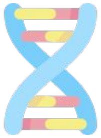
Disseminating string data ensuring privacy: new combinatorial models and algorithms

Giulia Bernardini, Università di Trieste

String data is versatile

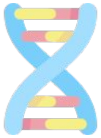


Letters = nucleobases

Letters = search query terms

Letters = locations

Why data dissemination?

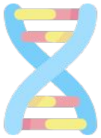


DNA sequence analysis

Product recommendation

Location-based service provision

Leakage of confidential information



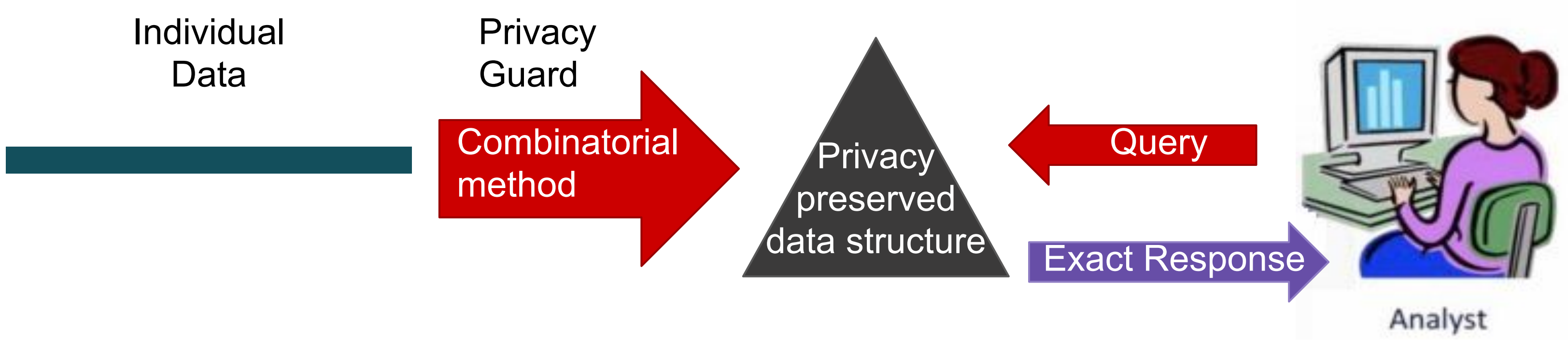
Genetic diseases

Political beliefs or sexual orientation

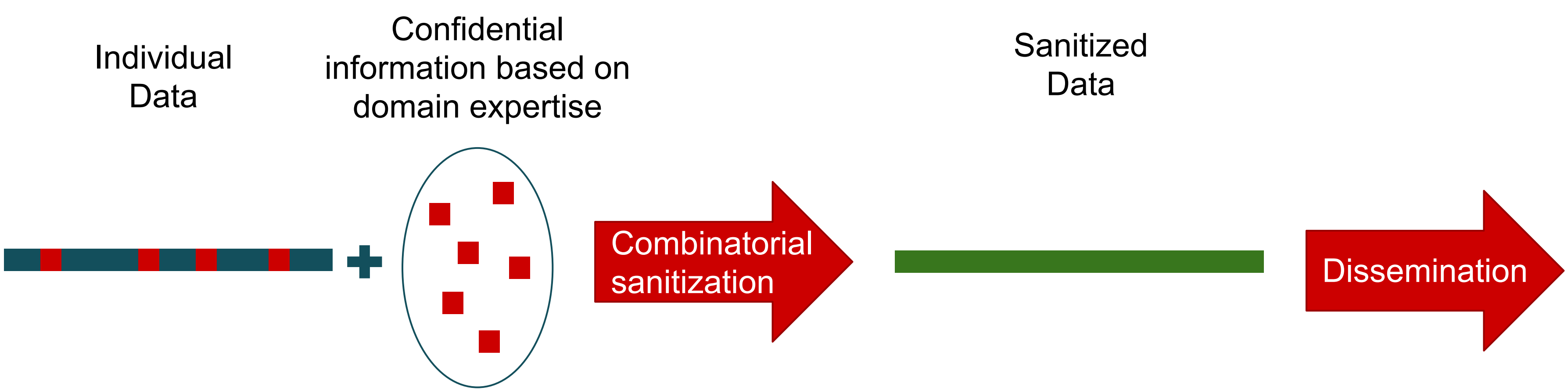
Trips to mental health clinics

Our models

Reverse-safe text indexing



Combinatorial string sanitization



Comparison with differential privacy

DP:

Aggregated Data

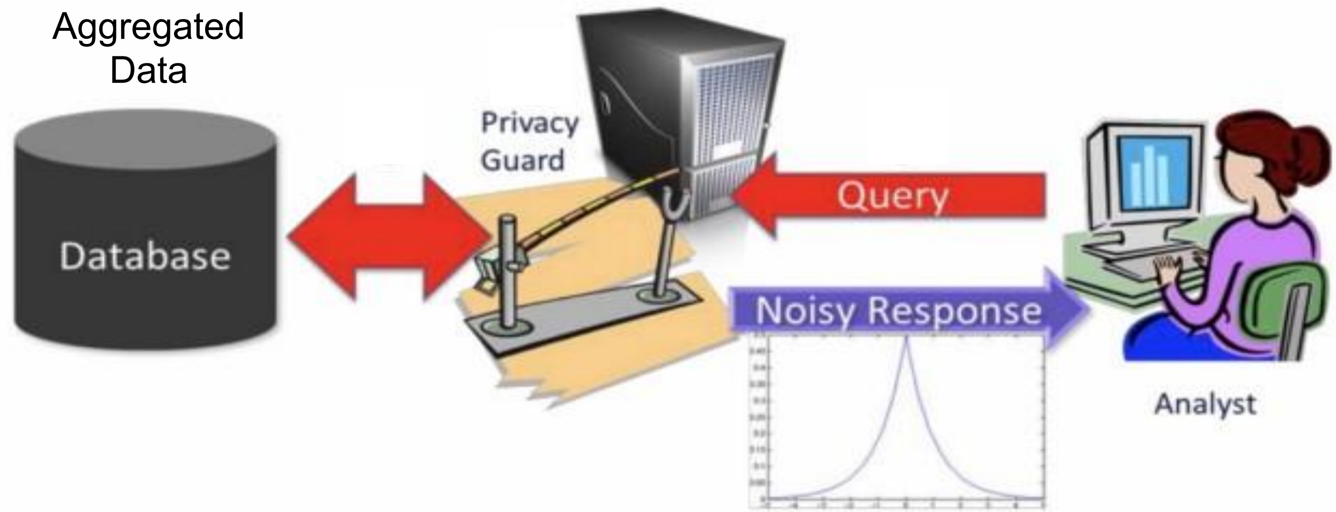
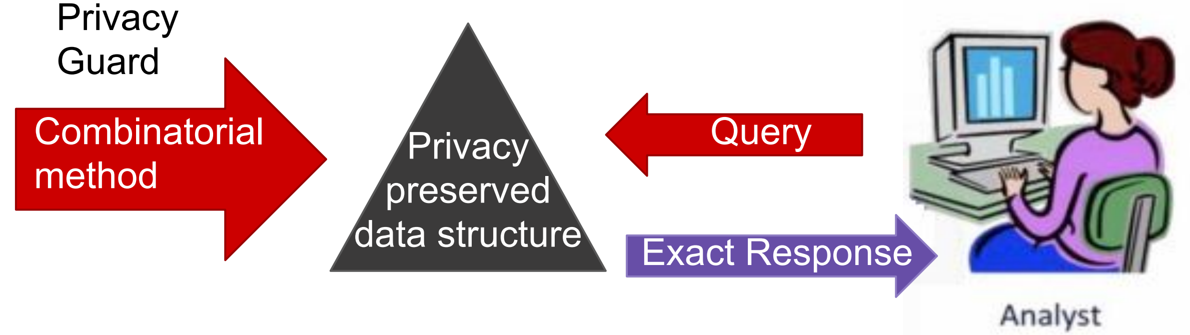


Image from Microsoft's "Differential Privacy for Everyone"

US:

Individual Data

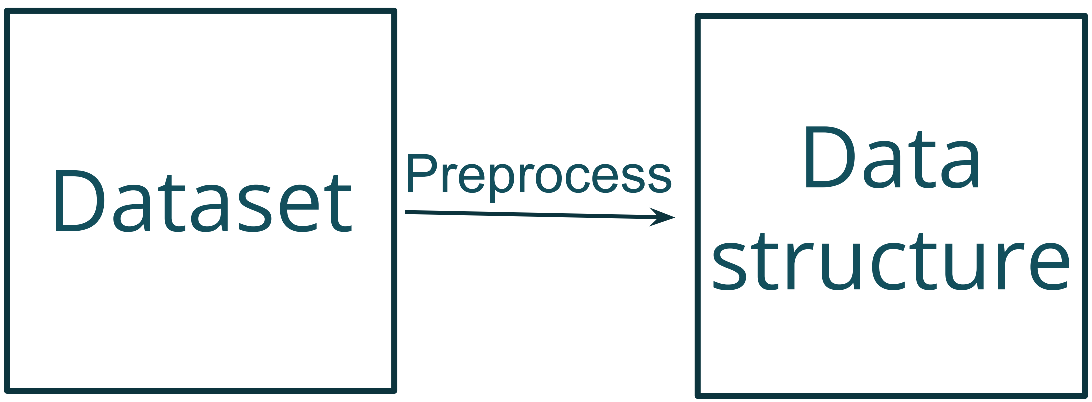
Privacy Guard



Main research question

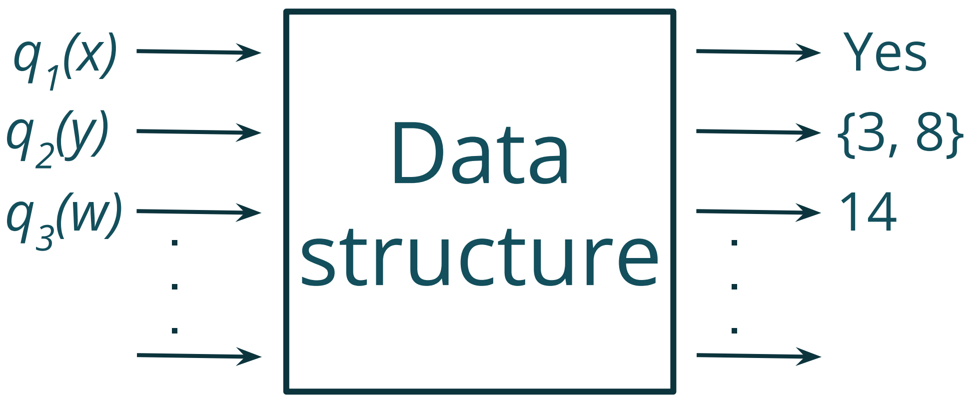
Can we provide provable **trade-offs** between **privacy** and **data utility** for individual data dissemination?

How do we view data structures?



Reverse-safe text indexing. G.B., H.Chen, G.Fici, G.Loukides, S.Pissis. ACM JEA, 2021

How do we view data structures?

