Catch a Blowfish Alive: A Demonstration of Policy-Aware Differential Privacy for Interactive Data Exploration

Jiaxiang Liu[1], Karl Knopf[1], Yiqing Tan[1], Bolin Ding[2], Xi He[1]

[1]University of Waterloo [2]Alibaba Group

Overview

Problem: Customized privacy policies can incur performance cost for sensitive dataset over a large domain.
Motivation: Answer queries accurately and efficiently with customized and provable privacy guarantees.
Contribution:

- A new privacy framework, Dynamic Blowfish privacy, that adaptively generates privacy policies at query time.
- A privacy-preserving system, BlowfishDB, that allows data curators to set policy and data analysts to query data.

Research Objective

Data curators wish to release statistics with privacy guarantees; Analysts wish to access sensitive information with accuracy and performance requirements.

	DP	Blowfish	Dynamic Blowfish
Accuracy	☆	☆☆☆	☆☆☆
Usability	☆☆☆	☆	☆☆
Performance	☆☆☆	☆	☆☆

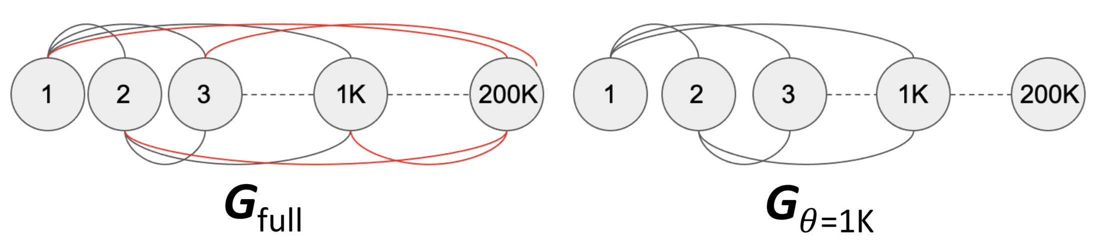
Background

Differential Privacy (DP) [DMN06]: A randomized algorithm $\boldsymbol{M}$ satisfies ϵ-DP if for all possible output sets, and any pair of *neighboring databases* (D_1, D_2) that differ in one record, we have

$$\mathcal{S} \in \mathrm{Range}(\mathcal{M}) : \Pr[\mathcal{M}(D_1) \in \mathcal{S}] \leq e^{\epsilon} \Pr[\mathcal{M}(D_2) \in \mathcal{S}]$$

Policy Graph: $\boldsymbol{G = (V, E)}$ is a secret discriminative graph over domain $\boldsymbol{T}$

- $\boldsymbol{V}$: a set of domain values $\boldsymbol{V = T}$
  e.g. Alice's salary is 10K
- $\boldsymbol{E}$: sensitive pairs of values $\boldsymbol{V}_{pair} \in \boldsymbol{V} \times \boldsymbol{V}$
  e.g. (Alice's salary is 10K, Alice's salary is 20K)



Blowfish Privacy [HMD14]: Let $\boldsymbol{G = (V, E)}$ be a policy graph. An algorithm $\boldsymbol{M}$ satisfies $(\epsilon, \boldsymbol{G})$-Blowfish Privacy if

$$\mathcal{S} \in \mathrm{Range}(\mathcal{M}) : \Pr[\mathcal{M}(D_1) \in \mathcal{S}] \leq e^{\epsilon} \Pr[\mathcal{M}(D_2) \in \mathcal{S}]$$

for neighboring databases (D_1, D_2) that differ in one record, such that $(u, v) \in \boldsymbol{E}$ where u is the record value in D_1 *and* v is the corresponding record value in D_2.
(α, β)-Accuracy: The distance between the noisy output and true query answer is no more than α with a high probability $(1\text{-}\beta)$.

Dynamic Blowfish Privacy

Matrix Representation of Query: Linear counting queries can be transformed into matrices:

- **x**: a histogram over the full domain $\boldsymbol{T}$
  e.g. the count over each possible salary in {1,…,200K}
- **W**: linear combinations of counts in **x**
  e.g. the workload of range [1, 100K) and [1, 200K) is

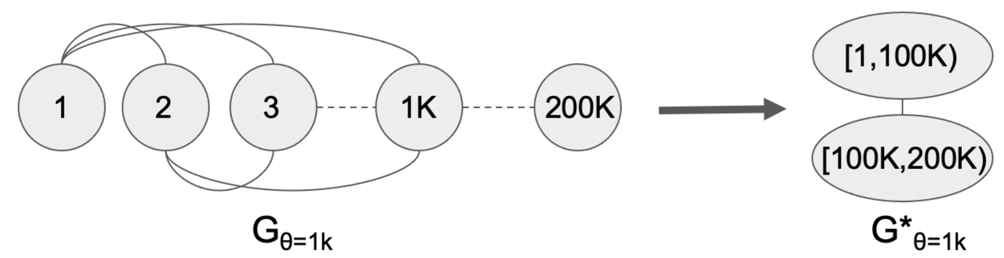
$$W = \begin{pmatrix} 1\cdots1 & 0\cdots0 \\ 1\cdots1 & 1\cdots1 \end{pmatrix}$$

Dynamic Blowfish Policy Projection: Partition vertices in $\boldsymbol{G}$ based on the partition matrix **T** generated at query time.

