网络测量中自适应数据采集方法*

王俊峰[1,2+], 杨建华[3], 周虹霞[4], 谢高岗[3], 周明天[1]

[1](电子科技大学 计算机科学与工程学院,四川 成都 610054)

[2](中国科学院 软件研究所,北京 100080)

[3](中国科学院 计算技术研究所 信息网络研究室,北京 100080)

[4](电子科技大学 电子工程学院,四川 成都 610054)

Adaptive Sampling Methodology in Network Measurements

WANG Jun-Feng[1,2+], YANG Jian-Hua[3], ZHOU Hong-Xia[4], XIE Gao-Gang[3], ZHOU Ming-Tian[1]

[1](College of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 610054, China)

[2](Institute of Software, The Chinese Academy of Sciences, Beijing 100080, China)

[3](Network Research Division, Institute of Computing Technology, The Chinese Academy of Sciences, Beijing 100080, China)

[4](College of Electronic Engineering, University of Electronic Science and Technology of China, Chengdu 610054, China)

+ Corresponding author: Phn: +86-28-83203300, E-mail: mailwangjf@yahoo.com.cn

Received 2003-04-21; Accepted 2003-06-20

Wang JF, Yang JH, Zhou HX, Xie GG, Zhou MT. Adaptive sampling methodology in network measurements. *Journal of Software*, 2004,15(8):1227~1236.

http://www.jos.org.cn/1000-9825/15/1227.htm

Abstract: Sampling methodologies are widely used in network measurements and other related fields. Most applications mainly focus on parent population statistical metrics estimation of interest. Recent researches reveal that many aspects of network characters present heavy-tailed distribution or self-similarity. These properties might cause a heavy passive effect on the estimation accuracy. In other circumstances, there exist demands on modeling the characteristics of a network in network operation. To develop an accurate model for network character is much

* Supported by the National High-Tech Research and Development Plan of China under Grant No.2002AA121032 (国家高技术研究发展计划(863)); the Institute of Computing Technology Youth Fund under Grant No.20026180-14 (计算技术研究所青年基金)

WANG Jun-Feng was born in 1976. He received the Ph.D degree (with honor) in Computer Application Technology from the College of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu, Sichuan. From July 2004, he holds a postdoctoral position in Institute of Software, the Chinese Academy of Sciences. His recent research interests include Internet measurements and traffic analysis, protocol modeling and performance evaluation. **YANG Jian-Hua** was born in 1978. She is a Ph.D. candidate at the Institute of Computing Technology, the Chinese Academy of Sciences. Her research areas include high speed network measurement and monitoring. **ZHOU Hong-Xia** was born in 1976. She is a postgraduate student at the College of Electronic Engineering, University of Electronic Science and Technology of China. Her current research is focused on digital signal processing. **XIE Gao-Gang** was born in 1974. He is an associate professor at the Institute of Computing Technology, the Chinese Academy of Sciences. His research is focused on high speed network technology, network measurement and monitoring, QoS. **ZHOU Ming-Tian** was born in 1939. He is a professor and doctoral supervisor at the College of Computer Science and Engineering, University of Electronic Science and Technology of China. His current research interests include computer networking information system, open distributed processing system, computer system and software, and computer supported collaboration work.

difficult. From a broader view, these applications are treated as special cases of fitting problems of planar data set or time series in applied mathematics. In the paper, a Fitting-based adaptive sampling methodology (FASM) is developed for reconstructing the evolution of some network characteristics (model). The contributions of the paper include: (1) Adopting a Piecewise Linear Function Approximation scheme to provide a more accurate approximation of the true character. (2) The statistical metric derived from the FASM provides a much more stable and accurate estimation than other popular methodologies under the same sampling size. Experiments based on two measurement traces show that the FASM can dramatically reduce the number of samples while retaining the same approximating residual error as others. (3) The variance of sampling size is more stable than those of other probability sampling schemes.

