抽样机制对报文长度分布测度影响的研究*

张巍，丁伟，龚俭

(CERNET华东(北)地区网络中心,东南大学计算机科学与工程学院,江苏南京,210096)

摘要：Claffy在1994年对1h的主干数据分析的基础上，提出了基于时间驱动技术比基于报文驱动技术得出的抽样结果要差一些，但是二者之间的差别很小。现在，这个结论是否依然适用呢？本文以CERNET某2.5G省网边界链路3个不同阶段连续1 h报文数据构成的Trace为对象，采用Claffy相似的方法，对报文长度分布测度进行了分析，获得了1组有关的结论，可以用于与网络测量和网络行为学有关的研究。

关键字：抽样；时间驱动；报文驱动；测度

网络测量的目的之一就是获取测度值，以此为网络行为的研究提供量化的依据。但是面对海量的 IP 数据，通常采用抽样方法，提取出少量的数据进行运算。

Claffy 在 1994 年基于对 4K 左右的主干 IP Trace 的分析，给出了不同的抽样方法对报文长度分布测度的影响。本文将采用相似的方法对近期 CERNET 主干网流量进行分析，主要得出以下结论：（1）系统抽样和分层随机抽样差别较小；（2）在显著性水平为 0.05 时，抽样的统计结果在报文大小的测度上与总体分布一致；(3)报文驱动和时间驱动差别较大，前者更准确一些；（4）随着网络的发展，长报文所占的比例越来越大。

1 数据对比和实验方法

1.1 数据比较和报文长度测度

表 1 展示和比较了 Claffy 和本文实验分别使用的 Trace。

报文长度测度定义为：根据 IP 报文头属性，记 P_i 表示第 i 个报文到达的长度。

为了更好的反映报文长度的分布情况，记样本容量为 n，m 为有效报文数，表示抽样报文的报文长度小于或者等于抽样粒度的个数。称测度值

$$EPP_F\text{（Efficient Packets Percentage）}=\frac{m}{n}$$

为 Trace F 的有效报文比。

1.2 实验方法

Claffy[1-2]研究抽样方法为：对同一段 Trace，分别采用系统抽样和分层随机抽样的方法，在时间驱动和报文驱动的方式下，进行实验，并按照特定评估标准来比较实验结果。本文采用的实验方法与 Claffy 类似，但作用于三段数据。

本文中所做的实验对比图均为比较 Trace 在不同粒度下的有效报文比。

1.3 评估标准

（1）抽样间隔和报文粒度的选取

根据 Cochran[3]提出的理论，在报文大小这个测度上，作者根据以下公式来选取合适的样本容量：

$$n=\left(\frac{100zs}{tm}\right)^2$$

Table1 Information of collected traces

	采集点	网络层次	采集日期	采集时间	数据总量(Pkt)
Claffy	Urbana-Champaign FDDI into NSFNET	主干网	1993-03	14: 11--15: 11	4019 k pkt
数据 1	CERNET 某 2.5G 省网边界链路	主干网	2005-11	14: 00--15: 00	2350 M pkt
数据 2			2007-01	14: 00--15: 00	2075 M pkt
数据 3			2008-05	14: 00--15: 00	458 M pkt

注:“pkt”表示报文(packet)，三段 Trace 的采集日均为工作日(the time of colleted traces is working day).

资金项目：国家 973 重点基础研究发展规划(2003CB314804)；国家科技支撑计划课题(2008BAH37B04)资助
收稿日期：2008-8-15
作者简介：张巍(1982-)，男，硕士研究生。E-mail:wzhang@njnet.edu.cn

其中 m 是样本均值， s 是样本标准差，在显著性水平为 0.05 时，z 的值取 1.96。对于文中的 Trace 集，作者以 2007 年的 Trace 为例：数据 2 报文个数大概有 $2.075*10^{10}$ 报文，报文大小的平均值为 596，报文大小的标准差为 925，当精度取 t =1%时，根据公式得出抽样大小为 92534，所以根据 2 的冥特性和特定报文大小的属性，抽样间隔分别取 8、128、256、1024、8192、32768、131072。

Claffy 的分析同时选择 41 和 180(bytes)2 个点，作为测度计算参数，这是因为这 2 个点是有效报文比为 1/3 和 2/3 的点，本文也通过实验分别计算了 3 组实验数据测度值为 1/3 和 2/3 的点。详见表 2，同时，为了获得更多的测度值，测度计算参数分布选取了 16、24、32、48、64、128、256、512、1024。

Table2 Three ranges of traces (bytes)

	1/3 点	2/3 点
Claffy（1993 年 3 月）	41	180
数据 1（2005 年 11 月）	51	1260
数据 2（2007 年 1 月）	56	1060
数据 3（2008 年 5 月）	66	1160

