



Wetenschappelijk Onderzoek- en
Documentatiecentrum

A review of Frequency Table Disclosure Control from a Microdata Perspective

Authors: Alexander Latenko, Mortaza S. Bargh, Susan van den Braak, Marco Vink

*Presented by: Alexander Latenko of the Research and Documentation Centre,
Ministry of Justice and Security, The Netherlands*



Contact: [a\(dot\)latenko\(at\)wodc\(dot\)nl](mailto:a(dot)latenko(at)wodc(dot)nl)



About me

- Background in computer science
- Algorithmic fairness
- Privacy
- Security





The Research and Documentation Centre

- Ministry of Justice and Security
- Social impact
- Applied Research
- One of multiple research departments
- Topics:
 - Big data
 - Machine learning
 - Data management
 - Privacy
 - Other topics



Research and Documentation Centre

[Home](#) [Research](#) [Figures and forecasts](#) [Publications](#) [Organization](#) [Themes](#) [Keyword Index](#)

Search

WODC (Research and Documentation Centre)



The WODC (Research and Documentation Centre) of the Ministry of Justice and Security can be characterised as an international knowledge centre on: security, police, criminal, civil and administrative justice and migration issues. "Excellence" and "customer-orientation" are the organisation's guiding principles. Its major output is knowledge for the benefit of policy development.

WODC News

Special research program

→ [Recidivism Monitor](#)



Content

1. Open data motivations
2. Statistical Disclosure Control (SDC)
3. Microdata
4. Tabular data
5. Practical differences
6. Sources of disclosure
7. Privacy models
8. Conclusions
9. Future work



Opening data

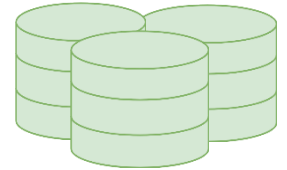
- Government setting
- Research
- Verification
- Transparency
- Public opinion