Key words: adaptive sampling; piecewise linear fitting; network measurement

摘 要: 抽样方法广泛地应用于网络测量与其他领域对被测总体的指标进行估计.研究表明,多种网络指标呈现重尾分布或自相似的特征.这些特性为准确估计总体指标带来了诸多困难.但同时,对被测网络指标进行建模也有着重要的应用.然而,建立精确网络模型是困难的.从时间序列拟合角度出发,提出了一种基于拟合的自适应抽样方法,对被测指标进行基于测量的建模.工作主要体现在: (1) 采用分段线性函数对被测指标进行逼近,建立基于测量的模型; (2) 与常用的抽样方法相比,在相同的样本数情况下,由拟合模型对指标进行的估计更准确、更稳定;通过对两个测量记录的分析表明,在与常用抽样方法保持相同的拟合误差时,自适应抽样方法明显地减少了所需采集的样本数量; (3) 与其他概率抽样方法相比,自适应抽样最终抽取的样本数更稳定、更可靠,并给出了最终样本数的概率分布.

关键词: 自适应抽样;分段线性拟合;网络测量

中图法分类号: TP393 **文献标识码**: A

Network measurements are prerequisite to a variety of application fields such as network fault diagnosis, performance evaluation, service level agreements (SLAs) validation and traffic engineering. Sampling provides an effective approach for operators to get a better understanding of the operating networks without causing heavy burden on the network bandwidth, CPU and storage resources. Many sampling methods and their applications concern on mining statistical metrics of interest from networks. Zseby deployed sampling methods to measure the one-way delay metric for SLA validation through estimating the proportion of packets that belong to a specific flow[1]. Nick *et al*. employed size-dependent sampling methodology to obtain the customer usage of network for accounting at a given accuracy[2]. Essentially, these methods and their applications provide an overview for some aspects of the network characteristics. In many situations, there exist demands on obtaining the evolution of some particular characters of network. Taking measurement-based traffic modeling for example, we are sometimes interested in not only the volumes of traffic observed at measurement point in a given time duration, but also a detailed description of the traffic dynamics. An accurate reflection of its variation in measuring is more crucial for establishing the empirical traffic model. From a broader point of view, many network measurement activities can be considered as the applications of some kind of sampling methods to reconstruct the evolution of characteristic of interest and then to infer the particular statistical metrics. Though those proposed sampling methods are useful to these applications, they are not effective, as illustrated later.

In this paper, we propose a Fitting-based adaptive sampling methodology (FASM) to deal with this characteristic reconstruction (measurement-based modeling) problem. Experiments show that the FASM could provide a more accurate approximation than other popular sampling methods. Under the same approximating residual error, FASM could reduce the number of samples dramatically while providing reasonable statistical metric estimation, thus the measurement costs are decreased considerably. In addition, the sample size of FASM is more deterministic than those of other probability based sampling methods.

The remainder of the paper is organized as follows. Section 1 formalizes the piecewise linear fitting (PLF) in the context of network measurement. Section 2 presents the FASM methodology. In Section 3, two datasets are used to evaluate the performance and compare with other methodologies. Section 4 concerns parameter estimation of the character. In Section 5, we focus on the generated sample size variation. Section 6 discuses the key issues in FASM application. Section 7 concludes the paper.

1 Formalization of Piecewise Fitting in Network Measurement

Constructing a model for a data set in plane field is a common problem that is always encountered in data warehouse mining, pattern recognition and applied mathematics. The main purpose of this sort is to develop the model for a given observed data set. Let $Y=f(X)+\varepsilon$ be a continuous process. X is an independent variable on which the metric Y to be modeled, ε denotes the minor random error with mean zero that might be introduced in observation. Suppose $S=\{(x_1,y_1),(x_2,y_2),...,(x_n,y_n)\}, x_1<x_2...<x_n$ be a sequence of observations of Y, and $s_i=(x_i,y_i)$ represent the i-th element of S. S is referred to as adjoint sequence of Y. As we do not know the true distribution of Y, S can be deemed equivalent to Y when $N=|S|$ is large enough.

Let $\mathbb{F}$ be a class of k-link piecewise linear functions (PLFs), and $\mathrm{E}(F_i), F_i\in\mathbb{F}$ the fitting error of F_i. To optimally approximate a data set S by a k-link piecewise linear function F is to find F_i that satisfy (1):

$$F=\{F_i \mid \min\{\mathrm{E}(F_i), F_i\in\mathbb{F}\}\} \tag{1}$$

$$\mathrm{E}(F_i)=\max_{i=1,2,...,n} |y_i-F(x_i)| \tag{2}$$

Goodrich defined the fitting error for F_i as Eq.(2). A comprehensive study of such work was conducted in Ref.[3] and its references. A plane sweep based computational geometry algorithm was developed to find the best k-link PLF. Pittman employed a genetic algorithm (GA) to optimize the number and location of PLF for a given data set[4].

