Sketching Sampled Data Streams

Florin Rusu[1], Alin Dobra[2]

CISE Department
University of Florida
Gainesville, FL, USA
[1]frusu@cise.ufl.edu
[2]adobra@cise.ufl.edu

***Abstract*—Sampling is used as a universal method to reduce the running time of computations – the computation is performed on a much smaller sample and then the result is scaled to compensate for the difference in size. Sketches are a popular approximation method for data streams and they proved to be useful for estimating frequency moments and aggregates over joins. A possibility to further improve the time performance of sketches is to compute the sketch over a sample of the stream rather than the entire data stream.**

In this paper we analyze the behavior of the sketch estimator when computed over a sample of the stream, not the entire data stream, for the size of join and the self-join size problems. Our analysis is developed for a generic sampling process. We instantiate the results of the analysis for all three major types of sampling – Bernoulli sampling which is used for load shedding, sampling with replacement which is used to generate i.i.d. samples from a distribution, and sampling without replacement which is used by online aggregation engines – and compare these particular results with the results of the basic sketch estimator. Our experimental results show that the accuracy of the sketch computed over a small sample of the data is, in general, close to the accuracy of the sketch estimator computed over the entire data even when the sample size is only 10% or less of the dataset size. This is equivalent to a speed-up factor of at least 10 when updating the sketch.

I. Introduction

Data streaming has received a lot of attention from the research community in the last decade. The requirement to process fast data streams motivates the need for approximation methods that make use of both small space and small time. AGMS sketches [1], [2] and their improved variant F-AGMS sketches [3], [4] proved to be a viable solution for estimating aggregates over joins. The main strengths of the sketching techniques are the simple and fast update procedure, the small memory requirement, and provable error guarantees. When the data streams that need to be processed are extremely fast, for example in the case of networking data or large datasets streamed over the Internet, it is desirable to further reduce the update time of sketches in order to achieve the required processing rates. Sampling is a universal method for data reduction and, in principle, it can be used to reduce the amount of data that needs to be sketched. If samples are sketched instead of the original data, an immediate update time reduction results. This is similar to the existing load shedding techniques employed in data stream processing engines [5]. The main concern when samples rather than the original data are sketched is how to extend the error guarantees sketches provide to this new situation. The formulas resulting from such an analysis could be used to determine how aggressive the load shedding can be without a significant loss in the accuracy of the sketch over samples estimator.

A seemingly unrelated, but, as shown in the paper, technically related, problem is analyzing streams of samples from unknown distributions. Samples from unknown distributions – the so called i.i.d. samples – are the input to most of the online data-mining algorithms [6]. In this case the samples are not used as a data reduction technique, but rather they are the only information available about the unknown distribution. A fundamental problem in this context is how to characterize the unknown distribution using only the samples. This is one of the fundamental problems in statistics [7]. If the samples are streamed, as is the case in online data-mining, the aim is to characterize the unknown distribution by using small space only, thus making sketches a natural candidate for computations that involve aggregates. It is a simple matter to use sketches in order to estimate aggregates over the samples. If predictions about the unknown distribution need to be made, the problem is significantly more difficult. Interestingly, this problem is mathematically similar to the load shedding problem in which sampling is used to reduce the update time of sketching. A third problem is sketching tuples that are processed by an online aggregation engine in order to compute statistics useful for decision making [8], [9].

In this paper we analyze the sketch over samples estimator for generic sampling. Then we instantiate the results for three different types of sampling. Our technical contributions are:

- We provide a generic analysis of the sketch over samples estimator. The analysis consists in expressing the first two frequency moments of the estimator in terms of the moments of the sampling frequency random variables.
- We instantiate the results for sketching Bernoulli samples. This immediately indicates how random load shedding for sketching data streams behaves.
- We instantiate the generic analysis for sketching samples with replacement from a large population. The analysis generalizes to sketching i.i.d. samples from an unknown distribution. The ability to sketch i.i.d. data is important if sketches are to be used for data-mining applications.
- We instantiate the generic analysis for sketching samples without replacement. Such samples are processed by online aggregation engines. By sketching the samples,

important statistics can be derived with little computational overhead.
- We present empirical evidence that the analysis is necessary since the error of the sketch over samples estimator is not simply the sum of the errors of the two individual estimators. The interaction, which is predicted by the analysis, plays a major role. The experiments also point out that in the majority of the cases a 10% sample results in minimal error degradation – the sketching of streams can thus be sped-up by a factor of 10.

In the rest of the paper, we introduce the formal problem in Section II. Section III gives an overview on sampling while Section IV introduces sketches. The formal analysis of the combined sketch over samples estimator is detailed in Section V. We discuss possible applications of sketching sampled data in Section VI. The empirical evaluation of the combined estimator is presented in Section VII, while Section VIII concludes the paper.

Related Work

There exists a large body of work on approximate query processing methods. The idea of combining two estimators to capitalize on the strengths of both is not new. F-AGMS sketches [3] are essentially a combination of random histograms and AGMS sketches. [10] presents a method to build incremental histograms from samples. To the best of our knowledge, sketching and sampling have not been combined in a principled fashion before. The main difficulty in characterizing sketches over samples is the fact that the sampling analysis [8], [11] is performed in the tuple domain while the sketch analysis [1] is performed in the frequency domain. This is the first obstacle we overcome in this paper. The work on sketching probabilistic data streams [12], [13] is somehow similar to our work. The important difference is the fact that sampling is part of the estimate in our work while it represents only a way to interpret the probabilistic data in the related work. The results in [12] do not characterize the sketch over sample estimator but approximate the probabilistic aggregates using sketches. The only overlap in terms of analysis seems to be the computation of the expected value of sketch over samples for the second frequency moment computation in [13]. [14] presents an alternative method to improve the sketching rate of a data stream by deterministically skipping stream items. In our work the items that are sketched are randomly selected through a sampling process.

II. Preliminaries

The general problem we discuss throughout the paper is how to approximate the *size of join* of two relations and the *self-join size* or second frequency moment of a relation. Let F and G be two streaming relations, each having a single attribute A, with domain I. Furthermore, let f_i and g_i be the frequency of the value i in F and G, respectively. With this, the size of join of relations F and G can be written as the dot-product of their frequency vectors:

$$|F \bowtie_A G| = \sum_{i \in I} f_i g_i \tag{1}$$

When relations F and G are identical, the quantity $\sum_{i \in I} f_i^2$ is known as the self-join size of F.

In order to compute the size of join exactly, the frequency vectors of the two relations have to be stored in full. This requires space proportional to the domain of the joining attribute A, which is infeasible for large domains, e.g., $|I| = 2^{64}$. Thus, randomized solutions with reduced space requirements and provable error guarantees have been proposed. The standard techniques [7], [15] to derive error guarantees or confidence intervals for an estimate is to compute the expected value and the variance and then to use either distribution-independent bounds given by Chernoff's and Chebyshev's inequalities, or to use distribution-dependent bounds. In the latter case, usually the Central Limit Theorem or one of its generalizations is used to argue that the distribution of the estimate has a particular shape, and then error bounds based on the assumed distribution with the same expected value and variance are computed. In order to simplify the exposition, we provide results in the form of expected values and variances throughout the paper. Actual error guarantees can be obtained straightforwardly using the mentioned techniques.

III. Sampling

Sampling as an approximation technique consists in obtaining samples F' and G' from relations F and G, respectively, computing the size of join aggregate over the samples, and applying a correction to ensure that the sampling estimator is unbiased. This method is generic and applies to all types of sampling. To simplify the theoretical exposition, we keep the treatment of sampling as generic as possible.

In Section II we expressed the size of join aggregate as a function of f_i and g_i, the frequencies of value i of the join attribute in relations F and G, respectively. If we define f'_i and g'_i to be the frequencies of i in F' and G', respectively, the size of join of the sample relations is:

$$|F' \bowtie_A G'| = \sum_{i \in I} f'_i g'_i \tag{2}$$

f'_i and g'_i are random variables that depend on the type of sampling and the parameters of the sampling process. Interestingly, a large part of the characterization of sampling can be carried out without specifying the type of sampling. This is also true for sketches over samples in Section V.

