

# Approximately Detecting Duplicates for Streaming Data using Stable Bloom Filters

Fan Deng  
*University of Alberta*  
fandeng@cs.ualberta.ca

Davood Rafiei  
*University of Alberta*  
drafie@cs.ualberta.ca

## ABSTRACT

Traditional duplicate elimination techniques are not applicable to many data stream applications. In general, precisely eliminating duplicates in an unbounded data stream is not feasible in many streaming scenarios. Therefore, we target at approximately eliminating duplicates in streaming environments given a limited space. Based on a well-known bitmap sketch, we introduce a data structure, Stable Bloom Filter, and a novel and simple algorithm. The basic idea is as follows: since there is no way to store the whole history of the stream, SBF continuously evicts the stale elements so that SBF has room for those more recent ones. After finding some properties of SBF analytically, we show that a tight upper bound of false positive rates is guaranteed. In our empirical study, we compare SBF to alternative methods. The results show that our method is superior in terms of both accuracy and time efficiency when a fixed small space and an acceptable false positive rate are given.

## 1. INTRODUCTION

Eliminating duplicates is an important operation in traditional query processing, and many algorithms have been developed [22]. A common characteristic of these algorithms is the underlying assumption that the whole data set is stored and can be accessed if needed. Thus, multiple passes over data are possible, which is the case in a traditional database scenario. However, this assumption does not hold in a new class of applications, often referred to as data stream systems [3], which are becoming increasingly important. Consequently, detecting duplicates precisely is not always possible. Instead, it may suffice to identify duplicates with some errors.

While it is useful to have duplicate elimination in a Data Stream Management System (DSMS)[3], some new properties of these systems make the duplicate detection problem more challenging and to some degree different from the one in a traditional DBMS. First, the timely response property of data stream applications requires the system to respond

quickly in real-time. There is no choice but to store the data in limited main memory rather than in huge secondary storage. Sometimes even main memory is not fast enough. For example, for network traffic measurement and accounting, ordinary memory (DRAM) is too slow to process each IP packet in time, and fast memory (on-chip SRAM) is small and expensive [19, 5].

Second, the potentially unbounded property of data streams indicates that it is not possible to store the whole stream in a limited space. As a result, exact duplicate detection is infeasible in such data stream applications.

On the other hand, there are cases where efficiency is more important than accuracy, and therefore a quick answer with an allowable error rate is better than a precise one that is slow. Sometimes there is no way to have a precise answer at all. Therefore, load shedding is an important topic in data stream system research [33, 4]. Next, we provide some motivating examples.

### 1.1 Motivating Examples

**URL Crawling.** Search engines regularly crawl the Web to enlarge their collections of Web pages. Given the URL of a page, which is often extracted from the content of a crawled page, a search engine must probe its archive to find out if the URL is in the engine collection and if the fetching of the URL can be avoided [8, 24].

One way to solve the problem is to store all crawled URLs in main memory and search for a newly encountered URL in it. However, the set of URLs can be too large to fit in memory. Partially storing URLs in the secondary storage is also not perfect because of the large volume of searches that is expected to be performed within a time unit.

In practice, detecting duplicates precisely may not be indispensable. The consequence of an imprecise duplicate detection is that some already-crawled pages will be crawled again, or some new URLs which should be crawled are missed. The first kind of errors may lead the crawler to do some redundant crawling. This may not have a great influence on performance as long as the error rate is not high. For the second kind of errors, since a search engine can archive only a small portion of the entire web, a small miss rate is usually acceptable. In addition, if a missed URL refers to a high quality page, it is quite likely that the URL will be listed in the content of more than one crawled page, and there is less chance of missing it again. The solution adopted by the Internet Archive crawler introduces the second kind of errors[24].

**Selecting distinct IP addresses.** In network monitoring and accounting, it is often important to understand

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

SIGMOD 2006 June 27–29, 2006, Chicago, Illinois, USA  
Copyright 2006 ACM 1-59593-256-9/06/0006 ...\$5.00.

the traffic and users on the network [25]. The following two queries, for example, may be interesting to network monitors: who are the users on the network within past one hour? Where do they go? The queries may be written as:

```
Select distinct Source/Destination IP
from IP_PacketsStream
Within Past 1 Hour
```

The result could be helpful for further analyzing the user profiles, interests and the network traffic. Because of the high throughput of Internet routers and limited amount of fast memory, currently it is hard to capture per package information precisely, and sampling is often used as a compromise [18]. We are not sure if sampling is good for this duplicate elimination problem.

**Duplicate detection in click streams.** Recently, Metwally et al. propose another application for approximate duplicate detection in a streaming environment [29]. In a Web advertising scenario, advertisers pay web site publishers for clicks on their advertises (or links). For the sake of profit, it is possible that a publisher fakes some clicks (using scripts), hence a third party, called an advertising commissioner, may want to detect those false clicks by monitoring duplicate user IDs. We discuss more about their work in the related work section.

## 1.2 Our Contributions

In this paper, we propose *Stable Bloom Filter (SBF)*, which extends and generalizes the regular Bloom filter, and accordingly, a novel algorithm which dynamically updates the sketch to represent recent data. We find and prove the stable properties of an SBF including stability, exponential convergence rate and monotonicity, based on which we show that using constant space the chance of a false positive can be bounded to a constant independent of the stream size, and this constant is explicitly derived. Furthermore, we show that the processing time of SBF for each element in the stream is also constant independent of the stream size. To make our algorithm readily applicable in practice, we provide detailed discussions of the parameter setting issues both in theory and in experiments. And we compare our method to alternative methods using both real and synthetic data. The result shows that our method is superior in terms of both accuracy and time efficiency when a fixed small space and an acceptable false positive rate are given.

## 1.3 Roadmap

In Section 2, we present the problem statement and some background on existing approaches. Our solution is presented and discussed in Section 3. In Section 4, we discuss how our algorithm can be used in practice. In Section 5, we verify our theoretical findings, experimentally evaluate our method and report the results of our comparisons to those from alternative methods. Related work is reviewed in Section 6, and conclusions and future work are discussed in Section 7.

## 2. PRELIMINARIES

This section presents the problem statement in a data stream model and some possible solutions.

### 2.1 Problem Statement

We consider a data stream as a sequence of numbers, denoted by  $S_N = x_1, \dots, x_i, \dots, x_N$ , where  $N$  is the size of the stream. The value of  $N$  can be infinite, which means that the stream is not bounded. In general, a stream can be a sequence of records, but it is not hard to transform each record to a number (e.g., using hashing or fingerprinting) and use this stream model.

Our problem can be stated as follows: given a data stream  $S_N$  and a certain amount of space,  $M$ , estimate whether each element  $x_i$  in  $S_N$  appears in  $x_1, \dots, x_{i-1}$  or not. Since our assumption is that  $M$  is not large enough to store all distinct elements in  $x_1, \dots, x_{i-1}$ , there is no way to solve the problem precisely. Our goal is to approximate the answer and minimize the number of errors, including both false positives and false negatives, where a false positive is a distinct element wrongly reported as duplicate, and a false negative is a duplicate element wrongly reported as distinct.

To address this problem, we examine two techniques that have been previously used in different contexts.

### 2.2 The Buffering Method

A straightforward solution is to allocate a buffer and fill the buffer with enough elements of the stream. For each new element, the buffer can be checked, and the element may be identified as distinct if it is not found in the buffer, and duplicate otherwise. When the buffer is full, a newly arrived element may evict another element out of the buffer before it is stored. There are many replacement policies for choosing an element to be dropped out (e.g. [22]). Clearly, buffering introduces no false positives. We will use this method in our experiments and compare its performance to that of our method.

### 2.3 Bloom Filters

Bloom [7] proposes a synopsis structure, known as the Bloom filter, to approximately answer membership queries. A Bloom filter,  $BF$ , is a bit array of size  $m$ , all of which are initially set to 0. For each element,  $K$  bits in  $BF$  are set to 1 by a set of hash functions  $\{h_1(x), \dots, h_K(x)\}$ , all of which are assumed to be uniformly independent. It is possible that one bit in  $BF$  is set multiple times, while only the first setting operation changes 0 into 1, and the rest have no effect on that bit. To know whether a newly arrived element  $x_i$  has been seen before, we can check the bits  $\{h_1(x_i), \dots, h_K(x_i)\}$ . If any one of these bits is zero, with 100% confidence we know  $x_i$  is a distinct element. Otherwise, it is regarded as a duplicate with a certain probability of error. An error may occur because it is likely that the cells  $\{h_1(x_i), \dots, h_K(x_i)\}$  are set before by elements other than  $x_i$ .

The probability of a false positive (*false positive rate*)  $FP = (1 - p)^K$ , where  $p = (1 - 1/m)^{Kn}$  is the probability that a particular cell is still zero after seeing  $n$  distinct elements. It is shown that when the number of hash functions  $K = \ln(2)(m/n)$ , this probability will be minimized to  $(1/2)^{\ln(2)(m/n)}$ , where  $m$  is the number of bits in  $BF$  and  $n$  is the number of distinct elements seen so far [30].

## 3. STABLE BLOOM FILTERS

The Bloom filter is shown to be useful for representing the presence of a set of elements and answering membership queries, provided that a proper amount of space is allocated according to the number of distinct elements in the set.

### 3.1 The Challenge to Bloom Filters

However, in many data stream applications, the allocated space is rather small compared to the size of the stream. When more and more elements arrive, the fraction of zeros in the Bloom filter will decrease continuously, and the false positive rate will increase accordingly, finally reaching the limit, 1, where every distinct element will be reported as a duplicate, indicating that the Bloom filter is useless.

Our general solution is to avoid the state where the Bloom filter is full by evicting some “elements” from it before the error rate reaches a predefined threshold. This is similar to the replacement operation in the buffering method, in which there are several possible policies for choosing a past element to drop. In many real world data stream applications, often the recent data is more important than the older data [13, 31]. However, for the regular Bloom filter, there is no way to distinguish the recent elements from the past ones, since no time information is kept. Accordingly, we add a random deletion operation into the Bloom filter so that it does not exceed its capacity in a data stream scenario.

### 3.2 Our Approach

To solve this problem, we introduce the *Stable Bloom Filter*, an extension of the regular Bloom filter.

