Contents

[Experiment Design 2](#_Toc391968707)

[Inputs 3](#_Toc391968708)

[Outputs 3](#_Toc391968709)

[Global Experiment Design Parameters 4](#_Toc391968710)

[Experimental Results 5](#_Toc391968711)

[BM25 MAP “Page One” Results 5](#_Toc391968712)

[MinDist\* “Page One” Results 7](#_Toc391968713)

[Attack Simulation: Obfuscations vs Accuracy 8](#_Toc391968714)

[Attack Simulation: Secrets vs Accuracy 10](#_Toc391968715)

[Attack Simulation: History Samples vs Accuracy 12](#_Toc391968716)

[Secrets vs Compression Ratio, Build Time, and Load Time 14](#_Toc391968717)

[False Positives vs BM25 MAP and Precision 16](#_Toc391968718)

[Obfuscations vs BM25 18](#_Toc391968719)

[Obfuscations vs MinDist\* 20](#_Toc391968720)

[Pages vs Secure Index Size 22](#_Toc391968721)

[Blocks vs Secure Index Size 25](#_Toc391968722)

[Documents (per corpus) vs Corpus Secure Index Size 26](#_Toc391968723)

[Pages vs Build Time 27](#_Toc391968724)

[Documents (per corpus) vs Build Time 29](#_Toc391968725)

[Pages vs Load Time 30](#_Toc391968726)

[Documents vs Load Time 31](#_Toc391968727)

[Pages vs MinDist\* Lag Time 32](#_Toc391968728)

[Simulation: Location Uncertainty vs Location Error 35](#_Toc391968729)

[Simulation: Location Uncertainty vs Minimum Pairwise Distance Error 36](#_Toc391968730)

[Location Uncertainty vs MinDist\* MAP 37](#_Toc391968731)

[Pages vs BM25 Lag Time 38](#_Toc391968732)

[Location Uncertainty vs BM25 MAP 41](#_Toc391968733)

[Pages vs Boolean Lag Time 43](#_Toc391968734)

[Compression Ratio (Secure Index Size to Document Size) vs MinDist\* MAP 45](#_Toc391968735)

[Background Information 46](#_Toc391968736)

[Precision and recall 46](#_Toc391968737)

[BM25 46](#_Toc391968738)

[MinDist\* 47](#_Toc391968739)

[Mean average precision (MAP) 49](#_Toc391968740)

[Secure indexes 50](#_Toc391968741)

[PSI 50](#_Toc391968742)

[PsiBlock (PSIB) 51](#_Toc391968743)

[PsiFreq (PSIF) 53](#_Toc391968744)

[PsiPost (PSIP) 53](#_Toc391968745)

[Bloom filter secure index (BSIB) 54](#_Toc391968746)

[Simulating an attack using query history 54](#_Toc391968747)

[Cryptographic Hash Attacks 55](#_Toc391968748)

[Maximum Likelihood Attack 56](#_Toc391968749)

[Simulating an attacker reconstructing documents from secure index information 58](#_Toc391968750)

[Problems with the block-based approach used by PSIB and BSIB 59](#_Toc391968751)

[An alternative solution that overcomes many of the problems for the block-based approach 59](#_Toc391968752)

[On the effect of false positives 60](#_Toc391968753)

[Secure index poisoning 61](#_Toc391968754)

[Query confidentiality – additional measures 61](#_Toc391968755)

[Letter n-grams and word n-grams 62](#_Toc391968756)

[Wild-card searching 62](#_Toc391968757)

[Approximate searching and error tolerance 62](#_Toc391968758)

[Boolean proximity searching 63](#_Toc391968759)

[Caching results 63](#_Toc391968760)

[Definition of a query 63](#_Toc391968761)

[Experiment Platforms 64](#_Toc391968762)

# Experiment Design

Experiments are intended to explore how one or more inputs relates to one or more outputs. To avoid over-complicating the experiments, techniques like multiple linear regression were not generally used. Instead, designs in which one input is changed (while the other inputs were held constant) and one output is observed was used.

## Inputs

* Secure index. The type of secure index. It is either PSIB, PSIF, PSIP, or BSIB. In most of the experiments, multiple secure indexes and their respective outputs are compared to one another. See Secure indexes on page 31 for more details.
* Documents (documents/corpus). Number of documents in the corpus. A variable corpus size should effect most outputs in a linear way, e.g., MinDist\* lag time should depend linearly on the number of documents (assuming documents are of fixed size). However, MinDist\* scoring and BM25 scoring may be effected in a non-linear way, thus I make this variable to see how such outputs respond.
* Pages. The number of pages in each document in the corpus. A variable page count will be used to see how each *secure index* scales with document size with respect to a number of parameters.
* Terms/query. The number of terms in a query, where a term is either a keyword or an exact phrase.
* Words/term. The number of words in each term.
* Secrets. Number of secrets that can be used to search for query terms in the secure index database.
* Obfuscations. This input is used in two different senses. In the context of the attack simulation, this input refers to the number of unique obfuscations; otherwise, it refers to the number of obfuscated terms added to a query.
* Obfuscation rate. In history attack simulations, obfuscation rate refers to the probability that a random term in the history set will be an obfuscated term.
* Location uncertainty. Unigram or bigrams in the document have exact positions. Exact positions reveal too much information about the contents of the document; thus, positions should only be known approximately. Simulating an attacker reconstructing documents from secure index information on page 39 for more information.
* False positive rate. A word not in a secure index document will have a probability (the false positive rate) of testing positively as belonging to it. This probability can be controlled by increasing or decreasing the input for false positive rate. On the one hand, a low false positive rate should improve search accuracy; on the other hand, a high false positive rate should improve confidentiality (e.g., more difficult for an attacker to reconstruct the document). See On the effect of false positives on page 41 for more information.

## Outputs

* Secure Index Size. The size (e.g., bytes) of the secure index database for a corresponding corpus (collection of documents).
* Build Time. Time taken to build the secure index database for a given corpus.
* Load Time. Time taken to load a secure index database for a given corpus.
* Boolean query precision. Proportion of retrieved documents relevant in a Boolean query, in which all of the query’s terms must exist in the document. See Precision and recall on page 43 for more details.
* Boolean query recall. Proportion of relevant documents were retrieved. See Precision and recall on page 43 for more details.
* BM25/MinDist\* ranking MAP (mean average precision). See Mean average precision (MAP) on page 45 for more details.
* BM25/MinDist\*/Boolean ranking lag. Time taken for the corresponding kind of query to complete.

# Global Experiment Design Parameters

* The number of unique words per corpus (corpus dictionary) is fixed at 10,000 words. The unique words in the dictionary follow a Zipf distribution, and they are randomly generated with an average length of 6.5 alphabetic characters. Such a dictionary is uniquely constructed for each trial of every experiment.
* For BM25 scoring, parameter b is set to 0.75 and parameter k1 is is set to 1.2. See Background Information for more details.
* Each document of size n (n words) in the corpus separately samples m= unique words from the corpus dictionary, conforming to Heap’s law with K=12 and β=0.5. Once m unique words are sampled, they are renormalized to make them into a proper distribution. It is this distribution that is used to generate the sequence of words for a document.  
    
  I did this with the intention of making each document approximately follow a Zipf distribution, but with a different subset of words to account for different authors with different but overlapping vocabularies. In hindsight, it would have been sufficient (and perhaps preferable) to have simply sampled n words directly from the corpus dictionary.
* When calculating the outputs for a given input, e.g., false positive rate, location uncertainty, etc., I always use a query set consisting of 30 queries. I then average the outputs over all of those queries where appropriate.
* For each query term in a query in a query set, I seed a document in the corpus with that term with probability p = 0.2 except where otherwise noted. Thus, for a query with k terms, the probability that one or more of its terms occurs in the document is . For k = 1, P[at least one] = p = 0.2; for k = 2, P[at least one] = . Thus, if a query consisting of k terms seeds a corpus of size N, then on average documents will be seeded with the query term.   
    
  In general, this should mean when doing a mean average precision calculation, the expected number of documents relevant to a query with k terms will be . The remainder should be non-relevant, i.e., MAP scores of 0, which means it is irrelevant how they are ranked and can thus be ignored in the mean average precision calculation. Of course, if the secure index does score them with a score of non-zero, that is a sign of a false positive, and should and does subsequently degrade the MAP score. This happens with probability for those documents which are nonrelevant—on average, for a query with k terms, this will happen . For k = 3, N = 1000, p = 0.2, and fp = 0.001, it is approximately 1.5 times.  
    
  To clarify, when calculating a MAP score, I include as many documents for scoring in the ranked list as there are non-zero scores for retrieved ranked documents.  
    
  Once a document is targeted to be seeded by a given query term, all such occurrences of the term will occur within a window size of w ~ U(min(size(doc), 2000), max(size(doc), 6000)) except where otherwise noted.  
  Finally, the number of occurrences of the term within that window will be n = }.
* When measuring precision, MinDist\* MAP, or BM25 MAP, each query in the query set (as previously mentioned, there are 30 queries per query set in total) is submitted 10 times for the block based indexes, and the average of those 10 is taken to be the output.

# Experimental Results

## BM25 MAP “Page One” Results

Figure

|  |  |  |
| --- | --- | --- |
| Experiment Details | | |
| Input | location uncertainty vs BM25 Top 10 MAP (first page of results) |
| Output | BM25 Top 10 MAP | |
| Constants | 1 secret  0 obfuscations  0.001 false positive rate  1000 documents (documents/corpus)  1 or 2 words/term  3 terms/query  16 pages | |
| Testbed | machine A | |

This is the “page one” Google test. Search users do not want to dig through multiple pages to find what they want. Indeed, studies have shown that Google’s second page of results only receives 1.5% of click-through rate.

In this experiment, we get the top 10 results according to the canonical index, and then get the top 10 results each secure index and restrict the mean average precision to only those top 10. This is a much more demanding measure than taking the mean average precision over all of the results.

In Figure 2, I randomly return the top 10 results for 250 documents (note that the real experiment is even more unforgiving since it draws the top 10 results from 1000 documents). Here is a histogram of the results; note that over 90% of the results have a MAP between 0.0 and 0.1.

