Privacy in Databases

Improving privacy in statistical disclosures

Maurice Dibbets (s1022543)
Arlin Kokkelmans (s1016578)
Max Pathuis (s1086382)
Sebastiaan Wortelboer (s1021918)
Patrick Lodeweegs (s1027584)

March 10, 2022

Privacy Seminar

Table of contents

1. Introduction
    - Attacks on databases
    - Legal regulation
2. Privacy enhancement techniques
    - Statistical Disclosure Control
    - Differential privacy
    - k-anonymization
3. Summary

Introduction

Introduction

Attacks on databases

AOL search data leak

Join TechCrunch+
Login

Search
Startups
TechCrunch+
Audio
Newsletters
Videos
Advertise
Events
More

AOL Proudly Releases Massive Amounts of Private Data

Michael Arrington @arrington?lang=en / 3:17 AM GMT+2 • August 7, 2006

Comment

Yet Another Update: AOL: "This was a screw up"

Further Update: Sometime after 7 pm the download link went down as well, but there is at least one mirror site. AOL is in damage control mode – the fact that they took the data down shows that someone there had the sense to realize how destructive this was, but it is also an admission of wrongdoing of sorts. Either way, the data is now out there for anyone that wants to use (or abuse) it.

Update: Sometime around 7 pm PST on Sunday, the AOL site referred to below was taken down. The direct link to the data is still live. A cached copy of the page is here.

AOL must have missed the uproar over the DOJ's demand for "anonymized" search data last year that caused all sorts of pain for Microsoft and Google. That's the only way to explain their release of data that includes 20 million web queries from 650,000 AOL users. [1]

[1] M. Arrington, *AOL Proudly Releases Massive Amounts of Private Data*, `https://techcrunch.com/2006/08/06/aol-proudly-releases-massive-amounts-of-user-search-data/`, Aug. 2006.

FBI watchlist exposed by misconfigured Elasticsearch cluster

A terrorist watchlist was found in an exposed database, and security researcher Bob Diachenko says there is no way of knowing just how long it was open to the public.

By Shaun Nichols   Published: **16 Aug 2021**

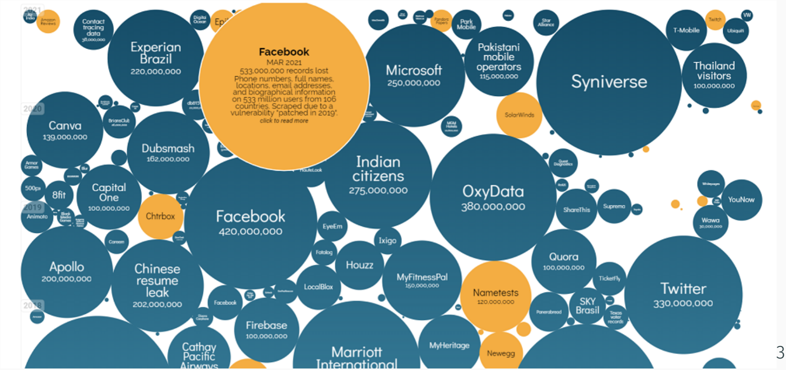
An apparent U.S. government terrorism watchlist was found exposed to the open internet. [2]

---

[2] S. Nichols, *FBI watchlist exposed by misconfigured Elasticsearch cluster*, `https://www.techtarget.com/searchsecurity/news/252505403/FBI-watchlist-exposed-by-misconfigured-Elasticsearch-cluster`, 2021.

Any more examples?

More data breaches



[3]

D. McCandless and T. Evans, *World's Biggest Data Breaches & Hacks*, https://www.informationisbeautiful.net/visualizations/worlds-biggest-data-breaches-hacks/, 2021.

Common attack vectors

- Lack of authentication
- Weak passwords
- SQL injection
- Command execution using malicious shared libraries
- CVEs
- **Statistical inference**
- …

So what is the problem?

- Privacy-sensitive information used for statistical research.
- Statistical research published for scientific purposes. It should not be possible to obtain private information.
- Privacy-preserving research allows participants to be honest in their responses.