**DEFINITION 1** (STABLE BLOOM FILTER (SBF)). *An SBF is defined as an array of integer  $SBF[1], \dots, SBF[m]$  whose minimum value is 0 and maximum value is  $Max$ . The update process follows Algorithm 1. Each element of the array is allocated  $d$  bits; the relation between  $Max$  and  $d$  is then  $Max = 2^d - 1$ . Compared to bits in a regular Bloom filter, each element of the SBF is called a **cell**.*

Concretely speaking, we change bits in the regular Bloom filter into cells, each consisting of one or more bits. The initial value of the cells is still zero. Each newly arrived element in the stream is mapped to  $K$  cells by some uniform and independent hash functions. As in a regular Bloom filter, we can check if a new element is duplicate or not by probing whether all the cells the element is hashed to are non-zero. This is the duplicate detection process.

After that, we need to update the SBF. We first randomly decrement  $P$  cells by 1 so as to make room for fresh elements; we then set the same  $K$  cells as in the detection process to  $Max$ . Our symbol list is shown in Table 1, and the detailed algorithm is described in Algorithm 1.

### 3.3 The Stable Property

Based on the algorithm, we find an important property of SBF both in theory and in experiments: after a number of iterations, the fraction of zeros in the SBF will become fixed no matter what parameters we set at the beginning.

We call this the *stable property* of SBF and deem it important to our problem because the false positive rate is dependent on the fraction of zeros in SBF.

**THEOREM 1.** *Given an SBF with  $m$  cells, if in each iteration, a cell is decremented by 1 with a probability  $p$  and set to  $Max$  with a probability  $k$ , the probability that the cells becomes zero after  $N$  iterations is a constant, provided that  $N$  is large enough, i.e.*

$$\lim_{N \rightarrow \infty} Pr(SBF_N = 0)$$

*exists, where  $SBF_N$  is the value of the cell at the end of iteration  $N$ .*

---

#### Algorithm 1: Approximately Detect Duplicates using SBF

---

**Data:** A sequence of numbers  $S = x_1, \dots, x_i, \dots, x_N$ .  
**Result:** A sequence of “Yes/No” corresponding to each input number.

```

begin
  initialize SBF[1]... SBF[m] = 0
  for each  $x_i \in S$  do
    Probe  $K$  cells  $SBF[h_1(x_i)] \dots SBF[h_K(x_i)]$ 
    if none of the above  $K$  cells is 0 then
      DuplicateFlag = “Yes”
    else
      DuplicateFlag = “No”
    Select  $P$  different cells uniformly at random
       $SBF[j_1] \dots SBF[j_P], P \in \{1, \dots, m\}$ 
    for each cell  $SBF[j] \in \{SBF[j_1], \dots, SBF[j_P]\}$ 
      do
        if  $SBF[j] \geq 1$  then
          DuplicateFlag = “Yes”
           $SBF[j] = SBF[j] - 1$ 
        for each cell  $\in \{SBF[h_1(x_i)], \dots, SBF[h_K(x_i)]\}$ 
          do
             $SBF[h(x_i)] = Max$ 
      Output DuplicateFlag
end

```

---

In our formal discussion, we assume that the underlying distribution of the input data does not change over time. Our experiments on the real world data show that this is not a very strong assumption, and the experimental results verify our theory.

**PROOF.** Within each iteration, there are three operations: detecting duplicates, decreasing cell values and setting cells to  $Max$ . Since the first operation does not change the values of the cells, we just focus on the other two operations.

Within the process of iterations from 1 to  $N$ , the cell could be set 0, ...,  $(N - 1)$  or even  $N$  times. Since the newest setting operation clears the impact of any previous operations, we can just focus on the process after the newest setting operation.

Let  $A_l$  denote the event that within the  $N$  iterations the most recent setting operation applied to the cell occurs at iteration  $N - l$ , which means that no setting happened within the most recent  $l$  iterations (i.e. from iteration  $N - l + 1$  to iteration  $N$ ,  $l < N$ ), and let  $\bar{A}_N$  denotes the event that the cell has never been set within the whole  $N$  iterations. Hence, the probability that the cell is zero after  $N$  iterations is as follows:

$$Pr(SBF_N = 0) = \sum_{l=Max}^{N-1} [Pr(SBF_N = 0 | A_l)Pr(A_l)] + Pr(SBF_N = 0 | \bar{A}_N)Pr(\bar{A}_N), \quad (1)$$

where

$$Pr(SBF_N = 0 | A_l) = \sum_{j=Max}^l \binom{l}{j} p^j (1-p)^{l-j} \quad (2)$$

$$Pr(A_l) = (1-k)^l k \quad (3)$$

$$Pr(SBF_N = 0 | \bar{A}_N) = 1 \quad (4)$$

**Table 1: The Symbol List**

Symbols	Meanings
$N$	Number of elements in the input stream
$M$	Total space available in bits
$m$	Number of cells in the SBF
$Max$	The value a cell is set to
$d$	Number of bits allocated per cell
$K$	Number of hash functions
$k$	The probability that a cell is set in each iteration
$P$	Number of cells we pick to decrement by 1 in each iteration
$p$	The probability that a cell is picked to be decremented by 1 in each iteration
$h_i$	The $i$ th hash function

$$Pr(\bar{A}_N) = (1 - k)^N. \quad (5)$$

We have Eq 2 because during those  $l$  iterations, there is no setting operation, and the cell becomes zero if and only if it is decremented by 1 no less than  $Max$  times. Clearly when  $l < Max$ , the cell is impossible to be decreased to 0, and  $l = N$  means  $\bar{A}_N$  happens, so we just consider the cases of  $Max \leq l \leq (N - 1)$  in Eq 2. When  $\bar{A}_N$  happens, the cell is 0 with a probability 1 because the initial value of the cell is 0 and it has never been set, therefore we have Eq 4. Having the above equations, we can prove that  $\lim_{N \rightarrow \infty} Pr(SBF_N = 0)$  exists. See Appendix for the details.  $\square$

Having Theorem 1, now we can prove our stable property statement.

**COROLLARY 1 (STABLE PROPERTY).** *The expected fraction of zeros in an SBF after  $N$  iterations is a constant, provided that  $N$  is large enough.*

**PROOF.** In each iteration, each cell of the SBF has a certain probability of being set to  $Max$  by the element hashed to that cell. Since the underlying distribution of the input data does not change, the probability that a particular element appears in each iteration is fixed. Therefore, the probability of each cell being set is fixed.

Meanwhile, the probability that an arbitrary cell is decremented by 1 is also a constant. According to Theorem 1, the probabilities of all cells in the SBF becoming 0 after  $N$  iterations are constants, provided that  $N$  is large enough. Therefore, the expected fraction of 0 in an SBF after  $N$  iterations is a constant, provided that  $N$  is large enough.  $\square$

Now we know an SBF converges. In fact this convergence is like the process that a buffer is filled by items continually. SBF is stable means that its maximum capacity is reached, similar to the case that a buffer is full of items. Another important property is the convergence rate.

**COROLLARY 2 (CONVERGENCE RATE).** *The expected fraction of 0s in the SBF converges at an exponential rate.*

**PROOF.** From Equations 1, 4 and 5, we can derive

$$\begin{aligned} & Pr(SBF_N = 0) - Pr(SBF_{N-1} = 0) \\ &= Pr(SBF_N = 0 | A_{N-1})Pr(A_{N-1}) \\ & \quad + Pr(\bar{A}_N) - Pr(\bar{A}_{N-1}) \\ &= k(1 - k)^{N-1}(Pr(SBF_N = 0 | A_{N-1}) - 1) \end{aligned} \quad (6)$$

Clearly, Eq. 6 exponentially converges to 0. i.e.  $Pr(SBF_N[c]=0)$  converges at an exponential rate, and this is true for all cells in the SBF. Therefore, the expected fraction of 0s in the SBF converges at an exponential rate.  $\square$

**LEMMA 1 (MONOTONICITY).** *The expected fraction of 0s in an SBF is monotonically non-increasing.*

**PROOF.** Since the value of Eq. 6 is always no greater than 0, the probability that a cell becomes zero is always decreasing or remains the same. Similar to the proof of Corollary 2, we can draw the conclusion.  $\square$

This lemma will be used to prove our general upper bound of the false positive rate where the number of iterations needs not to be infinity.

### 3.4 The Stable Point

Currently we know the fraction of 0s in an SBF will be a constant at some point, but we do not know the value of this constant. We call this constant the stable point.

**DEFINITION 2 (STABLE POINT).** *The stable point is defined as the limit of the expected fraction of 0s in an SBF when the number of iterations goes to infinity. When this limit is reached, we call SBF stable.*

From Eq 1, we are unable to obtain the limit directly. However, we can derive it indirectly.

**THEOREM 2.** *Given an SBF with  $m$  cells, if a cell is decremented by 1 with a constant probability  $p$  and set to  $Max$  with a constant probability  $k$  in each iteration, and if the probability that the cell becomes 0 at the end of iteration  $N$  is denoted by  $Pr(SBF_N = 0)$ ,*

$$\lim_{N \rightarrow \infty} Pr(SBF_N = 0) = \left( \frac{1}{1 + \frac{1}{p(1/k-1)}} \right)^{Max} \quad (7)$$

**PROOF.** The basic idea is to make use of the fact that SBF is stable, the expected fraction of 0,1,...,Max in SBF should be all constant. See Appendix for the details.  $\square$

The theorem can be verified by replacing the parameters in Eq. 1 with some testing values.

From Theorem 2 we know the probability that a cell becomes 0 when SBF is stable. If all cells have the same probability of being set, we can obtain the stable point easily. However, that requires the data stream to be uniformly distributed. Without this uniform distribution assumption, we have the following statement.

**THEOREM 3 (SBF STABLE POINT).** *When an SBF is stable, the expected fraction of 0s in the SBF is no less than*

$$\left( \frac{1}{1 + \frac{1}{P(1/K-1/m)}} \right)^{Max},$$

where  $K$  is the number of cells being set to  $Max$  and  $P$  is the number of cells decremented by 1 within each iteration.

