Long Range Mutual Information*

Nahur Fonseca
Boston University
111 Cummington St
Boston, Massachusetts, USA
nahur@cs.bu.edu

Mark Crovella
Boston University
111 Cummington St
Boston, Massachusetts, USA
crovella@cs.bu.edu

Kavé Salamatian
LIP6
P.O. Box 1212
Paris, France
kave@lip6.org

ABSTRACT

Network traffic modeling generally views traffic as a superposition of flows that creates a timeseries of volume counts (e.g. of bytes or packets). What is omitted from this view of traffic is the contents of packets. Packet contents (e.g. header fields) contain considerable information that can be useful in many applications such as change and anomaly detection, and router performance evaluation. The goal of this paper is to draw attention to the problem of modeling traffic with respect to the contents of packets. In this regard, we identify a new phenomenon: *long range mutual information* (LRMI), which means that the dependence of the contents of a pair of packets decays as a power of the lag between them. We demonstrate that although LRMI is hard to measure, and hard to model using the mathematical tools at hand, its effects are easy to identify in real traffic, and it may have a considerable impact on a number of applications. We believe that work in modeling this phenomenon will open doors to new kinds of traffic models, and new advances in a number of applications.

1. INTRODUCTION

Considerable effort has gone into characterizing and modeling network traffic. The vast majority of traffic modeling has been concerned with measures of traffic volume: the number of packets or bytes passing over a link. The resulting models of traffic volume are valuable for a variety of tasks, most notably performance evaluation of network elements. In the process of developing traffic volume models a number of important properties have been identified, including the observation that traffic volume measures generally show long-range dependence [6].

However, measures of traffic volume are not the only important property of traffic. A more detailed view of traffic might consider the information inside of packets, such as the values present in packet header fields. When thinking about the sequence of packet headers seen on a link, a natural abstraction is to treat them as a sequence of symbols drawn from a particular alphabet. That is, at a very basic level information carried in each packet is interpretable simply as a symbol.

This suggests treating a packet sequence as being emitted by a source that can be characterized by the sequence of symbols it generates. Models that capture the properties of packet header values would be useful for evaluating network elements that operate on headers, such as routers. Such models could also find use in characterizing "normal" packet header properties and thereby forming the basis of new traffic anomaly detection methods.

The goal of this paper is to draw attention to traffic modeling based on symbolic view of traffic. Although this is in some sense a natural view of traffic, relatively little work has been done to characterize traffic at this level. In this paper we help to fill this gap by studying the properties of network traffic when treated as a symbol sequence. In particular we study the sequence of headers found in traffic flowing over a single link.

The definition of what constitutes a symbol must be carefully chosen, in accord with the purpose of the model. It is desirable that the definition results in a manageable number of symbols, and yet be linked to the object under study. In this paper, we focus on a symbol definition which is derived from the packet header fields and is constant within a flow.

In this context we find and characterize a new phenomenon which we call *long range mutual information*. Informally, this property states that the dependence of symbols of two packets does not drop off sharply with the lag between them. The most important reason for packet symbols to show dependence is the fact that the sequence of symbols is the interleaving of symbols from a set of flows, and that symbols from the same flow can be the same from packet to packet.

Our first contribution is to show that that direct measurement of long range mutual information is quite difficult in typical traffic, because it requires a very large number of samples.

Secondly, we show the strong effects that LRMI has on the distribution of symbols seen in a collection of consecutive packets. This occurs because, although the mutual information between any two packets is quite small, the slow decline of mutual information over time means that its aggregate effect over a window of many packets is quite large.

In order to assess the effects of LRMI on the distribution of symbols seen in a window, we apply transformations

*This work was supported by NSF grants CCR-0325701 and ANI-0322990.

on the sequence of symbols and compare the resulting histograms. We show that long range mutual information has implications for, among other things, statistical anomaly detection methods based on the application of Sanov's Theorem to the distribution of header field values seen within a window of packets. Ideally we would like to obtain analytical relationships between the strength of long range mutual information, and the large deviation properties of the sampled histograms. However, as we explain in the body of the paper, current mathematical tools are insufficient for this purpose.

