ANON

Ein Tool zur Anonymisierung medizinischer Daten

Johann Eder, Marga Ciglic, Christian Koncilia
Institut für Informatik-Systeme
Alpen Adria Universität

Problem

- Anonymität der Individuen kann mit bloßem Entfernen der Identifikatoren (z.B. Name, SVNR, etc.) nicht erreicht werden.
  - **Tracker**
    - Charakteristische Formel, die mögliche Werte mit UND, ODER und NICHT verknüpft
    - Findet neue Informationen über ein Individuum
  - **Linking Attack**
    - Daten mit entfernten Identifikatoren werden mit externen Datenquellen verbunden

Tracker

PLZ	Alter	Geschl.	Diagnose
13053	28	M	Hepatitis
13068	29	F	Hepatitis
13068	21	M	Pneumonie
13053	23	F	Pneumonie
14853	50	M	Krebs
14853	55	M	Hepatitis
14850	47	M	Pneumonie
14850	49	M	Pneumonie
13053	31	F	Krebs
13053	37	M	Krebs
13068	36	M	Krebs
13068	35	M	Krebs

Anton möchte wissen, unter welcher Krankheit seine 29-jährige Nachbarin Berta leidet.

1) **COUNT(29 ∧ F) = 1**

   → Eine Abfrage, die 29-jährige Frauen zählt, liefert 1 zurück

   → Anton hat somit Berta identifiziert!

   Der Tracker für Berta lautet (29 ∧ F)

2) **COUNT(29 ∧ F ∧ Krebs) = 0**

   → negative Kompromittierung

   (Berta hat keinen Krebs)

3) **COUNT(29 ∧ F ∧ Hepatitis) = 1**

   → positive Kompromittierung

   (Berta hat Hepatitis)

Linking Attack

	Name	PLZ	Alter	Geschl.
1	Anton	13053	28	M
2	Berta	13068	29	F
3	Cäsar	13068	21	M
4	Dora	13053	23	F
5	Emil	14853	50	M
6	Friedrich	14853	55	M
7	Gustav	14850	47	M
8	Heinrich	14850	49	M
9	Ida	13053	31	F
10	Julius	13053	37	M
11	Konrad	13068	36	M
12	Ludwig	13068	35	M

LINK

PLZ	Alter	Geschl.	Diagnose
13053	28	M	Hepatitis
13068	29	F	Hepatitis
13068	21	M	Pneumonie
13053	23	F	Pneumonie
14853	50	M	Krebs
14853	55	M	Hepatitis
14850	47	M	Pneumonie
14850	49	M	Pneumonie
13053	31	F	Krebs
13053	37	M	Krebs
13068	36	M	Krebs
13068	35	M	Krebs

Wählerverzeichnis **Medizinische Daten**

k-Anonymität

	PLZ	Alter	Geschl.	Diagnose
1	130**	< 30	*	Hepatitis
2	130**	< 30	*	Hepatitis
3	130**	< 30	*	Pneumonie
4	130**	< 30	*	Pneumonie
5	148**	≥ 40	*	Krebs
6	148**	≥ 40	*	Hepatitis
7	148**	≥ 40	*	Pneumonie
8	148**	≥ 40	*	Pneumonie
9	130**	3*	*	Krebs
10	130**	3*	*	Krebs
11	130**	3*	*	Krebs
12	130**	3*	*	Krebs

4-anonyme Tabelle mit 3 Gruppen

Eine Tabelle erfüllt *k*-Anonymität, wenn jeder Datensatz in der Tabelle **von mindestens *k*-1 anderen Datensätzen** in Bezug auf die **Quasi-Identifikatoren nicht unterscheidbar** ist.

Personen können mit einem Tracker oder Linking Attack nicht mehr eindeutig identifiziert werden.

Homogenitätsangriff

	PLZ	Alter	Geschl.	Diagnose
1	130**	< 30	*	Hepatitis
2	130**	< 30	*	Hepatitis
3	130**	< 30	*	Pneumonie
4	130**	< 30	*	Pneumonie
5	148**	≥ 40	*	Krebs
6	148**	≥ 40	*	Hepatitis
7	148**	≥ 40	*	Pneumonie
8	148**	≥ 40	*	Pneumonie
9	130**	3*	*	Krebs
10	130**	3*	*	Krebs
11	130**	3*	*	Krebs
12	130**	3*	*	Krebs

Paula möchte wissen, nach woran ihr **35-jähriger** Nachbar **Ludwig** erkrankt ist. Beide wohnen im Ort mit der **PLZ 13068**.