Introduction

Legal regulation

CBS

CBS is one of the largest organisations that collects statistical information in the Netherlands.

CBS

CBS is one of the largest organisations that collects statistical information in the Netherlands.

They also published a paper about privacy preserving techniques [5].

Art. 4

Processing means any operation or set of operations which is performed on personal data or on sets of personal data, whether or not by automated means, such as collection, recording, organisation, structuring, storage ...

Art. 4

Processing means any operation or set of operations which is performed on personal data or on sets of personal data, whether or not by automated means, such as collection, recording, organisation, structuring, storage ...

Database owner is processor. Hence, needs to comply with GDPR!

GDPR - other relevant principles

Storage limitation

Art. 5.1e

Personal data shall be kept in a form which permits identification of data subjects for no longer than is necessary for the purposes for which the personal data are processed; personal data may be stored for longer periods insofar as the personal data will be processed solely for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes …

Integrity and confidentiality

Art. 5.1f

Personal data shall be processed in a manner that ensures appropriate security of the personal data, …

Impossibility result

- Information about an individual in a database could be combined with auxiliary information to infer new private information that is not available in the database.

Impossibility result

- Information about an individual in a database could be combined with auxiliary information to infer new private information that is not available in the database.

Can you think of an example?

Impossibility result

- Information about an individual in a database could be combined with auxiliary information to infer new private information that is not available in the database.

Can you think of an example?

- Sometimes, even information about an individual that is **not** in the database could be combined with auxiliary information to infer new private information that is not available in the database.

Impossibility result

- Information about an individual in a database could be combined with auxiliary information to infer new private information that is not available in the database.

Can you think of an example?

- Sometimes, even information about an individual that is **not** in the database could be combined with auxiliary information to infer new private information that is not available in the database.

Can you think of an example?

Impossibility result

- Information about an individual in a database could be combined with auxiliary information to infer new private information that is not available in the database.

Can you think of an example?

- Sometimes, even information about an individual that is **not** in the database could be combined with auxiliary information to infer new private information that is not available in the database.

Can you think of an example?

- This makes semantic security for databases difficult to guarantee!

Impossibility result

- Information about an individual in a database could be combined with auxiliary information to infer new private information that is not available in the database.

Can you think of an example?

- Sometimes, even information about an individual that is **not** in the database could be combined with auxiliary information to infer new private information that is not available in the database.

Can you think of an example?

- This makes semantic security for databases difficult to guarantee!
- In practice, semantic security for databases is **impossible**

Privacy enhancement techniques

Privacy enhancement techniques

- Statistical Disclosure Control
- Differential privacy
- k-anonymization

Privacy enhancement techniques

Statistical Disclosure Control

Statistical Disclosure Control

- Approaches to mitigate risk of disclosing sensitive data
- Statistical queries should not identify individual subjects in the database
- Many different controls are available
- Picking a suitable control depends on the data:
  1. Analyse sensitivity of data.
  2. Analyse use cases of data.
  3. Analyse disclosure risk.

Statistical Disclosure Control

- Sampling: Only part of a table is released.
- Cell suppression: Sensitive table cells are strategically removed.
- Table redesign: Re-code data to reduce sensitivity. For instance: merging several rows.
- Swapping: Table units are swapped.
- Rounding: All table cells are rounded to an integer multiple of a rounding base b.
- Simulation: Generation of synthetic data.

Sampling

Game title	Number of players
Assassin's creed	100,500
Watch dogs	200,000
Star Wars: Battlefront	50,123
Harry Potter and the Order of the Phoenix	144,122

Sampling

Game title	Number of players
Assassin's creed	100,500
Watch dogs	200,000
Star Wars: Battlefront	50,123
Harry Potter and the Order of the Phoenix	144,122

Game title	Number of players
Watch dogs	200,000
Harry Potter and the Order of the Phoenix	144,122

Cell suppression

Course	Grade
Privacy Seminar	7
Privacy Seminar	8
Privacy Seminar	2
Average	$5\frac{2}{3}$

Figure 1: Example primary suppression

Cell suppression

Course	Grade
Privacy Seminar	7
Privacy Seminar	8
Privacy Seminar	2
Average	$5^2/_3$

Figure 1: Example primary suppression

Do you know how such a suppression might still leak information?

