A Sum of Bernoulli Sources Approximation for Packet Switched Network Traffic in Backbone Links

Upeka Premaratne and Uthpala S. Premarathne.

***Abstract*—In this letter an intuitive measurement based approximation for packet switched network traffic in backbone links is presented. It consists of a sum of independent non-identical Bernoulli sources. The resulting sum has a Poisson binomial distribution which can be approximated to a skewed Gaussian distribution at the asymptote. It can be further approximated to a binomial distribution for predicting congestion by estimating the number of users of the network. The results are verified using samples of traffic data from 42 academic and research institutions.**

***Index Terms*—Network traffic, backbone links, Poisson binomial model, Das-Gupta and Rubin estimator**

I. Introduction

The ability to statistically describe the nature of network traffic sources in packet switched communication networks is essential for optimal network design in terms of capacity planning and flow management [1]. Initially a Poisson distribution was proposed [2]. Shortcomings of this model were shown in [3]. The self-similar model of network traffic [4] [5] [6] is currently the most widely accepted model based upon empirical statistical analysis (i.e., the results of the Hurst parameter). Network traffic has also been demonstrated to have Long Range Dependence (LRD) [7] [8].

With few exceptions such as industrial automation networks [9], network traffic is always human generated. Therefore, the human factor in network traffic source modeling cannot be excluded. One such example is the occurrence of flash events such as news, software upgrades and cyber attacks that cause a sudden surge in network traffic [10] [11]. Another main human factor is the circadian rhythm [12] which determines the time during which humans are most likely to interact with the network and in turn generate traffic. Previous studies have demonstrated the relationship between the circadian rhythm and cyber attacks [13].

A. Contribution

The main contribution of this letter is an intuitive and versatile approximate model for packet switched network traffic of backbone links. It is based on a sum of independent non-identical Bernoulli sources which approximates the traffic generated during the high activity phase of the human circadian rhythm. The resulting sum can be approximated to a binomial distribution from which the effective number of users of the network can be estimated for capacity planning. The asymptotic distribution for a large number of users can be shown to be approximately skewed Gaussian (from [14]). The existence of a circadian rhythm in network traffic is confirmed by the χ^2 periodogram method of [12] [15].

The main advantage of this model is that it is predominantly observation based, requiring the observed traffic density of the backbone. Such data is readily available from network monitoring tools such as RRDtool. In addition, the maximum capacity link within the network is also needed. When compared to recent individual on-off traffic models such as [16], individual traffic and capacity parameters have to be known. Also in the Hidden Markov Model of [17], a large volume of packet level data such as the inter-arrival time and packet size are required. The proposed model is also generic compared to models such as [1] which models video traffic. The proposed model is applicable to the main backbone of an enteprise network that connects it to the Internet as well as backbone links that connect local area networks within it.

II. Proposed Model

A. Sum of Bernoulli Traffic Sources

In this model, the basic unit of traffic is taken as a human user during active phase of the circadian rhythm. This unit is modeled as a Bernoulli source with a Bernoulli random variable $V_i \in \{0, 1\}$ which takes up bandwidth c_i when it transmits (i.e., $V_i = 1$) with $p(V_i = 1) = p_i$ where the probabilities p_i are independent but non-identical. Making p_i non-identical enables the model to incorporate different types of human users, such as those who do occasional web-browsing with small p_i and heavy users who frequently download large files or watch streaming videos. The total bandwidth used up at a given time instance, k is $s[k]$ and is given by the sum of all sources. Therefore,

$$s[k] = \sum_{i=1}^{n} c_i V_i \tag{1}$$

U. Premaratne is with the Department of Electronic and Telecommunication Engineering, University of Moratuwa, Katubedda, Moratuwa 10400, Sri Lanka e-mail:upeka@uom.lk

U. S. Premarathne is with the Department of Electrical and Computer Engineering, the Open University of Sri Lanka, Nawala, Nugegoda 11222, Sri Lanka e-mail:uspre@ou.ac.lk

The authors acknowledge the support from University of Moratuwa Senate Research Committee Grant SRC/ST/2016/10.

