Packet Sampling for Flow Accounting: Challenges and Limitations

Tanja Zseby[1], Thomas Hirsch[1], and Benoit Claise[2]

[1] Fraunhofer Institute FOKUS, Kaiserin-Augusta-Allee 31, 10589 Berlin, Germany
{Tanja.Zseby,Thomas.Hirsch}@fokus.fraunhofer.de
[2] Cisco Systems, De Kleetlaan 6a b, 1831 Diegem, Belgium
bclaise@cisco.com

Abstract. We investigate the applicability of packet sampling techniques to flow-based accounting. First we show by theoretical considerations how the achievable accuracy depends on sampling techniques, parameters and traffic characteristics. Then we investigate empirically which accuracy is achieved with typical flow characteristics by experiments with real traffic traces from three different networks. In a third step we illustrate how to support sampling-based accounting by providing an accuracy statement together with the measured data. We show which information is required for this and how an accuracy assessment can be approximated from information available after the sampling process using information elements of the IP flow information export protocol (IPFIX).

Keywords: packet sampling, accounting, IPFIX.

1 Introduction

Sampling aims at the reduction of measurement costs by estimating the metric of interest from a subset of data. It is important that the extent of potential estimation errors can be evaluated, especially if measurement results map to monetary values as it is the case for accounting. The achievable accuracy usually depends on characteristics of the population, i.e., in our case the traffic in the network. Since network traffic is extremely dynamic providing an up-to-date accuracy assessment is not trivial. It must be derived from the limited information available after the sampling process. It has to be calculated per flow and updated continuously.

Basic packet selection methods are currently standardized in the IETF PSAMP group [6]. A flow sampling scheme for accounting is introduced in [1]. Sample and Hold [2], Shared-state Sampling (S3) [3], and the Runs bAsed Traffic Estimator (RATE) [4] propose packet sampling methods that bias the selection process towards large flows in order to reduce resource consumption for flow caching and flow record transfer. This makes sense for accounting because in typical flow distributions a few large flows contribute to the majority to the overall traffic volume (e.g. [1]). Nevertheless, all those approaches require the classification of packets into flows before or during the sampling process. In contrast to this we investigate the effects of packet sampling that is applied *before* flow classification, so that only selected

M. Claypool and S. Uhlig (Eds.): PAM 2008, LNCS 4979, pp. 61–71, 2008.

packets need to be classified, which significantly reduces workload on routers [5]. We compare the achievable accuracy for basic PSAMP schemes and a stratified method used in Cisco NetFlow to accounting requirements. We show how the accuracy can be approximated from available information, using IPFIX information elements [11].

2 Flow Accounting Requirements

The accuracy of an estimate is assessed by ***bias*** and ***precision***. For accounting we should only use unbiased estimates. This is the case if the expectation of the estimated values equals the real value. The precision is derived from the variance (or its square root: the ***standard error***) of the estimate and expresses how far estimated values from sample runs would spread. The higher the standard error the lower is the precision. An accuracy statement can be presented to customers by a ***confidence interval***. Confidence boundaries define the area in which the real value should lie and can be expressed by the maximum tolerable estimation error. The confidence level (CL) gives the probability that the real value lies within this range. From this we can derive a maximum standard error that should not be exceeded if a given accuracy is required. Table 1 shows the maximum relative standard error for different accuracy requirements for a normal distributed estimate.

Table 1. Maximum Relative Standard Error for Different Accuracy Requirements

Rel. Est. Error	CL	Rel. StdErr
0.01 (1%)	99%	0.003876
0.01 (1%)	95%	0.005102
0.05 (5%)	99%	0.019380
0.05 (5%)	95%	0.025510

Rel. Est. Error	CL	Rel. StdErr
0.1 (10%)	95%	0.051020
0.15 (15%)	95%	0.076531
0.20 (20%)	95%	0.102041
0.30 (30%)	95%	0.1531

3 Accuracy Assessment in Theory

We here provide a theoretical assessment of bias and precision by providing formulas for expectation and standard error for the sampling schemes. We also give formulas for sampling after classification, but our focus is on sampling before classification. It is the more complex case, saves classification effort and is used in NetFlow.

Accuracy Assessment for n-out-of-N Sampling. In *n-out-of-N* sampling exactly n elements are selected from the population, which consists of N elements [6]. If there is only one flow ($N=N_f$) in the traffic mix or we apply sampling after classification, the number N_f of packets per flow is known. The number n_f of selected packets can be set per flow and is also known.