Paula beobachtet, dass sich Ludwig irgendwo in der **3. Gruppe** befindet. Nachdem alle Personen in der 3. Gruppe unter **Krebs** leiden, trifft das auch auf Ludwig zu.

ℓ-Diversität

	PLZ	Alter	Geschl.	Diagnose
1	1305*	≤ 40	*	Hepatitis
4	1305*	≤ 40	*	Pneumonie
9	1305*	≤ 40	*	Krebs
10	1305*	≤ 40	*	Krebs
5	1485*	> 40	*	Krebs
6	1485*	> 40	*	Hepatitis
7	1485*	> 40	*	Pneumonie
8	1485*	> 40	*	Pneumonie
2	1306*	≤ 40	*	Hepatitis
3	1306*	≤ 40	*	Pneumonie
11	1306*	≤ 40	*	Krebs
12	1306*	≤ 40	*	Krebs

4-anonyme, 3-diverse Tabelle

Eine Tabelle erfüllt ℓ-Diversität, wenn **jede Gruppe** für **jedes sensitive Attribut** mindestens ℓ **wohldefinierte Werte** beinhaltet.

ℓ-Diversität beugt Angriffen auf die k-Anonymität vor.

Erreichen von k-Anonymität (und ℓ-Diversität)

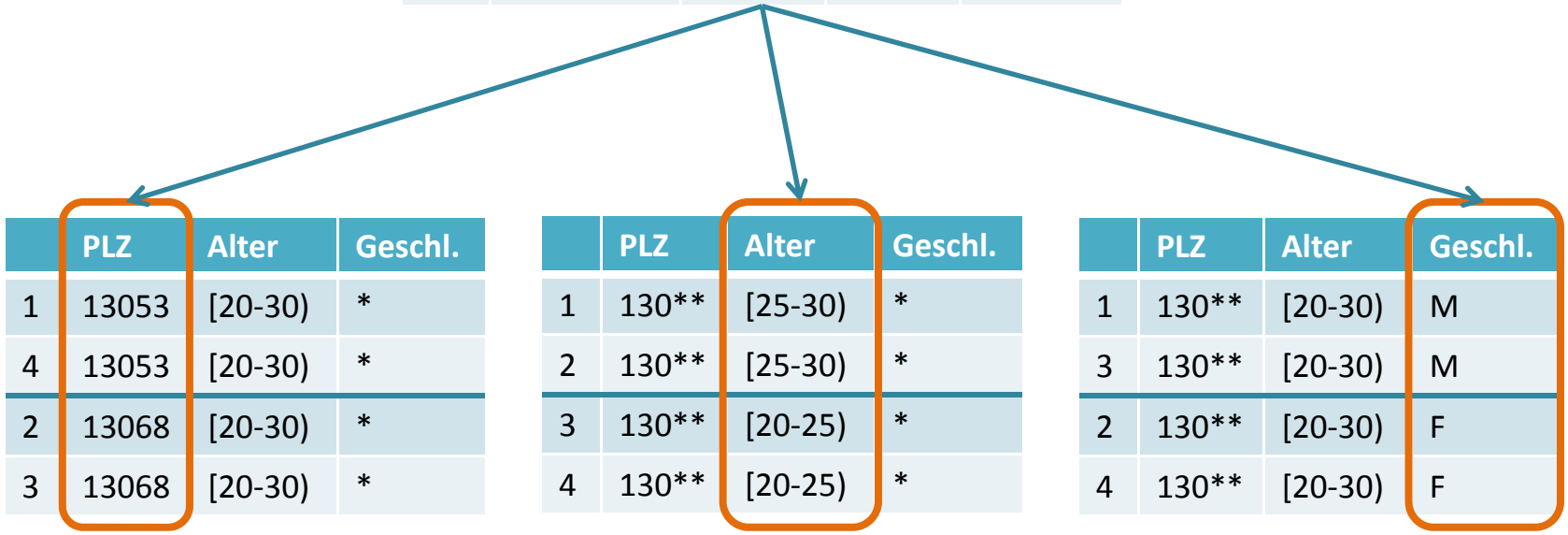
- Entfernung von expliziten Identifikatoren
- Transformation der Attributwerte der Quasi-Identifikatoren
  - Relative anstatt absolute Werte (Alter statt Geburtsjahr)
  - Generalisierung (Hierarchie)
  - Umschlüsselung (BMI statt Gewicht und Größe)
- Entfernen von Attributen
- Entfernen von Datensätzen