Cell suppression

row	A	B	C	D
1	x_1	4	2	x_2
2	x_3	0	6	2
3	6	2	x_4	8
4	1	7	9	5

Category	Average
A	$5^3/_4$
B	$3^1/_4$
C	6
D	$5^3/_4$

Row	Average
1	$5^1/_4$
2	$4^1/_4$
3	$5^3/_4$
4	$5^1/_2$

Cell suppression

row	A	B	C	D
1	7	4	2	8
2	9	0	6	2
3	6	2	7	8
4	1	7	9	5

Category	Average
A	$5^3/_4$
B	$3^1/_4$
C	6
D	$5^3/_4$

Row	Average
1	$5^1/_4$
2	$4^1/_4$
3	$5^3/_4$
4	$5^1/_2$

$$\begin{aligned} x_1 &= 21 - 4 - 2 - x_2 \\ &= 15 - (23 - 2 - 8 - 5) \\ &= 7 \end{aligned}$$

$$\begin{aligned} x_2 &= 4 \cdot 5^3/_4 - 2 - 8 - 5 \\ &= 23 - 2 - 8 - 5 \\ &= 8 \end{aligned}$$

$$\begin{aligned} x_3 &= 4 \cdot 4^1/_4 - 0 - 6 - 2 \\ &= 17 - 8 \\ &= 9 \end{aligned}$$

$$\begin{aligned} x_4 &= 4 \cdot 6 - 2 - 6 - 9 \\ &= 24 - 17 \\ &= 7 \end{aligned}$$

Table redesign

Game title	Number of players
Assassin’s creed	100,500
Watch dogs	200,000
Star Wars: Battlefront	50,123
Harry Potter and the Order of the Phoenix	144,122

Table redesign

Game title	Number of players
Assassin's creed	100,500
Watch dogs	200,000
Star Wars: Battlefront	50,123
Harry Potter and the Order of the Phoenix	144,122

Game publisher	Number of players
Ubisoft	300,500
EA	194,245

Swapping

Student	Length
Arlin	176
Maurice	196
Sebastiaan	195
Patrick	182

Swapping

Student	Length
Arlin	176
Maurice	196
Sebastiaan	195
Patrick	182

Student	Length
Arlin	195
Maurice	196
Sebastiaan	176
Patrick	182

Rounding

Student	Length
Arlin	176
Maurice	196
Sebastiaan	195
Patrick	182

Rounding

Student	Length
Arlin	176
Maurice	196
Sebastiaan	195
Patrick	182

Student	Length
Arlin	180
Maurice	200
Sebastiaan	200
Patrick	180

Rounding

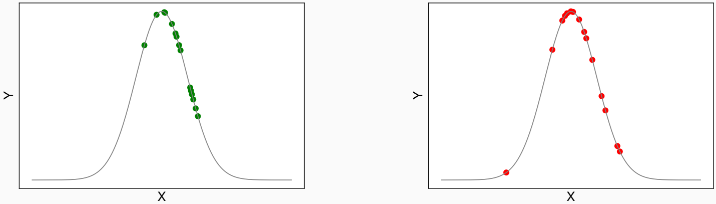
Student	Length
Arlin	176
Maurice	196
Sebastiaan	195
Patrick	182

Student	Length
Arlin	180
Maurice	200
Sebastiaan	200
Patrick	180

The entropy depends on the rounding base b as every original value is in an interval of the rounded value (n_i):

$$I_i = [n_i - 1/2b, n_i + 1/2b] \quad (1)$$

Simulation



(a) Original dataset

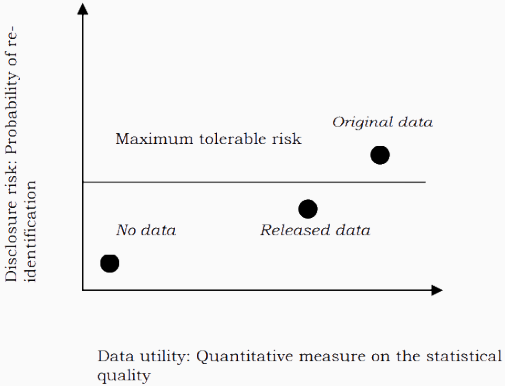
(b) Simulated dataset

Figure 2: Simulating points based on a standard distribution

Evaluating information utility

- Depends on variable type.
- Categorical variables: check equivalence
- Continuous variables: measure correlation