A. Generic Sampling

In general, $|F' \bowtie_A G'|$ is not an unbiased estimator for the size of join $|F \bowtie_A G|$. Fortunately, in the majority of the cases, a constant correction that *scales* for the difference in size between the samples and the original relations can be made to obtain an unbiased estimator. If we define the estimator as $X = C \sum_{i \in I} f'_i g'_i$, where C is the scaling factor, we can determine the value of C such that X is unbiased. In

order to derive error bounds for the estimator, the expectation $E[X]$ and the variance $\mathrm{Var}[X]$ have to be computed. It turns out that expressions for $E[X]$ and $\mathrm{Var}[X]$ can be written for generic sampling in terms of the moments of the frequency random variables f_i' and g_i'. There are two distinct cases that need separate treatment. The first case is when relations F and G are different and the samples are obtained independently from the two relations. The second case is when F and G are identical, thus only one sample is available. This situation arises in the case of self-join size.

When F' and G' are obtained independently, the random variables f_i' and g_i' are also independent.

Proposition 1 (Size of Join): Let $X = C\sum_{i\in I} f_i'g_i'$ be the estimator for the size of join defined over the generic samples F' and G'. Then:

$$E[X] = C\sum_{i\in I} E[f_i']\,E[g_i']$$
$$\mathrm{Var}[X] = C^2\left[\sum_{i\in I}\sum_{j\in I} E\left[f_i'f_j'\right]E\left[g_i'g_j'\right] - \left(\sum_{i\in I} E[f_i']\,E[g_i']\right)^2\right] \quad (3)$$

When F and G are identical and only the sample F' is available, the random variables f_i' and g_i' are also identical.

Proposition 2 (Self-Join Size): Let $X = C\sum_{i\in I} f_i'^2$ be the estimator for the self-join size defined over the generic sample F'. Then the expectation and the variance of X are given by:

$$E[X] = C\sum_{i\in I} E\left[f_i'^2\right]$$
$$\mathrm{Var}[X] = C^2\left[\sum_{i\in I}\sum_{j\in I} E\left[f_i'^2f_j'^2\right] - \left(\sum_{i\in I} E\left[f_i'^2\right]\right)^2\right] \quad (4)$$

A specific type of sampling determines the values of the expectations that appear in the formulas. These expectations are moments of the frequency random variables. They can be derived from the *moment generating function* corresponding to the sampling process. For the types of sampling we consider in this paper the moment generating function is well known (see for example [16]). Deriving final formulas for $E[X]$ and $\mathrm{Var}[X]$ and determining the constant C might seem just a matter of plugging in these quantities for a specific type of sampling, but the actual process is intricate because the frequency moments have to be separately computed before the algebraic manipulations are carried out.

The advantage of analyzing sampling in the frequency domain, as we do in this section, is that it allows the analysis to be extended to sketches over samples. Sampling estimators like the ones considered here have been analyzed in the published literature (the theory in [9] provides the analysis for all types of simple uniform sampling and for an arbitrary number of relations), but the analysis is in the tuple domain not the frequency domain. Interestingly, the analysis in the frequency domain is simpler than the analysis in the tuple domain since, as we show in this paper, the frequency random variables have easily identifiable distributions for which the moment generating functions are available. Deriving formulas for expected value and variance becomes just a matter of carrying the necessary algebraic manipulations.

We consider three types of sampling: Bernoulli sampling, sampling with replacement, and sampling without replacement. We derive the formulas for expectation and variance as a function of the sampling frequencies both for sampling and for sketching over samples (Section V). This parallel treatment simplifies the interpretation of the complex results obtained for sketches over samples.

B. Bernoulli Sampling

When the sampling process is Bernoulli, each tuple in F and G is selected independently in the sample F' and G' with probability p or q, $0 \le p \le 1$, $0 \le q \le 1$, respectively. Then, f_i' and g_i' are independent binomial random variables [16], $f_i' = \mathrm{Binomial}(f_i, p)$ and $g_i' = \mathrm{Binomial}(g_i, q)$, respectively, with expected values:

$$E[f_i'] = pf_i, E[g_i'] = qg_i \quad (5)$$

The scaling factor for the size of join estimator is in this case $C = \frac{1}{pq}$. The expectation and the variance for Bernoulli sampling can be derived in a straightforward manner using the frequency moments of the binomial random variables corresponding to the sampling frequencies.

Proposition 3 (Size of Join): Let $X = \frac{1}{pq}\sum_{i\in I} f_i'g_i'$ be the estimator for the size of join defined over the Bernoulli samples F' and G'. Then the expectation and the variance of X are given by:

$$E[X] = \sum_{i\in I} f_ig_i$$
$$\mathrm{Var}[X] = \frac{1-p}{p}\sum_{i\in I} f_ig_i^2 + \frac{1-q}{q}\sum_{i\in I} f_i^2g_i + \frac{(1-p)(1-q)}{pq}\sum_{i\in I} f_ig_i \quad (6)$$

The situation is more complicated for self-join size because the generic estimator $X = C\sum_{i\in I} f_i'^2$ has a bias that cannot be corrected by simply multiplying with a scaling factor. Nevertheless, the generic formula for the variance in Equation (4) is still applicable.

Proposition 4 (Self-Join Size): Let $X = \frac{1}{p^2}\sum_{i\in I} f_i'^2 - \frac{1-p}{p^2}\sum_{i\in I} f_i'$ be the estimator for self-join size defined over the Bernoulli sample F'. Then:

$$E[X] = \sum_{i\in I} f_i^2$$
$$\mathrm{Var}[X] = \frac{1-p}{p^3}\left[4p^2\sum_{i\in I} f_i^3 + 2p(1-3p)\sum_{i\in I} f_i^2 - p(2-3p)\sum_{i\in I} f_i\right] \quad (7)$$

C. Notation for Sampling Coefficients

In order to write compact formulas for sampling with and without replacement, we use the following notation throughout the rest of the paper:

$$\begin{aligned} \alpha &= \frac{|F'|}{|F|}, \quad \alpha_1 = \frac{|F'|-1}{|F|-1}, \quad \alpha_2 = \frac{|F'|-1}{|F|} \\ \beta &= \frac{|G'|}{|G|}, \quad \beta_1 = \frac{|G'|-1}{|G|-1}, \quad \beta_2 = \frac{|G'|-1}{|G|} \end{aligned} \tag{8}$$

α and β are the sampling fractions from relations F and G, respectively. $\alpha_1, \alpha_2, \beta_1, \beta_2$ are just small variations that appear in formulas.

D. Sampling with replacement

A sample of fixed size can be generated by repeatedly choosing a random tuple from the base relation for the specified number of times. If the same tuple can appear in the sample multiple times, the process is sampling with replacement. In this case the random variables corresponding to the frequencies in the sample, f_i' and g_i', respectively, are the components of a multinomial random variable [16] with parameters the size of the sample and the probability $\frac{f_i}{|F|}$ and $\frac{g_i}{|G|}$, respectively, where $|F|$ and $|G|$ are the size of F and G. Since each component of a multinomial random variable is a binomial random variable, the expectations in Equation 5 still hold but with different probabilities:

$$E\left[f_i'\right] = \alpha f_i, E\left[g_i'\right] = \beta g_i \tag{9}$$

The exact formulas for expectation and variance can be derived as for Bernoulli sampling. The moments of a multinomial random variable have to be used instead.