Compare the random results with the results returned from the secure indexes. They all do remarkably well; PSIP and PSIF, which can optionally preserve perfect frequency information for words (unigrams and bigrams) in the document, come out on top. Indeed, it achieves nearly 100%. Note that PSIP and PSIF are independent of location uncertainty—PSIF does not even store location information, and PSIP’s frequency information is independent of the location uncertainty.

The block-based indexes. PSIB and BSIB, also do quite well, although their scores expectedly trail off as the location uncertainty increases.

Figure

## MinDist\* “Page One” Results

Figure

|  |  |  |
| --- | --- | --- |
| Experiment Details | | |
| Input | location uncertainty vs BM25 Top 10 MAP (first page of results) |
| Output | BM25 Top 10 MAP | |
| Constants | 1 secret  0 obfuscations  0.001 false positive rate  1000 documents (documents/corpus)  1 or 2 words/term  3 terms/query  16 pages | |
| Testbed | machine A | |

Once again, we see PSIP pulling ahead. However, discouragingly, no matter which secure index is chosen, location uncertainty must be quite modest for MinDist\* to achieve competent scores.

The MinDist\* measure is more sensitive to location uncertainty than BM25 is to frequency uncertainties. This makes sense; when location uncertainty is modestly large, two terms in the same block, for instance, may not be relevant towards one another, e.g., whether terms are 3 pages apart or 40, in either case there is probably not much difference with respect to their relevance to one another. MinDist\* captures this intuition s.t. only documents in which the terms in the query are sufficiently close are relevant. Unfortunately, reasonably large location uncertainties are necessary for confidentiality—see Simulating an attacker reconstructing documents from secure index information on page 57 for more on this.

Note that this fact about proximity sensitivity is the primary motivation behind PSIM (PsiMinPair), which can (optionally) preserve perfect min-pairwise distance information for terms in the document that are up to words apart. (However, I do not conduct any experiments with PSIM in this document. It does work, but it is slow and memory-hungry for but the smallest values of .)

None of this is necessarily bad news. The first page of results may still relevant to the search user. Moreover, oblivious search users will probably be more sophisticated and thus more willing to dig deeper into the results to find what they are searching for.

## Attack Simulation: Obfuscations vs Accuracy

Figure

|  |  |  |
| --- | --- | --- |
| Experiment Details | | |
| Input | unique obfuscations (unique strings in the uniform distribution being sampled from) |
| Output | accuracy (proportion of hidden terms correctly mapped to the corresponding plaintext term) | |
| Constants | 1 secret  50 word search vocabulary (every query is composed from the same 50 unique search terms)  50,000 query term history (to be used as data points in MLE)  150,000 samples (in the Monte Carlo simulation to approximate MLE) | |

In this experiment, I am interested in seeing how accurately a hypothetical attacker, using maximum likelihood estimation, can learn a mapping from hidden terms to plaintext query terms with respect to the unique number of obfuscations for several obfuscation rates (which is just the probability that a random query term will be an obfuscated term). Whenever an obfuscation is injected into a query, I sample the obfuscated term from a discrete uniform distribution consisting of N unique strings. See Maximum Likelihood Attack on page 36 for background information on this.

From the graph, one may conclude that for a given obfuscation rate, there comes a point at which increasing the number of unique obfuscations has little effect on mitigating the attacker. The lower the obfuscation rate, the sooner this point is reached. Additionally, the lower the obfuscation rate, the larger the attacker’s limiting accuracy as n goes to infinity.

For high obfuscation rates, it is a mistake to let the number of unique obfuscations N be small. I imagine the reason for this is related to the fact that the obfuscations will tend to have a higher frequency than most of the real terms if is larger than the probability for most real terms. Thus, mapping a non-obfuscated hidden term to the obfuscation class will cause the likelihood of seeing the given history much lower. This presents another area to explore. Instead of sampling the obfuscated terms from a uniform distribution, sample them from a distribution which is designed to resemble, in some way, the real distribution s.t. incorrect mappings are less penalized in the MLE calculation.

Figure

In Figure 19, we see that (with the same fixed constants as before) the optimal combination of number of obfuscated terms and obfuscation rate—the combination that minimizes the attacker’s accuracy—is 50 obfuscated terms and 0.2 obfuscation rate.

## Attack Simulation: Secrets vs Accuracy

Figure

|  |  |  |
| --- | --- | --- |
| Experiment Details | | |
| Input | secrets |
| Output | accuracy (proportion of hidden terms correctly mapped to the corresponding plaintext term) | |
| Constants | 0 obfuscations  50 word search vocabulary (every query is composed from the same 50 unique search terms)  50,000 query term history (to be used as data points in MLE)  150,000 samples (in the Monte Carlo simulation to approximate MLE) | |

In this experiment, I am interested in seeing effective secrets are at mitigating the hypothetical MLE attacker. Arguably, it can be effective; moreover, its effectiveness comes at no cost to MAP accuracy and query lag time (except, possibly, for BSIB). However, it does cost in terms of inflating the secure index size.

Additionally, the marginal value of secrets has diminishing returns. Eventually, there comes a point where it hardly makes a difference at all, but you are likely to run out of memory space before that happens.

It is worth pointing out that the secrets for a given term are sampled from a discrete uniform distribution. This probably limits the effectiveness of having secrets. The attacker may be able to infer the underlying plaintext distribution using big data and statistics, but the attacker cannot know (just as with obfuscations) the distribution of an individual user’s secret distribution which may be randomly re-defined periodically (not only an unknown distribution, but a moving distribution). Indeed, each user can have their own way of sampling secrets (and the same is true for sampling obfuscated terms).

I expect that any experimental results on this would look quite promising.

## Attack Simulation: History Samples vs Accuracy

Figure

Figure

|  |  |  |
| --- | --- | --- |
| Experiment Details | | |
| Input | secrets, history size |
| Output | accuracy (proportion of hidden terms correctly mapped to the corresponding plaintext term) | |
| Constants | 0 obfuscations  50 word search vocabulary (every query is composed from the same 50 unique search terms)  150,000 samples (in the Monte Carlo simulation to approximate MLE) | |

As the number of query samples (history) increases, so too does the attacker’s accuracy. This, of course, is expected; the more data the attacker has to learn the model (the mapping from hidden terms to plaintext terms), the more accurate the model should be.

Indeed, for a history size of 512k, if only using one secret the attacker has a 93% accuracy rate. Increasing the number of secrets to 16 reduces the attacker’s accuracy rate to 36%. Lower is certainly better, but preferably it would be lower yet. Granted, this is a toy problem; there are only 50 words in the user’s search vocabulary, for instance. However, according to equation for the curve representing 512k history samples, we would need over 700 secrets to reduce the attacker’s accuracy to 10%. Since secrets inflate the size of the index, this is not a viable option.

However, as discussed elsewhere, if the secrets were not sampled uniformly, much better results could probably be realized.

## Secrets vs Compression Ratio, Build Time, and Load Time

Figure

Figure

Figure

|  |  |  |
| --- | --- | --- |
| Experiment Details | | |
| Input | secrets |
| Output | compression ratio (ratio of secure index size to document size), load time, build time | |
| Constants | 12 pages  256 location uncertainty | |

The only outputs secrets affected were build time, load time, and secure index size. For PSIB and PSIF, load time is nearly constant with respect to secrets. However, all of the secure indexes flatten out as the secrets increase (as they do for build time, also).

The compression ratio output is linear with respect to the number of secrets; this certainly makes sense, as each secret variation of each term will be dedicated a constant number of bits.

## False Positives vs BM25 MAP and Precision

Figure

|  |  |  |
| --- | --- | --- |
| Figure  Experiment Details | | |
| Input | pages (~250 words/page) |
| Output | secure index size (bytes) | |
| Constants | 12 pages  1 term/query  1 or 2 words/term  1 secret  0 obfuscations  128 location uncertainty  0.001 false positive rate  1000 documents (documents/corpus) | |
| Testbed | machine A | |

While not much space is saved by decreasing the false positive rate, the primary advantage in having a high false positive rate is its effect on confidentiality. The higher the false positive rate, the less certain the information in the secure index is. See On the effect of false positives on page 51 for more analysis on this.

Thus, Figure 9 paints a fairly positive picture for BM25 scoring. Indeed, false positives occurring even a quarter of the time on negative examples still result in a BM25 MAP of around 0.7. And this is only for 1 term/query and 1 or 2 words/term; BM25 tends to perform better when given more terms to work with, as other experiments have shown.

Figure 10 is less encouraging. If false positives occur a quarter of the time here, only 50% accuracy is achieved. Compared to precision, BM25 is far less sensitive to the false positive rate. This makes sense; if a false positive happens when measuring precision, it will admit a term that should not be included in the result set, which will certainly effect its precision negatively. However, BM25is ranking the documents. Thus, even if a document is falsely hitting on a search term, it is the order that counts—how it ranks the document. Other documents will also falsely hit on the term (with probability 0.25), which will cause the false hit to not be very discriminating—that is, it is not rare in the corpus since 25% of the documents have it. In practice, it may do better than this experiment suggests since I did not include particularly rare terms in the query set.

## Obfuscations vs BM25

Figure

Figure

Figure

|  |  |  |
| --- | --- | --- |
| Experiment Details | | |
| Input | Obfuscations (per query) |
| Output | BM25 MAP, BM25 lag | |
| Constants | 1 secret  256 location uncertainty  12 pages  0.001 false positive rate  1000 documents (documents/corpus)  1 term/query, 1 or 2 words/term *or* 6 terms/query, 6 words/term | |
| Testbed | machine B | |

In Figure 11, we plot obfuscations versus BM25 MAP on a rather large type of query—6 terms/query, 6 words/term. PSIP and PSIF perform very close to 100% and each additional obfuscation per query only reduces BM25’s MAP score by 0.005%. PSIB and BSIB also do well and remain largely unaffected by the obfuscated terms as well.

