International Journal of Population Data Science (2018) 3:3:134

International Journal of Population Data Science

Journal Website: www.ijpds.org

Evaluating hardening techniques against cryptanalysis attacks on Bloom filter encodings for record linkage

Ranbaduge, T[1], Vidanage, A[1], Vaiwsri, S[1], Schnell, R[2], and Christen, P[1]

[1]The Australian National University

[2]University Duisburg-Essen

Introduction

Due to privacy concerns personal identifiers used for linking data often have to be encoded (masked) before being linked across organisations. Bloom filter (BF) encoding is a popular privacy technique that is now employed in real-world linkage applications. Recent research has however shown that BFs are vulnerable to cryptanalysis attacks.

Objectives and Approach

Attacks on BFs either exploit that encoding frequent plain-text values (such as common names) results in corresponding frequent BFs, or they apply pattern mining to identify co-occurring BF bit positions that correspond to frequent encoded q-grams (sub-strings). In this study we empirically evaluated the privacy of individuals encoded in BFs against two recent cryptanalysis attack methods by Christen et al. (2017/2018). We used two snapshots of the North Carolina Voter Registration database for our evaluation, where pairs of records corresponding to the same voter (with name or address variations) resulted in files with 222,251 BFs and 224,061 plain-text records, respectively.

Results

We encoded between two and four of the fields first and last name, street, and city into one BF per record. For combinations of three and four fields all plain-text values and BFs were unique, challenging any frequency-based attack. For hardening BFs, different suggested methods (balancing, random hashing, XOR, BLIP, and salting) were applied.

Without any hardening applied up to 20.7% and 5% of plain-text values were correctly re-identified as 1-to-1 matches by both the pattern-mining and frequency-based attack methods, respectively. No more than 5% correct 1-to-1 re-identification matches were achieved with the frequency-based attack on BFs encoding two fields when either balancing, random hashing, or XOR folding was applied; while the pattern-mining based attack was not successful in any correct re-identifications for any hardening technique.

Conclusion/Implications

Given that BF encoding is now being employed in real-world linkage applications, it is important to study the limits of this privacy technique. Our experimental evaluation shows that although basic BFs without hardening technique are susceptible to cryptanalysis attacks, some hardening techniques are able to protect BFs against these attacks.

http://dx.doi.org/10.23889/ijpds.v3i4.725