Measure Correlation Analysis of Network Flow Based On Symmetric Uncertainty

Shi Dong[1 2], Wei Ding[1 2] and Liang Chen[2]
[1] Computer Science and Engineering, Southeast University,
Nan Jing, China
[2] Key Laboratory of Computer Network and
Information Integration, Southeast University
Nan Jing, China
[e-mail: {shdong, wding}@njnet.edu.cn]
*Corresponding author: Shi Dong

Received October 10, 2011; revised March 6, 2012; revised April 25, 2012; accepted May 16, 2012; published June 25, 2012

Abstract

In order to improve the accuracy and universality of the flow metric correlation analysis, this paper firstly analyzes the characteristics of Internet flow metrics as random variables, points out the disadvantages of Pearson Correlation Coefficient which is used to measure the correlation between two flow metrics by current researches. Then a method based on Symmetrical Uncertainty is proposed to measure the correlation between two flow metrics, and is extended to measure the correlation among multi-variables. Meanwhile, the simulation and polynomial fitting method are used to reveal the threshold value between different correlation degrees for SU method. The statistical analysis results on the common flow metrics using several traces show that Symmetrical Uncertainty can not only represent the correct aspects of Pearson Correlation Coefficient, but also make up for its shortcomings, thus achieve the purpose of measuring flow metric correlation quantitatively and accurately. On the other hand, reveal the actual relationship among fourteen common flow metrics.

Keywords: Network behavior，flow metrics， correlation，Pearson correlation coefficient，mutual information，symmetrical uncertainty

This project was supported in part by: State Scientific and Technological Support Plan Project of China under Grant No. 2008BAH37B04, National Basic Research Program of China (973) under Grant No. 2009CB320505, and National Natural Science Foundation of China under Grant No. 60973123.

DOhttp://dx.doi.org/10.3837/tiis.2012.06.009

1. Introduction

Research on network behavior can provide theoretical support for network management, network optimization, protocol design, QoS and SLA, It is important to understand the properties of such traffic for traffic monitoring and modeling purposes. Analysis of correlation between the metrics can help researcher to investigate the determinants of key variables, such as traffic identification and further better understand the behavior of network traffic.meanwhile,different evaluation methods have own application scope. so the effective method to evaluate correlation between flow metrics is becoming very important.

However, although some research have been focus on the network flow length, flow rate and flow arrival[1][2][3][4][5],and the flow metrics for evaluation on application layer protocol[6][7][8][9]. However, through the following analysis and studies, The analytical method has some limitations. Such as the analysis of the metric is also not comprehensive enough, understanding the correlation between current network flow metrics is not fully and do not meet the demands of filtering redundancy metric. It is also necessary to study further correlation between the IP flow metrics.

Therefore, this paper point out some problems existing in current study which methods on the correlation of flow metrics.The introduction of the mutual information method and symmetric uncertainty (SU) is based on information theory which was considered as a quantitative description tool of two metric correlation and proposed multi-metric relationship between the measurement methods which can improve accuracy and versatility of measurement methods of network traffic metric correlation.Meanwhile,based on method of simulation and polynomial fitting, Obtain the threshold of any metrics correlation which is evaluated by SU method. Finally, the actual use of multiple trace analyze the metrics of the current flow, In orde to deeply analyze symmetric uncertainty(SU) method.We compare accuracy of SU method with Pearson correlation coefficients which were used to do the correlation analysis of network traffic.The results show that SU method is better than others on the analysis of network traffic metrics correlation, and summed up status of the current network traffic metrics correlation. Compared to previous research, the contribution of this paper is:

(1)The use of SU method as tool to evaluate correlation between the standard flow metrics that is more accurate, and extend it to evaluate any dimension metrics correlation between the vectors.

(2)According to the threshold of the SU method,metric correlation is divided in different correlation levels (high/moderate/low correlation), which can provide quantitative analysis for flow metrics.

(3)This paper firstly analyzes correlation between the current IP flow metrics.

This paper is structured as follows: Section 2 analyzes research status of the correlation between flow metrics and points out defects of the Pearson correlation coefficient. Section 3 introduces and analyzes the SU method to evaluate the correlation between two metrics and any metrics and according to induction analysis based on SU method to obtain the threshold of metric correlation.In Section 4,through experiments and theoretical analysis,we compare the

Pearson correlation coefficient with SU to evaluate the accuracy of two methods, and analyze the correlation between usual flow metrics; Section 5 is summary of the paper

2. Related Work

2.1 Metric Correlation

With the development of Internet,Study on flow metrics has being a research focus, but only one metric feature is discussed and reseached,such as IP traffic flows length distribution, TCP flow arrival characteristics,flow duration, etc. only a small amount of paper discusses the correlation between the different flow metrics.

In the study of metric correlation on TCP/IP layer, Paper[3]showed that the TCP flow length had strong correlation with flow rate,The Longer TCP flow length is, the greater average flow rate. Paper[1] and [3] make use of the dot plot and Pearson correlation coefficient statistical to analyse the relationship among "heavy-hitter" flow length, flow duration, flow rate and flow burstiness.which also point out strong correlations between flow length, flow rate and flow burstiness.Meanwhile, such strong correlations between the flow metrics is limited to long TCP flows, for short TCP and UDP flows,which are more related to transport protocols (such as TCP) and high-level applications (such as DNS, HTTP, etc.).In paper[2],Zhou also adopted the same method to analyze the correlation between TCP and UDP flow duration, average flow rate and flow stability, also pointed out flow length and flow duration had a strong positive correlation, but correlation between flow length and flow rate is not obvious.

In recent years, various new application layer protocol are always being developed (such as various types of P2P protocols), and hope to meet a more detailed analysis demand of flow behavior, focus of behavioral measure gradually is shifted to the distribution of flow measure of the application protocols characterization. However, when related to methods of measures correlation, in all the paper, only adopted dot plot to depict relationship of two metrics. Such as S. Saroiu et al. [7] study Gnutella and Napster applications,he points out statistical relationship between downloading speed and network delay,the number of shared files and the number of shared bytes; K.Tutsvhku et al.[8] study eDonkey flow behavior to obtain the relationship chart of flow bytes and the duration .C.Dewes[6] and L.Plissonneau [9] respectively, studied and compared four P2P applications, and point out statistical relationship between the sending bytes number and the receiving bytes number .

According to discussion on the above metric correlation,although the prevalence study on metric correlation has been going on,the metric analysis methods and results are limitations, summarized as follows:

(1)Analysis method is only considered to adopt a dot plot or dot plot + Pearson correlation coefficient method. However, through the following analysis shows,Pearson correlation coefficient unable to adapt to some flow metrics for variable characteristics,and can not represent more complex relationship between the metrics, leading to results of the analysis for flow metric correlation to exist the bias and limitations.

(2)Most of the object of analysis metric only limited flow length,flow rate and flow duration,with little involved in the behavior of other metrics, such as the average packet length in each flow, ratio of the bidirectional packets number in each flow and so on.