Figure 12 reveals that every obfuscated term added to the query increases the BM25 lag time by 0.014 milliseconds. Note that this lag time cannot be directly compared with the lag time reported in other BM25 experiments, as this experiment was conducted by a different machine (see Experiment Platforms on page 61 for more information on the machines). However, it does not seem unreasonably slow. Also, note that every one of the queries is slow compared to smaller, more typical queries performed in other experiments.

Finally, Figure 13 shows BM25 MAP on a more modest set of queries consisting of 1 term/query and 1 or 2 words/term. This results in an across the board reduction in the BM25 MAP score. However, as discussed elsewhere, in that BM25 generally does better on more complicated queries. Unfortunately, each additional obfuscated term injected into the query also has a larger negative impact on the BM25 MAP score, i.e., -0.013 vs -0.005.

## Obfuscations vs MinDist\*

Figure

Figure

|  |  |  |
| --- | --- | --- |
| Experiment Details | | |
| Input | Obfuscations (per query) |
| Output | MinDist\* MAP | |
| Constants | 1 secret  256 location uncertainty  12 pages  0.001 false positive rate  1000 documents (documents/corpus)  6 terms/query, 1 or 2 words/term | |
| Testbed | machine B | |

Earlier attack simulations demonstrated the effectiveness of obfuscations in mitigating attacks. Judging by Figure 14, increasing obfuscations have almost no effect on MinDist MAP scores. This is certainly welcome news—we can exploit obfuscations without incurring much, if any, loss in MinDist\* accuracy.

Figure 15 shows a linear relationship between obfuscations and MinDist\* lag time. This makes sense; it is essentially the same increase in lag time expected from any additional query terms—obfuscated terms or otherwise.

Note that PSIP is pulling ahead in a majority of the benchmarks measuring lag time or mean average precision. This is not unexpected; subsequent experiments will expand on why this is happening.

## Pages vs Secure Index Size

Figure

Figure

|  |  |  |
| --- | --- | --- |
| Experiment Details | | |
| Input | pages (~250 words/page) |
| Output | secure index size (bytes) | |
| Constants | 1 secret  256 location uncertainty  0.001 false positive rate  1000 documents (documents/corpus) | |
| Testbed | machine A | |

As discussed elsewhere, PSIB is optimized for smaller documents. For PSIP and BSIB, every page is approximately 1700 and 1500 bytes respectively; their secure index sizes are linearly dependent upon their page counts.

However, as demonstrated by Figure 2 and Figure 3 , the sparse bit vector representation used in the PSIB does well for small to moderate pages, but explodes as the pages increase past a certain point (~100 pages). It is quadratic with respect to page count rather than linear. For small page counts, the squared component is dominated by the linear component, but for large page counts the squared component dominates.

The point of intersection between PSIB and BSIB is ~50 pages. This is the size of a relatively large document; for larger documents (e.g., books) with more than 50 pages, it may be advisable to (automatically) segment them into smaller chunks.

## Blocks vs Secure Index Size

Figure

Figure

|  |  |  |
| --- | --- | --- |
| Experiment Details | | |
| Input | blocks (block segments per document) |
| Output | secure index size (bytes); compression ratio (ratio of secure index size to document size) | |
| Constants | 1 secret  256 location uncertainty  0.001 false positive rate  1000 documents (documents/corpus) | |
| Testbed | machine A | |

In the previous experiment, we examined how page count affected secure index size while location uncertainty was held constant at 256. This had the effect of increasing the number of blocks per PSIB and BSIB as the page count increased. This motivates us to consider how the block count per PSIB and BSIB are affects secure index size.

In Figure 3, we see that the lines for PSIB and BSIB cross at ~47 blocks. If the primary metric of interest is secure index size (as measured by memory allocation size), this point of intersection represents a dividing line; to the left of the line PSIB is preferable, and to the right of the line BSIB is preferable. Note, however, that PSIB has other advantages that continue to exist regardless of the block count, e.g., query lag times.

In Figure 4, we see that for a small number of blocks, PSIB is a small fraction—a quarter—the size of the actual document. However, it grows linearly as the block count increases. On the other hand, BSIB converges (to a first approximation) to a constant factor of ~0.75 the size of the document as the block count increases.

A high block count is ideal for MinDist\* and BM25 MAP accuracy—it reduces the location uncertainty—but there is a trade-off between such accuracy and the amount of information leaked about the document.

PSIP does not represent a document as blocks; it represents a document as a postings list. Thus, location uncertainty can be adjusted to any desired value and PSIP’s file size (and query lag times) will remain the same.

Note that, as discussed later, it would be possible to use a more efficient representation for denoting which terms appear in which blocks. For example, since the bit vector becomes increasingly sparse as the block count increases, a compressed bit vector representation would result in significantly higher space efficiency at the expense of computational overhead.

## Documents (per corpus) vs Corpus Secure Index Size

Figure

|  |  |  |
| --- | --- | --- |
| Experiment Details | | |
| Input | documents (per corpus) |
| Output | corpus secure index size (bytes) | |
| Constants | 1 secret  16 pages (per document)  250 location uncertainty  0.001 false positive rate | |
| Testbed | machine A | |

For a reasonably large document consisting of 16 pages (4000 words, 250 words/page), we see that the average document is ~10.5 kilobytes for PSIB, ~16.8 kilobytes for PSIP, and ~23.3 kilobytes for PSIP. Note that, with that many pages and with that location uncertainty, the blocks per document is 16 for PSIB and BSIB; this is under the threshold of ~47 blocks under which PSIB is superior to BSIB.

For a corpus of nearly 20,000 documents, the total corpus size is a little over 200 MB; the original corpus was 593 MB.

## Pages vs Build Time

Figure

Figure

|  |  |  |
| --- | --- | --- |
| Experiment Details | | |
| Input | pages (~250 words/page) |
| Output | build time (milliseconds) | |
| Constants | 1 secret  256 location uncertainty  0.001 false positive rate  1000 documents (documents/corpus) | |
| Testbed | machine A | |

In this experiment, I am interested in seeing how page count (~250 words/page) affects secure index build time. The byte-size of the document is less important than its page count. BSIB is nearly twice as slow as PSIB and PSIB, but even BSIB is only 2.4 milliseconds per page.

None of the secure indexes are unreasonably slow; even a document consisting of ~300 pages takes only a fraction of a second to build. Indeed, the PSIB can build a ~800 page document in only a second. And, as discussed later, there are significant performance improvements that could be easily realized, e.g., replacing unnecessary cryptographic SHA256 re-hashes with non-cryptographic hash functions.

## Documents (per corpus) vs Build Time

Figure

|  |  |  |
| --- | --- | --- |
| Experiment Details | | |
| Input | documents (per corpus) |
| Output | corpus build time (milliseconds) | |
| Constants | 1 secret  16 pages (per document)  250 location uncertainty  0.001 false positive rate | |
| Testbed | machine A | |

The time to build a corpus consisting of reasonably large 16 page documents is 24 milliseconds per document for PSIB, 27 milliseconds for PSIP, and 36 milliseconds for BSIB.

In the previous experiment, we plotted page size vs build time. The line of best fit for PSIB build time ≈ 1.24 ∙ pages + 4.1; for PSIP, the line of best fit was build time ≈ 1.36 ∙ pages + 5.5; finally, for BSIB, the line of best fit was build time ≈ 2.4 ∙ pages - 2.4. Each one of these lines of best fit competently predicts the slopes in Figure 8.

## Pages vs Load Time

Figure

|  |  |  |
| --- | --- | --- |
| Experiment Details | | |
| Input | pages (~250 words/page) |
| Output | load time (milliseconds) | |
| Constants | 1 secret  256 location uncertainty  0.001 false positive rate  1000 documents (documents/corpus) | |
| Testbed | machine A | |

In this experiment, I am interested in seeing how page count affects secure index load time. Interestingly, PSIB (and to a greater extent although not shown here, PSIF) is nearly constant when representing documents from 1 page to 300 pages; 300 page documents take only 2.86 milliseconds to load raw from disk.

The other two perform less impressively. Concerning PSIP, I did not make much of an effort to optimize it. For instance, I load a term’s postings list as a vector of varints[[1]](#footnote-1), which incurs significant vector construction overhead as the number of terms in the document increases. A likely more efficient representation of a posting lists is a list consisting the word gaps between adjacent positions of a term. Since this list does not need to facilitate random access, i.e., operations on it can efficiently be performed on it in a sequential manner, the gaps may even be compressed, e.g., Huffman coded. Moreover, inspired by PSIB and its block-based approach, instead of storing the exact number of gaps N between adjacent positions of a term, store, where k is an integer denoting the block granularity size. And, of course, positions that map to the same block may be dropped to save even more space and processing time (but at the cost of a loss of frequency information, which may be desirable anyway).

For BSIB, as the document size increases, the overhead of de-serializing a larger number of Bloom filters may take a toll. However, the serialization seems fairly efficient.

## Documents vs Load Time

Figure

|  |  |  |
| --- | --- | --- |
| Experiment Details | | |
| Input | documents (per corpus) |
| Output | corpus load time (milliseconds) | |
| Constants | 1 secret  16 pages (per document)  250 location uncertainty  0.001 false positive rate | |
| Testbed | machine A | |

The time to load a corpus consisting of reasonably large 16 page documents is 0.26 milliseconds per document for PSIB, 2.13 milliseconds for PSIP, and 1.55 milliseconds for BSIB.

In the previous experiment, we plotted page size vs load time. The line of best fit for PSIB load time ≈ 2E-05 ∙ pages2 + 0.002 ∙ pages + 0.22. Plugging in pages = 16, we get a prediction of 0.26 milliseconds, which accurately matches the actual slope in Figure 10. The same is approximately the same for the other two secure indexes as well.

## Pages vs MinDist\* Lag Time

Figure

|  |  |  |
| --- | --- | --- |
| Experiment Details | | |
| Input | pages (~250 words/page) |
| Output | MinDist\* lag time (milliseconds) | |
| Constants | 1 secret  0 obfuscations  256 location uncertainty  0.001 false positive rate  2 terms/query  1 or 2 words/term | |
| Testbed | machine A | |

In this experiment, I am interested in seeing how page count affects MinDist\* lag time. BSIB is unique among the secure indexes in that MinDist\* lag time is linearly dependent upon page count; every additional page incurs ~0.0004 milliseconds.