Let the total number of users be n and the total observations of $s[k]$ be l. This results in an estimation problem of the form

$$\mathcal{V}C = \begin{pmatrix} 0 & 1 & \cdots & 0 \\ 0 & 0 & \cdots & 1 \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & \cdots & 0 \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \\ \vdots \\ c_n \end{pmatrix} = \begin{pmatrix} s[1] \\ s[2] \\ s[3] \\ \vdots \\ s[l] \end{pmatrix} = S \quad (2)$$

where $\mathcal{V}$ is a $l \times n$ sparse random matrix consisting of ones and zeros. It should be noted that the total number of sources n is unknown along with all p_i, requiring n and p_i to be estimated based upon the observation matrix S. Since utilization is calculated based upon the average usage of the network, from (1) and (2),

$$E[S] = \sum_{i=1}^{n} c_i p_i \quad (3)$$

Assumption 1. *For the given l samples, the number of users n and the individual probability p_i of an user remains constant.*

This assumption is required for (1)-(3) to hold. Though the individual capacities c_i can be known, in a network with a large number of users finding each individual c_i is impractical. This is resolved by scaling (3) by an arbitrary scalar γ such that $\max[C] \leq \gamma < \gamma_B$ (where γ_B is the backbone link capacity) which results in,

$$E[S] = \gamma \sum_{i=1}^{n} \left(\frac{c_i p_i}{\gamma} \right) = \gamma \sum_{i=1}^{n} \pi_i \quad (4)$$

Since all $p_i < 1$, this results in all $\pi_i \leq p_i < 1$ and the random variable $\mathcal{S} = \frac{S}{\gamma}$ becoming Poisson binomial i.e., $\mathcal{P}(n, \{\pi_i\})$ with $E[\mathcal{S}] = \sum_{i=1}^{n} \pi_i$ and variance, $\sigma^2 = \sum_{i=1}^{n} (1 - \pi_i)\pi_i$.

B. Asymptotic Behavior

From [14], when $n \to \infty$, $\mathcal{P}(n, \{\pi_i\})$ converges to a skewed Gaussian distribution. Once the maximum likelihood estimates for Gaussian μ_{NA} and σ^2_{NA} are obtained from $\mathcal{S}$, the discrete skewed CDF is obtained from

$$F_n(i) \approx \Phi(k) + \frac{\Gamma}{(1 - k^2)} \phi(k) \quad (5)$$

where $k = (i + 0.5 + \mu_{NA})/\sigma_{NA}$, $k \in \mathbb{Z}^+$ and $\Gamma = (E[S - \mu_{NA}])^3$ with $\Phi(\cdot)$ and $\phi(\cdot)$ being the CDF and PDF of the standard normal distribution respectively. Previous studies have shown the network traffic distribution of backbone links to be Gaussian [18] and in Section III-A, the proposed skewed Gaussian distribution is demonstrated to have a better empirical fit than the Gaussian distribution proposed in [18].

C. Congestion Prediction

From [19], $\mathcal{P}(n, \{\pi_i\})$ can be approximated to an equivalent binomial distribution $\mathcal{B}(m, p)$ such that,

$$\delta(\mathcal{P}, \mathcal{B}) \leq \frac{1 - p^{\tilde{m}} - q^{\tilde{m}}}{\tilde{m}pq} \left[2 \left(\lambda_3 - \frac{\lambda_2^2}{\lambda} \right) + \lambda \left| p - \frac{\lambda_2}{\lambda} \right| \right] \quad (6)$$

where $\tilde{m} = m + 1$, $p + q = 1$, $\lambda_r = \sum_{i=1}^{n} \pi_i^r$, $\lambda_1 = \lambda$ and $\delta(\cdot, \cdot)$ is the Total Variation Distance (TVD). Here, $m \neq n$ with $m > n$ in general. The method of binomial approximation of Roos [20] where $m = n$ cannot be applied because it requires $\{\pi_i\}$ to be explicitly known instead of a summation of order r as in [19]. Once approximated, it is possible to estimate $\bar{m}$ and $\bar{p}$ using the Das-Gupta and Rubin estimator [21] for confidence α given by