The estimate $\hat{Sum}_f$ for the number of bytes in flow f can be simply calculated from the packet sizes $x_{i,f}$ of the selected packets, by extrapolating with n_f and N_f. The expected bias is zero. The standard error can be calculated by the standard formula for an n-out-of-N selection [9] from sampling parameters and packet size variance $\sigma^2_{x_f}$.

$$\hat{Sum}_f = \frac{N_f}{n_f} \cdot \sum_{i=1}^{n_f} x_{i,f} \tag{1}$$

$$StdErr_{abs}\left[\hat{Sum}_f\right] = N_f \cdot \frac{\sigma_{x_f}}{\sqrt{n_f}} \cdot \sqrt{\frac{N_f - n_f}{N_f - 1}} \tag{2}$$

If we apply sampling before classification N_f and n_f are unknown. Extrapolation must be done with the overall population N and sample size n.

$$\hat{Sum}_f = \frac{N}{n} \cdot \sum_{i=1}^{n_f} x_{i,f} \tag{3}$$

In contrast to the case above (where n_f=n=*const*), here the number n_f of packets from flow f in the sample varies for each sampling run and has to be considered as random variable (r.v.) itself. The estimate contains two random variables, n_f and $x_{i,f}$. To assess the estimation quality we need to calculate expectation and variance of a sum of random variables, where the number of addends itself is a random variable. We model n_f as a discrete r.v. with a binomial distribution[1] $B(n, N_f/N)$. We denote the mean packet size of all packet sizes in flow f in the population by μ_{x_f} and their variance by $\sigma^2_{x_f}$. With $x_{i,f}$ we denote the number of bytes of the i^{th} selected packet[2]. Since we apply a random selection, the $x_{i,f}$ are independent identical distributed (i.i.d.).

With the assumption of the binomial distribution for n_f and independency for the $x_{i,f}$ we can derive the following formulas for expectation and variance for the estimated sum for flow f (see appendix):

$$E\left[\hat{Sum}_f\right] = N_f \cdot \mu_{x_f} = Sum_f \tag{4}$$

$$V\left[\hat{Sum}_f\right] = \frac{1}{n} \cdot \left(N \cdot N_f \cdot \left(\sigma^2_{x_f} + \mu^2_{x_f}\right) - N_f^2 \cdot \mu^2_{x_f}\right) \tag{5}$$

The expectation equals the real volume, i.e. the estimation is unbiased. The variance of the estimated flow volume, and with this the expected accuracy of the estimation depends on the parameters n, N, N_f, μ_{x_f} and $\sigma^2_{x_f}$. Sample size n and population size N are preconfigured sampling parameters. N_f, μ_{x_f} and $\sigma^2_{x_f}$ are flow characteristics. N_f denotes the number of packets in the population that belong to flow f. The packet size mean μ_{x_f} and the packet size variance $\sigma^2_{x_f}$ depend on the packet size distribution in flow f. If we take the square root of the variance we get the absolute standard error.

$$StdErr_{abs}\left[\hat{Sum}_f\right] = \sqrt{\frac{1}{n} \cdot \left(N \cdot N_f \cdot \left(\sigma^2_{x_f} + \mu^2_{x_f}\right) - N_f^2 \cdot \mu^2_{x_f}\right)} \tag{6}$$

A division by the flow volume provides the relative standard error (see appendix).

---

[1] If $f \leq 0.05$ and $0.1 < N_f/N < 0.9$ the hyper geometrical distribution $Hy(N, N_f, n)$ can be approximated by a binomial distribution $B(n, N_f/N)$ (see e.g., [10]).

[2] Note that the index i is used for the selected packets only and not for all packets in the flow.