Proposition 5 (Size of Join): Let $X = \frac{1}{\alpha\beta}\sum_{i\in I} f_i' g_i'$ be the estimator for the size of join defined over the samples with replacement F' and G'. Then the expectation and the variance of X are given by:

$$\begin{aligned} E\left[X\right] &= \sum_{i\in I} f_i g_i \\ \mathrm{Var}\left[X\right] &= \frac{1}{\alpha\beta}\left[\sum_{i\in I} f_i g_i + |F|\alpha\beta_2 \sum_{i\in I} f_i g_i^2 \right. \\ &\left. +|G|\alpha_2\beta \sum_{i\in I} f_i^2 g_i + (\alpha_2\beta_2 - \alpha\beta)\left(\sum_{i\in I} f_i g_i\right)^2\right] \end{aligned} \tag{10}$$

An unbiased estimator for self-join size defined over the sample with replacement F' is $X = \frac{1}{\alpha\alpha_2}\sum_{i\in I} f_i'^2 - \frac{1}{\alpha_2}|F|$. Notice that the estimator depends only on the size of the base relation and the size of the sample. $\mathrm{Var}\left[X\right]$ can be derived from the formula of the variance for generic sampling in Equation (4). We omit the actual formula here due to lack of space.

E. Sampling without replacement

A sample without replacement from a relation consists of a random subset of tuples selected from the relation. The difference between sampling with replacement and sampling without replacement is that a tuple can appear at most once in a sample without replacement while it can appear multiple times in a sample with replacement. In this case the random variables corresponding to the frequencies in the sample, f_i' and g_i', respectively, are the components of a multivariate hypergeometric random variable [16]. In order to derive the exact formulas for expectation and variance, the actual moments of the multivariate hypergeometric distribution have to be plugged in.

Proposition 6 (Size of Join): Let $X = \frac{1}{\alpha\beta}\sum_{i\in I} f_i' g_i'$ be the estimator for the size of join defined over the samples without replacement F' and G'. Then the expectation and the variance of X are given by:

$$\begin{aligned} E\left[X\right] &= \sum_{i\in I} f_i g_i \\ \mathrm{Var}\left[X\right] &= \frac{1}{\alpha\beta}\left[(1-\alpha_1)(1-\beta_1)\sum_{i\in I} f_i g_i + (1-\alpha_1)\beta_1 \sum_{i\in I} f_i g_i^2 \right. \\ &\left. +\alpha_1(1-\beta_1)\sum_{i\in I} f_i^2 g_i + (\alpha_1\beta_1 - \alpha\beta)\left(\sum_{i\in I} f_i g_i\right)^2\right] \end{aligned} \tag{11}$$

The only difference between sampling with replacement and sampling without replacement is the coefficients of the terms appearing in the variance formula. While the variance of sampling without replacement becomes 0 when the entire relation is sampled, the variance of sampling with replacement never becomes 0.

An unbiased estimator for self-join size defined over the sample without replacement F' is $X = \frac{1}{\alpha\alpha_1}\sum_{i\in I} f_i'^2 - \frac{1-\alpha_1}{\alpha_1}|F|$. The variance of X can be derived from the formula of the variance for generic sampling in Equation (4). We do not include the actual formula here.

IV. Sketches

While sampling techniques select a random subset of tuples from the input relation, sketching techniques summarize all the tuples as a small number of random variables. This is accomplished by projecting the domain of the input relation on a significantly smaller domain using random functions. Multiple sketching techniques are proposed in the literature for estimating the size of join and the second frequency moment (see [4] for details). Although using different random functions, i.e., $\{+1, -1\}$ or hashing, the existing sketching techniques have similar analytical properties, i.e., the sketch estimators have the same variance. For this reason we focus on the basic AGMS sketches [1], [2] throughout the paper.

The basic AGMS sketch of relation F consists of a single random variable S that summarizes all the tuples t from F. S is defined as:

$$S = \sum_{t\in F} \xi_{t.A} = \sum_{i\in I} f_i \xi_i \tag{12}$$

where ξ is a family of $\{+1, -1\}$ random variables that are 4–wise independent. Essentially, a random value of either $+1$ or -1 is associated to each point in the domain of attribute A.

Then, the corresponding random value is added to the sketch S for each tuple t in the relation. We can define a sketch T for relation G in a similar way and using the same family ξ.

Proposition 7 (Size of Join): The sketch-based estimator X defined as:

$$X = S \cdot T = \sum_{i \in I} f_i \xi_i \cdot \sum_{j \in I} g_j \xi_j \tag{13}$$

is an unbiased estimator for the size of join $|F \bowtie_A G|$. The variance of the sketch estimator is given by:

$$\text{Var}[X] = \sum_{i \in I} f_i^2 \sum_{j \in I} g_j^2 + \left(\sum_{i \in I} f_i g_i \right)^2 - 2 \sum_{i \in I} f_i^2 g_i^2 \tag{14}$$

Proposition 8 (Self-Join Size): The unbiased estimator for the self-join size is defined as:

$$X = S^2 = \sum_{i \in I} \sum_{j \in I} f_i f_j \xi_i \xi_j \tag{15}$$

The variance of the sketch estimator is given by:

$$\text{Var}[X] = 2 \left[\left(\sum_{i \in I} f_i^2 \right)^2 - \sum_{i \in I} f_i^4 \right] \tag{16}$$

A common technique to reduce the variance of an estimator is to generate multiple independent instances of the basic estimator and then to build a more complex estimator as the average of the basic estimators. While the expected value of the complex estimator is equal with the expectation of one basic estimator, the variance is reduced by a factor of n since $\text{Var}\left[\frac{1}{n} \cdot \sum_{k=1}^{n} X_k\right] = \frac{1}{n^2} \sum_{k=1}^{n} \text{Var}[X_k] = \frac{\text{Var}[X_k]}{n}$, where n is the number of basic estimators being averaged. This technique can be applied to reduce the variance of the sketch estimator if different families ξ are used for the basic estimators (see [1], [2] for details).

V. Sketches over Samples

Given the ability of sampling to make predictions about an entire dataset from a randomly selected subset and that sketches require the entire dataset in order to determine any of its properties, an interesting question that immediately arises is how to combine these two randomized techniques. Although the intuitive answer to this question seems to be simple – the sketch is computed over a sample of the data instead of the entire dataset – the behavior of the combined estimator is not the simple composition of the individual behavior of the ingredients. A careful analytical characterization of the estimator needs to be carried out. Furthermore, the sampling process can be either explicit and executed as an individual step before sketching is done, or implicit, situation in which the input dataset is assumed to be a sample from a large population. In the first case, a significant speed-up in updating the sketch structure can be obtained since only a random subset of the data is actually sketched. This process is essentially a load shedding technique for sketching extremely fast data streams that cannot be otherwise sketched. It can be implemented as an explicit Bernoulli sampling that randomly filters the tuples that update the sketch structure. In the second case, the data is assumed to be a sample from a large population and the goal is to determine properties of the population based on the sample. The sample itself is assumed to be large enough so it cannot be stored explicitly, thus sketching is required. If the population is infinite, the entire process can be seen as sketching i.i.d. samples from an unknown distribution. We provide the analysis both for sampling with replacement and sampling without replacement. The analysis straightforwardly extends to i.i.d. samples if all estimators are normalized by the size of the population and the limit, when the population size goes to infinity, is taken. In such a circumstance, the frequencies in the original unknown population become densities of the unknown population, but everything else remains the same.

In this section, we provide a generic framework for sketching sampled data streams in order to estimate the size of join and the self-join size. Then we compute the first two frequency moments of the combined estimator for the most common types of sampling – Bernoulli sampling, sampling with replacement, and sampling without replacement. This provides sufficient information to allow the derivation of confidence bounds for the combined estimator.

