m-Privacy for Collaborative Data Publishing

by

Slawomir Goryczka, Li Xiong, Benjamin C. M. Fung

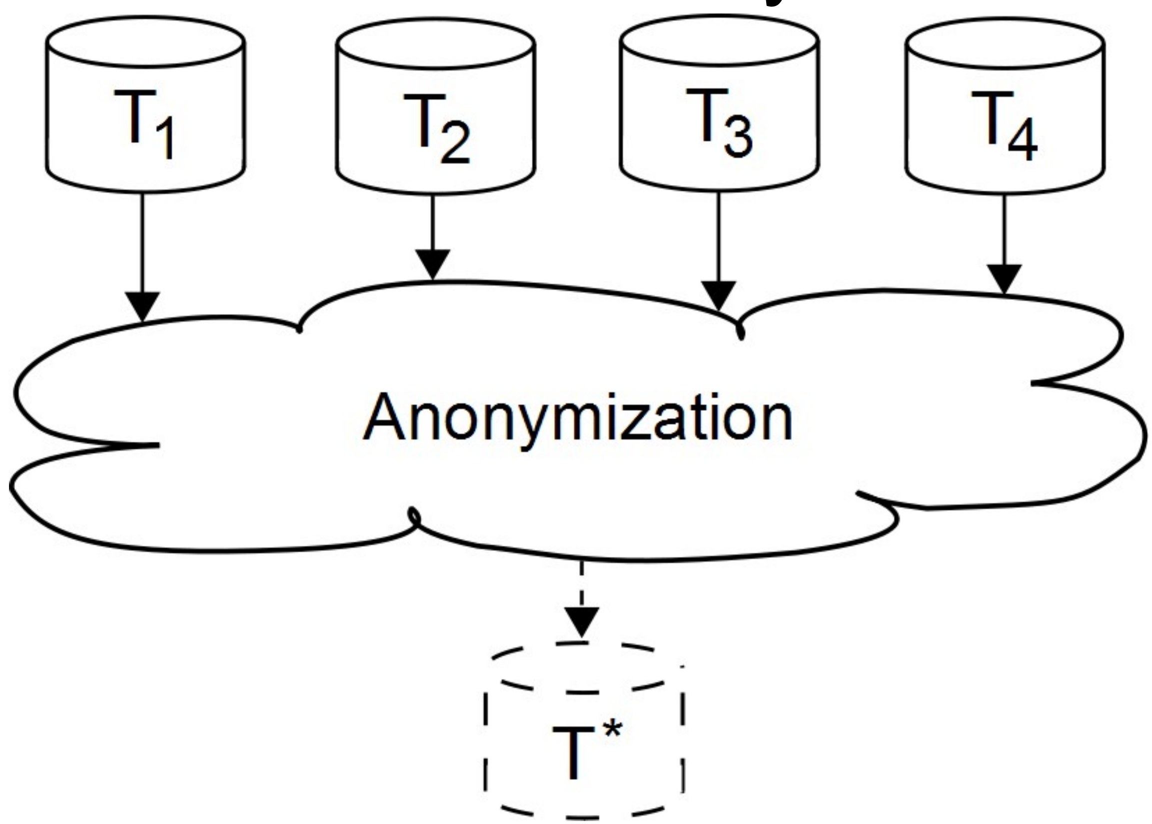
EMORY UNIVERSITY

EMORY UNIVERSITY

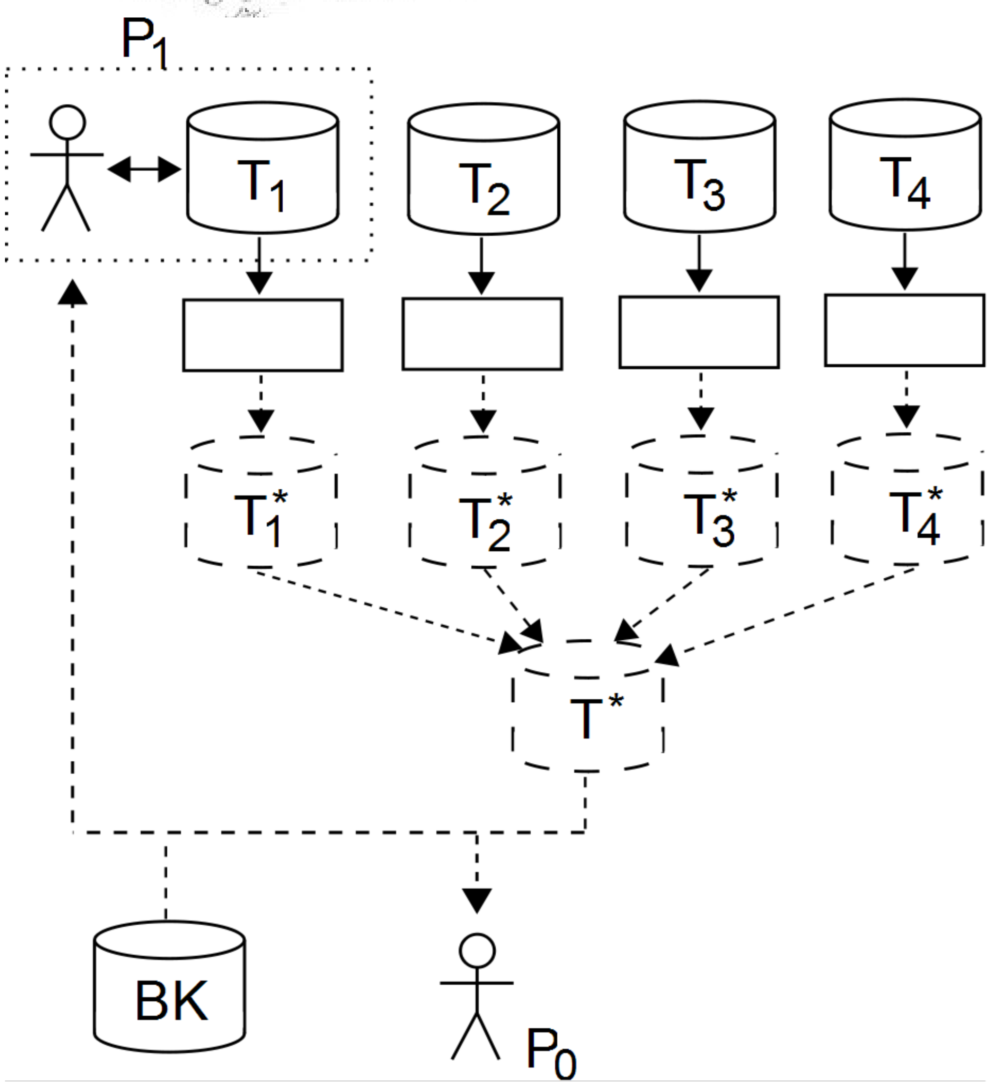
UNIVERSITÉ Concordia UNIVERSITY

Collaborative Data Publishing

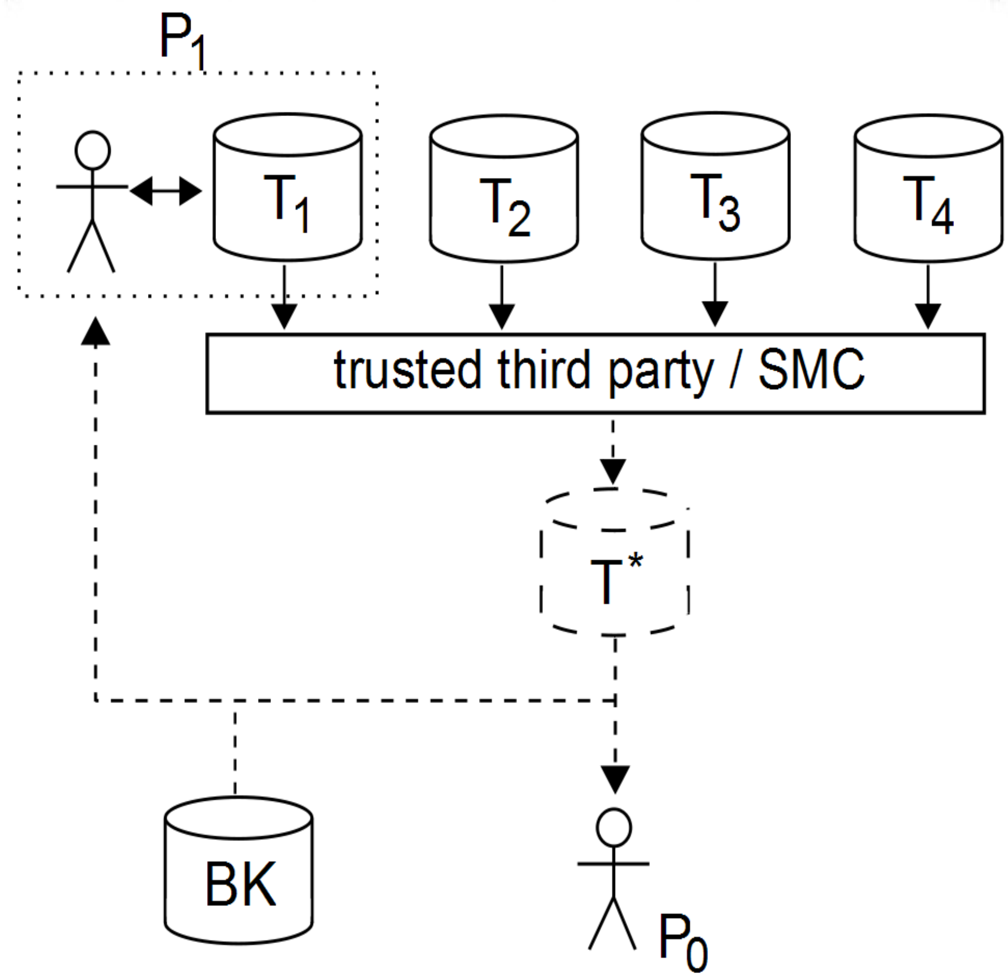
- Many data providers, e.g. hospitals, wish to publish anonymized view of their data,
- Different scenarios of anonymization.