Accuracy Assessment for 1-in-K Sampling (stratified). Cisco NetFlow implements a sampling scheme that we call 1-in-K sampling[3]. 1-in-K sampling is a count-based stratified n-out-of-N sampling. The selection process is done in two steps. First the measurement interval is divided into L subintervals of size K. Then one packet is randomly selected per subinterval. The measurement interval, i.e., the population for which a parameter should be estimated, still consists of N packets. The estimate is calculated from all n_f packets that were selected in all subintervals in the measurement interval.

$$\hat{Sum}_f = \frac{N}{n} \cdot \sum_{i=1}^{n_f} x_{i,f} \quad \text{with} \quad n_f = k_{f,1} + k_{f,2} + \ldots + k_{f,L} \tag{7}$$

The difference to n-out-of-N sampling is that here the number n_f of packets from flow f in the sample does not necessarily follow a binomial distribution. The sample size k within the subinterval is always 1. The number k_f of packets from flow f within this sample can be 0 or 1. The probability that k_f is 1 (i.e., the selected packet belongs to flow f) depends on the total amount of packets from flow f in the subinterval K_f. Therefore k_f can be considered as a Bernoulli distributed random variable with a probability of success $p_f=K_f/K$. So the distribution of n_f depends on those subinterval probabilities, which depend on the packets per flow in the subinterval.

If all packets in the measurement interval belong to one flow ($N_f=N$), the standard error for stratified sampling can be calculated as follows [see [9], following equation 5.9]:

$$StdErr[\hat{Sum}]_{strat} = \sqrt{\sum_{l=1}^{L} K_l \cdot (K_l - k_l) \cdot \frac{\sigma_{x,l}^2}{k_l}} \tag{8}$$

In the 1-in-K sampling implemented in NetFlow all strata have the same size ($K_l=N/L$) and only one packet is selected per stratum ($k_l=1$). Furthermore, if $K_l>>k_l$ we can approximate $K_l-k_l \approx K_l$. With this we get

$$StdErr[\hat{Sum}]_{strat} = \sqrt{\sum_{l=1}^{L} K_l^2 \cdot \frac{\sigma_{x,l}^2}{k_l}} = \sqrt{\frac{N^2}{L^2} \cdot \sum_{l=1}^{L} \sigma_{x,l}^2} = N \cdot \sqrt{\frac{1}{L^2} \cdot \sum_{l=1}^{L} \sigma_{x,l}^2} \tag{9}$$

The accuracy depends on the number L of strata and on the packet size variances $\sigma_{x,l}^2$ in the subintervals.

If the packets in the measurement interval belong to different flows ($N_f<N$), one has to consider not only the distribution of packet sizes over the subintervals but also the distribution of flow IDs. The calculation of the standard error becomes more complex because the variances have to be calculated per strata. The standard error now depends on the per-flow characteristics (number of packets K_f, packet size variance $\sigma_{x_f,l}^2$, and mean $\mu_{x_f,l}$) within each subinterval.

$$StdErr[\hat{Sum}_f]_{strat} = \sqrt{\sum_{l=1}^{L} \left(K_{f,l} \cdot K \cdot \left(\sigma_{x_f,l}^2 + \mu_{x_f,l}^2 \right) - \mu_{x_f,l}^2 \cdot K_{f,l}^2 \right)} \tag{10}$$

[3] To avoid confusion with the interval length N we call the scheme 1-in-K instead of 1-in-N.

The vigilant reader may miss the sampling parameters n and N in the formula. But for 1-in-K sampling the population size N is formed by the stratum size K and the number of strata L ($N=K*L$). The sample size n equals the number of strata L.

Theoretical Comparison of Schemes. A scheme provides a higher estimation accuracy if the standard error is smaller. That means 1-in-K sampling performs better if the following condition holds:

$$StdErr[\hat{Sum}]_{strat} < StdErr[\hat{Sum}]_{rand} \tag{11}$$

If we consider only one flow a stratification gain can be achieved if:

$$N \cdot \sqrt{\frac{1}{L^2} \cdot \sum_{l=1}^{L} \sigma_{x,l}^2} < N \cdot \sqrt{\frac{\sigma_x^2}{n}} \tag{12}$$

Since $n=L$, this can be simplified to.

$$\frac{1}{L} \cdot \sum_{l=1}^{L} \sigma_{x,l}^2 < \sigma_x^2 \tag{13}$$

That means we get a higher accuracy with 1-in-K sampling if the mean of the variances per subinterval (over all subintervals) is smaller than the variance within the whole measurement interval.

For multiple flows the formula gets more complex, because per-flow characteristics need to be taken into account. With the formulas for the standard error for n-out-of-N and stratified sampling for case II we get:

$$\sqrt{L\sum_{l=1}^{L}\left(K_{f,l}K\left(\sigma_{x_f,l}^2+\mu_{x_f,l}^2\right)-\mu_{x_f,l}^2K_{f,l}^2\right)} < \sqrt{\frac{1}{n}\left(NN_f\left(\sigma_{x_f}^2+\mu_{x_f}^2\right)-N_f^2\mu_{x_f}^2\right)} \tag{14}$$

In order to assess the accuracy for 1-in-K sampling one would need information about per flow characteristics for each subinterval. In contrast to n-out-of-N sampling those parameters cannot be approximated for 1-in-K sampling.

