1 Signature Schemes for the Join Size

First we study the obvious signature scheme we obtain from randomly selecting a small number of elements. Let the number of possible item types be \( n \), and assume that our sample has size \( r \). See what we get if we estimate the join size of two data bases, \( F \) and \( G \) by \((n/r)^2 \sum f'_i g'_i\), where \( n \) is the size of each data base.

Think of the items in \( F \) and \( G \) as nodes in the two sides of a bipartite graph \( \Gamma \). We connect two nodes iff they have the same type. Then \(|E(\Gamma)| = J(F,G)\), where \( J(F,G) \) is the join size of \( F \) with \( G \). Assume that we select a random sample items from \( F \) and \( G \) by picking elements randomly from both data bases with probability \( p = r/n \). The join size of the samples equals to the number of edges spanned in \( \Gamma \) by the samples.

Lemma 1 Let \( \Gamma \) be any graph on \( n \) nodes. Assume we select nodes of \( \Gamma \) randomly, each with probability \( p \). Let \( X \) denote the random variable whose value is the number of edges that are spanned by the nodes in the sample. Then:

\[
E(X) = |E|p^2
\]

\[
Var(X) \leq |E|p + \sum_{i=1}^{n} \frac{d_i^2}{2} p^3,
\]

where \( d_i \) is the degree of node \( i \).

Since \( \sum_{i=1}^{n} \frac{d_i^2}{2} \leq |E|n \) we can estimate \( Var(X) \) in Lemma 1 by \( O(|E|np^3) \).

The error of the sampling method is tolerable if the variance is smaller than the square of the expected value (since in this case we can use the Chebyshev inequality to bound the error we make in the estimate.) We must have \( |E|np^3 \leq |E|^2p^4 \), i.e., \( p \geq n/|E| \). This shows that a sample of size \( n^2/J(F,G) \) is sufficiently large. We conclude:

Lemma 2 Suppose we have an a priory lower bound \( B \) on the join size. The naive sampling signature scheme estimates the join size with constant relative error with high probability if our random sample has size at least \( n^2/B \).
2 Lower Bounds on Signature Schemes for the Join Size

We prove that the naive algorithm in the previous section cannot be improved with no further assumptions.

Theorem 3 Let \( S \) be any scheme which assigns bit strings to data bases, so that there is a random or deterministic pairing function \( D \) such that given two data bases \( F \) and \( G \) the formula \( D(S(F), S(G)) \) gives a good estimate on the join size of \( F \) and \( G \) with high probability, when an a-priory lower bound \( B \geq 10n \) is given on the join size. Then the length of the bit string that \( S \) assigns to data bases of size \( n \) must be at least \( n^2/B \).

We use a standard lower bound technique developed by Yao for a wide range of randomized models. Define \( t = 10n^2/B \) and fix a set \( T \) of \( t \) types. Let \( D_1 \) be the uniform probability distribution on trivial data bases over \( T \), that is, with probability \( 1/t \) we take the data base all whose items are of type \( i \), where \( 1 \leq i \leq t \). We define another distribution \( D_2 \) in the following way: Let \( S \) be a family of subsets of \( \{1, 2, \ldots, t\} \) such that:

- All sets in \( S \) have size \( n^2/B = t/10 \)
- \( |S| = 2n^2/B = 2^{t/10} \)
- For all \( S_1, S_2 \in S \), \( S_1 \neq S_2 \) we have \( |S_1 \cap S_2| \leq n^2/2B = t/20 \).