Distributed Anonymization



anonymize-and-aggregate



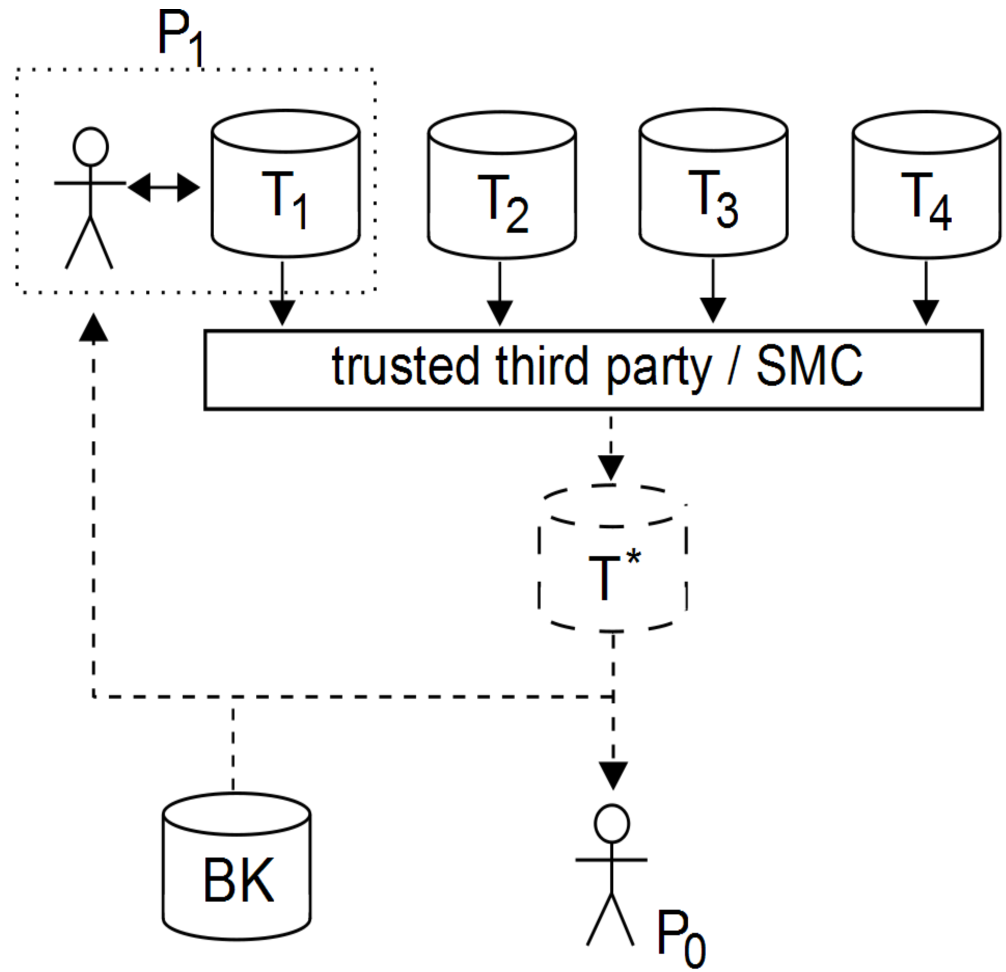
aggregate-and-anonymize

Privacy Concerns in Collaborative Data Publishing

- Potential attackers:
  - Data recipients P_0

    Data are private due to privacy of T^*
  - Data providers P_1

    Data may not be private due to instance level knowledge of P_1



aggregate-and-anonymize

Anonymization Example (data)

- Data attributes:
  - Identifiers, e.g. Name
  - Quasi-Identifiers (QI), e.g. Age, Zip
  - Sensitive, e.g. Disease

BK

Name	Age	Zip
Alice	2*	98***
Bob	3*	12***
Sara	2*	12***
Dorothy	3*	98***
...		

Voters registration list

T_1

Name	Age	Zip	Disease
Alice	24	98745	Cancer
Bob	35	12367	Asthma
Emily	22	98712	Asthma

T_2

Name	Age	Zip	Disease
Dorothy	38	98701	Cancer
Mark	37	12389	Flu
John	31	12399	Flu

T_3

Name	Age	Zip	Disease
Sara	20	12300	Epilepsy
Cecilia	39	98708	Flu

T_4

Name	Age	Zip	Disease
Olga	32	12337	Cancer
Frank	33	12388	Asthma

Anonymization Example (attack)

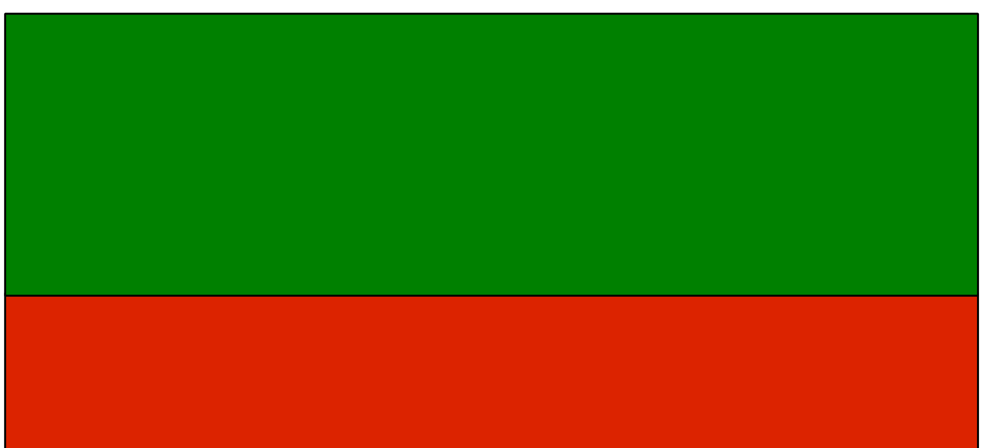
- Privacy is defined as *k*-anonymity and (simple) *l*-diversity (*k* = 2, *l* = 2).

T_a^*

Provider	Name	Age	Zip	Disease
P_1	~~*Alice*~~	~~[20-30]~~	~~*****~~	~~Cancer~~
P_1	~~*Emily*~~	~~[20-30]~~	~~*****~~	~~Asthma~~
P_3	***Sara***	**[20-30]**	*********	**Epilepsy**
P_1	~~*Bob*~~	~~[31-35]~~	~~*****~~	~~Asthma~~
P_2	*John*	[31-35]	*****	Flu
P_4	*Olga*	[31-35]	*****	Cancer
P_4	*Frank*	[31-35]	*****	Asthma
P_2	*Dorothy*	[36-40]	*****	Cancer
P_2	*Mark*	[36-40]	*****	Flu
P_3	*Cecilia*	[36-40]	*****	Flu

m-Privacy

An equivalence group of anonymized records is *m*-private with respect to a privacy constraint *C* if any coalition of *m* parties (*m*-adversary) is not able to breach privacy of remaining records.



Private records provided by other parties.

Records provided by *m*-adversary

Anonymization Example

- An attacker is a single data provider (1-privacy)

T_b^*

Provider	*Name*	**Age**	**Zip**	**Disease**
P_1	*Alice*	[20-40]	*****	Cancer
P_2	*Mark*	[20-40]	*****	Flu
P_3	*Sara*	[20-40]	*****	Epilepsy
P_1	*Emily*	[20-40]	987**	Asthma
P_2	*Dorothy*	[20-40]	987**	Cancer
P_3	*Cecilia*	[20-40]	987**	Flu
P_1	*Bob*	[20-40]	123**	Asthma
P_4	*Olga*	[20-40]	123**	Cancer
P_4	*Frank*	[20-40]	123**	Asthma
P_2	*John*	[20-40]	123**	Flu

Anonymization Example

- An attacker is a single data provider (1-privacy)

T_b^*

Provider	Name	Age	Zip	Disease
P_1	*Alice*	[20-40]	*****	Cancer
P_2	*Mark*	[20-40]	*****	Flu
P_3	*Sara*	[20-40]	*****	Epilepsy
P_1	*Emily*	[20-40]	987**	Asthma
P_2	*Dorothy*	[20-40]	987**	Cancer
P_3	*Cecilia*	[20-40]	987**	Flu
P_1	*Bob*	[20-40]	123**	Asthma
P_4	*Olga*	[20-40]	123**	Cancer
P_4	*Frank*	[20-40]	123**	Asthma
P_2	*John*	[20-40]	123**	Flu

Anonymization Example

- An attacker is a single data provider (1-privacy)

T_b^*

Provider	*Name*	Age	Zip	Disease
~~P_1~~	~~*Alice*~~	~~[20-40]~~	~~*****~~	~~Cancer~~
P_2	*Mark*	[20-40]	*****	Flu
P_3	*Sara*	[20-40]	*****	Epilepsy
P_1	*Emily*	[20-40]	987**	Asthma
P_2	*Dorothy*	[20-40]	987**	Cancer
P_3	*Cecilia*	[20-40]	987**	Flu
P_1	*Bob*	[20-40]	123**	Asthma
P_4	*Olga*	[20-40]	123**	Cancer
P_4	*Frank*	[20-40]	123**	Asthma
P_2	*John*	[20-40]	123**	Flu