$$\bar{m} = \frac{(\max[\mathcal{S}])^{\alpha+1} (V[\mathcal{S}])^{\alpha}}{(E[\mathcal{S}])^{\alpha} (\max[\mathcal{S}] - E[\mathcal{S}])^{\alpha}} \quad (7)$$

$$\bar{p} = \frac{(E[\mathcal{S}])^{\alpha+1} (\max[\mathcal{S}] - E[\mathcal{S}])^{\alpha}}{(\max[\mathcal{S}])^{\alpha+1} (V[\mathcal{S}])^{\alpha}} \quad (8)$$

These can be subsequently used to estimate λ, λ_2 and λ_3 by minimizing (6), a feasible albeit computationally intensive task.

Assumption 2. *For the given l samples, the number of users m, remains constant.*

Assumption 2 is needed for the estimates (7) and (8) to be valid since m cannot change. It is reasonable only for small time periods since typically the number of users will change in an academic institution over the course of a day due to the nature of academic activities. In the case of research institutes and offices this can extend to typical working hours.

Assumption 3. *During the estimation no anomalous flash events where the proposed statistical model of (1) no longer applies occur.*

In other words, Assumption 3 is required to ensure normal traffic conditions for which the proposed model is valid. This excludes the congestion caused by anomalous traffic during flash events where the bottleneck results in significant traffic delays that compromise the service [10]. Such delays are not incorporated in the proposed model making it invalid for such traffic conditions.

For the backbone links considered, congestion ($\mathcal{C}$) is defined as the bandwidth utilization exceeding 75%. By defining $N = \gamma_B/\gamma$ such that $N \in \mathbb{Z}^+$ and selecting γ accordingly, congestion can be inferred from the satisfaction of $\bar{m} > 0.75N$. This results in the inference

$$\text{If } [\bar{m} > 0.75N] \rightarrow \mathcal{C} = \text{true} \quad (9)$$

for predicting congestion from the Das-Gupta Rubin Estimate (DRE) of $\bar{m}$.

III. Empirical Results

The proposed model is verified using real network traffic data. The data is obtained from RRDtool logs of Lanka Education and Research Network (LEARN), the educational Internet service provider of 73 locations of 42 national universities and research institutes of Sri Lanka. The data consists of average bandwidth usage of the downlink for 1, 6, 24 and 288 minute slots. The backbone link capacities range from 5 to 1000Mbps. The scalar $\gamma = 125000$ bytes. Samples with at least 400 data points are considered.

A. Comparison of Distributions

In this section, the skewed Gaussian and binomial distributions of the proposed model are compared with the previously reported Gaussian [18] and log-normal [22] distributions obtained for backbone traffic. Despite the large body of literature for network traffic, only limited work analyzes the distribution itself. For the binomial model, the number of users $\bar{m}$ and probability $\bar{p}$ are first obtained. Since $\max[\mathcal{S}] \leq \bar{m}$, the resulting binomial distribution need only be considered up to $\lceil \bar{m} \rceil$ if $\bar{m} < N$ or N if $\bar{m} \geq N$. The other three distributions are obtained up to N by maximum likelihood estimation of the distribution parameters from $\mathcal{S}$. Samples with a trivial estimation result of $\bar{m} = 1$ are not considered.

The four distributions are compared in terms of the Kullbeck-Liebler Divergence (KLD) and TVD. The results (Table I) show that the proposed skewed Gaussian approximation is the best fit according to the TVD which is valid for all results. In the case of the KLD however, the log-normal distribution fares better for the results that can be computed. Another result of significance is the fact that both skewed Gaussian and Gaussian distributions exist for all valid data samples but the log-normal distribution does not exist for 4.56-7.30% of samples across the timescales due to invalid maximum likelihood estimates. For both metrics, the proposed skewed Gaussian distribution provides a better fit than the Gaussian distribution. The results are also consistent along all time scales.