4 Accuracy Assessment in Practice

As we have seen we need the flow characteristics to calculate the accuracy. Since those are unknown, they have to be estimated from sampled values. A second problem is the amount of data that needs to be stored to provide an accuracy statement. Storing per-packet information results in too much data even if only sampled packets are stored. Therefore we here show how to calculate the accuracy from aggregated information. In addition we show how IPFIX Information Elements (IEs) can be utilized to export the required values needed for the accuracy assessment.

Accuracy Assessment from Sampled Packets. With the sampling parameters, the number of the sampled packets and their packet sizes we can provide estimates for the relevant parameters for n-out-of-N sampling.

$$\hat{N}_f = \frac{N}{n} \cdot n_f \quad (15) \quad \hat{\mu}_{x_f} = \overline{x}_f = \frac{1}{n_f} \cdot \sum_{i=1}^{n_f} x_{i,f} \quad (16) \quad \hat{\sigma}^2_{x_f} = s^2_{x_f} = \frac{1}{n_f - 1} \cdot \sum_{i=1}^{n_f} (x_{i,f} - \overline{x}_f)^2 \quad (17)$$

Using those estimates in formula (5) results in the following equation:

$$\hat{V}[\hat{Sum}_f] = \frac{N^2}{n} \cdot \left(\frac{n_f}{n} \cdot \left(s^2_{x_f} + \overline{x}^2_{x_f} \right) - \overline{x}^2_{x_f} \cdot \frac{n_f^2}{n^2} \right) \tag{18}$$

For 1-in-K sampling the assessment from sampled values is problematic. As can be seen from the formulas in section 3 we would need to estimate $K_{f,l}$, $\mu_{f,l}$ and $\sigma^2_{f,l}$ per subinterval. Since we select only one packet per subinterval, it is not possible to calculate acceptable estimates for mean and variance. As a consequence we cannot provide a practical accuracy statement from the sampled values for 1-in-K sampling. In empirical investigations we have seen that for many flows the accuracy for 1-in-K is close to the n-out-of-N model with current packet size distributions. Therefore the n-out-of-N accuracy often provides a good approximation.

Accuracy Assessment from Aggregated Information and IPFIX. Cisco currently stores for each flow the number n_f of packets in the sample and the sum of packet sizes from the sampled packets. With these two values and the sampling parameters n and N, one can easily calculate the estimates $\hat{N}_f$ and $\overline{x}_f$ ((15),(16)). But the calculation of the estimated variance $s^2_{x_f}$ is not possible with the stored values. A calculation of $s^2_{x_f}$ using (17) would require knowledge about all packet sizes in the sample. In order to avoid the storage of all packet sizes from the sampled packets, one can use an alternative variance calculation based on the sum and the square sum of the selected packet sizes.

$$s^2_{x_f} = \frac{1}{n_f - 1} \cdot \sum_{i=1}^{n_f} x^2_{i,f} - \frac{1}{n_f \cdot (n_f - 1)} \cdot \left(\sum_{i=1}^{n_f} x_{i,f} \right)^2 \tag{19}$$

Sum and square sum can be updated when a packet is selected and the packet sizes themselves do not need to be stored. If we insert (19) into formula (18) one can easily derive the accuracy from the stored aggregated values (sum and square sum). We recommended the storage of the square sum to Cisco. It has been added as an information element to the flow information export protocol IPFIX [12], and therefore will be available in Cisco routers in future. Table 2 shows the IPFIX and PSAMP information elements ([11], [13]) that provide the required values for calculating an accuracy statement.

If sampling is applied those values are calculate from the sampled packets and can be used to derive the required estimates. For count-based measurement intervals the number of packets in the measurement interval is preconfigured and can be reported with the samplingPopulation IE. For time-based measurement intervals one can report the number by defining an IPFIX flow that comprises all packets on the link

Table 2. IPFIX/PSAMP Information Elements

Parameter	IPFIX/PSAMP IEs
Number *N* of packets in measurement interval	`samplingPopulation`
Number *n* of packets in sample	`samplingSize`
Number of packets from flow *f* in sample	`packetTotalCount`
Sum (bytes in sampled packets)	`octetTotalCount`
Square sum (bytes in sampled packets)	`octetTotalSumOfSquares`

and use the packetTotalCount information element for this flow. An alternative is to use link packets counters from SNMP.

