JURNAL INFOTEL
VOL. 17, NO. 2, MAY 2025, PP. 268–278. DOI:10.20895/INFOTEL.V17I2.1359

RESEARCH ARTICLE

Investigating Synthetic Traffic Generators for Zipf Distribution Simulation Accuracy

Feri Fahrianto[1,*], Viva Arifin[2], Imam Marzuki Shofi[3], Hendra Bayu Suseno[4], Victor Amrizal[5], Muhamad Azhari[6], and Anggy Eka Pratiwi[7]

[1,2,3,4,5,6]Faculty of Science and Technology, Syarif Hidayatullah State Islamic University Jakarta, 15412, Indonesia

[7]Indian Institute of Technology Jodhpur, India

*Corresponding email: feri.fahrianto@uinjkt.ac.id

Received: May 05, 2025; Revised: May 31, 2025; Accepted: June 05, 2025.

Abstract: Accurate traffic generation is essential for realistic network simulations in systems such as Content Delivery Networks (CDNs), Information-Centric Networks (ICNs), and the Internet of Things (IoT). These environments handle various types of data traffic, from web pages and videos to sensor data and software updates. A well-designed traffic generator enables researchers to simulate real-world workloads, test scalability, and evaluate the performance of caching and routing under realistic conditions. Each traffic class has unique characteristics, including object size distributions and access patterns. Capturing these differences is the key to producing meaningful simulation results. For instance, CDNs require simulation of content popularity and delivery latency, ICNs focus on content retrieval and caching efficiency, while IoT simulations demand modeling of device behavior and intermittent communication. To support such complex scenarios, a traffic generator must not only mimic real user behavior but also allow flexible scaling, combination, and modification of traffic patterns. This paper presents a method for evaluating synthetic traffic generators by comparing their output with the statistical properties of the Zipf distribution. The focus is on assessing whether synthetic traffic accurately reflects the heavy-tailed nature of real-world traffic as modeled by Zipf's law. By analyzing the frequency distribution of requests generated by the traffic model and comparing it to theoretical Zipf curves, the study provides insight into the fidelity of the traffic generator. We measure the discrepancy between the simulated network traffic and the theoretical model to evaluate the accuracy and realism of the traffic generation approach.

Keywords: CDN, ICN, RMSE, traffic generator, Zipf-like distribution

 ISSN: 2085-3688; e-ISSN: 2460-0997

1 Introduction

1.1 Background

Network simulation has become an indispensable tool in the research, development, and evaluation of modern computer networks. As the complexity and scale of networks continue to grow, from global Content Delivery Networks (CDN) and Information-Centric Networks (ICNs) to highly distributed and resource-constrained Internet of Things (IoT) systems, accurate and reliable simulation environments are critical. These simulations enable researchers and engineers to test new protocols, optimize system architectures, and predict performance under various conditions without the need for expensive or logistically challenging real-world deployments.

One of the most crucial components of any network simulation is the traffic generator. Traffic generators are responsible for producing synthetic traffic that mimics the behavior of real-world users and applications. The effectiveness of a simulation depends heavily on how accurately the traffic generator can reproduce realistic workloads. Inaccurate or overly simplistic traffic patterns can lead to misleading conclusions about network performance, scalability, and reliability.

Different types of network generate highly diverse traffic profiles. For instance, CDNs typically handle large volumes of data, including high-definition videos, dynamic web content, and software distributions. These networks are sensitive to content popularity and latency, requiring the simulation of varying demand patterns and caching strategies. ICNs, on the other hand, focus on data-centric communication, where the network retrieves data based on content names rather than host addresses. This model introduces unique caching and forwarding behaviors that must be accurately reflected in simulated traffic. Meanwhile, IoT networks often involve thousands or even millions of low-power devices transmitting small, irregular data packets with sporadic activity. Simulating such scenarios requires fine-grained modeling of device behavior, timing, and data generation patterns.