Statistical Disclosure Control



Privacy enhancement techniques

Differential privacy

Differential privacy

- Algorithm $\mathcal{A}$ analyses a dataset D and computes statistics (mean, variance, mode, median, etc)

Differential privacy

- Algorithm $\mathcal{A}$ analyses a dataset D and computes statistics (mean, variance, mode, median, etc)

Differential Privacy

An algorithm $\mathcal{A}$ is differentially private if by looking at the output, it is impossible to determine whether any individual's data is included in the dataset or not.

Statistical inference - an example

- Consider a dataset D_2 that differs from D_1 in only **one** row
- Statistical differences between D_2 and D_1 can leak information about this row. **How?**

Statistical inference - an example

- Consider a dataset D_2 that differs from D_1 in only **one** row
- Statistical differences between D_2 and D_1 can leak information about this row. How?

D_1	Name	Age
	Arlin	23
	Maurice	25
	Max	24
	Sebastiaan	37
	Patrick	20

Statistical inference - an example

- Consider a dataset D_2 that differs from D_1 in only **one** row
- Statistical differences between D_2 and D_1 can leak information about this row. **How?**

D_1 Name	Age
Arlin	23
Maurice	25
Max	24
Sebastiaan	37
Patrick	20

$\mathcal{A}(D_1)$

Average :	25.8
Median :	24
Variance :	34.16

Statistical inference - an example

- Consider a dataset D_2 that differs from D_1 in only **one** row
- Statistical differences between D_2 and D_1 can leak information about this row. **How?**

D_1 Name	Age
Arlin	23
Maurice	25
Max	24
Sebastiaan	37
Patrick	20

D_2 Name	Age
Arlin	23
Maurice	25
Max	24
Sebastiaan	37

$$\mathcal{A}(D_1)$$

$$\begin{aligned} \textit{Average}: &\quad 25.8 \\ \textit{Median}: &\quad 24 \\ \textit{Variance}: &\quad 34.16 \end{aligned}$$

Statistical inference - an example

- Consider a dataset D_2 that differs from D_1 in only **one** row
- Statistical differences between D_2 and D_1 can leak information about this row. **How?**

D_1	Name	Age
	Arlin	23
	Maurice	25
	Max	24
	Sebastiaan	37
	Patrick	20

D_2	Name	Age
	Arlin	23
	Maurice	25
	Max	24
	Sebastiaan	37

$\mathcal{A}(D_1)$

$$\begin{aligned} Average: &\quad 25.8 \\ Median: &\quad 24 \\ Variance: &\quad 34.16 \end{aligned}$$

$\mathcal{A}(D_2)$

$$\begin{aligned} Average: &\quad 27.25 \\ Median: &\quad 24.5 \\ Variance: &\quad 42.917 \end{aligned}$$

Query set control

What if we introduce lower and upper bounds to the query size?

- Consider a threshold t such that any query must involve at least a set of $\geq t$ rows.
- For a database of N entries, only allow queries on a subset size between t and N - t
- Don't allow successive queries of sets K and L if $K \subseteq L$ and $|L| - |K| < t$

Query set control

What if we introduce lower and upper bounds to the query size?

- Consider a threshold t such that any query must involve at least a set of $\geq t$ rows.
- For a database of N entries, only allow queries on a subset size between t and N - t
- Don't allow successive queries of sets K and L if $K \subseteq L$ and $|L| - |K| < t$

Does this solve the problem?

Query set control

What if we introduce lower and upper bounds to the query size?

- Consider a threshold t such that any query must involve at least a set of $\geq t$ rows.
- For a database of N entries, only allow queries on a subset size between t and N - t
- Don't allow successive queries of sets K and L if $K \subseteq L$ and $|L| - |K| < t$

Does this solve the problem?