5 Experiments

We investigate the achievable accuracy for different schemes, classification rules and interval lengths with real traffic traces from 3 different networks. We show how many flows conform to given accuracy requirements.

Traces. The first trace set is from a large European operator (denoted as OP). The second set we collected at CIRIL [17], a regional network provider that interconnects universities and research institutes with the French Research and Education Network RENATER. Measurements were taken on a 1 Gbit multimode Ethernet access link to the national research network. As a third source we used the 6 hour traces *NZIX07m06d12h* (NZIX1) and *NZIX07m06d06h* (NZIX2) from [14]. We performed experiments with two different classification schemes. S24D24 distinguishes flows with respect to source and destination network both with a 24 bit netmask. S24D00 distinguishes flows only with respect to the source network. If packets of the same flow are observed in different measurement intervals they are counted as separate flows. Table 3 shows the number of flows observed for different classification rules and interval lengths (in number of packets). We use a letter per setting as identifier.

Table 3. Trace Characteristics

Setting	Trace	Size	#packets	Classification	MI	#flows
A	OP1	15 GB	122,800,288	S24D00	10M	852,593
B	OP1	15 GB	122,800,288	S24D24	10M	5,354,933
C	OP2	92 GB	766,071,712	S24D00	10M	69,001
D	CIRIL	2 GB	34,324,092	S24D00	10M	3,588,520
E	NZIX1	2 GB	65672186	S24D00	10M	8,569
F	NZIX2	39 GB	770,842,909	S24D00	10M	4,093
G	NZIX1	2 GB	65672186	S24D00	1M	79,383
H	NZIX1	2 GB	65672186	S24D24	1M	53,7138

Fig. 1 (left) shows a summarized representation of all flows in the CIRIL trace (setting D). Each dot represents a flow. The dimensions are the three flow characteristics that are relevant for the estimation accuracy: number of packets, packet size mean and variance (represented by the standard deviation). With settings D the

trace contains 3,588,520 flows. The majority of flows are small. Only 4,624 flows consist of more than 200,000 packets (not shown in graph). The peak at the standard deviation of zero and small means is caused by flows with packets of equal sizes. Several flows consist of only one packet. Those also have a standard deviation of zero. For the other traces and settings we observed similar flow distributions. Especially the existence of a majority of small flows was observed for all traces.

Conformance to Accuracy Requirements. First we calculate the achievable accuracy using the observed real flow characteristics and formula (6). Table 4 shows how many flows in the traces conform to given accuracy requirements for a sampling fraction of *f=5%*. The accuracy is given by the threshold *t* for the standard error.

Table 4. Conformant Flows for n-out-of-N, *f=5%*

ID	Number of Conformant Flows for rel. StdErr ≤ t				
	t=0.003876	*t=0.005102*	*t=0.019380*	*t=0.025510*	*t=0.051020*
A	0	1	1330	3,316	25,746
B	0	0	8	38	659
C	2	5	30	66	310
D	300	578	12,475	19,984	56,904
E	0	0	63	98	425
F	7	21	276	437	1,414
G	0	0	64	72	421
H	0	0	0	0	311

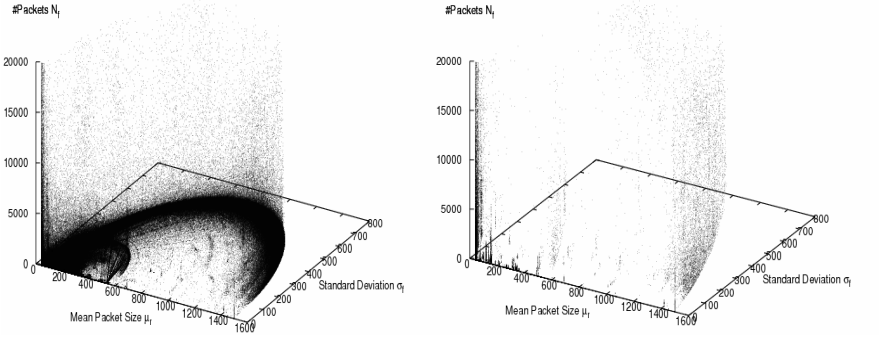

Fig. 1. Setting D: All Flows (left) and Conformant Flows (right)

Common accuracy requirements for accounting are a maximum relative error of *0.01* or *0.05* with a confidence level of at least *95%*. With a sampling fraction of *5%* the achievable accuracy is too low for the vast majority of flows for all settings. Fig. 1 (right) shows the flows conformant to *StdErr ≤0.05*. Only flows with a large number of packets N_f achieve an acceptable accuracy.