Routes for releasing data

- Generic or specific
- more eyes or more data
- Contracts

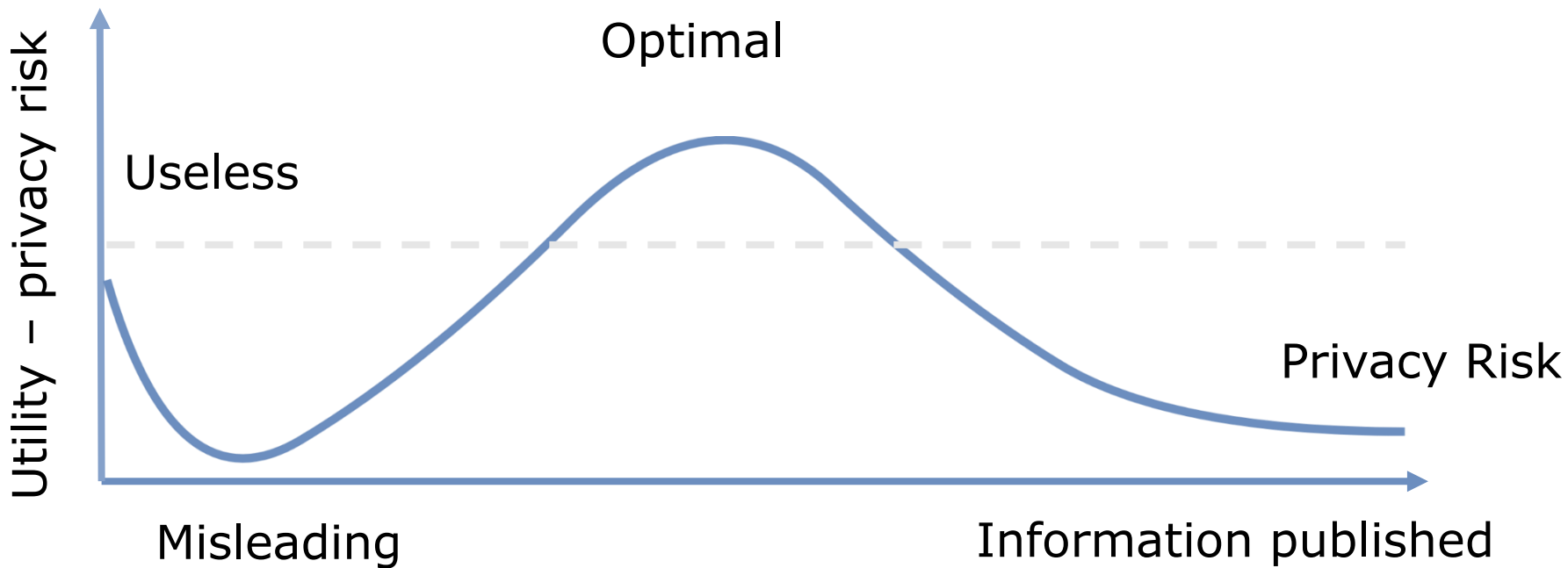


TNO

TU Delft
The University of
Technology



Why SDC?





Why SDC?

SDC is a step between the government data and the public

Research remains important due to advances in:

- Accessibility of (big) data
- Legislation (GDPR)





Microdata

One entity per row

High dimensionality

More difficult to protect

- Netflix challenge
- Taxi data

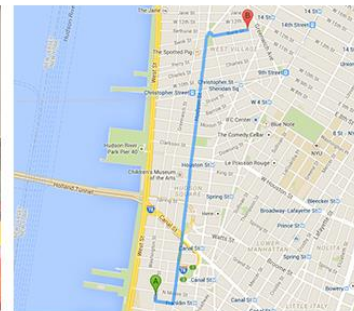
Name	Job	Sex	Age	Disease
Bob	Engineer	Male	35	Hepatitis
Fred	Engineer	Male	38	Hepatitis
Doug	Lawyer	Male	38	HIV
Alice	Writer	Female	30	Flu
Cathy	Writer	Female	30	HIV
Emily	Dance	Female	30	HIV

EID



BRADLEY COOPER

QIDs



JULY 8, 2013 • 7:34 PM - 7:44 PM
376 GREENWICH ST. TO 13 BANK ST.
\$9.00 FARE • CASH: UNKNOWN TIP • @SPLASH

SAT



Tabular data

- Derived from microdata
- Describe groups not individuals
- Sufficient for most transparency and verification cases.
- Still suffers from privacy risks

	Number of holdings (1 000)	Utilised agricultural area — UAA (¹) (1 000 hectares)	Livestock units — LSU (1 000 LSU)	Labour force (²) (1 000 annual work units)	Standard output (EUR million)	Average area of holdings (hectares)
EU-28	12 248.0	174 115.6	135 212.3	9 945.8	307 889.5	14.2
BE	42.9	1 358.0	3 798.7	81.6	7 247.8	31.7
BG	370.5	4 475.5	1 148.5	406.5	2 536.7	12.1
CZ	22.9	3 483.5	1 722.5	108.0	3 852.2	152.4
DK	42.1	2 646.9	4 919.4	52.3	8 430.8	62.9
DE	299.1	16 704.0	17 792.6	545.5	41 494.1	55.8
EE	19.6	940.9	306.3	25.1	594.6	48.0
IE	139.9	4 991.4	5 787.4	165.4	4 297.7	35.7
EL	723.0	3 477.9	2 406.5	429.5	6 700.0	4.8
ES	989.8	23 752.7	14 830.9	889.0	34 173.1	24.0
FR	516.1	27 837.3	22 674.2	779.7	50 733.2	53.9
HR	233.3	1 316.0	1 020.2	184.5	2 114.7	5.6
IT	1 620.9	12 856.1	9 911.5	953.8	49 460.3	7.9
CY	38.9	118.4	200.8	18.6	458.9	3.0
LV	83.4	1 796.3	474.6	85.2	777.2	21.5
LT	199.9	2 742.6	900.1	146.8	1 526.3	13.7
LU	2.2	131.1	167.7	3.7	268.6	59.6
HU	576.8	4 686.3	2 483.8	423.5	5 241.0	8.1
MT	12.5	11.5	41.7	4.9	95.9	0.9
NL	72.3	1 872.4	6 711.5	161.7	18 930.0	25.9
AT	150.2	2 878.2	2 517.2	114.3	5 879.3	19.2
PL	1 506.6	14 447.3	10 377.2	1 897.2	18 987.1	9.6
PT	305.3	3 668.2	2 205.0	363.4	4 639.7	12.0
RO	3 859.0	13 306.1	5 444.2	1 610.3	10 420.3	3.4
SI	74.7	482.7	518.5	76.7	913.2	6.5
SK	24.5	1 895.5	668.3	56.1	1 731.0	77.5
FI	63.9	2 291.0	1 121.1	59.7	3 097.6	35.9
SE	71.1	3 066.3	1 751.9	56.9	3 733.3	43.1
UK	186.8	16 881.7	13 308.4	266.3	19 555.0	90.4
IS	2.6	1 595.7	161.0	4.2	237.1	616.1
NO	46.6	1 005.9	1 229.3	46.4	3 156.2	21.6
CH	59.1	1 047.8	1 793.8	96.0	5 717.1	17.7
ME	48.9	221.3	118.4	47.9	127.1	4.5