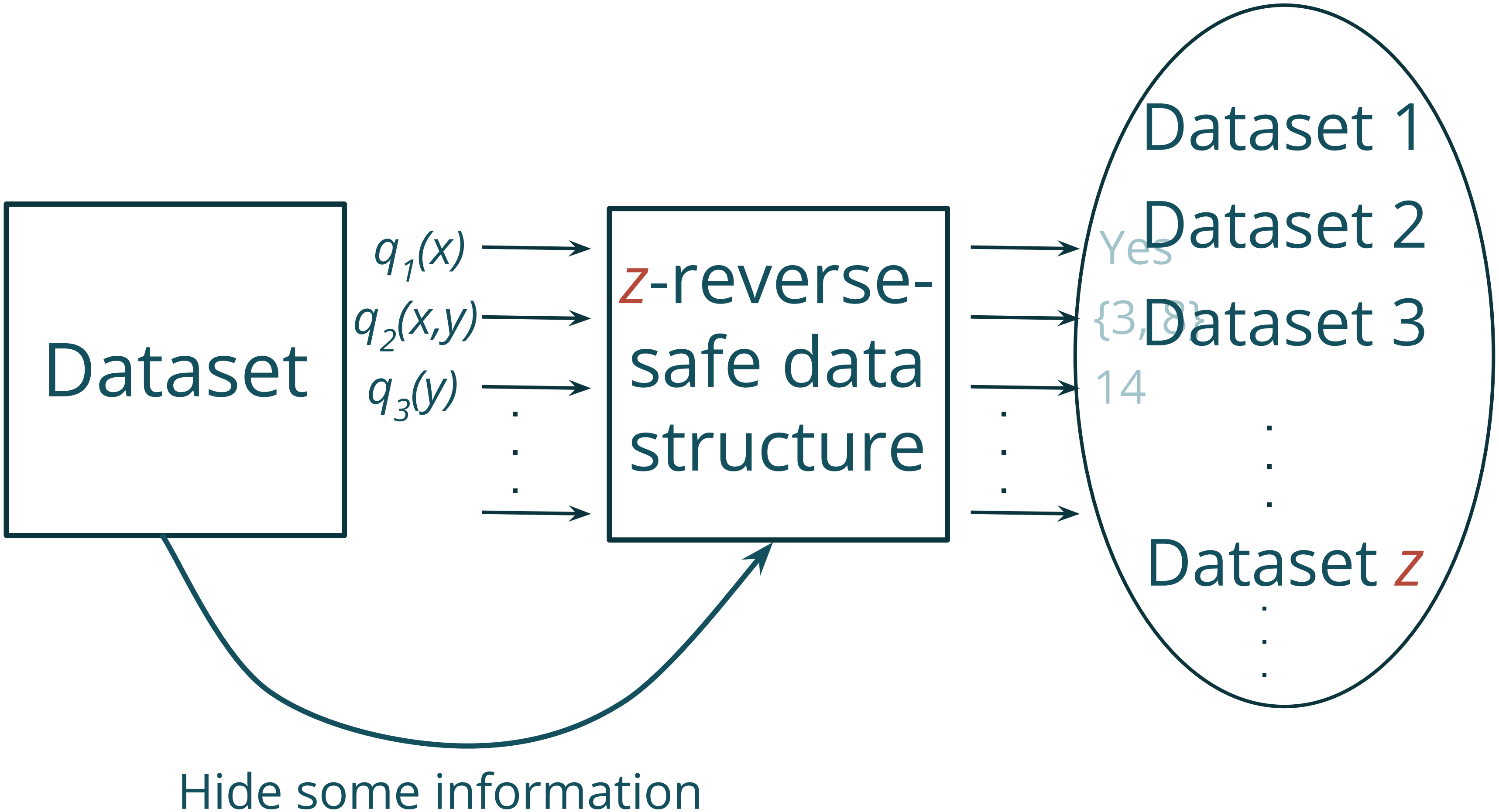


Reverse-safe text indexing. G.B., H.Chen, G.Fici, G.Loukides, S.Pissis. ACM JEA, 2021

How do we view data structures?

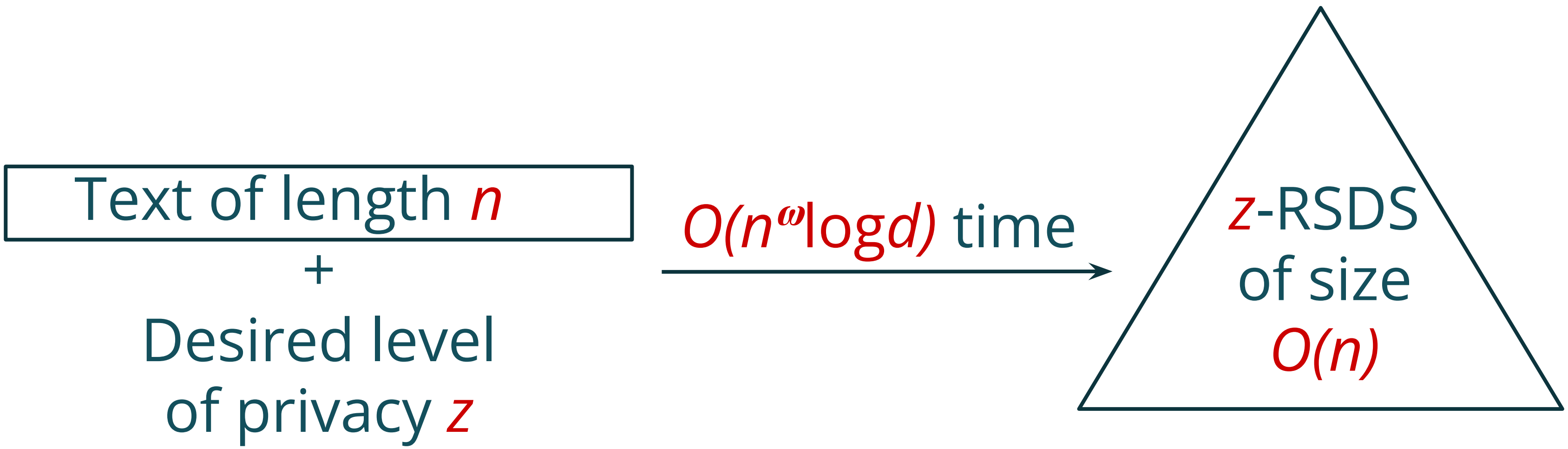
Reverse-safe text indexing. G.B., H.Chen, G.Fici, G.Loukides, S.Pissis. ACM JEA, 2021

z-reverse-safe data structures



Reverse-safe text indexing. G.B., H.Chen, G.Fici, G.Loukides, S.Pissis. ACM JEA, 2021