For large documents (more specifically, for documents with a large number of block segments), BSIB performs poorly on this measure. This is the expected outcome. For a fixed location uncertainty, as the page count increases the document must be segmented into more blocks and therefore, because every block is assigned a Bloom filter, more Bloom filters must be queried (i.e., more hash functions must be evaluated). Since each Bloom filter hash function evaluation requires a constant amount of time, all query lag times—MinDist\* included—are dependent upon the number of hash functions that must be evaluated per document.

For a ~300 page document, the lag time is nearly 0.14 milliseconds. If the corpus consists of a million such documents, this operation would require nearly 140 seconds to complete. This is certainly impractical.

The PSI-based secure indexes, to a first approximation, take only a small constant amount of time with respect to page count. However, the constant—while small—will have scalability issue as corpus size grows to many thousands of documents. For instance, a corpus consisting of a million documents would require nearly ~20 seconds to complete. Even PSIP, the fastest secure index on this benchmark, would require ~7 seconds to complete the query.

None of them are below the 1 second mark for the million document example; one second response times are often considered the maximum delay a typical user will tolerate, and the network latency time is not even being factored into this measure. Of course, secure indexes do not represent the typical use case—the services provided by secure indexes are not free. Moreover, the slowed response times are far better than the alternative of downloading the entire corpus, decompressing the documents, and conducting local searches on them.

There are a two immediately obvious things that could significantly speed up the query operations like MinDist\*. First and foremost, each secure index in the database re-hashes the hidden query’s unigram and bigram terms with SHA256. While the re-hashing operation is desirable to ensure that a cryptographic hash of a term in one secure index looks nothing like the cryptographic hash of the same term in any other secure index, using SHA256 to perform the re-hashing is overkill; after all, the hidden query itself has already been transformed using SHA256.

According to my benchmarks, each evaluation of SHA256 takes ~0.0024 milliseconds on machine A. This is fairly significant; a single SHA256 hash around one-third the total time taken, on average, to complete a PSIP MinDist\* query (per secure index) consisting of two terms per query and one or two words per term. Moreover, ~0.0024 milliseconds is required for each hidden unigram or bigram in the hidden query. While the secure indexes short-circuit processing queries where appropriate (in this experiment consisting of two terms per query, they can at most avoid processing one term per query per document), it is clear that significant savings could be realized by using an orders-of-magnitude faster non-cryptographic hash function without any loss in confidentiality.

The second way to speed up query processing is through parallel programming techniques. Each query can be independently queried so this is an *embarrassingly parallel* problem and performance scales linearly with core count. Given N cores, PSIP MinDist\* query lag time would be ~0.0078/N milliseconds. It could complete the MinDist\* query operation against a million documents in less than a second given N=8 cores.

BSIB would require around N=140 cores to get under the one-second mark on a million documents consisting of ~300 pages per document. Using Bloom filters in the ways that have been previously proposed does not seem to be practical at scale. When we include some optimizations, like replacing the slow cryptographic SHA256 re-hash with a fast hash and incorporate memorization or caching (as discussed in **Background Information**), it may still work at scale in practice.

## Simulation: Location Uncertainty vs Location Error

Figure

|  |  |  |
| --- | --- | --- |
| Experiment Details | | |
| Input | location uncertainty |
| Output | average absolute location error | |
| Testbed | machine A | |

MinDist\*, discussed at length on page 30, relies on a secure index’s approximate location information. The less uncertain the location is, all things else being equal, the more accurate MinDist\* output will be (compared to the canonical index with perfect information).

However, if the location information is too precise, a hypothetical attacker will have more success at infer the contents of the document. Thus, the reported positions for a word must be uncertain—e.g., only reporting that a word falls within some range (block), as PSIB and BSIB do, or scrambling the positions in some random way, as PSIP does. See *simulating an attacker reconstructing documents from secure index information* on page 35 for a thorough analysis on this.

In this experiment, I am interested in observing the expected location error for block-based secure indexes and scrambled postings in which the scrambled positions of terms are randomly offset from their true position.

Figure 12 clearly shows that the block-based approach results in the greatest loss of accuracy. This was expected since each term is assigned a block range that is not centered on its true position (except for the term with a true position in the middle of the assigned block range). The other two (used by PSIP), centered\_uniform and centered\_triangular, do effectively do this by offsetting a term from its true position. They are named after the PDFs they sample their position offsets from: centered\_uniform uniformly samples an integral offset from , where is equal to the location uncertainty, and centered\_triangular samples from the triangular distribution with a mean and mode equal to the true position and a base also of length . The triangular distribution has a higher probability of being closer to the mean, so naturally it should have the least variance of the three.

The block-based approach has around 100% more error than the triangular approach for a given location uncertainty, and nearly 33% the error than the uniform approach. This seems fairly significant, especially in light of my analysis on page 35 which suggests that the block-based approach is also easier for the hypothetical attacker to compromise.

## Simulation: Location Uncertainty vs Minimum Pairwise Distance Error

Figure

|  |  |  |
| --- | --- | --- |
| Experiment Details | | |
| Input | location uncertainty |
| Output | average minimum pairwise distance error | |
| Testbed | machine A | |

In this simulation, instead of estimating the expected absolute location error for a term, I calculate the expected average minimum pairwise distance error. The error is now expected to be greater since there are two terms with uncertain locations instead of one, but the ratio of the errors between the three outputs are the same.

## Location Uncertainty vs MinDist\* MAP

Figure

Figure

|  |  |  |
| --- | --- | --- |
| Experiment Details | | |
| Input | location uncertainty |
| Output | MinDist\* MAP | |
| Constants | 1 secret  0 obfuscations  1000 documents (per corpus)  0.001 false positive rate  3 terms/query  1 or 2 words/term | |
| Testbed | machine A | |

As location uncertainty converges to 0, all of the secure indexes converge to the same point. However, as location uncertainty increases, PSIP quickly begins to diverge from the other two. These results reinforce the analysis in Simulation: Location Uncertainty vs Minimum Pairwise Distance Error.

Moreover, PSIB and BSIB do not scale well to large numbers of blocks (recall that location uncertainty is inversely proportional to number of blocks in PSIB and BSIB).

## Pages vs BM25 Lag Time

Figure

Figure

|  |  |  |
| --- | --- | --- |
| Experiment Details | | |
| Input | pages (~250 words/page) |
| Output | BM25 lag time (milliseconds) | |
| Constants | 1 secret  0 obfuscations  256 location uncertainty  0.001 false positive rate  2 terms/query  1 or 2 words/term | |
| Testbed | machine A | |

In this experiment, I am interested in seeing how page count affects BM25 query lag time. Pages vs BM25 lag time shows a similar pattern to pages vs MinDist\* lag time. However, note that BM25 is even slower. The reason for this is related to a technicality in the implementation of the BM25 algorithm: the implementation queries secure indexes twice for each query term; once for calculating the number of documents which contain a given query term, and once for calculating the frequency of the given query term per document. The implementation of the BM25 algorithm can be streamlined to only require a single frequency per query term request per document, rather than the two different queries used presently. Moreover, as mentioned in the section on caching, many of these computations can be memoized or cached. This could in practice be expected to save a significant amount of time.

As shown by Figure 16 and Figure 17, BM25 lag time for PSIB and PSIP are, to a first approximation, independent of page count.

Both MinDist\* and BM25 are scoring functions used to rank documents according to their relevance to a given query so that the results returned to the user can be ordered from most relevant to least relevant. As discussed in **Background Information** on MinDist\*, MinDist\* is most appropriately used as a way to give proximity sensitivity to established scoring measures like BM25. A straightforward way to do this, as discussed later, is to linearly combine them:

This calculation can be efficiently done with a single expensive operation used by MinDistScore—*GetLocations*. In light of convenient but inefficient implementation of BM25Score, it is plausible that this combined scoring function could be nearly as fast as *MinDistScore* by itself.

## Location Uncertainty vs BM25 MAP

Figure

Figure

|  |  |  |
| --- | --- | --- |
| Experiment Details | | |
| Input | location uncertainty |
| Output | BM25 MAP | |
| Constants | 1 secret  0 obfuscations  0.001 false positive rate  3 terms/query (Figure 16), 2 terms/query (Figure 17)  1 or 2 words/term  1000 documents (per corpus)  16 pages (per document) | |
| Testbed | machine A | |

In this experiment, I am interested in seeing how location uncertainty affects BM25 MAP. The results are not unexpected.

In Figure 16 and Figure 17, PSIP and PSIF track each other perfectly, as do PSIB and BSIB. PSIP and PSIF preserve, if desired, perfect frequency information (for unigrams and bigrams) for positive examples (false positives on negative examples are still possible), and this is independent of location uncertainty in this two secure index types.

However, PSIB and BSIB approximate a term’s frequency by counting how many of the blocks it appears in. The more blocks (the lesser the location uncertainty), the more accurately it approximates the true frequency. Indeed, in Figure 16, we see that as location uncertainty converges to 2, all of the secure indexes converge to the same BM25 score.

Also, notice that compared to Figure 16, Figure 17 is less accurate for a given location uncertainty for all of the secure index types. The reason for this has to do with the fact that Figure 16 has more terms per query, and thus when a document has all three of the terms the BM25 scoring algorithm is able to more accurately score the documents despite any frequency approximation errors. In other words, it seems to be the case that the more terms per query, potentially the less sensitive BM25 is to frequency approximation errors.

## Pages vs Boolean Lag Time

Figure

Figure

|  |  |  |
| --- | --- | --- |
| Experiment Details | | |
| Input | pages (~250 words/page) |
| Output | Boolean query lag time (milliseconds) | |
| Constants | 1 secret  0 obfuscations  256 location uncertainty  0.001 false positive rate  2 terms/query  1 or 2 words/term | |
| Testbed | machine A | |

As with every other output related to lag time, BSIB performs comparatively poorly. Boolean queries do not rank documents; a document is either relevant to the query (in this case, for a document to be relevant it must contain all of the terms in the query—a Boolean AND operation) or it is non-relevant. This is the quickest kind of query. (Indeed, the time required to perform the excessive SHA256 re-hashing operation takes up the most significant portion of time.)