（2） c^2 检验和 f 系数

本文沿用 Claffy 的评估标准 Pearson's 的 c^2 检验：

$$c^2 = \sum_{i=1}^{B} \frac{(O_i - E_i)^2}{E_i}$$

其中，B 表示可能出现的属性的总数，在这里表示报文的区间数， O_i 表示第 i 个属性出现的次数，在本文中表示落在第 i 个区间的报文数，即观察值， E_i 表示理论上落在第 i 个区间的报文数，即理论值。

由于 Pearson's 的 c^2 检验对样本数量比较敏感，不能准确比较出不同抽样比的实验结果，Fleiss[4]提出了另一种与样本容量无关的 f 系数：

$$f = \sqrt{\frac{c^2}{n}}$$

其源于 c^2 检验， $n = \sum_{i=1}^{B} (O_i + E_i)$ 。

f 系数的意义在于： f 值越接近 0，表示抽样精度越高，抽样方法越能反映出总体的分布规律，与之越贴切，反之， f 值越大，表示抽样方法越差，与总体的分布规律分歧也越大。

2 实验结果与分析

2.1 报文驱动下的系统抽样

由于 3 组数据的结果图形基本相似，本文仅以数据 1 的实验结果为例。

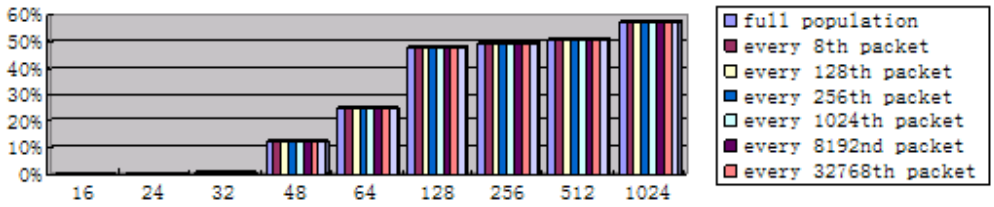

Figure1 Distribution of packet sizes as a function of sampling granularities (packet-triggered ,systematic sampling)

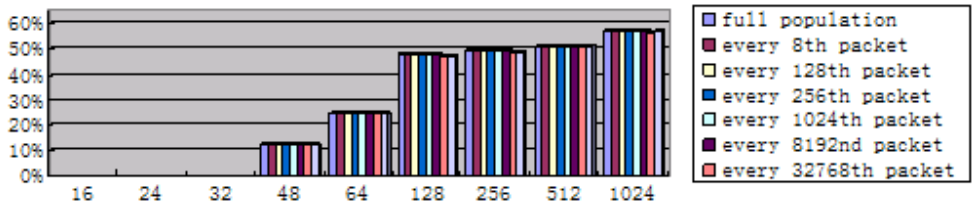

Figure2 Distribution of packet sizes as a function of sampling granularities (packet-triggered ,random sampling)

2.2 报文驱动下的随机抽样

图 2 表示的方法是在报文驱动方式下，随机抽样的统计图。从图 1 和图 2 的对比可以看出，在报文驱动的方式下，系统抽样和随机加，抽样间隔对实验结果的影响也不是很明显，均能较好的反映出总体报文大小的分布特征，但前者的结果更好。

2.3 时间驱动下的系统抽样

图 3 表示的是在时间驱动方式下，系统抽样的统计图。其中，时间的抽样间隔与报文驱动时的抽样间隔尽量保持一致。对比图 1 和图 3 可以看出，报文驱动要比时间驱动更精确一些。

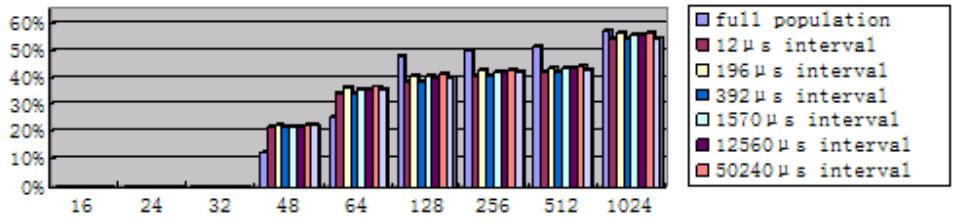

Figure3 Distribution of packet sizes as a function of sampling granularities (time-triggered ,systematic sampling)

2.4 f 值对比

图 4、图 5 是基于报文驱动下，在 3 组数据上分别进行系统抽样和随机抽样的 f 值对比图。可以看出在抽样间隔为 1024 之前 6 组数据都比较接近，随着抽样粒度的增加，f 值也有增大的趋势。但是 f 值总体比较小，为 10^{-4} ～10^{-2} 之间。

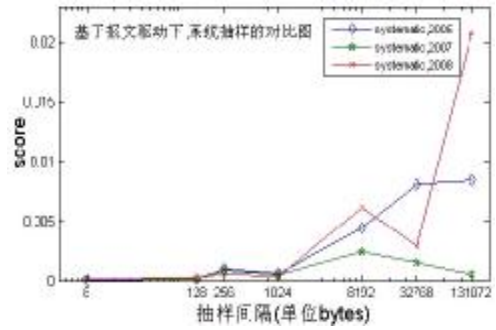

Figure4 f -value scores as a function of sampling fraction for packet size distribution (packet-triggered ,systematic sampling)