Once the optimal k-link PLF F is obtained, we get a subset $S', S'\subseteq S$ with k+1 elements, which locates the start and stop points for each segment of F. However, these approximation algorithms are not very suitable for network measurements for the following reasons: (1) we could not get the full data S in advance for approximating evaluation during the process of measurement in that our main goal is to model the network characters accurately with least samples. (2) As S is not known to us, we could not get the optimal PLF F in approximating. Sampling provides potential schemes for this type of fitting purposes because two consecutive samples could be treated as consecutive elements in S'. In network measurement, those widely used sampling methods include systematic sampling, random sampling and stratified sampling. In Ref.[1], Zseby presented a detailed performance analysis on these sampling methods. With a-priori information about the serial correlation, Zseby pointed out that stratified sampling could get a more accurate estimation of the parent population. However, in a practical online measurement, we cannot subgroup the parent population into subsets according to a given characteristic. Therefore, we cannot get the serial correlation to guide us in sampling process directly.

To facilitate the analysis of the adaptive sampling methodology in next section, we formalize this approximating problem in network measurement context. Let $Y=f(X)$ be the true model of interest, X represent the time elapse of measurement procedure t, S' denote a collection of samples by any sampling or selection schemes (each element is an observation of Y e.g. (x_i,y_i)). PLF F' is constructed from S'. We define the approximating residual error E' as:

$$E'(t)=\int_0^t |f(x)-F'(x)|\,dx \tag{3}$$

Therefore, the mission of network character approximation is to find the PLF F' which minimizes $E'(t)$ with a given sample size n. Equation (3) provides the criterion for F' evaluation. Though the true $f(x)$ is still unknown, we use it to evaluate the fitting performance of the FASM algorithm with others in the paper.

2 Fitting-Based Adaptive Sampling Methodology

2.1 Nomenclature and definitions

Consider a time series T, which is divided into k consecutive segments S_T^k with segment error E_T^k. S_T^{k+1} is the k+1 consecutive segments based on S_T^k. We get Eq.(4) as:

$$E_T^{k+1}\le E_T^k \tag{4}$$

The more segments, the smaller segment error E_T^k. This property of time series segmentation is performed in Refs.[5,6]. Thus we can control the time at which a sample is drawn and the number of elements sampled to limit the residual error at a given level. The core of the FASM is that if the variance of Y in an interval Δt is acute, then the number of samples collected in this interval should be increased, e.g. the time interval between two consecutive samples should be decreased, conversely the number of samples should be decreased in a smoother period, which increases the time interval between two consecutive samples. Let Var_i^j demonstrates the absolute variance level (AVL) of Y when the i-th samples is obtained from Eq.(5):

$$Var_i^j=\frac{\sum_{l=i-j+1}^{i}|\overline{f_i^j}(x_l)-y_l|}{\frac{1}{j}(x_i-x_{i-j+1})\sum_{l=i-j+1}^{i}y_l},\quad 1\le j\le i \tag{5}$$

where $\overline{f_i^j}(x)$ represents the linear regression for fitting the $f(x)$ from x_{i-j+1} to x_i and j is the regression order. Let R_i^j measures the relative variance level (RVL) which is defined as:

$$R_i^j=\frac{Var_i^j}{Var_{i-1}^j},\ j\ge 1, i\ge j+1 \tag{6}$$

Assume τ_i to be the time interval between (i−1)-th and i-th samples, then the time τ_{i+1} for the (i+1)-th sample is determined by Eq.(7):

$$\tau_{i+1}=\frac{1}{\lambda_{i+1}}(D_{i+1}^1 EXP+D_{i+1}^2) \tag{7}$$

where
$$(D_{i+1}^1,D_{i+1}^2)=\begin{cases}(d_{11},d_{12}), & \text{if } R_i^j\ge C, C>1\\ (d_{21},d_{22}), & \text{if } R_i^j<1\\ (0,1), & \text{otherwise}\end{cases}$$

and variable EXP is subservient to exponential distribution with mean 1 and standard deviation 1. The constant C represents the threshold for adjust the sampling interval. $0<d_{kl}<1$ $k,l=1,2$ are parameters for sampling interval control. $1/\lambda_{i+1}$ is the interval expectation for the following n-i samples defined as Eq.(8), which is a minor variation of the sampling probability definition introduced in Ref.[7].

$$\lambda_{i+1}=\frac{n-i}{x_n-x_i},\ 1<i<n \tag{8}$$

Therefore from Eq.(7), the time interval τ_{i+1} is determined by two components: a stochastic part which is subservient to exponential distribution $E(\lambda_{i+1}/D_{i+1}^1)$ and a deterministic part D_{i+1}^2/λ_{i+1}. By carefully choosing the weight of the two parts, the quantities of D_{i+1}^1 and D_{i+1}^2, we may control the sampling process according to the evolution of Y.