**PROOF.** The basic idea is to prove the case of  $m = 2$  first, and generalize it to  $m \geq 2$ . See Appendix for the details.  $\square$

### 3.5 False Positive Rates

In our method, there could be two kinds of errors: false positives (FP) and false negatives (FN). A false positive happens when a distinct element is wrongly reported as duplicate; a false negative happens when a duplicate element is wrongly reported as distinct. We call their probabilities *false positive rates* and *false negative rates*.

**COROLLARY 3 (FP BOUND WHEN STABLE).** *When an SBF is stable, the FP rate is a constant no greater than FPS,*

$$FPS = \left(1 - \left(\frac{1}{1 + \frac{1}{P(1/K - 1/m)}}\right)^{Max}\right)^K. \quad (8)$$

**PROOF.** If  $Pr_j(0)$  denotes the probability that the cell  $SBF[j] = 0$  when the SBF is stable, the FP rate is

$$\begin{aligned} & \left(\frac{1}{m}(1 - Pr_1(0)) + \dots + \frac{1}{m}(1 - Pr_m(0))\right)^K \\ &= \left(1 - \frac{1}{m}(Pr_1(0) + \dots + Pr_m(0))\right)^K \end{aligned}$$

Please note that  $\frac{1}{m}(Pr_1(0) + \dots + Pr_m(0))$  is the expected fraction of 0s in the SBF. According to Theorems 1 and 3, the FP rate is a constant and Eq. 8 is an upper bound of the FP rate.  $\square$

This upper bound can be reached when the stream elements are uniformly distributed.

**COROLLARY 4 (THE CASE OF REACHING THE FP BOUND).** *Given an SBF with  $m$  cells, if the stream elements are uniformly distributed when the SBF is stable, the FP rate is FPS (Eq. 8).*

**PROOF.** Because elements in the input data stream are uniformly distributed, each cell in the SBF will have the same probability to be set to *Max*. According to Theorem 1 and the proof of Theorem 3 we can derive this statement.  $\square$

**COROLLARY 5 (GENERAL FP BOUND).** *Given an SBF with  $m$  cells, FPS (Eq. 8) is an upper bound for FP rates at all time points, i.e. before and after the SBF becomes stable.*

**PROOF.** This can be easily derived from Lemma 1 and Corollary 3.  $\square$

Therefore, the upper bound for FP rates is valid no matter the SBF is stable or not.

From Eq. 8 we can see that  $m$  has little impact on *FPS*, since  $1/m$  is negligible compared to  $1/K$  ( $m \gg K$ ). This means the amount of space has little impact on the FP bound once the other parameters are fixed. The value of  $P$  has a direct impact on *FPS*: the larger the value of  $P$ , the smaller the value of *FPS*. This can be seen intuitively: the faster the cells are cleared, the more 0s the SBF has, thus the smaller the value of *FPS* is. Oppositely, increasing the value of *Max* results in the increase of *FPS*. In contrast to  $P$  and *Max*, from the formula we can see the impact of the value of  $K$  on *FPS* is twofold: intuitively, using more hash functions increases the distinguishing power for duplicates (decreases *FPS*), but ‘fills’ the SBF faster (increases *FPS*).

### 3.6 False Negative Rates

A false negative (FN) is an error when a duplicate element is wrongly reported as distinct. It is generated only by duplicate elements, and is related to the input data distribution,

especially the distribution of gaps. A gap is the number of elements between a duplicate and its nearest predecessor.

Suppose a duplicate element  $x_i$  whose nearest predecessor is  $x_{i-\delta_i}$  ( $x_i = x_{i-\delta_i}$ ) is hashed to  $K$  cells,  $SBF[C_{i1}] \dots SBF[C_{iK}]$ . An FN happens if any of those  $K$  cells is decremented to 0 during the  $\delta_i$  iterations when  $x_i$  arrives. Let  $PR0(\delta_i, k_{ij})$  be the probability that cell  $C_{ij}$  ( $j = 1 \dots K$ ) is decremented to 0 within the  $\delta_i$  iterations. This probability can be computed as in Eq 1:

$$\begin{aligned} PR0(\delta_i, k_{ij}) &= \sum_{l=Max}^{\delta_i-1} [Pr(SBF_{\delta_i} = 0 | A_l)Pr(A_l)] \\ &+ Pr(SBF_{\delta_i} = 0 | \bar{A}_{\delta_i})Pr(\bar{A}_{\delta_i}), \end{aligned} \quad (9)$$

where

$$Pr(SBF_{\delta_i} = 0 | A_l) = \sum_{j=Max}^l \binom{l}{j} p^j (1-p)^{l-j}, \quad (10)$$

$$Pr(A_l) = (1 - k_{ij})^l k_{ij}, \quad (11)$$

$$Pr(SBF_{\delta_i} = 0 | \bar{A}_{\delta_i}) = \sum_{j=Max}^{\delta_i} \binom{\delta_i}{j} p^j (1-p)^{\delta_i-j}, \quad (12)$$

$$Pr(\bar{A}_{\delta_i}) = (1 - k_{ij})^{\delta_i}, \quad (13)$$

and  $k_{ij}$  is the probability that cell  $C_{ij}$  is set to *Max* in each iteration. The meanings of the other symbols are the same as those in the proof of Theorem 1. Also, most of above equations are similar, except that Eq. 12 is different from Eq. 4. This is because the initial value of the cell in the case of Theorem 1 is 0, but it is *Max* here.

Furthermore, The probability that an FN occurs when  $x_i$  arrives can be expressed as follows:

$$Pr(FN_i) = 1 - \prod_{j=1}^K (1 - PR0(\delta_i, k_{ij})). \quad (14)$$

When  $\delta_i < Max$ ,  $PR0(\delta_i, k_{ij})$  is 0, which means the FN rate is 0. Besides, for distinct elements who have no predecessors, the FN rates are 0. The value of  $\delta_i$  depends on the input data stream. In the next section, we discuss how to adjust the parameters to minimize the FN rate under the condition that the FP rate is bounded within a user-specified threshold.

## 4. FROM THEORY TO PRACTICE

In the previous section we propose the SBF method and analytically study some of its properties: stability, convergence rate, monotonicity, stable point, FP rates (upper bound) and FN rates. In this section, we discuss how SBF can be used in practice and how our analytical results can be applied.

### 4.1 Parameters Setting

Since FP rates can be bounded regardless of the input data but FN rates cannot, given a fix amount of space, we can choose a combination of *Max*,  $K$  and  $P$  that minimizes the number of FNs under the condition that the FP rate is within a user-specified threshold. Meanwhile we take into account the time spent on each element, which is crucial in many data stream applications.

**The expected number of FNs.** Since our goal is to minimize the number of FNs, we can compute the expected number of FNs,  $E(\#FN)$ , as the sum of FN rates for each duplicate element in the stream:  $E(\#FN) = \sum_{i=1}^{\tilde{N}} Pr(FN_i)$ , where  $\tilde{N}$  is the number of duplicates in the stream. Combining it with Eq. 14 we have

$$E(\#FN) = \sum_{i=1}^{\tilde{N}} [1 - \prod_{j=1}^K (1 - PR0(\delta_i, k_{ij}))], \quad (15)$$

where  $\delta_i$  is the number of elements between  $x_i$  and its predecessor, and  $k_{ij}$  is the probability that cell  $C_{ij}$  is set to  $Max$  in each iteration.  $C_{ij}$  is the cell element  $x_i$  is hashed to by the  $j$ th hash function. Since the function  $PR0(\delta, k)$  is continuous, for each  $x_i$  there must be a  $\bar{k}_i$  such that

$$(1 - PR0(\delta_i, \bar{k}_i))^K = \prod_{j=1}^K (1 - PR0(\delta_i, k_{ij})).$$

For the same reason, there must be an ‘‘average’’  $\hat{\delta}$  and an ‘‘average’’  $\hat{k}$  such that

$$\tilde{N} [1 - (1 - PR0(\hat{\delta}, \hat{k}))^K] = \sum_{i=1}^{\tilde{N}} [1 - (1 - PR0(\delta_i, \bar{k}_i))^K] = E(\#FN).$$

Let  $f(\hat{\delta}, \hat{k})$  be the average FN rate, i.e.

$$f(\hat{\delta}, \hat{k}) = 1 - (1 - PR0(\hat{\delta}, \hat{k}))^K. \quad (16)$$

Our task then becomes setting the parameters to minimize this average FN rate,  $f(\hat{\delta}, \hat{k})$ , while bounding the FP rate within an acceptable threshold.

**The setting of  $P$ .** Suppose users specify a threshold  $FPS$ , indicating the acceptable FP rate. This threshold establishes a constraint between the parameters:  $Max$ ,  $K$ ,  $P$ ,  $m$  and  $FPS$  according to Corollary 5. Thus, users can set  $P$  based on the other parameters:

$$P = \frac{1}{\left(\frac{1}{(1 - FPS^{1/K})^{1/Max}} - 1\right)(1/K - 1/m)}. \quad (17)$$

Since  $m$  is usually much larger than  $K$ ,  $1/m$  is negligible in the above equation, which means that the setting of  $P$  is dominated only by  $FPS$ ,  $Max$ ,  $K$ , and is independent of the amount of space.

**The setting of  $K$ .** Since the FP constraint can be satisfied by properly choosing  $P$ , we can set  $K$  such that it minimizes the number of FNs. From the above discussions we know the relationship between the expected number of FNs and the probabilities that cells are set to  $Max$ . Next, we connect these probabilities with our parameters  $K$ ,  $m$  and the input stream.

