Our research contributes to Encrypted Search in a few different ways. We designed and implemented several different new types of secure indexes—PSIB, PSIP, PSIF, and PSIM—based on a probabilistic set we call the Perfect filter.

On various metrics, we compared PSIB, PSIP, and PSIP to each other and to BSIB, a previously proposed secure index based on the popular Bloom filter probabilistic set. We compared them with respect to lag time, compression ratio, build time, load time, precision, recall, BM25, and MinDist\*. On most benchmarks, the PSI-based secure indexes compared favorably to BSIB, especially with respect to lag time. Moreover, given the flexibility of the Perfect filter, we were able to use more sophisticated representations, like PSIP, which performed in some cases significantly better than both PSIB and BSIB in terms of accuracy and significantly better than BSIB in terms of lag time.

Moreover, in PSIP location uncertainty is independent of every other observed output except MinDist\*—e.g., location uncertainty is independent of compression ratio, build time, BM25, etc. This flexibility makes it possible to choose a location uncertainty independent of concerns over these other outputs, and thus the choice of a location uncertainty becomes exclusively a trade-off between location accuracy (e.g., MinDist\* accuracy) and confidentiality—i.e., confidentiality and location accuracy are inversely proportional. However, it performed generally worse on the compression ratio metric, although suggestions for ways to significantly improve this outcome were discussed.

We also explored the use of these standard information retrieval scoring techniques while paying close attention to confidentiality concerns. In general, we discovered that query privacy is greatly improved by using obfuscated queries and multiple secrets. Obfuscations had an insignificant impact on BM25 and MinDist\*, and thus Encrypted Search is free to use obfuscations without significantly degrading relevancy of search results. However, obfuscations did significantly negatively affect Boolean search, e.g., if a user submits a search to find documents containing all of the terms in a query, and the query has obfuscations (fake terms), then very few documents (depending on the false positive rate) will both have the terms of interest to the user, and the obfuscated terms. Secrets, however, had no impact on the quality of search results, but they do degrade the compression ratio.

Experimentally, higher rates of obfuscation did not necessarily improve query privacy. Indeed, there was a global optimum s.t. confidentiality becomes progressively worse as you move away from it in either direction. We speculate that this was due to the distribution of obfuscation terms (uniformly sampled) being significantly different than the distribution of real terms (Zipf). If we chose a distribution for obfuscations that more accurately resembled the distribution of real terms, we believe there would be no such sweet spot; the higher the obfuscation rate, the better.

Including multiple secrets for each searchable atomic term (i.e., multiple trapdoors) also had a huge impact on query privacy. In this case, the more secrets there are, the stronger the confidentiality is (with respect to a simulated adversary using maximum likelihood attacks). However, there comes a point of diminishing returns in which the advantage of including one more secret is very unlikely to outweigh its cost in other respects, namely compression ratio.

Analytically, we also discovered that the location uncertainty should be quite large to preserve document confidentiality against an adversary, who has access to the raw contents of the secure index, employing jig-saw-like attacks. However, increasing the location uncertainty has a significant negative impact on MinDist\* MAP accuracy[[1]](#footnote-1). In response to this insight, we designed and implemented the PSIM secure index, and suggested using it as a way to make any of the other secure indexes more sensitive to the MinDist\* metric without revealing too much about the confidential document (i.e., the adversary would not have as much success with the proposed jig-saw-like attacks against PSIM).

While we did not perform any simulations of an adversary trying to compromise the confidentiality of secure indexes using such techniques, we did provide a detailed theoretical treatment to motivate experiments on secure index poisoning and high false positive rates.

Increasing the false positive rate had little impact on compression ratio and only a small impact on BM25 and MinDist\* MAP scores. However, increasing the false positive rate had a huge impact on precision, as expected. This expectation was one of the motivations for exploring the use of standard degree of relevancy scoring techniques for Encrypted Search—these scoring techniques seem to work well despite the presence of approximation errors and false positives when using secure indexes.

Also, secure index poisoning, especially in the form of adding fake terms, encouragingly had little impact on BM25 MAP, MinDist\* MAP, and precision, and only a modest impact on compression ratio. An Encrypted Search user is relatively free to poison a secure index to whatever desired level and still be justified in expecting good search performance in terms of accuracy and speed.

1. On top of that, PSIB and BSIB do not scale well to large numbers of blocks—which is the ratio of word count to location uncertainty—and thus it may not even be a reasonable option to use small location uncertainties (unless the documents are reasonably small). PSIP, as just discussed, does not possess this problem. [↑](#footnote-ref-1)