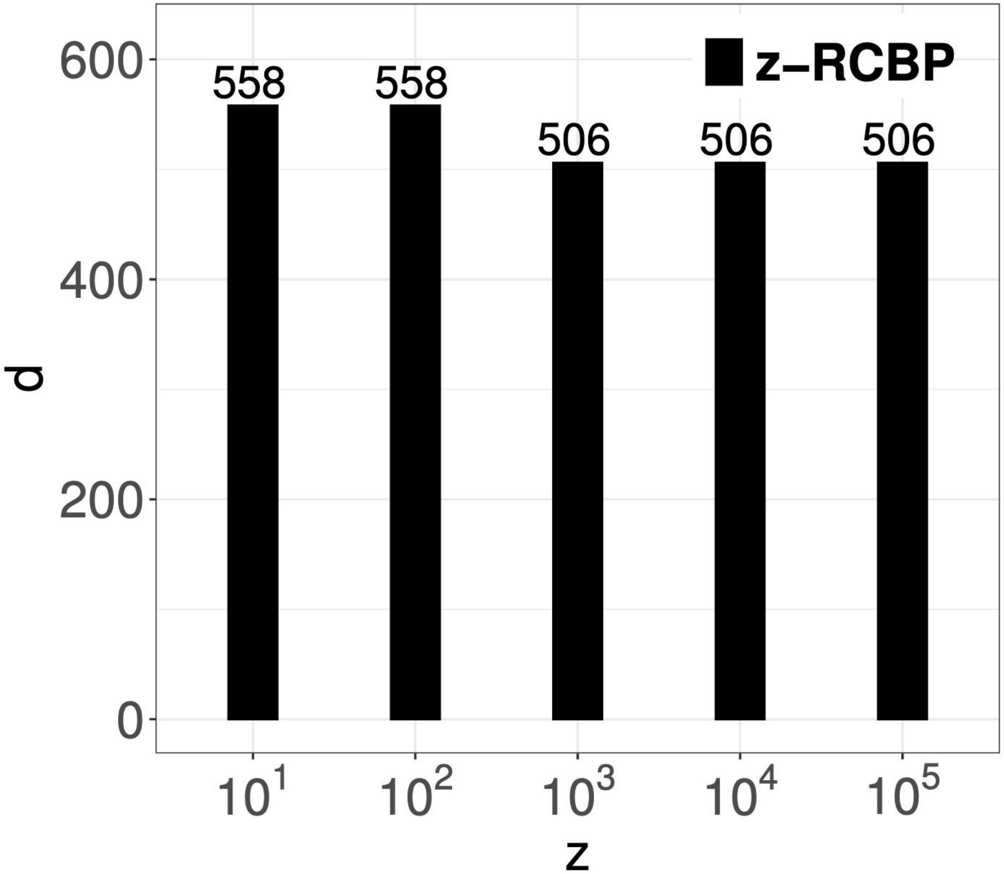
z-RSDS for text indexing: main result



Answering pattern matching queries of length $m \leq d$ in $O(m)$ time, where d is maximal for the input z.

Reverse-safe text indexing. G.B., H.Chen, G.Fici, G.Loukides, S.Pissis. ACM JEA, 2021

Experiments: utility



MSN dataset: page categories visited in 24h

n>4.6 million, alphabet size (categories) 17

Experiments: runtime

Dataset	*z*-RCB	*z*-RCE	*z*-RC	*z*-RCBP
MSN	438.49	421.96	659.17	**347.34**
EC	364.84	725.26	571.8	**339.18**
KAS	710.55	1022.59	2555.8	**649.09**

Runtime (in seconds) for different implementations of the algorithm and different datasets.

EC: genomic data, n>4.6 millions, alphabet size 4

KAS: e-commerce data, n>15.8 millions, alphabet 94

Combinatorial String Dissemination

INPUT:

- A *string W* of data to be disseminated
- A set of *constraints* to capture privacy
- A set of *properties* to capture data utility

OUTPUT: A string *X* satisfying the properties subject to the constraints

The Minimal String Length setting

Constraints: for $k > 0$, no given length-k *sensitive pattern,* modelling confidential knowledge, occurs in X

Properties: the order of all the other length-k patterns is the same in W and in X;

Goal: produce the **shortest** string X that satisfies the properties subject to the constraints

Combinatorial algorithms for string sanitization. G.B., H.Chen, A.Conte, R.Grossi, G.Loukides, N. Pisanti, S.Pissis, G.Rosone, M.Sweering. ACM TKDD, 2021

The MSL setting: an example

W = aabaaaababbbaab

k=4; sensitive patterns={aaaa, abab, abbb}

A solution:

aabaaa#aaaba#babb#bbbaab

The **shortest** solution:

aabaaaba#babb#bbbaab

The MSL setting: an example

W = aabaaaababbbbaab

k=4; sensitive patterns={aaaa, abab, abbb}

A solution:

aab**aaa**#**aaa**ba#babb#bbbaab

The **shortest** solution:

aab**aaa**ba#babb#bbbaab

The MSL setting: main result

We are able to solve the problem $O(k|W|)$ time, which is worst-case optimal. An $O(|W|)$-sized representation of X can be built in $O(|W|)$ time.