Suppose there are  $N$  elements in the stream of which  $n$  are distinct, and the frequency for each distinct element  $x_l$  is  $f'_l$ . Clearly  $\sum_{l=1}^n f'_l = N$ . Assuming that the hash functions are uniformly at random, for a cell that element  $x_i$  is hashed to, the number of times the cell is set to  $Max$  after seeing all  $N$  elements is a random variable,  $f_i + \sum_{l=1}^{n-1} f'_l I_l$ , where  $f_i$  is the frequency of  $x_i$  in the stream, and each  $I_l (l = 1 \dots n-1)$  is an independent random variable following the Bernoulli distribution, i.e.

$$I_l = \begin{cases} 1, & Pr(I_l = 1) = \frac{K}{m}, \\ 0, & Pr(I_l = 0) = 1 - \frac{K}{m}. \end{cases}$$

Thus,  $k_i = \frac{1}{N} f_i + \frac{1}{N} \sum_{l=1}^{n-1} f'_l I_l$  is also a random variable. For the  $K$  cells an element  $x_i$  is hashed to, the probabilities that those cells are set to  $Max$  in each iteration can be considered as  $K$  trials of  $k_i$ . Since the mean and the variance of each  $I_l$  are  $\mu_{I_l} = \frac{K}{m}$  and  $\sigma_{I_l}^2 = \frac{K}{m}(1 - \frac{K}{m})$  respectively, it is not hard to derive that the mean and variance of  $k_i$ :  $\mu_{k_i} = \frac{1}{N} f_i + \frac{1}{N} \frac{K}{m} \sum_{l=1}^{n-1} f'_l = \frac{1}{N} f_i + \frac{K}{m}(1 - \frac{1}{N} f_i)$  and  $\sigma_{k_i}^2 = \frac{1}{N^2} \frac{K}{m}(1 - \frac{K}{m}) \sum_{l=1}^{n-1} f_l'^2$ . Let

$$\phi_i = \frac{k_i - \mu_{k_i}}{\sqrt{\frac{K}{m}(1 - \frac{K}{m})}} = \frac{k_i - \frac{1}{N} f_i - \frac{K}{m}(1 - \frac{1}{N} f_i)}{\sqrt{\frac{K}{m}(1 - \frac{K}{m})}} \quad (18)$$

be a transformation on  $k_i$ . Then  $\phi_i \in [-\frac{\frac{1}{N} f_i + \frac{K}{m}(1 - \frac{1}{N} f_i)}{\sqrt{\frac{K}{m}(1 - \frac{K}{m})}}, \frac{1 - \frac{1}{N} f_i - \frac{K}{m}(1 - \frac{1}{N} f_i)}{\sqrt{\frac{K}{m}(1 - \frac{K}{m})}}]$  is a random variable whose mean and variance are:  $\mu_{\phi_i} = 0$  and  $\sigma_{\phi_i}^2 = \frac{1}{N^2} \sum_{l=1}^{n-1} f_l'^2$ . Note that  $\sigma_{\phi_i}^2 \leq \frac{1}{N^2} \sum_{l=1}^{n-1} (f'_l f'_{max}) < \frac{1}{N^2} \sum_{l=1}^{n-1} (f'_l f'_{max}) = \frac{f'_{max}}{N}$ , where  $f'_{max}$  is the frequency of the most frequent element in the stream. Since the mean and the variance of the random variable  $\phi_i$  are independent of  $K$  and  $m$ , we may consider  $\phi_i$  independent of  $K$  and  $m$  in practice. In other words,  $\phi_i$  can be seen as a property of the input stream. Similar to  $\bar{k}_i$  we can obtain a  $\bar{\phi}_i$  such that

$$1 - (1 - PR0(\delta_i, \bar{\phi}_i))^K = 1 - \prod_{j=1}^K (1 - PR0(\delta_i, \phi_{ij})) = Pr(FN_i) \quad (19)$$

where  $\bar{\phi}_i \in [Min(\phi_{ij}), Max(\phi_{ij})]$ , and  $\phi_{ij}$  are  $K$  trials of  $\phi_i (j = 1 \dots K)$ . Since the standard deviation of  $\phi_i$  is very small compared to the range of its possible values, and  $\phi_i$  is considered independent of  $K$  and  $m$ ,  $\bar{\phi}_i$  can be approximately considered independent of  $K$  and  $m$  as well. For example, when  $\frac{f'_{max}}{N} = 0.01$ ,  $\frac{f_i}{N} = \frac{1}{10^6}$ ,  $\frac{m}{K} = 10^6$ , the value range of  $\phi_i$  is approximately  $[0, 1000]$ , while  $\sigma_{\phi_i} \leq 0.1$ .

To set  $K$ , keeping all other parameters fixed we vary the values of  $K$  and compute the FN rate based on Equations 19, 9, 17 and 18. By trying different combinations of parameter settings ( $Max = 1, 3, 7, 15$ ,  $FPS = 0.2, 0.1, 0.01, 0.001$ ,  $m = 1000, 10^5, 10^7, 10^9$ ,  $\delta_i = 10, 100, 1000, 10^5, 10^7, 10^9$ ,  $f_{xn} = \frac{f_i}{N} = 0.5, 0.1, 0.01, 0.0001, 0.000001$  and  $\bar{\phi}_i = 0.001, 0.1, 1, 10, 100, 1000, \dots$ ), we find that once the values of  $FPS$  and  $Max$  are fixed, the value of the optimal or near optimal  $K$  is independent of the values of  $\delta_i$ ,  $f_i/N$ ,  $\bar{\phi}_i$  and  $m$ .

**OBSERVATION 1.** *The value of the optimal or near optimal  $K$  is dominated by  $Max$  and  $FPS$ . The input data and the amount of the space have little impact on it. Furthermore, the value is small ( $< 10$  in all of our testing).*

For example, when  $FPS = 0.2$  and  $Max = 1$ , the value of the optimal or near optimal  $K$  is always between 1 and 2; when  $FPS = 0.1$  and  $Max = 3$ , it is always between 2 and 3; when  $FPS = 0.01$  and  $Max = 3$ , it is always between 4 and 5. Therefore, without considering the input data stream we can pre-compute the FN rates for different values of  $K$  based on  $Max$  and  $FPS$  and choose the optimal one. Our experimental results reported in the next section are consistent with this observation.

Figure 1 shows an example of how the FN rates change with different values of  $K$  under different parameter settings based on Eq. 19 and Eq. 9. From the figure we can see that in the case of  $Max = 1$  and  $FPS = 0.1$ , we can set  $K$  to

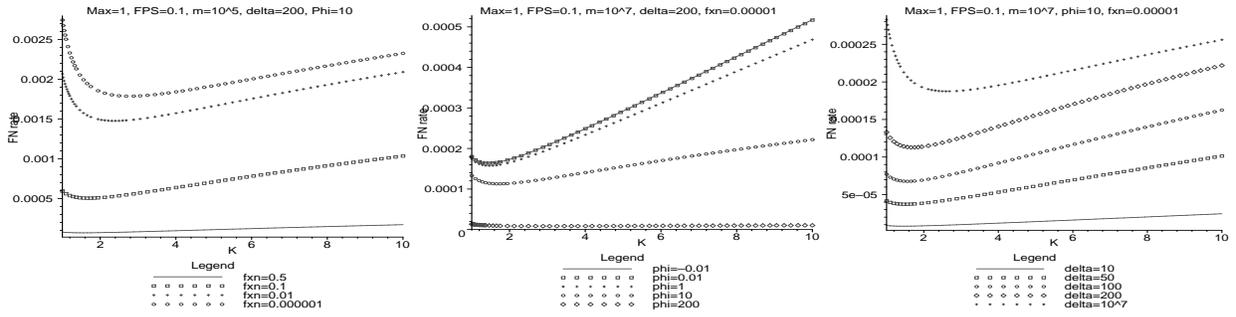


Figure 1: FN rates vs.  $K$

2 regardless of the input stream and the amount of space. Therefore, in practice we can set  $\phi_i$ ,  $\delta_i$  and  $f_i/N$  to some testing values (e.g. 0, 200, 0.00001 respectively) and find the optimal or near optimal  $K$  using the formulas.

**The setting of  $Max$ .** Based on the above discussion, we can set  $K$  regardless of the input data, but to choose a proper value of  $Max$ , we need to consider the input. More specifically, to minimize the expected number of FNs, we need to know the distributions of gaps in the stream to try different possible values of  $Max$  on Eq. 16 and 9. Since the expected value of  $\phi_i$  is 0 and its standard deviation is very small compared to its value domain, we set  $\hat{\phi}$  to 0 in the formulas.

To effectively use the space we only set  $Max$  to  $2^d - 1$  ( $d$  is the number of bits/cell), otherwise the bits allocated for each cells are just wasted. Furthermore, in terms of the time cost,  $Max$  should be set as small as possible, because the larger  $Max$  is set, the larger  $P$  will be (see Eq.17, assuming  $K$  is a constant). For example, when  $Max = 1$ ,  $FPS = 0.01$  and  $K = 3$ (the optimal  $K$ ), the computed value of  $P$  is 10; while  $Max = 15$ ,  $FPS = 0.01$ , and  $K = 6$  (the optimal  $K$ ), the value of  $P$  computed is 141 (the value of  $P$  is not sensitive to  $m$ ). In practice, we limit our choice of  $Max$  to 1, 3 and 7 (if higher time cost can be tolerated, larger values of  $Max$  can be tried similarly).

To choose a  $Max$  from these candidates, we try each value on Eq. 16 and Eq. 9, and find the one minimizing the average FN rate.

Figure 2 depicts the difference of average FN rates between  $Max = 3$  and  $Max = 1$  based on Eq.16 and Eq.9. We set  $\frac{f_i}{N} = 0$  because we are considering the entire stream rather than a particular element in this case. The figure shows that if the values of gaps( $\delta$ ) are smaller than a threshold,  $Max = 3$  is a better choice. When the gaps become larger,  $Max = 1$  is better. If the gaps are large enough, there is not much difference between the two options. The figure shows the cases under different settings of  $\hat{\phi}$ , space sizes and acceptable FP rates.

We also tested the FN rate difference between  $Max = 7$  and  $Max = 3$ , and observe the same general rule: a larger value of  $Max$  is better for smaller gaps, and a smaller  $FPS$  suggests a larger setting of  $Max$ . Similarly, we find no exceptions under other combinations of settings:  $FPS = 0.2, 0.1, 0.01, 0.001$  and  $m = 1000, 10^5, 10^7, 10^9$ .