PSIB, PSIP, and PSIF (PSIF is not shown but it tracks PSIP) can complete this operation in approximately 0.003 milliseconds, at least 0.0024 milliseconds of which is consumed by computing unnecessary SHA256 re-hashes. In other words, this is a ~500 nanosecond operation. This should allow a corpus of two million secure indexes to be searched by queries of this form in a second (not including the round-trip network delay).

With some of the other proposed performance enhancers, like caching and parallel computing, the simple Boolean query operation is extremely scalable. For documents of a typical length, e.g., less than 50 pages, even BSIB performs acceptably.

## Compression Ratio (Secure Index Size to Document Size) vs MinDist\* MAP

|  |  |  |
| --- | --- | --- |
| Experiment Details | | |
| Input | compression ratio: secure index size (bytes) to document size (bytes) |
| Output | MinDist\* MAP | |
| Constants | 1 secret  0 obfuscations  0.001 false positive rate  3 terms/query  1 or 2 words/term  1000 documents (per corpus) | |
| Testbed | machine A | |

In this experiment, I am interested in seeing how the ratio of the secure index size (in bytes) to the original document size (also in bytes) affects MinDist\* MAP accuracy.

PSIP has the highest unconditional MinDist\* MAP score. However, it does not come close to PSIB and BSIB in terms of having the smallest compression ratio. In fact, PSIP is a fairly constant size—the only way to change its size is by changing the false positive rate or by poisoning it. Depending on how the poisoning is done, it can be made smaller (e.g., replacing multiple positions with a single mean position in a postings list) or larger (adding positions in a postings list or adding fake terms). This is both a positive and a negative, as PSIP is nearly (in these examples) always the same fraction of the original document’s size but is consistently the highest performer on MAP accuracy and lag time.

# Background Information

## Precision and recall

Precision and recall are relevant metrics for Boolean searches; they do not rank retrieved documents like BM25 or MinDist\*; a document is either considered relevant (contains all of the terms in a query) or non-relevant.

Precision measures the proportion of retrieved documents which are relevant to the query. It is defined as:

Precision has a range of [0, 1]. Recall measures the proportion of relevant documents that were retrieved. It is defined as:

Recall also has a range of [0, 1]. It is trivial to achieve a recall of 1 (100%) by retrieving every document in the corpus. This, however, comes at the cost of decreased precision. Thus, in general, there is a trade-off between precision and recall.

## BM25

BM25 ranks documents according to their relevance to a specified query. It is a term weighting heuristic based on two fundamental insights. First, some of the terms in a query will occur more frequently in one relevant document compared to another relevant document. When scoring the relevancy of a document, if a query term appears in it frequently, the given document should be given more weight all things else being equal.

The second insight is that some query terms will be in a large proportion of the documents in the corpus. These terms, therefore, carry very little meaning—they have little discriminatory power since they appear in a large fraction of the documents. Conversely, many query terms will be very rare or even unique in a corpus, and thus they have significant discriminatory power. For example, the term “the” is in nearly every document—it serves as linguistic glue— but the term “acatalepsy” will be found in very few. The more discriminatory power a query term has, the more weight it should be given when scoring a document’s relevancy to the query.

Mathematically, these intuitions are realized using the following standard IR functions:

where Q is the query (and |Q| is the number of terms in Q), D is the document to be ranked by *BM25Score*, C is the corpus (collection of documents D is being scored against), qi is the ith term in query Q, and avgdl is the average size (in words) of documents in corpus C.

Parameters b and k1 are free parameters. In the experiments, b is set to 0.75 and k1 is set to 1.2. These are typical values, although ideally these parameters would be automatically tuned for each secure index. *TermFrequency* is a function which simply returns the number of times query term qi appears in document D. *InverseDocumentFrequency* is defined as:

where |C| is the number of documents in corpus C and *count* is a function which returns the number of documents in C which had one or more occurrences of qi.

When using BM25 to rank search results, each document D in corpus C is ranked according to query Q by the function *BM25Score.* After giving every document a BM25 rank, the results are sorted in descending order of rank as the final output.

Note that in the experiments, no standard short-cut techniques were used to avoid processing every query term in Q for every document in C.

## MinDist\*

MinDist, like BM25, ranks documents according to their proximity relevance to a given query. It is a less established ranking heuristic than BM25, but in experiments [1] it had performed well compared to other proximity heuristics. In our experiments, we add additional tunable parameters to MinDist and call it MinDist\*.

It is a proximity heuristic in which the minimum distance between each existent pair of query terms in the document is summed over. Thus, it needs location information. So, for example, if query Q = {A, B, C}, where A, B, and C are the terms of Q, and document D = “A B D D A D C”, then the minimum pair-wise distances are: (A, B) = 1, (A, C) = 2, and (B, C) = 5. The summation of these distances is simply s = 1+2+5=8. MinDist\* is a scoring function of *s*.

The intuition behind the min-pairwise summation is, the more concentrated the query terms are in a document, the more relevant the document is to the query, but only up to a certain point. For example, consider a query Q consisting of two keywords, Q = {“computer”, “science”}. Given two documents, A and B, where A contains both “hello” and “doctor” on page 7, and B contains “computer” on page 7 and “science” on the page 20. It is obvious that A should be considered much more relevant to Q since the two keywords of interest are much closer together. However, consider a third document, C, in which “hello” appears on page 7 and “science” appears on page 100. Intuitively, this is not much worse than B; both documents are simply not that relevant, and B is at best only marginally more relevant.

Mathematically, these intuitions are implemented in the following way. Let *Q* be the set of query terms, *Q’* be the subset of *Q* that exist in the given document, and *s* be the sum of the minimum pair-wise distances between terms in *Q’* as previously described.

where .

To see if this function matches our expected intuition—a strictly decreasing function that flattens out as *s* increases—it may be instructive to consider the limits and partial derivative of *MinDistScore* with respect to *s*.

As *s* converges to 0, *MinDistScore* converges to . As *s* converges to ∞, *MinDistScore* converges to . To see if these end points are the maximum and minimum values respectively, let us consider the partial derivative with respect to *s*.

This function, for all positive values of *s*, is negative. The smaller *s* is, the more negative it is; the larger *s* is, the less negative it is—it approaches as s approaches 0 and it asymptotically approaches 0 as *s* approaches ∞. This matches the desired intuition; for small *s*, a small increase in *s* corresponds to a large decrease in *MinDistScore*; and, for large *s*, a small increase in *s* corresponds to small decrease in *s.* In other words, its graph flattens out as *s* increases.

It is also reassuring to note that for large |Q’|, the function will decrease less rapidly than for small |Q’|, which is the desired behavior. Recall that |Q’| corresponds to matching more of the terms in the query. Thus, we do not wish to unduly penalize a document which contains more of the query’s terms but spread out over a larger region.

In the experiments, for the secure indexes we bind the parameters to *MinDistScore*(Q’, s; . For the canonical index (index with perfect information), all but the β parameter is the same; β has been set to instead of . This causes the canonical index’s *MinDistScore* to decrease more rapidly as *s* increases. However, since the experiments measuring MinDist\* output do not interact with any other scoring functions, like BM25, the ordering of how documents are ranked is the same for any MinDistScore binding as long as β > 0, γ > 0, and they have the same value for θ. Since this is the only aspect of the output that matters in the MinDist\* MAP experiments, this difference for β has no effect on the results.

MinDist\* is most appropriately used as a way to add proximity sensitivity to already established scoring methods, like BM25. For example, a linear combination of their scores can be used as the final output of a scoring function that is both sensitive to proximity and term frequencies:

Since training data is abundant—it is the output from the canonical index for a set of queries on a given corpus—ideally the free parameters (e.g., the parameters in MinDist\*) for each secure index would be independently optimized using a supervised learning algorithm to minimize some error measure on actual output versus the expected output. This would make for an interesting area of future research.

## Mean average precision (MAP)

MAP is a way to measure a secure index’s BM25 and MinDist\* output scores. It does this by measuring how closely its outputs matches a canonical (expected) output, as determined by a non-secure index that provides perfect location and frequency information.

The more approximately a secure index represents a document, the less information one can infer about the document from the secure index. Thus, to what extent the secure index can approximate a document while still achieving high MAP scores is an important question.

MAP is calculated by taking the mean of the average precisions on over 30 queries. The precision at k is:

The average precision for the top n documents is:  
The mean average precision (MAP) for the top n documents over Q queries is:  
Consider the following. Suppose the ranked list of relevant documents to a query is [3, 0, 1, 2, 4], and the retrieved ranked list (by a secure index) is [2, 4, 3, 0, 1]. The precision at is ; the precision at k=2 is ; the precision at k=3 is , the precision at is , and the precision at k=5 is. Thus, the average precision is . The mean average precision would simply be the mean of the average precisions for queries.

Note that the average precision for the last value of is necessarily 1 if, by that iteration of , the relevant set and the retrieved set contain the same elements. However, in general, this is not the case; for instance, if the relevant ranked list of documents to a query is (A, B), and the retrieved ranked list is (D, C, B, A), then if the mean average precision goes from to , the average precision is 0. In my simulation, I do a variation of this.

Suppose the relevant ranked list of documents to a query is (A=0.9, B=0.85, C=0, D=0), and the retrieved ranked list is (A=0.9, C=0.85, D=0.5, B=0). Then, I calculate the average precision for the top instead of the top or top . In this example, document B is not included in any of the precision at to calculations.

Finally, in one of the experiments, I conduct a “page one” MAP test, i.e., I find the mean average precision using only the top 10 results. The randomized algorithm does much more poorly in this instance, e.g., with over 85% probability, the mean average precision will be less than or equal to 0.05.

## Secure indexes

There are several different kinds of secure indexes explored in my research. With the exception of the BSIB, none of them have been proposed as far I am aware.

### PSI

**P**erfect-hash **S**ecure **I**ndex. A secure index only capable of answering approximate Boolean queries, i.e., do these terms exist in the given document?

