// NO.25-036 | 07/2025

DISCUSSION PAPER

// PATRICK BREITHAUPT AND SANDRA GOTTSCHALK

Disclosure Risk for Firms in Combined Firm-Datasets

Leibniz Association

ZEW

Disclosure Risk for Firms in Combined Firm-Datasets

Patrick Breithaupt[1,*], Sandra Gottschalk[1]

[1] ZEW – Leibniz Centre for European Economic Research, L7 1, 68161 Mannheim, Germany

Abstract

Past studies have successfully shown that the level of anonymisation of scientific use files (SUF) is sufficiently high to protect against disclosure attacks that use data from traditional firm databases. However, with the increasing availability of online data about firms, new challenges for the provision of SUFs arise. In this paper, we therefore focus on a scenario, where an attack against the Mannheim Innovation Panel SUF is performed using data from the Mannheim Enterprise Panel and Mannheim Web Panel. We find that the disclosure risk of our attack is small and increases only slightly if data from the Mannheim Web Panel are considered. Data protection officers may use our findings when researchers want to publish SUFs.

Keywords: matching, firms, text as data, scientific use files
JEL-Classification: L00, C8, C55, C61

**Corresponding author:* Patrick Breithaupt, Digital Economy Department, ZEW Mannheim, P.O. Box 103443, 68034 Mannheim, Germany. E-mail: patrick.breithaupt@zew.de.

Acknowledgments: The authors would like to thank the DFG for the funding (project name: BERD@NFDI for Business, Economic and Related Data, DFG funding id: NFDI 27/1). We thank Stefan Bender, Irene Bertschek, Jan Kinne, Thomas Niebel, Dominik Rehse, Luca Sandrini, Pascal Siegers, Christian Rammer, the participants of ZEW's Digital Economy seminar, and the BERD@NFDI team for valuable feedback. All remaining errors are ours alone.

1 Motivation and Introduction

Empirical research in economic and social sciences requires information about households and firms, which are collected by statistical offices and public or private research institutions in form of microdata. This applies to both basic research and evidence-based policy advice. European and German law (DSGVO and BDSG) provide that microdata of households, individuals and firms from official statistics are allowed to be passed on for scientific purpose only and if disclosure limitations are in effect guaranteed. Hence, disclosure should not be possible without unusually high costs and waste of time and energy. The same holds for data assembled by private or public research institutes or universities, especially if confidentiality is promised to the respondents. One problem that data providers ace when releasing micro-data sets for researchers n the form of scientific use files (SUF) is the preservation of confidentiality. Even business data are at risk because disclosure is more likely than for personal data as additional information are easier obtainable and the population size is substantially smaller.

Due to the steadily increasing availability of data on individuals and firms in the Internet, data protection issues were discussed in the last decades. Lawyers as well as computer scientists agree that data protection issues have to be faced with when disseminating research data because data sets of individuals, households, and firms can be easier disclosed when further information is available in the world wide web. Disclosure (or de-anonymisation) of observational units (individuals or firms) of SUFs are possible when relevant additional information on these units is (semi-) publicly available and matching techniques are effective. These circumstances give rise to reconsider traditional anonymisation techniques.[1] Further, the new European data protection law (EU 2018) postulates that providers of SUFs have to make a risk assessment and to document anonymisation techniques when disseminating individual data (compliance regulation).

Disclosure risk of firms in SUFs

This paper focuses solely on the dissemination of firm data SUFs and firms' disclosure risk. Traditional methods to avoid disclosure which are used by the official

[1]Data protection involves techniques like pseudonymisation and anonymisation, which aim to safeguard sensible information. In practice, various models within these frameworks are employed to enhance data protection, e.g. differential privacy (Dwork 2008), k-anonymity (e.g. Samarati and Sweeney 1998), uniqueness analysis (e.g. Bandara et al. 2020), l-diversity (e.g. Machanavajjhala et al. 2007) and artificial intelligence (e.g. Yoon et al. 2020). On the attacker side, there are attack methods such as probabilistic matching using linear optimisation or statistical models.

statistics and further scientific data providers to produce firm data SUFs[2] do not take into account that disclosure risk may have risen a lot due to the availability of these additional data. Further, mathematical and statistical algorithms, text analyzing software and machine learning techniques ("artificial intelligence") to analyze big structured and also unstructured data have been developed and improved so that the probability of disclosing research data via additional public data increased. Additionally, hardware capabilities (e.g. cloud computing infrastructure) are available for a wider range of people which also simplifies disclosure.[3]

Besides considering disclosure risk of firm data SUFs, which are currently available, this study thinks through the possibility of combining existing research data sets with publicly available firm information. In recent years, firm-level data are increasingly enriched with publicly available data on firms such as data from websites or other public data bases, in order to allow more detailed data analysis (Rammer and Es-Sadki 2023). The linkage of individual firm data sets is beneficial when the combined information opens up new research agendas and helps to answer research questions which could not have been responded to with only a single data set. But, combining SUFs with such external information results in a new form of disclosure risk, since one part of the information included in the data set comes from publicly accessible data bases.

This paper aims to assess how high disclosure risk is for SUFs that are enriched with publicly available data. For this purpose, we use firm-level data from the Mannheim Innovation Survey (MIP), which is the German part of the European Commission's Community Innovation Surveys that collect innovation-related data from firms across Europe. For the MIP, which is designed as a panel survey and conducted annually since 1992 (see Peters and Rammer 2023 and Section 3), the ZEW[4] has been offering SUFs to researchers for more than 20 years.

Adding publicly available information - ISO norm standards - to the MIP SUF

To be concrete, we plan to generate a SUF, which combines an existing SUF of the MIP of the survey year 2020 with an ISO norm indicator for the information security management system standard. This additional information enables scientists, who

[2] An overview is found in e.g. Brand 2000; Gottschalk 2004; Höhne et al. 2003; Müller et al. 1991; Wirth 1992.

[3] However, the disclosure scenarios are purely hypothetical, as no incidents of SUF disclosure attacks are known yet. Instead, the results of this paper should be primarily used by data protection officers in their legal arguments when releasing data.

[4] ZEW - Leibniz Centre for European Economic Research in Mannheim

are studying the innovation behaviour of firms, to measure the effects of this specific type of organisational innovation on firm performance. To make this information available, ZEW developed a web scraping and mining algorithm. We took advantage of the fact that in recent years computer capability increased and thus demand for microdata. Further, the world wide web and social networks handle with a big amount of individual data which are quite easily retrievable for every internet user. So-called web-scraping techniques were developed to search for specific information on more or less each topic and gather them in an unstructured data conglomerate (see e.g. Kinne and Axenbeck 2020; Kinne and Lenz 2021; Shigapov et al. 2021). To make these unstructured data usable for empirical studies they must be transformed into a usable, structured form. Kinne and Axenbeck (2020) and Kinne and Lenz (2021) demonstrate how to measure firms' innovation activities by using information from firms' websites. The authors describe the web scraping (or web mining) process and the transformation method they apply for generating a structured database from an unstructured source. The validity of their innovation indicator was evaluated and the model was trained with data of the MIP.[5]

Further, Mirtsch et al. (2021) analyse the adoption of the international information security management system standard ISO/IEC 27001 with the help of this web mining approach. The output of our study is applied to this project when we combine the constructed ISO norm indicator with the firm data of the MIP. This can easily be done because the authors of the ISO norm project used the same firm identifier as the MIP. This identifier originates from the firm data base of the credit rating agency Creditreform which forms the basis of ZEW's Mannheim Enterprise Panel (MUP, Bersch et al. 2014). The MUP is the most comprehensive micro database of firms in Germany and is also the sampling frame of the MIP. The underlying study of Mirtsch et al. (2021) used firms' internet addresses which are part of the MUP. Therefore, we can be quite confident that we are observing a significant proportion of firms that operate websites. We postulate that we observe almost all firms that registered for a managerial ISO norm certificate. Mirtsch et al. (2021) found a total of 47,919 firms that refer to at least one of the different ISO managerial system standards on their website.

Combining their ISO norm indicator with the MIP yields a new research data set with these additional publicly available pieces of information. This creates a growing disclosure risk for the individual firms participating in the MIP survey. Therefore, it is necessary to estimate the disclosure risk before disseminating the data to external

[5]The new approach has e.g. been applied for measuring the effects of the Corona pandemic on firm behaviour, which was reflected on the firms' websites (Dörr et al. 2022)

researchers. We design different attack scenarios to disclose individual firms where the MIP SUF is linked with external data: (a) the MUP and (b) the Mannheim Web Panel which will be described in Section 3. Our approach uses graph theory for solving the so-called bipartite matching problem. We calculate similarity scores and use other heuristics to identify suitable candidates to disclose firms contained in the combined MIP and ISO data set. Our results show that the provision of web indicators in the MIP SUF increases the disclosure risk of firms in specific attack scenarios, but the disclosure risk remains at a low level.

The remainder of this paper is structured as follows: In section 2 the related strands of literature are summarised. Section 3 presents the necessary data sets. Section 4 gives an overview of the methods. In Section 5, the results of the matching are presented. Section 6 discusses our findings and presents paths for future work. Lastly, Section 7 summarises and concludes this paper.