Trying different value of  $Max$  on Eq. 16 and Eq. 9, we set  $\hat{\phi}$  to 0 and assume that the distribution of the gaps are known. If the assumption cannot be satisfied in practice, we suggest setting  $Max$  to 1, because this setting often benefits

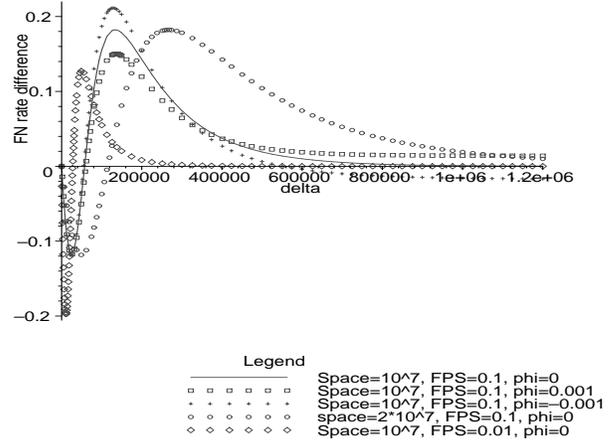


Figure 2: FN rates difference between  $Max = 3$  and  $Max = 1$  ( $Max3 - Max1$ ) vs. gaps.  $K$  is set to the optimal value respectively under different settings.

a larger range of gaps in the stream. And our experiments also show that in most cases setting  $Max$  to 1 achieves better improvements in terms of error rates compared to the alternative method, LRU buffering. In fact, buffering performs well when gaps are small, which is similar to the cases that  $Max$  is larger. The behavior of our SBF becomes closer to the buffering method when the value of  $Max$  is set larger.

**Summary of parameters setting and FP/FN trade-off.** In practice, given an  $FPS$ , the amount of available space and the gap distribution of the input data, to set the parameters properly, we first establish a constraint for  $P$ , which means  $P$  can be computed based on  $FPS$ ,  $m$ ,  $Max$  and  $K$ ; then find the optimal values of  $K$  for each case of  $Max(1, 3, 7)$  by trying limited number( $\leq 10$ ) of values of  $K$  on the FN rate formulas; Last, we estimate the expected number of FNs for each candidate of  $Max$  using its corresponding optimal  $K$  and some prior knowledge of the stream, and thus choose the optimal value of  $Max$ . In the case that no prior knowledge of the input data is available, we suggest setting  $Max = 1$ . The described parameter setting process can be implemented within a few lines of codes. In addition, based on the above analysis, we obtain a constraint function between FN rates,  $FPS$  and other parameters, where the tradeoff between FPs and FNs can be clearly seen: when the other parameters are fixed, the larger the FP bound, the smaller the FN rate is.

## 4.2 Time Complexity

Since our goal is to minimize the error rates given a fixed amount of space and an acceptable FP rate, we do not discuss space complexity, and just focus on time complexity. There are several parameters to be set in our method:  $K$ ,  $Max$  and  $P$ . Within each iteration, we firstly need to probe  $K$  cells to detect duplicates. After that we pick  $P$  cells and decrement 1 from them. Last we set the same  $K$  cells as probed in the first step to  $Max$ .

Therefore, the time cost of our algorithm for handling each element is dominated by  $K$  and  $P$ .

**THEOREM 4 (TIME COMPLEXITY).** *Given that  $K$  and  $Max$  are constants, processing each data stream element needs  $O(1)$  time, independent of the size of the space and the stream.*

**PROOF.** From Eq.17 we know the constraint among  $K$ ,  $P$ ,  $m$ ,  $Max$  and  $FPS$  (the user-specified upper bound of false positive rates). If  $K$ ,  $Max$  and  $FPS$  are constants, the relationship between  $P$  and  $m$  is inversely proportional, which means  $m$  has no impact on the processing time. Since  $Max$ ,  $K$  and  $FPS$  are all constants, the time complexity is  $O(1)$ .  $\square$

Based on the discussion of parameter settings, we know that the selection of  $K$  is insensitive to  $m$ . Furthermore, the value of  $m$  and the stream size have little impact on the selection of  $Max$  based on our testing on Eq. 16. Therefore, our algorithm needs  $O(1)$  time per element, independent of the size of space.

## 5. EXPERIMENTS

In this section, we first describe our data set and the implementation details of 4 methods: SBF, Bloom Filter (BF), Buffering and FP Buffering (a variation of buffering which can be fairly compared to SBF). We then report some of the results on both real and synthetic data sets. Last, we summarize the comparison between different methods.

### 5.1 Data Sets

**Real World Data.** We simulated a web crawling scenario [8] as discussed in Section 1, using a Web crawl data set obtained from the Internet archive [2]. We hashed each URL in this collection to a 64-bit fingerprint using Rabin's method [32], as was done earlier [8]. With this fingerprinting technique, there is a very small chance that two different URLs are mapped to the same fingerprint. We verified the data set and did not find any collisions between the URLs. In the end, we obtained a 28GB data file that contained about 700 million fingerprints of links, representing a stream of URLs encountered in a Web crawling process.

**Synthetic Data.** We used 2 steps in generating the synthetic data. First, we generated a sequence of positive integers in which there was only one frequent integer appearing multiple times, and all others were distinct integers. Second, we scanned through the integer sequence generated in the first step, and randomly converted the only frequent integer into another integer with certain probability. We describe the details next.

In the first step, we used 2 different models to simulate the arrival positions of the only frequent integer in the sequence: the Poisson model and the b-model. The Poisson model is widely used in modeling real world traffic [15]; it is also

shown that when the traffic has a bursty or a self-similar property, the b-model is more accurate [36].

The Poisson model had a parameter  $\lambda (\leq 1)$ , indicating the number of arrivals of the frequent integer within a unit time interval. Since we used a tuple-based stream model, there were no explicit timestamps for elements in our case. Instead, we used the element arrival positions as timestamps. As a result, when the arrivals of the only frequent integer followed a Poisson process with a parameter  $\lambda$ , the fraction of the frequent integer among an arbitrary number of elements is  $\lambda$  on average. It was also shown that the inter-arrival time (the number of elements in our case) between two consecutive arrivals followed an exponential distribution with a parameter  $1/\lambda$  [15].

To simulate bursty traffics we used the b-model. The b-model had 3 parameters: bias  $b (\geq 0.5)$ , aggregation level  $l$ , and total volume  $v$ . When  $b = 0.7$  and  $l = 1$ , the entire time interval under consideration was divided into 2 equal sub-intervals:  $v * b$  traffic volume in the first sub-interval, and  $v * (1 - b)$  in the second sub-interval. In other words, 70% of the total volume  $v$  was in the first sub-interval and 30% in the second sub-interval (or 30% in the first and 70% in the second, but we assumed the former case for the ease of presentation). If  $l = 2$ , we could further divide the first sub-interval into two equal finer sub-intervals, and allocate 70% of the volume in the higher level sub-interval into the first finer sub-interval and the remaining 30% into the second finer sub-interval. This process could be continued recursively for a larger  $l$ .

In our tuple-based stream model, there were no explicit time intervals. Instead, we used the element arrival positions as timestamps. The traffic volume  $v$  in our case was the fraction of duplicates. As an example,  $b = 0.7$  and  $l = 1$  meant that 70% of the arrivals of the frequent integer were in the first half of the sequence, and 30% in the second half. However, it was possible that the predicted number of arrivals of the frequent integer in the first half was greater than the total number of arrivals in the first half. For example, if  $v = 0.9$ , the number of duplicates in the first half of the sequence was  $0.9 * 0.7 = 0.63$ , which meant that 63% of the total number of elements in the sequence should be allocated in the first half of the sequence. This was obviously infeasible. Observing this, we obtained a constraint between  $v$ ,  $b$  and  $l$ :  $l \leq \ln(v) / \ln(\frac{1}{2b})$ . This constraint indicated that to obtain a larger aggregation levels  $l$ , the value of  $v$  and  $b$  should be smaller. For example, when  $v = 0.1$  and  $b = 0.7$ ,  $l \leq 6$ . But when  $v = 0.5$  and  $b = 0.8$ ,  $l \leq 1$ . When  $b = 0.5$ , there was no constraint because the recursive division process was always possible.

In the second step, we scanned through the integer sequence generated in the first step using the Poisson model or the b-model, and randomly converted the only frequent integer into another number  $i$  with a probability  $q_i$ , which followed an exponential distribution with a parameter  $u$ . After the second step, there were more than one frequent integer in the sequence, and the new arrival process of any frequent number  $i$  still followed the Poisson model or b-model, with a new parameter different from the old one by a factor of  $q_i$  (the Poisson case) or with the same parameter (the b-model case). By varying the parameters of the Poisson model or b-model in the first step, we were able to tune the arrival rate or the burstiness of duplicates. By varying the parameters of the exponential distribution in the second



rate (number of FPs/number of distincts), FPS, to 10%, and Max, K,P to 1, 2, 4 respectively. The results under different FPS settings will be shown in the next experiment. For FPbuffering, we set the FP rates to the same number as SBF so that both generated exactly the same number of FPs, and we can just compare their FN rates. Please note that buffering and BF only generate FNs and FPs respectively, and FPbuffering reduces the FN rates of buffering substantially in most cases by introducing a certain amount of FPs.

76% Distinct(100K elements)					
Space (bits)	FPBuffering FN Rate	SBF FN Rate	SBF&FPBuffering FP Rate	Buffering FN Rate	BF FP Rate
16384	46%	35%	8.4%	50%	83.8%
65536	35%	23%	6.7%	37%	43.4%
262144	24%	11%	3.0%	25%	7.2%
1048576	9%	4%	0.4%	9%	0.6%
4194304	0.1%	1%	0.1%	0.1%	0.1%

49% Distinct (10M elements)					
Space (bits)	FPBuffering FN Rate	SBF FN Rate	SBF&FPBuffering FP Rate	Buffering FN Rate	BF FP Rate
16384	60%	54%	8.1%	65.7%	99.7%
262144	48%	40%	6.6%	51.5%	95.9%
4194304	30%	23%	4.5%	31.7%	43.5%
67108864	11%	5%	0.5%	11.0%	0.7%
1073741824	0.0%	0.4%	0.1% (only SBF)	0.0%	0.1%

15% Distinct (695M elements)					
Space (bits)	FPBuffering FN Rate	SBF FN Rate	SBF&FPBuffering FP Rate	Buffering FN Rate	BF FP Rate
16384	71%	68%	8.2%	78%	99.99%
262144	65%	60%	7.0%	70%	99.81%
4194304	55%	50%	5.7%	58%	96.93%
67108864	43%	36%	3.5%	45%	54.08%
1073741824	17%	13%	1.6%	17%	2.65%
4294967296	2%	5%	1.6%	2%	1.87%

Figure 4: Error rates comparison between SBF, FP-Buffering, Buffering and BF

**Comparison between different methods.** The tables in Figure 4 show that when the space is relatively small, SBF is better. SBF beats FPbuffering by 3-13% in terms of FN rate on different data sets, when their FP rates are the same. For the problem we are studying, we think this amount of improvement is nontrivial for 2 reasons. First, Broder et al. [8] implemented a theoretically optimal buffering algorithm called MIN, which assumes "the entire sequence of requests is known in advance", and accordingly chooses the best replacement strategy. Even this obviously impractical and ideal algorithm can only reduce the miss rates (FN rates in our case) of the LRU buffering, by no more than 5% in about 2/3 region (different buffer sizes). Second, from the tables we can see that even increasing the amount of the space by a factor of 4, the FN rates for buffering can be decreased by around 10-20%, which means the improvement from SBF may be equivalent of that from doubling the amount of space. The FP rates of BF is much higher than the acceptable FP rates in the first 2-3 rows of each table. Since buffering only generates FNs, it is not comparable to SBF here. But we can see that the FN rates of FPbuffering also decrease by introducing FPs into it.