An attacker could make many, many more queries to eventually circumvent this limitation. Query set control might not be the right approach.

Introducing randomness

What if we add random noise to $\mathcal{A}$ to (slightly) distort results.

A simple protocol to determine if a row has a certain property:

1. Flip a coin.
2. If tails, respond truthfully.
3. If heads, flip a second coin.
   a. If heads again, respond 'Yes'.
   b. If tails again, respond 'No'.

Introducing randomness

What if we add random noise to $\mathcal{A}$ to (slightly) distort results.

A simple protocol to determine if a row has a certain property:

1. Flip a coin.
2. If tails, respond truthfully.
3. If heads, flip a second coin.
   a. If heads again, respond 'Yes'.
   b. If tails again, respond 'No'.

Does this solve the problem?

Introducing randomness

What if we add random noise to $\mathcal{A}$ to (slightly) distort results.

A simple protocol to determine if a row has a certain property:

1. Flip a coin.
2. If tails, respond truthfully.
3. If heads, flip a second coin.
    a. If heads again, respond 'Yes'.
    b. If tails again, respond 'No'.

Does this solve the problem?

- Random noise allows for *refutability*.

Introducing randomness

What if we add random noise to $\mathcal{A}$ to (slightly) distort results.
A simple protocol to determine if a row has a certain property:

1. Flip a coin.
2. If tails, respond truthfully.
3. If heads, flip a second coin.
   a. If heads again, respond 'Yes'.
   b. If tails again, respond 'No'.

Does this solve the problem?

- Random noise allows for *refutability*.
- The accuracy is not always ideal, but if ρ rows contain the attribute, we can expect $(\frac{1}{4})(1-\rho)+(\frac{3}{4})\rho = \frac{1}{4}+\frac{\rho}{2}$ positive responses.
- Since ρ can be estimated, sufficiently large datasets can have significant statistics.

Calibrating noise to sensitivity

- A transcript is the interaction between a user and a privacy mechanism
$$t = [Q1, a1, Q2, a2..., Qd, ad] \tag{2}$$

- Ideally, noise should be optimised to acceptable margin of error

- Use random noise function with a carefully chosen distribution

- **Sensitivity** - the maximum amount that any single argument to a function can change its output

$$\Delta\mathcal{A} = \max_{D_i, D_j \in \mathcal{S}} (||\mathcal{A}(D_i), \mathcal{A}(D_j)||_1) \tag{3}$$

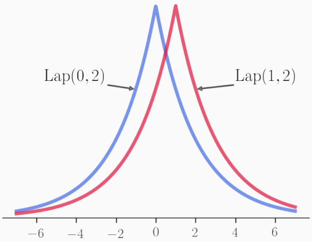
Here, $\mathcal{S}$ denotes the set of all pairs of databases that differ from each other in at most one row, and $|| \cdot ||_1$ denotes the ℓ_1 norm (Manhattan distance)

Additive noise mechanisms

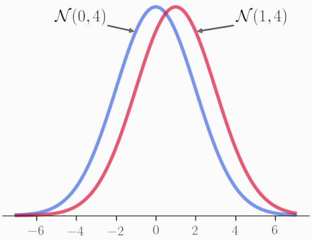
- For each query, the server either refuses to answer, or answers $f_i(x)$ + the desired amount of noise, where $f_i(x)$ is the requested information.
- Ideally, these have a *low sensitivity* (≤ 1)
- Controlled noise based on a carefully chosen probability distribution

Additive noise mechanisms

- For each query, the server either refuses to answer, or answers $f_i(x)$ + the desired amount of noise, where $f_i(x)$ is the requested information.
- Ideally, these have a *low sensitivity* (≤ 1)
- Controlled noise based on a carefully chosen probability distribution



(a) LaPlace



(b) Gaussian

Exponential mechanism

- Adding random noise doesn't work for some types of data

Can you think of some examples?

Exponential mechanism

- Adding random noise doesn't work for some types of data

Can you think of some examples?