2.2 Fitting-based adaptive sampling methodology (FASM) algorithm

During a measurement time duration T, the PLF is constructed by n samples including the start and end observation of Y, i.e. $(0, y_0)$ and (T, y_T). Equation (7) indicates that the time interval of consecutive samples τ_i is a stochastic variable as variable *EXP* introduced. Thus the ultimate sample size is also a variable depending on τ_i. Figure 1 illustrates the FSAM algorithm in pseudo code.

As the measuring duration T is finite, algorithm will stop in limited steps. For every sample $i, i > j$, step 10 requires calculating Var_i^j and Var_{i-1}^j, and each of them is of order $O(j)$. Commonly $j \ll n$, the computational complexity of the FASM is of order $O(Jn)$, where factor J depends on j, the number of calculations, and the ultimate sample size of S.

```
1.  PROC FASM: IN( n, j, T ), OUT( S )
2.      t_1 = 0; τ_1 = 0; S = {(t_1, y_{t_1})}; i = 2
3.      // select the first j + 1 samples
4.      WHILE( i ≤ j + 1 AND t_{i-1} < T )
5.          λ_i = (n - (i - 1)) / (T - t_{i-1}); τ_i = (1/λ_i) EXP; t_i += τ_i
6.          S = S ∪ {(t_i, y_{t_i})}; i++   // get a sample
7.      END // end of WHILE
8.
9.      WHILE( t_{i-1} < T )
10.         calculate R_i^j from Eq.(6)
11.         calculate t_i += τ_i from Eq.(7)
12.         IF( t_i < T )
13.             S = S ∪ {(t_i, y_{t_i})}; i++   // get a new sample
14.         ELSE
15.             BREAK
16.     END // end of WHILE
17.     S = S ∪ {(T, y_T)}   // get the last sample
18.     OUTPUT S
19. END PROC
```

Fig.1 Pseudo code for FASM algorithm

3 Algorithm Performance Evaluation

In this section, we compare the performance of FASM with other popular sampling schemes in network measurements. We use traces as those used in Refs.[1,2,8] to simulate the scenario of the true network model of interest for reference, i.e. let the traces delegate the parent population for sampling. The first one is a one-hour wide-area traffic load trace (simply packet counts in every second) that is derived from the passive traffic measurement experiment mentioned in Ref.[9]. Another dataset is a one-way delay trace generated by the active measurement from Ref.[5]. The parent population sizes of N are 3496 and 12000 respectively.

Zesby[1] and Claffy[8] summarized the sampling methods of systematic sampling, simple random sampling and stratified sampling, and assessed the performance by their particular standards. As the stratified sampling method requires a prior knowledge for dividing the parent population into subgroups, we exclude this method and include Poisson sampling that is recommended by IETF IPPM in Ref.[10]. In the paper, we implement packet-driven simple sampling and systematic sampling, timer-driven Poisson sampling and FASM sampling for the evaluation of their goodness-of-fit.