TABLE I
COMPARISON O F DISTRIBUTIONS

Size (No.)	Metric	Distribution			
		Bin.	Log N.	Norm.	Skew N.
1 min (219)	KLD Valid (%)	49.772	95.434	100.000	77.626
	KLD (μ)	1.762	**0.309**	0.925	0.552
	KLD (σ^2)	28.375	**0.070**	0.425	0.230
	TVD Valid (%)	100.000	95.434	100.000	100.000
	TVD (μ)	0.347	0.167	0.296	**0.159**
	TVD (σ^2)	0.045	**0.011**	0.050	**0.011**
6 min (227)	KLD Valid (%)	45.815	94.273	100.000	78.414
	KLD (μ)	3.125	**0.295**	0.738	0.426
	KLD (σ^2)	62.984	**0.044**	0.234	0.113
	TVD Valid (%)	100.000	94.273	100.000	100.000
	TVD (μ)	0.308	0.153	0.253	**0.130**
	TVD (σ^2)	0.034	0.012	0.036	**0.008**
24 min (236)	KLD Valid (%)	46.186	93.220	100.000	77.542
	KLD (μ)	2.700	**0.243**	0.669	0.352
	KLD (σ^2)	39.139	**0.029**	0.182	0.064
	TVD Valid (%)	100.000	93.220	100.000	100.000
	TVD (μ)	0.302	0.139	0.245	**0.119**
	TVD (σ^2)	0.034	0.010	0.035	**0.007**
288 min (219)	KLD Valid (%)	56.164	92.694	100.000	82.192
	KLD (μ)	1.500	**0.198**	0.420	0.248
	KLD (σ^2)	15.704	**0.024**	0.085	0.043
	TVD Valid (%)	100.000	92.694	100.000	100.000
	TVD (μ)	0.244	0.105	0.163	**0.081**
	TVD (σ^2)	0.022	0.008	0.020	**0.003**

B. Congestion Prediction

Though the binomial distribution performs poorly in terms of the KLD and TVD (Table I), the resulting DRE is capable of predicting congestion of the backbone link with the highest accuracy. The inference of (9), is used along with

$$\text{If } [P(Y > 0.75N) > 0] \rightarrow \mathcal{C} = \text{true}$$

for a distribution of random variable Y. The results of each predictor (Table II) show that the DRE has the highest accuracy, precision and specificity compared to predictors that use a statistical distribution. This can be attributed to the fact that the skewed Gaussian distribution (since asymptotic) requires n to be large and unquantifiable, loosing the causal relationship between the network traffic measurement and the individual sources of (1). Furthermore, there is no established causal relationship between the log-normal distribution and the generation process. However, the DRE can estimate $\bar{m}$, retaining the causal relationship.

The only performance metric where the DRE lags is the sensitivity where prediction using the binomial distribution (the only other distribution that retains the causal relationship) leads. Due to the existence of $0.75N$ in the tail of the Gaussian and skewed Gaussian distributions where the skew "correction" of (5) has little influence, the performance of the two distributions is near identical. When it comes to the actual correlation between the observed $P(\mathcal{S} > 0.75N)$ and analytically obtained $P(Y > 0.75N)$ of the distribution (Table III along with sample Figs. 1 and 2), the Gaussian distribution performs best followed by the skewed Gaussian distribution both having correlations in excess of 0.9. Due to its high spread, the DRE has the lowest correlation.

