

Overcoming the limitations of k -anonymity through association rule hiding

SAP RESEARCH

Enriching the Quality of Data for Business Analytics

Marek Zielinski

University of Pretoria / SAP Research CEC Pretoria
South Africa



Overcoming the limitations of k -anonymity through association rule hiding

SAP RESEARCH

Enriching the Quality of Data for Business Analytics

- Types of statistical data
- k -Anonymity and its limitations
- Association rule hiding
- Proposed anonymisation approach
- Conclusion



Types of statistical data

SAP RESEARCH

Summarised tables

Disease	Gender		TOTAL	Disease	Age				TOTAL
	Male	Female			< 20	21-40	41-60	> 60	
Cancer	50	60	110	Cancer	20	40	30	20	110
Hypertension	10	5	15	Hypertension	10	0	5	0	15
Heart disease	15	20	35	Heart disease	3	7	15	10	35
TOTAL	75	85	160	TOTAL	33	47	50	30	160

Microdata

Date of Birth	Race	Gender	Zipcode	Disease
1967/01/01	Black	Male	40121	Cancer
1967/02/02	Black	Male	40121	Hypertension
1967/03/03	Black	Male	40121	Cancer
1968/04/04	White	Male	40242	Heart disease
...



k - Anonymity

SAP RESEARCH

k -Anonymity: every record in the anonymised microdata set must be indistinguishable from at least $(k - 1)$ other records within the same data set ($k \geq 1$)

	Date of Birth	Race	Gender	Zipcode	Disease
1	1967/01/01	Black	Male	40121	Cancer
2	1967/02/02	Black	Male	40121	Hypertension
3	1967/03/03	Black	Male	40121	Cancer
4	1968/04/04	White	Male	40242	Heart disease
5	1968/05/05	White	Male	40242	Heart disease
6	1968/06/06	White	Male	40242	Heart disease
7	1969/07/07	Black	Female	40373	Cancer
8	1969/08/08	Black	Female	40373	Hypertension
9	1969/09/09	Black	Female	40373	Hypertension
10	1970/10/10	White	Female	40404	Heart disease
11	1970/11/11	White	Female	40404	Cancer
12	1970/12/12	White	Female	40404	Hypertension



k - Anonymity

SAP RESEARCH

k -Anonymity: every record in the anonymised microdata set must be indistinguishable from at least $(k - 1)$ other records within the same data set ($k \geq 1$)

	Date of Birth	Race	Gender	Zipcode	Disease
1	1967	Black	Male	40121	Cancer
2	1967	Black	Male	40121	Hypertension
3	1967	Black	Male	40121	Cancer
4	1968	White	Male	40242	Heart disease
5	1968	White	Male	40242	Heart disease
6	1968	White	Male	40242	Heart disease
7	1969	Black	Female	40373	Cancer
8	1969	Black	Female	40373	Hypertension
9	1969	Black	Female	40373	Hypertension
10	1970	White	Female	40404	Heart disease
11	1970	White	Female	40404	Cancer
12	1970	White	Female	40404	Hypertension



Association Rule Hiding

SAP RESEARCH

A rule of the form $X \Rightarrow Y$

{Date of Birth = 1968 &
Race = White &
Gender = Male &
Zipcode = 40242} \Rightarrow {Disease = Heart disease}

$$Supp(X \Rightarrow Y) = \frac{\text{Number of Tuples containing both } X \text{ and } Y}{\text{Total Number of Tuples}} \times 100$$


$$Conf(X \Rightarrow Y) = \frac{\text{Number of Tuples containing both } X \text{ and } Y}{\text{Total Number of Tuples containing } X} \times 100$$



Proposed anonymisation approach

SAP RESEARCH

1. A de-identified microdata table is k -anonymised
2. Sensitive association rules with a confidence value above a minimum threshold value c are found
3. The sensitive association rules found in step 2 are "hidden" by decreasing the support of the rule's consequent




THE BEST-RUN BUSINESSES RUN SAP

Proposed anonymisation approach – Step 1

SAP RESEARCH

A de-identified microdata table is k -anonymised (let $k = 3$)

	Date of Birth	Race	Gender	Zipcode	Disease
1	1967/01/01	Black	Male	40121	Cancer
2	1967/02/02	Black	Male	40121	Hypertension
3	1967/03/03	Black	Male	40121	Cancer
4	1968/04/04	White	Male	40242	Heart disease
5	1968/05/05	White	Male	40242	Heart disease
6	1968/06/06	White	Male	40242	Heart disease
7	1969/07/07	Black	Female	40373	Cancer
8	1969/08/08	Black	Female	40373	Hypertension
9	1969/09/09	Black	Female	40373	Hypertension
10	1970/10/10	White	Female	40404	Heart disease
11	1970/11/11	White	Female	40404	Cancer
12	1970/12/12	White	Female	40404	Hypertension




THE BEST-RUN BUSINESSES RUN SAP

Proposed anonymisation approach – Step 1

SAP RESEARCH

A de-identified microdata table is k -anonymised (let $k = 3$)

	Date of Birth	Race	Gender	Zipcode	Disease
1	1967	Black	Male	40121	Cancer
2	1967	Black	Male	40121	Hypertension
3	1967	Black	Male	40121	Cancer
4	1968	White	Male	40242	Heart disease
5	1968	White	Male	40242	Heart disease
6	1968	White	Male	40242	Heart disease
7	1969	Black	Female	40373	Cancer
8	1969	Black	Female	40373	Hypertension
9	1969	Black	Female	40373	Hypertension
10	1970	White	Female	40404	Heart disease
11	1970	White	Female	40404	Cancer
12	1970	White	Female	40404	Hypertension