- Consider a set $\mathcal{R}$ of possible outputs we are interested in
- Design a scoring function $u : D \times \mathcal{R} \rightarrow \mathbb{R}$ with sensitivity Δu
- Output $r \in \mathcal{R}$ will have a probability proportional to:

$$Pr[r] = \exp(\frac{\epsilon \cdot u(d, r)}{2 \cdot \Delta u}) \tag{4}$$

- This is the probability defined in r, which is the possibility for a single r to be selected.

A mathematical notion of differential privacy

- Ideally, $\mathcal{A}(D_1)$ should be hard to distinguish from $\mathcal{A}(D_2)$
- Consider ϵ the maximum distance between a query on D_1 and the same query on D_2.
- Then, $\exp(\epsilon)$ provides us with the dilation of the probability.

$$\Pr[\mathcal{A}(D_1) \in S] \leq \exp(\epsilon) \cdot \Pr[\mathcal{A}(D_2) \in S] \tag{5}$$

Extended for group privacy: instead of difference in **one** row, consider difference of **c** rows

$$\Pr[\mathcal{A}(D_1) \in S] \leq \exp(\epsilon \cdot \mathbf{c}) \cdot \Pr[\mathcal{A}(D_2) \in S] \tag{6}$$

Privacy enhancement techniques

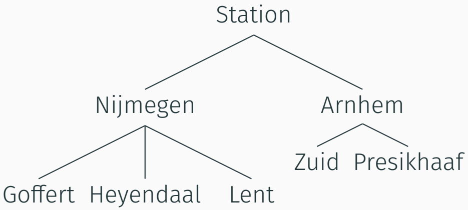
k-anonymization

k-anonymization

- Each data point is indistinguishable from $k - 1$ other data points
- Trade-off between equivalence class size and minimal loss of data utility
- Three main steps:
  1. Partition data into clusters
  2. Re-assign attributes to ensure each cluster has at least k points
  3. Anonymization of the original data values to something useful:
     - Numerical values: centeroids
     - Categorical values: common ancestor
- An optimal solution is NP-hard

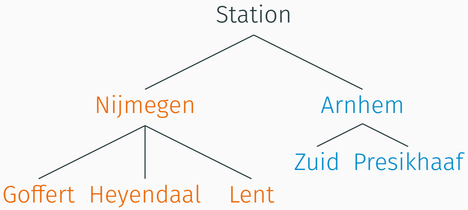
Example dataset: Public transport

Departure station	Distance
Nijmegen	8km
Nijmegen Goffert	5km
Nijmegen Heyendaal	73km
Nijmegen Lent	9km
Arnhem	6km
Arnhem zuid	14km
Arnhem Presikhaaf	24km



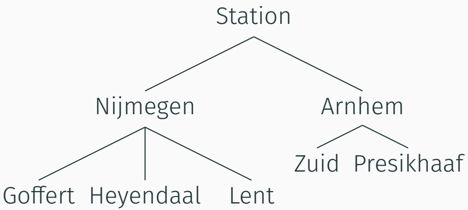
Example dataset: Public transport

Departure station	Distance
Nijmegen	8km
Nijmegen Goffert	5km
Nijmegen Heyendaal	73km
Nijmegen Lent	9km
Arnhem	6km
Arnhem zuid	14km
Arnhem Presikhaaf	24km



Example dataset: Public transport

Departure station	Distance
Nijmegen	$23^3/_4$
Nijmegen	8km
Nijmegen Goffert	5km
Nijmegen Heyendaal	73km
Nijmegen Lent	9km
Arnhem	$14^2/_3$
Arnhem	6km
Arnhem zuid	14km
Arnhem Presikhaaf	24km



In total: $2^c - c - 1 = 2^8 - 8 - 1 = 247$ options

What are limitations of k-anonimity?

Attacks on k-anonimity

- Background knowledge attack: use demographics and public records to increase probability of identifying records.
- Homogeneity attack: attack reveals private information when all values of sensitive attributes are the same in a equivalence class.

l-diversity: distinctness

Any generalized attribute should consist of sufficiently many different sensitive values.