2 Related Literature

2.1 Research on firm data disclosure risk

Ronning et al. 2005 conducted one of the projects measuring disclosure risk for firms in official firm data sets. Further, the project team developed anonymisation techniques or adapted existing ones and found a trade-off between data protection and data quality. Official firm data sets are available within the research data centers of the German official statistics system. Scientific-use-files can also externally be used by researchers when firms' micro data had been anonymised in such a way that individual firms cannot be identified by the data users anymore. Common anonymisation techniques were tested. Typically, the techniques aggregate or even disturb information to avoid unique variable combinations which may facilitate disclosure of specific firms. Consider R&D-intensive pharmaceutical firms in a sparsely populated rural area which could be identified if firm information were published without applying anonymisation techniques. Different matching techniques were developed or existing ones were used to link the firms contained in these anonymized scientific-use files with publicly available firm information in order to estimate their disclosure risk (e.g. Gottschalk 2004; Höhne et al. 2003; Müller et al. 1991; Wirth 1992). The additional information about firms originated from commercial firm registers (e.g. Bureau van Dijk's MARKUS). It became obvious that some previously common applied anonymisation techniques had to be adjusted to ensure data protection issues on the one hand and to retain data quality on the other. Research

data centers of official statistical offices in Germany applied the results of this studies when generating scientific-use-files of firm data. Research data centers outside the official statistics – e.g. ZEW-FDZ - also made use of the conclusions and adjusted the generating process for firm data scientific-use-files. Ever afterwards, the amount of firm information has increased steadily via the Internet. Hence, disclosure risk for firms in scientific-use-files might have been risen, too.

More recent studies focus on other scenarios such as disclosure attacks on individuals (e.g., Carey et al. 2023; Dankar et al. 2012; Li et al. 2023; Rocher et al. 2019; Sondeck and Laurent 2025), importers in foreign trade statistics (e.g. Favato et al. 2022), social networks (e.g. Lee et al. 2017), patient data (e.g. Benitez and Malin 2010), public platform data (e.g. Narayanan and Shmatikov 2006) or present a disclosure attack on anonymised texts using machine learning (e.g. Manzanares-Salor et al. 2024). Others give an overview of the research field (e.g. Gadotti et al. 2024).

2.2 Graph theory and the bipartite matching problem

Graph theory is a part of discrete mathematics and theoretical computer science (e.g. Bondy and Murty 1976; Diestel 2005). The basic principles go back to the work of Leonhard Euler and the famous *Königsberg bridge problem*. Nowadays, the field of social work analysis is also strongly related to graph theory as it often uses graph theory as a fundamental (e.g. Scott 2017; Wasserman and Faust 1994). The possible applications of graph theory are numerous and highly interdisciplinary.

The building blocks of graph theory are the so-called nodes and edges. The nodes are capable to model a variety of entities, such as cities or firms. The edges model relationships between the nodes, e.g. road connections or cooperations. There are various special forms of graphs which are adapted to application examples, e.g. in economics. For example, Breithaupt et al. (2023) model employee flows between firms as graphs and create a Linked-Employer-Employee (LEE) data set. Similarly, Abbasiharofteh et al. (2021) model firms and their relationships that are extracted from hyperlinks on firm websites. However, there are many comparable studies.

Bipartite graphs are a special form of graphs as the nodes are divided into two different types (Zha et al. 2001), e.g. firms in West and East Germany. A second example could be two sets of firms that have a yet unidentified overlap. More generally, in this study we want to find a solution to a bipartite graph matching problem (e.g. Doherr 2023; Riesen and Bunke 2009; Tanimoto et al. 1978). The objective is to determine a good or optimal matching between two sets of nodes, whereby additional constraints often have to be met. This type of matching problems is of-

ten solved using methods from the field of linear optimisation (e.g. Padberg 1999). Since we are interested in hard assignments between the node sets, the problem to be solved is in the subarea of (mixed) integer programming (e.g. Greenberg 1971).

Unfortunately, the methods for the exact solution of the matching problem have a high level of complexity. The complexity is usually described with the Big-O notation (e.g. Knuth 2005), which is a specific case of the more general Bachmann–Landau notation. The required run-times to solve matching problems are often very high and usually belong to the complexity class NP (e.g. Ladner 1975). Therefore, they can often not be solved with integer programming in a timely manner, e.g. in the case of very large bipartite graphs. For this reason, so-called heuristics are used. These aim to solve difficult problems well, but not optimally. For example, constraints may be broken or the maximum value of the objective function may not be reached (e.g. Gilli and Winker 2009). There are various heuristics for bipartite graph matching, e.g. based on a reduction to a *minimum cost flow* problem (Schwartz et al. 2005).

Our paper makes a contribution to the *firm data disclosure risk* literature by using a novel web-based data set and methods from computer science and mathematics. By that, we provide research data centers and legislative bodies with helpful and up-to-date information on the underlying risks when publishing scientific-use-files.

3 Data

In the following, the MUP data (Section 3.1), MWP data (Section 3.2), MIP data (Section 3.3), as well as some data insights (Section 3.4) are presented.

3.1 Mannheim Enterprise Panel (MUP)

The Mannheim Enterprise Panel (MUP) is a long-running panel data set on business activity in Germany (Bersch et al. 2014). Almost all economically active German firms are included in the MUP, e.g. about 3.1 million in 2023. The MUP data are provided by Creditreform (Creditreform 2024), which is a German credit rating agency. The data are utilised in a variety of studies, e.g. the Mannheim Innovation Panel and Mannheim Web Panel. Firms with less than five employees are discarded because they are not part of the MIP. In this study, we use the MUP data for the year 2019, i.e. the processed firm data are available for about 2.5 million firms. Each firm has several characteristics: We use the number of employees, NACE codes, and district identifiers. Missing employee counts are filled with data about the

firm from the previous and following three years. Firms with less than five employees are neglected in the subsequent steps. The NACE code is used to determine the industry sector (21 groups), the employee count is implemented to identify size classes (3 groups) and the district identifier is used to determine whether the firm is located in East or West Germany (2 groups). Missing rates are varying, e.g. data on employee counts are more often missing. The MUP is selected for this study as comparable data sets, such as ORBIS, are accessible by the public using a fee-based licence (BVD 2024). ORBIS and MUP share firm identifiers because they both stem from Creditrefrom.

3.2 Mannheim Web Panel (MWP)

The Mannheim Web Panel (MWP) has been collected since 2018 (ZEW 2024b). For this purpose, a subset of the complete MUP is used. The subset contains firms with data on the firm websites. We have access to the website texts and selected meta data of the websites, e.g. hyperlinks, languages, timestamps, errors, and titles of sub pages. We create six types of indicators: First, we use a digitalisation indicator that was trained on news articles from four providers and then applied on the firm websites. The digitalisation scores lie between zero and one (Axenbeck and Breithaupt 2022). Second, we create an East/West Germany indicator. For this, we look at all five-digit numbers on the firm websites and examine whether they might be postcodes from the West or the East of Germany. The result indicates the percentage of postal codes from the West. However, this approach is a simple heuristic and does, therefore, not claim to be error-free. Third, we generate an industry indicator using the texts on firm websites. To do this, we search for keywords associated with each industry. So far, we only search for a list of industry-specific German keywords. Since matches can be found for different industries, we standardise the result per firm by dividing with the total number of keyword matches. This results is an indicator between zero and one and the industry variables sum up to one per firm. Forth to sixth, we search for ISO codes, R&D activity and indications of internationalism on the firm websites. If at least one of the keywords per indicator is found, then the binary indicator is set to one. Seventh, we use a real-valued web-data based indicator measuring the product innovativeness of firms (Kinne and Lenz 2021). The processed data are available for about 1.1 million firms in 2020 (in total, there are 1.7 million firms in the MWP for the year 2017). Table A.1 gives an overview of the firm characteristics derived from the MWP data and details on their extraction.

3.3 Mannheim Innovation Panel (MIP)

We use a scientific use file (SUF) of the Mannheim Innovation Panel (ZEW 2024a). The MIP SUF refers to the year 2019 and makes use of the questions about digital business models that were so far only included once in the survey. The data set comprises data about 5,500 firms and is anonymised using different methods so that the data can be made available to researchers outside of research data centres. The following anonymisation techniques are used in the data set: a) error overlay, b) reporting of intensities and ratios, c) truncating, d) grouping, e) aggregation of information, and f) non-disclosure of information. In addition, users of the MIP scientific use file get access to a documentation file. The anonymisation techniques are described in more detail in Appendix A.5. A description of the MIP variables for product innovation, digitalisation, industry, R&D, employees, internationalism, Eastern/Western Germany, and the web-based ISO code indicator are presented in Table A.2. The SUF is the target file (or estimation sample) that should be disclosed by an attacker using data from the MUP and MWP. For this purpose, the eight MIP SUF variables are used in a matching approach using different data accessibilities (Case 1: MWP, Case 2: MUP, and Case 3: MWP and MUP). By doing so, we simulate attackers with different levels of data access.