However, we also notice that when the space is relatively large (the last row of each table), SBF performs not as good as buffering and BF. This is because when the space is large, BF might be able to hold all the distincts and keep a reasonable FP rates. We can directly compute the amount of

space required based on the FP rates desired and the number of distincts in the data set according to the formula in Section 2.3. In this case, there is no need to evict elements out of the BF, which means SBF is not applicable. If we can afford even more space, which is large enough to hold all the distincts in the data set using a buffer, there will be no errors at all. The last row of the second table shows this scenario. But in many data stream applications, a fast storage is needed to satisfy real time constraints and the size of this storage is typically less than the universe of the stream elements as discussed in Introduction.

Another fact is that in both SBF and buffering, we can refresh the storage and bias it towards recent data; they both evict stale elements continuously and keep those fresh ones. While BF is not applicable in this case since BF can be only used to represent a static data set. Thus, it is not useful in many data stream scenarios that require dynamic updates.

**Varying acceptable FP rates.** Another experiment we ran was to test the effect of changing the acceptable FP rates. The results are shown in Figure 5. In this experiments, we set Max=3, K=4 when acceptable FP rates are set to 0.5% and 1%, and set Max=1, K=2 when acceptable FP rates are set to 10% and 20%. The bar chart depicts the FN rate difference between FPbuffering and SBF. Again, the FP rates of both methods are set to the same number. Clearly, it shows that the more FPs are allow, the better SBF performs.

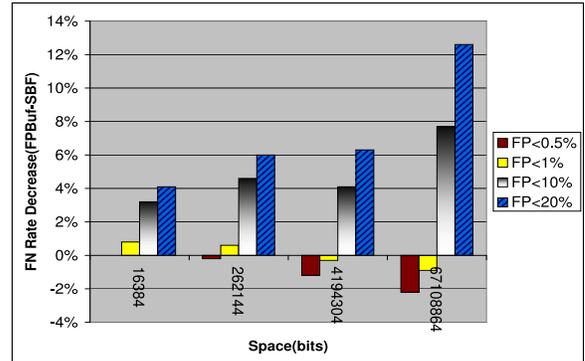


Figure 5: FN rate differences between FPBuffering and SBF varying allowable FP rate(695M elements)

**Comparison on different data sets.** To see the impact of different data sets on both the SBF and FPBuffering methods, we ran similar experiments on the synthetic data. In general, the results were consistent with those from the real data. See Appendix for the details.

## 5.5 Time Comparison

As discussed in the implementation section, SBF and BF need  $O(1)$  time to process each element. The exact time depends on the the parameter settings. For example, when  $K=2$  and  $P=4$ , SBF needs less than 10 operations within each iteration.

For buffering and FPbuffering, their processing time is the same. It depends on 2 processes: element searching and element evicting. Searching can be quite expensive without an index structure. Both our experiments and those of Broder et al.[8] used a hash table to accelerate the search

process. The extra space that is needed for a hash table to keep the search time constant is linear in the number of elements stored. The process of maintaining the LRU replacement policy (finding the least recently used element) is also costly, and extra space is needed to make it faster. This extra space can be quite large for LRU. However, this cost can be reduced to 2 bits per elements by using the Clock approximation of LRU [8].

Therefore, buffering and FPbuffering need extra space linear in the number of buffer entries to reach a similar  $O(1)$  processing time. But in our error rate comparison, we did not count this extra space for buffering and FPbuffering.

## 5.6 Methods Comparison Summary

We compared 4 methods in this section: SBF, BF, FPbuffering and buffering. Among them, BF and buffering have only FPs and FNs respectively, and SBF and FPbuffering have errors of both sides.

BF is a space efficient data structure which has been studied in the past and is widely used. It is good for representing a static set of data provided that the number of distinct elements is known. However, in data stream environments, the data is not static and it keeps changing. Usually it is hard to know the number of distinct elements in advance. Moreover, BF is not applicable in cases where dynamic updates are needed since elements can only be inserted into BF, but cannot be dropped out. Consequently, BF is not suitable for many data stream applications.

SBF, buffering and FPBuffering can be all applied to data stream scenarios. SBF is better in terms of accuracy and time when certain amount of FP rates are acceptable and the space is relatively small, which is the case in many data stream applications due to the real-time constraint. When the space is relatively large or only tiny FP rates are allowed, buffering is better.

## 6. RELATED WORK

The recent work of Metwally et al.[29]also study the duplicate detection problem in a streaming environment based on Bloom filters(BF) [7]. They consider different window models: Landmark windows, sliding windows and jumping windows. For the landmark window model, which is the scenario we consider, they apply the original Bloom filters without variations to detect duplicates, and thus do not consider the case that the BFs become “full”. For the sliding window model, they use counting BFs [20] (change bits into counters) to allow removing old information out of the Bloom filter. However, this can be done only when the element to be removed is known, which is not possible in many streaming cases. For example, if the oldest element needs to be removed, one has to know that which counters are touched by the oldest element, but this information cannot be found in counting BFs, and maintaining this knowledge can be quite expensive. For the jumping window model, they cut a large jumping window into multiple sub-windows, and represent both the jumping window and the sub-windows with counting BFs of the same size. Thus, the jumping window can “jump” forward by adding and removing sub-window BFs.

Another solution for detecting duplicates in a streaming environment is the buffering or the caching method, which has been studied in many areas such as database systems [22], computer architecture [23], operating systems [34], and

more recently URL caching in Web crawling [8]. We compare our method with those of Broder et al.[8] in the experiments.

The problem of exact duplicate elimination is well studied, and there are many efficient algorithms(e.g. see [22] for details and references). For the problem of approximate membership testing in a non-streaming environment, the Bloom filter has been frequently used (e.g. [27, 26]) and occasionally extended [20, 30]. Cohen and Matias[14] extend the Bloom filter to answer multiplicity queries. Counting distinct elements using the Bloom filter is proposed by Whang et al.[38]. Another branch of duplicate detection focus on fuzzy duplicates [11, 1, 6, 37], where the distinction between elements is not straightforward to see.

A related problem to duplicate detection is counting the number of distinct elements. Flajolet and Martin[21] propose a bitmap sketch to address this problem in a large data set. The same problem is also studied by Cormode et al. [16], and a sketch based on stable random variables is introduced. Besides, the sticky sampling algorithm of Manku and Motwani [28] also randomly increment and decrement counters storing the frequencies of stream elements, but the decrement frequency is varying and not for each incoming element. Their goal is to find the frequent items in a data stream.

As for data stream systems [3, 9, 35, 10, 17, 12], as far as we know, most of them divide the potentially unbounded data stream into windows with limited size and solve the problem precisely within the window. For example, Tucker et al. introduce punctuations into data streams, and thus duplicate eliminations could be implemented within data stream windows using traditional methods[35].

Since there is no way to store the entire history of an infinite data stream using limited space, our SBF essentially represents the most recent information by discarding those stale continuously. This is useful in many scenarios where the recent data is more important and this importance decays over time. A number of such kinds of applications are provided in [13] and [31]. Our motivating example of web crawling also has this property, since it may not matter that much to redundantly fetch a Web page that have been crawled a long time ago compared to fetching a page that have been crawled more recently.

## 7. CONCLUSIONS AND FUTURE WORK

In this paper, we propose the SBF method to approximately detect duplicates for streaming data. SBF extends the Bloom filter to allow outdated information to be removed from the summary structure so as to have enough space to represent fresh data. We formally and experimentally study the properties of our algorithm including stability, exponential convergence rates, monotonicity, stable points, bounded FP rates, FN rates dependent on input data stream,  $O(1)$  time independent of the stream size and space.

We empirically evaluate SBF and report the conditions under which our method outperforms the alternative methods, in particular buffering and FPBuffering methods using LRU replacement policy, both of which need extra space linear in the buffer size to obtain constant or nearly constant processing time.

Our SBF method works in a “landmark” window environment where the starting point of the window is fixed. SBF implicitly favors recent data, but no explicit sliding windows

can be specified by users to indicate the “recentness” range. We are working on the sliding version now.