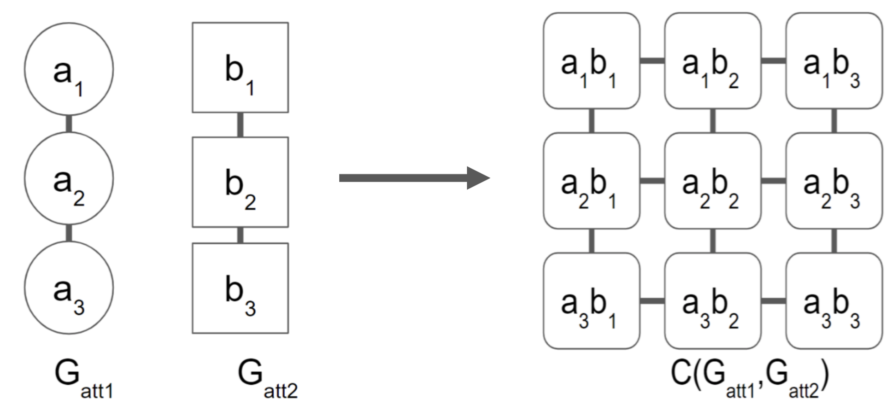
- **x***: the corresponding data vector partitioned by **T**
  e.g. {1,…,200K} into {[1, 100K), [100K, 200K)}
- **W***: the corresponding workload partitioned by **T**

$$T = \begin{pmatrix} 1\cdots1 & 0\cdots0 \\ 0\cdots0 & 1\cdots1 \end{pmatrix} \longrightarrow W^* = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}$$

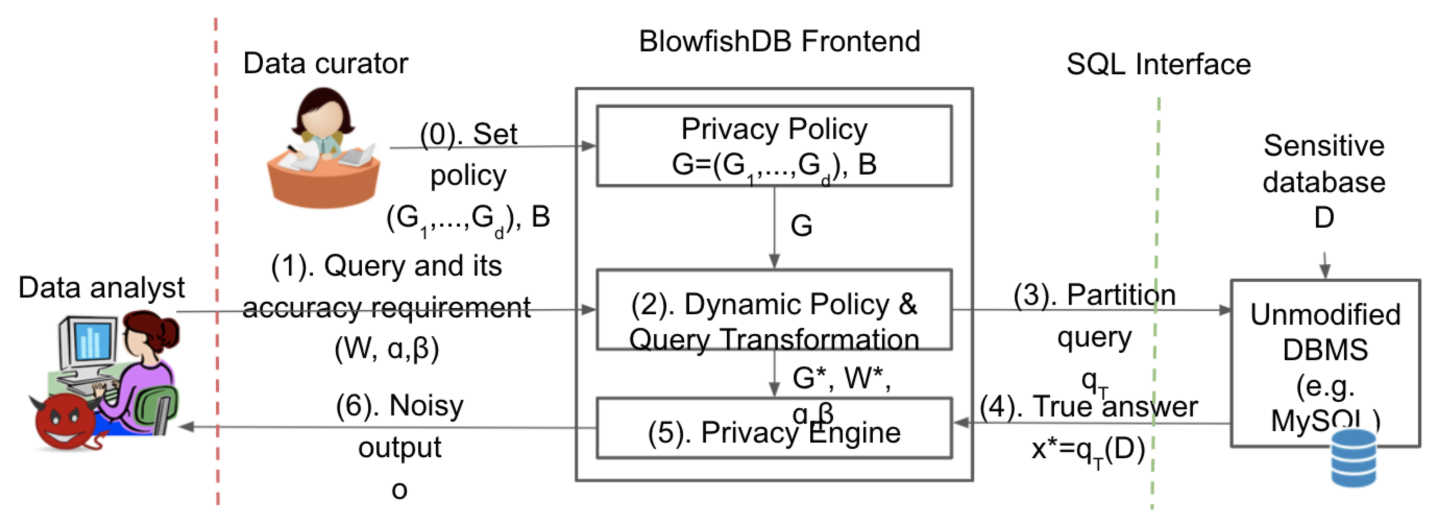
- **G***: The policy graph $\boldsymbol{G}$ projected onto the partitioned domain



Privacy and Accuracy: Let $\boldsymbol{M}$ be a matrix mechanism that answers **W** under $(\epsilon, \boldsymbol{G})$-Blowfish Privacy, applying an optimal ϵ-DP mechanism $\boldsymbol{M}^*$ that answers **W*** also satisfies $(\epsilon, \boldsymbol{G})$-Blowfish Privacy on **W** with the same error.
Attribute Policy Composition: We create higher dimensional privacy policies by composing single attribute privacy definitions.