(3)The study on definition of the metric are not the same as in paper[1] and paper[4] which the number of bytes in each flow is defined as flow length.however,Paper[2] points out that packets number in each flow is defined as the flow length. For theoretical analysis, unless proved that the two metrics have height correlation by themselves, the above research does not have the comparative study.

2.2 Pearson correlation coefficient

The current study on quantitative description of flow metric correlation only used all the Pearson correlation coefficient as main tool. Pearson correlation coefficient (hereinafter referred to as P coefficient) [10] is described as the relationship between two variables in the mathematical statistics and statistical methods which can evaluate correlation. The formula is as follows:

$$\rho_{XY} = \rho_{YX} = \frac{Cov(X,Y)}{\sqrt{Var(X)\cdot Var(Y)}} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{i=1}^{n}(y_i-\bar{y})^2}} \qquad \text{F1}$$

Where X and Y are two random variables, (x_i, y_i) (i = 1, 2, …, n) is n pairs observed value of the two variables,and $\bar{x}$ and $\bar{y}$ were respectively the mean of n observed value. $|\rho_{XY}| \in [0, 1]$, the greater value is,the stronger correlation is;otherwise is weaker. Although the Pearson correlation coefficient can be expressed as the correlation between the distribution of variables, but its variable distribution and the results that have certain requirements and limitations. The previous study did not further analyze flow metric variable, so it did not find the following problems,when the Pearson correlation coefficient was applied to evaluate correlation between network traffic metrics.

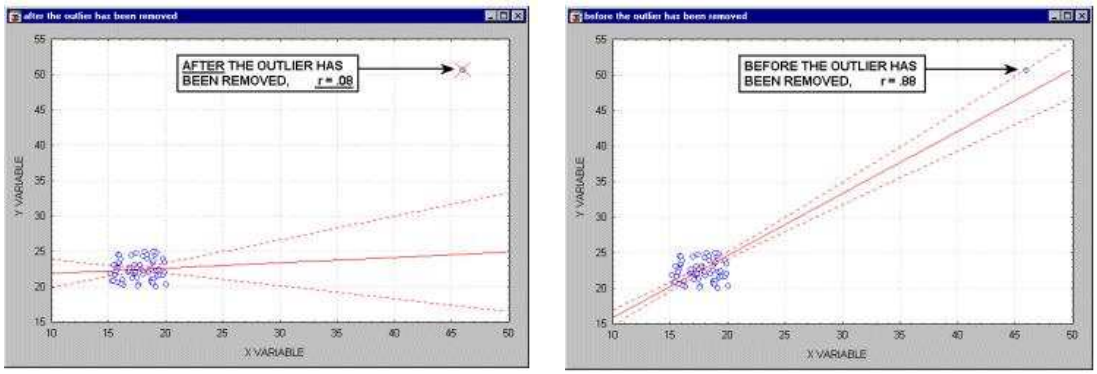
Question 1. Nonlinear correlation of network traffic metric

The potential problem of pearson correlation coefficient is related to the type of variable. P coefficient is only used to evaluate the positive and negative linear relationship between the two variables, linear deviation will increase the total square sum of linear regression line, even if which can reflect the real relationship between two variables. For example, there is a strict functional relationship between the metric "the average packet arrival interval (IAT) in a flow" and the metric "packets per second (PPS)": IAT = 1 / PPS, however, it can not be used. because the relationship between the two metrics can not be discribed by linear equation, P coefficient is small between the two metrics, and not as a fixed value.Further proof can be seen in section 4.1.

Question 2. Heavy-tailed distribution of network traffic metric

Another drawback of Pearson correlation coefficient is the extreme value problem (or singular value problems). The so-called extreme value is observations value away from centralized data in variable distribution. In **Fig. 1**, (X,Y) is concentrated in the point (17, 22) and there is only an isolated point (47, 50) called extreme values. Since regression line is not decided by the minimum distance sum but decided by the minimum distance squared sum, so extreme values will make greater impact on the regression line slope and P coefficient value size. As long as there is an extreme value ,which is likely to dramatically change the slope of the

regression line and correlation coefficient; and extreme values may not only artificially reduce the correlation coefficient, may also increase the correct correlation. **Fig. 1(a)**, when not considering the extreme values, P coefficient of X and Y is 0.08, after adding an extreme values (47, 50), as shown from the **Fig. 1 (b)**, P coefficient of X and Y increases from 0.08 to 0.88. So the existence of only an extreme value change the correlation coefficient,originally close to 0,to strong correlation. It is impossible to caculate metric correlation only based on the Pearson correlation coefficient.



(a) coefficient when delete extreme values (b) coefficient when add extreme values

Fig. 1. Extreme value of variable distribution

Heavy-tailed distribution of network flow metric also has problem of extreme values. The so-called heavy-tailed distribution is are probability distributions whose tails are not exponentially bounded: that is, they have heavier tails than the exponential distribution. **Fig. 2** shows the relationship points chart of flow packets number(pkts) and flow duration (duration) for trace CERNET_b(discribled in section 4),As is shown from **Fig. 2**, two metrics value concentrated in small observations, Pearson (pkts<= 50) = 0.02933; and tail, then there exists a strong linear positive correlation, Pearson (pkts>50) = 0.76239 . Although the number of pkts>50 only occupied about 2% in the total sample, largely increased the coefficient of P between the two metrics, making the whole Pearson (pkts, duration) = 0.66903, thus the disorder two variables becomes stronger approximate correlation.

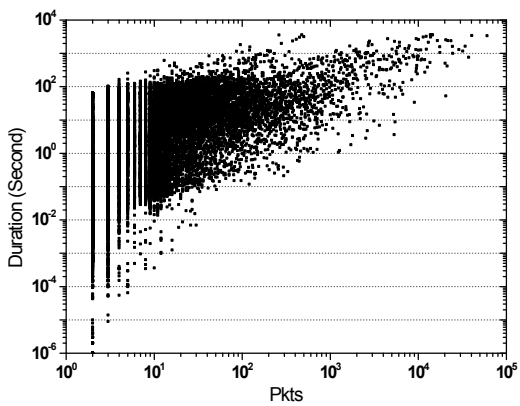


Fig. 2. Relationship chart of trace CERNET_b flow length and flow during time

Question 3. Network traffic multi_metric correlation

Pearson correlation coefficient limited study on metric correlation of current network flows to two single-related metrics,not involving the discussion on multiple metric correlation. However, only study on the correlation between two metrics can not fully reveal the inherent laws of flow behavior, if further analyze complex multi_metric correlation, you can see that multiple factors will affect one metric distribution, which will further reveal the inherent of network traffic. But only simple Pearson correlation coefficients will be not able to discrible any dimension statistical correlation. Multiple correlation based on P coefficient can only evaluate a limited multi-linear relationship. Therefore, it is necessary to further analysis for P-factor method, or choose another way to evaluate the correlation.