A. Sketches over Generic Sampling

Consider F' to be a generic sample obtained from relation F. Sketching the sample F' is similar to sketching the entire relation F and consists in summarizing the sampled tuples t' as follows:

$$S = \sum_{t' \in F'} \xi_{t'.A} = \sum_{i \in I} f_i' \xi_i \tag{17}$$

where ξ is a family of $\{+1, -1\}$ random variables that are 4–wise independent. A sample G' from relation G can be sketched in a similar way using the same family ξ:

$$T = \sum_{t' \in G'} \xi_{t'.A} = \sum_{i \in I} g_i' \xi_i$$

Size of Join

We define the estimator X for the size of join $|F \bowtie_A G|$ based on the sketches computed over the samples as follows:

$$X = C \cdot ST = C \cdot \sum_{i \in I} f_i' \xi_i \cdot \sum_{j \in I} g_j' \xi_j \tag{18}$$

Notice that the estimator is similar to the sketch estimator computed over the entire dataset in Proposition 7 multiplied with a constant scaling factor C that compensates for the difference in size.

Self-Join Size

The self-join size or second frequency moment of a relation is the particular case of size of join between two instances of the same relation. One way of analyzing the sketches over samples estimator for the self-join size problem is to build two independent samples and two independent sketches from the same base relation and then to apply the results corresponding to size of join. Although sound from an analytical point of

view, this solution is inefficient in practice. In the following we consider a practical solution that requires the construction of only one sample and one sketch from the base relation. A new estimator for the self-join size has to be defined instead, but the analysis is closely related to the analysis of the size of join estimator. With S defined in Equation 17, we define the self-join size estimator X as follows:

$$X = S^2 = C \cdot \left(\sum_{i \in I} f'_i \xi_i \right)^2 = C \cdot \sum_{i \in I} f'_i \xi_i \cdot \sum_{j \in I} f'_j \xi_j \quad (19)$$

where C is the same scaling factor compensating for the difference in size. Notice that the difference between the size of join estimator and the self-join size estimator is only at the sampling level since the same family of ξ random variables is used for sketching in both cases. For this reason we carry out the analysis for the two estimators in parallel and make the distinction only when necessary.

In order to derive confidence bounds for the estimator X, the first two moments, expected value and variance, have to be computed. Intuitively, the scaling factor C should compensate for the difference in size and make the estimator unbiased. Since the two processes, sampling and sketching, are independent and sequential, the interaction between them is minimal and the sum of the two variances should be a good estimator for the variance of the combined estimator. In the following, we derive the exact formulas for the expectation and the variance in the generic case. The independence of the families of random variables corresponding to sampling and sketching, f'_i, g'_i and ξ, respectively, plays an important role in simplifying the computation. This independence is due to the independence of the two random processes. While the computation of the expectation is straightforward, the computation of the variance is more intricate since the interaction between sketching and sampling is more complex and it can be characterized only through a detailed analysis.

The first step in our analysis is to derive the formulas for the moments of the basic estimator. Proposition 9 and 10 characterize the behavior of the basic sketch over samples estimator.

Proposition 9 (Size of Join): Let the sketch over samples estimator for the size of join to be defined as $X = C \cdot \sum_{i \in I} f'_i \xi_i \cdot \sum_{j \in I} g'_j \xi_j$. Then, the expectation and the variance of X are given by:

$$E[X] = C \cdot \sum_{i \in I} E[f'_i] E[g'_i]$$

$$\text{Var}[X] = C^2 \cdot \left[\sum_{i \in I} E\left[f'^2_i\right] \sum_{j \in I} E\left[g'^2_j\right] + 2 \cdot \sum_{i \in I} \sum_{j \in I} E\left[f'_i f'_j\right] E\left[g'_i g'_j\right] - 2 \cdot \sum_{i \in I} E\left[f'^2_i\right] E\left[g'^2_i\right] - \left(\sum_{i \in I} E[f'_i] E[g'_i] \right)^2 \right] \quad (20)$$

Proposition 10 (Self-Join Size): Let the sketch over samples estimator for the self-join size to be defined as $C \cdot \sum_{i \in I} f'_i \xi_i \cdot \sum_{j \in I} f'_j \xi_j$. Then, the expectation and the variance of X are given by:

$$E[X] = C \cdot \sum_{i \in I} E\left[f'^2_i\right]$$

$$\text{Var}[X] = C^2 \left[3 \cdot \sum_{i \in I} \sum_{j \in I} E\left[f'^2_i f'^2_j\right] - 2 \cdot \sum_{i \in I} E\left[f'^4_i\right] - \left(\sum_{i \in I} E\left[f'^2_i\right] \right)^2 \right] \quad (21)$$

Notice that the variance of sketching over generic sampling is an expression depending only on the properties of the sampling process. More precisely, in order to evaluate the variance, only expectations of the form $E[f'_i]$ and $E\left[f'_i f'_j\right]$ have to be computed, where f'_i and f'_j are random variables corresponding to the frequencies in the sample.

The averaging technique applied to reduce the variance of basic sketches in Section IV cannot be used straightforwardly in the case of sketches computed over samples. This is the case since, although the basic sketch estimators are built independently using different ξ families of random variables, they are computed over the same sample and this introduces correlations between any two estimators. The variance of the average estimator is in this case:

$$\text{Var}\left[\frac{1}{n} \cdot \sum_{k=1}^{n} X_k \right] = \frac{1}{n} \left[\text{Var}[X_k] + (n-1) \cdot \text{Cov}_{k \neq l}[X_k, X_l] \right] \quad (22)$$

where n is the number of basic estimators being averaged and $\text{Cov}[X_k, X_l] = E[X_k X_l] - E[X_k] E[X_l]$ is the covariance between any two basic estimators.

The next step in our analysis is to derive the formulas for the variance of the average sketch over samples estimator. Proposition 11 and 12 contain the derived formulas.

Proposition 11 (Size of Join): The variance of the average sketch over samples size of join estimator is given by:

$$\text{Var}\left[\frac{1}{n} \cdot \sum_{k=1}^{n} X_k \right] = C^2 \cdot \left[\sum_{i \in I} \sum_{j \in I} E\left[f'_i f'_j\right] E\left[g'_i g'_j\right] - \left(\sum_{i \in I} E[f'_i] E[g'_i] \right)^2 + \frac{1}{n} \left(\sum_{i \in I} E\left[f'^2_i\right] \sum_{j \in I} E\left[g'^2_j\right] + \sum_{i \in I} \sum_{j \in I} E\left[f'_i f'_j\right] E\left[g'_i g'_j\right] - 2 \cdot \sum_{i \in I} E\left[f'^2_i\right] E\left[g'^2_i\right] \right) \right] \quad (23)$$

Essentially, the variance of the average estimator is the sum of the variance of the generic sampling estimator in Equation 3, the variance of the sketch estimator in Equation 14, and a

term corresponding to the interaction between the two random processes. For the particular types of sampling considered in this work, we derive the exact formula of the interaction term. Notice that the improvement obtained by averaging is less significant than a factor of n obtained in the case of independent estimators.

Proposition 12 (Self-Join Size): The variance of the average sketch over samples self-join size estimator is given by:

$$\text{Var}\left[\frac{1}{n}\cdot\sum_{k=1}^{n}X_k\right] = C^2\cdot\left[\sum_{i\in I}\sum_{j\in I}E\left[f_i'^2f_j'^2\right] - \left(\sum_{i\in I}E\left[f_i'^2\right]\right)^2 + \frac{2}{n}\left(\sum_{i\in I}\sum_{j\in I}E\left[f_i'^2f_j'^2\right] - \sum_{i\in I}E\left[f_i'^4\right]\right)\right] \tag{24}$$

The independence of sketching and sampling plays an important role in deriving the formulas for expectation and variance. The independence has the effect of factorizing the expectations over products of random variables corresponding to sketching and sampling. Thus, only expectations involving the sampling frequency random variables appear in the final formulas since the expectations corresponding to sketches are either 0 or 1, i.e., $E\left[\xi_i\xi_j\right] = E\left[\xi_i\right]\cdot E\left[\xi_j\right] = 0$ whenever $i \neq j$ due to the 4–wise independence of the family ξ, and $E\left[\xi_i^2\right] = 1$. Using these equalities and some complex algebraic manipulations, the given formulas are obtained. The dominant factor that simplifies the analysis is the modeling of sampling as frequency random variables. This is our main contribution.