Anonymization Example

- An attacker is a single data provider (1-privacy)

T_b^*

Provider	Name	Age	Zip	Disease
P_1	*Alice*	[20-40]	*****	Cancer
P_2	*Mark*	[20-40]	*****	Flu
P_3	*Sara*	[20-40]	*****	Epilepsy
P_1	*Emily*	[20-40]	987**	Asthma
P_2	*Dorothy*	[20-40]	987**	Cancer
P_3	*Cecilia*	[20-40]	987**	Flu
P_1	*Bob*	[20-40]	123**	Asthma
P_4	*Olga*	[20-40]	123**	Cancer
P_4	*Frank*	[20-40]	123**	Asthma
P_2	*John*	[20-40]	123**	Flu

Anonymization Example

- An attacker is a single data provider (1-privacy)

T_b^*

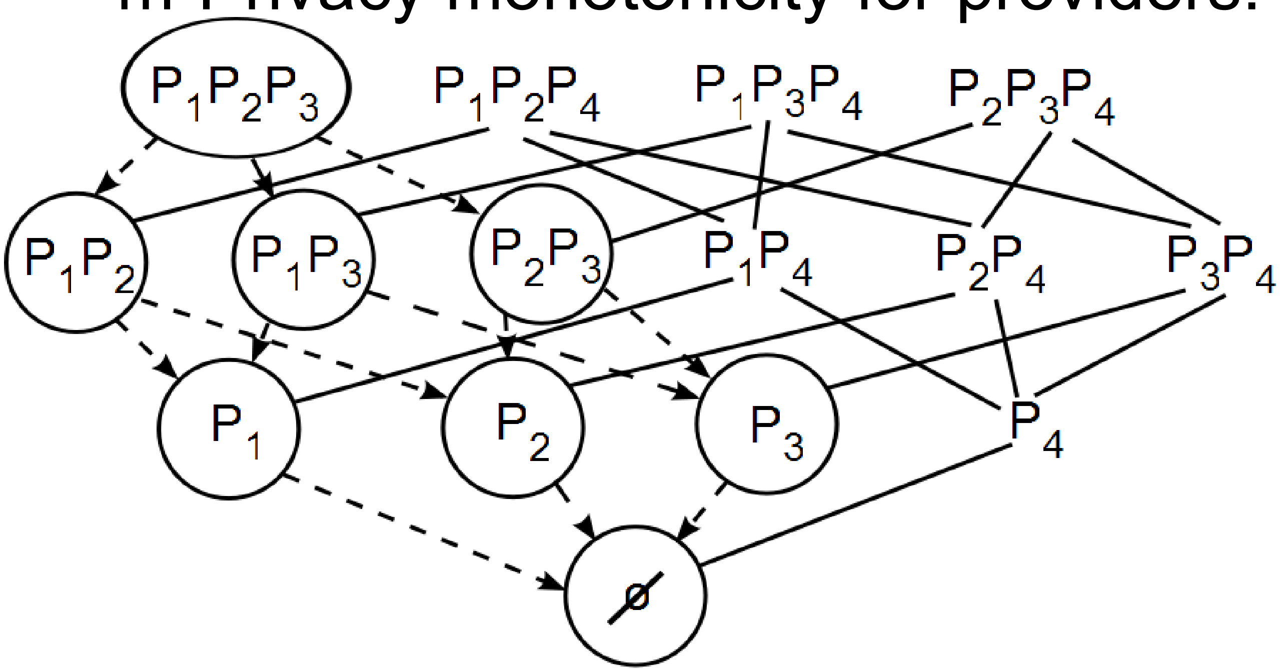
Provider	*Name*	Age	Zip	Disease
P_1	*Alice*	[20-40]	*****	Cancer
P_2	*Mark*	[20-40]	*****	Flu
P_3	*Sara*	[20-40]	*****	Epilepsy
P_1	*Emily*	[20-40]	987**	Asthma
P_2	*Dorothy*	[20-40]	987**	Cancer
P_3	*Cecilia*	[20-40]	987**	Flu
P_1	*Bob*	[20-40]	123**	Asthma
P_4	*Olga*	[20-40]	123**	Cancer
P_4	*Frank*	[20-40]	123**	Asthma
P_2	*John*	[20-40]	123**	Flu

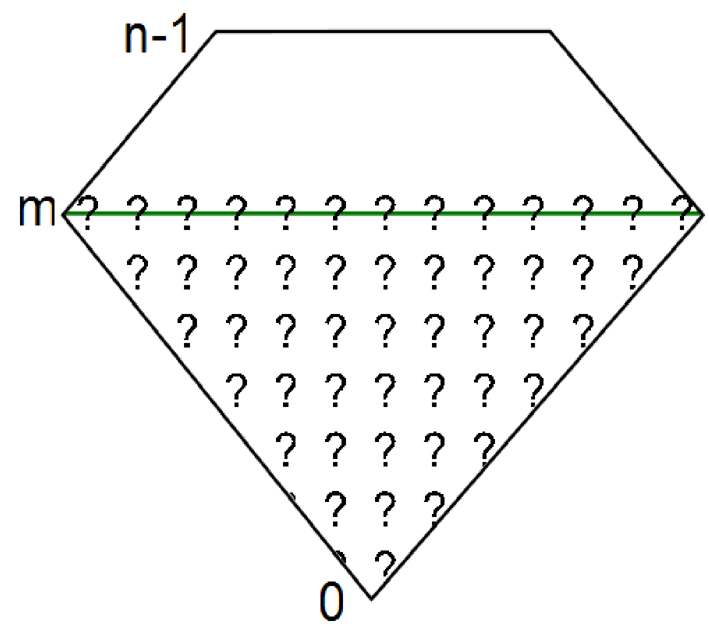
Parameters *m* and *C*

- Number of malicious parties: m
  - $m = 0$ (0-privacy) is when the coalition of parties is empty, but each data recipient can be malicious
  - $m = n$-1 means that no party trusts any other (anonymize-and-aggregate)
- Privacy constraint C:
  - with conditional *BK* (0-privacy), e.g. *k*-anonymity, *l*-diversity
  - with unconditional *BK* ((n-1)-privacy), e.g. differential privacy
  - m-privacy is orthogonal to C and inherits all its advantages and drawbacks

m-Adversary Modeling

- Domain space is exponential!
- If a coalition of attackers cannot breach privacy of records, then any its subcoalition will not be able to do so as well.
- If a coalition of attackers breaches privacy of records, then all its supercoalitions will do that as well.
- m-Privacy monotonicity for providers: up and down.

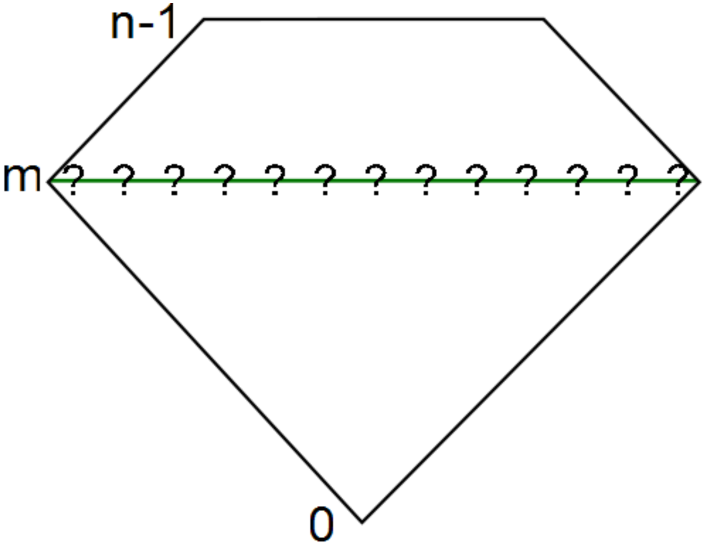




Equivalence Group Monotonicity

Adding new records to a private *T** will not change the privacy fulfillment!