- Distinctness can be ensured with well-represented groups
- An attacker needs information about $l-1$ data points to infer a specific data point

Departure station	Distance
Nijmegen	27
Nijmegen Goffert	6km
Nijmegen Heyendaal	73km
Nijmegen Lent	9km
Arnhem	$14^2/_3$
Arnhem Zuid	6km
Arnhem Zuid	14km
Arnhem Zuid	24km

Entropy l-diversity

A measure to determine if there is sufficient distinctness

- q a generalized nonsensitive value ("Arnhem" as departure station instead of "Arnhem Zuid")
- s a possible value of a sensitive attribute S
- $p(q, s)$ fraction of data points white nonsensitive value q and sensitive value s.
- l the protection against l data points of background knowledge

$$-\sum_{s \in S} p(q, s) \ln(p_{(q,s')}) \geq \ln(l) \tag{7}$$

Entropy l-diversity

Q	S
Arnhem	$\mathcal{F}$
Arnhem	$\mathcal{T}$
Arnhem	$\mathcal{F}$
Arnhem	$\mathcal{F}$
Arnhem	$\mathcal{F}$

Entropy l-diversity

Q	S
Arnhem	$\mathcal{F}$
Arnhem	$\mathcal{T}$
Arnhem	$\mathcal{F}$
Arnhem	$\mathcal{F}$
Arnhem	$\mathcal{F}$

$$
\begin{aligned}
p(A, \mathcal{F}) &= \frac{4}{5} \\
p(A, \mathcal{T}) &= \frac{1}{5} \\
-\sum_{s \in S} p_{(q,s)} \ln(p_{(q,s')}) &= -\left(p_{(A,\mathcal{F})} \ln p_{(A,\mathcal{T})} + p_{(A,\mathcal{F})} \ln p_{(A,\mathcal{T})}\right) \\
&= -\left(\frac{4}{5} \cdot \ln \frac{1}{5} + \frac{1}{5} \cdot \ln \frac{4}{5}\right) \\
&\approx 0.18 + 0.32 \\
&\approx 0.5
\end{aligned}
$$

$$
-\sum_{s \in S} p(q, s) \ln(p_{(q,s')}) \geq \ln(l)
$$

Entropy l-diversity

Q	S
Arnhem	$\mathcal{F}$
Arnhem	$\mathcal{T}$
Arnhem	$\mathcal{F}$
Arnhem	$\mathcal{F}$
Arnhem	$\mathcal{F}$

$$p(A, \mathcal{F}) = \frac{4}{5}$$

$$p(A, \mathcal{T}) = \frac{1}{5}$$

$$\begin{aligned}
-\sum_{s \in S} p_{(q,s)} \ln(p_{(q,s')}) &= -\left(p_{(A,\mathcal{F})} \ln p_{(A,\mathcal{T})} + p_{(A,\mathcal{F})} \ln p_{(A,\mathcal{T})}\right) \\
&= -\left(\frac{4}{5} \cdot \ln \frac{1}{5} + \frac{1}{5} \cdot \ln \frac{4}{5}\right) \\
&\approx 0.18 + 0.32 \\
&\approx 0.5
\end{aligned}$$

$$-\sum_{s \in S} p(q, s) \ln(p_{(q,s')}) \geq \ln(l)$$

$$0.5 \not\geq 0.69$$

This does not hold for $l = 2$!

Summary

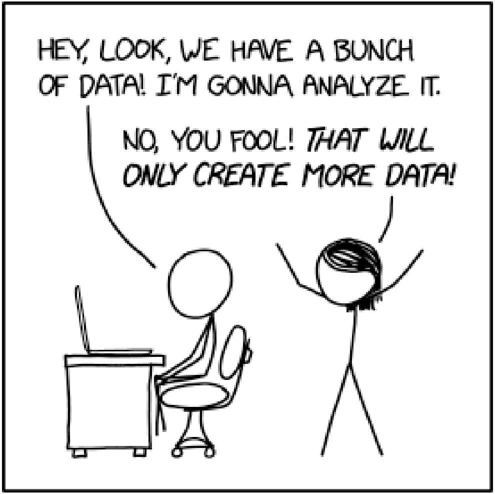
Summary

Statistical Disclosure Control	Differential Privacy	K-anonymity
Easy to implement	Some privacy guarantees	Prerequisite for privacy protection
Prerequisite for other approaches	Refutability	k-anonymity is NP-hard
Protection only for accounted attacks	Too much noise reduces utility	Too large k reduces utility

Summary

- Statistical disclosure methods can help.
- Disclosure risk vs data utility.
- Combination of methods provides most protection.