The remainder of the paper is organized as follows. In Section 2 we introduce LRMI, by means of an analytical model, and demonstrate the difficulty of directly measuring LRMI. In Section 3 we show evidence of the presence of LRMI in network traffic, and how LRMI affects the distribution of symbols in a trace. In Section 4 we place our results in the context of prior work. Finally in Section 5 we conclude and present directions for future work.

2. A NEW KIND OF LONG MEMORY

Turning to a symbolic view of network traffic requires the selection of a mapping function that transforms each packet to a symbol. We deffer the discussion of different mappings for the future, and assume for the purposes of this section that such function exists and maps each packet to a symbol from an alphabet $\mathscr{X}$.

In this context, we are interested in characterizing the memory structure between packets at different lags. A natural way of quantifying this dependence is by measuring the mutual information. We first define what mutual information is, and extend this definition to network traffic. Then, we develop an analytical model of mutual information and use it to show how hard it is to directly measure mutual information in real traces.

2.1 Mutual Information of Random Variables

The amount of information contained in a random variable X, with distribution $p(x)$ can be measured by the Shannon Entropy, defined as

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

Similarly, the joint entropy measures the information of a pair of random variables X and $Y \sim p(x,y)$, and is given by

$$H(X,Y) = -\sum_{x\in\mathscr{X},y\in\mathscr{Y}} p(x,y)\log p(x,y)$$

If the variable Y was totally dependent on X, then $p(x,y)$ would collapse to being simply $p(x)$, and $H(X,Y)$ would be equal to $H(X)$. On the other hand, if X and Y were totally independent, $p(x,y)$, would be equal to the product of the marginals for X and Y, $p(x)p(y)$, and $H(X,Y)$ would be equal to $H(X)+H(Y)$[1].

The mutual information of X and Y captures the dependence between X and Y with a number in the range $[0, H(X)]$ (from independence to total dependence). Mutual information is symmetric, and is defined as

$$I(X;Y) \quad = \sum_{x\in\mathscr{X},\ y\in\mathscr{Y}} p(x,y)\log\frac{p(x,y)}{p(x)p(y)}$$

2.2 Auto Mutual Information

One can define auto mutual information at lag l for a stationary discrete-valued stochastic process $X_1, X_2, \ldots, X_n$ as the mutual information between random variables X_i and X_{i+l}.

$$I(X_i; X_{i+l}) = \sum_{x_i\in\mathscr{X}} \sum_{x_{i+l}\in\mathscr{X}} P(x_i, x_{i+l}) \log\frac{P(x_i,x_{i+l})}{P(x_i)P(x_{i+l})}$$

Since the process is stationary, $I(X_i; X_{i+l})$ is independent of i and so we can simply refer to the mutual information at lag l, as $I(l)$.

At the heart of the calculation of auto mutual information, lies the estimation of the joint distribution of symbols at different lags. We first develop an analytical model to compute mutual information and then show evidence of how hard it is to measure it directly in real traces.

2.3 A Simple Model of Symbolic Traffic

How does mutual information vary as a function of l? For a n^{th}-order Markovian process, where each symbol depends only on the n previous ones, we expect the mutual information to decline exponentially fast as a function of l when $l > n$. However, this is not necessarily the case for network traffic.

To analyze the relationship between $I(l)$ and l we model network traffic as follows. We consider a model where m concurrent flows generate symbols over an alphabet $\mathscr{X}$. The symbol sequence emitted is the (random) interleaving of the symbols associated with the packets of each flow.