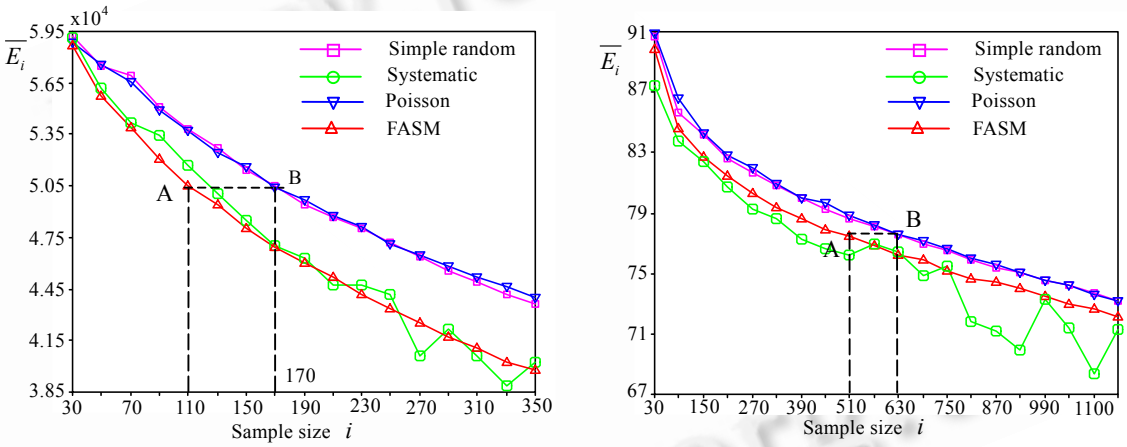
1000 independent sampling rounds are taken for each sampling scheme under the same expected sampling size n, which ranges from nearly 0.5% to 10% of the parent population N (30 to 350 for first trace and 30 to 1200 for the second trace). Let $(E_n^1, E_n^2, ..., E_n^{1000})$ denote the approximating residual error of n samples, E_n the real residual error, and $\overline{E_n}$ the estimation of E_n obtained from Eq.(9):

$$\overline{E_n} = \frac{1}{1000}\sum_{i=1}^{1000} E_n^i \tag{9}$$

Figure 2 illustrates the performance for each sampling scheme of the two traces in approximation. It shows that the Poisson sampling and simple random sampling have nearly the same performance in this scenario. The systematic sampling appears to perform as well as the FASM, but the stability of fitting residual error is subjected to the variance of sampling size. The FASM method is more robust than the systematic sampling. The residual error of the FASM is much smaller than those of the Poisson and simple random sampling, for example in the traffic load trace, the residual error of the FASM with sample size 110 equals to the Poisson or simple random sampling with sample size 170. Therefore, from approximation point of view, we could reduce the sample size by 35.3% in this situation and nearly 19% at points A and B in the one-way delay trace.

Table 1 Parameters selection for the two-trace approximation

Param	j	C	d_{11}	d_{12}	d_{21}	d_{22}
Value	20	1.1	0.8	0.2	0.16	0.64



(a) Traffic load trace (b) One-Way delay trace

Fig.2 Comparison of approximating residual error between different sampling methodologies

As systematic sampling is a biased sampling method that suffers from two potential problems described in Ref.[10], therefore, it is excluded from our following analysis. Now, we compare the sample efficiency of FASM and simple random sampling methods. Let F_F^i, F_{SR}^i denote the PLF with sample size i under FASM and simple random sampling. E_F^i and E_{SR}^i represent the fitting error respectively and n_F is defined as $E_F^{n_F} = E_{SR}^n$. We calculate the ratio of the saved samples $R_{F|SR}$ by Eq.(10):

$$R_{F|SR} = \frac{\sum_{i=l}^{u}(i - i_F)}{\sum_{i=l}^{u} i} \tag{10}$$

where l and u are the lower and upper bounds of the sample size respectively.

The $R_{F|SR}$ is 33% and 17.3% for the first and second trace respectively. Clearly, the FASM significantly

reduces the number of samples compared with the simple random sampling method.

4 Parameter Estimation

The most important application of sampling methodologies is to estimate the mean, variance or distribution of parent population. In this paper, based on the first trace, we concentrate on mean estimation by FASM sampling, comparing it with the unbiased simple random sampling to show that the change of sampling method from simple random sampling to FASM does not affect this statistic estimation significantly.

Obviously, compared with Poisson sampling or simple random sample method, FASM is a biased sampling method. It is unreasonable to estimate the statistic metric of parent population as others do. We define the expectation value of parent population under sample size n as:

$$\overline{Y_F^n} = \frac{1}{T}\int_0^T F^n(x)\mathrm{d}x \tag{11}$$

As in Section 3, we also perform 1000 rounds independent experiments on the first traffic load trace with the same sample size n, $30 \le n \le 350$. According to the central limit theorem, suppose that traffic load trace is subject to $N(\mu_F^n, \sigma^2)$ by FASM sampling and $N(\mu_{SR}^n, \sigma^2)$ by simple random sampling with the same sample size n. We take following hypothesis testing to determine the significant effect of different sampling methodologies under the significant level α:

$$H_0 : \mu_F^n = \mu_{SR}^n (H_1 : \mu_F^n \ne \mu_{SR}^n),$$

let $\overline{Y_F^n}$ and $\overline{Y_{SR}^n}$ be the estimation of μ_F^n, μ_{SR}^n, we construct:

$$T_n = \frac{\overline{Y_F^n} - \overline{Y_{SR}^n}}{S_w\sqrt{\frac{1}{m}+\frac{1}{n}}} \sim t(m+n-2) \tag{12}$$

where $\alpha = 0.01, n = m = 1000$, $t_{0.995}(1998)$ and $|t_n|$ are illustrated in Fig.3. The ratio of $|t_n| \le t_{0.995}(1998)$ is about 86.3%. This result reveals that in most cases, we cannot reject the hypothesis that the two sampling methods do not take significant difference in mean estimation under the same sample size.

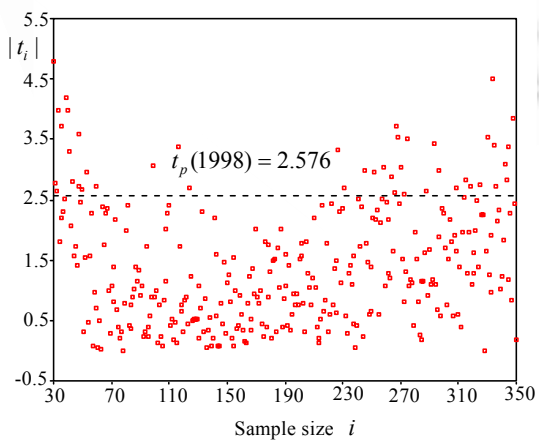


Fig.3 Distribution of $|t_i|$ under the same sample size between FASM and simple random sampling

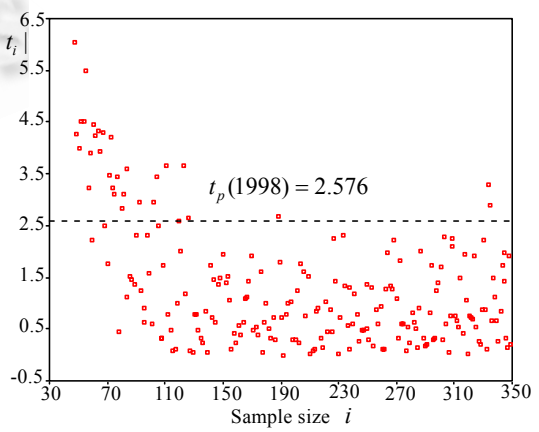


Fig.4 Distribution of $|t_i|$ under the same approximation accuracy between FASM and simple random sampling

Section 3 also mentions that FASM could significantly reduce the number of samples for a given approximation accuracy compared to simple random sampling and Poisson sampling. We test the validation of parent population estimation with different sample sizes in the situation of retaining the fitting accuracy. Take points A and B in Fig.2(a) for example, $|t|=1.18<t_{0.995}(1998)$. So we can accept the mean estimation by FASM compared to simple random sampling. Similarly, an equal expectation value testing hypothesis is taken under the same approximating error between the two sampling methods as depicted in Fig.4. It shows that 85.6% of $|t_n|$ is below 2.576. This reveals that the reduced samples generated by FASM could also provide a reasonable estimation of parent population as the simple random sampling does.

5 Sample Size Stability

Poisson sampling, sample random sampling and FASM are all probability samples. The resulting sample size $\underline{n}$ is a random variable with an expected value n. If the obtained sample size is far from the expectation, then the sample will be rejected because of the concern of precision or resource occupation. This section analyzes the FASM sample size stability and compares it with those of the Poisson sampling and simple random sampling.

Suppose $N_t, t\geq 0$ denote the realized sample size within time t by the sampling methodologies. Apparently, N_t is a counting process. As to the Poisson sampling and sample random sampling, Poisson process and Bernoulli process can be used to calculate the probability $P(N_t=k), k=0,1,...,n$.