(¹) Excluding common land in Greece.

(²) Labour force directly employed on the farm.

Source: Eurostat (online data code: ef_kvaareg)



Practical differences

The information is similar but the difference in usage is vast.

The are difference in:

- Target audience
- Purpose
- Dimensionality
- Relations

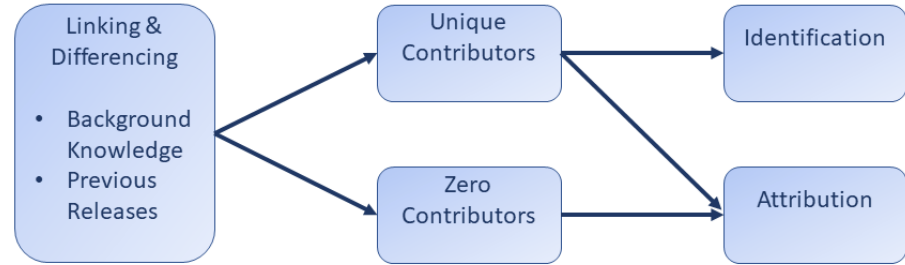
These differences have split the existing research on protection for these data forms.



Sources of disclosure

Intruders seek to identify an individual or small group or learn something new about them (attribution).

- Background knowledge
- Tabular Disclosure scenario's
- Microdata linkage attacks





Tabular disclosure example

	Detention days<30		Detention days 30-60	
	Gender F	Gender M	Gender F	Gender M
Age=18	2	2	1	0
Age=19	10	10	2	2

	Crime=Assault	Crime=Petty theft	Crime=Murder
	Detention days=10	0	7
Detention days=10-30	0	7	5
Detention days>30	5	0	2