2.3 Correlation analysis

There are some solutions in the field of mathematical statistical analysis,such as variable transformation,non-parametric correlation coefficient [12] which can solve the first two defects of pearson correlation coefficient that are mentioned in the previous section, Generally speaking,based on variable distribution characteristics in the points map,these methods choose the appropriate functions and parameters to change correlation from curve to linear, then use P coefficient analysis. However, for many of the network metrics, this method requires to analyze the actual distribution characteristics of each metric variable,moreover, the different functions and parameters have to be selected to deal with this problems, so it will be heavy workload and not a generic.

On the other hand, correlation analysis is widely used as an optional sub-process of feature selection(technique of selecting a subset of relevant features for building robust learning models), to determine the importance of features by evaluating whether feature's correlation with class. SU-based,Relief (based on Gini index) and minimum description length (MDL) method have been proposed.Paper [12][13] compared three methods, and pointed out the MDL approach works best when number of samples were sufficient, but the SU method is the most stable. P. Mitra et al.[14] proposed a new fast method based on measuing similarity between features whereby redundancy therein is removed, but its research object only focus on the linear correlation. In order to make up for the deficiency of linear correlation, Jing Yuan et al. [15][16][17] introduced entropy and mutual information in information theory, put forward their own fast feature selection method, and obtain good classification results. Qu, G et al.[18] still was based on the entropy theory, and pointed out the previous method of calculating correlation metric (including the SU) could not accurately represent the correlation when there exists decision-making variables, and propose a calculation method of correlation to be applied to feature selection of network anomaly detection when decision variable exists. T Ganchev et al. [19] select the most important features and feature vector obtain verification, and evaluate the pros and cons of five kinds of feature selection which include information gain, gain ratio, symmetric uncertainty, correlation-based feature selection, support vector machines, Results show that the SVM is less effective, the other four methods were similar. Shuang Hong Yang et al.[24]presents a theoretically optimal criterion, namely the discriminative optimal criterion (DoC) for feature selection. Compared with the existing representative optimal criterion (RoC,which retains maximum information for modeling the relationship between input and output variables. Huawen Liu et.al [25] propose a new feature selection algorithm based on dynamic mutual information,which is only estimated on unlabeled instances. The experimental results indicate that the algorithm achieved better

results than other methods in most cases.

Through the above analysis can be found,entropy theory [20] is increasingly being introduced and use. Because calculation of conditional entropy and mutual information does not require to make any assumptions about the variable distribution, it can be adapted to the variables correlation analysis.Therefore, this paper introduced the entropy into metric correlation between the network flow, adopt the same method with paper [17][18], to eliminate defects produced by the past pearson correlation coefficient, to achieve the goal of accurate analysis on current commonly correlation between flow metrics.

3. Uncertainty Measure Based On Symmetric Correlation Analysis

3.1Relationship Between The Two Analytical Methods Measure

Entropy represents the uncertainty of random variables, that is how much the amount of information contained in the metrics. Let X be a random variable, its entropy is defined as:

$$H(X) = -\sum_i P(x_i)\log_2(P(x_i))$$

Which P (x_i) is prior probability of the random variable X ,that is P (x_i) = P (X = x_i). H (X) is greater and entropy of variable X is greater, that X, the greater is the uncertainty, the greater is the amount of information carried by the self.

If values of another random variable Y observed have been confirmed in the case, the variable X, conditional entropy is:

$$H(X|Y) = -\sum_j P(y_j)\sum_i P(x_i|y_j)\log_2(P(x_i|y_j))$$

P (x_i | y_j) shows that value of random variable Y is y_j confirmed,that probability of a random variable X value is x_i and P (x_i | y_j) is called the posterior probability. Because H (X) shows the uncertainty of X, and H (X | Y) shows the variable X remains the uncertainty when values of random variable Y are confirmed,H (X)-H (X | Y) mean that the random variable Y provide the information of variable X, that is called the mutual information of X and Y in information theory, expressed as

I (X ; Y):

$$I(X;Y) = H(X) - H(X\,|\,Y) = H(X) + H(Y) - H(X,Y) \qquad \text{F2}$$

$H(X,Y) = -\sum_i \sum_j P(x_i, y_j)\log_2(P(x_i, y_j))$ is called the joint entropy of two variables

I (X; Y) shows that the average information of the variable X when variable Y value has been confirmed, it also show statistical constraints of two random variables. And, the mutual information of two variables have symmetry. Just as type F2, and obtain the equation:

I (X; Y) = I (Y; X). If the variables X and Y are not related, then I (X; Y) = 0; otherwise I (X; Y)> 0, and I (X; Y) is greater, indicating that the correlation of X and Y is the stronger. That is, if I (X; Y)>I (Z; Y), then Y and X is more relevant than Y and Z. Therefore, we can use mutual information I (X; Y) to quantitatively evaluate the correlation of two metrics. However, results of I (X; Y) were affected by variable values and units, so it will be further homogenization,SU (Symmetrical Uncertainty) [22] is defined as follows:

$$SU(X;Y)=SU(Y;X)=2\times\left[\frac{I(X;Y)}{H(X)+H(Y)}\right] \qquad \text{F3}$$

SU range in [0,1] is the same to Pearson correlation coefficient ρ; and which is a monotone increasing function of the mutual information I (X; Y)[21], the value is the greater, correlation of two variables is the stronger , and vice versa, show correlation is the weaker; if the value takes 0 ,two variables are mutally independent, if takes 1 show that two variables exist strict function relationship. Since SU has a high accuracy and versatility, this paper will introduce it into field of the network flow metric correlation, as a quantitative evaluation standard of two flow metrics correlation.

3.2 Relationship analysis between multi-metrics

Considering the defects that pearson correlation coefficient for the metric can not deal with more complex relationship, and current study is lack of effective method to weigh relationship between the multi-metric, so the section 3.1 propose to expand definition mutual information of two variables, and obtain the following proposition 1 shows that mutual information of the any dimension random vectors, that means it can evaluate correlation of the multiple flow metrics of statistical information:

Proposition 1. $I(\vec{X};\vec{Y})=H(\vec{X})+H(\vec{Y})-H(\vec{X},\vec{Y})$ $\vec{X}$, $\vec{Y}$ is random vector of ***m, n*** dimension

$$H(\vec{X})=-\sum_i P(\vec{x}_i)\log_2(P(\vec{x}_i)),$$

$$H(\vec{X},\vec{Y})=-\sum_i\sum_j P(\vec{x}_i,\vec{y}_j)\log_2(P(\vec{x}_i,\vec{y}_j))。$$

Proof: Combined random vector values of all the variables are mapped into a single random variable value.that is formula F2