There are a total of $K^m \times \mathscr{X}^m$ states, where K is the maximum flow size allowed. Each state has associated with it a tuple $(\mathbf{R}, \mathbf{S})$, where

$\mathbf{R} = (R_1, \ldots, R_i, \ldots, R_m)$, $R_i \in [1, K]$ is the number of remaining packets of flow i, K is the maximum flow size; and

$\mathbf{S} = (S_1, \ldots, S_i, \ldots, S_m)$, $S_i \in \mathscr{X}$ is the symbol assigned to flow i.

The transition probabilities are given by:

$$\begin{array}{lll} \{(R_1,\ldots,R_i,\ldots,R_m), & \rightarrow & \{(R_1,\ldots,R_i-1,\ldots,R_m), \\ (S_1,\ldots,S_i,\ldots,S_m)\}_n & & (S_1,\ldots,S_i,\ldots,S_m)\}_{n+1} \\ \text{if } R_i > 1 \quad \text{w.p.} & & (1/m) \end{array}$$

$$\begin{array}{lll} \{(R_1,\ldots,R_i,\ldots,R_m), & \rightarrow & \{(R_1,\ldots,k,\ldots,R_m), \\ (S_1,\ldots,S_i,\ldots,S_m)\}_n & & (S_1,\ldots,s,\ldots,S_m)\}_{n+1} \\ \text{if } R_i = 1 \quad \text{w.p.} & & \kappa(k)\phi(s)(1/m) \end{array}$$

where $\kappa(k)$ is the probability that any given flow contains k packets and $\phi(s)$ probability that any given flow has symbol s. On each transition the process emits symbol S_i.

Given this model, we compute the mutual information between emitted symbols that are separated by lag l. A complete derivation of the formula of mutual information at lag l for this model is given in Appendix A.

The question we ask is how mutual information declines with lag. To answer this, we show results in Figure 1. The figure presents numerical results of mutual information as a function of lag for cases with $|\mathscr{X}| = \{16, 64\}$, and multiplexing level $m = \{200, 3000\}$.

We consider a variety of different distributions of flow lengths $\kappa(\cdot)$. The distributions corresponding to empirically

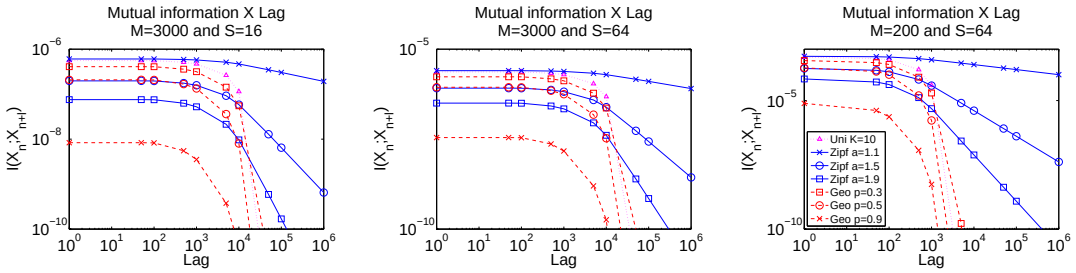

Figure 1: Numerical evaluation of Mutual Information, using Normal approximation of the Binomial

observed traffic are the heavy-tailed Zipf distributions (i.e., the discrete Pareto distributions). As mentioned in Section 4, there is considerable empirical evidence for this distribution as a good model of flow lengths. For comparison purposes we also show a range of Geometric distributions (discrete exponential) and a single Uniform distribution.

The figure shows that distributions like the light-tailed Geometric and Uniform show a sharp decline in mutual information past a given lag threshold. This threshold corresponds roughly to the product of the multiplexing level and the mean flow size.

In contrast, for the more realistic (heavy-tailed) flow length distributions, mutual information declines slowly and does not exhibit any particular threshold. This slow decline is analogous to the polynomial decline of auto-correlation in a long-range dependent process, and so we refer to it as *long range mutual information* (LRMI).