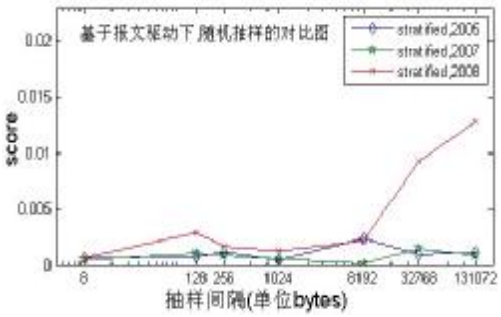

Figure5 f -value scores as a function of sampling fraction for packet size distribution (packet-triggered ,random sampling)

可以看出，与图 4、图 5 不同的是，图 6 表示在时间驱动下，f 值波动比较大，而且在大多数都在10^{-1}以上。通过 3 幅图的对比可以看出：在报文驱动下，抽样结果差别比较小；而时间驱动下，差别比较大。这是因为由于报文到达的突发性所致，采用时间驱动的方法，由于依赖报文的到达间隔，就会导致抽样的精确性降低，甚至丢失采样报文。

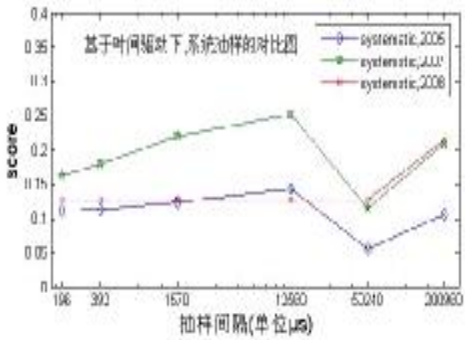

Figure6 f -value scores as a function of sampling fraction for packet size distribution (time-triggered ,systematic sampling)

3 结论与展望

针对网络测量的分析需要和系统资源的限制，本文研究了抽样机制对报文长度分布测度影响。通过实验比较了各种抽样机制，实验证明，在报文长度分布测度上，基于报文驱动下系统抽样精确度最高，而且代价最小。

本文还通过对 3 组 Trace 的观察，得到以下结论：

(1) 系统抽样和分层随机抽样差别较小。

(2) 在显著性水平为 0.05 时，对试验数据进行抽样后的测度计算，得出的结果与总体基本一致。

(3) 通过 f 值的对比，得出基于报文驱动技术比基于时间驱动技术的抽样精确性要高一些，但是二者之间的差别不大。

(4) 由于网络发展的变化，Trace 信息中报文大小的比例也发生了很大的变化，长报文所占的比例越来越大。

本文提出的抽样机制在一定程度上能刻画出样本与总体之间的特性，但并不是绝对的。由于时间驱动下，随机抽样的复杂性以及与其他三种的抽样比不同，本文没有考虑，希望今后的研究能有所完善。其次，就目前 Trace 采集系统而言，微秒的到达间隔精度还是有所欠缺，由于目前网络的突发性的增多，如何更有效的刻画网络行为也是需要解决的问题之一。

参考文献：

[1] KC.Claffy, GC.Polyzos, Hans-Werner Braun. Application of sampling methodologies to network traffic characterization. [J]ACM SIGCOMM Computer Communication Review, 1993,23(4):194-203

[2] KC.Claffy, GC.Polyzos, Hans-Werner Braun. A parameterizable methodology for internet traffic flow profiling. [J]IEEE Journal on Selected Areas in Communications, 1995, 12 (8): 1481 - 1494.

[3] W.Cochran. Sampling Techniques. [M]John Wiley & Sons, 1987.

[4] J.Fleiss. Statistical Methods for Rates and Proportions. [M]John Wiley & Sons, 1981.

[5] 王远,丁伟,龚俭. TCP 数据流超时研究[J]. 厦门大学学报, 2007, 46 (2) : 192 –195.

Study on Influence of Sampling Methodologies to the Metrics of Packet Size

ZHANG Wei, DING Wei, GONG Jian

(CERNET Eastern China (North) Regional Network Center, School of Computer Science and Engineering, Southeast University, Nanjing 210096, China)

Abstract: Since drawn by Claffy in 1994, it has been well-accepted so far that the time-triggered techniques did not perform as well as the packet-triggered ones. Furthermore, the performance differences within each class (packet-based or time-based techniques) are small. Is this conclusion still suitable now? Base on three one-hour packet traces collected from a CERNET 2.5 G province net border link recently, we did methodology as Claffy's and analyzed on metrics of packet size. Some interesting characters behind the traces are appeared, which can be used for network measure and network behavior research.

Key words: sampling; time-triggered; packet-triggered; metrics