Theorem 1. $I(\vec{X};\vec{Y})$ The following show its recursive relation

$$I(\vec{X};\vec{Y})=\sum_{i=1}^{n} I(\vec{X};Y_i|Y_{i-1},Y_{i-2},\cdots Y_1)$$

Proof:

$$\begin{aligned}
&I(\mathbf{X};\mathbf{Y})=I[\mathbf{X};(Y_1,Y_2,\cdots,Y_n)]=I[(Y_1,Y_2,\cdots,Y_n);\mathbf{X}]\\
&=H(Y_1,Y_2,\cdots,Y_n)-H(Y_1,Y_2,\cdots,Y_n|\mathbf{X})\\
&=\sum_{i=1}^{n}H(Y_i|Y_{i-1},\cdots,Y_1)-\sum_{i=1}^{n}H(Y_i|Y_{i-1},\cdots,Y_1,\mathbf{X})\\
&=\sum_{i=1}^{n}[H(Y_i|Y_{i-1},\cdots,Y_1)-H(Y_i|Y_{i-1},\cdots,Y_1,\mathbf{X})]\\
&=\sum_{i=1}^{n}I(\mathbf{X};Y_i|Y_{i-1},Y_{i-2},\cdots Y_1)
\end{aligned}$$

The meaning of theorem 1 is that average information of $\vec{X}$ provided by random variables Y_1, ..., Y_n is equal to the average information of $\vec{X}$ provided by Y1 and accumulation of the average information of $\vec{X}$ provided by Yi when Y_1, ..., Y_{i-1} (i = 2, ... , n) are known, Therefore,

when the actual flow metrics correlation are calculated, It can enable less mutual information of metrics to push out mutual information between multiple metrics, simplify the calculation process.

Theorem 2. $I(\vec{X};\vec{Y}) \geq I(\vec{X};Y_i)$, $i = 1,2,\cdots,n$. Enquation conditions is satisfied , if and only if for all meet $P(\vec{x},\vec{y}) > 0$. $(\vec{x},\vec{y})$, and exists $P(\vec{x} \mid \vec{y}) = P(\vec{x} \mid y_i)$ 。

Proof: from theorem 1 we can see ,

$$I(\vec{X};\vec{Y}) = \sum_{i=1}^{n} I(\vec{X};Y_i | Y_{i-1}, Y_{i-2}, \cdots Y_1)$$

$$= I(\vec{X};Y_i) + I(\vec{X};Y_1 \mid Y_i) + \cdots + I(\vec{X};Y_{i-1} \mid Y_1,\cdots,Y_{i-2}) + I(\vec{X};Y_{i+1} \mid Y_1,\cdots Y_i) + \cdots + I(\vec{X};Y_n \mid Y_1,\cdots Y_{n-1})$$

$$\because I \geq 0$$

$$\therefore I(\vec{X};\vec{Y}) \geq I(\vec{X};Y_i)$$

when variables $Y_1,\cdots,Y_{i-1},Y_{i+1},\cdots,Y_n$ are independent with $\vec{X}$, then The conditional mutual information values are 0, $I(\vec{X};\vec{Y}) = I(\vec{X};Y_i)$ 。

Theorem 2 has a very intuitive meaning of statistics: information provided by random vector $\vec{Y}$ is no less than information by any component Y_i .That is, the correlation between the random vectors is stronger than any component between vectors.

Similarly, in order to overcome the effect of variable unit for metric correlation, Reference formula F3 enable $I(\vec{X};\vec{Y})$ homogenization, and obtain definition of the expansion of symmetrical uncertainty correlation between any dimensional random vectors:

$$\text{Definite}: SU(\vec{X};\vec{Y}) = 2 \times \left[\frac{I(\vec{X};\vec{Y})}{H(\vec{X}) + H(\vec{Y})} \right], \ \vec{X}, \vec{Y}$$

Respectively, $\vec{X}$, $\vec{Y}$ is m, n dimensional random vector.

According to definition of $I(\vec{X};\vec{Y})$ in Proposition 1, the symmetry uncertainty of multi-variables $SU(\vec{X};\vec{Y})$ exists the same in the [0, 1], and the result value is greater, correlation of $\vec{X}$, $\vec{Y}$ is the stronger. In particular, when m = 1, n = 2, then

$$SU(X;YZ) = 2 \times \left[\frac{I(X;YZ)}{H(X) + H(YZ)} \right] = 2 \times \left[1 - \frac{H(XYZ)}{H(X) + H(YZ)} \right] \quad \text{F4}$$

Formula F4 shows the complex correlation between three metrics. And there is:

$$SU(X;YZ) \geq SU(X;Y)$$

$$SU(X;YZ) \geq SU(X;Z)$$

3.3 SU method to determine the relevance threshold

When actual measuring the correlation between different objects, generally there is no strict correlated and uncorrelated, the correlation of objects are somewhere between them. Therefore, when we actual analyze and study correlation degree of variables that is usually divided into the following three areas:

1 Low correlation: the joint distribution of variables is irregular.

2 Moderate correlation: the joint distribution of variables have a more significant trend.

3 Highly correlation: the basic can infer from a variable to another approximation variable.

The use of SU method can weigh the correlation between variables, but SU only can weigh the order (from strong to weak) of multi-metric correlation. For example, if the $SU(\mathbf{X};\mathbf{Y})$ = 0.95, $SU(\mathbf{X};Z)$ = 0.60, then the correlation between X and Y is stronger than X and Z. However, there is no corresponding research on the own absolute sense of SU's results, and SU value range of the influence on the metric correlation. For example, if the SU values between the random variables X and Y has reached 0.95, but the current study did not have confirm degree of correlation.

To solve this problem, this section make use of the Pearson correlation coefficient to divide value range of linear correlation degrees, board threshold of the SU method is fitted out through experiment, and considered it as the critical value which is evaluated by SU method. Two assumptions will be given as follows:

Assumption 1: Ability of SU evaluation methods to value of extreme linear relationship is not weaker than the Pearson correlation coefficient.

Assumption 2: SU evaluation method to linear and nonlinear correlation between variables has the same ability.

The two ***Assumptions*** have their rationality: First, since SU is more accurate than the Pearson correlation coefficient on variables correlation, so when weighing any variable, SU's ability is at least fairly with Pearson correlation coefficient. Second, through the above analysis, SU only judges based on the values of variables, does not need to distinguish between types of variables.

On the other hand, when Pearson correlation coefficient weigh the linear relationship, existing the following correlation degree [10]:

$$\begin{cases} 0.0 \le SU(X;Y) \le 0.4 & low \\ 0.4 < SU(X;Y) < 0.8 & \text{mi}\,ddle \\ 0.8 \le SU(X;Y) \le 1.0 & high \end{cases}$$

Since the Pearson correlation coefficient weighing the linear relationship of no extreme values can achieve accurate results, therefore, according to the assumption 1, can be based on SU and the Pearson correlation coefficient weighing the linear correlation of no extreme values correspond to the correlation of calculated value, and relevant level threshold of the Pearson correlation coefficient, summarized classification threshold when SU weighing the relevance of linear correlation; and further based on the assumption 2, the threshold value will make use of SU analysis to extend to all types of correlation between the level of classification threshold.