## 8. REFERENCES

- [1] R. Ananthkrishna, S. Chaudhuri, and V. Ganti. Eliminating fuzzy duplicates in data warehouses. In *Proc. of VLDB*, 2002.
- [2] Internet Archive. <http://www.archive.org/>.
- [3] B. Babcock, S. Babu, M. Datar, R. Motwani, and J. Widom. Models and issues in data stream systems. In *Proc. of PODS*, 2002.
- [4] B. Babcock, M. Datar, and R. Motwani. Load shedding for aggregation queries over data streams. In *Proc. of ICDE*, 2004.
- [5] F. Baboescu, S. Singh, and G. Varghese. Packet classification for core routers: Is there an alternative to cams? In *Proc. of INFOCOMM*, 2003.
- [6] M. Bilenko and R. J. Mooney. Adaptive duplicate detection using learnable string similarity measures. In *Proc. of KDD*, 2003.
- [7] B. H. Bloom. Space/time trade-offs in hash coding with allowable errors. In *CACM*, 1970.
- [8] A. Z. Broder, M. Najork, and J. L. Wiener. Efficient url caching for world wide web crawling. In *Proc. of WWW*, 2003.
- [9] D. Carney, U. Cetintemel, M. Cherniack, C. Convey, S. Lee, G. Seidman, M. Stonebraker, N. Tatbul, and S. B. Zdonik. Monitoring streams - a new class of data management applications. In *Proc. of VLDB*, 2002.
- [10] S. Chandrasekaran, O. Cooper, A. Deshpande, M. J. Franklin, J. M. Hellerstein, W. Hong, S. Krishnamurthy, S. Madden, V. Raman, F. Reiss, and M. A. Shah. Telegraphicq: Continuous dataflow processing for an uncertain world. In *Proc. of CIDR*, 2003.
- [11] S. Chaudhuri, V. Ganti, and R. Motwani. Robust identification of fuzzy duplicates. In *Proc. of ICDE*, 2005.
- [12] J. Chen, D. J. DeWitt, F. Tian, and Y. Wang. Niagaraq: A scalable continuous query system for internet databases. In *Proc. of SIGMOD*, 2000.
- [13] E. Cohen and M. Strauss. Maintaining time-decaying stream aggregates. In *Proc. of PODS*, 2003.
- [14] S. Cohen and Y. Matias. Spectral bloom filters. In *Proc. of SIGMOD*, 2003.
- [15] R. B. Cooper. *Introduction to Queueing Theory(2nd edition)*. Elsevier, 1981.
- [16] G. Cormode, M. Datar, P. Indyk, and S. Muthukrishnan. Comparing data streams using hamming norms (how to zero in). *IEEE Trans. Knowl. Data Eng.*, 15(3):529–540, 2003.
- [17] C. Cranor, T. Johnson, O. Spatscheck, and V. Shkapenyuk. Gigascope: A stream database for network applications. In *Proc. of SIGMOD*, 2003.
- [18] C. Estan, K. Keys, D. Moore, and G. Varghese. Building a better netflow. In *Proc. of SIGCOMM*, 2004.
- [19] C. Estan and G. Varghese. Data streaming in computer networks. In *Proc. of Workshop on Management and Processing of Data Streams(MPDS) in cooperation with SIGMOD/PODS*, 2003.
- [20] L. Fan, P. Cao, J. Almeida, and A. Z. Broder. Summary cache: a scalable wide area web cache sharing protocol. *IEEE/ACM Trans. Netw.*, 8(3):281–293, 2000.
- [21] P. Flajolet and G. N. Martin. Probabilistic counting algorithms for data base applications. *J. Comput. Syst. Sci.*, 31(2):182–209, 1985.
- [22] H. Garcia-Molina, J. D. Ullman, and J. Widom. *Database System Implementation*. Prentice Hall, 2000.
- [23] J. L. Hennessy and D. A. Patterson. *Computer Architecture: A Quantitative Approach, 3rd Edition*. Morgan-Kaufmann, 2002.
- [24] A. Heydon and M. Najork. Mercator: A scalable, extensible web crawler. *World Wide Web*, 2(4), 1999.
- [25] Cisco System Inc. Cisco network accounting services. [http://www.cisco.com/warp/public/cc/pd/iosw/prodlit/nwact\\_wp.pdf](http://www.cisco.com/warp/public/cc/pd/iosw/prodlit/nwact_wp.pdf), 2002.
- [26] L. F. Mackert and G. M. Lohman. R\* optimizer validation and performance evaluation for distributed queries. In *Proc. of VLDB*, 1986.
- [27] U. Manber and S. Wu. An algorithm for approximate membership checking with application to password security. *Information Processing Letters*, 50(4), 1994.
- [28] G. S. Manku and R. Motwani. Approximate frequency counts over data streams. In *Proc. of VLDB*, 2002.
- [29] A. Metwally, D. Agrawal, and A. E. Abbadi. Duplicate detection in click streams. In *Proc. of WWW*, 2005.
- [30] M. Mitzenmacher. Compressed bloom filters. *IEEE/ACM Trans. Netw.*, 10(5):604–612, 2002.
- [31] T. Palpanas, M. Vlachos, E. J. Keogh, D. Gunopulos, and W. Truppel. Online amnesic approximation of streaming time series. In *Proc. of ICDE*, 2004.
- [32] M. O. Rabin. Fingerprinting by random polynomials. Technical Report TR-15-81, Center for Research in Computing Technology, Harvard University, 1981.
- [33] N. Tatbul, U. Cetintemel, S. Zdonik, M. Cherniack, and M. Stonebraker. Load shedding in a data stream manager. In *Proc. of VLDB*, 2003.
- [34] A. Tenenbaum. *Modern Operating Systems, 2nd Edition*. Prentice Hall, 2001.
- [35] P. A. Tucker, D. Maier, and T. Sheard. Applying punctuation schemes to queries over continuous data streams. *IEEE Data Eng. Bull.*, 26(1):33–40, 2003.
- [36] M. Wang, N. H. Chan, S. Papadimitriou, and T. M. Madhyastha C. Faloutsos. Data mining meets performance evaluation: Fast algorithms for modeling bursty traffic. In *Proc. of ICDE*, 2002.
- [37] M. Weis and F. Naumann. Dogmatix tracks down duplicates in xml. In *Proc. of SIGMOD*, June 2005.
- [38] K. Whang, B. T. V. Zenden, and H. M. Taylor. A linear-time probabilistic counting algorithm for database applications. *ACM Trans. Database Syst.*, 15(2):208–229, 1990.

## APPENDIX

### Theorem 1

$$\lim_{N \rightarrow \infty} Pr(SBF_N = 0)$$

exists.

PROOF. (The last step.) From Eq 4 and Eq 5 we can

clearly see the limit of the second part of  $Pr(SBF_N = 0)$

$$Pr(SBF_N = 0 | \bar{A}_N) * Pr(\bar{A}_N)$$

is 0. So we just need to focus on the first part of Eq 1, i.e.

$$\sum_{l=Max}^{N-1} (Pr(SBF_N = 0 | A_l) * Pr(A_l)),$$

which is denoted by  $Pr_{Part1}(0)$  in the rest of this proof. Now let us look at the first part of  $Pr_{Part1}(0)$ :

$$\begin{aligned} Pr(SBF_N = 0 | A_l) &= \sum_{j=Max}^l \binom{l}{j} p^j (1-p)^{l-j} \\ &= 1 - \sum_{j=0}^{Max-1} \binom{l}{j} p^j (1-p)^{l-j} \leq 1 - \binom{l}{0} p^0 (1-p)^{l-0} \\ &= 1 - (1-p)^l \end{aligned}$$

$$\begin{aligned} \therefore Pr_{Part1}(0) &\leq \sum_{l=Max}^{N-1} (1 - (1-p)^l)(1-k)^l k \\ &= \sum_{l=Max}^{N-1} (k(1-k)^l) - \sum_{l=Max}^{N-1} (k(1-k)^l(1-p)^l) \\ &= (1-k)^{Max} - (1-k)^N \\ &\quad - \frac{k}{k+p-kp} ((1-k)^{Max}(1-p)^{Max} - \\ &\quad (1-k)^N(1-p)^N) \end{aligned}$$

$$\begin{aligned} \therefore \lim_{N \rightarrow \infty} Pr_{Part1}(0) \\ \leq (1-k)^{Max} - \frac{k}{k+p-kp} ((1-k)^{Max}(1-p)^{Max} \end{aligned}$$

$\therefore Pr_{Part1}(0)$  monotonically increase when  $N \in [Max, \infty)$   
 $\therefore \lim_{N \rightarrow \infty} Pr_{Part1}(0)$  exists.  $\therefore$  The limit of the second part of  $Pr(SBF_N = 0)$  also exists.  $\therefore$  The limit of  $Pr(SBF_N = 0)$  exists.  $\square$

## Theorem 2

$$\lim_{N \rightarrow \infty} Pr(SBF_N = 0) = \left( \frac{1}{1 + \frac{1}{p(1/k-1)}} \right)^{Max}$$

PROOF. Similar to the proof of Theorem 1, we can prove that

$$\lim_{N \rightarrow \infty} Pr(SBF_N = v)$$

exists, where  $0 \leq v \leq Max$ . In other words, when  $N$  is large enough, the probability that the cell becomes  $v$  is a constant. We do not provide the details here. To simplify our notation, we will refer to  $\lim_{N \rightarrow \infty} Pr(SBF_N = v)$  as  $Pr(v)$  in this proof.

The key idea of this proof is as follows: since  $Pr(v)$  is a constant, when the SBF is stable, the probability that the cell changes to  $v$  from some other values (e.g.  $v+1$ ) in each iteration must be the same as the probability that the cell changes from  $v$  to other values. Otherwise,  $Pr(v)$  will either increase or decrease. Note that the probabilities of those changes caused by the decrementing and setting operations are all constants ( $p$  and  $k$ ).

For example, we can consider the case of  $Max = 3$  and  $Pr(1)$ . In each iteration, we decrement 1 from the cell with

a constant probability  $p$ , and set the cell to  $Max$  with a constant probability  $k$ . Since  $Pr(1)$  is a constant after the SBF becomes stable, the probability that the cell changes from 2 to 1 should be the same as the probability that the cell changes from 1 to 0 and  $Max$ . This can be expressed as

$$\begin{aligned} Pr(2 \rightarrow 1 | SBF = 2) * Pr(2) \\ = Pr(1 \rightarrow 0 | SBF = 1) * Pr(1) \\ + Pr(1 \rightarrow Max | SBF = 1) * Pr'(1) \end{aligned} \quad (20)$$

Note that  $Pr'(1)$  is the probability of  $SBF = 1$  after the decrementing but before the setting operation. Accordingly, we have

$$p * Pr(2) = p * Pr(1) + k * (p * Pr(2) - p * Pr(1) + Pr(1)).$$

The left hand side of the above equation is the probability of gaining 1s from the decrementing operation, and right hand side is the probability of losing 1s from both the decrementing and setting operations.