Questions?



It's important to make sure your analysis destroys as much information as it produces.

[4]

[4] https://xkcd.com/2582/

References

[1] M. Arrington, *AOL Proudly Releases Massive Amounts of Private Data*, https://techcrunch.com/2006/08/06/aol-proudly-releases-massive-amounts-of-user-search-data/, Aug. 2006.

[2] S. Nichols, *FBI watchlist exposed by misconfigured Elasticsearch cluster*, https://www.techtarget.com/searchsecurity/news/252505403/FBI-watchlist-exposed-by-misconfigured-Elasticsearch-cluster, 2021.

[3] D. McCandless and T. Evans, *World's Biggest Data Breaches & Hacks*, https://www.informationisbeautiful.net/visualizations/worlds-biggest-data-breaches-hacks/, 2021.

References

[4] C. McNab, *Network Security Assessment*. 2016, vol. 3, ch. 15.

[5] C. B. voor de Statistiek, *Privacy preserving techniques and statistical disclosure control*, https://www.cbs.nl/-/media/imported/onze-diensten/methoden/gevalideerde-methoden/output/documents/2012/28/2012-statistical-disclosure-control-art.pdf, 2012.

[6] E. Parliament, *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)*, https://eur-lex.europa.eu/eli/reg/2016/679/oj, Apr. 2016.

References

[7] C. Dwork, “Differential privacy,” in *33rd International Colloquium on Automata, Languages and Programming*, 2006, pp. 1–12. DOI: 10.1007/11787006.

[8] J. Domingo-Feffer, A. Oganian, and V. Torra, “Information-theoretic disclosure risk measures in statistical disclosure control of tabular data,” in *Proceedings 14th International Conference on Scientific and Statistical Database Management*, 2002, pp. 227–231. DOI: 10.1109/SSDM.2002.1029724.

[9] T. Dalenius and S. P. Reiss, “Data-swapping: A technique for disclosure control,”, Elsevier, 1978. DOI: 10.1016/0378-3758(82)90058-1.

References

[10] A. Hundepool, J. Domingo-Ferrer, L. Franconi, *et al.*, *Handbook on Statistical Disclosure Control*. 2010, pp. 9–14. [Online]. Available: `https://ec.europa.eu/eurostat/cros/system/files/SDC_Handbook.pdf`.

[11] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in *Theory of Cryptography*, S. Halevi and T. Rabin, Eds., Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 265–284, ISBN: 978-3-540-32732-5.

[12] R. Bayardo and R. Agrawal, "Data privacy through optimal k-anonymization," in *21st International Conference on Data Engineering (ICDE'05)*, 2005, pp. 217–228. DOI: `10.1109/ICDE.2005.42`.

References

[13] B. S. Bhati, J. Ivanchev, I. Bojic, A. Datta, and D. Eckhoff, "Utility-driven k-anonymization of public transport user data," *IEEE Access*, vol. 9, pp. 23 608–23 623, 2021. DOI: 10.1109/ACCESS.2021.3055505.

[14] J. Domingo-Ferrer and V. Torra, "Ordinal, continuous and heterogeneous k-anonymity through microaggregation," en, *Data Min. Knowl. Discov.*, vol. 11, no. 2, pp. 195–212, Sep. 2005.

[15] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam, "L-diversity: Privacy beyond k-anonymity," in *22nd International Conference on Data Engineering (ICDE'06)*, 2006, pp. 24–24. DOI: 10.1109/ICDE.2006.1.

$l$diversity: recursiveness

Recursive definition:

q* as defined earlier, l being the number of sensitive values

A q*-block is (c, 2)-diverse if $r_1 < c(r_2 + \cdots + r_m)$ for chosen constant c

For $l > 2$:

(c, l)-diversity if we can eliminate one sensitive value and (c,l-1)-diversity still holds