Knowledge of the age and detention days

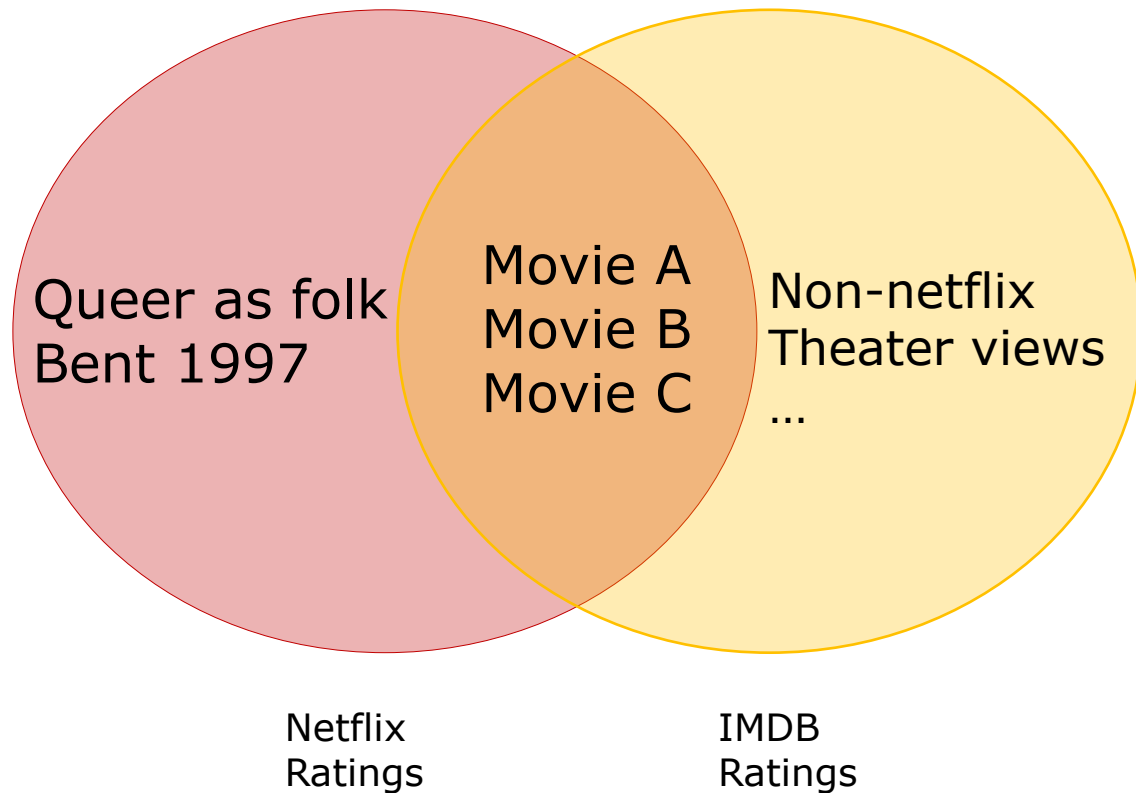
Deriving in a chain

	Gender=F	Gender=M
Crime=Assault	3	2
Crime=Petty theft	10	10
Crime=Murder	0	2



Microdata disclosure example

- Record linkage
- Mapping Quasi identifiers
- Linkage to EIDs on IMDB
- Requires uniqueness for identification





Privacy models

- K-anonymity
 - Minimum group size for each QID is at least k
- Frequency rule
 - Minimum cell size is at least n
- Stricter rules for tables
- Margin inclusions

Name	Job	Sex	Age	Disease
Bob	Engineer	Male	35	Hepatitis
Fred	Engineer	Male	38	Hepatitis
Doug	Lawyer	Male	38	HIV
Alice	Writer	Female	30	Flu
Cathy	Writer	Female	30	HIV
Emily	Dance	Female	30	HIV

EID

QIDs

SAT

	Total Crimes	Section 1	Section 2	Section 3	Section 4
Original	10	2	2	2	2
Rounded	10	0	0	0	0

Range	8-12	0-2	0-2	0-2	0-2
--------------	------	-----	-----	-----	-----

Audited Range	8	2	2	2	2
----------------------	---	---	---	---	---



Privacy models

- L-diversity
 - Every QID group has at minimum l SAT values.
- Entropy L-diversity
- T-closeness

- Zero cells prevention
- Prevention skewed distribution
- No tabular mapping (again)

Job	Sex	Age	Disease
Engineer	Male	35	Hepatitis
Engineer	Male	38	Hepatitis
Lawyer	Male	38	HIV
Writer	Female	30	Flu
Writer	Female	30	HIV
Dance	Female	30	HIV