2.4 LRMI is not LRD

A stochastic process with the property of *long range dependence* can be characterized by an auto-correlation function that decays as a power of the lag. Therefore, even though the absolute values of the auto-correlation function may be small at large lags, their sum diverges. The consequences are important in many settings; for a study of the effects of LRD in queue performance, see [4].

Long range mutual information is also a property of a stochastic process with dependencies among random variables at large lags. However LRMI and LRD are not the same phenomenon.

LRMI does not imply LRD. Given a network traffic trace that has LRMI and LRD, it could be passed through a traffic shaper to remove LRD, but the sequence of symbols would be unchanged, and therefore, LRMI would be preserved.

LRD does not imply LRMI. Depending on the definition of symbol used, a network traffic trace with LRD may have no mutual information at all (for example, when using the check-sum field of IP headers); or it may be fully dependent (for example, when using the IP version field).

The presence of LRMI may have implications in many aspects that are open to investigation. In general, LRMI implies that a histogram of symbols measured over any window in a trace will show more variability then predicted by a Markov model.

2.5 Estimating LRMI in Real Traffic

To measure the presence of long range mutual information in real traffic, the straightforward approach would be to estimate the joint probability function of symbols at each lag l. However, in this section we argue that such an approach presents considerable difficulties.

The model used in the previous section provides a helpful framework for understanding the problem. In that model, two symbols are independent if they arise from different flows. This makes the explicit assumption that the effect of higher level structure (e.g. sessions) is negligible. Thus the joint probability distribution to be estimated is concerned with two values: the joint probability of two symbols that are different and the joint probability of two symbols that are the same.

In the former case, we are concerned with

$$P(x_i, x_{i+l} \,|\, x_i \neq x_{i+l})$$

and in the latter case we are concerned with

$$P(x_i, x_{i+l} \,|\, x_i = x_{i+l}).$$

Accurate estimation of the mutual information at lag l requires accurate estimation of both these quantities, and in particular, accurate estimation of the *difference* between these two quantities. This is because different symbols come from different flows and so are necessarily independent, but when two symbols are the same they may have come from the same flow and so are dependent. It is this latter dependence that induces mutual information.

Because of the assumption of independence of flows, the joint probability at lag l of two symbols which are different is given by the product of the symbol distribution $\phi(x)$ of the two symbols

$$P(x_i, x_{i+l} \,|\, x_i \neq x_{i+l}) = \phi(x_i)\phi(x_{i+l}) \qquad (1)$$

and the joint probability of two symbols at lag l that are the same depends on the probability $p(l)$ that the two symbols belong to the same flow

$$P(x_i, x_{i+l} \,|\, x_i = x_{i+l}) = \phi(x_i)p(l) + \phi(x_i)^2(1 - p(l)) \qquad (2)$$

$p(l)$ is given by the following equation, of which a derivation is provided in Appendix A

$$p(l) \;=\; \frac{1}{mE_\kappa} \sum_{\forall k} \kappa(k) \sum_{i=1}^{k-1} \Big(1 - \sum_{j=k-i}^{l-1} B_{l-1,\frac{1}{M}}(j)\Big)$$

	$M = 9000$		$M = 15000$	
l	1.5	1.9	1.5	1.9
2	$1.044e9$	$1.678e9$	$1.739e9$	$2.796e9$
8	$1.044e9$	$1.678e9$	$1.740e9$	$2.797e9$
32	$1.045e9$	$1.680e9$	$1.740e9$	$2.799e9$

Table 1: Minimum number of samples to estimate mutual information. Four settings shown: multiplexing $M = \{9000, 15000\}$, and Zipf parameter $\alpha = \{1.5, 1.9\}$.

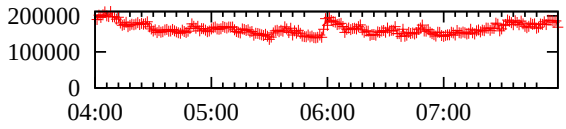

Figure 2: Timeseries of packet counts per minute. Timeseries of byte counts and multiplexing level also show similar stability characteristics.