THE BEST-RUN BUSINESSES RUN SAP

Proposed anonymisation approach – Step 2


SAP RESEARCH

Sensitive association rules with a confidence value above a minimum threshold value c are found (let $c = 70\%$)

Sensitive rules with confidence above 70%:

{Date of Birth = 1968 &
Race = White &
Gender = Male &
Zipcode = 40242} \Rightarrow {Disease = Heart disease}

(Confidence of 100%)




THE BEST-RUN BUSINESSES RUN SAP

Proposed anonymisation approach – Step 3a

SAP RESEARCH

The sensitive association rules found in step 2 are "hidden" by decreasing the support of the rule's consequent

	Date of Birth	Race	Gender	Zipcode	Disease
1	1967	Black	Male	40121	Cancer
2	1967	Black	Male	40121	Hypertension
3	1967	Black	Male	40121	Cancer
4	1968	White	Male	40242	Heart disease
5	1968	White	Male	40242	Heart disease
6	1968	White	Male	40242	Heart disease
7	1969	Black	Female	40373	Cancer
8	1969	Black	Female	40373	Hypertension
9	1969	Black	Female	40373	Hypertension
10	1970	White	Female	40404	Heart disease
11	1970	White	Female	40404	Cancer
12	1970	White	Female	40404	Hypertension




THE BEST-RUN BUSINESSES RUN SAP

Proposed anonymisation approach – Step 3a

SAP RESEARCH

The sensitive association rules found in step 2 are "hidden" by decreasing the support of the rule's consequent

	Date of Birth	Race	Gender	Zipcode	Disease
1	1967	Black	Male	40121	Cancer
2	1967	Black	Male	40121	Hypertension
3	1967	Black	Male	40121	Cancer
4	1968	White	Male	40242	*
5	1968	White	Male	40242	Heart disease
6	1968	White	Male	40242	Heart disease
7	1969	Black	Female	40373	Cancer
8	1969	Black	Female	40373	Hypertension
9	1969	Black	Female	40373	Hypertension
10	1970	White	Female	40404	Heart disease
11	1970	White	Female	40404	Cancer
12	1970	White	Female	40404	Hypertension



THE BEST-RUN BUSINESSES RUN SAP

Proposed anonymisation approach – Step 3b

SAP RESEARCH

Determine the actual number of tuples that should contribute to the confidence of a rule



Sensitive rules with confidence above 70%:

{Date of Birth = 1968 &
Race = White &
Gender = Male &
Zipcode = 40242} ⇒ {Disease = Heart disease}

(Confidence of 100%)

Only 3 tuples contributed to the high (100%) confidence value

That's why the confidence value has been reduced so significantly by suppressing the sensitive cell of only one tuple



THE BEST-RUN BUSINESSES RUN SAP

Proposed anonymisation approach – Step 3b

SAP RESEARCH

Determine the actual number of tuples that should contribute to the confidence of a rule (e.g. at least 2 tuples)

	Date of Birth	Race	Gender	Zipcode	Disease
1	1967	Black	Male	40121	Cancer
2	1967	Black	Male	40121	Hypertension
3	1967	Black	Male	40121	Cancer
4	1968	White	Male	40242	*
5	1968	White	Male	40242	Heart disease
6	1968	White	Male	40242	Heart disease
7	1969	Black	Female	40373	Cancer
8	1969	Black	Female	40373	Hypertension
9	1969	Black	Female	40373	Hypertension
10	1970	White	Female	40404	Heart disease
11	1970	White	Female	40404	Cancer
12	1970	White	Female	40404	Hypertension



THE BEST-RUN BUSINESSES RUN SAP

Proposed anonymisation approach – Step 3b

SAP RESEARCH

Determine the actual number of tuples that should contribute to the confidence of a rule (e.g. at least 2 tuples)

	Date of Birth	Race	Gender	Zipcode	Disease
1	1967	Black	Male	40121	Cancer
2	1967	Black	Male	40121	Hypertension
3	1967	Black	Male	40121	Cancer
4	1968	White	Male	40242	*
5	1968	White	Male	40242	*
6	1968	White	Male	40242	Heart disease
7	1969	Black	Female	40373	Cancer
8	1969	Black	Female	40373	Hypertension
9	1969	Black	Female	40373	Hypertension
10	1970	White	Female	40404	Heart disease
11	1970	White	Female	40404	Cancer
12	1970	White	Female	40404	Hypertension






THE BEST-RUN BUSINESSES RUN SAP

Conclusion

SAP RESEARCH

- **Related work:** *l*-diversity
- **Unanswered questions:**
 - How to determine what is the "best" value for the confidence of an association rule that should be used?
 - How to determine what is the "best" value for the actual number of records that should support both the antecedent and the consequent of a rule?
 - How to balance the needs of privacy and information utility?



THE BEST-RUN BUSINESSES RUN SAP

Overcoming the limitations of *k*-anonymity through association rule hiding

SAP RESEARCH

Questions & Comments

marek.zielinski@sap.com

THE BEST-RUN BUSINESSES RUN SAP