QIDs

SAT

	Hepatitis	Hiv	Flu
Male	2	1*	0
Female	0	2	1

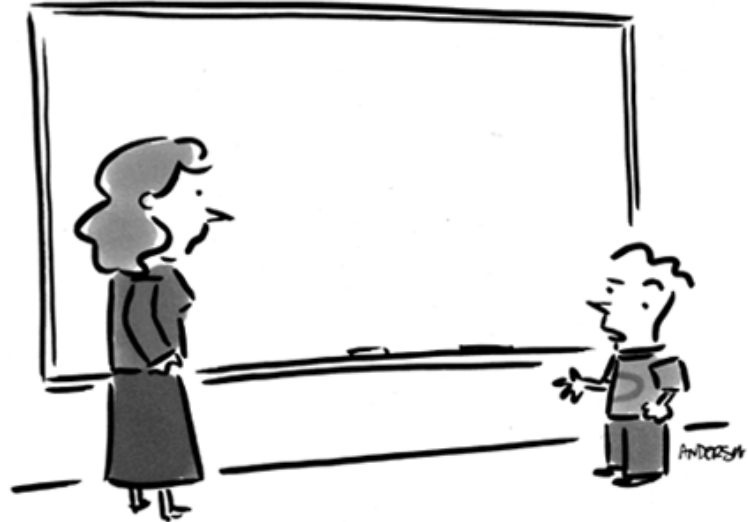


Conclusions

- Differential privacy
- Simplification of microdata models
- Differences in likely sources of disclosure
- More evaluation of microdata models for tabular data is needed

© MARK ANDERSON

WWW.ANDERSTOONS.COM



"Before I write my name on the board, I'll need to know how you're planning to use that data."



Future work

- Empirical comparisons
- Protection methods comparison
- Of interest are how SDC techniques for structured data types relate to those for unstructured data, such as textual data.

F.B.I. Agent **Peter Strzok PERSON** . **Who Criticized Trump PERSON** in Texts, Is **Fired GPE** - **The New York Times ORG** SectionsSEARCHSkip to contentSkip to site indexPoliticsSubscribeLog InSubscribeLog InToday's **PaperAdvertisementSupported ORG** byF.B.I. Agent **Peter Strzok PERSON** , **Who Criticized Trump PERSON** in Texts, Is FiredImagePeter Strzok, a top **F.B.I. GPE** counterintelligence agent who was taken off the special counsel investigation after his disparaging texts about President **Trump PERSON** were uncovered, was fired. **CreditT.J. Kirkpatrick PERSON** for **The New York TimesBy Adam Goldman ORG** and **Michael S. SchmidtAug PERSON** . **13 CARDINAL** , **2018WASHINGTON CARDINAL** — **Peter Strzok PERSON** , the **F.B.I. GPE** senior counterintelligence agent who disparaged President **Trump PERSON** in inflammatory text messages and helped oversee the **Hillary Clinton PERSON** email and **Russia OPE** investigations, has been fired for violating bureau policies. Mr. **Strzok PERSON** 's lawyer said **Monday DATE** Mr. Trump and his allies seized on the texts — exchanged during the **2016 DATE** campaign with a former **F.B.I. GPE** lawyer, **Lisa Page — in PERSON** assailing the **Russia GPE** investigation as an illegitimate “witch hunt.” Mr. **Strzok PERSON** , who rose over **20 years DATE** at the **F.B.I. GPE** to become one of its most experienced counterintelligence agents, was a key figure in **the early months DATE** of the inquiry Along with writing the texts, Mr. **Strzok PERSON** was accused of sending a highly sensitive search warrant to his personal email account The **F.B.I. GPE** had been under immense political pressure by Mr. **Trump PERSON** to dismiss Mr. **Strzok PERSON** , who was removed **last summer DATE** from the staff of the special counsel, **Robert S. Mueller III PERSON** . The president has repeatedly denounced Mr. **Strzok PERSON** in posts on **Twitter EVENT** , and on **Monday DATE** expressed satisfaction that he had been sacked.Mr. **Trump's ORG** victory traces back to **June DATE** , when Mr. **Strzok PERSON** 's conduct was laid out in a wide-ranging inspector general's report on how the **F.B.I. GPE** handled the investigation of **Hillary Clinton's PERSON** emails in the run-up to the **2016 DATE** election. The report was critical of Mr. **Strzok PERSON** 's conduct in sending the