In the case when $\phi(x) \sim$ Uniform, the difference between (2) and (1) is

$$p(l)(\phi(x_i) - \phi(x_i)^2)$$

which is in general quite small. Note that $p(l)$ is upper bounded by $1/m$. Thus accurate estimation of mutual information becomes more difficult at higher levels of traffic aggregation.

Compounding the difficulty is that there are $|\mathscr{X}|$ symbols for which this difference must be measured, and that some number k observations are needed for numerical accuracy. Thus we have an instance of the coupon collecting problem, and we expect that we need to observe a sequence of length approximately $k * |\mathscr{X}| \log(k * |\mathscr{X}|)/p(l)$ to witness k such events for each symbol.

In Table 1 we show typical values of the sequence lengths needed to accurately estimate mutual information for a range of lags, and multiplexing levels, and for alphabets of size 64. The figures show that the number of samples needed is quite large even for moderate multiplexing levels, and that the number of samples grows with increasing levels of multiplexing.

As a confirmation of the difficulty of measuring mutual information directly, we have explored the possibility of training both Markov models and Hidden Markov Models to capture the dependence structure in real traffic. These efforts have been ineffective due to the characteristics of the problem presented in this section: low mutual information at any given lag combined with the high order of the model needed.

3. THE IMPORTANCE OF LRMI

The preceding subsections would suggest that since mutual information in traffic is generally quite small, that it is insignificant. However, this is most definitely not the case.

In order to understand why, it is important to keep in mind that although mutual information is small, it declines very slowly as well.

We illustrate this phenomenon by examining the distribution of symbols observed in windows of $20,000$ packets from real traffic. We use 24-hr unsampled and non-anonymized packet traces from the WIDE backbone [7]. We divide the trace into 4-hr intervals. Each interval has a volume of approximately 20GBytes and 40 million packets. We chose an interval of 4 hours because statistics such as the multiplexing level and the number of packets per minute were stable over this period (see Figure 2), and because in such a long period there usually are more than $2,000$ window samples.

As mapping function, we use the AS number[1] of the destination IP address. (We have also used the AS number of the source IP address and obtained similar results.) The result is a sequence of AS numbers for a period of 4 hours. We compute the empirical histogram of AS numbers of the entire trace to use as the reference distribution in the rest of the experiment. We believe this distribution to be an approximation of the typical distribution of symbols induced by the normal usage of the network. To rule out the interference of non-stationarity effects in our observations, we repeated these experiments breaking the entire trace into intervals of 1-hour as well, and found that the results of 1-hour traces were consistent with each other and with the 4-hour trace.

In order to measure the deviation of each window to the reference distribution, we compute the Küllback-Leibler distance between the histogram of a window $p(x)$ and the reference distribution $q(x)$, obtained from the entire sequence. The KL distance is defined as

$$D(p||q) = \sum_{x \in \mathscr{X}} p(x) \log \frac{p(x)}{q(x)}$$

Sanov's Theorem shows that the probability of observing a histogram $p(x)$ in a window of i.i.d. samples from a reference distribution $q(x)$ declines exponentially in the KL distance [2]

$$Prob[D(p||q) > x] \propto 2^{-nx}$$

That is, Sanov's Theorem establishes KL distance as the natural metric for measuring the variability of histograms.

In the leftmost plot of Figure 3, we show the distribution of KL distance for the original sequence, and for transformations of the original sequence. Each transformation consists of scrambling the symbols from a number n of adjacent windows. Clearly, the smallest possible value for n is 2, since scrambling the symbols within a single window will not change its histogram. After the sequence has been transformed in this way, we compute again the distribution of KL distance for each window of $20,000$ packets. We also include the distribution of KL distance for an i.i.d. sequence with the same reference distribution $q(x)$. It is evident from these curves that for small values of n, the difference in the distribution of KL distance to the original case is also small. It is only when almost the entire trace is scrambled that we get close to the i.i.d. case. This is a simple way to observe the presence of LRMI in the sequence of symbols.