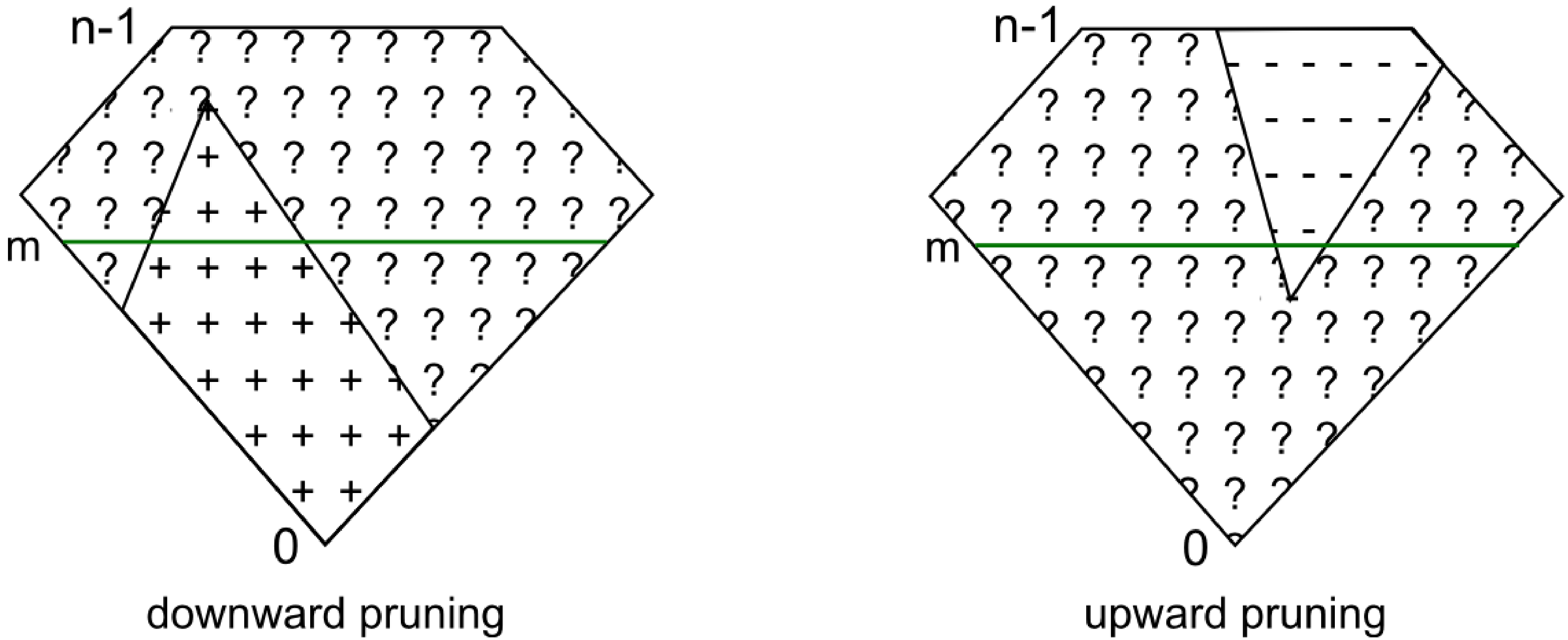
- To verify *m*-privacy it is enough to determine privacy fulfillment <u>only</u> for *m*-adversaries,



- EG monotonic privacy constraints: *k*-anonymity, simple *l*-diversity, …
- Not EG monotonic constraints: entropy *l*-diversity, *t*-closeness, ...

Pruning Strategies

- Number of coalitions to verify: exponential to number of providers, but with efficient pruning strategies!



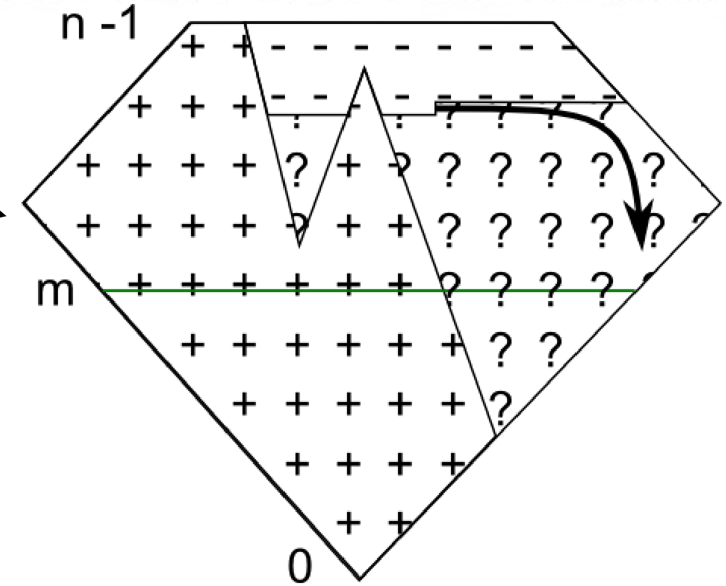
Efficient Pruning - Adaptive Ordering

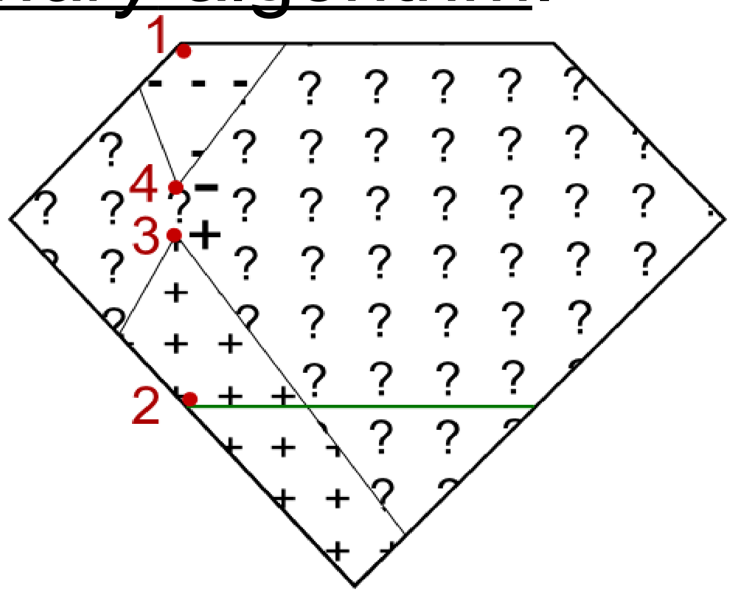
- To speed up verification pruning strategies should be used as early as possible and as frequent as possible.
  - For downward pruning, *m*-adversaries with limited attack power should be checked first.
  - For upward pruning, *m*-adversaries with significant attack power should be checked first.
- Privacy fitness score is a measure of the privacy fulfillment with values greater or equal to 1 only if records are private, i.e. it measures attack power. Example:

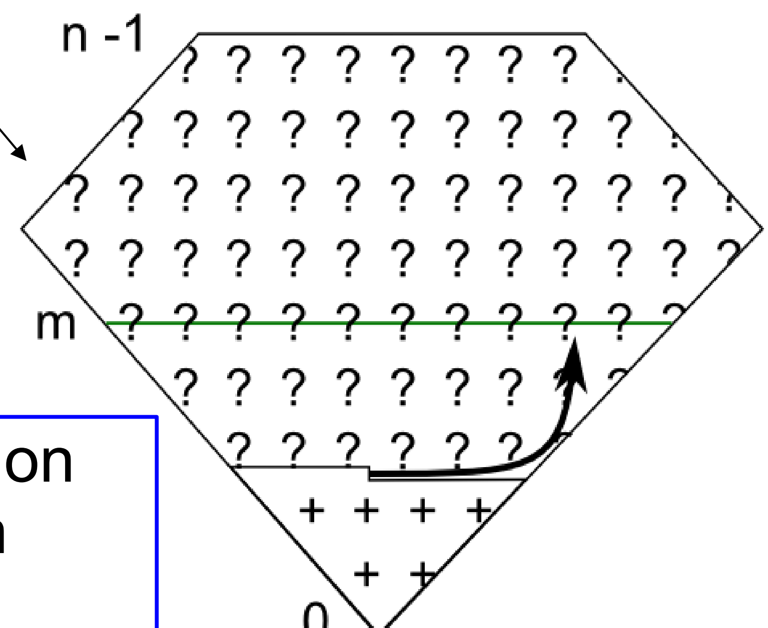
$$score_{F_{C_1 \wedge C_2}}(T^*) = (1-\alpha) \cdot \frac{|T^*|}{k} + \alpha \cdot \frac{|\{t[A_S] : t \in T^*\}|}{l}$$

Verification Algorithms

- <u>top-down algorithm</u>,
- <u>bottom-up algorithm</u>,
- <u>*binary* algorithm</u>.



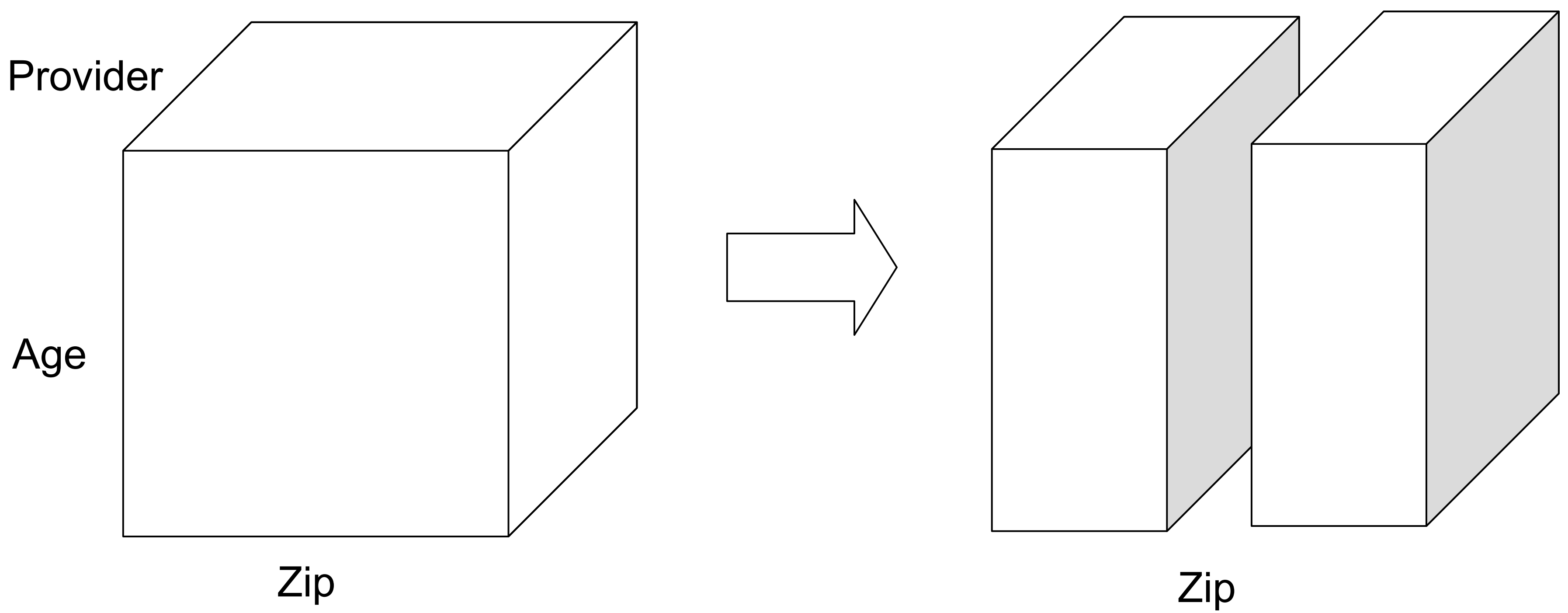




Choosing a single the most efficient verification algorithm is hard, we adaptively (based on privacy fitness scores) select one of them.