Flow Conformance from Empirical Tests. In a second step we investigate the standard error empirically from sampling tests. For this we performed *R=1,000* sampling runs for each scheme. Table 5 shows the results from experiments with setting G and different schemes.

Table 5. Conformant Flows for Setting G (*NZIX1, S24D00, f=5%*)

Max rel. StdErr	Error/CL	n-of-N	1-in-K	Systematic
0.003876	0.01/99%	0	0	0
0.005102	0.01/95%	0	0	0
0.019380	0.05/99%	64	64	62
0.025510	0.05/95%	72	72	83
0.051020	0.1/95%	473	475	567
0.076531	0.15/95%	1406	1425	1580
0.102041	0.2/95%	2316	2568	2860
0.1531	0.3/95%	5146	5397	5799
>0.1531	-	79383	79383	79383

The numbers for n-out-of-N sampling correspond quite well to those derived from the formula Table 4. For 1-in-K sampling we get quite similar numbers. This is in line with previous tests we performed about the scheme differences. Systematic sampling performs a little bit better, but the standard errors in the tests differed much from those of n-out-of-N. A theoretical prediction is problematic. Again, only few flows get accuracies sufficient for accounting. In order to achieve higher accuracies per flow one can increase the sample fraction, work with more coarse grained classifications or modify the measurement interval length. When modifying the measurement interval length it is relevant how flow characteristics evolve in order to assess the accuracy (see section 3).

6 Conclusion

We investigated the applicability of packet sampling to flow accounting. We analyzed basic PSAMP schemes and a stratified scheme used in Cisco NetFlow and showed how the accuracy depends on flow parameters and measurement settings. Theoretical considerations were supplemented by experiments with traffic traces from three different networks. The accuracy for sampling before classification was very poor. The main reason is the high number of small flows in the traces. Longer observation periods, coarse grained classification or the aggregation of flows results in larger flows and higher accuracies. A further option is to use a biased flow selection based on the expected accuracy. In addition we showed how the accuracy can be derived from sampled values and aggregated information stored in routers during run-time. For this, Cisco has included the storage of the square sum of the packet sizes in NetFlow.

References

[1] Duffield, N., Lund, C., Thorup, M.: Charging from Sampled Network Usage. In: ACM Internet Measurement Workshop IMW 2001, San Francisco, USA, November 1-2 (2001)

[2] Estan, C., Varghese, G.: New Directions in Traffic Measurement and Accounting: Focusing on the Elephants, Ignoring the Mice. ACM Transactions on Computer Systems (August 2003)

