

**Hochschule für
Technik und Wirtschaft
des Saarlandes**

University of
Applied Sciences


**Fakultät für
Ingenieur-
wissenschaften**

School of
Engineering



Implementierung eines Prototypen für ein an SAFE angelehntes Anonymisierungsverfahren mit Hilfe von Clusteringalgorithmen

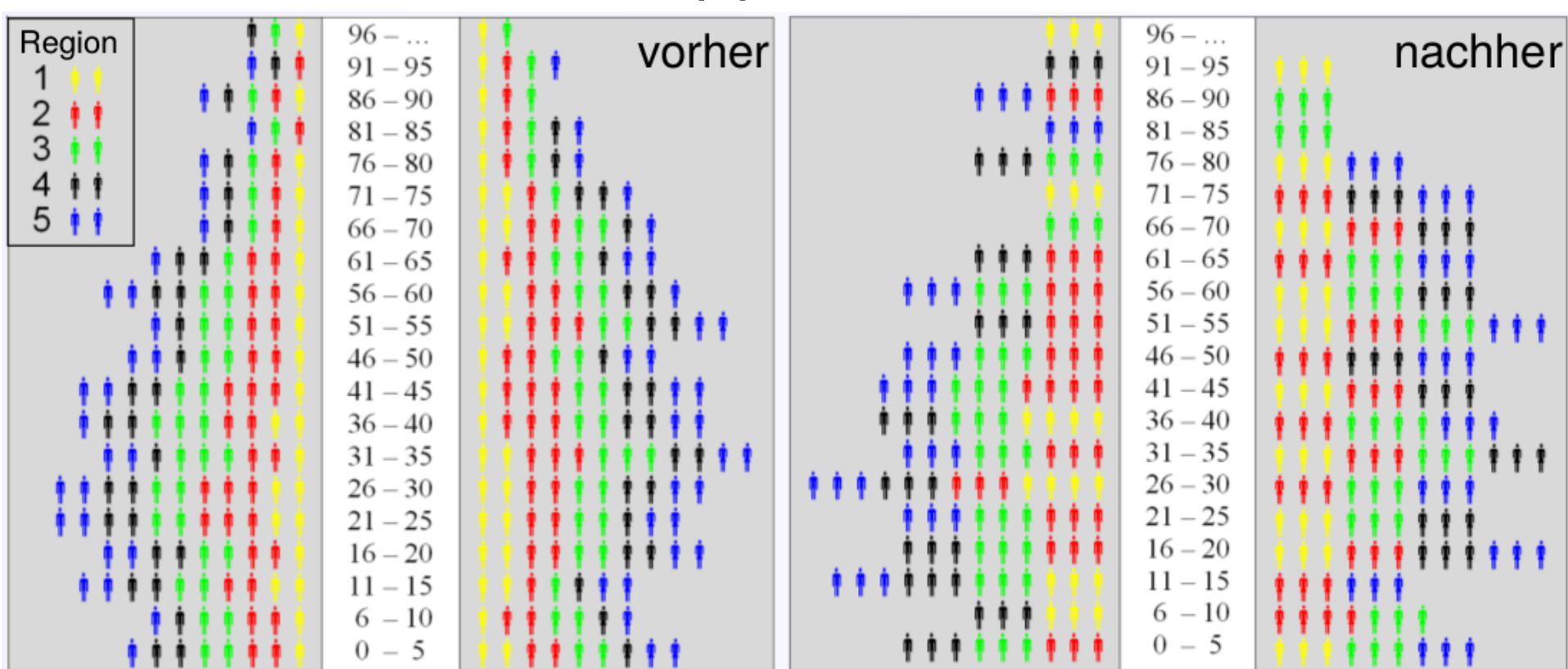
Boris Wiegand
SS 2015 – 26.06.2015
Seminar Anonymisierung von Mikrodaten
Prof. Dr. rer. nat. habil. Rainer Lenz
Dipl.-Vw. Emanuel Weiß



Inhalt

- SAFE
 - Idee
 - Mathematisches Modell
- Entwickeltes Verfahren
 - Konzept
 - Verwendete Clusteringalgorithmen
 - Bewertung
 - Ausblick
- Livedemo nach der Präsentation

SAFE-Grundidee (1)



Höhne, Jörg. "Methoden und Potenziale des Zensus 2011." (2012). Vortrag auf den Statistik-Tagen Bamberg 2012. Folien verfügbar unter https://www.statistik.bayern.de/medien/wichtigethemen/st_vortrag_hoehne_27072012.pdf (letzter Aufruf am 17.06.15)

SAFE-Grundidee (2)

Merkmal 1	Merkmal 2		Merkmal 1	Merkmal 2
A	10		A	11
B	12		A	11
A	12		A	11
A	15		A	11
B	12	SAFE →	A	11
A	11		A	11
C	10		B	12
A	25		B	12
A	2		B	12