B. Bernoulli Sampling

We instantiate the formulas derived for generic sampling in Section V-A with the moments of the binomial random variables corresponding to the Bernoulli sampling frequencies.

Proposition 13 (Size of Join): Let the unbiased sketch over Bernoulli samples size of join estimator to be defined as $X = \frac{1}{pq}\sum_{i\in I}f_i'\xi_i\cdot\sum_{j\in I}g_j'\xi_j$. Then, the variance of the average estimator is given by:

$$\begin{aligned}\text{Var}\left[\frac{1}{n}\cdot\sum_{k=1}^{n}X_k\right] = {}& \frac{1-p}{p}\sum_{i\in I}f_ig_i^2 + \frac{1-q}{q}\sum_{i\in I}f_i^2g_i + \frac{(1-p)(1-q)}{pq}\sum_{i\in I}f_ig_i \\ & + \frac{1}{n}\left[\sum_{i\in I}f_i^2\sum_{j\in I}g_j^2 + \left(\sum_{i\in I}f_ig_i\right)^2 - 2\sum_{i\in I}f_i^2g_i^2\right] \\ & + \frac{1}{n}\left[\frac{1-p}{p}\sum_{i\in I}\sum_{j\in I,j\neq i}f_ig_j^2 + \frac{1-q}{q}\sum_{i\in I}\sum_{j\in I,j\neq i}f_i^2g_j + \frac{(1-p)(1-q)}{pq}\sum_{i\in I}\sum_{j\in I,j\neq i}f_ig_j\right]\end{aligned} \tag{25}$$

Proposition 14 (Self-Join Size): The variance of the average unbiased self-join size estimator $X = \frac{1}{p^2}\left(\sum_{i\in I}f_i'\xi_i\right)^2 - \frac{1-p}{p^2}\sum_{i\in I}f_i'$ is given by:

$$\begin{aligned}\text{Var}\left[\frac{1}{n}\cdot\sum_{k=1}^{n}X_k - \frac{1-p}{p^2}\sum_{i\in I}f_i'\right] = {}& \frac{1-p}{p^3}\left[4p^2\sum_{i\in I}f_i^3 + 2p(1-3p)\sum_{i\in I}f_i^2 - p(2-3p)\sum_{i\in I}f_i\right] \\ & + \frac{2}{n}\left[\left(\sum_{i\in I}f_i^2\right)^2 - \sum_{i\in I}f_i^4\right] \\ & + \frac{2}{n}\left[\frac{(1-p)^2}{p^2}\sum_{i\in I}\sum_{j\in I,j\neq i}f_if_j + \frac{2(1-p)}{p}\sum_{i\in I}\sum_{j\in I,j\neq i}f_i^2f_j\right]\end{aligned} \tag{26}$$

The variance of the average estimator is, as derived for generic sampling, the sum of the average sketch estimator individual variance, the Bernoulli sampling estimator individual variance, and an interaction term. Since deriving an exact analytical relation between these terms is a daunting task, we consider some extreme scenarios that allow a partial characterization. First notice that n, the number of estimators being averaged, can be ignored when comparing the sketch variance and the interaction variance because it has the same effect on both (the variance is reduced by a factor of n). When the distribution of the frequencies is uniform, the interaction variance is the dominant term whenever the unique frequency has a smaller value than the size of the domain $|I|$. At the other extreme, when the distribution of the frequencies is skewed, the sketch variance is the dominant term by far. These results suggest that the interaction variance could represent a problem for uniform-like data. This is not necessarily the case because the value of the variance for uniform distributions is significantly smaller than for skewed data, thus, although the interaction variance is the dominant term, its absolute value is not large.

To better understand the exact significance of each of the terms appearing in the variance and to confirm the analysis for the extreme cases, we designed a set of simulations to determine the relative contribution of each of the terms. The experimental setup is described in Section VII. Figure 1 and 2 depict the relative contribution of each of the three terms appearing in the variance of the average estimator over Bernoulli samples (Equation 26 and 25). The relative contribution is represented as a function of the data skew for different sampling probabilities. A common trend both for size of join and self-join size is that the interaction term is highly significant for low skew data. This completely justifies the analysis we develop throughout the paper since an analysis assuming that the variance of the composed estimator is the sum of the variances of the basic estimators would be incorrect. At the same time, this suggests that the accuracy of the sketch over samples estimator can be significantly

worse than the accuracy of the sketch estimator for non-skewed data. As already explained, this is not necessarily true. Moreover, the experimental results we provide in Section VII show that this is not the case for practical scenarios. As expected, the impact of the variance of the sampling estimator is more significant as the size of the sample is smaller. For self-join size (Figure 2), the variance is dominated by the term corresponding to the sampling estimator, while for size of join (Figure 1) the variance of the sketch estimator quantifies for almost the entire variance irrespective of the sampling probability. This is entirely supported by the existing theoretical results which show that sketches are optimal for estimating the second frequency moment while sampling is optimal for the estimation of size of join [2].

C. Sampling with replacement

In a similar way to Bernoulli sampling, we instantiate the formula for the size of join estimator derived for generic sampling with the moments of the multinomial random variables corresponding to the sampling frequencies. The interaction between two different sampling frequencies makes the derivation of the formula more complicated. We do not provide the formula for self-join size variance due to space constraints.

Proposition 15 (Size of Join): Let the unbiased sketch over samples with replacement size of join estimator to be defined as $X = \frac{1}{\alpha\beta}\sum_{i\in I} f_i'\xi_i \cdot \sum_{j\in I} g_j'\xi_j$. Then, the variance of the average estimator is given by:

$$\begin{aligned}\text{Var}\left[\frac{1}{n}\cdot\sum_{k=1}^{n}X_k\right] = \frac{1}{\alpha\beta}\Bigg[&\sum_{i\in I} f_i g_i + |F|\alpha\beta_2\sum_{i\in I} f_i g_i^2 \\ &+|G|\alpha_2\beta\sum_{i\in I} f_i^2 g_i + (\alpha_2\beta_2-\alpha\beta)\left(\sum_{i\in I} f_i g_i\right)^2\Bigg] \\ &+\frac{1}{n}\frac{\alpha_2}{\alpha}\frac{\beta_2}{\beta}\left[\sum_{i\in I} f_i^2\sum_{j\in I} g_j^2 + \left(\sum_{i\in I} f_i g_i\right)^2 - 2\sum_{i\in I} f_i^2 g_i^2\right] \\ &+\frac{1}{n}\frac{1}{\alpha\beta}\Bigg[\sum_{i\in I}\sum_{j\in I, j\neq i} f_i g_j + |F|\alpha\beta_2\sum_{i\in I}\sum_{j\in I, j\neq i} f_i g_j^2 \\ &\qquad +|G|\alpha_2\beta\sum_{i\in I}\sum_{j\in I, j\neq i} f_i^2 g_j\Bigg]\end{aligned} \tag{27}$$

D. Sampling without replacement

We apply the same procedure for sampling without replacement. We obtain similar results to the other types of sampling – the terms in the variance are the same, only the coefficients are dependent on the sampling procedure. The formula for self-join size is not provided due to space limitations.