We also show in Figure 3 two horizontal cuts of the previous curve, at 90^{th} and 99^{th}-percentiles for two more traces. These cuts show the logarithmic decline of KL distance with the decrease in the correlation in the sequence of packets.

[1]Multi-homed addresses were resolved by choosing the smallest AS number.

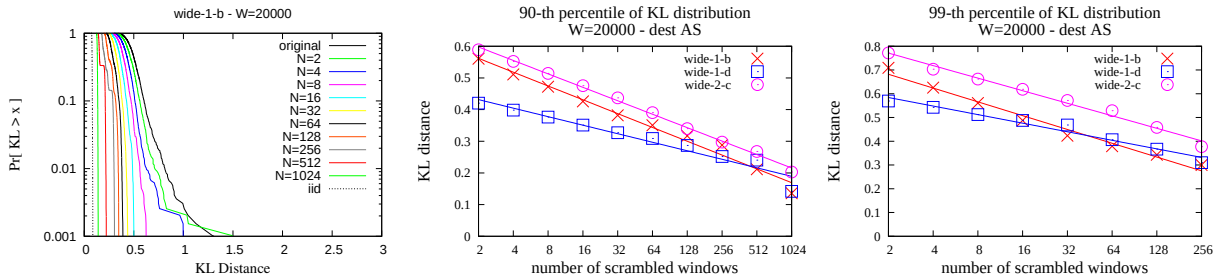

Figure 3: Evidence of Long Range Mutual Information

These figures illustrate the importance of accurately capturing LRMI when modeling real traffic. For instance, a simple anomaly detection method would be to plot the distribution of KL distance for anomaly free traffic, and then chose a false alarm rate, say 1%, to set a threshold for detecting anomalies. Inspecting these figures it is clear that models that ignore mutual information, despite its small values at any given lag, give highly inaccurate results for real traffic.

4. RELATED WORK

The seminal papers of [4] and [8] have demonstrated the presence of LRD in network traffic and its effects on performance of network elements. In this paper we introduce the concept of LRMI. Both LRD and LRMI have origins in the heavy tailed nature of distributions involved in each domain. Whereas LRD is concerned with the temporal distribution of volume counts, LRMI is concerned with the spatial distribution of volume counts. The consequences of one and the other are different as well, since they expose the high variability of traffic in time or space. For instance, a router that performs table look ups based on destination IP address prefixes may be more concerned with the variability in space than in time.

We have illustrated the importance of LRMI in anomaly detection and change detection (without presenting a full-blown method). We point the interested reader to [3] and [5] for practical frameworks that use KL distance for change and anomaly detection in settings like the one we consider here.

5. CONCLUSIONS

In this paper we have investigated the nature of dependence in packet traffic when the traffic is viewed as a sequence of symbols. This is an area that has not been extensively explored, but can yield insights useful in synthetic traffic generation and statistical anomaly detection.

We have shown that the mutual information between packet headers tends to decline slowly over time, in a manner analogous to the slow decline of correlation in long-range dependent processes. For this reason, we use the term *long range mutual information* to describe the dependence structure between packets. We show that long range mutual information has a very strong effect on the distribution of symbols seen in a given window, and yet in real traffic it is rather difficult to measure directly.

We have demonstrated the presence of LRMI in real traces. In particular, we showed the effect of LRMI on the distribution of KL distance of histograms over windows of fixed length. The distribution of KL distance of network traffic is significantly different from what is expected of i.i.d. sequences. Furthermore, we show that scrambling adjacent windows is not enough to eliminate the effect of LRMI. Only when almost the entire sequence of millions of symbols is randomly reorganized that the resulting histograms become approximately uncorrelated.