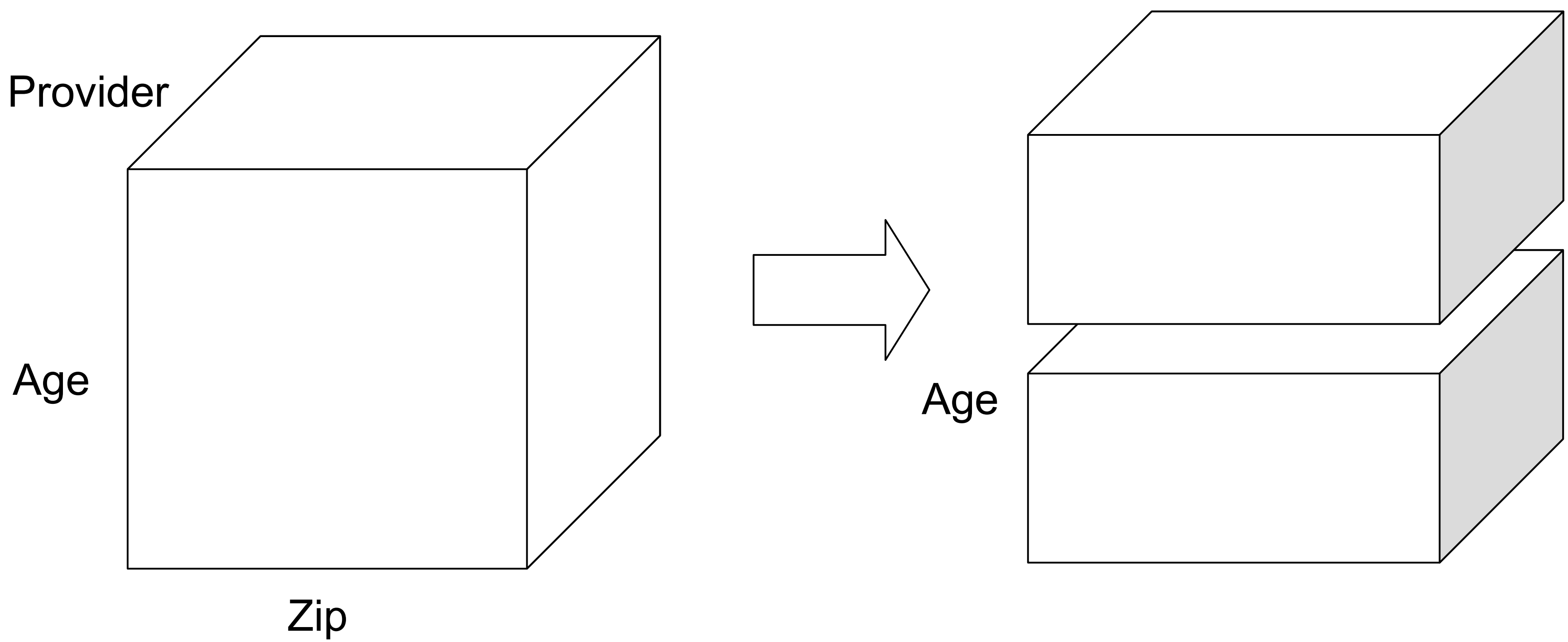
Anonymizer for *m*-Privacy

- We add one more attribute – data provider, which is used as any other attribute in splitting data records.



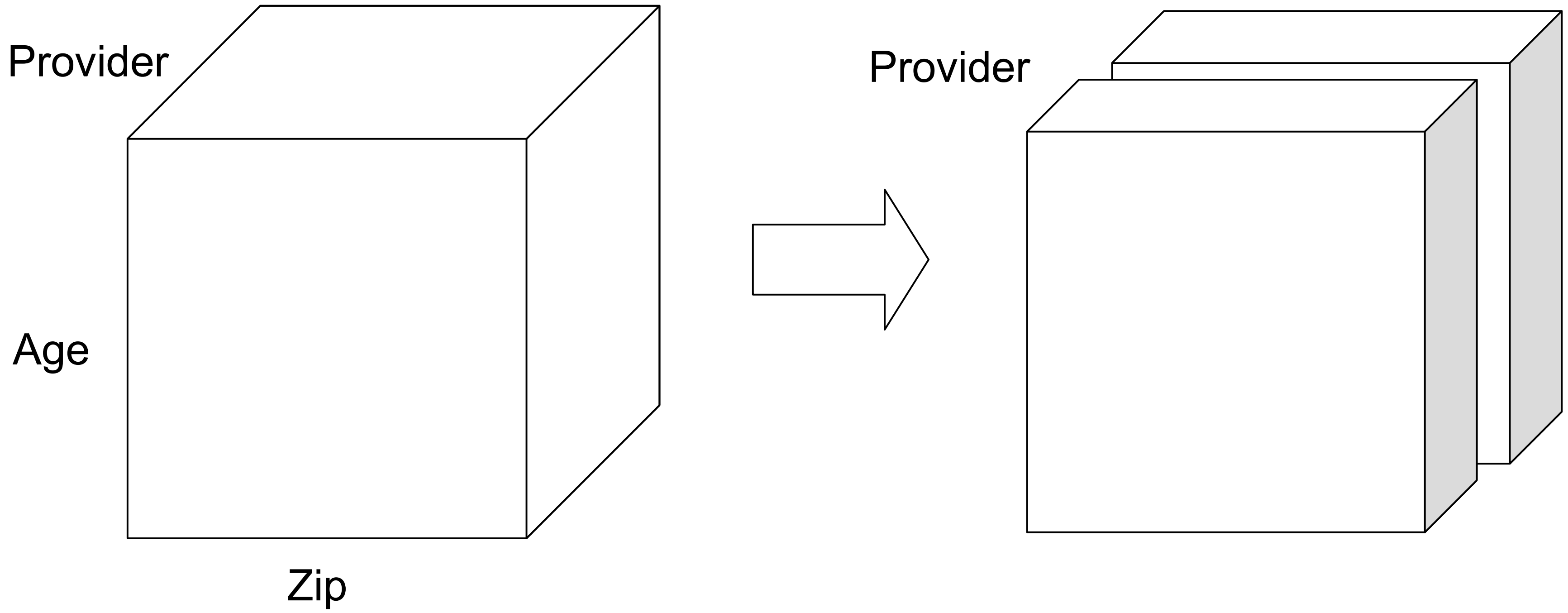
Anonymizer for *m*-Privacy

- We add one more attribute – data provider, which is used as any other attribute in splitting data records.