The pdf (Probability density function) for i-th sample time interval $f(\tau_i)$ is derived from Eq.(13):

$$f(\tau_i)=\frac{\lambda_i}{D_i^1}e^{-\frac{\lambda_i\tau_i-D_i^2}{D_i^1}},\tau_i\geq\frac{D_i^2}{\lambda_i} \tag{13}$$

The probability of $P(N_T=n)$ can be inferred from Fig.5:

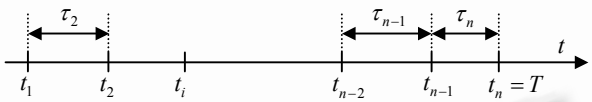


Fig.5 Sampling event description

$$\begin{aligned}P(N_T=n)&=P(\tau_n\geq T-t_{n-1},t_{n-1}<T)\\&=P(\tau_n\geq T-t_{n-1}\,|\,t_{n-1}<T)P(t_{n-1}<T)\\&=P(\tau_n\geq T-t_{n-1})P(\tau_{n-1}<T-t_{n-2},t_{n-2}<T)\\&=P(\tau_n\geq T-t_{n-1})P(\tau_1<T)\prod_{i=1}^{n=2}P(\tau_{n-i}<T-t_{n-i-1})\end{aligned} \tag{14}$$

Assume that for the i-th sample $1<i<n$, $(D_i^1,D_i^2)=(0,1)$, i.e. $\tau_i=1/\lambda_i$, then $P(\tau_i<T-t_{i-1})=1$. Such samples are referred to as fixed samples. Let $P'(N_T=n)$ represent the probability of $N_T=n$ while no fixed samples are included, therefore $P'\leq P$. Applying Eq.(13) to (14), P' is given by:

$$P'(N_T=n)=e^{-\frac{1-D_n^2}{D_n^1}}\prod_{i=2}^{n-1}(1-e^{-\frac{(n-i+1)-D_i^2}{D_i^1}}) \tag{15}$$

Eq.(15) indicates that the probability is independent of the measurement time duration T. Let $P'(n)$ denote the $P'(N_T=n),T>0$. Figure7 shows two extreme cases of $P'(n)$, e.g. $P_1'(n)$ and $P_2'(n)$ when $(D_i^1,D_i^2)=(d_{11},d_{12})$, $(D_i^1,D_i^2)=(d_{21},d_{22}),i=2,...,n$. $d_{kl},k,l=1,2$ are taken from Table 1. The probability $P'(n)$ becomes stable if the

expected sample size is larger than $n' = \max\{n_1, n_2\}$.

Based on the second trace, 1000 rounds of sampling are performed for each $n, 30 \le n \le 1200$. Figure 7 illustrates the proportion of the generated sample size that equals to the expected size n. FASM presents a much more stability on the sample size. From Eq.(15), if the systematic sampling is adopted for the last $n', n' \ll n$ samples, then $P'(n) \approx 1$.

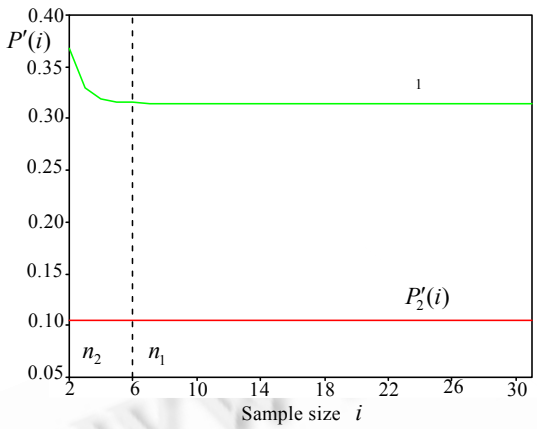


Fig.6 P' evolution

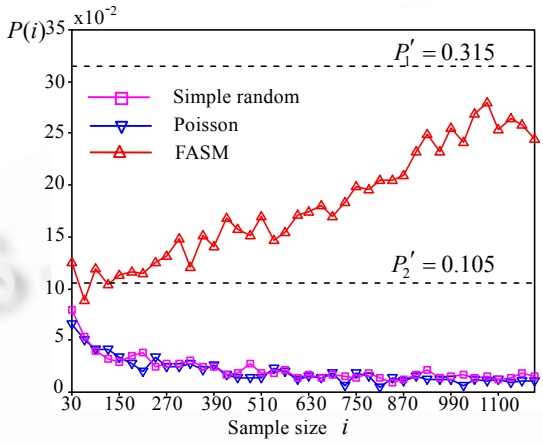


Fig.7 Comparisons of the proportion of the generated sample size that equals to expectation i

6 Discussions

Apparently, parameters $C, j, d_{11}, d_{12}, d_{21}, d_{22}$ will significantly affect the performance of the FASM, but some basic rules underlie the selection of them. Threshold C is commonly chosen between 1.0 and 2.0. The smaller the threshold, the higher the change frequency for the time interval of samples. The regression window size j reflects the sensitivity to the fluctuation of the measured characteristic. A shortened j makes the FASM take immediate actions in approximation, while a wider window size smoothens the true property, causing more information loss. In experiments, a window size from 5 to 50 is suggested and a smaller window has a higher priority. $d_{11} > d_{12}, d_{21} < d_{22}$ indicates that during a smoother period of character evolution, FASM generates a longer sampling time interval with higher probability of short sampling interval for an acute evolution period.