Given these variations, a single traffic generator must be flexible and sophisticated enough to adapt to different network contexts. More importantly, its output must be validated to ensure that the generated traffic resembles real-world behavior. One effective method for such validation is comparing the synthetic traffic against a known theoretical traffic model. In this study, we focus on the Zipf distribution, a statistical model that has been widely observed in real-world network traffic, especially in content-centric systems. The Zipf model captures the principle that a small number of popular items account for the majority of access requests, a phenomenon commonly known as the "long tail" effect.

1.2 Traffic Generator

A traffic generator like TRAGEN is a tool to generate content requests such as website names. It is developed in Python and comprises roughly 2000 lines of code. Users can interact with it via either a graphical user interface (GUI) or through a command line interface (CLI). Figure 1 displays an example of the GUI layout. To configure the tool, users are required to provide several inputs:

- Choose hit rate mode: Decide whether the synthetic trace should replicate the original's Request Hit Rate (RHR) or Byte Hit Rate (BHR).

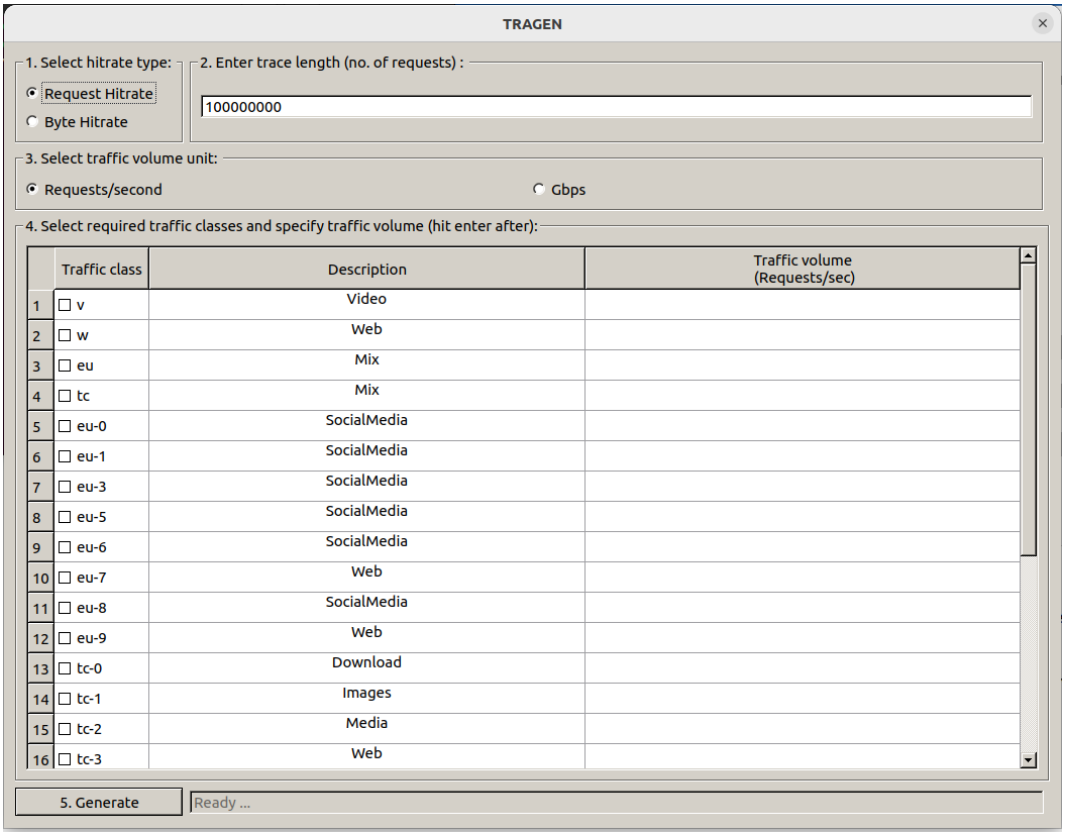

Figure 1: GUI of TRAGEN.