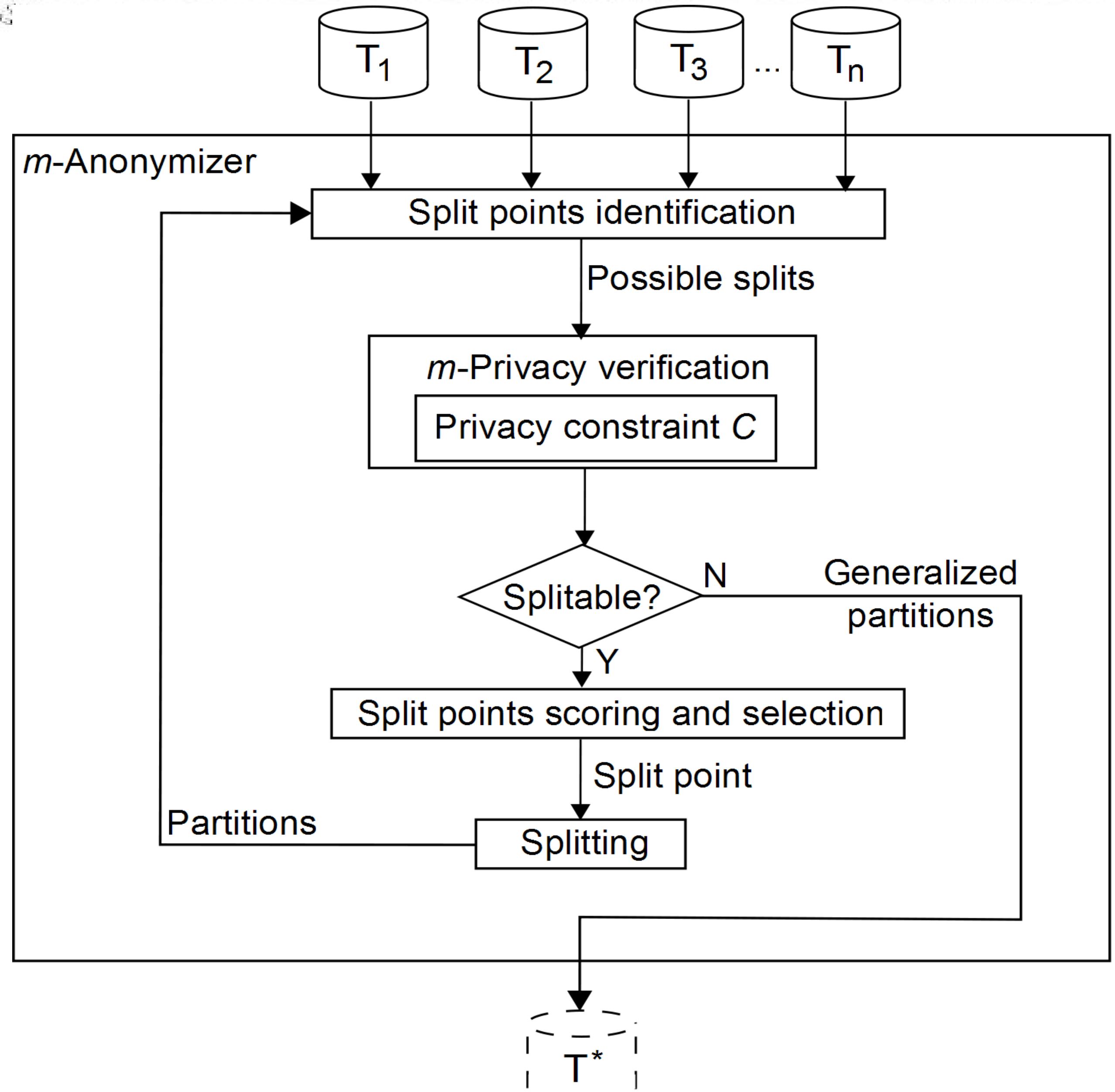


Anonymizer for *m*-Privacy

- We add one more attribute – data provider, which is used as any other attribute in splitting data records.



m-Anonymizer (diagram)

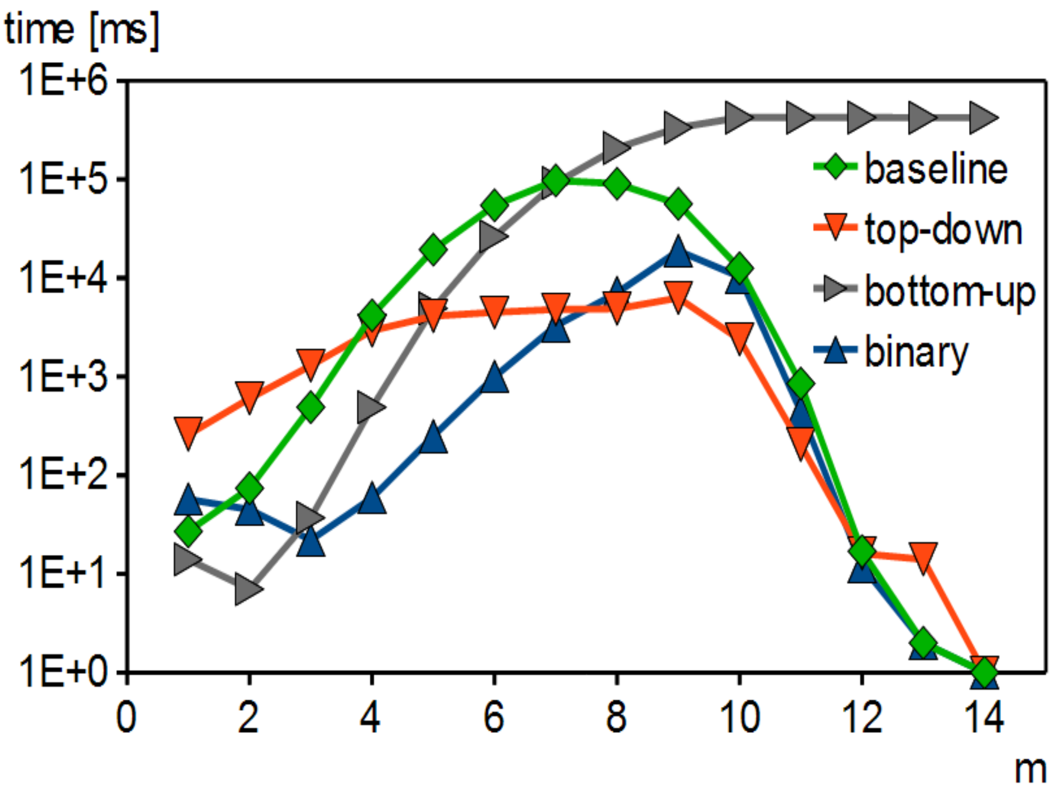


Experiments Setup

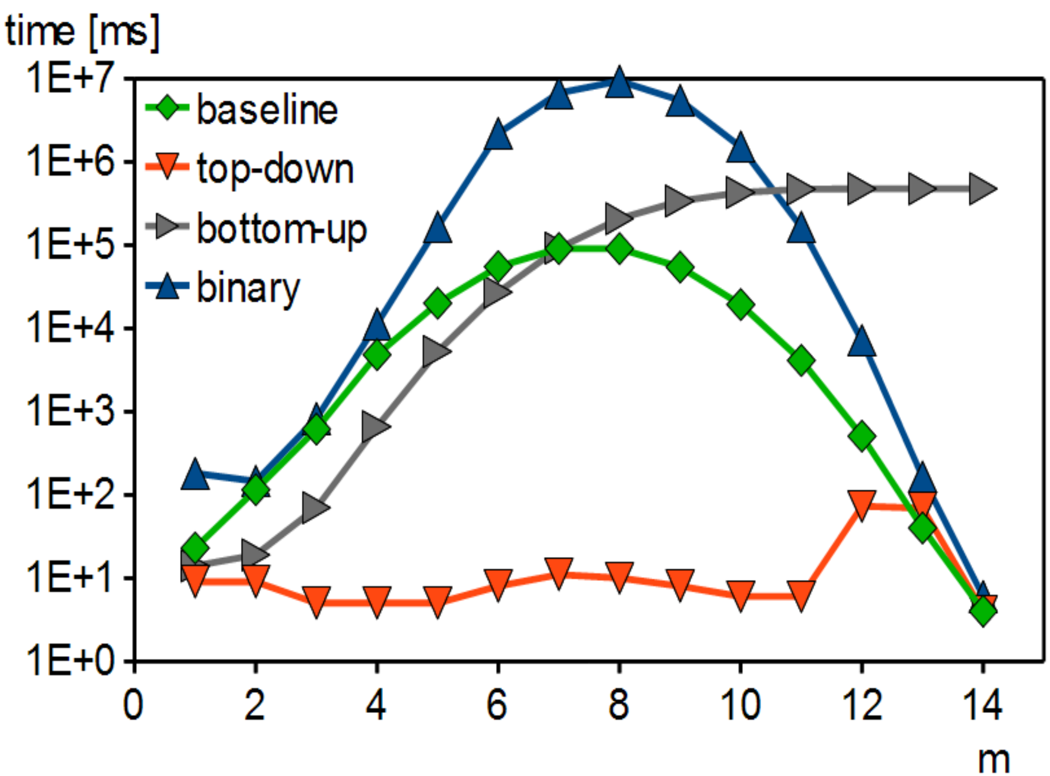
- Dataset: the *Adult* dataset has been prepared using the Census database from 1994.
- Attributes: age, workclass, education, marital-status, race, gender, native-country, <u>occupation</u> (sensitive attribute with 14 possible values).
- Privacy defined as a conjunction of *k*-anonymity and *l*-diversity.
- Metrics:
  - Runtime
  - Query error