Probleme:

- Informationsverlust
- Datenqualität

Transformationsmöglichkeiten

	Name	PLZ	Alter	Geschl.
1	Anton	13053	28	M
2	Berta	13068	29	F
3	Cäsar	13068	21	M
4	Dora	13053	23	F



	PLZ	Alter	Geschl.
1	13053	[20-30)	*
4	13053	[20-30)	*
2	13068	[20-30)	*
3	13068	[20-30)	*

	PLZ	Alter	Geschl.
1	130**	[25-30)	*
2	130**	[25-30)	*
3	130**	[20-25)	*
4	130**	[20-25)	*

	PLZ	Alter	Geschl.
1	130**	[20-30)	M
3	130**	[20-30)	M
2	130**	[20-30)	F
4	130**	[20-30)	F

Nutzwertmaximierung

Nutzwertmaximierung wird mit Hilfe von Berechnung des gewichteten Informationsverlusts ermöglicht.

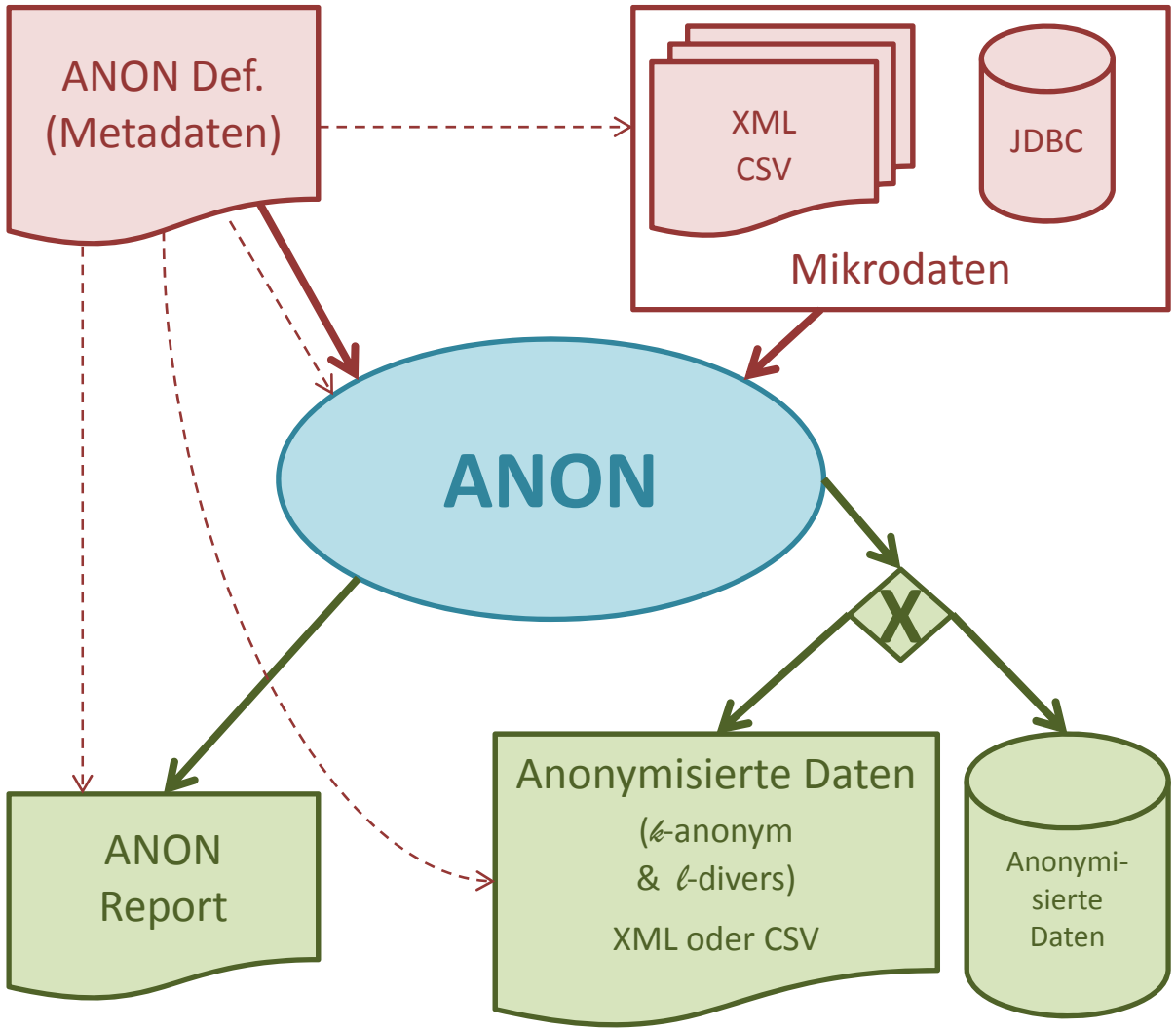
Gewichteter Informationsverlust:

$$\sum_{i=1}^{n} P_{\alpha_i} * V_{\alpha_i}^{L}$$

P_{α_i} ..... Priorität des Attributs α_i

$V_{\alpha_i}^{L}$ ..... anwendungsspezifischer Informationsverlust des Attributs α_i, wenn dieses in den Level L transformiert wird

Anonymisierungstool ANON



Anonymisierungstypen (Beispiel)

Attributname	Anonymisierungstyp
Patient No.	ignore
Name	ignore
Address	ignore
ZIP	k-attribute
Age	k-attribute
Sex	k-attribute
Education	k-attribute
Pseudonym	dontcare
Topology	l-attribute
Staging	l-attribute
Grading	l-attribute
R	l-attribute
V	l-attribute

Eindeutige Identifikatoren

Quasi-Identifikatoren

Weder Identifikatoren, noch sensitiv

Sensitive Attribute

ANON - Inputparameter

- kValue
- Threshold
- SearchType
- WorkReport

```
<Parameters>
  <kValue>5</kValue>
  <Threshold>0.1</Threshold>
  <SearchType>best first</SearchType>
  <WorkReport generateReport="true">
    <writeToURI>../Output/report.xml</writeToURI>
  </WorkReport>
</Parameters>
```

- Input (Mikrodaten)
  - CSV, XML und/oder JDBC
- Output (Anon. Daten
  - CSV, XML oder JDBC

```
<DatasourceDefinition>
  <source>
    <XMLSource>
      <URI>../Input/synth10000.xml</URI>
    </XMLSource>
  </source>
</DatasourceDefinition>
```

ANON - Attribute

- anonymizationType
- ID
- HierarchyID
- SQLName
- Limit
- Priority
- Ldiversity
  - lValue

```
<AttributesDefinition>
  <Attribute type="int" ID="ZIP"
  anonymizationType="k-attribute"
  useGeneralizationHierarchyWithID="GH_ZIP">
    <Label>ZIP</Label>
    <SQLName>zip</SQLName>
    <Limit>4</Limit>
    <Priority>0.1</Priority>
  </Attribute>
  <Attribute type="string" ID="Topology"
  anonymizationType="l-attribute"
  useGeneralizationHierarchyWithID="GH_Topology">
    <Label>Topology</Label>
    <SQLName>topology</SQLName>
    <LDiversity>
      <DistinctLD>
        <lValue>4</lValue>
      </DistinctLD>
    </LDiversity>
  </Attribute>
</AttributesDefinition>
```

ANON - Generalisierungshierarchien

- NumericalHierarchy
  - Levels
    - levelNumber
    - informationLoss
    - Intervals
- CategoricalHierarchy
  - ID
  - Levels
    - levelNumber
    - informationLoss
  - GHTree

```
<NumericalHierarchy id="GH_Age">
  <simpleNumericalHierarchy>
    <GHInfo>
      <Description>Age</Description>
      <Levels>
        <Level informationLoss="0.05"
          stepSize="5" levelNumber="1"/>
        <Level informationLoss="0.1"
          stepSize="10" levelNumber="2"/>
        <Level informationLoss="0.2"
          stepSize="20" levelNumber="3"/>
        <Level informationLoss="0.4"
          stepSize="40" levelNumber="4"/>
        <Level informationLoss="1.0"
          stepSize="100" levelNumber="5"/>
      </Levels>
      <startValue>0</startValue>
      <maxValue>100</maxValue>
    </GHInfo>
  </simpleNumericalHierarchy>
</NumericalHierarchy>
```