- Input trace size: Indicate the total number of requests that the generated synthetic trace should contain.
- Set traffic volume format: Specify the measurement unit for traffic volume in the third column—either as requests per second or in Gigabits per second (Gbps).
- Select traffic categories and assign volumes: Pick the desired traffic categories from the first column of the table and define the corresponding volume in the third column. The resulting synthetic traffic will reflect the specified distribution. The second column provides a brief explanation for each category, which may represent a single traffic type (e.g., video, web, social media) or a predefined traffic mix (e.g., "EU," which aggregates traffic handled by a cache server located in Europe).
- Initiate generation: Click the "Generate" button to allow TRAGEN to begin the synthetic trace creation process.

TRAGEN is capable of generating multiple types of network traffic, making it a versatile tool for traffic simulation and analysis. Specifically, based on its available menu options, TRAGEN can generate eight distinct categories of traffic that are sourced from various servers located in Europe and the United States. These categories include video traffic,

web traffic, mixed traffic from the United States, mixed traffic from Europe, as well as several specialized subsets: video traffic from the U.S., image traffic from the U.S., download traffic from the U.S., social media traffic from the U.S., and social media traffic from Europe. All of the selected traffic traces are derived from datasets provided by Akamai Technologies, one of the world's largest and most widely deployed content delivery network (CDN) providers. Consequently, the synthetic traffic generated by TRAGEN is designed to closely replicate the characteristics of real-world CDN traffic patterns, thereby ensuring the relevance and validity of simulation results for research and performance evaluation in network environments.

1.3 Contribution

This research investigates the ability of a synthetic traffic generator to accurately replicate network traffic patterns predicted by the Zipf model. We aim to analyze how closely the generated traffic aligns with the expected distribution and to identify any discrepancies that may affect the validity of the simulation. The evaluation includes statistical comparison, distribution fitting, and error measurement to assess the realism of synthetic traffic.

By exploring the gap between theoretical traffic models and synthetic traffic generation, this study contributes to the development of more accurate and reliable network simulation frameworks. The findings will be valuable for researchers working on network performance analysis, protocol design, and system optimization in diverse network architectures such as CDN, ICN, and IoT.

2 Related Works

This section provides an overview of prior research and developments relevant to the field of traffic generation and modeling. It highlights significant studies, methodologies, and technologies that have shaped this area, identifying both strengths and limitations in existing approaches.

The synthetic traffic model introduced by [1] was designed to support cache evaluation in Content Delivery Networks (CDNs), using real-world data from Akamai servers in the USA and Europe. While the model incorporates multiple traffic types, it lacks configurable request distributions, limiting its utility in accurately simulating network conditions in a controlled test-bed environment.

Breslau et al. [2] demonstrated that web request traffic exhibits a Zipf-like distribution, characterized by a content popularity skewness parameter typically ranging between 0.6 and 0.85. This distribution was further validated in the context of video traffic by Cha [3], and corroborated by subsequent studies such as [4], [5], and [6,7]. These findings underline the importance of aligning synthetic traffic generators with realistic demand distributions to ensure credible simulations.

Various synthetic traffic generators have been proposed since then. For example, ForTT-Gen [8] focuses on generating traffic for malware forensic analysis training. Similarly, Nguyen-An et al. [9] introduced an IoT traffic generator tailored for anomaly detection, while Shi et al. [10] proposed a trace-based approach to Internet traffic generation. More recent innovations include PAC-GPT [11], which leverages generative pre-trained trans-

formers to synthesize network traffic with high fidelity, and FPGA-based traffic generators designed for speed and extensibility [12].

Patil et al. introduced UTGen [13], a traffic generator classified based on its synthetic traffic generation mechanisms, and also provided a comprehensive survey of existing tools in the domain [14]. Additionally, Matoušek and Korček [15] developed a precise packet generator for both IPv4 and IPv6 networks using the NetCOPE platform. These tools reflect a growing diversity of approaches in synthetic traffic generation, addressing various network conditions, protocol types, and application scenarios.