Combinatorial algorithms for string sanitization. G.B., H.Chen, A.Conte, R.Grossi, G.Loukides, N. Pisanti, S.Pissis, G.Rosone, M.Sweering. ACM TKDD, 2021

The Minimal Edit Distance setting

Constraints: for *k > 0*, no length-*k sensitive pattern,* modelling confidential knowledge, occurs in *X*

Properties: the order of all the other length-*k* patterns is the same in *W* and in *X*

Goal: a string *X* that satisfies the properties subject to the constraints and is at **minimum edit distance** from *W*

Combinatorial algorithms for string sanitization. G.B., H.Chen, A.Conte, R.Grossi, G.Loukides, N. Pisanti, S.Pissis, G.Rosone, M.Sweering. ACM TKDD, 2021

The MED setting: an example

W = babaaaaabbbab

k=3; sensitive patterns={aba, baa, aaa, aab, bba}

The **shortest** solution:

babbb#bab

A solution at minimum edit distance from *W*:

bab#aa#abbb#bab

The MED setting: an example

W = babaaaaabbbab

k=3; sensitive patterns={aba, baa, aaa, aab, bba}

A **shortest** solution ($d(W,X)$=6):

babbb#bab

A solution at minimum edit distance from W:

bab#aa#abbb#bab

($d(W,X)$=4)

The MED setting: main result

The problem can be solved in $O(k|W|^2)$ time, and it cannot be solved in $O(|W|^{2-\delta})$ time, for any $\delta > 0$, unless the strong exponential time hypothesis is false.

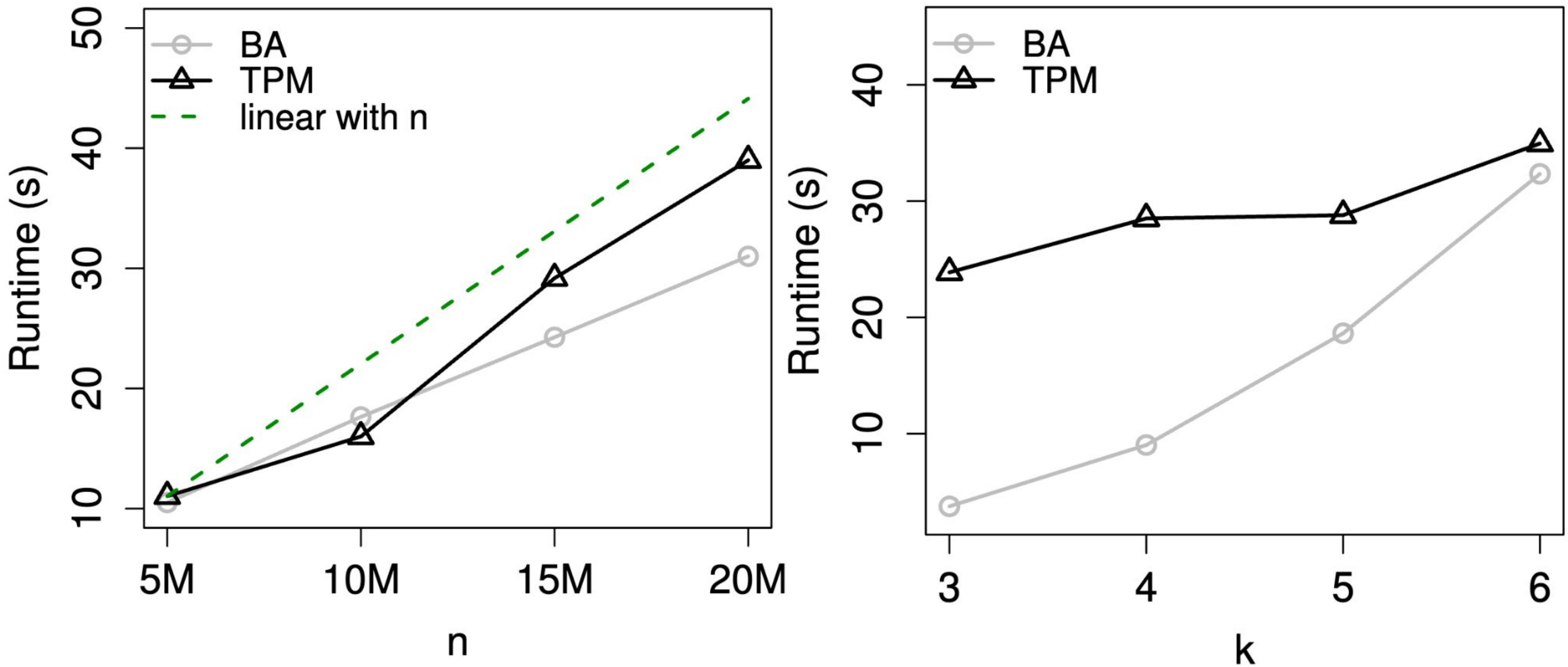
String sanitization under edit distance. G.B., H.Chen, A.Conte, R.Grossi, G.Loukides, N. Pisanti, S.Pissis, G.Rosone, M.Sweering. CPM, 2020

The MED setting: main result

The problem can be solved in $O(\log^2 k |W|^2)$ time, and it cannot be solved in $O(|W|^{2-\delta})$ time, for any $\delta > 0$, unless the strong exponential time hypothesis is false.