One can show the existence of such a set system using the probabilistic method. We define \( D_2 \) as the uniform distribution on data bases of the following type: Let \( S \in S \). For every color in \( S \) the data base contains \( B/n \) items of that color. In addition, to make sure all our join sizes are at least \( B \), we add to each data base the item 0, appearing \( \sqrt{B} \) times. (Thus the total length of each data base is \( n + O(\sqrt{n}) \) and the total number of items is at most \( n + 1 \); the facts that these numbers are not precisely \( n \) is clearly not crucial here). Let \( F \) be a data base randomly chosen from \( D_1 \) and let \( G \) be a data base randomly chosen from \( D_2 \). The join size of \( F \) and \( G \) is either \( B \) or \( B + n(B/n) = 2B \). Applying Yao's technique it is enough to show that any deterministic signature scheme which assigns strings of length at most \((n^2/B) - 1 \) estimates the join size incorrectly with probability bounded away
from 0 for a random pair \( F \in D_1, G \in D_2 \). Let us put data bases into classes according to the signature they get. Since the number of different signatures is at most \( 2^{n^2/B^2} / 2 \), and in each class there can be at most one element in the support of \( D_2 \) for which the answer is right for more than 95% of the elements from \( D_1 \), we get that for a constant fraction of the pairs \( F \in D_1, G \in D_2 \) the algorithm must fail.

3 The Tug of War Signature Scheme

We propose a signature scheme, called the Tug of War, which gives an estimator for the join size which (with high probability) makes an error at most \( \sqrt{2S_F^2 S_G^2} \), where \( S_F^2 \) and \( S_G^2 \) are the self join sizes of \( F \) and \( G \). If the join size is relatively large compared to the self join sizes of \( F \) and \( G \), our estimator works well. A particularly nice feature of the Tug of War scheme is its space efficiency: it requires only \( \log N \) space for data bases of length \( N \).

Let \( F \) and \( G \) be two data bases of size \( N \) each with frequencies \( f_i, g_i \) \((1 \leq i \leq n)\). The join size \( J(F, G) = \sum_{i=1}^n f_i g_i \).

Let \( \{\epsilon_i\}_{i=1}^n \) be four-wise independent \( \{-1, 1\} \)-valued random variables. For \( F \) and \( G \) we create the signatures \( S(F) = \sum_{i=1}^n \epsilon_i f_i, S(G) = \sum_{i=1}^n \epsilon_i g_i \).

The estimator for \( J(F, G) \) is simply \( S(F)S(G) \).

**Theorem 4**

\[
\begin{align*}
E(S(F)S(G)) &= J(F, G), \\
\text{Var}(S(F)S(G)) &\leq 2S_F^2 S_G^2,
\end{align*}
\]

where \( S_F^2 \) and \( S_G^2 \) are the self join sizes of \( F \) and \( G \).

**Proof:**

\[
E(S(F)S(G)) = E(\sum_{i=1}^n f_i g_i + \sum_{1 \leq i \neq j \leq n} \epsilon_i \epsilon_j f_i g_j) = \sum_{i=1}^n f_i g_i = J(F, G),
\]

since \( E(\epsilon_i \epsilon_j) = 0 \) for \( 1 \leq i \neq j \leq n \). To prove Equation (4) define \( X = S(F)S(G) - E(S(F)S(G)) = \sum_{1 \leq i \neq j \leq n} \epsilon_i \epsilon_j f_i g_j \). We have:

\[
\text{Var}(S(F)S(G)) = E(X^2) = \sum_{1 \leq i \neq j \leq n} f_i^2 g_j^2 + \sum_{1 \leq i \neq j \leq n} f_i g_i f_j g_j.
\]
Now from

\[ \sum_{1 \leq i \neq j \leq n} f_i^2 g_i^2 = \sum_{1 \leq i \leq n} f_i^2 \sum_{1 \leq i \leq n} g_i^2 - \sum_{1 \leq i \leq n} f_i^2 g_i^2; \quad (7) \]

\[ \sum_{1 \leq i \neq j \leq n} f_i g_i f_j g_j = \left( \sum_{1 \leq i \leq n} f_i g_i \right)^2 - \sum_{1 \leq i \leq n} f_i^2 g_i^2 \]

\[ \leq \sum_{1 \leq i \leq n} f_i^2 \sum_{1 \leq i \leq n} g_i^2 - \sum_{1 \leq i \leq n} f_i^2 g_i^2; \quad (8) \]

and Equation (6) we get that

\[ \text{Var}(S(F)S(G)) \leq 2 \sum_{1 \leq i \leq n} f_i^2 \sum_{1 \leq i \leq n} g_i^2 - \sum_{1 \leq i \leq n} f_i^2 g_i^2 \leq 2S_F^2 S_G^2. \]