[3] Raspall, F., Sallent, S., Yufera, J.: Shared-state sampling. In: Proceedings of the 6th Internet Measurement Conference (IMC 2006), Rio de Janeiro, Brazil (2006)
[4] Kodialam, M., Lakshman, T.V., Mohanty, S.: Runs bAsed Traffic Estimator (RATE): A Simple, Memory Efficient Scheme for Per-Flow Rate Estimation. In: IEEE INFOCOM 2004, Hong Kong (2004)
[5] NetFlow Performance Analysis, Cisco white paper (2005), http://www.cisco.com/en/US/products/ps6601/products_white_paper0900aecd802a0eb9.shtml
[6] Zseby, T., Molina, M., Duffield, N., Niccolini, S., Raspall, F.: Sampling and Filtering Techniques for IP Packet Selection. Internet Draft <draft-ietf-psamp-sample-tech-10.txt> (work in progress, June 2007)
[7] Quittek, J., Zseby, T., Claise, B., Zander, S.: Requirements for IP Flow Information Export (IPFIX). In: RFC 3917 (October 2004)
[8] Zseby, T.: Stratification Strategies for Sampling-based Non-intrusive Measurements of One-way Delay. In: Proceedings of Passive and Active Measurement Workshop (PAM 2003) April 6-8 (2003)
[9] Cochran, W.G.: Stichprobenverfahren. Walter de Gruyter &Co, Berlin, New York (1972)
[10] Schwarz, H.: Stichprobenverfahren. Oldenbourg Verlag, GmbH (1975)
[11] Quittek, J., Bryant, S., Claise, B., Aitken, P., Meyer, J.: Information Model for IP Flow Information Export. In: RFC 5102 (January 2008)
[12] Claise, B. (ed.): Specification of the IP Flow Information Export (IPFIX) Protocol for the Exchange of IP Traffic Flow Information. In: RFC 5101 (January 2008)
[13] Dietz, T., Dressler, F., Carle, G., Claise, B., Aitken, P.: Information Model for Packet Sampling Exports, Internet-Draft draft-ietf-psamp-info-07.txt (work in progress, October 2007)
[14] Waikato Internet Traffic Storage (WITS), http://wand.cs.waikato.ac.nz/wand/wits/
[15] Fisz, M.: Probability Theory and Mathematical Statistics, 3rd edn. Robert E. Krieger Publishing Company Inc, Malabar, Florida (1963)
[16] Wentzel, E.S., Owtscharow, L.A.: Aufgabensammlung zur Wahrscheinlichkeitsrechnung. Akademieverlag, Berlin (1975)
[17] Centre Interuniversitaire de Ressources Informatiques de Lorraine (CIRIL), http://www.ciril.fr/

Appendix: Expectation and Variance for n-out-of-N Sampling

The random variable $x_{i,f}$ denotes the packet size of the i^{th} selected packet from flow f. Since a random selection is applied, we can assume that the $x_{i,f}$ are statistically independent. Since n_f follows a binomial distribution, the expectation and variance of n_f is given by formulas for a binomial distribution:

$$E\left[n_f\right] = n \cdot \frac{N_f}{N} \qquad (20) \qquad\qquad V\left[n_f\right] = n \cdot \frac{N_f}{N} \cdot \left(1 - \frac{N_f}{N}\right) \qquad (21)$$

With these considerations, the task is reduced to the calculation of expectation and variance of a r.v. Z, where Z is the sum of independent identical distributed (i.i.d.) random variables X and the number of summands Y is a binomial distributed random variable. The expectation of such a r.v. is given in [15].

$$E[Z] = E[X] \cdot E[Y] \quad \text{for} \quad Z = \sum_{i=1}^{Y} X_i \qquad (22)$$

With this the expectation of the estimated volume is calculated as follows:

$$E[Z] = \frac{N}{n} \cdot E\left[\sum_{i=1}^{n_f} x_{i,f}\right] = \frac{N}{n} \cdot E[x_{i,f}] \cdot E[n_f] = \frac{N}{n} \cdot \mu_{x_f} \cdot n \cdot \frac{N_f}{N} = N_f \cdot \mu_{x_f} = Sum_f \quad (23)$$

The expectation of the estimate equals the real volume, i.e. the estimation is unbiased. A formula to calculate the variance for this special case, but for continuous random variables is derived in [16]. This formula can be also applied for discrete variables.

$$V[Z] = E[Y] \cdot V[X] + E[X]^2 \cdot V[Y] \text{ for } Z = \sum_{i=1}^{Y} X_i \quad (24)$$

With this the variance of the estimated flow volume can be expressed as follows:

$$V[\hat{Sum}_f] = \frac{N^2}{n^2} \cdot V\left[\sum_{i=1}^{n_f} x_{i,f}\right] = \frac{N^2}{n^2} \cdot \left(E[n_f] \cdot V[x_{i,f}] + E[x_{i,f}]^2 \cdot V[n_f]\right)$$

$$= \frac{N^2}{n^2} \cdot \left(E[n_f] \cdot V[x_{i,f}] + E[x_{i,f}]^2 \cdot V[n_f]\right) \quad (25)$$

The relative standard error can be easily derived from the variance.

$$StdErr_{rel}[\hat{Sum}_f] = \frac{StdErr_{abs}[\hat{Sum}_f]}{Sum_f} = \frac{\sqrt{\frac{1}{n} \cdot \left(N \cdot N_f \cdot \left(\sigma_{x_f}^2 + \mu_{x_f}^2\right) - N_f^2 \cdot \mu_{x_f}^2\right)}}{N_f \cdot \mu_{x_f}} \quad (26)$$