To choose better sampling parameters for the FASM, an effective way is to tune them by a training data set of the interested network character. An on-line self-learning mechanism can be adopted to adjust these parameters for a more accurate approximating and statistical metrics estimation.

7 Conclusions

In this paper, we propose a novel sampling mechanism, FASM, from a piecewise linear approximation point of view. The main advantage is that it could significantly reduce the number of samples while providing a reasonable evolution reflection, e.g. modeling of parent population. We apply this sampling method to network character extraction according to two real network traces. The experiments show that FASM could drastically reduce the number of samples by 33% and 17.3% compared to other popular sampling methods for each trace. Though this algorithm is developed based on the network character extraction, it provides a framework for other application fields of data mining.

Acknowledgement The authors would like to thank Vern Paxson at Lawrence Berkeley National Laboratory for making data sets available on the web.

References:

[1] Zseby T. Deployment of sampling methods for SLA validation with non-intrusive measurements. In: Proc. of the Passive and Active Measurements Workshop 2002 (PAM2002). Fort Collins, 2002.

[2] Nick D, Carsten L, Mikkel T. Charging from sampled network usage. In: Proc. of the Internet Measurement Workshop. San Diego: ACM Press, 2001.

[3] Goodrich MT. Efficient piecewise-linear function approximation using the uniform metric. In: Proc. of the 10th Annual Symp. on Computational Geometry. 1994. 322~331.

[4] Pittman J, Murthy CA. Fitting optimal piecewise linear function using genetic algorithms. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2000,7:701~718.

[5] Wang JF, Yang JH, Zhou HX, Xie GG, Zhou MT. Detecting clock dynamics in one-way delay measurement. Journal of Software, 2004,15(4):584~593 (in English with Chinese abstract). http://www.jos.org.cn/1000-9825/15/584.htm

[6] Vasko KT, Toivonen Hannu TT. Estimating the number of segments in time series data using permutation tests. In: Proc. of IEEE Int'l Conf. on Data Mining 2002 (ICDM 2002). Maebashi: IEEE Computer Science Press, 2002. 466~473.

[7] Vitter JS. An efficient algorithm for sequential random sampling. ACM Trans. on Mathematical Software, 1987,3:58~67.

[8] Claffy KC, Polyzos GC, Braun HW. Application of sampling methodologies to network traffic characterization. Technical Report CS93-275, UCSD.

[9] Paxson V, Floyd S. Wide-Area traffic: the failure of poisson modeling. IEEE/ACM Trans. on Networking, 1995,6:226~244.

[10] Paxson V, Almes G, Mathis M. Frameworks for IP performance metrics, RFC 2330, 1998.

附中文参考文献：

[5] 王俊峰,杨建华,周虹霞,谢高岗,周明天.单向延迟测量中时钟动态性检测算法.软件学报,2004,15(4):584~593. http://www.jos.org.cn/1000-9825/15/584.htm

****

第 2 届全国搜索引擎和网上信息挖掘学术研讨会

征文通知

为促进国内外相关领域科研人员的学术和工作交流，研讨本领域的最新技术进展和发展趋势，以推动搜索引擎和 Web 挖掘技术在中国的发展，2004 年“全国搜索引擎和网上信息挖掘学术研讨会”将于 2004 年 11 月 12 日~13 日在广州华南理工大学举行。经专家评审录用的会议论文，将在《华南理工大学学报》（自然科学版）增刊正式出版。会议还将举行“中文 Web 检索竞赛”，欢迎单位或个人组队参加。

欢迎高等院校教师、科研院所和企业的业界专家、科研人员、研究生以及业界专业人士踊跃投稿，国家和省部级资助项目支持的论文将优先录用。详情请浏览我们的网站：http://www.scut.edu.cn/sewm2004