Based on this collective body of work, it becomes evident that synthetic traffic generators must not only reproduce volume and timing characteristics but also replicate statistical properties—such as Zipf-like distributions—to enhance simulation realism and reliability.

3 Relevant Theory

In this section, we discuss the related theory that was used in this paper. The request model that models the request pattern by previous research is briefly described. Furthermore, the mean square error (MSE) is used to analyze the gap between theory and traffic generator.

3.1 Request Model

The distribution of requested content followed a Zipf distribution with skewed parameter α due to some studies reporting the request distribution of various types of digital content. Websites and user-generated videos followed the Zipf distribution as reported in [2] and [16]. Breslau et al. said that the request count of web-pages obeyed the Zipf distribution with a parameter α between 0.64 and 0.83 in [2]. Moreover, Mahanti et al. showed that it was between 0.74 and 0.84 in [16]. The request count of YouTube videos obeyed the Zipf distribution with a parameter α about 0.8 in [3]. Therefore, we assumed that α was between 0.2 and 1.5 in our test-bed to cover all possible request distribution regarding content popularity on the internet.

The Zipf distribution is defined as:

$$P(k;\alpha,N) = \frac{1/k^{\alpha}}{\sum_{n=1}^{N} 1/n^{\alpha}} \tag{1}$$

where:

- k is the rank of the element (e.g., 1, 2, 3, ...),
- α is the exponent characterizing the distribution,
- N is the total number of elements,
- The denominator is the generalized harmonic number up to N.

Figure 2 presents the probability density function (PDF) of the Zipf distribution, which is characterized by the content skewness parameter, α. This parameter, α, determines the shape of the distribution's tail. A higher α value indicates that a small number of content objects account for the majority of the access frequency, reflecting a highly skewed distribution. In contrast, a lower α value implies that access frequencies are more evenly distributed across a larger set of content objects, resulting in a flatter PDF curve.

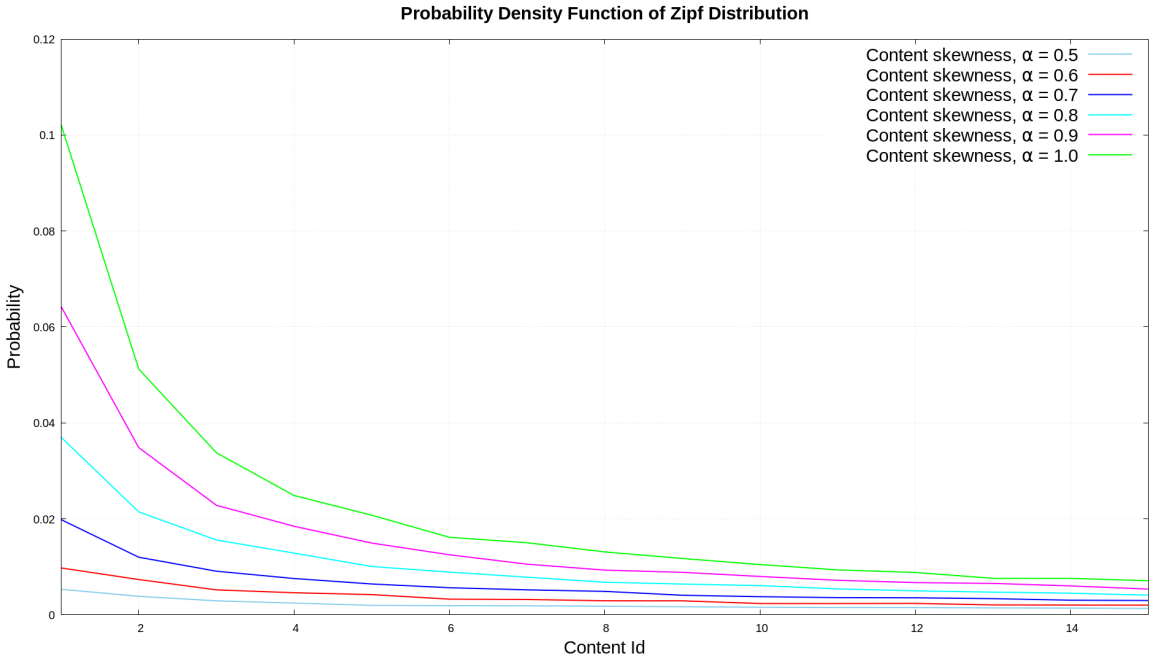

Figure 2: Zipf distribution.