The experimental data is randomly generated by MATLAB tool, a total 57,270 groups of data and every group of data contain 10,000 <X, Y> value, and there exists no extreme values. For two random variables X and Y in each group we calculated the Pearson correlation coefficient and SU values, the results between both shown in **Fig. 3**.

Seen from **Fig. 3**, when weighing linear correlation of no extreme values, Values of SU represent a monotone increasing function of Pearson correlation coefficient.

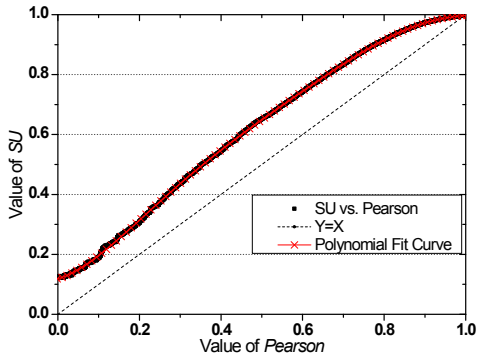


Fig. 3. Relationship chart of Pearson and *SU*

In order to accurately adopt SU method to obtain threshold based on the Pearson correlation coefficient of variables, firstly, this paper proposed polynomial fitting of order n based on least squares for the experimental data, that is.

$$SU = a_0 + a_1\rho + a_2\rho^2 + a_3\rho^3 + \cdots a_n\rho^n$$

As is shown in **Table 1**, the coefficient of determination (COD) and standard deviation (SD) is obtained.

Table 1. Pearson Correlation coefficien and *SU* standard deviation

Order number	Coefficient	COD	SD
1	a_0= 0.12983, a_1=0.94608.	0.9883	0.0354
2	a_0= 0.0789, a_1=1.33674, a_2=-0.39194.	0.99811	0.01423
3	a_0=0.10189, a_1=0.98362, a_2=0.5325, a_3=-0.61679.	0.99964	0.00621
4	a_0= 0.10428, a_1=0.92434, a_2=0.81773, a_3=-1.0686, a_4=0.22567.	0.99965	0.0061
5	a_0= 0.1159, a_1=0.50967, a_2=3.94061, a_3=-9.65003, a_4=9.9739, a_5=-3.89679.	0.99989	0.0034
6	a_0= 0.11999, a_1=0.31216, a_2=6.07593, a_3=-18.53704, a_4=26.95102, a_5=-18.93909, a_6=5.01375.	0.99992	0.00298

Table 1, SD representing standard deviation of polynomial fitting, which should be as small as possible; COD curve fitting represent determination coefficient based on the residuals, as follows (where e represents the fitting residual):

$$COD = 1 - \frac{\sum_i e_i^2}{\sum_i (Y_i - \overline{Y})^2} = 1 - \frac{\sum_i (Y_i - \hat{Y}_i)^2}{\sum_i (Y_i - \overline{Y})^2}$$

0≤COD≤1, and the greater is the value of COD, the better is fit.

In the polynomial fitting method, the higher is the order of the polynomial, the better (**Table 1**)is curve fitting, but higher-order polynomials represent high computational complexity; Therefore, for general applications, third-order and four-order polynomial can completely meet accurate fit requirements. To reflect this relationship as accurately as possible between SU and Pearson correlation coefficient, using a six-order polynomial fitting, curve shown in Figure 3 × line shown in red, the error probability is less than 10^{-4} (1-COD = 1-0.99992 = 0.00008).

We further use the Kolmogorov-Smirnov method to fit in order to demonstrate that the sixth-order polynomial can characterize the relationship between Pearson correlation coefficient and SU. Kolmogorov-Smirnov (KS) goodness fit test [23] is a testing methods of distribution function based on experience (ECDF), and test whether the continuous distribution $F(x)$ is subject to a known distribution function $S_n(x)$. In short, shows as follows:

$$D_n = \max_{-\infty<x<+\infty} \left| S_n(x) - F(x) \right|$$

Where 95% confidence interval, when the sample number n is greater than 35, if the value is less than $1.36/\sqrt{n}$, then that hypothesis H0: $S_n(x) = F(x)$ was established. Therefore, when n = 583 samples was randomly selected from the experimental data, then D_n = 0.012639; take $\alpha = 0.05$, get $D_{n,\alpha} = 0.05633$. for $D_n < D_{n,a}$, accept the hypothesis. And the six order polynomial shown in **Table 1** can be an accurate representation of relationship curves between SU and the Pearson correlation coefficient. Therefore, critical threshold of related level is divided by the Pearson correlation coefficient, we will let it substitute polynomials, and based on assumption 1 and assumption 2, we draw the following conclusions:

Conclusion 1. Division range of the Symmetrical uncertainty on the variable degree of correlation

$$\begin{cases} 0.0 \le SU(X;Y) \le 0.55 & low \\ 0.55 < SU(X;Y) < 0.90 & \text{mi}\,ddle \\ 0.90 \le SU(X;Y) \le 1.0 & high \end{cases}$$

4 Experiment

Section 3 introduced the SU method to evaluate correlation between two random variables, and enable to extend it to weigh the correlation between random vectors of any dimension, and fitting out threshold level of the experimental correlation adopting the SU method. This section will adopt the statistical and theoretical analysis on the multiple actual trace to test and verify accuracy of SU method; meanwhile we do further study on commonly the present study of flow metrics correlation. Compared to previous studies, this section has the following two differences:

(1) According to the study of network flow behavior at home and abroad, the analysis was extended to 14 flow behavior metrics, and do detailed study on the correlation;

(2) Refining metric, consider the basic metric for the analysis.

4.1 Experiment Data And Flow Metric

To avoid the accidental result of a single experimental data, and the accuracy of each method presented in this paper will be correctly reflected, rationality, this paper selected a series of experiments trace collected in a variety of different times to do analysis and testing, and to point out the current network flow metric correlation. Experiments Trace consists of the following two groups of data, the same groups of trace data are collected from the same location, and the main difference is that the group collected at different times:

(1) CERNET East (North) Regional Network Center collected trace;

(2) National Laboratory United States Internet Research (NLANR) publicly available traces [22];
Trace overview of the specific distribution shown in **Table 2**.

Table 2. Experiment Trace

Trace	Start time	During time	Bandwidth	bps(M)	pps(K)
CERNET_a	Nov. 10, 2005 14:00	1 hour	1G*2*3	3120	640
CERNET_b	Aug. 20, 2008 15:00	1 hour	1G*2	338.0	69.2
CERNET_c	Aug. 20, 2008 16:00	1 hour	1G*2	380.1	77.9
NLANR_a	Aug. 14, 2002 09:00	2 hours	2.5Gbps*2	846.0	138.1
NLANR_b	Jun. 01, 2004 19:31	50 min	10Gbps*2	1622.1	220.3

The analysis of network 14 flow metric feature includes the following as shown in **Table 3**.