Proposition 16 (Size of Join): Let the unbiased sketch over samples without replacement size of join estimator to be defined as $X = \frac{1}{\alpha\beta}\sum_{i\in I} f_i'\xi_i \cdot \sum_{j\in I} g_j'\xi_j$. Then, the variance of the average estimator is given by:

$$\begin{aligned}\text{Var}\left[\frac{1}{n}\cdot\sum_{k=1}^{n}X_k\right] = &\\ \frac{1}{\alpha\beta}\Bigg[&(1-\alpha_1)(1-\beta_1)\sum_{i\in I} f_i g_i + (1-\alpha_1)\beta_1\sum_{i\in I} f_i g_i^2 \\ &+\alpha_1(1-\beta_1)\sum_{i\in I} f_i^2 g_i + (\alpha_1\beta_1-\alpha\beta)\left(\sum_{i\in I} f_i g_i\right)^2\Bigg] \\ &+\frac{1}{n}\frac{\alpha_1}{\alpha}\frac{\beta_1}{\beta}\left[\sum_{i\in I} f_i^2\sum_{j\in I} g_j^2 + \left(\sum_{i\in I} f_i g_i\right)^2 - 2\sum_{i\in I} f_i^2 g_i^2\right] \\ &+\frac{1}{n}\frac{1}{\alpha\beta}\Bigg[(1-\alpha_1)(1-\beta_1)\sum_{i\in I}\sum_{j\in I, j\neq i} f_i g_j \\ &+(1-\alpha_1)\,\beta_1\sum_{i\in I}\sum_{j\in I, j\neq i} f_i g_j^2 + \alpha_1\,(1-\beta_1)\sum_{i\in I}\sum_{j\in I, j\neq i} f_i^2 g_j\Bigg]\end{aligned} \tag{28}$$

E. Discussion

The result we derived in this section is somewhat surprising: the variance of the combined sketch-sampling estimator can be written as the sum of the variance of the sketch estimator, the variance of the sampling estimator, and an interaction term. This separation of the variance formula was accomplished for all three types of sampling and both size of join and self-join-size problems. Although the sketch variance seems to be the dominant term, the exact significance of each of the terms is dependent on the actual distribution of the data (see the experimental results for Bernoulli sampling). When multiple sketch estimators are averaged in order to decrease the variance, the covariance must also be considered since the sketch estimators are computed over the same sample, thus they are correlated. Our results show that the variance of the combined estimator does not decrease by a factor equal to the number of averages, only the sketch variance and the interaction term do.

VI. Applications

In this section, we identify applications for sketching sampled data for each of the types of sampling discussed in the paper. We also delve into the algorithmic issues corresponding to the combined randomized process.

A. Bernoulli Sampling

Even though sketching can be implemented for fairly high-speed data streams, the update time could still become the limiting factor if all the tuples need to be sketched. In order to alleviate this problem, some of the tuples need to be dropped. Bernoulli sampling provides a principled approach to drop tuples and still be able to characterize the result. Sketching Bernoulli samples is a method to further increase the rate of the data streams that can be sketched. Used along hashing, sketching over Bernoulli samples allows the processing of

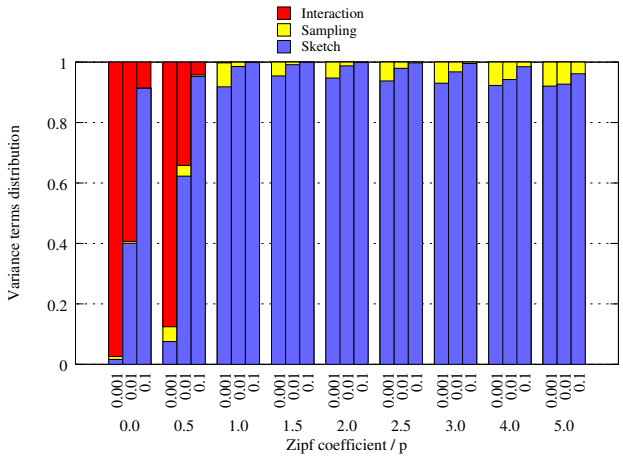

Fig. 1. Size of join variance

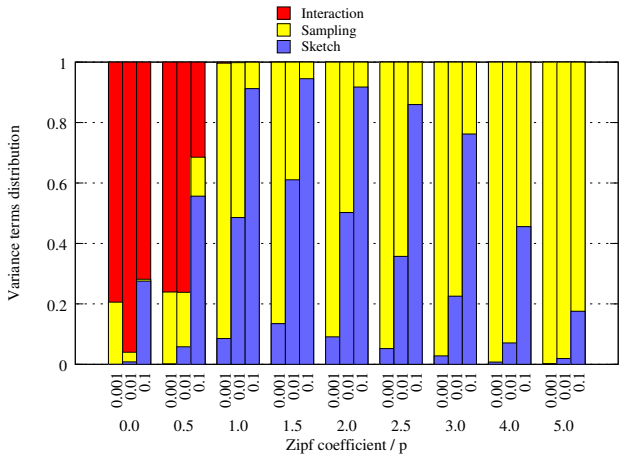

Fig. 2. Self-join size variance

data streams having arrival rates of billions of tuples per second encountered in the existing networking equipment. The analysis in Section V-B provides a complete characterization of the error behavior for the overall process. The experimental results in Section VII-A show that for a sampling fraction of 1% the decrease in accuracy over sketching the entire data stream is insignificant.

The basic Bernoulli sampling algorithm consists in generating a random number for each tuple in the dataset. The tuples for which the random value is smaller than the sampling probability are included in the sample. The main drawback of Bernoulli sampling is that the size of the sample is unknown prior to running the process. This is not a problem anymore when the sample is sketched instead of being explicitly stored.

The algorithm for sketching Bernoulli samples is a simple extension of the basic Bernoulli sampling algorithm. The tuples that are selected in the sample are inserted into the sketch data structure instead of being explicitly stored. Since updating the sketch is an extremely fast operation [17], it could be doubtful that the extra sampling is indeed beneficial for speeding the entire process. Fortunately, there exist algorithms that avoid the coin tossing for each tuple by generating the intervals between the tuples that are selected in the sample [18]. This way there is work to be done only for the tuples that are actually sampled and subsequently sketched. In this situation the speedup over sketching the entire data stream is clearly proportional with the sampling fraction.

B. Sampling with replacement

Consider a generative model that draws samples from a finite population. The samples are draw with replacement. A data stream containing all the samples is generated. The objective is to determine properties, e.g., the second frequency moment, of the generative model, or correlations between two different generative models, e.g., the size of join, from the stream of samples. An additional requirement is that the stream is large enough that it cannot be stored in full. These kind of scenarios are frequent in data-mining applications [6].

Sketching the stream of samples obtained with replacement from the finite population whose size is known represents a solution to determining properties of the generative model. The input stream is the actual sample in this case, so no explicit sampling of the stream is required. Thus, the standard updating algorithm for sketches can be used in this case. The estimation algorithm is though different because it has to take into consideration that the stream is only a sample. The formulas derived in Section V-C have to be used for estimation in this case. The important question in this case is how accurate is the combined estimator. And how large has to be the sketched sample in order to obtain accurate estimations. The experimental results in Section VII-B show that for a sampling fraction of 10% the estimation is accurate and stable. No significant increase in accuracy is obtained if the sample size is larger.

C. Sampling without replacement

The application we have in mind for sketching samples obtained without replacement is online aggregation. In a traditional database engine, the exact answer to a query is provided only after the entire data is processed. The user does not get any clues about the result during the query execution. This may take a long time for complex queries over a datawarehouse. Online aggregation has a different strategy. Partial approximate answers are provided to the user while the query is processed by executing equivalent queries on a smaller fraction of the entire data and then scaling the result. As more data is processed, the accuracy of the approximate result increases to the point where the exact answer is returned (when the entire data is processed). The fundamental requirement for the partial results to be estimates of the final result is that the portions of the data the equivalent queries are executed on to represent random samples without replacement from the entire data. More details on online aggregation can be found in the literature[8], [11], [9].

Sketching samples obtained without replacement represents a fast and inexpensive method to gather some of the statistics

(second frequency moment, correlation between attributes) used by an online aggregation engine to take decisions and to compute the approximate results. The idea is to build sketches for the desired statistics while the relations (materialized or intermediary results) are scanned. The fraction of the relation seen at each point during the scan represents a sample without replacement of the entire relation as long as the order of the tuples is random. More accurate estimates for the computed statistics are available as the scanning advances. The goal is to obtain stable estimators as early as possible such that the online aggregation engine takes the optimal decisions and provides estimates as accurate as possible. The analysis in Section V-D provides a complete characterization of the combined estimators. In Section VII-C we show experimental results for which accurate estimates are available after only 10% of the relations is scanned.