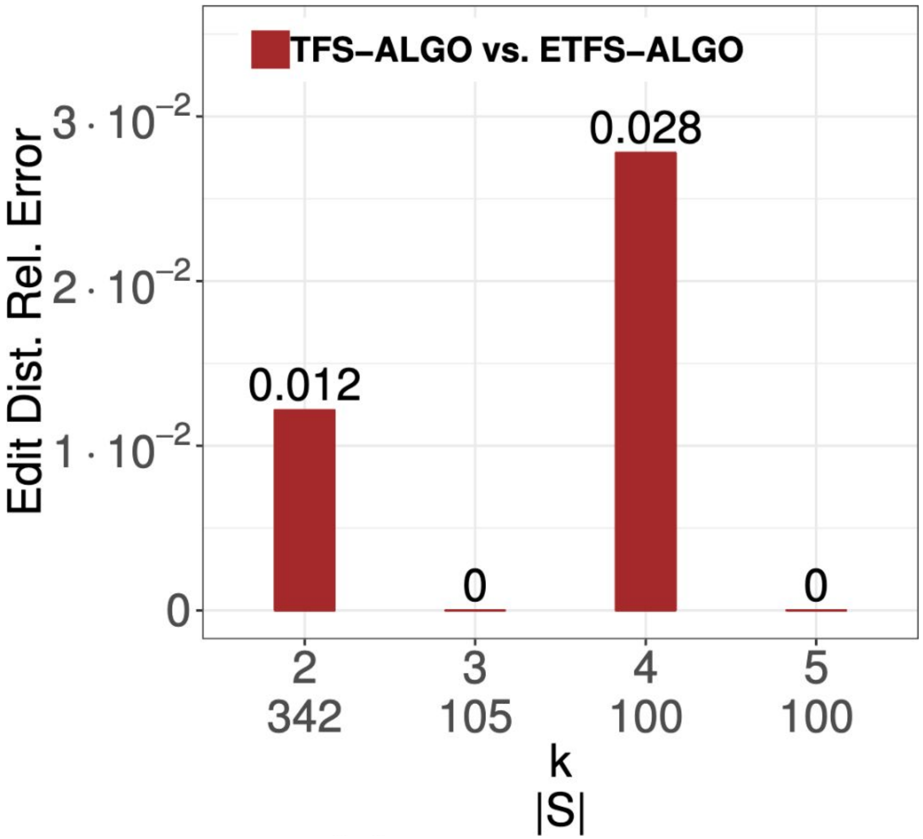
String sanitization under edit distance: improved and generalized. T.Mieno, S.P.Pissis, L.Stougie, M.Sweering. CPM, 2021

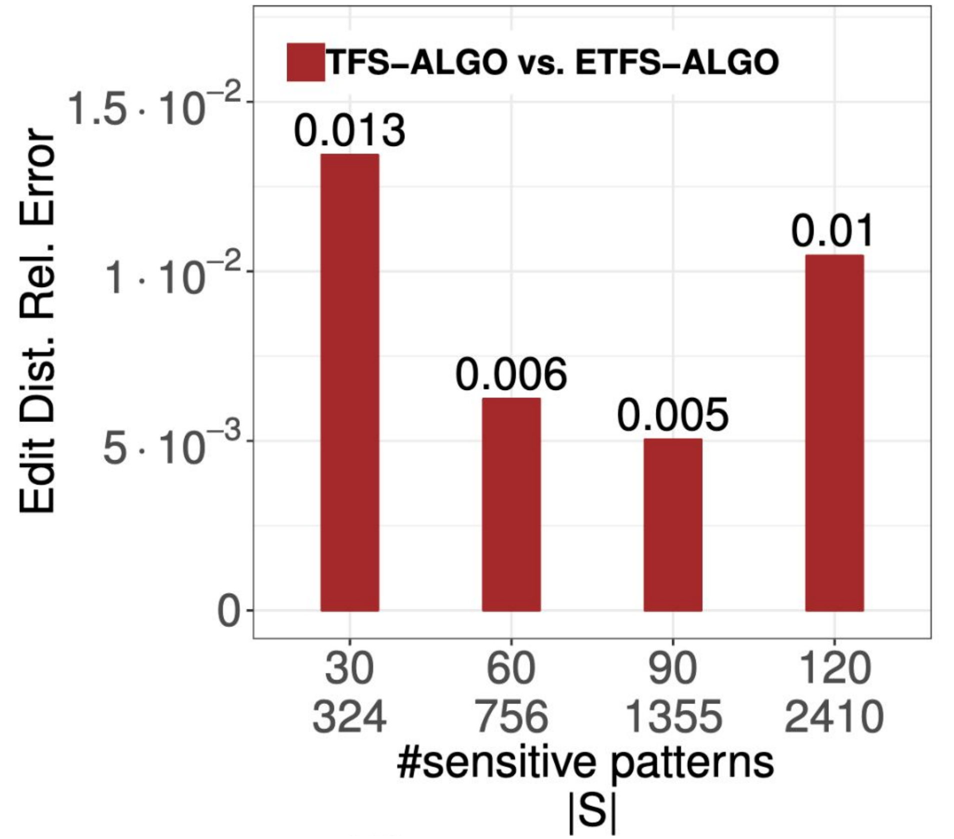
Experiments for MSL: runtime



SYN dataset: uniformly random string of length 20 millions, with 1000 sensitive patterns that occur ~20000 times, alphabet of size 10

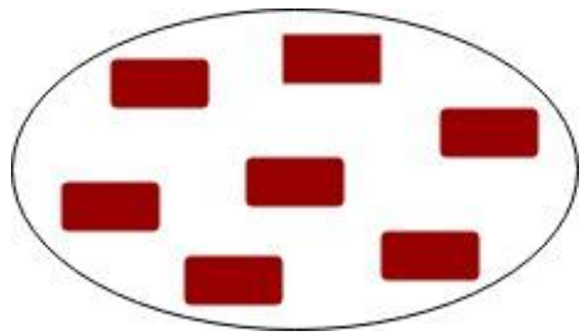
MSL is a heuristic for MED



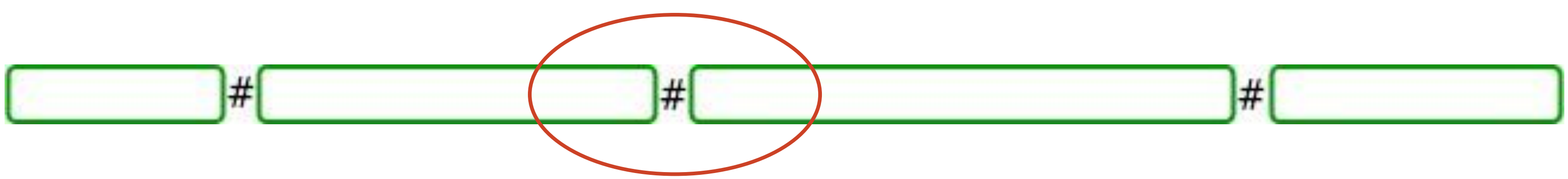


TRU dataset: transportation data of length ~6000 and alphabet of size 100. |S| is the number of occurrences of sensitive patterns.