3.4 Bringing the data sets together

Firms in the MIP, MWP and MUP share an identifier, i.e. the *crefo identifier* assigned by Creditreform. However, the identifier has been deleted from the MIP SUF as one part to assure anonymity. The venn diagram data (Table 1) presents the observation counts of the data sets and their intersections. MIP data are neither completely part of the MWP nor the MUP. Reasons for the differences are that firms do not have a publicly accessible websites or are not listed as economically active in the MUP. Further, only a subset of the MIP observations is part of the MWP and MUP (4.1 thousand). Next, we show firm characteristics of the MIP SUF observations to validate that all types of firms are part of the data set (all industries, etc.). Appendix A.3 shows the summary statistics of the MIP SUF, MUP, and MWP data. Missing data are filled with mean values or uniform data to allow a matching between the data sources. The industry and firm size classes of the MIP are calculated for the MWP and MUP using the same definition. The MIP data covers firms with at least 5 employees and the following industries: mining, manufacturing, energy and water supply, disposal, wholesale trade, transport and postal services, information and communication, financial services, and business-

Data Set	Observations	Explanation
MIP	5.5 thousand	MIP SUF data set for 2019 (survey 2020).
MWP	1.1 million	MWP data set for 2020 (1.7 million). About 1.4 million of firms have texts. 1.5 million firms marked as German.
MUP	2.5 million	MUP data set for 2019 (1.2 million firms fall into industries and firm siz groups surveyed in the Mannheim Innovation Panel).
MIP ∩ MUP	4.9 thousand	The match between both data sets was not possible for some firms, e.g. because of different data collection time points.
MIP ∩ MWP	4.1 thousand	Some MIP firms have no websites or the match between both data sets was not possible, e.g. the identifier has changed.
MWP ∩ MUP	0.7 million	Some MUP firms have no websites or the match between data sets was not possible, e.g. different observation time points.
MIP ∩ MWP ∩ MUP	4.1 thousand	All of the explanations above apply.

Table 1: Number of observations in the data sets (MIP, MWP, and MUP) and their respective intersections. The presented intersections are identified using a match based on the website URLs of firms and are not known to attackers.

related services. However, retail trade, construction, and consumer-oriented services are not part of the sampling frame. The different industry shares of the total sample lie between 2 (industry: glas) and 9 (industry: transport) percent. Most of the firms have less than fifty employees and only ten percent have at least 250 employees. Most of the firms are located in Western Germany and 34 percent were product innovators within the last three years. The average digitalisation score is 30 percent and about 19 percent of the firms mention at least one of the relevant ISO codes on their website. Almost half of the MIP firms have international business activities. The MWP data set describes a large share of firms in the German economy. There are some differences to the MIP estimation sample for the industries and other characteristics (see Kinne and Axenbeck 2020; Kinne and Lenz 2021). For example, larger differences are found for the ISO code indicator, i.e. 12% for the processed MWP data vs. 19% in the MIP SUF. One explanation is the different composition (regions, industries, firm sizes, etc.) of the MIP SUF and the MWP data. The MUP data set is representative for Germany because almost all firms are included (Bersch et al. 2014). In summary, the presented statistics appear plausible and the data from the MUP and MWP can be used for an attack on the MIP.

4 Methods

This section presents the calculation of similarity scores for firms (Section 4.1), and heuristics for the selection of suitable candidates (Section 4.2).

4.1 Calculation of similarity scores

Every firm x has a set of characteristics $x_1, ..., x_k$ that consists of natural and real-valued numbers, e.g. the number of employees or the industry affiliation. Categorical variables are 1-hot encoded (a set of binary dummy variables is created) so that they can be represented with vectors consisting of numbers. For each firm, the list of characteristics is transformed to a vector $x = (x_1, ..., x_k)^T$ so that we can compare k firm characteristics for pairs of firms. Some firm characteristics exist for the MIP as well as for the MUP firms making comparisons across data sets possible. However, the data source and sources for errors of the characteristics might be different. For example, the MIP employee count is recorded in the annual survey (ZEW 2024a). The employee count in the MUP is determined by Creditreform (Creditreform 2024).

Similarity functions measure the likeness of two firms, modelled with the two firm-specific vectors x and y. The similarity score is usually real-valued, positive, and between zero and one. In our case, we use the well-established cosine similarity to measure the angle between the two vectors (e.g. Pedregosa et al. 2011). The closer the angles between two vectors are to each other, the more similar are the cosine similarity scores (see Equation 1). As a result, the scores can be calculated for pairs of firms within one or across two data sets, e.g. MIP and MUP. The scores are used as an indication of whether two firms in different data sets describe the same firm. Furthermore, we also experiment with the Manhattan, Mahalanobis and Euclidian distance measures. Since we calculate distances instead of similarity measures, we have to reverse the scale of the functions (e.g. V. Kumar et al. 2014).

$$Cosine(x, y) = \frac{\sum_{i=1}^{n} x_i * y_i}{\sqrt{\sum_{i=1}^{n} (x_i)^2} * \sqrt{\sum_{i=1}^{n} (y_i)^2}} \tag{1}$$

4.2 Selection of suitable candidates

We use a bipartite graph to model the data about firm similarity. In this paper, the set of nodes consists of two disjoint sets V_1 and V_2 (see e.g. Diestel 2005). First,

the MIP firms and, second, the MUP firms[6]. The edges between two nodes and their weights represent the pairwise similarity score, i.e. $w_{x,y}$ for the nodes (or firms) x and y. Figure 1 shows an example of a bipartite graph. The graphs that we consider in the subsequent analysis are significantly larger. Furthermore, the node set of the MUP firms is considerably larger. The different node colors illustrate the disjoint sets consisting of MIP or MUP firms. For each node in the red node set, all scores in combination with the blue nodes are calculated. Because a high similarity score

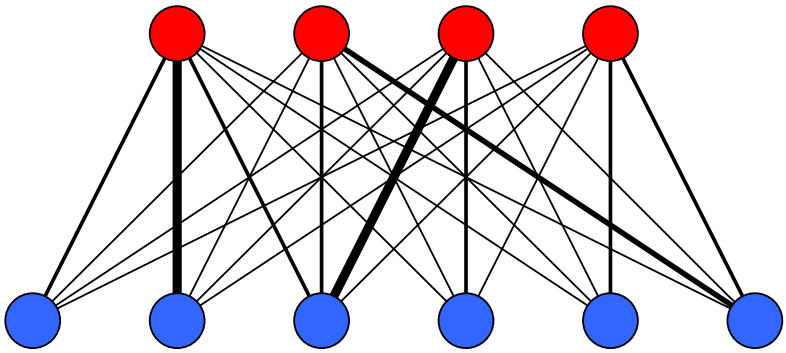

Figure 1: Bipartite graph (example): Red nodes refer to MIP firms; blue nodes to MUP firms. Edge weights are modelled as the edge thicknesses. The figure was created with "yED graph editor" (https://www.yworks.com/products/yed).

between one MIP firm and a large number of MUP firms might be calculated, the similarity scores can be standardised (optional). These might be firms with widely distributed characteristics in the German population of firms. To do this, we divide the edge weight of a each MIP node by the sum of all edge weights adjacent to that node (see Equation 2). By doing so, we penalise high and ambiguous similarity scores between one MIP firm and many MUP firms. As a side effect, the edge weights for each MIP node in V_1 sum up to exactly one making the interpretation of the results in some cases easier. Further changes can be made to favor other graph properties, e.g. favoring links to certain industries because we know the distributions of the MIP and complete population of German firms.

$$w_{x,y} = \frac{w_{x,y}}{\sum_{z \in V_2} w_{x,z}}; x \in V_1, y \in V_2 \tag{2}$$

Next, we want to find the best or at least a good assignment between nodes of the

[6]The MWP firms are a subgroup of the MUP. Hence, we generalize the description of the matching procedure. The procedure is repeated by using only MWP firms, hence, with additional information stem from web scraping.

set V_1 and V_2 by using the bipartite graph as a foundation for the optimisation. For this, a linear optimisation problem needs to be solved. An overview on this research field is provided by Papadimitriou and Steiglitz (1998). More specifically, we need to solve an integer programming (IP) problem, because either an assignment between two nodes should take place or not. Solving the problem is NP-hard and, therefore, usually associated with long run times to find the optimal solution. Equation 3 presents the formal definition of the IP problem. We try to find the edges that maximise the sum of edge weights. The variable $e_{x,y}$ indicates if an edge is selected and takes the values zero or one, i.e. edges are either selected or not (constraint 1). By that, we avoid soft or fuzzy assignments, i.e. assignments between both node sets modelling probabilities. By adding this constraint, the problem is not solvable using methods like gradient descent. Further, we make sure that for each node of V_1 exactly one adjacent edge leading to V_2 is selected (constraint 2). Not every node of V_2 needs to be linked to a node of V_1 (constraint 3). This translates to a assignment of the MIP firms to one MUP firm each, but not every MUP firm needs to be linked to a MIP firm. In doing so, we implicitly assume that there are no duplicates within the two node sets. The case that a MIP firm cannot be linked to the MUP is not explicitly modeled at this point. Instead, it is possible to eliminate assignments in a subsequent step that do not meet minimum standards.