Table 3. Metric feature

NO	metric	Metric discrible
1	l_port	low port number
2	h_port	high port number
3	pkts	Packets of flow
4	bytes	Bytes of flow
5	pkt_size	the average packet length
6	head_size	the average packet header length
7	payload_size	the average packet payload length
8	IAT	the average packet arrival interval of flow
9	duration	flow duration
10	pps	packets number per second
11	Bps	Byte number per second
12	pkts_ratio	packets number ratio of bidirectional flow
13	bytes_ratio	byte number ratio of bidirectional flow
14	pkt_size_ratio	packet length ratio of bidirectional flow

4.2 SU Method Accuracy Analyses

Experiment makes use of formula F1 and F3 and obtains Pearson correlation coefficient and SU values in every two metrics, shown in **Table 4**. values of the first line of each entry represent corresponding Pearson correlation coefficient of two-flow metric, the value of the second line represent its SU value; each row represent the four results values of trace CERNET_a / CERNET_b / NLANR_a / NLANR_b, Calculation result of Trace CERNET_c is similar to CERNET_b,so it not in table list .the result will be multiplied by 100

Table 4. Flow measurement between the two measurement conditions (×100)

	h_port	*pkts*	*bytes*	*pkt_size*	*head_size*	*payload_size*	*IAT*
l_port	32/26/20/23	5/3/1/1 4/12/20/18	7/2/3/2 21/26/24/2	4/10/3/1 21/22/24/2	2/1/5/7 7/10/15/19	4/10/3/2 20/22/30/3	7/9/3/1 22/23/28/2

	29/26/23/27		3	5		2	7
h_port		3/1/1/1 5/16/10/12	6/1/2/1 26/31/38/37	5/8/9/8 25/32/27/28	3/2/6/5 8/12/16/18	5/8/7/8 23/32/25/26	7/1/9/8 33/19/50/49
pkts			71/96/69/65 33/43/50/48	5/6/3/4 18/24/45/47	2/14/1/1 53/54/60/67	5/6/3/4 19/24/47/48	10/1/2/1 34/38/25/19
bytes				84/77/78/81 90/80/90/91	6/14/3/1 35/38/60/58	84/76/79/81 86/80/84/89	11/1/1/1 35/33/41/49
pkt_size					26/44/23/26 39/21/61/64	99/99/98/98 95/94/90/89	11/1/7/8 28/15/48/49
head_size						27/40/25/28 18/22/57/56	5/2/8/10 17/31/26/25
payload_size							11/1/7/9 28/15/46/46

	duration	*Bps*	*pps*	*pkts_ratio*	*bytes_ratio*	*pkt_size_ratio*
l_port	8/11/1/1 29/23/27/29	4/3/2/2 39/25/34/36	2/3/1/1 30/20/26/24	1/1/1/1 12/10/8/9	1/1/1/1 15/19/10/10	5/6/4/2 16/18/8/7
h_port	4/3/3/3 27/22/40/40	2/12/2/2 36/40/50/48	2/12/2/1 29/18/55/53	2/1/3/2 10/14/8/9	1/1/4/4 27/35/15/17	2/4/5/6 27/33/20/21
pkts	67/57/65/62 40/44/27/23	6/3/2/1 43/32/26/27	1/3/2/1 35/38/25/20	5/3/2/4 25/20/15/18	4/2/1/2 31/29/12/16	9/8/7/7 30/24/20/19
bytes	68/52/72/70 37/38/45/44	6/3/8/10 49/63/51/56	4/3/2/1 38/35/45/46	1/0/1/2 30/35/28/30	2/2/2/3 49/64/33/37	5/8/10/13 51/60/35/36
pkt_size	8/10/8/7 29/18/45/44	13/44/23/21 41/41/57/55	13/44/11/11 30/18/44/46	4/3/2/1 23/20/18/17	22/18/9/11 41/44/30/28	78/85/69/70 47/42/36/37
head_size	3/18/2/1 11/20/29/28	11/38/4/2 18/27/39/35	12/38/34/4/1 11/32/28/27	6/9/3/4 39/52/30/27	10/15/8/6 27/24/19/20	50/66/37/33 27/20/18/16
payload_size	8/8/7/7 28/18/43/41	13/43/23/21 40/40/55/52	13/43/12/11 28/18/43/44	6/4/2/1 25/20/22/20	21/18/10/11 42/44/29/30	79/84/70/71 45/41/35/37
IAT	49/18/31/34 33/52/74/75	18/27/25/3/4 93/89/90/87	2/27/1/1 98/98/100/99	2/1/1/1 21/27/25/24	2/1/2/1 21/19/23/20	8/6/6/4 18/26/20/19
duration		7/12/3/1 89/81/75/76	1/12/2/1 86/45/76/72	1/2/5/7/ 19/18/25/26	2/2/6/7 25/24/29/31	5/8/12/10 22/20/27/26
Bps			84/97/82/77 91/94/88/84	5/6/2/1 24/25/17/19	6/6/2/2 37/44/23/21	28/30/25/23 35/41/21/20
pps				4/6/2/2 26/28/20/25	5/6/3/2 23/22/21/20	26/30/25/21 19/20/18/16
pkts_ratio					49/45/50/51 32/25/35/36	3/1/4/6 25/18/29/32
bytes_ratio						25/21/33/29 67/61/75/72

Table 4 shows that calculation result of Pearson correlation coefficient and SU for various

metric does not show exactly the same, and their results in the correlation degree range are not the same, such as: Pearson (pkts, bytes) = 0.7 ~ 0.96, SU (pkts, bytes) = 0.33 ~ 0.5, Pearson> SU; Pearson (duration, Bps) ≈ 0.1, SU (duration, Bps) ≈ 0.8, Pearson <SU. This inconsistency, this section analyze defects which are mentioned by section 2.2 for the Pearson correlation coefficient, indicating that its representation ability of SU method improve accuracy;

Through the above discussion on Pearson correlation coefficients and SU we can draw the following conclusions:

(1) If SU coefficient and P values of the two metrics both are great in **Table 4**.it showed a linear correlation. Generally speaking, packet header has a fixed length, so the metric pkt_size and payload_size have a strong linear correlation, as shown in Figure 4. Calculation results of Pearson correlation coefficient and the SU respectively were 0.98 and 0.93, from the conclusion 1 we can see, which can accurately represent strong linear correlation between metric pkt_size and payload_size. Therefore, if the Pearson correlation coefficient can accurately reflect correlation between the metrics, SU can achieve the same effect.

(2) If the coefficient P and SU values of the two measures both are small, the correlation between two measures is weak. **Fig. 5** shows relation points chart of metric l_port and bytes for the trace CERNET_b, from **Table 4** we can see, Pearson correlation coefficient and SU of two metrics were approximately 0.05 and 0.2,both are low correlation in the range. Therefore, itself only with the metric pairs of low correlation, which of Pearson correlation coefficient and SU values are close to 0, the results show, two methods both can accurately weigh.