3.2 Root Mean Squared Error

The **Root Mean Squared Error (RMSE)** is a commonly used metric to evaluate the accuracy of predictive models. It is defined as the square root of the average of the squared differences between the actual values y_i and the predicted values $\hat{y}_i$. Mathematically, it is given by:

The Root Mean Squared Error (RMSE) is defined as:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{2}$$

where:

- y_i is the actual value,
- $\hat{y}_i$ is the predicted value,
- n is the number of observations.

RMSE emphasizes larger errors due to the squaring of the differences, making it particularly useful when large deviations are more significant. The result is expressed in the same units as the target variable, which makes interpretation straightforward. A lower RMSE indicates better model performance, while a higher RMSE suggests poorer prediction accuracy. RMSE is widely used in regression analysis, time series forecasting, and other quantitative modeling contexts.

4 Numerical Evaluation

We conduct a simulation test-bed to compare the traffic characteristics generated by two distinct sources: TRAGEN, a synthetic traffic generator, and a Zipf-based traffic generator implemented in the Python language in the same environment. The objective was to evaluate the difference between these two traffic models in terms of how well they replicate real-world content request distributions. The Zipf generator was constructed using Python 3.7, based on the commonly accepted Zipf distribution, which is known to represent content popularity on various digital platforms such as websites and video-sharing services. The machine and software specifications used in this simulation are shown in Table 1.

Table 1: Hardware and software setup

Paramater	Value
Number of unique content	10000
Number of request	100000
Request rate	100 req/s
Request skewness (α)	0.2 - 1.5
CPU	Ryzen 7
OS	Ubuntu 24.04 LTS
Programming language	Python 3.7

Both generators were configured to produce interest packets asynchronously at a rate of approximately 100 packets per second, simulating high-throughput conditions. The Zipf generator produced traffic based on a skew parameter α, which was varied between 0.2 and 1.5, in line with previous studies. For example, Breslau et al. [2] and Mahanti et al. [16] found α values ranging from 0.64 to 0.84 for web traffic, while α values around 0.8 have been observed in YouTube traffic [3].

To assess the accuracy and similarity of TRAGEN-generated traffic in comparison to the theoretical Zipf distribution, we collected and analyzed packet request frequency data from both generators. Each simulation run captured approximately one hundreds thousand packets, and the experiment was repeated ten times at random intervals to ensure statistical robustness. From the collected data logs, we extracted the request frequencies and compared them using the Root Mean Squared Error (RMSE) metric. This allowed us to quantitatively evaluate the deviation between the synthetic traffic produced by TRAGEN and the expected distribution from the Zipf model.

Through this simulation-based evaluation, we aim to highlight the effectiveness and limitations of TRAGEN in emulating realistic traffic patterns observed in modern networks.

5 Results and Discussion

As shown in Figure 3, the distribution patterns generated by Tragen for the nine different traffic types generally follow the Zipf distribution, although the degree of tail skewness varies across traffic categories. For instance, the mixed traffic from the United States and Europe exhibits similar distributional characteristics, suggesting comparable access patterns in these regions. This observation highlights the ability of Tragen to realistically emulate distinct traffic behaviors that reflect real-world content popularity distributions.