Replacing the spurious characters



Forbidden strings of length *k* over Σ



Replace with a letter from Σ such that

1. No forbidden strings are introduced and
2. The accuracy of frequent pattern mining is preserved

Frequent pattern mining problem

IN: a string W, an integer $k>0$, a frequency threshold $\tau>0$

OUT: the set of length-k substrings of W whose frequency is $\geq \tau$

$$W \rightarrow X \rightarrow Z$$

A τ-ghost is a substring of Z whose frequency in Z is $\geq \tau$ and whose frequency in W is $< \tau$

τ-ghosts

W = GACAAAAACCCAT

k=3; τ=2

A sanitized version of W:

GAC#AA#ACCC#CAT

Replacing the second occurrence of # with G makes GAC a τ-ghost:

GACGAAGACCCGCAT

Hide and Mine problem

IN: an integer $k>0$, a string $X=X_0\#X_1\#\ldots\#X_d$ with all X_i over Σ, a set of forbidden strings of length k over Σ, a frequency threshold $\tau>0$

OUT: a replacement function $g : [d] \rightarrow \Sigma$ such that $Z=X_0 g(1) X_1 g(2) \ldots g(d) X_d$ is such that

1. No forbidden strings occur in Z
2. The number of τ-ghosts is minimized

Hide and Mine in Strings: Hardness, Algorithms, and Experiments. G.Bernardini, A.Conte, G.Gourdel, R.Grossi, G.Loukides, N.Pisanti, S.Pissis, G.Punzi, L.Stougie, M.Sweering. TKDE, 2022

Hide and Mine is hard

The decision version of Hide and Mine is strongly NP-complete, via a reduction from the bin packing problem.

Hide and Mine itself is hard to approximate.

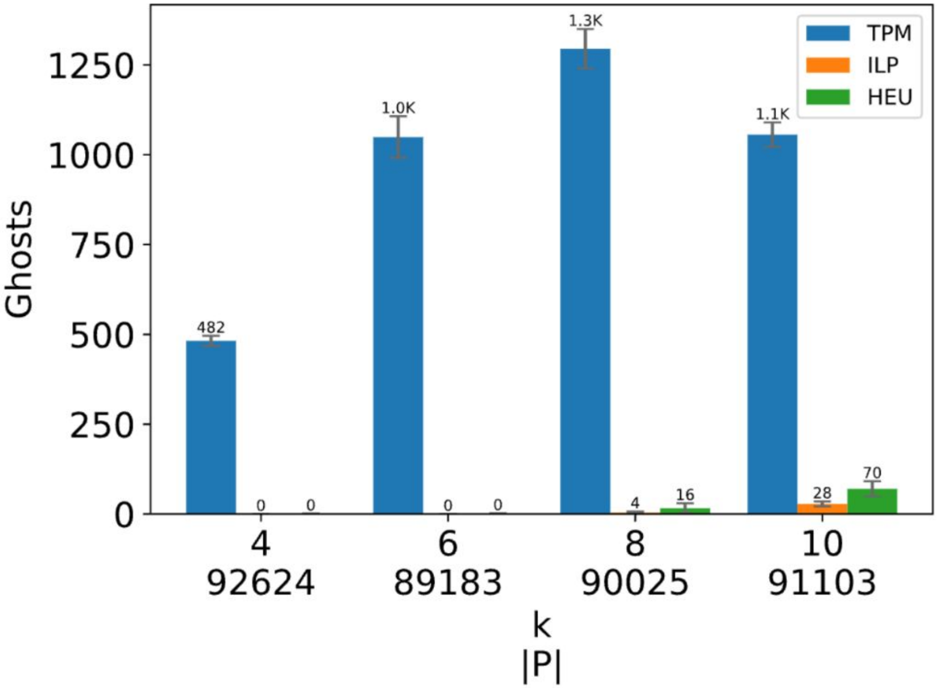
Hide and Mine in Strings: Hardness, Algorithms, and Experiments. G.Bernardini, A.Conte, G.Gourdel, R.Grossi, G.Loukides, N.Pisanti, S.Pissis, G.Punzi, L.Stougie, M.Sweering. TKDE, 2022

Algorithms for Hide and Mine

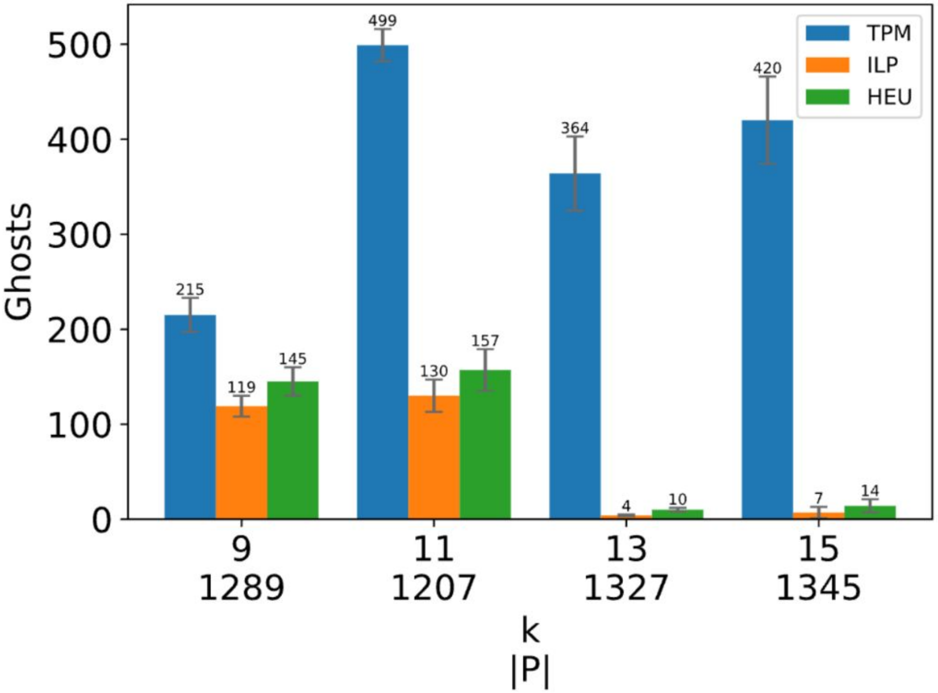
An ILP formulation of the problem is fixed-parameter tractable for many realistic parameter combinations: e.g., when both $|\Sigma|$ and k are O(1).