TABLE II
CONGESTION PREDICTION PERFORMANCE

Size (No.)	Predictor	Model Performance (%)			
		Accu.	Prec.	Sens.	Spec.
1 min (219)	DRE	**90.868**	**100.000**	80.769	**100.000**
	Binomial	70.776	23.810	**100.000**	67.839
	Log-normal	63.636	90.476	52.778	87.692
	Normal	81.279	58.333	89.091	78.659
	Skew Norm.	81.279	58.333	89.091	78.659
6 min (227)	DRE	**89.868**	**99.048**	82.540	**99.010**
	Binomial	70.044	35.238	**100.000**	64.211
	Log-normal	60.748	93.137	55.233	83.333
	Normal	78.414	59.048	91.176	72.956
	Skew Norm.	78.414	59.048	91.176	72.956
24 min (236)	DRE	**88.559**	**100.000**	79.389	**100.000**
	Binomial	68.220	28.846	**96.774**	63.902
	Log-normal	61.364	96.000	54.237	90.698
	Normal	81.780	62.500	94.203	76.647
	Skew Norm.	81.356	61.538	94.118	76.190
288 min (219)	DRE	**87.671**	**100.000**	75.455	**100.000**
	Binomial	74.886	33.735	**100.000**	71.204
	Log-normal	58.621	97.500	48.750	95.349
	Normal	85.388	65.060	94.737	82.099
	Skew Norm.	84.932	63.855	94.643	81.595

C. Circadian Rhythm

The longer 24 and 288 minute slots are used for testing for the circadian rhythm. The peak of the χ^2 periodogram is detected. For 24 minute slots, 60 will correspond to a single day. So $P = 60$ while for 288 minute slots, $P = 5$ with $P = 35$ also being used to test for the existence of a weekly

TABLE III
CONGESTION PREDICTION CORRELATION

Size	Model Correlation				
	DRE	Binom.	Log Norm.	Normal	Skew N.
1 min	0.495	0.523	0.752	**0.946**	0.938
6 min	0.401	0.725	0.808	**0.981**	0.974
24 min	0.441	0.738	0.837	**0.975**	0.970
288 min	0.363	0.761	0.747	**0.963**	0.956

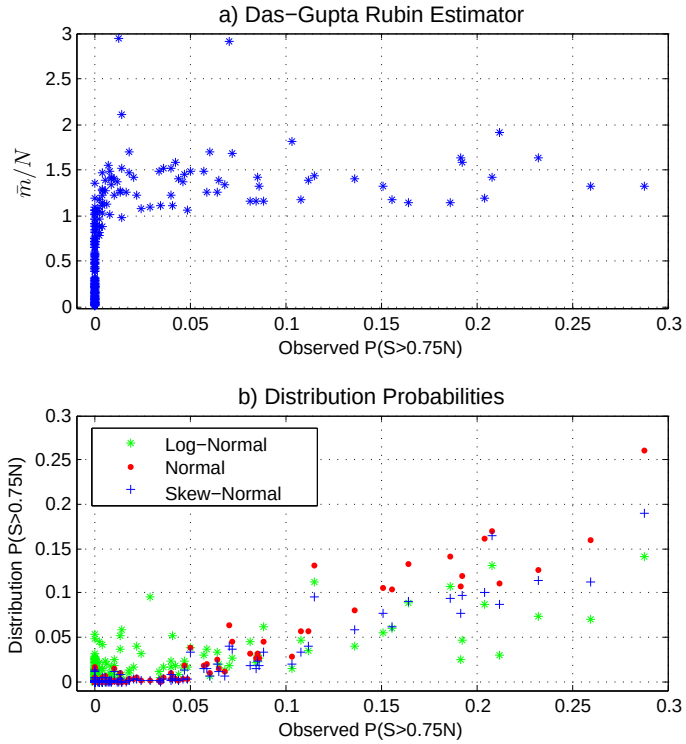

Fig. 1. Congestion Prediction Result Comparison (1 min Samples)

cycle of high activity weekdays and low activity weekends. The results (Table IV) confirm this with $P = 60$ and $P = 5$ peaks being detected in respectively 84.1% and 98.1% of the valid samples. The weekly cycle is detected in 88.8% of samples. The existence of a circadian rhythm in network traffic is intuitive, but to the best of the authors knowledge not previously verified from empirical evidence.