6. REFERENCES

[1] T. M. Cover and J. A. Thomas. *Elements of information theory*. Wiley-Interscience, New York, NY, USA, 1991.

[2] T. M. Cover and J. A. Thomas. *Elements of information theory*, chapter 12. Wiley-Interscience, New York, NY, USA, 1991.

[3] T. Dasu, S. Krishnan, S. Venkatasubramanian, and K. Yi. An information-theoretic approach to detecting changes in multi-dimensional data streams. In *Interface 2006, 38th Symposium on the interface of statistics, computing science, and aplications*, 2006.

[4] A. Erramilli, O. Narayan, and W. Willinger. Experimental queueing analysis with long-range dependent packet traffic. *IEEE/ACM Transactions on Networking*, 4(2):209–223, April 1996.

[5] Y. Gu, A. McCallum, and D. Towsley. Detecting anomalies in network traffic using maximum entropy estimation. In *IMC'05, Internet Measurement Conference*, 2005.

[6] W. Leland, M. Taqqu, W. Willinger, and D. Wilson. On the self-similar nature of Ethernet traffic (extended version). *IEEE/ACM Transactions on Networking*, 2:1–15, 1994.

[7] MAWI Working Group. Traffic archive. `http://mawi.wide.ad.jp/mawi/`.

[8] W. Willinger, M. S. Taqqu, R. Sherman, and D. V. Wilson. Self-similarity through high-variability: Statistical analysis of Ethernet LAN traffic at the source level. *IEEE/ACM Transactions on Networking*, 5(1):71–86, February 1997.

APPENDIX

A. DERIVATION OF MUTUAL INFORMATION FORMULA

In this section, we derive the formula of mutual information at lag l for the sequence of symbols generated by the

model described in Section 2.

Remember that mutual information at lag l for a sequence drawn from alphabet $\mathscr{X}$, is defined as

$$I(X_i; X_{i+l}) = \sum_{x_i \in \mathscr{X}} \sum_{x_{i+l} \in \mathscr{X}} \phi(x_i, x_{i+l}) \log \frac{\phi(x_i, x_{i+l})}{\phi(x_i)\phi(x_{i+l})}$$

$\phi(x)$ is the symbol distribution, which supposedly is known already. For the uniform case $\phi(x) = 1/|\mathscr{X}|$. Thus the only unknown part in the equation above is the joint distribution of symbols at lag l.

Let $f(X_i)$ denote the flow associated with the random variable X_i, and denote by $\mathcal{S}_l$ the event $f(X_i) = f(X_{i+l})$, and by $p(l)$, its probability. Using a total probability argument, we can write the joint probability as

$$\begin{aligned}
\phi(x_i, x_{i+l}) &= \phi(x_i, x_{i+l}|\mathcal{S}_l)p(l) \\
&+ \phi(x_i, x_{i+l}|\neg\mathcal{S}_l)(1 - p(l)) \\
&= \phi(x_i, x_{i+l}|\mathcal{S}_l)p(l) \\
&+ \phi(x_i)\phi(x_{i+l})(1 - p(l)) \\
&= \delta(x_i, x_{i+l})\phi(x_i)p(l) \\
&+ \phi(x_i)\phi(x_{i+l})(1 - p(l))
\end{aligned}$$

where in step 2, we use the fact that if two packets belong to different flows, then the distribution of symbols is independent, by construction of our model; and in step 3, notice that the joint distribution of symbols from the same flow is zero, if the symbols are different, since a flow must have the same symbol in all packets, and $\phi(x)$ otherwise, since the choice of the second symbol is totally dependent on that of the first. Thus we use the delta function, defined as $\delta(x, y) = 1$, if $x = y$; and 0 otherwise, as a result of this analysis.