SAFE – Mathematisches Modell (1)

- Matrix metrischer Werte

Alter	Gehalt	Ort	Geschl
51	1500	A-Stadt	m
21	550	B-Dorf	w
32	3500	C-Weiler	m

$$X = \begin{pmatrix} 51 & 1500 \\ 21 & 550 \\ 32 & 3500 \end{pmatrix}$$

SAFE – Mathematisches Modell (2)

- Zuordnungsmatrizen

Alter	Gehalt	Ort	Geschl
51	1500	A-Stadt	m
21	550	B-Dorf	w
32	3500	C-Weiler	m

$$Z_1 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

$$Z_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \end{pmatrix}$$

SAFE – Mathematisches Modell (3)

- Kombination von Zuordnungsmatrizen

$$Z_1 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \quad Z_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \end{pmatrix}$$

$$Z_{1,2} = \begin{pmatrix} Z_{1_{\text{Zeile 1}}} \otimes Z_{2_{\text{Zeile 1}}} \\ Z_{1_{\text{Zeile 2}}} \otimes Z_{2_{\text{Zeile 2}}} \\ Z_{1_{\text{Zeile 3}}} \otimes Z_{2_{\text{Zeile 3}}} \end{pmatrix} = \begin{pmatrix} 1 \cdot (1 \ 0) & 0 \cdot (1 \ 0) & 0 \cdot (1 \ 0) \\ 0 \cdot (1 \ 0) & 1 \cdot (1 \ 0) & 0 \cdot (0 \ 1) \\ 0 \cdot (1 \ 0) & 0 \cdot (1 \ 0) & 1 \cdot (1 \ 0) \end{pmatrix}$$

$$= \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix}$$

SAFE – Mathematisches Modell (4)

- Auswertung kategoriale Merkmale

$$A_j := Z_j^T \cdot Z_j$$

$$A_2 = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix} \cdot \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix}$$

SAFE – Mathematisches Modell (5)

- Auswertung kategoriale und metrische Merkmale

$$T_j := Z_j^T \cdot X$$

$$T_2 = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix} \cdot \begin{pmatrix} 51 & 1500 \\ 21 & 550 \\ 32 & 3500 \end{pmatrix} = \begin{pmatrix} 83 & 5000 \\ 21 & 550 \end{pmatrix}$$

SAFE – Mathematisches Modell (6)

- Minimierungsproblem:

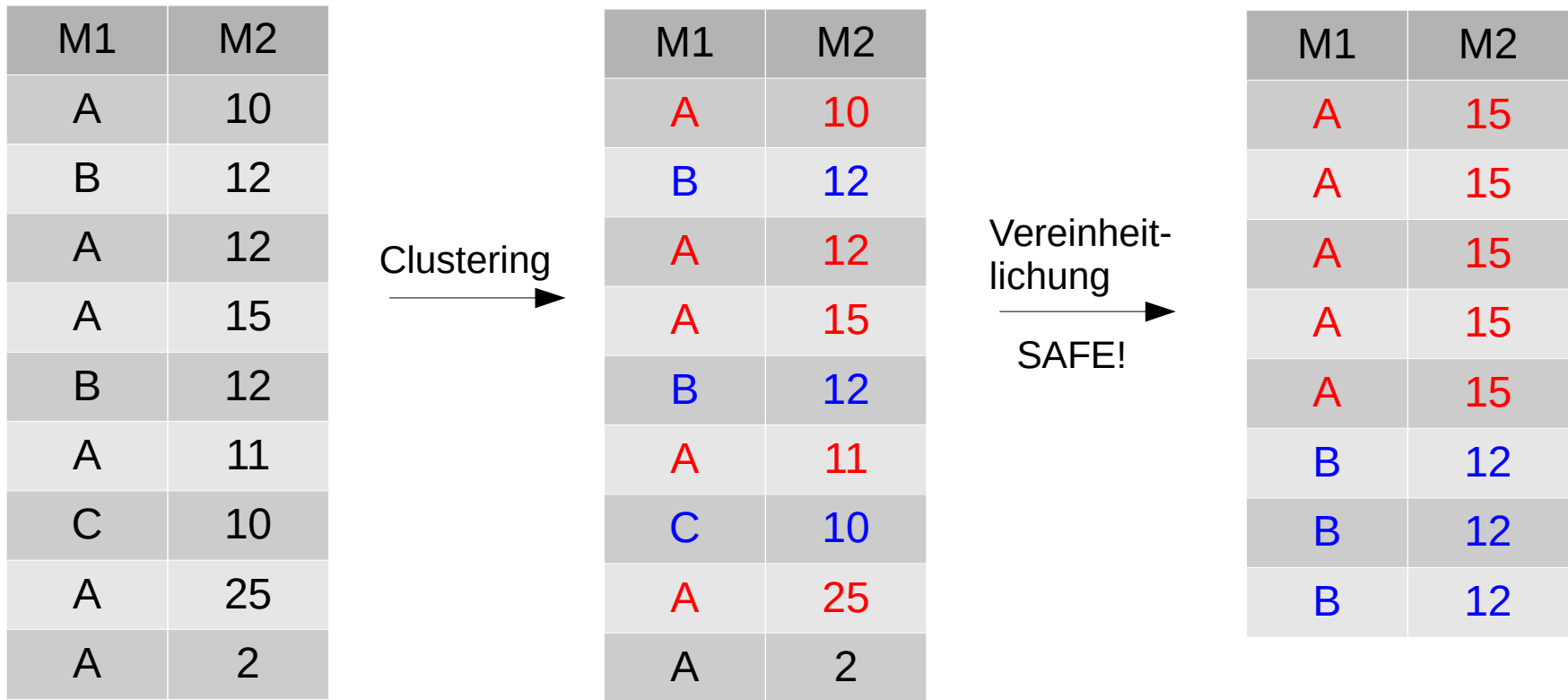
Gegeben: X^0, Z_1^0 bis Z_k^0

Gesucht: X^a, Z_1^a bis Z_k^a , sodass

Differenz aller A_j^a und A_j^0 minimal und

Differenz aller T_j^a und T_j^0 minimal

Entwickeltes Verfahren - Konzept



k-means++

- Verbesserter k-means Algorithmus (bessere Auswahl der Startpunkte)
- Idee von k-means:
 - Minimiere φ mit
 - X Menge aller Punkte
 - C Menge aller Clusterzentren, $|C| = k$
 - NP \Rightarrow Näherungslösung

$$\varphi = \sum_{x \in X} \min_{c \in C} \|x - c\|^2$$

Arthur, David, and Sergei Vassilvitskii. "k-means++: The advantages of careful seeding." Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms. Society for Industrial and Applied Mathematics, 2007.

DBSCAN

- Idee: Punkte in Cluster liegen dicht beieinander
- A ist von B **erreichbar**, wenn $d(A,B) < \epsilon$
- A ist **Kernpunkt**, wenn A *minPts* erreichbare Nachbarn besitzt
- A ist **Dichte-erreichbar**, wenn A von einem Kernpunkt erreichbar ist, selbst aber kein Kernpunkt ist
- Ansonsten ist A **Rauschpunkt (Noise)**
- Kernpunkte und Dichte-erreichbare Punkte bilden einen **Cluster**

Ester, Martin, et al. "A density-based algorithm for discovering clusters in large spatial databases with noise." Kdd. Vol. 96. No. 34. 1996.

Bewertung (1)

- umfassendere Analyse notwendig
- bisher eine Auswertung für k-means++ mit folgenden Parametern:
 - $k = 50$
 - minimale Gruppengröße 3
 - metrische Merkmale werden mit arithmetischem Mittelwert vereinheitlicht
 - kategoriale Merkmale werden durch Ersetzung mit häufigstem Merkmal innerhalb einer Gruppe vereinheitlicht

Bewertung (2)

- Kategoriale Merkmale

	West	Ost
Original	19118	4720
Anonymisiert	18655	4256
Abweichung	463	464

	Erwerbs- tätig	Erwerbs- -los	Arbeit- suchend	Sonstige (z.B.<15)
Original	11101	806	130	11337
Anonymisiert	11806	586	52	10930
Abweichung	705	220	78	407

Bewertung (3)

- Metrische Merkmale

Alter	Mittelwert	Varianz	Stand.-Abw.	Min	Max	Summe
Original	43,95	523,82	22,89	0	95	1027218
Anonymisiert	43,45	359,64	18,96	14	80	1015685
Abweichung	0,49	164,18	3,92	14	15	11533

Anzahl Personen im Haushalt	Mittelwert	Varianz	Stand.-Abw.	Min	Max	Summe
Original	2,78	5,02	2,24	1	40	64982
Anonymisiert	2,35	0,85	0,92	1	40	54989
Abweichung	0,43	4,17	1,32	0	0	9993

Ausblick

- Weitere Clusterverfahren (z.B. Single-Linkage)
- Behandlung von Missings
- Speichereffizienz (Zuordnungsmatrizen!)
- Beste Parameter für Clusterverfahren
- Laufzeit (Analyse, evtl. Parallelisierung)
- Weitere Auswertungen



<http://winfwiki.wi-fom.de/images/f/fd/Ausblick.jpg> - letzter Aufruf am 19.06.2015

