CWI-Inria
INTERNATIONAL LAB

CWI

Inria
inventeurs du monde numérique

Workshop 2019
September 17-18, 2019, Amsterdam

Solon Pissis (CWI)
ERABLE project

Title: « String Sanitization: A Combinatorial Approach »

Abstract:

String data are often disseminated to support applications such as location-based service provision or DNA sequence analysis. This dissemination, however, may expose sensitive patterns that model confidential knowledge (e.g., trips to mental health clinics from a string representing a user's location history). In this talk, we consider the problem of sanitizing a string by concealing the occurrences of sensitive patterns, while maintaining data utility.

First, we propose a time-optimal algorithm, TFS-ALGO, to construct the shortest string preserving the order of appearance and the frequency of all non-sensitive patterns. Such a string allows accurately performing tasks based on the sequential nature and pattern frequencies of the string.

Second, we propose a time-optimal algorithm, PFS-ALGO, which preserves a partial order of appearance of non-sensitive patterns but produces a much shorter string that can be analyzed more efficiently. The strings produced by either of these algorithms may reveal the location of sensitive patterns. In response, we propose a heuristic, MCSR-ALGO, which replaces letters in these strings with carefully selected letters, so that sensitive patterns are not reinstated and occurrences of spurious patterns are prevented. We implemented our sanitization approach that applies TFS-ALGO, PFS-ALGO and then MCSR-ALGO and experimentally show that it is effective and efficient