(3) As shown in **Table 4** the correlation between metric IAT and pps is calculated, Pearson (IAT, pps) = 0.01~0.27, SU (IAT; pps) = 0.98~1.0, we can see the papers mentioned in Section 2.2 that Pearson correlation coefficient has the defects of not accurately reflect non-linear high correlation, while SU can be accurately weighed.

(4) For the defect of P factor extreme values, and metric pkts and during exist heavy-tailed distribution, very few metric observations (2%) make the overall sample of unrelated P reach a high correlation coefficient of approximation (Pearson = 0.55~0.67). However, the calculated value of the SU is only about 0.4 +. From the conclusion 1 we can see, through the SU method calculated both the overall measure pkts and duration is a low degree of correlation. Therefore, SU can reflect the actual overall correlation better than the Pearson correlation coefficient. The Pearson correlation coefficient will deviate from the exact value of the case for extreme values, while SU results can be corrected.

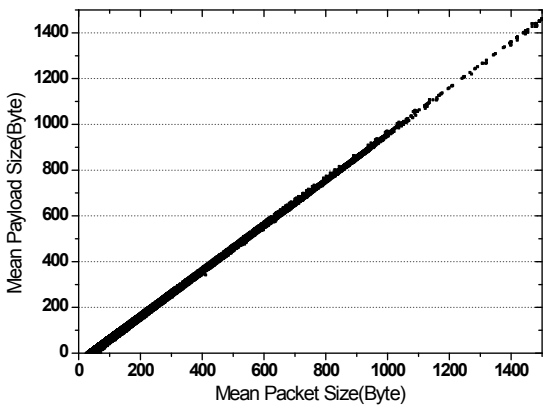


Fig. 4. Distribution of trace CERNET_b mearn packet size and payload size

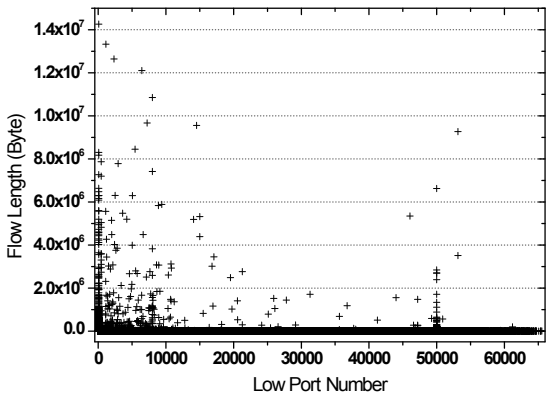


Fig. 5. Distribution of trace CERNET_b low port number and flow length

Analysis results show that, the Pearson correlation coefficient can accurately weigh the correlation between the flow metrics, it is the same as SU; the Pearson correlation coefficient has the essential deficiency which can not accurately weigh the correlation, SU can overcome its shortcomings, which achieves a comprehensive and accurate purposes to weigh correlation between network flow metrics, moreover, it can provide ways and means for further analysis of the network traffic behaviors.

4.3 Measure Correlation Analysis between Multi-Flows

Above two experiments focused on analysis of the correlation between the flow metrics. For the correlation of multiple metric, because there are too many possible combinations, this section only lists the SU method results of three measures for an example. Since the Pearson correlation coefficient cannot describe the correlation between multi-metrics, **Table 5** only lists some metrics' values of Trace CERNET_a / CERNET_b / NLANR_a / NLANR_b and their respective correlation degree.

Table 5. Relationship of flow measure (×100)

	l_port	*pkt_size*	*duration*	*pps*
pkts & *IAT*	26/21/27/28 low	65/59/70/72 middle	73/75/76/77 middle	98/99/98/100 high
bytes & *IAT*	25/31/21/20 low	90/81/93/93 middle	73/59/76/78 middle	99/98/98/100 high
IAT & *Bps*	39/25/32/36 low	42/60/55/57 middle	89/83/79/80 middle	99/100/99/100 high
pkts & *head_size*	20/24/25/23 low	55/43/60/64 middle	40/49/44/39 middle	35/37/41/39 low
pkts & *bytes*	24/29/26/25 low	94/96/96/96 high	41/47/45/45 low	42/45/51/52 low
pkts & *Bps*	40/32/39/40 low	52/57/69/65 middle	89/87/77/80 middle	91/97/95/91 high
bytes & *Bps*	40/30/35/39 low	91/81/92/94 high	89/89/78/81 middle	91/96/96/92 high

First, through the comparison of the three metrics in **Table 5** and the analysis of SU values correlation between the two metrics in **Table 4**, which can verify the validity of Theorem 2: If

X, Y, Z are three random variables, the SU (X;YZ)≥SU (X;Y) and SU (X; YZ) ≥SU (X;Z), the co-correlation degree between X and Y, Z is greater than X with a single Y or Z. On the other hand, multiple correlation degree between the metrics can also be divided into highly correlation, moderate correlation and low correlation which is same to correlation between the two metrics. For example pkt_size = bytes / pkts, SU (pkt_size; pkts, bytes) ≈ 0.96 indicates when metric pkts and bytes are known, metric information of pkt_size is redundancy. For example bytes & Bps and pkt_size have highly correlation, because metric pkt_size and *bytes* have a strong correlation, therefore the three metrics is more closely. Low port numbers and the other metrics still have only low correlation degree, and further indicate that the port number protocol is independence. Meanwhile, as shown in **Table 5**, respectively the metric correlation between pkt_size ,IAT and pkts are weak (the SU values of the trace reach between 0.15 to 0.49), while **Table 5** shows correlation degree between the joint information of IAT and pkts with pkt_size has strengthened, SU (pkt_size; IAT, pkts) reached between 0.6 to 0.72. This shows when flow rate and flow length are known, the average flow length can have higher statistical probability to be obtained through previous experience estimates.

4.4 Summary

In summary, we believe that our work still provides some useful insights and a first step toward understanding the relationship between different characterizations of Internet flows.Nowadays because feature selection method mainly is based on the correlation between metrics. so the use of the SU expansion of the definition 1 can not only accurately weigh the correlation between multiple metrics, and it is also a complement to weigh the correlation between the two metrics. The method can be applied to delete high correlation between flow metrics and beneficial to reduce redundancy. So traffic identification can be optimized.

5. Conclusion

This article firstly analyzed network traffic based on the characteristics of random variables and pointed out defects using the Pearson correlation coefficient to weigh the current network traffic, make use of information theory, mutual information and symmetrical uncertainty SU as a standard to weigh metric correlation between two flows and extend the method to weigh correlation between two any dimensional metric vectors. Theoretical analysis and experimental results show that: SU not only can accurately represent the Pearson correlation coefficient to reflect the linear correlation between the flow metrics and make up for the deficiency of Pearson correlation coefficient, and can further analyze the problem that Pearson correlation coefficient cannot handle more accurate metric correlation and improve the accuracy and versatility of network flow analysis methods on metric correlation.