PSI is based upon the perfect hash[[2]](#footnote-4).

1. Each unigram or bigram in the target document D is concatenated with *n* secrets—. Every unigram and bigram in the document will thus be searchable with *k* different secrets.
2. Each is cryptographically hashed—[[3]](#footnote-5)
3. Each is concatenated with the document’s identifier (e.g., hash of its filename) and is then re-hashed—. This prevents the same cryptographically hashed term in different documents from mapping to the same hash value[[4]](#footnote-7).
4. Each is uniquely hashed by a perfect hash function s.t. . That is, no collisions among any of the cryptographic hashes are possible. If using a minimum perfect hash, then
5. Let U be a bit array with at least indices, where each index maps to M contiguous bits. Then, .

Note that because an M bit hash of is stored in U rather than the actual value, false positives on non-member strings *x* are possible, i.e., if , then with probability , . This equality denotes a false positive. Therefore, PSI represents an approximate set in which false positives occur with conditional probability .

There is a different way in which a false positive may occur when processing n-gram query terms, n > 2 Consider the following. If a document contains the words “A B B D”, then the PSI will conceptually represent it as the set {“A”, “B”, “D”, “A B”, “B B”, “B D”}. To determine if “A B” exists in the document, a single set membership test will suffice.

However, determining whether the query term “A B B” exists in the document is more complicated (it does not exist in the set if only unigrams and bigrams are members). To support exact phrase searches larger than bigrams, as in the trigram “A B B”, a biword model is used in which n-gram query terms, , are decomposed into a set of *n-1* bigram tests, e.g., testing if “A B B” exists is transformed into a a conjunction of membership tests for “A B” and “B B”. If all bigrams test as positive, the n-gram term is said to exist in the document. In this case, “A B B” will correctly test positively. But if the term is “A B D”, then it will test positively both for “A B” and “B D” but nowhere in the document is the trigram “A B D” found. Therefore, this query term would cause a false positive to occur.

Notes:

1. In my PSI-based experiments, I allow for any false positive rate of the form , where M can be any positive integer. In a more practical implementation, it would be sensible to optimize the special case where M represents a byte-aligned number of bits, e.g., 8-bits or 16-bits, to take advantage of much faster parallel bit-wise operations.
2. PSI is not used in isolation in any of the experiments. Instead, PSIB, PSIF, and PSIP—which build on top of PSI—are tested. In hindsight, it is regrettable I did not include any experiments exclusively for it.
3. I also regret not including a PSI-based index more directly comparable to BSIB, i.e., a secure index which constructs a perfect hash for each block-of-word segments as is done in BSIB.

### PsiBlock (PSIB)

PSIB uses the PSI interface to, for instance, check for the existence of query terms, and on top of that provides an interface capable of answering approximate frequency and location requests for query terms.

To construct a PSIB, first a PSI is constructed. Then, the document is segmented into blocks, and a bit vector of size is constructed such that there are N bits assigned to each unigram and bigram in the document. Finally, if a term resides in a block, the corresponding index representing that block in the bit vector for the given term is set to 1. Otherwise, it is set to 0. The larger the N, the more precisely PSIB can locate terms.

Notes:

1. Bit vectors are used so that byte-alignment in memory is not necessary. For all of the PSI and PSI-based secure indexes, this approach is used to minimize the size of the secure index resident in memory without compression.
2. The bit vector representing the blocks a term resides in can become very sparse as the number of blocks increase. This can be very inefficient. However, sparse bit vectors can be easily compressed—see compressed bit vectors. I elected not to do this in my experiments, some of which do clearly reveal the need for compressed bit vectors or some other representation.   
     
   Moreover, I allow for an arbitrary number of blocks per document. However, like with the PSI, a more practical implementation could see a significant performance boost if byte-aligned sizes were used instead, e.g., parallel bit-wise AND operations could determine which blocks contain all the bigrams in a k-gram query term.

The frequency for a query term is approximated by summing the binary digits of the bit vector representing that term’s approximate block locations. Note that for unigram and bigram query terms, the pre-computed bit vector may be used, but if the term is an n-gram, , then the bit vector representing locations for the term is derived from an AND operation on the corresponding bit vector entries for all of the bigrams of the n-gram query term.

The location for a query term is approximated in much the same way as the frequency, except instead of summing the binary digits of the bit vector, a list of approximate locations is returned. For example, if each block is of size m (each block has m words, except the last which may have fewer) and query term *t* exists in and , then two locations will be returned, one in the range and the other in the range .

On false negatives  
Unlike the PSI, false negatives are possible because we use the approximate location information to eliminate positive matches that are most likely false positives. However, if an occurrence of a true positive spans two blocks, that occurrence will be eliminated by the algorithm. One solution is to check for whether the bigrams of an n-gram query term exist in adjacent blocks, e.g., for the query term “A B C”, if “A B” exists in , check for “B C” in either or . For sufficiently long query terms, a chain of adjacent blocks may also be acceptable.

### PsiFreq (PSIF)

PSIF uses the PSI interface to, for instance, check for the existence of query terms, and on top of that provides an interface capable of answering approximate frequency requests for query terms.

To construct a PSIF, first a PSI is constructed. Then, the frequency of each member (unigram and bigram) is calculated. These frequency counts are then stored in a bit vector in a memory efficient way.

If exact frequency information for positive examples of unigrams and bigrams is not desirable, then during the construction phase an approximation of the exact frequency may be used instead, e.g., .

To service frequency requests for unigrams or bigrams, the PSI interface is used to index into the PSIF’s frequency bit vector. If the query term is not a unigram or bigram, then the frequency is considered to be the minimum frequency of all of the bigrams making up the query term.

### PsiPost (PSIP)

PSIP uses the PSI interface to, for instance, check for the existence of query terms, and on top of that provides an interface capable of answering approximate frequency and location requests for query terms.

To construct a PSIP, first a PSI is constructed. Then, a postings list (a list of positions) for each unigram and bigram in the document is constructed. Finally, the positions in the postings list are randomly changed according some random variate, i.e., . In the experiments, I use a triangular distribution (with a mode equal to the exact position) instead of a uniform distribution. The triangular distribution has less variance and therefore preserves more information about location information.

Frequency requests are serviced in the same way as PSIF using the size of a term’s postings list as the frequency. Note that this means the PSIP does not exploit location information to eliminate false positives. This is done for the sake of speed.

If exact frequency information for positive examples of unigrams and bigrams is not desirable, then insert random positions into the postings lists and/or calculate the average position of adjacent positions for a term and use that average position in place of the two positions. Do this as many times as necessary to achieve the desired level of approximation, e.g., if a term appears times, repeating this times would result in storing its mean position. Of course, this also effects location accuracy.

To service location requests for unigram or bigrams, the PSI interface is used to index into the PSIP postings lists. For an n-gram query term, a greedy algorithm is used to construct non-overlapping sets each with a diameter less than or equal to some constant that depends on the way in which the postings list was changed by the random variate (e.g., when a triangular distribution is used with a maximum offset of positions in either direction, then is the diameter). The positions of the query term are taken to be the center of each such non-overlapping set.

Note that since a greedy algorithm is used, this operation is fairly quick as evidenced by experiments consisting of queries with a large terms. Moreover, the use of a more sophisticated algorithm, e.g., an algorithm which produces the maximum number of such non-overlapping sets, is not obviously an improvement in the context of greater accuracy.

### Bloom filter secure index (BSIB)

Bloom filter secure index. A secure index capable of answering approximate frequency and location requests for query terms. BSIB uses a Bloom filter rather than a PSI.

Similar to PSIB, BISB is a block-based secure index. A document is segmented into N blocks, and for each block a Bloom filter (another kind of approximate set) is constructed such that the unigrams and bigrams residing in that block are inserted into it.

The frequency for a query term is approximated by counting how many of the N blocks it tests as positive in. If the query term is a unigram or a bigram, a test for membership in the block’s Bloom filter is performed directly. Otherwise, a test for the presence of all the n-gram query’s n-1 bigrams is performed, e.g., if the query term is “A B C D”, then test for “A B”, “B C”, and “C D”. If any do not test as positive, that block does not contain the query term. Otherwise, it is said to contain it.

The location for a query term is approximated by storing which blocks it tests as positive in, and returning a random location for each such block range. For example, if each block is of size m words and query term *t* exists in blocks 0 and 5, then two locations will be returned; one in the range and the other in the range .

## Simulating an attack using query history

There are many possible ways an attacker could compromise the confidentiality of the secure indexes and the hidden queries. There is the obvious case where a secret is disclosed to a would-be attacker. Once a secret is known and the attacker has acquired authorization to query the secure indexes, it may, for instance, systematically probe the secure indexes in such a way as to try to classify the secret documents using a trigram language model.

However, I limit my attention to attacks on one of the more vulnerable parts of the system: hidden queries[[5]](#footnote-9). I consider below two general strategies to compromise hidden query confidentiality, cryptographic hash attacks and maximum likelihood attacks. I only consider cryptographic hash attacks, but I conduct experiments using maximum likelihood attacks.

### Cryptographic Hash Attacks

Cryptographic hash functions take as input an arbitrary-length string and output a fixed-length string (hash value).

In general, cryptographic hash functions have the following properties:

1. Pre-image resistance.  
     
   Given a hash string , finding a string *m* s.t. should be intractable. Lacking this property, an attacker can observe and find one or more candidate *.*  
     
   On the one hand, in the context of hidden queries, lacking pre-image resistance, an attacker may be able to discern what a target is searching for. On the other hand, effective collision resistance means that any collisions found suggests both that (a) the guessed secret is correct (if not already known) and (b) implies that the searcher was indeed looking for m.  
     
   Thus, there is a case to be made that pre-image resistance is undesirable in this context. Indeed, since most queries will consist of common terms, if the attacker knows any secrets he can hash a dictionary of common terms to discover what other users are searching for.[[6]](#footnote-10)  
     
   However, if too many collisions on legitimate queries occur, then this may have a negative effect on the accuracy of search results, like BM25 ranking of documents. So, collision resistance represents a trade-off between privacy and accuracy of search results for the kind of attack mentioned above.  
     