$$
\begin{aligned}
&\max \textstyle\sum_{x \in V_1} \sum_{y \in V_2} w_{x,y} e_{x,y} \\
&\\
&\text{so that:} \\
&(1)\ e_{x,y} \in \{0,1\} \\
&(2)\ \textstyle\sum_{x \in V_2} e_{j,x} = 1;\ \forall j \in V_1 \\
&(3)\ \textstyle\sum_{x \in V_1} e_{x,i} \in \{0,1\};\ \forall i \in V_2
\end{aligned}
\tag{3}
$$

Solving the IP is difficult because of the size of the graph and required computing power. Therefore, we use two simple heuristics that determine an approximation to the optimal solution. However, we can not calculate a lower bound for the quality of the approximation. Algorithm 1 presents the pseudo code for the first heuristic. First, the similarity scores W are sorted in descending order and an empty list for the matches R are created. Second, we loop over the ordered list of similarity scores and iteratively choose the largest similarity score W_i. We take the edge and store it in the result list. After that, we delete all edges from W that start in the node x of V_1 or end in the node y of V_2. The heuristic yields results that fulfill the constraints 1 to 3, but might not find the maximum sum of the edge weights. Algorithm 2

presents the pseudo code for the second heuristic. The heuristic selects the highest weight $w_{x,y}$ for the MIP firm x. If there are multiple hits of the same quality, we select the oldest MUP firm y. This fulfills constraints 1 and 2, but an MUP firm can be assigned to several MIP firms, i.e. condition 3 is not fulfilled.

Algorithm 1 Heuristic to solve IP problem (greedy algorithm).

$W = \{w_{x,y} | x \in V_1, y \in V_2\}$.
Sort W descending.
Create empty result list R.
for $W_i \in W$ **do**
    Assume that W_i refers to the entry $w_{x,y}$ ($x \in V_1, y \in V_2$).
    Store $w_{x,y}$ in R.
    Delete the entries $\{w_{x,m} | m \in V_2\}$ and $\{w_{m,y} | m \in V_1\}$ from W.
end for
Return R.

Algorithm 2 Heuristic to solve IP problem (best matching oldest firm).

$W = \{w_{x,y} | x \in V_1, y \in V_2\}$.
Keep for each firm x the highest $w_{x,y}$ and store them in R.
If multiple $w_{x,y}$ exist for an firm x (same score), keep the entry for oldest firm y.
Return the filtered set R consisting of assignments.

Figure 2 presents the result of the heuristic, i.e. a visualisation of the selected edges. For reasons of clarity, the edge weights are not shown in the figure. The edges correspond to the content of the list R. In this example, each node from V_1 was connected to a node from V_2 by means of an edge. The weights were maximised using one of the two heuristics, but the edge sum does not have to be the global maximum. Opportunities for model extensions are, for example, to add a minimum similarity score and to weaken constraint 2. For this, we have to adjust the constraints of the IP, so that not every node from V_1 necessarily gets assigned a node from V_2.

5 Results

Before we match the MIP to the MUP data, we define the importance of firm characteristics. By doing so, we assign a higher or lower importance to the variables. We assign a high importance to the industry, ISO-code, and location (East/West). Medium importance is assigned to the employee count. The other variables receive a low importance. We implement this by multiplying the variables with 1 (low im-

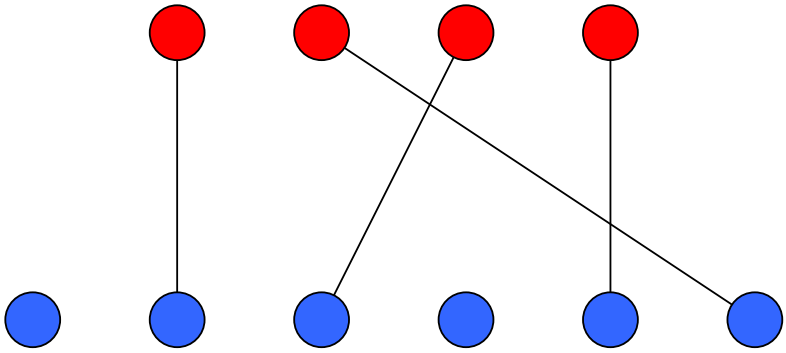

Figure 2: Selected matches (example): The red nodes refer to MIP firms; blue nodes to the MUP firms. Edge weights are not shown. The figure was created with "yED graph editor" (https://www.yworks.com/products/yed).

portance), 2 (medium importance), or 3 (high importance) before matching. This is similar to scaling the respective dimensions of the vectors. The variable importance can also be modified if necessary. For computational reasons, not all similarity scores can be calculated. We restrict the calculations to plausible subsets of candidates. For a MIP SUF firm in a certain industry, only candidates that are with at least a twenty percent probability in the same industry are considered. Next, we consider three cases for the firm matching: (1) only the MWP, (2) only the MUP and (3) both the MUP and MWP data. Depending on the case, we use different variables and data sources for the matching. Table 2 shows the utilisied variables by case. Case 1 and 2 are straightforward because only one data source is used. Case 3 makes use of MWP and MUP data. The ISO code, the internationalism, product innovation, R&D as well as digitalisation indicator are based on the MWP data. The other variables rely on the MUP data (see brackets). The industry and Western/Eastern Germany indicators exist in the MUP and MWP data.

Cases	Variables
Case 1: MWP	ISO Code, Industry, Digitalisation score, R&D, Product innovation, Internationalism, West/East Germany.
Case 2: MUP	Industry, West/East Germany, Employee count.
Case 3: MUP & MWP	ISO Code (MWP), Industry (MUP), Employee count (MUP), West/East Germany (MUP), Internationalism (MWP), Product Innovation (MWP), R&D (MWP), Digitalisation score (MWP).

Table 2: List of variables differentiated by cases (MWP, MUP, MUP and MWP). The three cases represent the attackers' access to the firm-specific data.

Table 3 shows the results of the attack using the cosine similarity function and (not) performing a standardisation of the similarity scores. The number of correct

matches between the MIP SUF and the MUP firms is small and lies between 0 and 21. Our success rate increases when data from additional sources, such as the MUP, are included. The number of correct matches is larger than the results of a baseline model that uses a random selection of MUP firms as matching strategy, i.e. about $5,448$ of $2,500,000 \approx 0.002$ firms of the MIP SUF would be disclosed. If we do not standardise the similarity scores, the results improve slightly. The second algorithm provides in all cases better matching results. The results also do not improve in the MWP case if the East-West variable is neglected. The summary statistics show a particularly large difference for this particular variable. Table A.4 shows the results if individual variables are discarded. This highlights the relevance of firm characteristics in the disclosing procedure. Firm size and the East/West indicator appear to be the most important. The ISO indicator and the international orientation of the firms are in the middle of the field. Digitalisation and R&D activities have hardly any effect. The inclusion of data for product innovation activity is apparently even counterproductive in the disclosing of firms.

A similar pattern emerges when we reduce the number of MUP candidates using a filter. We take advantage of the fact that the MIP survey only contains active firms with at least 5 employees and certain industries are not covered. By doing this, we reduce the size of the candidate list. The results indicate that only few firms are identified in the MUP. The disclosure risk is not explicitly shown in the following table because the numbers are similar to the main results of our paper.

Setup of the data attack			**Correct matches in target data set**	
			MUP	
Cases	Standard.	Algorithms	N	Percent
MUP	No	Algorithm 1	3	0.06%
MWP	No	Algorithm 1	0	0.0%
MUP&MWP	No	Algorithm 1	21	0.39%
MUP	Yes	Algorithm 1	2	0.04%
MWP	Yes	Algorithm 1	1	0.02%
MUP&MWP	Yes	Algorithm 1	9	0.17%
MUP	No	Algorithm 2	5	0.09%
MWP	No	Algorithm 2	0	0.0%
MUP&MWP	No	Algorithm 2	16	0.29%
MUP	Yes	Algorithm 2	5	0.09%
MWP	Yes	Algorithm 2	1	0.02%
MUP&MWP	Yes	Algorithm 2	11	0.2%

Table 3: Matching results for three cases and the baseline model. The results are based on the cosine similarity. Baseline: Pick MWP/MUP firms at random.

In the next step, we vary the level of uncertainty by considering not only the best candidate but a sorted list of the best MUP candidates per MIP firm. Figure 3 shows the results for candidate lists of increasing sizes. Candidate lists consist of the MUP candidates with the X highest scores for each MIP SUF firm. As expected, the percentage of correct matches in the list increases with the size of the candidate list. The candidate list of size one is strongly related to the results from Table 3.