From an algorithmic point of view, sketching a relation while it is scanned incurs almost no additional cost. The advantage over using the samples to provide estimates is that the samples do not have to be explicitly stored and processed. There is extra memory required only to store the sketches. Sketching can be done for arrival rates of tens of millions of tuples per second without any time penalty. Or it can be executed as a separate thread in parallel with scanning, which is necessary. On the modern multi-core processors, sketching can be done essentially for free.

VII. Experimental Evaluation

We pursue two main goals in the experimental evaluation of the sketching over samples estimators. First, we want to determine the behavior of the error of the sketch over samples estimator when compared with the error of the sketch estimator. And second, we want to identify what is the behavior of the estimation error as a function of the sample size. We design experiments to determine these relations for all three types of sampling presented in the paper. And both for the size of join and the self-join size problems. In order to accomplish these goals, we designed a series of experiments over both synthetic datasets and the TPC-H dataset. Synthetic datasets allow a better control of the important parameters that affect the results, while the TPC-H dataset validates the results for large scales.

The synthetic datasets used in our experiments contain either 10 or 100 million tuples generated from a Zipfian distribution with the coefficient ranging between 0 (uniform) and 5 (skewed). The domain of the possible values is 1 million. In the case of size of join, the tuples in the two relations are generated completely independent. For the experiments over the TPC-H dataset, we used the scale 1 benchmark data. We used F-AGMS sketches [3] in all of the experiments due to their superior performance both in accuracy and update time (see [4] for details on sketching techniques). The number of buckets is either $5,000$ or $10,000$. This is equivalent to averaging $5,000$ or $10,000$ basic estimators. In order to be statistically significant, all the results presented in this section are the average of at least 100 independent experiments.

A. Bernoulli Sampling

The experimental relative error, i.e., $\frac{|\text{estimation}-\text{true result}|}{\text{true result}}$, of the sketch over Bernoulli samples estimator is depicted in Figure 3 and 4 as a function of the data skew for different sampling probabilities. Probability $p = 1.0$ corresponds to sketching the entire dataset. These experimental results show that, with some exceptions, the sampling rate does not significantly affect the accuracy of the sketch estimator. For Zipf coefficients smaller than 1, in the case of self-join size, and smaller than 3, in the case of size of join, the error of the sketch estimator is almost the same both when the entire dataset is sketched or when only one tuple out of a thousand is sketched. The impact of the sampling rate is significant only for high skew data in the case of self-join size. This is to be expected from the theoretical analysis (Figure 2). What cannot be explained from the theoretical analysis is the effect of the sampling rate for skewed data in the case of size of join. As shown in [4], the experimental behavior of F-AGMS sketches is in some cases orders of magnitude better than the theoretical predictions, thus although the theoretical variance is dominated by the variance of the sketch estimator, the empirical absolute value is small when compared to the variance of the sampling estimator. In the light of [4], the empirical results for high sampling rates are much better than the theoretical predictions, increasing thus the significance of the sampling rate for highly skewed data.

B. Sampling with replacement

In Figure 5 and 6 we depict the experimental relative error as a function of the sample size for sampling with replacement. Since the actual size of the sample is different for different Zipf coefficients, we represent on the x axis the size of the sample as a fraction from the population size, with 1 corresponding to a sample with replacement of size equal to the population size. As expected, the error is decreasing as the sample size becomes larger, but it stabilizes after a certain sample size (a 0.1 fraction of the population size for the included figures). Thus, sketching more samples does not provide any increase in the accuracy after a certain point. For the situations depicted in Figure 5 and 6, the edge sampling fraction is around 0.1.

C. Sampling without replacement

We used the scale 1 TPC-H data for our experiments on sketching samples without replacement. Figure 7 depicts the error as a function of the sampling rate for the size of join between the relations *lineitem* and *orders*. In Figure 8 we plot the error of the second frequency moment of relation *lineitem* on the *l_orderkey* attribute. As expected, the error of the self-join size estimator decreases while increasing the sample size and it becomes stable for sampling rates larger than 10%. For size of join, the behavior of the error is somehow unexpected. The smallest error in Figure 7 is obtained for a sampling rate of 10%. Then, the error starts to increase while increasing the sampling rate. This behavior is due to F-AGMS sketches.

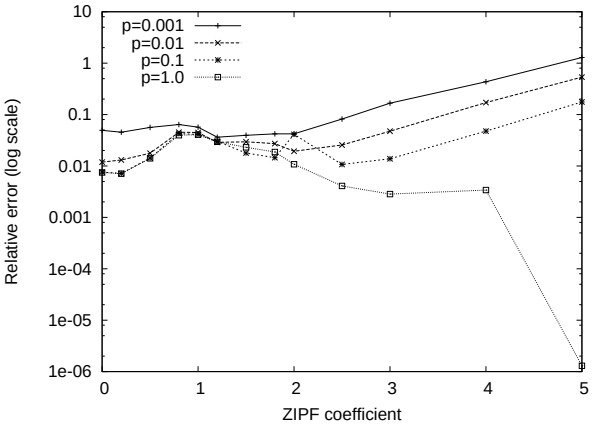

Fig. 3. Size of join error (Bernoulli)

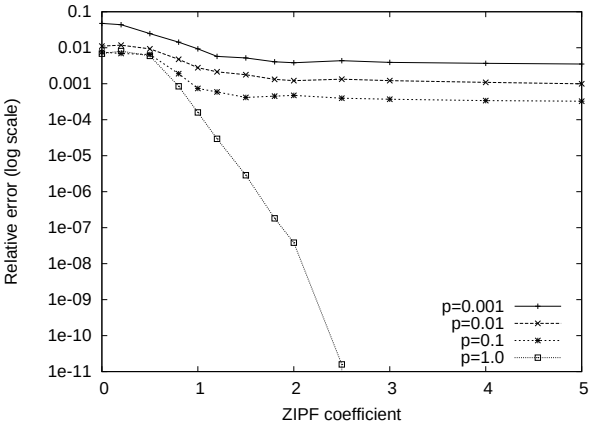

Fig. 4. Self-join size error (Bernoulli)

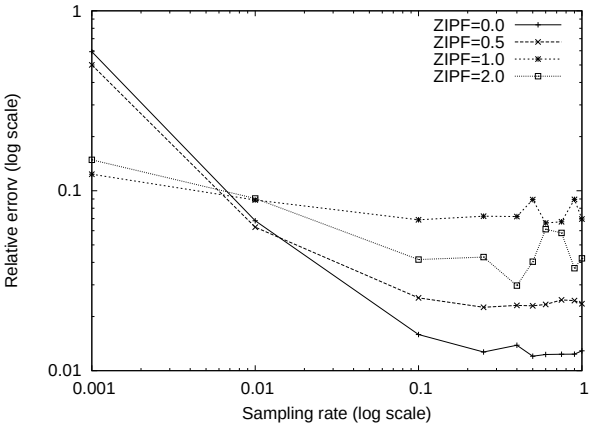

Fig. 5. Size of join sample size (WR)

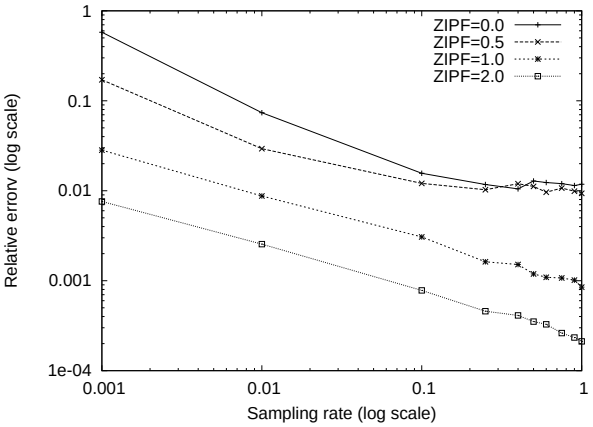

Fig. 6. Self-join size sample size (WR)