   I do not explore this trade-off in my experimental design, but a simple and effective approach to exploring it consists of changing the size of the fixed-length output of the hash; if a hash function maps all input to n bits, then a smaller n corresponds to a larger collision rate. For example, if there are 64 query terms which are mapped to 4 bits each, then on average (assuming a good uniform hash function) each term will collide with 64/24 = 4 other terms in the population. How this will in practice effect outputs of interest is difficult to estimate without performing experiments.
2. Collision resistance.  
     
   Finding strings m1 and m2 s.t. hash(m1) = hash(m2). A general purpose cryptographic hash function should make this kind of search infeasible, but in the context of hidden queries, it is irrelevant. That is, it does not compromise the hidden query stream nor the contents of the secure indexes if such collisions are discovered. Thus, for instance, birthday attacks on the hidden terms are not useful even if feasible.

### Maximum Likelihood Attack

Instead of mounting an attack that depends on finding collisions, I simulate a different kind of attack.

Suppose there is a sample of independent and identically distributed observations – that is, a history of query terms – coming from some distribution , where f is a pmf denoting how probable a randomly sampled query term is.

Thus, the probability of seeing a particular history of terms t1, t2, …, tn is[[7]](#footnote-11):

Each term t is mapped to a hidden term h. The objective of the attacker is to find a function which maps each hidden term h to a term .

To accomplish this goal, I simulate an attacker for which the distribution f(t) is known (and, in the simulation, is a zipf distribution); since this distribution can be estimated by examining queries in an IR system that does not hide the query terms, assuming the attacker can arrive at a reasonable approximation of the true distribution is not unreasonable.

For a given in g, the probability of seeing a particular history of hidden terms is:

To discover the most likely mapping function in g, the attacker will use maximum likelihood estimation; that is, it will explore the space of g and choose a which maximizes the probability[[8]](#footnote-12) of seeing .

Since the space of g is O(n!), a subset of the space must be explored which has a high likelihood of finding local maxima. In my simulation, I use a hill-climbing algorithm, in which the neighbors to a point in this space are defined as the interchange any two and in (in other words, a swap).

Note that an excellent initial starting point in this space, especially given a sufficient number of samples, is to collect all of the hidden terms, sort them by frequency, and pair them up to the terms t in f sorted by probability. However, I do not use this initial estimator in my simulation. If I did, this would give an even greater advantage (with respect to mitigating maximum likelihood attacks) to simulations involving multiple secrets or obfuscations.

When we add m secrets per term, i.e., hidden terms for term *t* consist of the set {h(*t*|*secret1*, *secret2*, …, *secretm*}, then g has the form s.t. each plaintext term maps to m hidden terms. Thus, the space of g is now (nm!) instead of O(n!), and it is expected that more samples of hidden terms will be needed for a given level of accuracy (where accuracy is defined as the percentage of hidden terms which have been correctly mapped).

Finally, when we add k obfuscations to the vocabulary of hidden terms (without secrets), g takes the form:

In the above function g, hn+1 to hn+k do not actually map to any plaintext term; they all map to class *obfuscation*. Thus, if a specific set of k hidden terms map to class *obfuscation*, there is only one way for each of those hidden terms to be mapped to it. The space for g, then, is ). This is equal to or larger than n! for all non-negative integer values of n and k; as a degenerate case, when k = 0 (no obfuscations), it reduces to n!.

Obfuscations introduce additional unknowns that either must be given or estimated. As with the distribution of plaintext terms being given, the probability that a random hidden term is an obfuscation term will also be given, i.e., P[obfuscation] = c, 0 < c < 1.

There are many ways to complicate matters for the attacker when dealing with obfuscations, e.g., making it so that the distribution of individual obfuscation hidden terms are similar to the distribution of non-obfuscated hidden terms. However, in the design experiment, each obfuscation term has a uniform probability.

When combining both obfuscations and secrets, the space of g is ). In any case, the space of g explodes as m or k grows (the original space was already exponential with respect to n). I only consider increasing m or k separately, i.e., if I increase k, I fix m at 1. Alternatively, if I increase m, I fix k at 0.

## Simulating an attacker reconstructing documents from secure index information

Once the attacker has successfully decoded (or has been provided with) the mapping from hidden terms to plaintext terms, it may be possible to reconstruct a document from its secure index.

A secure index contains an approximation of a document’s word[[9]](#footnote-13) frequencies without revealing which words are in the document. In addition, it may contain location information about each of the words.

If the secure index only provides approximate frequency information, then a line of attack may consist of the following steps. First, sample from a distribution to automatically determine some fraction of the unigrams or bigrams in the given document and their respective frequencies (multiplicities). Then, using a bag-of-words model, classify the document, e.g., P[medical document | word histogram]. More specific classes are possible also. For example, a single plaintext document can represent a class.

However, more sophisticated attacks may be done. For instance, note that bigrams are more informative feature classifiers than unigrams. Moreover, trigrams are more informative than bigrams. Indeed, the larger the n-gram, the more informative it may be as a feature. The limiting case for this is an n-gram the size of an entire plaintext document. Finding a match on this in a secure index would indeed be very informative.

So, with this insight as motivation, an attacker could use a secure index’s bigrams in conjunction with a language model to partially reconstruct a document from the information contained its secure index. First, the attacker can use a generative language model, like the trigram language model, to probe the secure index for plausible n-grams. For instance, if the attacker finds a positive hit on the bigram “A B”, this information can be used to generate plausible trigram phrases, e.g., sample word *x* from the conditional distribution, P[*x* | “A”, “B”]. If “C” is plausible given the previous two words were “A” and “B”, then check for a hit on the trigram phrase “A B C”. If this this trigram tests positively in the secure index, then generate and test plausible 4-gram phrases by sampling from the distribution, P[x | “B”, “C”].

Repeating the above steps over and over again, a large set of n-gram phrases (that test positive in the secure index) may be constructed. Furthermore, some of the discovered n-grams may overlap in some way, in which case the attacker can automatically stitch the pieces together in various ways. The plausibility of a stitching can be estimated using the language model, especially if multiple consistent stitchings of the same size are possible.

Already, this may reveal a significant amount of details about the document. However, if the secure index also provides location information, the attacker has a much easier job.

If exact location information is provided for each word, then as the attacker finds words as previously described, he puts them in their proper place. As words are placed, larger and larger n-grams are constructed, and the language model can be used to generate plausible candidates for the missing words. Or, since the problem has been vastly simplified, the attacker may choose to exhaustively check every word in a dictionary.

Since false positives are possible, each position may have multiple candidates. To deal with this eventuality in a reasonably straightforward way, the attacker can find an assignment of candidates that tries to maximize the likelihood (given a language model) of a given assignment of candidates to each position.

Given the likely effectiveness of this attack, the reported positions for a word should have some degree of uncertainty—e.g., only reporting that a word falls within some range (block), as PSIB and BSIB do, or scrambling the positions in some random way, as PSIP does.

### Problems with the block-based approach used by PSIB and BSIB

The block-based approach used in PSIB and BSIB reduces the problem to treating each block as a small document, and solving each one independently without location information using the techniques described at the beginning of this section. Since the document is much smaller, the reconstruction effort will be significantly easier than trying to do this for the entire document.

To paint a clearer picture, if the document consists of N words, and the words are segmented into k blocks, then there are n = N/k words per block. If all n words in the block are discovered (and ignoring word multiplicities), then there are n! = N!/k! ways to order them. So, for each permutation, the attacker calculates its likelihood given the chosen language model (e.g., trigram language model), and saves the permutations with the highest likelihoods. Note that the space of permutations is a factor 1/k! of the original space; for sufficiently large k the search space may even be exhaustively explored.

### An alternative solution that overcomes many of the problems for the block-based approach

A significant problem with the block-based approach is the ability for the attacker to treat each block as a separate, independent problem—this has the effect of reducing the attacker’s curse of dimensionality. PSIP is designed, in part[[10]](#footnote-14), to overcome this problem. In PSIP, there is no such block delineation—instead, words are offset from their true position according to some random variate. This makes it harder to treat the document as a set of smaller independent problems.

For instance, suppose document D = “A B C D E F G H” and to simply matters suppose we can swap any word in D with any other word in D as long as the words final position is within two units of its starting position. Then, let scrambled document D’ = “B A E D C G F H”. Is it possible to break this larger problem down into two smaller independent problems?

d1’ = “B A E D” and d2’ = “C G F H” will not work since, in the original document, the first set should contain elements from {A, B, C, D}, but it is missing B and has an additional E. Any 4-gram ordering on these two sub-problems cannot match the ordering in the original document; indeed, in this case, the only sub-ordering that matches the original ordering is “A B” and “F G H”. These two sets cannot be stitched together since they have no overlapping components.

Another possible division is d1’ = “B A”, d2’ = “E D C”, and d3’ = “G F H”. This is a legitimate way to reduce the larger problem into a set of smaller independent problems, but the attacker has no way of knowing this beforehand. For instance, if instead A had been swapped with C, this would no longer be a legitimate partition.

This, I would argue, will significantly blow up the search space for the attacker. The attacker may still use the location information to do things like eliminate impossible stitchings, but it is more difficult to use the location information to create independent sub-problems. Of course, it may be acceptable to reduce the original document into sub-problems with a size dependent upon the location uncertainty and settle for more erroneous solutions. It is also possible to parameterize the segmentation points and include those as additional parameters to optimize, but this has the effect of blowing up the search space.

### On the effect of false positives

As shown in the previous section, false positives create a problem for a hypothetical attacker attempting to (at least partially) reconstruct a document from the information in its secure index.

Given a secure index of with N words, each unique (in order to simplify the discussion), there are permutations. The attacker wishes to find some words (which will test as positive in the index) and then find a permutation that maximizes the likelihood of observing that sequence of words given a chosen language model.

An exact solution is already computationally intractable—. Adding false positives complicates matters even more for the attacker, although it is still in . Suppose false positives occur at a rate of , and the attacker wishes to perform an exhaustive search on the secure index by iterating through a dictionary consisting of words, where the words in the document are a subset of the words in the dictionary. Then, to find the words in the document, of those words from the dictionary will necessarily be true positives and it is expected that there will be false positives.