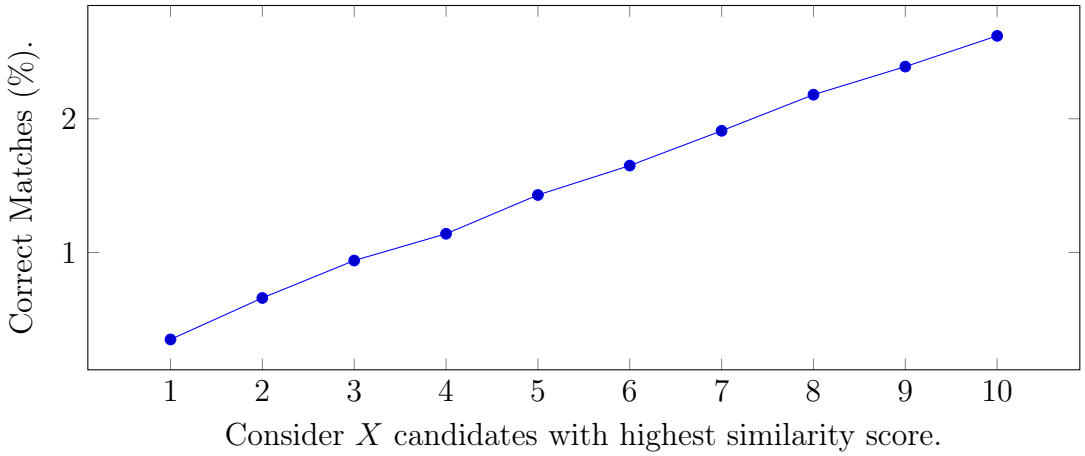

Figure 3: Size of candidate lists vs. correct matches (%). Setup: MUP and MWP features (Case 3), cosine similarity, no standardisation. Reading help: We find the correct match in 2.62% of the cases in top 10 candidate list per MIP firm.

Finally, we consider the problem that the number of candidates could be further reduced by skillful filtering or using external data. For this purpose, we create synthetic MUP data of different sizes. The data sets contain all firms of the MIP estimation sample and a variable number of other MUP firms that are randomly selected. Figure 4 shows the size of the MUP candidate list in relation to the matching score. If the candidate list is reduced, then the matching result is substantially better. From 1,000 additional MUP candidates onwards, it becomes apparent that the quality of the matches decreases rapidly. The most left dot is strongly related to a search restricted to the MIP SUF and the most right dot corresponds to a search on the entire MUP. Our findings are comparable with results in Table 3 (case 3).

The results can be partly explained by the fact that the characteristics of firms are not or only partially sufficient to separate them in the MIP SUF. For example, if all firms are grouped according to the characteristics with a high or medium weight (industry, ISO code, employee count, and East/West Germany), then only less than one percent of firms are alone in a group. In the other cases, the matching is, by definition, ambiguous. The described problem is mitigated by firm characteristics with low weights, but these are also much more prone to errors or uncertainties

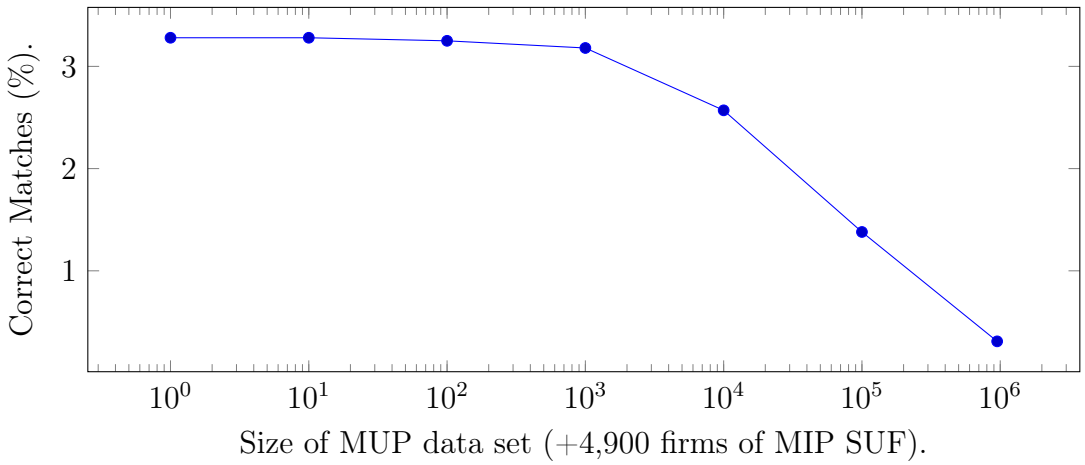

Figure 4: Size of MUP data set vs. correct matches. Setup: MUP and MWP features (Case 3), cosine similarity, no standardisation. Reading help: We find the correct match in 1.4% of cases if we search in 104,900 MUP firms.

because they are calculated on MWP data. MUP data has the same problem, i.e. almost no group contains only one firm. Therefore, the exact match of one MIP SUF firm to one MUP firm is, in many cases, not possible by definition. One possible way of solving this problem is to expand the set of firm characteristics, e.g. using external data. In summary, our proposed matching approach can not be used to disclose a large share of the MIP SUF in a third party data set such as the MUP. Our results hold for disclosure attacks using information from the MUP and MWP.

6 Discussion and Future Work

In addition to the anonymisation efforts for the SUF, the other data sets presented also have various types of inaccuracies or errors that make matching between the data sets even more difficult. For example, in the MUP certain information such as the number of employees is often missing or partly approximated. Handling missing values is a large problem and different methods were tested to tackle this issue, e.g. removing the observations or imputing zeros, sample averages or uniform data. Firms usually report very positively about business activities on their websites. One well-known example is corporate greenwashing or overstating digitalisation efforts (e.g. Vos 2009). The time delay in recording the observation is another source of error. Survey data in the MIP relates to the last three years and website information, on the other hand, are snapshots of the most current state.

Furthermore, we only consider one specific attack scenario. This cannot provide a comprehensive and conclusive assessment of the actual risk, but rather serves as a guide. Strictly speaking, the study only provides a lower bound for the disclosure

risk. However, our starting point is quite good, as we know exactly which data sets form the basis of the MIP SUF and thus enable the construction of a high-quality data set for the matching. Such high-quality data are not available for all SUFs. In addition, previous studies were also unable to determine an exact upper bound for the disclosure risk. Furthermore, the attacked data set plays an important role. Some data sets are very specific, which might severely limit the search space and generally facilitates the quality of the results. The disclosure risk for personal data might be different, as the pool of candidates is often larger. Another dimension that has not yet been considered is the fact that some SUFs might be available as a panel which could change the disclosure risk.

Our paper brings up the question of a reasonable trade-off between anonymisation and dislosure risk. Ideally, researchers should be provided with realistic data without risking the disclosure of larger parts of the data. Our paper gives evidence that the anonymisation level could possibly be reduced (at least for certain groups of firms). We showed that not all firms in the data set are subject to a notable disclosure risk. Larger firms and companies from the manufacturing industry are most frequently disclosed. Immediately, the question arises whether these groups of firms should be anonymized to a greater extent and others to a lesser extent. However, the aspiration to totally avoid disclosure would disproportionately destroy the quality of the data. Hence, in our view, a certain degree of uncertainty should be accepted.

Future work consists of two fields. First, defining a smaller subproblem covering fewer firms that can be solved exactly using the integer programming methodology. Second, an attack against the use of the differential privacy (e.g. Kearns and Roth 2019) searching for a better trade-off between anonymisation and disclosure risk.

Our paper makes a contribution to the *disclosure risk of firm data* literature by using MUP and MWP data to disclose MIP SUF firms. By that, we provide research data centers and legislative bodies with helpful and up-to-date information on the underlying risks when publishing scientific use files. However, there is still plenty of potential for future research. This includes further data sets from which additional attributes can be extracted for the matching or other matching approaches, e.g. based on machine learning or using other similarity measures.

7 Conclusion

In this paper, we perform a disclosure attack against the Mannheim Innovation Panel (MIP) SUF using data from the Mannheim Enterprise Panel (MUP) and

Mannheim Web Panel (MWP). In a carefully selected set of data attack scenarios, we show that there is only a low disclosure risk, even if additional information from firms' websites is added (here: ISO norm indicator). In all presented cases, less than half a percent of firms are disclosed. The risk changes slightly if data from the MWP are considered in addition to MUP data. Hence, large-scale disclosure is not possible without unusually high costs and waste of time and energy. Thereby, our paper makes a contribution to the firm data disclosure risk literature and provides research data centers and legislative bodies with helpful and up-to-date information on the underlying risks when publishing scientific use files in the future.