Hide and Mine in Strings: Hardness, Algorithms, and Experiments. G.Bernardini, A.Conte, G.Gourdel, R.Grossi, G.Loukides, N.Pisanti, S.Pissis, G.Punzi, L.Stougie, M.Sweering. TKDE, 2022

Experiments



(c) MSN



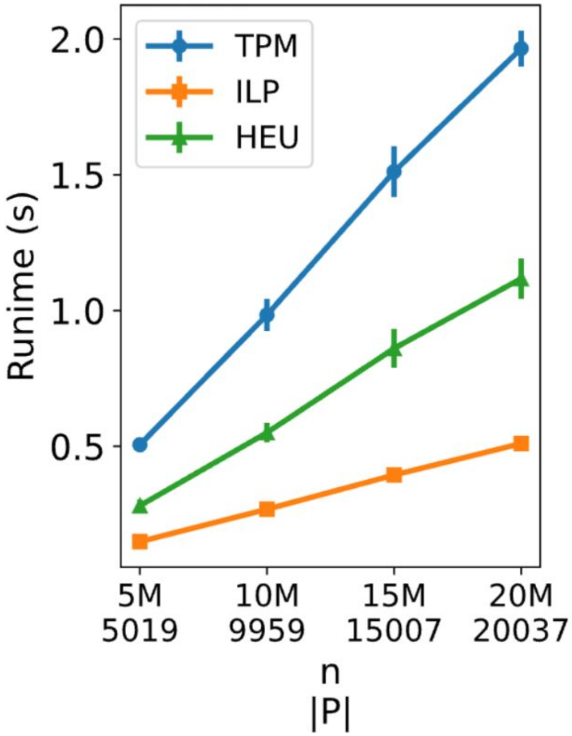
(d) DNA

MSN: clickstream data, n>4.6 millions, |Σ|=17, τ=200

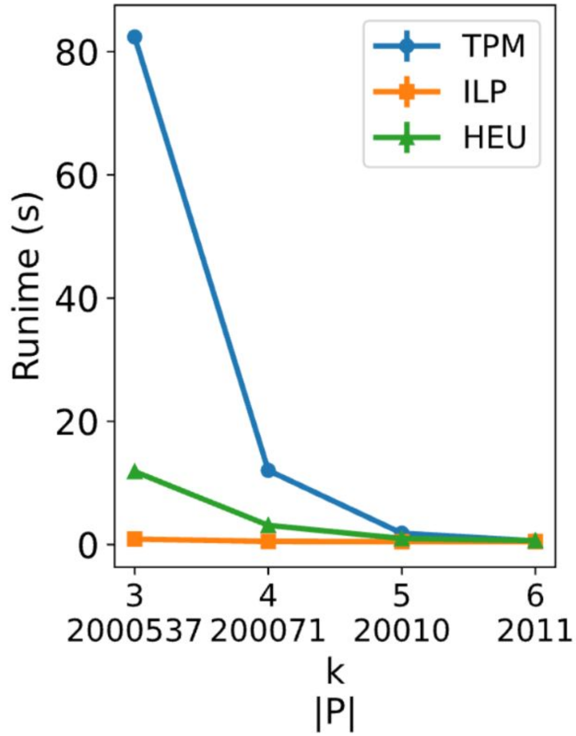
DNA: genomic data, n>4.6 millions, |Σ|=4, τ=20

|P|=number of occ. of sensitive patterns in X

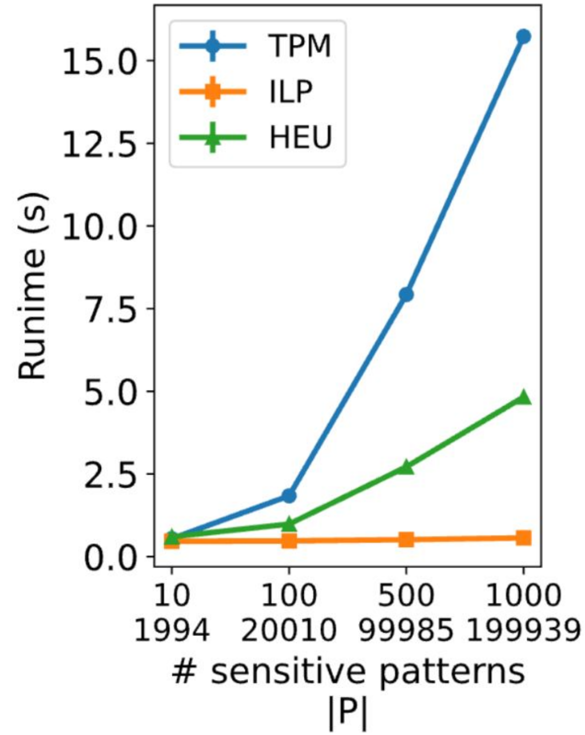
Experiments



(a) Substr. of SYN1



(b) SYN1



(c) SYN1

SYN1: uniformly random string, n=20 millions, k=5, |S|=100, τ=10

|P|=number of occ. of sensitive patterns in X

Thank you for your attention