Anonymisierungstool ANON

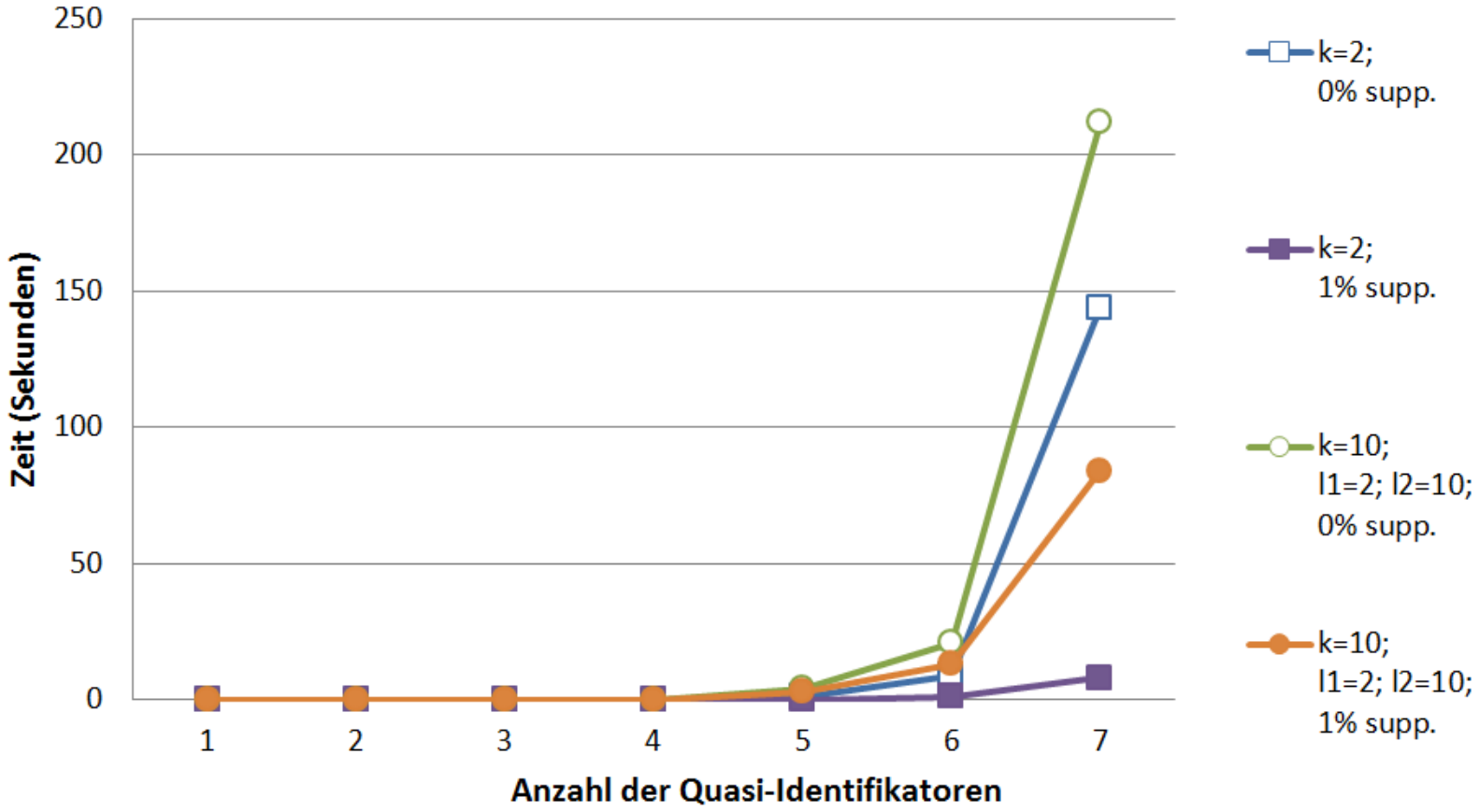
```
<GHTree>
    <rootMember Description="ALL" Name="ALL">
        <member Description="" Name="null_____">
        <member Description="Bestimmte infektiöse und parasitäre Krankheiten" Name="A00-B99">
            <member Description="Infektiöse Darmkrankheiten" Name="A00-A09">
                <member Description="Cholera" Name="A00__">
                    <member Description="Cholera" Name="A00_">
                        <member Description="Cholera" Name="A00">
                            <member Description="Cholera durch Vibrio cholerae O:1, Biovar cholerae" Name="A00.0"/>
                            <member Description="Cholera durch Vibrio cholerae O:1, Biovar eltor" Name="A00.1"/>
                            <member Description="Cholera, nicht näher bezeichnet" Name="A00.9"/>
                        </member>
                    </member>
                </member>
                <member Description="Typhus abdominalis und Paratyphus" Name="A01__">
                    <member Description="Typhus abdominalis und Paratyphus" Name="A01_">
                        <member Description="Typhus abdominalis und Paratyphus" Name="A01">
                            <member Description="Typhus abdominalis" Name="A01.0"/>
                            <member Description="Paratyphus A" Name="A01.1"/>
                            <member Description="Paratyphus B" Name="A01.2"/>
                            <member Description="Paratyphus C" Name="A01.3"/>
                            <member Description="Paratyphus, nicht näher bezeichnet" Name="A01.4"/>
                        </member>
                    </member>
```