8 Literature

Abbasiharofteh, M., Kinne, J., & Krüger, M. (2021). *The strength of weak and strong ties in bridging geographic and cognitive distances* (ZEW Discussion Paper No. 21-049) (Preprint (Online); last accessed 26.03.2024). ZEW - Leibniz Centre for European Economic Research. Mannheim, Germany. http://ftp.zew.de/pub/zew-docs/dp/dp21049.pdf

Axenbeck, J., & Breithaupt, P. (2022). *Measuring the Digitalisation of Firms – A Novel Text Mining Approach* (ZEW Discussion Paper No. 22-065) (Preprint (Online); accessed 26.03.2024). ZEW - Leibniz Centre for European Economic Research. Mannheim, Germany. http://ftp.zew.de/pub/zew-docs/dp/dp22065.pdf

Bandara, P. K., Bandara, H. D., & Fernando, S. (2020). Evaluation of re-identification risks in data anonymization techniques based on population uniqueness [Preprint (Online); last accessed 22.05.2024]. *2020 5th International Conference on Information Technology Research (ICITR)*, 1–5. https://ieeexplore.ieee.org/document/9310884

Benitez, K., & Malin, B. (2010). Evaluating re-identification risks with respect to the HIPAA privacy rule. *Journal of the American Medical Informatics Association : JAMIA*, *17*, 169–77. https://doi.org/10.1136/jamia.2009.000026

Bersch, J., Gottschalk, S., Müller, B., & Niefert, M. (2014). *The Mannheim Enterprise Panel (MUP) and Firm Statistics for Germany* (ZEW Discussion Paper No. 14-104) (Preprint (Online); last accessed 26.03.2024). ZEW - Leibniz Centre for European Economic Research. Mannheim, Germany. https://ftp.zew.de/pub/zew-docs/dp/dp14104.pdf

Bondy, J. A., & Murty, U. S. R. (1976). *Graph Theory With Applications*. Elsevier Science Ltd/North-Holland, ISBN: 0444194517.

Brand, R. (2000). Anonymität von Betriebsdaten: Verfahren zur Erfassung und Maßnahmen zur Verringerung des Reidentifikationsrisikos (BeitrAB 237). In *Beiträge zur Arbeitsmarkt- und Berufsforschung*. www.statistischebibliothek.de/mir/servlets/MCRFileNodeServlet/DEMonografie_derivate_00000258/5106155-9783824606993.pdf

Breithaupt, P., Hottenrott, H., Rammer, C., & Römer, K. (2023). *Mapping employee mobility and employer networks using professional network data* (ZEW Discussion Paper No. 23-041) (Preprint (Online); last accessed 26.03.2024). ZEW - Leibniz Centre for European Economic Research. Mannheim, Germany. https://ftp.zew.de/pub/zew-docs/dp/dp23041.pdf

BVD. (2024). ORBIS data set. Link to Moody's Analytics website (data provider). https://www.moodys.com/web/en/us/capabilities/company-reference-data/orbis.html

Carey, C., Dick, T., Epasto, A., Javanmard, A., Karlin, J., Kumar, S., Muñoz Medina, A., Mirrokni, V., Nunes, G. H., Vassilvitskii, S., et al. (2023). Measuring re-identification risk. *Proceedings of the ACM on Management of Data*, *1*(2), 1–26. https://dl.acm.org/doi/10.1145/3589294

Creditreform. (2024). Mannheim Enterprise Panel data set. Creditreform website (data provider). https://www.creditreform.de

Dankar, F. K., El Emam, K., Neisa, A., & Roffey, T. (2012). Estimating the re-identification risk of clinical data sets. *BMC medical informatics and decision making*, *12*, 1–15. https://pubmed.ncbi.nlm.nih.gov/22776564/

Diestel, R. (2005). *Graph Theory (3rd ed.)* Springer-Verlag, Heidelberg, Germany, ISBN: 978-3-662-53621-6.

Doherr, T. (2023). *The SearchEngine: A Holistic Approach to Matching* (ZEW Discussion Paper No. 23-001) (Preprint (Online); last accessed 26.03.2024). ZEW - Leibniz Centre for European Economic Research. Mannheim, Germany. http://ftp.zew.de/pub/zew-docs/dp/dp23001.pdf

Dörr, J. O., Kinne, J., Lenz, D., Licht, G., & Winker, P. (2022). An integrated data framework for policy guidance during the coronavirus pandemic: Towards real-time decision support for economic policymakers. *PLOS ONE*, *17*(2), e0263898. https://doi.org/10.1371/journal.pone.0263898

Dwork, C. (2008). Differential privacy: A survey of results. *International conference on theory and applications of models of computation*, 1–19. https://link.springer.com/chapter/10.1007/978-3-540-79228-4_1

EU. (2018). General Data Protection Regulation (website of the Federal Ministry for Economic Affairs and Climate Action). www.bmwk.de/Redaktion/DE/Artikel/Digitale-Welt/europaeische-datenschutzgrundverordnung.html

Favato, D. F., Coutinho, G., Alvim, M. S., & Fernandes, N. (2022). *A novel reconstruction attack on foreign-trade official statistics, with a Brazilian case study* (arXiv preprint arXiv:2206.06493) (Preprint (Online); last accessed 22.05.2024). https://www.researchgate.net/publication/364083769_A_novel_reconstruction_attack_on_foreign-trade_official_statistics_with_a_Brazilian_case_study

Gadotti, A., Rocher, L., Houssiau, F., Creţu, A.-M., & De Montjoye, Y.-A. (2024). Anonymization: The imperfect science of using data while preserving privacy.

Science Advances, *10*(29), eadn7053. https://www.science.org/doi/10.1126/sciadv.adn7053

Gilli, M., & Winker, P. (2009). Heuristic optimization methods in econometrics. In Belsley & Kontoghiorghes (Eds.), *Handbook of computational econometrics* (pp. 81–119). Wiley Online Library, Chichester, England. https://onlinelibrary.wiley.com/doi/book/10.1002/9780470748916

Gottschalk, S. (2004). *Unternehmensdaten zwischen Datenschutz und Analysepotenzial*. Nomos, Baden-Baden, Germany, ISBN: 978-3-8329-1459-2.

Greenberg, H. (1971). *Integer programming (1st Edition, Volume 76)*. Elsevier Science, New York/London, ISBN: 9780080955858.

Höhne, J., Sturm, R., & Vorgrimler, D. (2003). Konzept zur Beurteilung der Schutzwirkung faktischer Anonymisierung. *Wirtschaft und Statistik*, *2003 (4)*, 287–292. https://www.destatis.de/DE/Methoden/WISTA-Wirtschaft-und-Statistik/2003/04/faktische-anonymisierung-042003.pdf

Kearns, M., & Roth, A. (2019). *The ethical algorithm: The science of socially aware algorithm design*. Oxford University Press.

Kinne, J., & Axenbeck, J. (2020). Web mining for innovation ecosystem mapping: a framework and a large-scale pilot study. *Scientometrics*, *125*, 2011–2041. https://doi.org/10.1007/s11192-020-03726-9

Kinne, J., & Lenz, D. (2021). Predicting innovative firms using web mining and deep learning. *PLOS ONE*, *16*(4), e0249071. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0249071

Knuth, D. E. (2005). *The Art of Computer Programming: Fundamental Algorithms (Volume 1)*. Addison-Wesley Longman, Amsterdam, ISBN: 9780201896831.

Kumar, V., Chhabra, J. K., & Kumar, D. (2014). Impact of distance measures on the performance of clustering algorithms. *Intelligent Computing, Networking, and Informatics: Proceedings of the International Conference on Advanced Computing, Networking, and Informatics, India, June 2013*, 183–190. https://www.researchgate.net/profile/Vijay-Chahar/publication/284938375_Impact_of_Distance_Measures_on_the_Performance_of_Clustering_Algorithms/links/599da590aca272dff12fc948/Impact-of-Distance-Measures-on-the-Performance-of-Clustering-Algorithms.pdf

Ladner, R. E. (1975). On the structure of polynomial time reducibility. *Journal of the ACM (JACM)*, *22*(1), 155–171. https://dl.acm.org/doi/10.1145/321864.321877

Lee, W.-H., Liu, C., Ji, S., Mittal, P., & Lee, R. (2017). *Quantification of de-anonymization risks in social networks* (arXiv) (Preprint (Online); last accessed 22.05.2024). https://arxiv.org/abs/1703.04873

Li, S., Schneider, M. J., Yu, Y., & Gupta, S. (2023). Reidentification risk in panel data: Protecting for k-anonymity. *Information systems research*, *34*(3), 1066–1088. https://pubsonline.informs.org/doi/10.1287/isre.2022.1169

Machanavajjhala, A., Kifer, D., Gehrke, J., & Venkitasubramaniam, M. (2007). l-diversity: Privacy beyond k-anonymity. *Acm transactions on knowledge discovery from data (tkdd)*, *1*(1), 3–es. https://dl.acm.org/doi/10.1145/1217299.1217302

Manzanares-Salor, B., Sánchez, D., & Lison, P. (2024). Evaluating the disclosure risk of anonymized documents via a machine learning-based re-identification attack. *Data Mining and Knowledge Discovery*, *38*(6), 4040–4075. https://link.springer.com/article/10.1007/s10618-024-01066-3

Mirtsch, M., Kinne, J., & Blind, K. (2021). Exploring the adoption of the international information security management system standard ISO/IEC 27001: a web mining-based analysis. *IEEE Transactions on Engineering Management*, *68*(1), 87–100. https://doi.org/10.1109/TEM.2020.2977815

Müller, W., Blien, U., Knoche, P., & Wirth, H. (1991). *Die faktische Anonymität von Mikrodaten* (Vol. 19). Forum der Bundesstatistik 19, Statistisches Bundesamt, Stuttgart, Germany. https://www.statistischebibliothek.de/mir/servlets/MCRFileNodeServlet/DEMonografie_derivate_00000304/5106155-9783824602315.pdf

