Entropy Based P2P Flow Detection
Ji-yan SHI¹, Zong-liang YANG², Yan LIU³, and Dong-ying LIU⁴,*

¹Logistics Service Center of Yunnan Power Grid Co., LTD, Kunming
²Logistics Service Center of Yunnan Power Grid Co., LTD, Kunming
³Logistics Service Center of Yunnan Power Grid Co., LTD, Kunming
⁴Project Research and Development Department, Kunming Nengxun Technology Co., LTD, Kunming.

*Corresponding author

Keywords: P2P traffic, information entropy, P2P flow

Abstract. As an efficient distribution mechanism, peer-to-peer technology has providing users with cheap and powerful communicating facilities. With the growth of P2P applications, Statistics suggest that P2P flow accounts for a significant fraction of the Internet traffic. Existing approaches to identify P2P traffic have well known drawbacks. In this paper, a novel approach to identify P2P traffic based on information entropy theory has been proposed. We investigate the problem of estimating the entropy in a P2P streaming computation model and find that using entropy can aid P2P stream monitoring. Experimental results show that the characteristic entropy changes can be used to construct a threshold based detector for P2P stream applications.

Introduction
In both academia and industry, Peer-to-peer (P2P) applications have recently received much attention with successful commercial systems showing its viability in the internet. With the growth of P2P applications, Statistics suggest that P2P flow accounts for a significant fraction of the Internet traffic[1]. ISP and enterprise network administrators would like to be able to effectively identify P2P traffic for network planning and management.

There has been a great deal of researches in the area of P2P track detection[2], [3]. However, existing approaches to identify P2P flow have well known drawbacks. Port based analysis is the most basic and straightforward method to detect P2P users in network traffic. It is based on the simple concept that many P2P applications have default ports on which they function. Traced service port numbers in TCP/UDP packet headers are compared with known port numbers of P2P applications. Nevertheless such technique becomes no longer effective since new version P2P applications use dynamic port numbers. Application payload method can achieve higher accuracy. It assumes that P2P applications have a unique signature located in the data portion of the packet. But constraints like hardware resource limitation, payload encryption and privacy make these methods in effective. With dynamic behaviors, P2P applications’ precise characteristics are usually unknown beforehand. This makes direct matching P2P flow difficultly. Hence, It is desirable to derive new parameters that describe P2P traffic flow in such a way that details are hidden, but the type of changes associated P2P behaviors are visible. This method can be able to effectively differentiates P2P flow from other applications.

Related Work
Recently information entropy theory has been used in worm detection and adaptive flow aggregation, and shows its merit[4,5]. As far as we know that no previous working focuses on P2P stream detection. In this letter, by estimating the P2P stream identification in a computation entropy model, we show that the variances of entropy changes can be used to construct a threshold based detector for P2P stream
Problem Formulation

In information theory, entropy is a measure of the uncertainty associated with a random variable. Consider a random variable X that may take $N_X$ discrete values. If we observe a random variable X which may take $N_X$ discrete values for m times, then the empirical probability distribution on X is $p(x_i) = m_i / m$, $x_i \in X$, where $m_i$ is the number of times that $x_i$ takes. The empirical entropy of X is then defined by Eq.1.

$$H(X) = -\sum_{x_i \in X} p(x_i) \log p(x_i)$$

(1)

Define Maximum Entropy as $H_{\text{max}}(X)$, it is clear that $0 \leq H(X) \leq H_{\text{max}}(X)$. Then we have normal entropy as

$$H(X) = \frac{\log m}{H_{\text{max}}(X)}.$$

If all probabilities of the observations are equal which means variable X have the highest degree uncertainty, $H(X) = 1$. In another extreme situation where observations of variable X are same, $H(X) = 0$.