ANON - Report

- ANONExceptions
- Status
- kParameter
- lParameter
- resultLevels
- numRemovedTuples
- resultLocation

```
<ANONReport xmlns="report">
  <ANONExceptions>
    <ANONException>
      <code>4102</code>
      <description>lValue is higher than the k-parameter or lower than 1.
      This is not allowed. The default value is k/2.</description>
      <details>Check attribute with the ID Occupation</details>
      <timestamp>08.03.2013 at 17:53:33</timestamp>
    </ANONException>
  </ANONExceptions>
  <resultInformation>
    <status>solution found</status>
    <searchStrategy>best first</searchStrategy>
    <kParameter>5</kParameter>
    <lParameters>salary_class: 2; occupation: 2; </lParameters>
    <resultLevels>age: 1; marital_status: 0; race: 0; sex: 0; </resultLevels>
    <anonymizationDuration>0 seconds</anonymizationDuration>
    <numVisitedNodes>2</numVisitedNodes>
    <numSourceTuples>135666</numSourceTuples>
    <numRemovedTuples>5304</numRemovedTuples>
    <numGroups>278</numGroups>
    <resultLocation>file://C:/Users/user/anondata.csv</resultLocation>
  </resultInformation>
</ANONReport>
```

ANON - Performanz



Adult Data Set, UCI, 45.222 Datensätze

ANON Live Demo

Zusammenfassung

- Informed consent für medizinische Forschung: Anonymität ist eines der wichtigsten Anliegen der Patienten
- Angriffe müssen verhindert werden
- Entfernung der Identifikatoren reicht nicht aus, um mögliche Angriffe zu verhindern
- ANON (k-Anonymität und ℓ-Diversität) verhindert die vorgestellten Angriffe und maximiert den Nutzwert der Daten

Literatur

- Dorothy E. Denning and Peter J. Denning. The tracker: a threat to statistical database security. ACM Trans. Database Syst., 4(1):76-96, 1979.
- J Eder, H Gottweis, and K Zatloukal. It solutions for privacy protection in biobanking. Public Health Genomics, 15(5):254-262, 2012.
- J. Eder, Konrad Stark, and K. Zatloukal. Achieving k-anonymity in datamarts used for gene expressions exploitation. Journal of Integrative Bioinformatics, 4(1):483-495, 2007.
- G. J. Matthews and O. Harel. Data condentiality: A review of methods for statistical disclosure limitation and methods for assessing privacy. Statist. Surv., 5:129, 2011
- A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam. l-diversity: Privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, *1*(1), 2007
- K. Stark, J. Eder, and K. Zatloukal. Priority-based k-anonymity accomplished by weighted generalisation structures. In Proceedings of the 8th international conference on Data Warehousing and Knowledge Discovery, DaWaK'06, pages 394-404, Springer-Verlag, 2006.
- P. Samarati, L. Sweeney. *Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression*. Technical report, SRI International, 1998.
- Latanya Sweeney. Guaranteeing anonymity when sharing medical data, the datay system. In Proceedings of AMIA Annual Fall Symposium, 1997.