

Geocoding of the locations from Table 11 using the R package `ggmap` and calculation of the mutual distances via the command `spDists` from the package `sp` [4] yields the distance matrix D_2:

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