Given a stream S that contains m kinds of packets with sizes (number) $a_1$, $a_2$, ..., $a_m$ and the totally number of packets in S is $N = \sum_{i=1}^{m} a_i$. Respectively, its empirical entropy $H(S)$ is defined as

$$H(N) = -\sum_{i=1}^{m} \frac{a_i}{N} \log \frac{a_i}{N}.$$

To compute the entropy,

$$H(N) = -\sum_{i=1}^{m} \frac{a_i}{N} \log \frac{a_i}{N} = -\frac{1}{N} \left[ \sum_{i=1}^{m} a_i \log a_i - \sum_{i=1}^{m} a_i \log N \right]$$

(2)

$$= \log N - \frac{1}{N} \sum_{i=1}^{m} a_i \log a_i.$$

Intuitively, the $H(N)$ is a measure of the diversity of the data set over the stream. The entropy attains its minimum value of zero when all the items coming over the stream are the same and its
maximum value when all the items in the stream are distinct. So for the equation (2),

\[
\text{Min } (H(N)) = 0, \text{ when } \alpha_1 = \alpha_2 = \ldots = \alpha_N.
\]

Max \( (H(N)) = \log N - \frac{2}{N} \sum_{i=1}^{N} \log 1 = \log N \), when all the packets in the stream S are distinct.

Denote by \( H(N) \) the absolute entropy differential value between two different dataset divisions over the same stream S. We have

\[
H(N) = |H_i(N_i) - H_j(N_j)| = \left| - \sum_{t=1}^{N_i} \frac{\alpha_t}{N_i} \log \frac{\alpha_t}{N_i} + \sum_{t=1}^{N_j} \frac{\alpha_t}{N_j} \log \frac{\alpha_t}{N_j} \right|
\]

\[
= \frac{1}{N} \left| \sum_{t=1}^{N_i} \alpha_t \log \alpha_t - \sum_{t=1}^{N_j} \alpha_t \log \alpha_t \right|
\]

Suppose there are m kinds of packets in the stream S with totally N numbers.

**Experimental Findings**

Our dataset use flow captured in June 2009 at Yunnan University. The dataset consist of traffic to and from thousands of active IP addresses involving several P2P applications. We develop a software which can automatically trace the traffic packets developed with c++. The time slot of each trace file varies from 10 minutes to 1 hour. Trace records are defined by the 5-tuple of source/destination pairs for the IP address, port, packet count, and byte count. It also includes the connection time, protocol used, connection state, and flow direction. After pre-processing of the traces, a srcIP-dstIP pair analysis was conducted on the resulting bi-directional traces. Since our goal is to profile P2P traffic by entropy mining of P2P communication pattern, we consider the essential five-dimensional feature space consisting of srcIP, dstIP, and dstPrt, packet count, and byte count.

![Figure 1. IP entropy of different applications.](image-url)
Fig. 1 plots the IP entropy of different P2P applications in a node with the separating packet size N=1000. We can observe that the IP entropy of Non-P2P stream will significantly lower than other P2P applications. In some extremely states, IP entropy of Non-P2P application can reach zero, which means all the 1000 packets will come from the same IP address, while such results are never repeated in P2P applications.

![Correlation coefficient](image)

Figure 2. Different entropy-based statistical Correlation.

Linear regression is a well-known statistical approach to model the relationship between two variables by fitting a linear equation to observed dataset. It can be expressed as $Y = bX + a$.

$X$ is considered to be an explanatory variable and $Y$ is a dependent variable. In order to numerical measure the association between $X$ and $Y$ we consider correlation coefficient $(r)$ where $|r| \leq 1$ and $r$ indicate the strength of the association. It can be calculated by

$$r = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \sum_{i=1}^{n} (y_i - \bar{y})^2}}$$

We use linear regression to model the statistical relationship between IP-based entropy and Length-based entropy of different P2P applications. Correlation coefficient $(r)$ indicates the strength of the association. In Fig2 the average $r$ of P2P applications will larger than 0.5 and some cases can even reach 0.9. Non-P2P application is less than 0.4.

![Comparison of ΔH(N) and AP2](image)

Figure 3. Comparing $\Delta H(N)$ and APZ.
Fig 3 plots the normalized IP-based $\Delta H(N)$ and APZ( the average packets length of P2P). We can see that combined with the APZ dramatic shift, which means that P2P applications transfer its behaviors, the Fig. 3 tell us they are different.

The above results indicate P2P applications have different flow characteristics in contrast to Non-P2P applications with higher IP-based entropy and higher correlation coefficient of different entropy. In addition there lative change $\Delta H(N)$ can be use to analyze the P2P behaviors pattern. Thus by exploiting the entropy variances in the flow we can construct a threshold based detector for P2P stream applications.

Conclusions

Existing approaches to identify P2P traffic have well known drawbacks due to its dynamic behaviors. In this letter, we have proposed a novel entropy-based method to identify P2P traffic. Experimental results on multiple flow traces demonstrate that our new method is a promising approach.

References