In order to compute $p(l)$, we need to define two new events. First denote by $(X_i \leftarrow k)$ the event of a random packet X_i being associate with a flow of size k. The probability of such event can be calculated by

$$P[X_i \leftarrow k] = \frac{k\kappa(k)}{\sum_{\forall j} j\kappa(j)} = \frac{k\kappa(k)}{E_\kappa}$$

where $\kappa(\cdot)$ is the flow size probability distribution function, and E_κ, its mean. Imagine that there are F flows in a sequence, then sort all the packets in the sequence by flow, and then, sort all the flows by size. There are $kF\kappa(k)$ packets which belong to a flow of size k, in a total of $\sum_{\forall j} jF\kappa(j)$ packets.

Second let $ord(X_i)$ be the order of X_i in the sequence of packets of flow $f(X_i)$, the probability that a random packet has $ord(X_i) = \{1, \ldots, k\}$, where k is the size of its flow is $1/k$.

Now we are going to use the total probability argument twice, first conditioning on all events $(X_i \leftarrow k)$, and then, on all possible orders of a packet within a flow, to derive the formula for $p(l)$.

$$\begin{aligned}
p(l) &= \sum_{\forall k} P[\mathcal{S}_l | X_i \leftarrow k] P[X_i \leftarrow k] \\
&= \sum_{\forall k} P[\mathcal{S}_l | X_i \leftarrow k] \frac{k\kappa(k)}{E_\kappa} \\
&= \sum_{\forall k} \sum_{i=1}^{k-1} P[\mathcal{S}_l | ord(X_i) = i, X_i \leftarrow k] \frac{1}{k} \frac{k\kappa(k)}{E_\kappa} \\
&= \frac{1}{mE_\kappa} \sum_{\forall k} \kappa(k) \sum_{i=1}^{k-1} \Big(1 - \sum_{j=k-i}^{l-1} B_{l-1, \frac{1}{M}}(j)\Big)
\end{aligned}$$

In the last step, we use the Binomial distribution ($B_{n,r}(i) = \binom{n}{i} r^i (1-r)^{n-i}$) to capture the fact that two packets with lag l will belong to the same flow only if the remaining $(k-i)$ packets of the flow are not consumed in $(l-1)$ steps or less, and a packet from the same flow is selected with probability $1/m$, where m is the number of active flows, which is fixed in our model.

Finally, putting all this together, we can derive a formula for mutual information.

$$\begin{aligned}
I(X_i, X_{i+l}) &= \sum_{\forall x_i, x_{i+l}} \Bigg\{ \delta(x_i, x_{i+l})\phi(x_i)p(l) \\
&+ \phi(x_i)\phi(x_{i+l})\Big[1 - p(l)\Big] \Bigg\} \\
&\log \Bigg\{ \frac{\delta(x_i, x_{i+l})}{\phi(x_{i+l})} p(l) \\
&+ \Big[1 - p(l)\Big] \Bigg\}
\end{aligned}$$

This formula is computationally expensive, since it has to evaluate l^2 binomial sums. Therefore we present another formulation of $p(l)$ using the normal approximation of the binomial distribution.

$$\begin{aligned}
p(l) &= \frac{E_\kappa - 1}{mE_\kappa} - \frac{(l-1)F_\kappa^c(l-1)}{m^2 E_\kappa} - \\
&- \sum_{i=2}^{l-1} \Bigg[\Big(erf(\frac{l - 0.5 - \mu}{\sqrt{2}\sigma}) - erf(\frac{i - 1.5 - \mu}{\sqrt{2}\sigma}) \Big) \\
&\times \Big(\frac{F_\kappa(l-1) - F_\kappa(i-1)}{2mE_\kappa} \Big) \Bigg]
\end{aligned}$$

where F_κ is the cumulative flow size distribution function, and F_κ^c is its complement; and erf is the Normal cumulative distribution function, with mean $\mu = \frac{l-1}{m}$ and variance $\sigma^2 = \frac{(l-1)(m-1)}{m^2}$.