Narayanan, A., & Shmatikov, V. (2006). *How to break anonymity of the netflix prize dataset* (arXiv) (Preprint (Online); last accessed 22.05.2024). https://arxiv.org/abs/cs/0610105

Padberg, M. (1999). Linear optimization and extensions. In *Algorithms and Combinatorics (AC, volume 12)*. Springer Berlin, Heidelberg, Germany. https://link.springer.com/book/10.1007/978-3-662-12273-0

Papadimitriou, C. H., & Steiglitz, K. (1998). *Combinatorial Optimization: Algorithms and Complexity*. Dover Publications Inc, Mineola, New York, ISBN: 0486402584.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., & Dubourg, V. (2011). Scikit-Learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830. https://arxiv.org/abs/1201.0490

Peters, B., & Rammer, C. (2023). Innovation panel surveys in germany: The mannheim innovation panel. In *Handbook of innovation indicators and measurement* (pp. 54–87). Edward Elgar Publishing. https://www.elgaronline.com/edcollchap/book/9781800883024/book-part-9781800883024-14.xml

Rammer, C., & Es-Sadki, N. (2023). Using big data for generating firm-level innovation indicators-a literature review. *Technological Forecasting and Social Change*, *197*, 122874. https://www.sciencedirect.com/science/article/pii/S0040162523005590

Riesen, K., & Bunke, H. (2009). Approximate graph edit distance computation by means of bipartite graph matching [7th IAPR-TC15 Workshop on Graph-based Representations (GbR 2007)]. *Image and Vision Computing*, *27*(7), 950–959. https://doi.org/https://doi.org/10.1016/j.imavis.2008.04.004

Rocher, L., Hendrickx, J. M., & De Montjoye, Y.-A. (2019). Estimating the success of re-identifications in incomplete datasets using generative models. *Nature communications*, *10*(1), 3069. https://www.nature.com/articles/s41467-019-10933-3

Ronning, G., Sturm, R., Höhne, J., Lenz, R., Rosemann, M., Scheffler, M., & Vorgrimler, D. (2005). *Handbuch zur Anonymisierung wirtschaftsstatistischer Mikrodaten* (Vol. 4). Statistisches Bundesamt, Wiesbaden, Germany. www.statistischebibliothek.de/mir/servlets/MCRFileNodeServlet/DEMonografie_derivate_00000312/Band4_AnonymisierungMikrodaten1030804059004.pdf

Samarati, P., & Sweeney, L. (1998). *Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression* (Technical report, SRI International) (Preprint (Online); last accessed 22.05.2024). https://epic.org/wp-content/uploads/privacy/reidentification/Samarati_Sweeney_paper.pdf

Schwartz, J., Steger, A., & Weißl, A. (2005). Fast algorithms for weighted bipartite matching. *International Workshop on Experimental and Efficient Algorithms (conference paper)*, 476–487. https://link.springer.com/chapter/10.1007/11427186_41

Scott, J. (2017). *Social Network Analysis (4th Edition)*. SAGE Publications Ltd, ISBN: 9781473952126.

Shigapov, R., Mechnich, J., & Schumm, I. (2021). RaiseWikibase: Fast inserts into the BERD instance. *The Semantic Web: ESWC 2021 Satellite Events, pp. 60–64 (conference paper)*. https://doi.org/10.1007/978-3-030-80418-3_11

Sondeck, L.-P., & Laurent, M. (2025). *Practical and Ready-to-Use Methodology to Assess the re-identification Risk in Anonymized Datasets* (arXiv preprint

arXiv:2501.10841) (Preprint (Online); last accessed 22.05.2024). https://arxiv.org/abs/2501.10841

Tanimoto, S. L., Itai, A., & Rodeh, M. (1978). Some matching problems for bipartite graphs. *Journal of the ACM (JACM)*, *25*(4), 517–525. https://dl.acm.org/doi/10.1145/322092.322093

Vos, J. (2009). Actions Speak Louder than Words: Greenwashing in Corporate America. *Notre Dame Journal of Law, Ethics & Public Policy*, *23*(2), 673–697. http://scholarship.law.nd.edu/ndjlepp/vol23/iss2/13

Wasserman, S., & Faust, K. (1994). *Social Network Analysis: Methods and Applications*. Cambridge University Press, ISBN: 9780511815478.

Wirth, H. (1992). Die faktische Anonymität von Mikrodaten: Ergebnisse und Konsequenzen eines Forschungsprojektes. *ZUMA Nachrichten*, *16*(30), 7–65. https://www.ssoar.info/ssoar/handle/document/20967

Yoon, J., Drumright, L. N., & Van Der Schaar, M. (2020). Anonymization through data synthesis using generative adversarial networks (ADS-GAN). *IEEE journal of biomedical and health informatics*, *24*(8), 2378–2388. https://pubmed.ncbi.nlm.nih.gov/32167919/

ZEW. (2024a). Mannheim Innovation Panel Data (website of research data center). https://kooperationen.zew.de/en/zew-fdz/provided-data/mannheim-innovation-panel

ZEW. (2024b). Mannheim Web Panel Data (website of research data center). https://kooperationen.zew.de/zew-fdz/datenangebot/mannheimer-webpanel

Zha, H., He, X., Ding, C., Simon, H., & Gu, M. (2001). Bipartite graph partitioning and data clustering. *Proceedings of the tenth international conference on Information and knowledge management*, 25–32. https://arxiv.org/abs/cs/0108018

A Appendix: Additional Figures and Tables

A.1 Feature Generation from the Mannheim Web Panel

Indicator	Description
Digitalisation	Digitalisation indicator using labeled news articles and firm websites (Axenbeck and Breithaupt 2022).
East / West	Check leading numbers of five digit numbers (= postal codes) on websites. Regular expressions: • West: ' 2[0-9][0-9][0-9][0-9] | 3[0-8][0-9][0-9][0-9] | 9[0-7][0-9][0-9][0-9] ' • East: ' [0-1][0-9][0-9][0-9][0-9] | 39[0-9][0-9][0-9] | 98[0-9][0-9][0-9] | 99[0-9][0-9][0-9] '
Industry	Regular expression checking the text on websites for industry keywords. Regular expressions: • energie|bergbau|mineralöl • nahrungsmittel|getränke|tabak • textil|bekleidung|leder • holz|papier • chemie|pharma • gummi|kunststoff • glas|keramik|stein • metall • elektro • maschinenbau • fahrzeugbau • möbel|spielwaren|medizintechnik|reparatur • wasser|entsorgung|recycling • großhandel|grosshandel • transport|post • medien • edv|telekommunikation • finanz • fue|forschung und entwicklung|f&e • unternehmensberatung|werbung • unternehmensdienste
ISO code	Regular expression checking for different ISO codes: • 9001, 14001, 27001, 50001, 13485 (see Mirtsch et al. 2021)
R&D	Regular expression checking for R&D keywords: • fue|forschung und entwicklung|f&e|r&d|research and development
Internationalism	Regular expression checking the text on websites for keywords indicating internationalism: • ausland|international
Innovation	Innovation indicator based on firm website content (see Kinne and Lenz 2021).

Table A.1: List of firm characteristics extracted from MWP data.

A.2 Anonymisation and Aggregation of the Mannheim Innovation Panel

MIP SUF variables	New variable	Description	Anonymisation method
pd	Innovator	MIP 2020. We use the original MIP data.	None.
gmdig1, ..., gmdig8	Digitalisation	MIP 2020. Recode data (none to 0, low to 1, middle to 2, high to 3). Sum up gmdig variables for each firm; divide by the maximum sum.	None.
branche	Industry	MIP 2020. The NACE codes are aggregated to 21 industries.	Aggregation of NACE codes.
fueint, fueext	R&D	MIP 2020. If at least one MIP SUF variable is larger than zero, set R&D variable to one; otherwise zero.	None.
exs	Internationalism	MIP 2020.	Truncation.
gk3n	Firm size	MIP 2020. Firm counts are grouped to three classes.	Aggregation of employee counts.
ost	East / West	MIP 2020. The firm locations are aggregated to binary East-West data.	Aggregation of location data.
iso_norm	ISO code	Extension of MIP 2020. The indicator is extracted from the MWP data by searching for the iso codes 9001, 14001, 27001, 50001, and 13485 (see Mirtsch et al. 2021).	None.

Table A.2: List of original and processed firm characteristics of the extended MIP SUF 2020 data set.