BlowfishDB Overview



- Privacy policy exploration from data curator
- High-level language with accuracy requirements from data analysts
- Query-dependent partition matrix **T** for efficient query processing
- Overall privacy loss is bounded by privacy set by the data curator

Evaluation

Dataset: Adult (Income) dataset
Query: Histogram/Cumulative histogram with varying granularities G5/G10
Privacy Policy: DP, (Dynamic) Blowfish Privacy (distance threshold policies with varying thresholds, T1, T5, and T10)

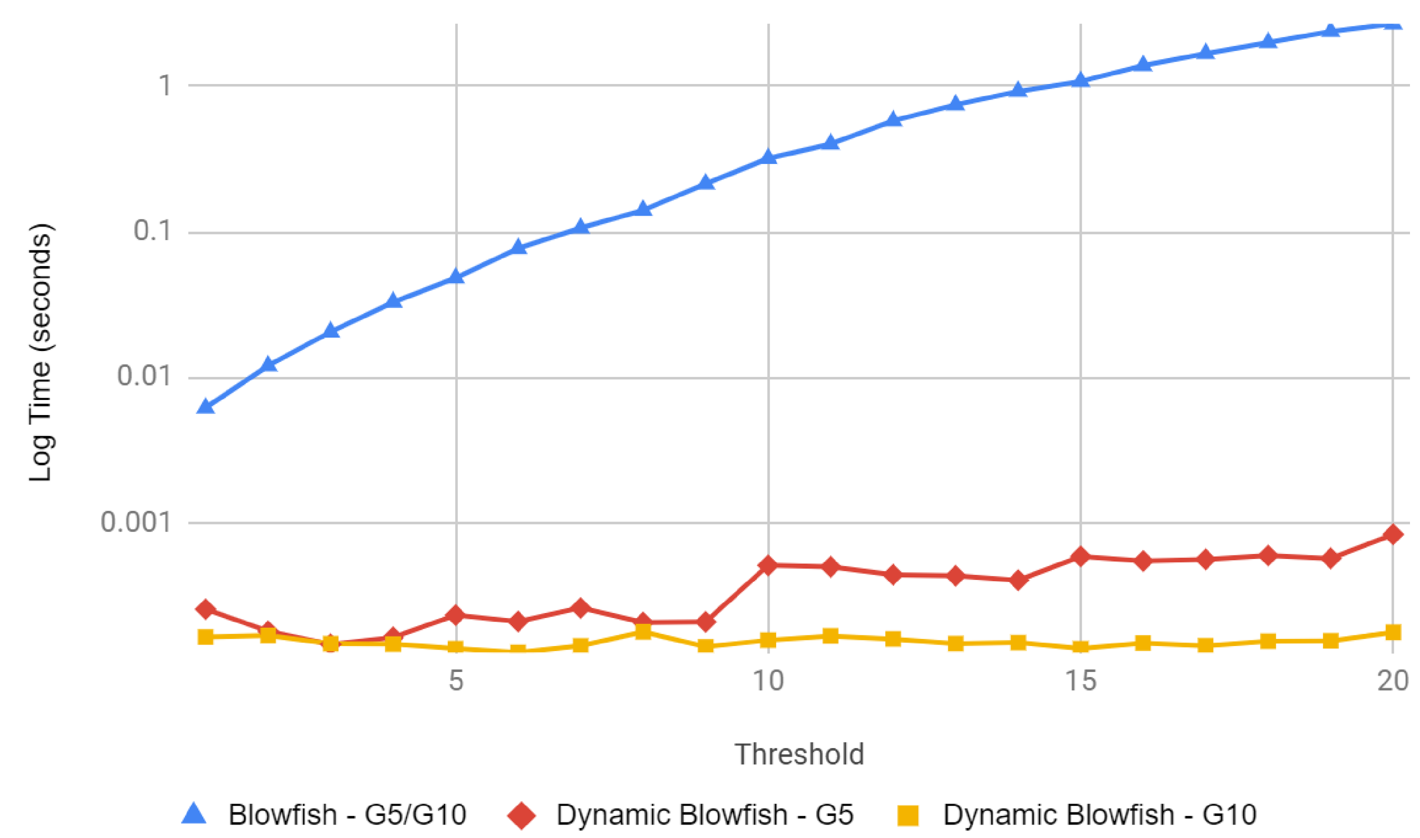


Figure 1: **Performance vs Privacy Guarantee**. For privacy guarantee characterized by each threshold, runtime performance is reported for significant thresholds.