Experiments

- *m*-Privacy verification runtime for different algorithms vs *m*



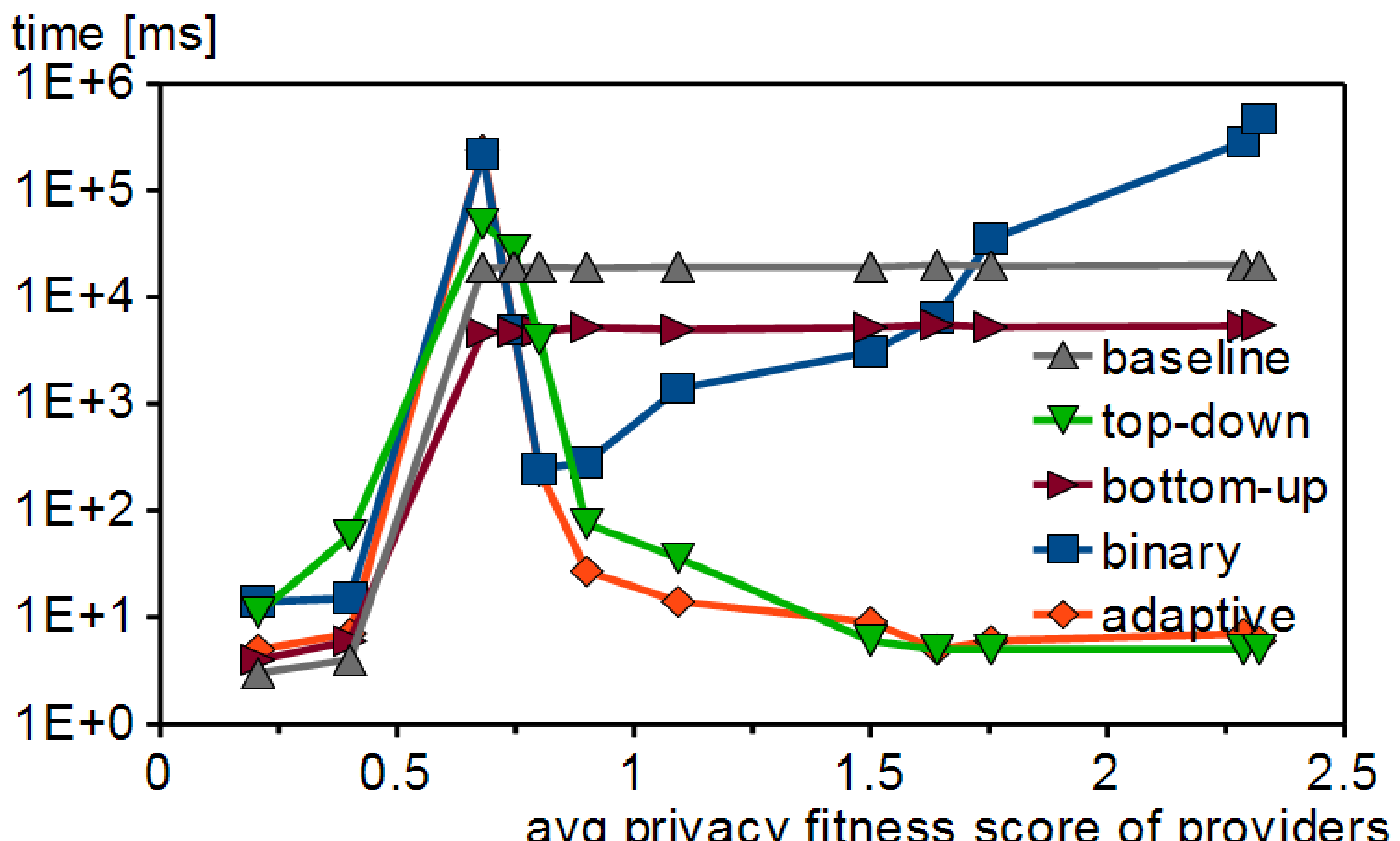
Average privacy fitness score per provider = 0.8



Average privacy fitness score per provider = 2.3

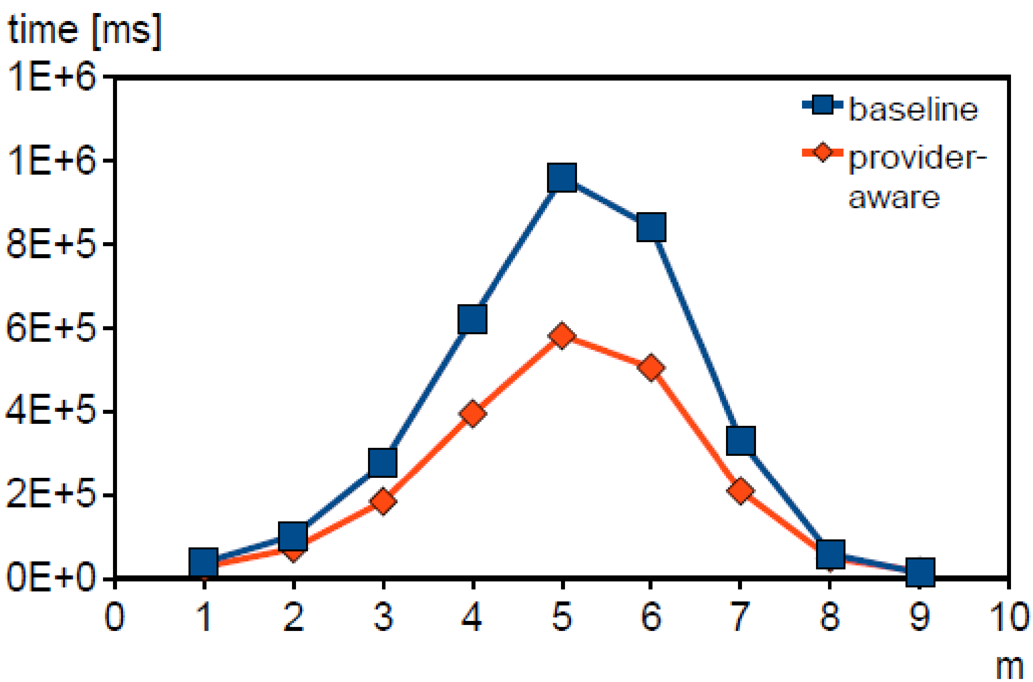
Experiments

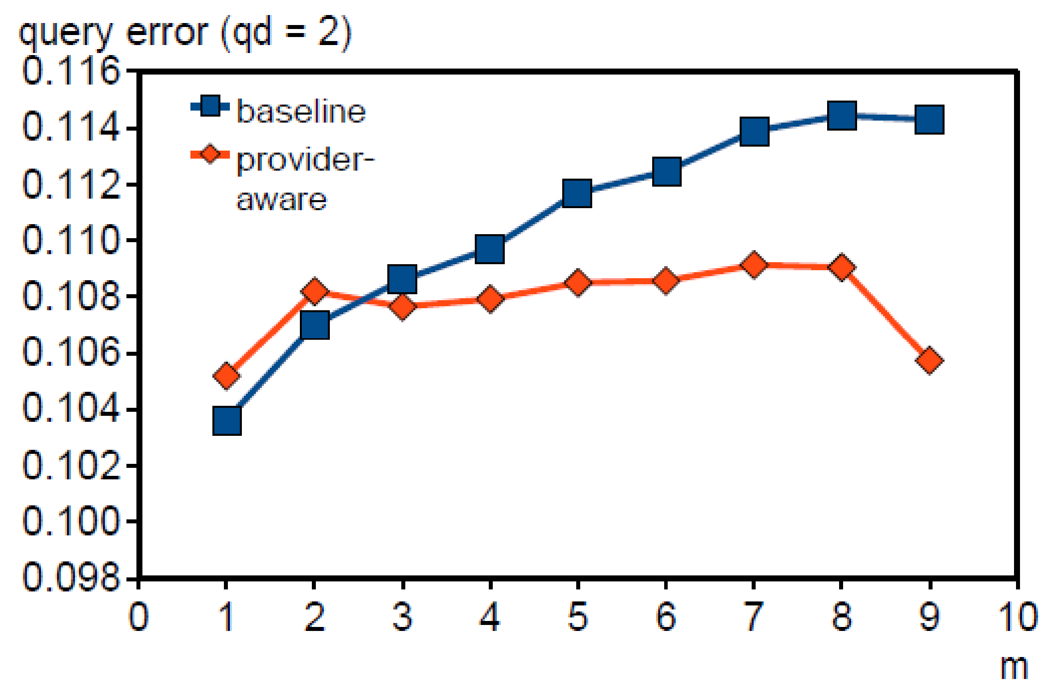
- *m*-Privacy verification runtime for different algorithms vs the average privacy fitness score per provider records (average attack power)



Experiments

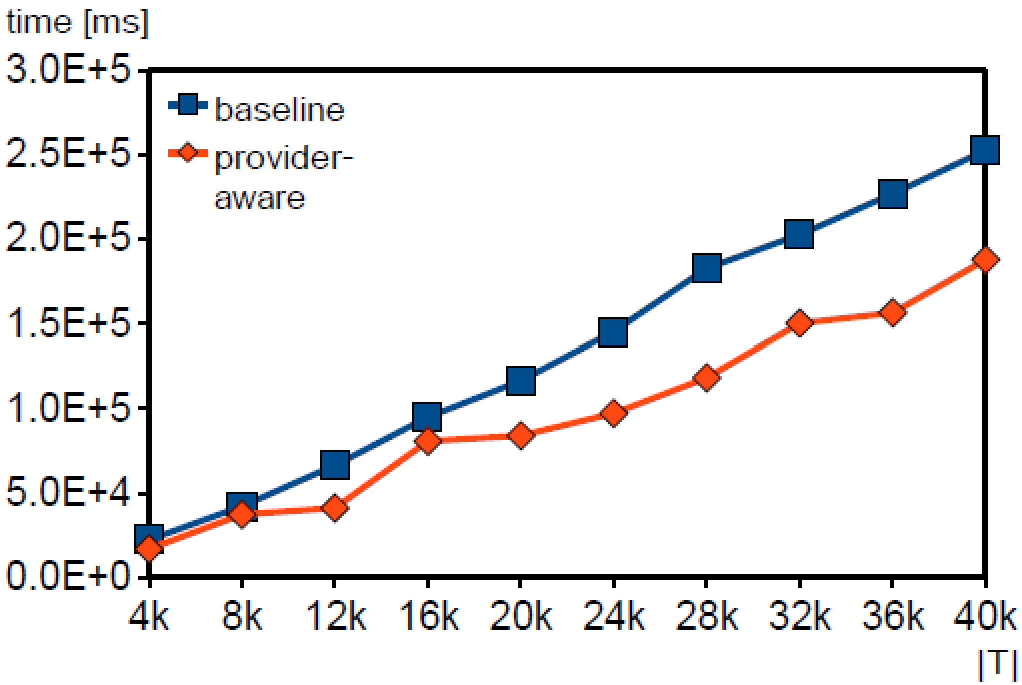
- *m*-Anonymizer runtime and query error for different anonymizers vs size of attacking coalitions *m*