A.3 Summary Statistics of the MIP SuF, MUP and MWP

Variable name	mean	std	min	max
Digitalisation score	0.296330	0.220190	0.000000	1.000000
East Germany	0.353520	0.478110	0.000000	1.000000
Firm size: 50-249	0.264130	0.440910	0.000000	1.000000
Firm size: <50	0.634540	0.481600	0.000000	1.000000
Firm size: >= 250	0.099300	0.299100	0.000000	1.000000
ISO code	0.185200	0.370160	0.000000	1.000000
International	0.4563451	.4981425	0.0000000	1.000000
Product innovator	0.342330	0.474530	0.000000	1.000000
R&D	0.353710	0.478160	0.000000	1.000000
Industry: Consulting	0.054150	0.226330	0.000000	1.000000
Industry: Electronics	0.063140	0.243240	0.000000	1.000000
Industry: Metal	0.076360	0.265590	0.000000	1.000000
Industry: Rubber	0.032670	0.177790	0.000000	1.000000
Industry: Chemistry	0.033410	0.179710	0.000000	1.000000
Industry: Energy	0.040570	0.197300	0.000000	1.000000
Industry: Finance	0.037810	0.190760	0.000000	1.000000
Industry: Firm services	0.055070	0.228130	0.000000	1.000000
Industry: Food	0.044970	0.207260	0.000000	1.000000
Industry: Furniture	0.064240	0.245210	0.000000	1.000000
Industry: Glas	0.022760	0.149150	0.000000	1.000000
Industry: IT	0.043690	0.204410	0.000000	1.000000
Industry: Machinery	0.040200	0.196440	0.000000	1.000000
Industry: Media	0.049010	0.215910	0.000000	1.000000
Industry: R&D	0.060390	0.238230	0.000000	1.000000
Industry: Textile	0.034510	0.182550	0.000000	1.000000
Industry: Trade	0.038180	0.191650	0.000000	1.000000
Industry: Transport	0.085540	0.279700	0.000000	1.000000
Industry: Vehicle	0.024050	0.153200	0.000000	1.000000
Industry: Water	0.066260	0.248760	0.000000	1.000000
Industry: Wood	0.033040	0.178760	0.000000	1.000000

Table A.3: Summary statistics of Mannheim Innovation Panel (MIP) scientific use file. The data set consists of 5.5 thousand observations.

Variable name	mean	std	min	max
East Germany	0.159400	0.345000	0.000000	1.000000
Firm size: 50-249	0.019380	0.137860	0.000000	1.000000
Firm size: <50	0.193550	0.395080	0.000000	1.000000
Firm size: >= 250	0.003960	0.062830	0.000000	1.000000
Industry: Consulting	0.062460	0.242000	0.000000	1.000000
Industry: Electronics	0.006200	0.078490	0.000000	1.000000
Industry: Metal	0.016390	0.126960	0.000000	1.000000
Industry: Rubber	0.002540	0.050360	0.000000	1.000000
Industry: Chemistry	0.002230	0.047120	0.000000	1.000000
Industry: Energy	0.010240	0.100670	0.000000	1.000000
Industry: Finance	0.041320	0.199040	0.000000	1.000000
Industry: Firm services	0.088010	0.283300	0.000000	1.000000
Industry: Food	0.009590	0.097450	0.000000	1.000000
Industry: Furniture	0.013910	0.117100	0.000000	1.000000
Industry: Glas	0.003310	0.057410	0.000000	1.000000
Industry: IT	0.034820	0.183330	0.000000	1.000000
Industry: Machinery	0.006540	0.080620	0.000000	1.000000
Industry: Media	0.012580	0.111470	0.000000	1.000000
Industry: R&D	0.030560	0.172130	0.000000	1.000000
Industry: Textile	0.003820	0.061680	0.000000	1.000000
Industry: Trade	0.065440	0.247310	0.000000	1.000000
Industry: Transport	0.043780	0.204590	0.000000	1.000000
Industry: Vehicle	0.001900	0.043560	0.000000	1.000000
Industry: Water	0.003800	0.061540	0.000000	1.000000
Industry: Wood	0.007080	0.083850	0.000000	1.000000

Table A.4: Summary statistics of Mannheim Enterprise Panel (MUP). The data set consists of 1.2 million observations. Some variables have missing values. Industry and firm size numbers do not add up to one because of missing values.

Variable name	mean	std	min	max
Digitalisation	0.234370	0.140630	0.000840	0.926620
East Germany	0.556750	0.353260	0.000000	1.000000
ISO code	0.122660	0.328050	0.000000	1.000000
International	0.359920	0.479980	0.000000	1.000000
Product Innovation	0.269180	0.142520	0.038260	0.901070
R&D	0.063540	0.243930	0.000000	1.000000
Industry: Consulting	0.173000	0.238290	0.000000	1.000000
Industry: Electronics	0.085120	0.155820	0.000000	1.000000
Industry: Metal	0.026050	0.092260	0.000000	1.000000
Industry: Rubber	0.020260	0.068820	0.000000	1.000000
Industry: Chemistry	0.016680	0.070010	0.000000	1.000000
Industry: Energy	0.044670	0.120640	0.000000	1.000000
Industry: Finance	0.052620	0.140460	0.000000	1.000000
Industry: Firm services	0.005990	0.015810	0.000000	0.564100
Industry: Food	0.026520	0.098220	0.000000	1.000000
Industry: Furniture	0.056340	0.144200	0.000000	1.000000
Industry: Glas	0.105310	0.196440	0.000000	1.000000
Industry: IT	0.016660	0.066410	0.000000	1.000000
Industry: Machinery	0.011050	0.047240	0.000000	1.000000
Industry: Media	0.047290	0.119880	0.000000	1.000000
Industry: R&D	0.013650	0.061440	0.000000	1.000000
Industry: Textile	0.028090	0.103060	0.000000	1.000000
Industry: Trade	0.008800	0.037360	0.000000	1.000000
Industry: Transport	0.117920	0.206320	0.000000	1.000000
Industry: Vehicle	0.006820	0.023550	0.000000	1.000000
Industry: Water	0.072040	0.155590	0.000000	1.000000
Industry: Wood	0.065140	0.151750	0.000000	1.000000

Table A.5: Summary statistics of Mannheim Web Panel (MWP). The raw data set consists of 1.0 million observations. Some variables have missing values. Industry and firm size numbers do not add up to one because of missing values.

A.4 Discarding Firm Information

Discarded variable	Number of matched firms
ISO indicator	10
Employee count	7
East/West Germany	11
Product innovator	20
Digitalisation	16
Internationalism	15
R&D	10

Table A.6: Matching results for seven robustness checks that discard firm information. One variable is removed in each case/row. The industry variables can not be removed because of the search space reduction method. Setup: Algorithm 1 and without standardisation, i.e. best setup from main results.

A.5 Anonymisation Techniques in Case of the Mannheim Innovation Panel

Scientific-Use-Files are factually anonymised data sets that allow researchers to work with ZEW data outside the ZEW-FDZ, at their home institutions. Factually anonymised means that the data sets were manipulated in such an extent that re-identifying the surveyed participants would require an excessive amount of time, money and work (§3(7) BDSG). It is thus predictably impossible to re-identify firms or persons when using scientific-use-files. Different processes are used to achieve a low re-identification risk for individual firms. These processes depend on the type of variable and imply "destroying" or "corrupting" information in the dataset. These processes are:

Disturbance by a multiplicative error: In this method, the value specified by a firm is multiplied by a random number. This random number represents a firm-specific, time-invariant constant, i.e. each randomised variable is multiplied by the same number. This guarantees that the firm can no longer be recognised by the absolute figures it provides. This procedure was used for the turnover data and for data on the number of employees (in full-time equivalents). However, the quotient of these two variables (turnover per full-time employee) remains unchanged.

Reporting of intensities and ratios: All other quantitative variables are shown in relation to turnover or employees (in the case of training costs in relation to total personnel costs). These intensities or ratios are then listed in the data set. For example, total innovation expenditure, R&D expenditure, foreign turnover, investments, personnel costs and material costs are shown in relation to turnover, while the number of R&D employees and employees by qualification structure are shown in relation to the number of employees. The proportions of the various types of innovation expenditure are shown in relation to total innovation expenditure (in intervals). If required, users can calculate (randomised) absolute values or other intensities such as R&D expenditure per employee by conversion.

Truncation of intensities: In individual cases, firms may exhibit "extreme" intensities, e.g. an R&D intensity of 25%. In order to prevent firms from being re-identified by these intensities, these extreme cases, which are rarely found in the population, were truncated. Depending on the distribution of the respective intensities, different upper limits were used. For example, the upper limit for R&D intensity (R&D expenditure/turnover) is 0.15. If a firm has an R&D intensity of 0.25, the R&D intensity is truncated to 0.15. To allow the user to recognise that

the respective variable has been truncated, an additional variable was included in the data set to indicate the truncation.

Grouping: For some variables, only the interval in which the characteristic values of the firms lie is specified (by means of an ordinal variable).

Aggregation of information: Information is summarised or coarsened in a new characteristic value.

Non-disclosure of data: Some sensitive information is not included in the scientific-use-files, e.g. information on tax incentives for R&D.

Download ZEW Discussion Papers:

https://www.zew.de/en/publications/zew-discussion-papers

or see:

https://www.ssrn.com/link/ZEW-Ctr-Euro-Econ-Research.html
https://ideas.repec.org/s/zbw/zewdip.html

IMPRINT

ZEW – Leibniz-Zentrum für Europäische Wirtschaftsforschung GmbH Mannheim
ZEW – Leibniz Centre for European Economic Research

L 7,1 · 68161 Mannheim · Germany
Phone +49 621 1235-01
info@zew.de · zew.de

Discussion Papers are intended to make results of ZEW research promptly available to other economists in order to encourage discussion and suggestions for revisions. The authors are solely responsible for the contents which do not necessarily represent the opinion of the ZEW.