TABLE IV
χ^2 PERIODOGRAM RESULTS

Size (min)	Range $P \pm \delta$	Samples	Mean	Mode P (%)	Comment
24	60±10	258	60.3	84.1	Circadian Rhythm
288	5±3	268	5.1	98.1	Circadian Rhythm
288	35±10	240	35.7	88.8	Weekly Cycle

IV. CONCLUSION

In this letter, the traffic of packet switched backbone links is shown to be approximate to a sum of independent non-identical Bernoulli sources during the high activity phase of the human circadian rhythm. The resulting sum has a Poisson binomial distribution with a skewed Gaussian distribution at

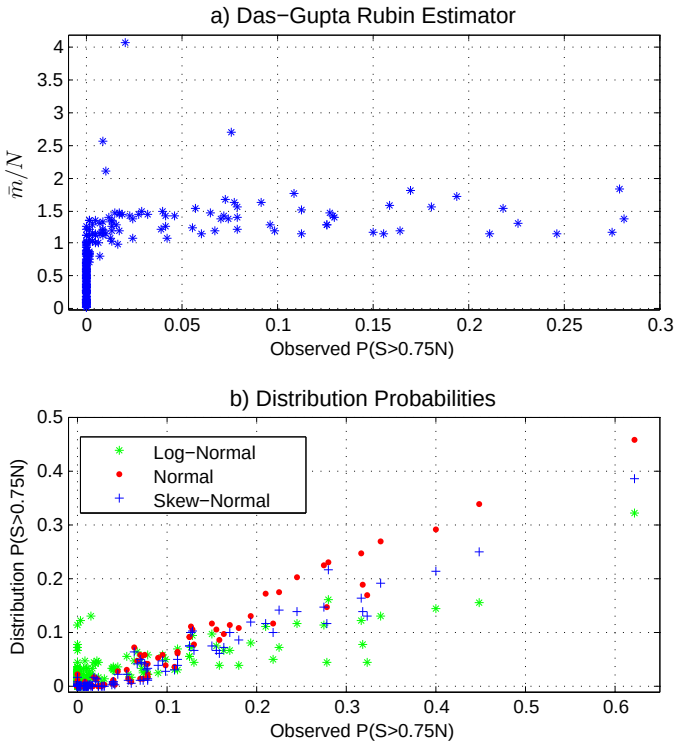

Fig. 2. Congestion Prediction Result Comparison (6 min Samples)

the asymptote. For the given dataset, the skewed Gaussian distribution has the best empirical fit in terms of the TVD. In order to preserve the latent relationship between the individual traffic sources and measured backbone traffic, the Poisson binomial distribution can be approximated to a binomial distribution from which the effective number of users can be estimated using the Das-Gupta Rubin estimator. This results in a highly accurate predictor for network congestion which is applicable to the main backbone of an enterprise network and the backbones between local area networks within it. The presence of a circadian rhythm in network traffic is also empirically confirmed.

Overall, the results of this letter favor the LRD model of network traffic over self-similarity. Future directions of this work include the modeling of flash events and investigation into the possibility of combined flash event and the circadian rhythm giving rise to the apparent self-similarity of network traffic. Finding a causal relationship between the traffic source and log-normal distribution is another interesting open problem. A further direction of investigation is the applicability and possible extension of this model for data communication protocols in 4G cellular networks.

V. ACKNOWLEDGEMENTS

The authors would like to thank R. G. Regal, S. Herath and D. Gunawardena of LEARN for providing the backbone network traffic data of the academic and research institutions of Sri Lanka.

REFERENCES

[1] W. Abbessi and H. Nabli, "General approach for video traffic: from modeling to optimization," *Multimedia Syst.*, pp. 1–17, 2018.

[2] F. A. Tobagi, M. Gerla, R. W. Peebles, and E. G. Manning, "Modeling and measurement techniques in packet communication networks," *Proc. IEEE*, vol. 66, no. 11, pp. 1423–1447, 1978.

[3] V. Paxson and S. Floyd, "Wide area traffic: the failure of Poisson modeling," *IEEE/ACM Trans. Netw.*, vol. 3, no. 3, pp. 226–244, 1995.

[4] W. E. Leland, M. S. Taqqu, W. Willinger, and D. V. Wilson, "On the self-similar nature of Ethernet traffic (extended version)," *IEEE/ACM Trans. Netw.*, vol. 2, no. 1, pp. 1–15, 1994.