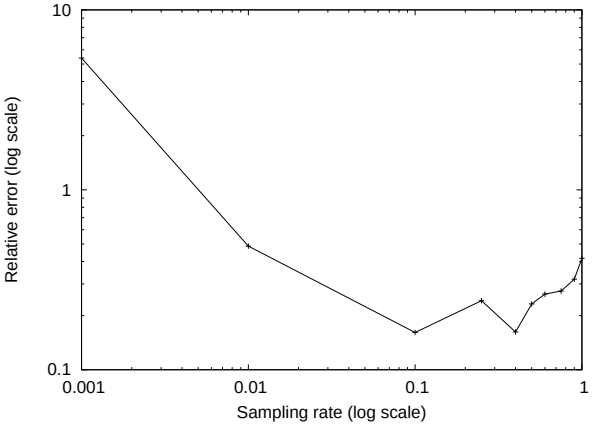

Fig. 7. Size of join sample size (WOR)

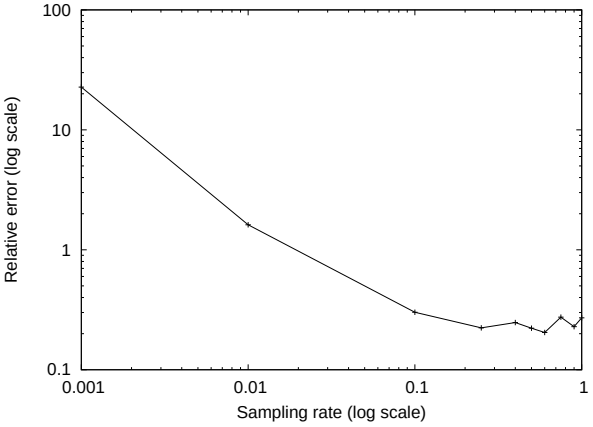

Fig. 8. Self-join size sample size (WOR)

D. Extreme behavior of F-AGMS sketches

An interesting trend in Figure 7 is the increase of the error as the sampling rate increases over 10%. This is counter intuitive since we would expect smaller error as more data is sketched. This behavior is not predicted by the theory and it must be due to the fact that the theoretical formulas are derived for AGMS sketches, but F-AGMS sketches are used in the experiments. To understand why this is happening we have to reffer to the analysis of F-AGMS sketches in [4]. F-AGMS sketches use a combination of hashes and AGMS sketches within each hash bucket. As more data is sketched, the contention in buckets increases and this produces a wider variance of the estimates. This suggests that in some situations it is better to sketch only a sample of the data rather than the entire data for F-AGMS sketches.

E. Discussion

We provide experimental results for each of the types of sampling presented in the paper. The goal is to compare the combined sketch over samples estimator with the sketch estimator and to determine their relation. Our experiments for Bernoulli sampling show that a significant speed-up (a factor of 10 in general and a factor of up to 1000 in some cases) can be obtained by sketching only a small sample of the data instead of the entire data. The decrease in accuracy due to sampling seems to be insignificant. For sampling with replacement, the difference in accuracy between sketching only a small sample (a fraction of 0.1 or less from the population size) and sketching the entire data is minimal. Moreover, the error becomes stable and it does not decrease anymore if the sample size is increased above a certain threshold (10% in our results). Although the situation is almost similar for sampling without replacement, we also observed some unexpected results for this type of sampling. The smallest error is obtained when only a sample of the data (10%) is sketched, not the entire dataset. This is due to the extreme behavior of F-AGMS sketches in some particular situations [4]. Essentially, sketching more data can actually decrease the accuracy of F-AGMS sketch estimators if the sketched sample captures well enough the distribution of the entire dataset. Summarizing, the experimental results show that the sketch over samples estimator has almost similar accuracy to the sketch estimator starting with sampling rates of 10% or smaller.

VIII. Conclusions

In this paper we introduce the sketch over samples estimator for size of join and self-join size. We provide a detailed analysis of the error of this estimator for generic sampling based on the moment generating function of the sampling frequencies. Then, we instantiate the general results for three different types of sampling: Bernoulli sampling, sampling with replacement, and sampling without replacement. Our results show that it is possible to express the variance of the combined estimator as the sum of the variance of the sampling estimator, the variance of the sketch estimator, and an interaction term. Although the sketch variance seems to be the dominant term, the exact importance of each of the terms is highly dependant on the exact distribution of the actual data. We provide experimental results that show that the accuracy of the combined estimator is almost similar (and sometimes better) to the accuracy of the sketch estimator even for small sampling rates of 10%. We also identify possible applications of the sketch over samples estimator for each type of sampling. In conclusion, we believe that the sketch over samples estimator can be used instead of the sketch estimator without a significant degradation in accuracy and with a clear gain in processing time as long as the sample rate is around 10%.

IX. Acknowledgments

Material in this paper is based upon work supported by the National Science Foundation under Grant No. 0448264. We would like to thank the anonymous referees for their useful reviews that helped us to improve the quality of the paper.

References

[1] N. Alon, Y. Matias, and M. Szegedy, "The space complexity of approximating the frequency moments," in *STOC 1996*, pp. 20–29.
[2] N. Alon, P. Gibbons, Y. Matias, and M. Szegedy, "Tracking join and self-join sizes in limited storage," *J. Comput. Syst. Sci.*, vol. 64, no. 3, pp. 719–747, 2002.
[3] G. Cormode and M. Garofalakis, "Sketching streams through the net: distributed approximate query tracking," in *VLDB 2005*, pp. 13–24.
[4] F. Rusu and A. Dobra, "Statistical analysis of sketch estimators," in *SIGMOD 2007*, pp. 187–198.
[5] N. Tatbul, U. Çetintemel, S. Zdonik, M. Cherniack, and M. Stonebraker, "Load shedding in a data stream manager," in *VLDB 2003*, pp. 309–320.
[6] P. Domingos and G. Hulten, "Mining high-speed data streams," in *KDD 2000*, pp. 71–80.
[7] J. Shao, *Mathematical Statistics*. Springer-Verlag, 1999.
[8] P. Haas and J. Hellerstein, "Ripple joins for online aggregation," in *SIGMOD 1999*, pp. 287–298.
[9] C. Jermaine, S. Arumugam, A. Pol, and A. Dobra, "Scalable approximate query processing with the dbo engine," in *SIGMOD 2007*, pp. 725–736.
[10] P. Gibbons, Y. Matias, and V. Poosala, "Fast incremental maintenance of approximate histograms," in *VLDB 1997*, pp. 466–475.
[11] C. Jermaine, A. Dobra, S. Arumugam, S. Joshi, and A. Pol, "The sort-merge-shrink join," *ACM TODS*, vol. 31, no. 4, pp. 1382–1416.
[12] G. Cormode and M. Garofalakis, "Sketching probabilistic data streams," in *SIGMOD 2007*, pp. 281–292.
[13] T. S. Jayram, A. McGregor, S. Muthukrishnan, and E. Vee, "Estimating statistical aggregates on probabilistic data streams," in *PODS 2007*, pp. 243–252.
[14] S. Bhattacharyya, A. Madeira, S. Muthukrishnan, and T. Ye, "How to scalably and accurately skip past streams," in *ICDE Workshops 2007*, pp. 654–663.
[15] R. Motwani and P. Raghavan, *Randomized Algorithms*. Cambridge University Press, 1995.
[16] N. L. Johnson, S. Kotz, and N. Balakrishnan, *Discrete Multivariate Distributions*. John Wiley & Sons, Inc., 1996.
[17] F. Rusu and A. Dobra, "Pseudo-random number generation for sketch-based estimations," *ACM TODS*, vol. 32, no. 2, p. 11, 2007.
[18] F. Olken, *Random Sampling from Databases – Ph.D. Thesis*. UC Berkeley, 1993.