In total, it is expected that words will positively match. Each one of these words is a candidate, and thus instead of the attacker needing to explore a space consisting of possibilities, the attacker must explore a space of possibilities. The degenerate case evaluates to , but as grows it quickly diverges from for a given .

Thus, we see that on the one hand, a high false positive rate mitigates reconstruction attacks. On the other hand, as some of the experiments were designed to probe, a high false positive rate may cause the searching apparatus to return unacceptably poor results if it is affected unduly by the false positive hits. Ultimately, this is trade-off; the least disclosing is , and the most accurate (for searching) is .

### Secure index poisoning

The intent of poisoning a secure index is to make it so that the hypothetical attacker will be less successful at (at least partially) reconstructing a document from the information in its secure index.

Experiments have explored poisoning in three general ways:

1. Inserting fake terms (unigrams and bigrams) into a secure index. They seem to have a similar effect to increasing the false positive rate, but unlike increasing the false positive rate, it can be done in a way that should theoretically not affect search accuracy (as experimentation has demonstrated).  
     
   Poisoning often represents a trade-off between accuracy and space efficiency (and time, for some secure indexes). However, perfect hash-based secure indexes may be able to accomplish this at very little (if any) cost since the choice between having fake terms and not having fake terms is essentially a choice between using a minimum perfect hash or a perfect hash. A perfect hash can be far more space efficient—indeed, in hindsight, by default a perfect hash is preferable to the minimum perfect hash even if poisoning is not desired.
2. Adjusting frequencies. For a given term, alter the secure index s.t. the frequency is altered. See secure indexes on page 27.
3. Adjusting locations. For a given term, alter the secure index s.t. the locations are altered. See secure indexes on page 27.

## Query confidentiality – additional measures

In my experimental design, I only entertain query confidentiality as it goes from the user to the server. Note, however, that the response from the server may also reveal (leak) information about the user and the document.

To mitigate information leaks about what the search user may be interested in, we can look to Oblivious RAM for inspiration. For instance, instead of the user’s search client sending a single query for each query a user is interested in, the client may submit a packet of N queries, only one (or even none) of which is related to the user’s actual query. The N-1 fake queries are strictly intended to obfuscate the user’s actual query of interest, much like the query obfuscation discussed elsewhere is designed to obfuscate the actual terms of interest in a single query.  
  
The N-1 other queries in the query packet may consist of terms chosen from a distribution that is likely to result in a plausible distribution of hits (i.e., relevant to a large set of documents), e.g., sample the queries from a plausibly realistic Zipf distribution. In this way, it may be impossible for an attacker to infer which queries are actually of interest to the user. But, of course, this comes at the cost of increased resource consumption, e.g., the server must process more queries per legitimate query thus increasing processing and network transmission requirements.

## Letter n-grams and word n-grams

In my experiments, I use word unigrams and word bigrams as the atomic, indivisible units in my secure indexes. However, this is not necessary; I could go in either direction, supporting larger word n-grams (e.g., inserting trigrams) to support faster searching on larger phrases, or go in the opposite direction and support units smaller than whole words using letter n-grams.

Consider the string "hello world". If storing letter trigrams, then the following transformation takes place, where \* denotes whitespace.

To search for the word “hello”, check for the existence of "hel", "ell", and "llo" in the set. If all three letter trigrams exist, that word is said to exist. As in the biword model, false positives are possible.

With letter n-grams, partial word matches are automatically possible. For instance, if the user wishes to find any words matching “ello”, then simply check for the existence of “ell” and “llo”. In fact, any substring that is three characters or larger can be matched.

## Wild-card searching

It is also possible to support wild-card searching, e.g., match any pattern “hello \* doctor”, where \* represents a word wildcard. To support word wild-cards, In addition to inserting word unigrams and word bigrams, strings of the form “first \* third” must also be inserted, i.e., skip the word in the middle. For example:

This increases the size of the secure index fractionally while facilitating rapid word wildcard searching. This can, at increasing cost to size, be extended to support multiple word wildcards, e.g., “hello \* \* doctor”. Character wildcards can be supported by letter n-grams in the same way that word wildcards are supported by word n-grams. All of this, of course, may inflate the secure index beyond what is desirable.

## Approximate searching and error tolerance

Approximate searching is a superset of wild-card searching. Other ways approximate searching can be accommodated include hashing the terms of the document in such that like or similar terms hash to the same value, e.g., locality sensitive hashing, phonetic coding (e.g., Soundex, which tends to map words that sound alike to the same hash), or stemming (reducing words to their stem or root form, e.g., {“computer”, “computing”, “computes”, “computed”, “compute”, “computation”, “computations”, “computational”} all map to the same stem, “comput”.

These kind of transformations generally take place during a preprocessing step so that the secure index constructor only sees the final forms of terms. Note that there is nothing preventing a preprocessor from including any combination of these transformations.

Another form of approximate searching can be accomplished through online query expansion, e.g., finding synonyms for words, fixing spelling errors, and so forth. However, these tricks—stemming, synonym expansion, etc.—must be used with caution lest they come at too high a cost in terms of trading precision for recall. Ideally, both recall and precision will be improved, since documents that are relevant to the user’s information need (represented by the query) will be retrieved where otherwise they would not have been.

## Boolean proximity searching

In our experiments, we use MinDist\* as a way to rank documents according to a minimum pair-wise distance metric. However, another perhaps more useful—and far more straightforward—way to use the proximity information in secure indexes is to require that all of the terms in a query be within a maximum proximity of each other. This is an extension of Boolean searching in which all of the keywords and phrases in a query must be in the document.

An algorithm to enable this functionality has already been essentially implemented (for the MinDist\* heuristic). It is less complicated (both conceptually and computationally) than the MinDist\* scoring function. Furthermore, MinDist\* and/or BM25 can be used in tandem with Boolean proximity requirements s.t. only those documents which contain all the terms in the query within a maximum proximity are ranked and everything else is excluded.

## Caching results

Caching previously calculate results would result in significant savings, e.g., whenever a query is mapped to a list of ranked documents, store the mapping in a cache; or, whenever query term *t*’s BM25 inverse document weight is computed, store the mapping in a cache. I had initially planned on using an LRU cache to memoize computations like the above, but then I reasoned that I wanted to actually test the outputs of the various secure indexes without the effect of a cache.  
  
In a practical implementation, a simple LRU cache or some other memorization technique could be used to avoid recalculating results. Since queries will be heavily biased towards a small subset of terms, this would result in significant savings. Since the secure index database server could do this without the user’s permission,

Definition of a query

A query represents an *information need*. In practice, it a string of terms, where a term is either a keyword or an exact phrase surrounded by quotes. Consider the following query: <QUERY BEGIN>“doctors without borders” volunteer<QUERY END>. This query consists of two terms: the keyword “volunteer” and the exact phrase “doctors without borders”. When conducting a search on a secure index, it will look for that exact phrase and that exact keyword. It will not count “doctors with borders” as a hit since the phrase must exactly match.

So, each query consists of one or more terms and each term consists of one or more words. The actual secure index itself only stores representations of the unigrams (keywords) and bigrams found in the document it represents. Thus, any phrase term larger than two words must be converted into a chain of bigrams. More details on this are provided elsewhere.

## Experiment Platforms

|  |  |
| --- | --- |
| *Machine A* |  |
| Operating System | Windows 7 Service Pack 1 |
| Processor | AMD A6-6400K APU 3.9GHz |
| Installed memory (RAM) | 8.00 GB |
| Storage Device | Kingston SSDNow V300 Series SV300S37A/60G 2.5" 60GB SATA III SSD |
| Compiler | Visual Studio 2013; 32-bit target; command line = ” /MP /GS /GL /W3 /Gy /Zi /Gm- /O2 /fp:precise /GF /GT /WX- /Zc:forScope /arch:SSE2 /Gd /Oy- /Oi /MD” |

1. A way of storing small integers in fewer bytes. This is sort of a cheat, since in general I attempted to ensure size(secure index data structure in memory) ~ size(secure index serialization on disk), but a varint is converted into an unsigned integer once loaded into memory. However, the cheat is justified in that it is designed to mimic a more efficient solution that does meet the size(serialization) = size(secure index data structure) objective. [↑](#footnote-ref-1)
2. A perfect hash function for a set S maps distinct elements in S to a set of integers with no collisions. [↑](#footnote-ref-4)
3. Default implementation uses SHA256 and non-invertibly maps the hashes to N hexadecimal digits. [↑](#footnote-ref-5)
4. Limits statistical inference to sampling from a single secure index rather than an entire corpus of secure indexes since each secure index has a unique and random way of mapping its unigrams and bigrams to hashes. [↑](#footnote-ref-7)
5. It is one of the more vulnerable parts of the system because it is more susceptible to substitution cipher attacks; the secure index itself has many safe-guards against such attacks: (1) each secure index uniquely hashes its members with a salt such that the same unigram or bigram in one secure index will look completely different than in any other, (2) each Psi-based secure index maps every string in the universe of strings to a much small bit string (e.g., 10 bits), (3) frequency and location information is only approximate, and (4) entries in the secure index may be fake. [↑](#footnote-ref-9)
6. Note that this assumes the attacker has access to the hidden query stream. If the hidden query steam is taking place over a secret channel, the secure index server and the attacker must share information to make this an effective kind of attack. [↑](#footnote-ref-10)
7. Since order is irrelevant (i.i.d. distribution), the actual probability is , where ki represents how many times ti appears in the query history set. However, is a constant for a given history of n and k, so we can safely ignore it in our maximum likelihood attack. [↑](#footnote-ref-11)
8. When simulating the attacker, the log of the maximum likelihood will be used instead. [↑](#footnote-ref-12)
9. Actually, a secure index using a biword search model stores the unigrams (words) and bigrams in a document. [↑](#footnote-ref-13)
10. It is also designed to provide more accurate location information by allowing the mean error of approximate word positions to be 0—i.e., PSIP changes (or at least it is possible) each words position about its true mean. [↑](#footnote-ref-14)