At the same time, this article using polynomial fitting, based on accuracy of Pearson correlation coefficient method to weigh non-linear correlation between the extreme values, summed up threshold of the SU approach to partition the variable high / moderate / low correlation degree.

Further, in this paper, study and analysis on flow metric distribution and correlation by SU method, the following a few general conclusions, more detailed results, as shown in **Table 4** and **Table 5**.

(1)There is a high correlation between metrics in the each type of metrics. Such as flow length type metric:pkts and bytes, the average packet length of flow type metric:pkt_size and

payload_size. flow rate type metric: Bps and pps have reached or nearly reached a high correlation.
(2)There exists a moderate correlation between metrics of the average packet length and flow length type, the bidirectional flow throughput metric (such as bytes ratio).
(3)Flow duration and flow rate type metric have the moderate correlation.
(4)As when the number of packets in the flow is less, flow duration and flow length type metrics distribution is broader, so the whole only have a low correlation; but with the observations increases, the correlation degree gradually increased, when larger is metric values can be reached strong correlation degree.

References

[1] Kun-chan Lan, and John Heidemann, "A measurement study of correlations of Internet flow characteristics," *Computer Networks*, vol.50, no.1, pp.46-62, Jan.2006. Article (CrossRef Link)

[2] Zhou Mingzhong. Study of Large-scale network IP flows behavior characteristics and measurement algorithms. Jiangsu: Southeast University, 2006 Nanjing. Article (CrossRef Link)

[3] Gregor Maier, Anja Feldmann, Vern Paxson, and Mark Allman. "On dominant characteristics of residential broadband internet traffic," in *Proc. of the 9th ACM SIGCOMM conference on Internet measurement conference* ACM, pp.90-102,2009.

[4] Felix Hernandez Campos, J. S. Marron, Sidney I. Resnick, and Kevin Jeffay. "Extremal dependence: Internet Traffic Applications," *Stochastic Models*, vol.21, no.1, pp.1-35, 2005. Article (CrossRef Link)

[5] Cheolwoo Park, Felix Hernandez-Campos, J. S. Marron, Kevin Jeffay, and F. Donelson Smith. "Analysis of dependence among size, rate, and duration in internet flows," *Annals of Applied Statistics*. 2010. Article (CrossRef Link)

[6] C.Dewes, A.Wichmann, A.Feldmann. "An analysis of internet chat systems," in *Proc. of ACM SIGCOMM* , pp.51-64*, 2003.* Article (CrossRef Link)

[7] S. Saroiu, P. K. Gummadi, and S. D. Gribble. "A measurement study of peer-to-peer file sharing systems," in *Proc. of Multimedia Computing and Networking 2002*, pp.156-170, 2002. Article (CrossRef Link)

[8] Kurt Tutschku. "A measurement-based traffic profile of the edonkey file-sharing service," in *Proc. of the 5th annual Passive and Active Measurement Workshop,* pp.12-21, 2004. Article (CrossRef Link)

[9] Louis Plissonneau, Jean-Laurent Costeux, Patrick Brown. "Analysis of Peer-to-Peer Traffic on ADSL," In: *Proc. of the 6th annual Passive and Active Measurements Workshop (PAM'05)*. Boston, USA, pp.69-82, 2005. Article (CrossRef Link)

[10] Weinstein, Eric W. Correlation Coefficient.http://mathworld.wolfram.com/CorrelationCoefficient.html. 2005-5-25/2006-4-15. Article (CrossRef Link)

[11] StatSoft, Inc. Basic Statistics. http://www.statsoft.com/textbook/basic-statistics/. 2009. Article (CrossRef Link)

[12] M A Hall. "Correlation-based feature selection for discrete and numeric class machine learning," in *Proc. of the 17th International Conference on Machine Learning,* pp.359–366, 2000. Article (CrossRef Link)

[13] Mark A. Hall. Correlation-based Feature Selection for Machine Learning. New Zealand: The University of Waikato. 1999. Article (CrossRef Link)

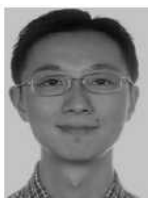
[14] P. Mitra, C. A. Murthy, and S. K. Pal. "Unsupervised Feature Selection Using Feature Similarity," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.24, no.3, pp.301–312,2002. Article (CrossRef Link)

[15] Jing Yuan, Zhu Li, and Ruixi Yuan. "Information entropy based clustering method for unsupervised internet traffic classification", in *Proc. of IEEE International Conference on Communications*, pp.1588-1592, 2008. Article (CrossRef Link)

[16] D. A. Bell and H. Wang "A formalism for relevance and its application in feature subset selection," *Machine Learning*, vol.41, no.2, pp.175–195, 2000. Article (CrossRef Link)
[17] Lei Yu, and Huan Liu. "efficient feature selection via analysis of relevance and redundancy," *Journal of Machine Learning Research*, vol.5, no.2, pp.1205–1224, 2004. Article (CrossRef Link)
[18] Qu, G., Hariri, S., and Yousif, M. "A new dependency and correlation analysis for features", *IEEE Transactions on Knowledge and Data Engineering,* vol.17, no.9, pp.1199-1207, Sep.2005. Article (CrossRef Link)
[19] T Ganchev, P Zervas, N Fakotakis, and G Kokkinakis. "Benchmarking feature selection techniques on the speaker verification task," in *Proc. of CSNDSP'06*, pp.314-318, 2006. Article (CrossRef Link)
[20] Qu Wei, Zhu Shibing et al. *the information theory and applications*, Beijing: Tsinghua University press. 2005.Article (CrossRef Link)
[21] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. *Numerical Recipes in C*, London: Cambridge University Press, 1988. Article (CrossRef Link)
[22] CAIDA. NLANR PMA. http://pma.nlanr.net. 2002-09-11/2005-04. Article (CrossRef Link).
[23] Ye Cinan, and Cao Weili. *Mathematical statistics with applications*, Mechanical industry press. 2004. Article (CrossRef Link).
[24] Shuang Hong Yang, and Bao-Gang Hu, "discriminative feature selection by nonparametric bayes error minimization," *IEEE Transactions on Knowledge and Data Engineering*, Apr.2011. Article (CrossRef Link).
[25] Huawen Liu, Jigui Sun, Lei Liu, and Huijie Zhang, "Feature selection with dynamic mutual information," *Pattern Recognition*, vol.42, no.7, pp.1330-1339, Jul.2009. Article (CrossRef Link).

Shi Dong is Ph.D. candidate in school of computer science and engineering at Southeast University, His major research interests include network security, network management and network measurement.

Wei Ding is a Professor of Southeast University, doctoral supervisor. Her major research interests include network management and network measurement



Liang Chen received the Ph.D. degree in school of computer science and engineering from Southeast University in 2010. His major research interests include network management and network measurement