Using the same idea, the relationship among  $Pr(0), Pr(1), \dots, Pr(Max)$  can be expressed as follows (when  $Max \geq 3$ ): when  $v = 0$ ,

$$p * Pr(v+1) = k * (p * Pr(v+1) + Pr(v)), \quad (21)$$

when  $0 < v < Max$ ,

$$p * Pr(v+1) = p * Pr(v) + k * (p * Pr(v+1) - p * Pr(v) + Pr(v)), \quad (22)$$

when  $v = Max$ ,

$$k * (1 - Pr(v) + p * Pr(v)) = p * Pr(v). \quad (23)$$

From the above equations, we can derive

$$Pr(0) = \left( \frac{1}{1 + \frac{1}{p(1/k-1)}} \right)^{Max}.$$

Currently this equation only holds when  $Max \geq 3$  since we have this assumption earlier. But it is not hard to verify that it holds when  $Max = 1, 2$  as well using the same idea.  $\square$

**Theorem 3.** Given an SBF with  $m$  cells,  $SBF[c_1], \dots, SBF[c_m]$ , when the SBF is stable, the expected fraction of 0s in the SBF is no less than

$$\left( \frac{1}{1 + \frac{1}{P(1/K-1/m)}} \right)^{Max},$$

where  $K$  is the number of cells being set and  $P$  is the number of cells being decremented by 1 within each iteration.

PROOF. To simplify our notation, we refer to  $\lim_{N \rightarrow \infty} Pr(SBF_N[c_i] = 0)$  as  $Pr_i(0)$ , ( $i = 1, \dots, m$ ) and  $\left( \frac{1}{1 + \frac{1}{P(1/K-1/m)}} \right)^{Max}$  as  $LBound$  in this proof.

We know that when all cells have the same probability being set, the expected fraction of 0s in the SBF is  $LBound$ .

Now we prove that when cells have different probabilities being set, the expected fraction of 0s in the SBF is no less than  $LBound$ . In other words, we need to prove when  $k_1 = k_2 = \dots = k_m = K/m$ , the average probability

$$\frac{1}{m} \sum_{i=1}^m Pr_i(0) = \frac{1}{m} \sum_{i=1}^m \left( \frac{1}{1 + \frac{1}{p(1/k_i-1)}} \right)^{Max},$$

reaches the minimum value, where  $k_i$  is the probability that  $SBF[c_i]$  is set. Note that  $\sum_{i=1}^m k_i = K$ .

To simplify the problem, we first prove when  $m = 2$ ,  $k_1 + k_2 = \alpha$ , ( $0 \leq \alpha \leq K$ ), the above statement holds, i.e. when  $k_1 = k_2 = \alpha/2$ ,  $\frac{1}{1 + \frac{1}{p(1/k_1 - 1)}} + \frac{1}{1 + \frac{1}{p(1/k_2 - 1)}}$  is minimized. If  $F(k_1)$  denote the above formula and  $f(k_1)$  denote the first part,  $F(k_1) = f(k_1) + f(\alpha - k_1)$ . For  $f(k_1)$ , we can verify that the derivative of  $f(k_1)$  w.r.t.  $k_1$ ,  $df/dk_1 < 0$ , and  $d^2f/dk_1^2 > 0$  when  $k_1 \in [0, \alpha]$ . Because of the symmetric property between  $f(k_1)$  and  $f(\alpha - k_1)$ , the decreasing rate of  $f(k_1)$  is faster than the increasing rate of  $f(\alpha - k_1)$  when  $k_1 \in [0, \alpha/2)$ , thus  $F(k_1)$  monotonically decreases within this period. When  $k_1 = \alpha/2$ , those two rates are equal. Symmetrically it increases monotonically when  $k_1 \in (\alpha/2, \alpha]$ . Therefore, when  $k_1 = \alpha/2$ ,  $F(k_1)$  is minimized.

Now we prove when  $m > 2$ , this conclusion also holds. First we assume  $m = 2^j$ , where  $j$  is an integer greater than zero. Later we will discuss the general case without this assumption. Since we have proved the case of two cells, we can group all cells into  $m/2$  pairs arbitrarily. For each pair, we know if those two cells are set with an equal probability, the sum of  $Pr_i(0)$  will decrease. For example, if we group  $SBF[1]$  and  $SBF[2]$  together, we can minimize  $Pr_1(0) + PR_2(0)$  simply by changing both  $k_1$  and  $k_2$  into  $(k_1 + k_2)/2$ . This is the case for all other pairs. After the first grouping, we have at most  $m/2$  different  $k_i$ s. We then group those  $m/2$  cells into pairs and minimized their sums again by balance those pairs. We can keep doing this until all cells are all balanced, i.e. all of them have the same probability being set. Now the sum is minimized because in each step the partial sum is minimized.

When the assumption  $m = 2^j$  does not hold, we can divide one cell into two, where both have a same  $k_x$  and the sum does not change. For example, when  $m = 3$ , we can pick one cell arbitrarily, say  $SBF[3]$ , whose probability being set is  $k_3$ . We can find a proper  $k_x$ , such that  $f(k_3) = 2f(k_x)$ . Therefore, we can transform all cases into what we have proved, i.e.  $m = 2^j$ .  $\square$

**Experimental results from the synthetic data.** The acceptable FP rate was set to 10%, and the parameters for SBF were as follows:  $Max = 1$ ,  $K = 2$  and  $P = 4$ . In these data sets, the stream sizes were all fixed to  $2^{24}$  (about 16 million elements). The parameters generating the data sets are shown in Figure 6.

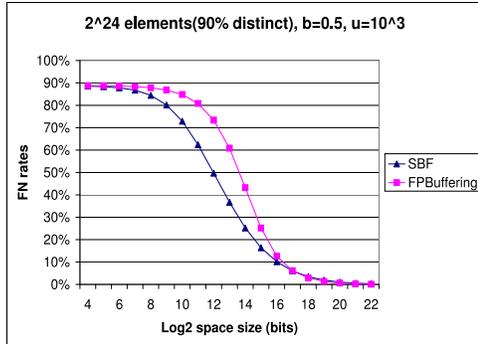
Between the two data sets shown in Figure 6(a) and 6(b), the only difference is the bias parameter  $b$ , which indicates the burstness of the data sets. These two sub-figures show that the FN rate difference between SBF and FPBuffering becomes smaller when the data set has the bursty property ( $b = 0.7$ ) compared to the case that the data set is not bursty ( $b = 0.5$ ). But the impact of this burstness is not significant.

We also set  $b$  to 0.7 and change the exponential distribution parameter  $u$ , which controls the frequency distribution of duplicates, from  $10^3$  to  $10^6$ . In the data set generated under this setting, the gaps between duplicates are increased in general, because increasing  $u$  means that the exponential distribution becomes less skewed, and the number of distinct elements in the set of frequent elements is increased (we consider the first time that a frequent element appears as a distinct element), and thus the probability of generating two identical integers within a short period becomes smaller. Figure 6(c) shows the results obtained from exper-

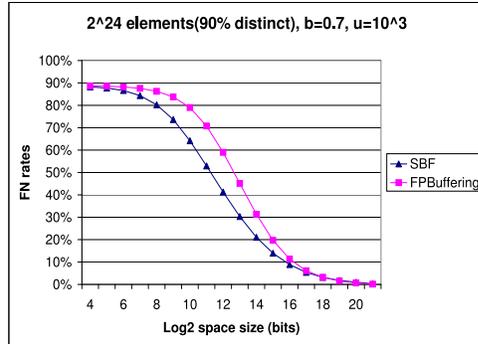
iments run on this data set. It demonstrates that both SBF and FPBuffering need more space to reach a particular FN rate when the gaps are increased. The region that the FN rate is 90% means that the space allocated to both methods is too small and does not help much, because without any space cost one can obtain similar results by randomly predicting each element as a duplicate at a probability 0.1.

Figure 6(d) depicts the results from the experiments run on the Poisson data set.  $u$  is set to the same value as in the previous sub-figure. Since there are more duplicates in this data set, the gaps in general are smaller compared to the previous data set. As a result, to reach a certain FN rate, both methods need less space. From this and some others experiments whose results are not shown here we find that SBF consistently outperforms FPBuffering on datasets generated using both the Poisson and the b- models.

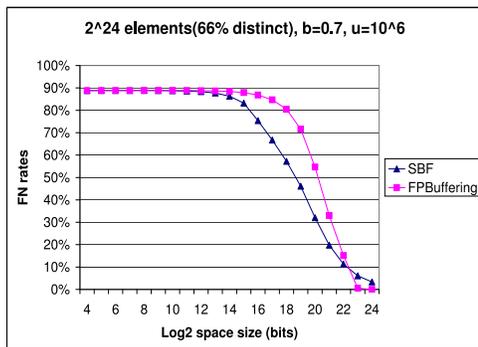
Besides, all sub-figures in Figure 6 also show that SBF can beat FPBuffering up to 30% in terms of FN rates when both methods have exactly the same FP rate ( $\leq 10\%$ ). When the allocated space is large (but not large enough to store all distinct elements), FPBuffering may be more accurate than SBF. But in this case, as discussed in the comparison on the real world data, the original BF method (which is also a special case of SBF) is usually the best since it only generates FPs under the acceptable threshold and no FNs. Of course, if the available space is large enough to store all distinct elements, buffering is the best in terms of accuracy, since it is precise.



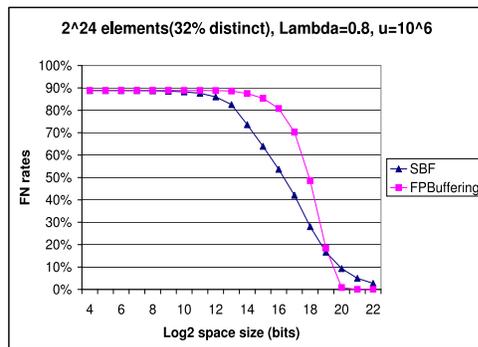
(a) B-model dataset 1 (aggregation level  $l = 6$ ).



(b) B-model dataset 2 (aggregation level  $l = 6$ ).



(c) B-model dataset 3 (aggregation level  $l = 6$ ).



(d) Poisson dataset

**Figure 6: FN rate comparisons between SBF and FPBuffering on synthetic data sets (FP rates  $\leq 10\%$ , in the SBF method  $Max = 1$ ,  $K = 2$  and  $P = 4$ ).**