[5] H. Chen, H. Jin, and S. Wu, "Minimizing inter-server communications by exploiting self-similarity in online social networks," *IEEE Trans. Parallel Distrib. Syst.*, no. 4, pp. 1116–1130, 2016.

[6] D. Jiang, L. Huo, and Y. Li, "Fine-granularity inference and estimations to network traffic for SDN," *PloS one*, vol. 13, no. 5, p. e0194302, 2018.

[7] J. Beran, R. Sherman, M. S. Taqqu, and W. Willinger, "Long-range dependence in variable-bit-rate video traffic," *IEEE Trans. Commun.*, vol. 43, no. 234, pp. 1566–1579, 1995.

[8] J. Domańska, A. Domańska, and T. Czachórski, "A few investigations of long-range dependence in network traffic," in *Information Sciences and Systems 2014*, T. Czachórski, E. Gelenbe, and R. Lent, Eds. Cham: Springer International Publishing, 2014, pp. 137–144.

[9] S. Vitturi and F. Tramarin, "Energy efficient Ethernet for real-time industrial networks," *IEEE Trans. Autom. Sci. Eng.*, vol. 12, no. 1, pp. 228–237, 2015.

[10] S. Bhatia, D. Schmidt, G. Mohay, and A. Tickle, "A framework for generating realistic traffic for distributed denial-of-service attacks and flash events," *Comput. Secur.*, vol. 40, pp. 95–107, 2014.

[11] S. Yu, W. Zhou, and R. Doss, "Information theory based detection against network behavior mimicking DDoS attacks," *IEEE Commun. Lett.*, vol. 12, no. 4, pp. 318–321, 2008.

[12] R. Refinetti, "Non-stationary time series and the robustness of circadian rhythms," *J. Theor. Biol.*, vol. 227, no. 4, pp. 571–581, 2004.

[13] U. Premaratne, J. Samarabandu, T. Sidhu, R. Beresh, and J. C. Tan, "Security analysis and auditing of IEC61850-based automated substations," *IEEE Trans. Power Del.*, vol. 25, no. 4, pp. 2346–2355, 2010.

[14] A. Y. Volkova, "A refinement of the central limit theorem for sums of independent random indicators," *Theory Probab. Appl.*, vol. 40, no. 4, pp. 791–794, 1996.

[15] P. G. Sokolove and W. N. Bushell, "The chi square periodogram: its utility for analysis of circadian rhythms," *J. Theor. Biol.*, vol. 72, no. 1, pp. 131–160, 1978.

[16] H. Schwefel, I. Antonios, and L. Lipsky, "Understanding the relationship between network traffic correlation and queueing behavior: A review based on the n-burst on/off model," *Perform. Evaluation*, vol. 115, pp. 68–91, 2017.

[17] A. Dainotti, A. Pescapé, P. Salvo Rossi, F. Palmieri, and G. Ventre, "Internet traffic modeling by means of Hidden Markov Models," *Comput. Netw.*, vol. 52, no. 14, pp. 2645–2662, 2008.

[18] R. V. De Meent, M. Mandjes, and A. Pras, "Gaussian traffic everywhere?" in *2006 IEEE International Conference on Communications*, vol. 2, June 2006, pp. 573–578.

[19] S. Y. Soon, "Binomial approximation for dependent indicators," *Stat. Sinica*, pp. 703–714, 1996.

[20] B. Roos, "Binomial approximation to the Poisson binomial distribution: The Krawtchouk expansion," *Theory Probab. Appl.*, vol. 45, no. 2, pp. 258–272, 2001.

[21] A. DasGupta and H. Rubin, "Estimation of binomial parameters when both n, p are unknown," *J. Stat. Plan. Infer.*, vol. 130, no. 1-2, pp. 391–404, 2005.

[22] T. Li and J. Liu, "Cluster-based spatiotemporal background traffic generation for network simulation," *ACM T. Model Comput. S.*, vol. 25, no. 1, p. 4, 2015.