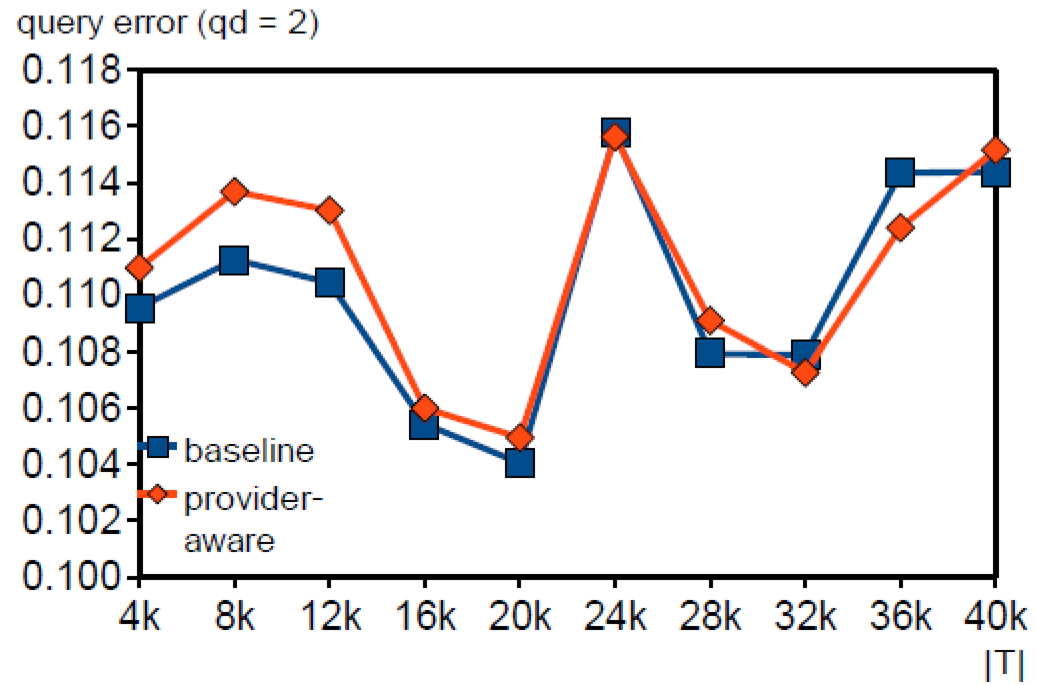




Experiments

- *m*-Anonymizer runtime and query error for different anonymizers vs number of data records





Summary

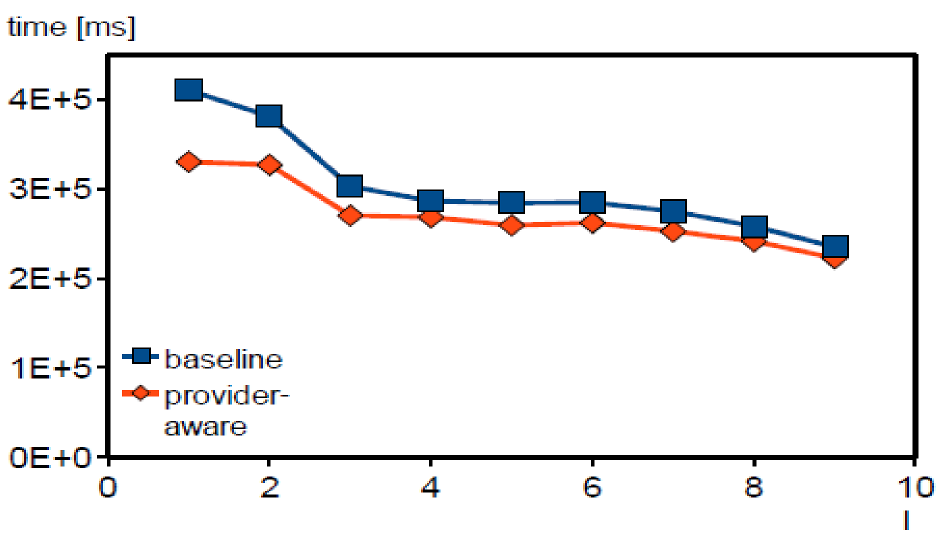
- Identify and model privacy threats for collaborative data provider settings by *m*-privacy,
- Introduce and implement efficient strategies for *m*-privacy verification,
- Propose an *m*-privacy verification algorithm that adapts its strategy to input data,
- Design and implement *m*-anonymizer that anonymizes data with respect to *m*-privacy.

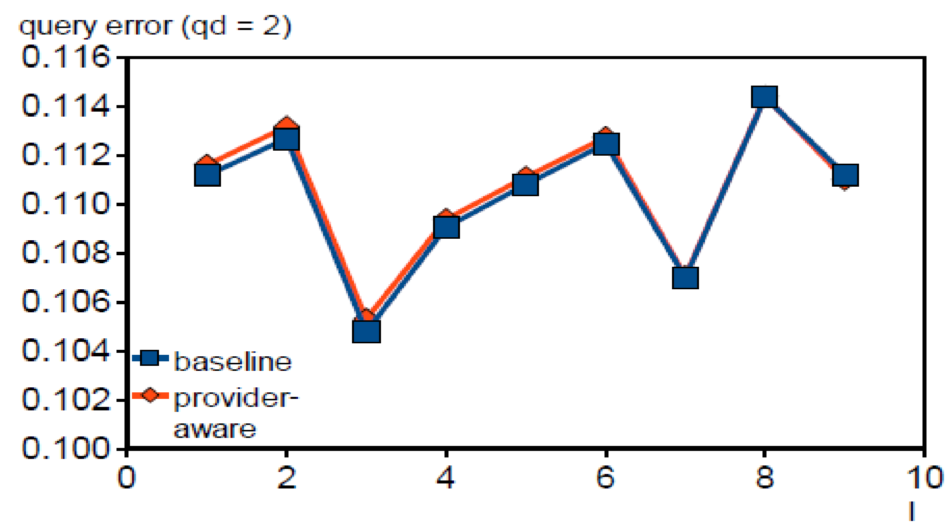
Thank you!

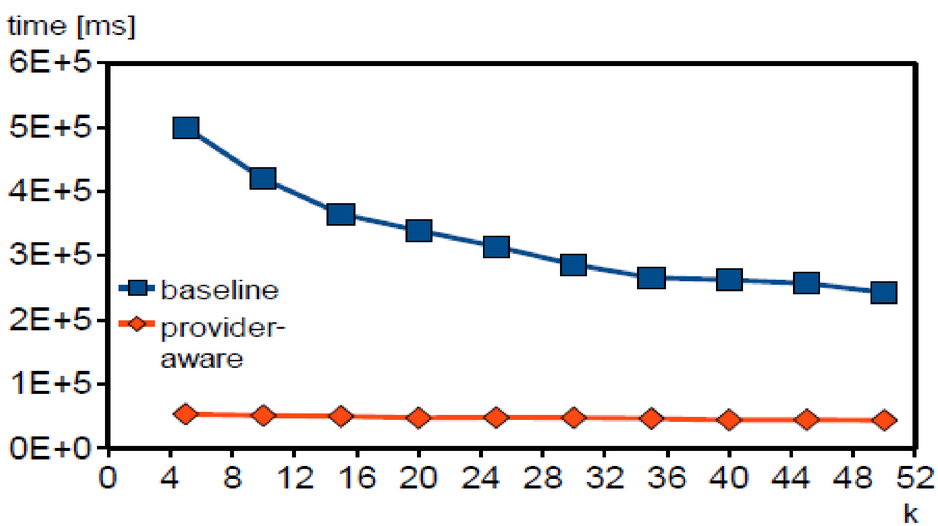
Q & A

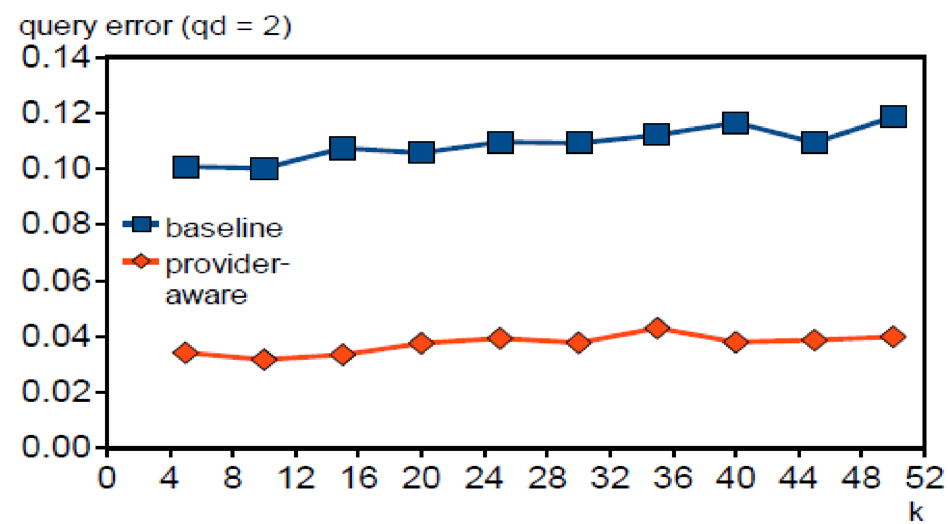
More Experiments

- Experiments for different k and l









Equivalence Group Monotonicity

- A privacy constraint *C* is EG monotonic if and only if any equivalence group of records *T** satisfies *C*, then all its supersets satisfy *C* as well.
- Properties:
  - *m*-Privacy with respect to a constraint *C* is EG monotonic if and only if *C* is EG monotonic,
  - If a constraint *C* is EG monotonic, then the definition of *m*-privacy w.r.t. *C* may be simplified and requires only determining privacy of records only for coalitions of *m* attackers.