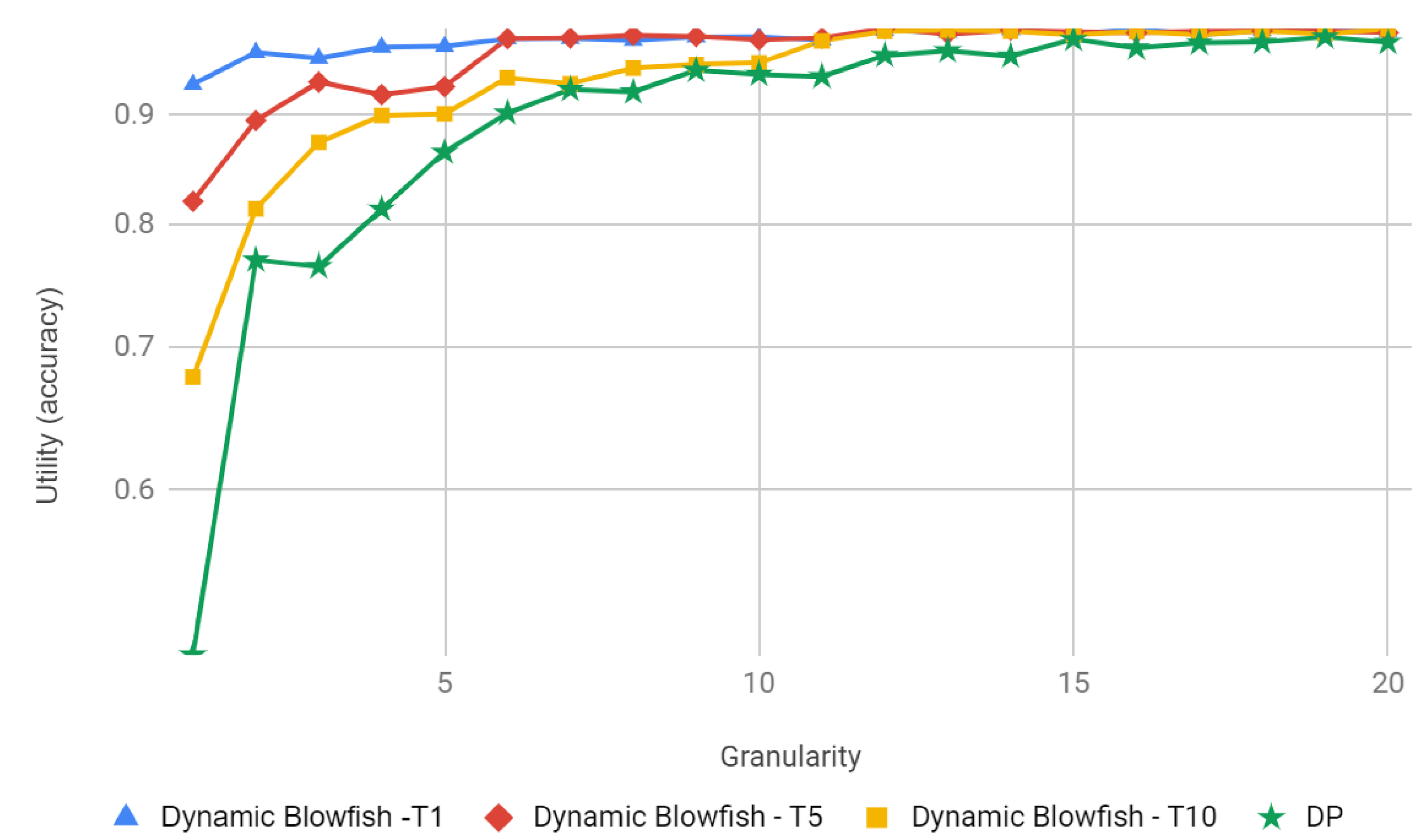


Figure 2: **Utility vs Query Granularity (Cumulative Histogram)**. For partitioning matrix **T** characterized by each granularity, accuracy is reported for significant thresholds.

Conclusions

Conclusion:

- Dynamic Blowfish achieves better utility versus DP and better performance than Blowfish.
- BlowfishDB provides a way for data curators to set better policies and data analysts to retrieve more accurate results.

Future work:

- Incorporate more complex privacy policies and queries.