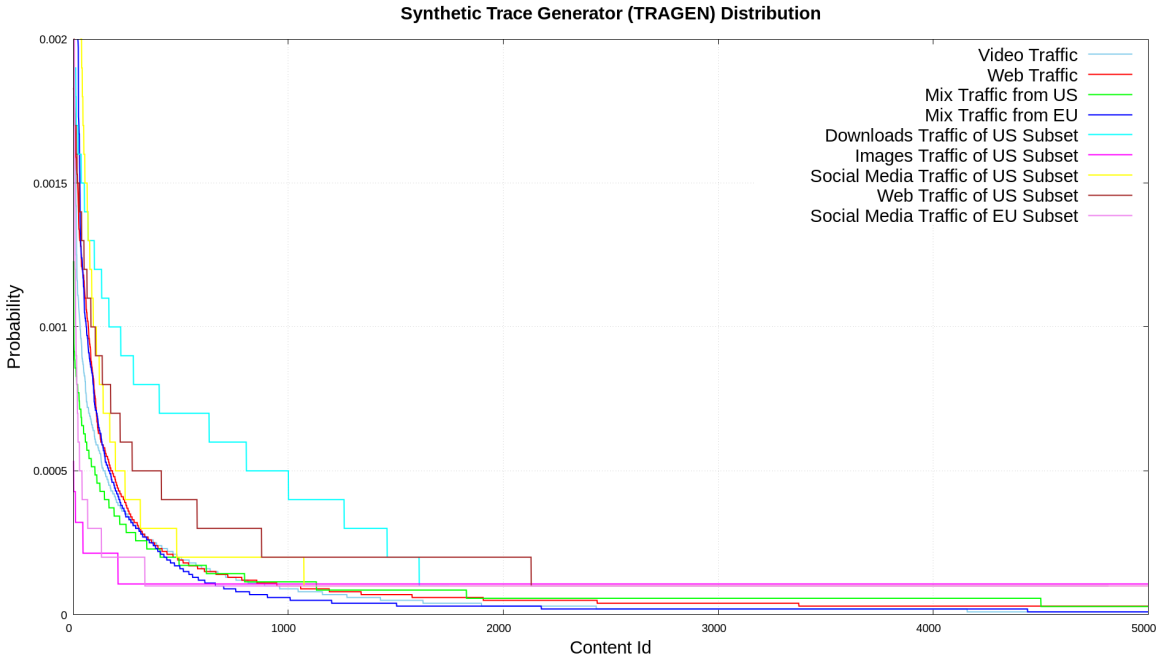

Figure 3: Tragen traffic distribution.

To further assess the accuracy of the simulated traffic and to quantify the degree of fit between the empirical data and the theoretical Zipf distribution, we apply the Root Mean Square Error (RMSE) as an evaluation metric. The RMSE enables us to systematically identify the skewness parameter, α, that best aligns with the theoretical model proposed in prior studies. By employing this approach, we can rigorously evaluate how closely the generated traffic traces approximate the expected statistical properties, thereby enhancing the validity of subsequent simulation and analysis efforts.

Figure 4 displays the relationship between Content Skewness, α and RMSE Values across various traffic types. As the content skewness increases, the RMSE values rise for all traffic types, indicating that larger content skewness leads to higher prediction errors. The Video Traffic (cyan) and Web Traffic (red) exhibit the lowest RMSE values, suggesting that these traffic types are less affected by changes in content skewness. In contrast, traffic types like Social Media Traffic of US Subset (orange) and Social Media Traffic of EU Subset (purple) show the highest RMSE values, meaning they experience larger prediction errors as skewness increases. Other traffic types, such as Mix Traffic from EU (blue), fall somewhere in between, highlighting that different traffic categories exhibit varying sensitivities to content skewness.

Moreover, the results indicate that 75% of the mixed traffic and specific content types, including video and web traffic, are concentrated around an α value of approximately 0.6. In contrast, the remaining traffic types, particularly those from regional subsets, exhibit content skewness α values that deviate significantly from the 0.6 to 0.8 range. This disparity is clearly described in Table 2, which highlights the differences in content skewness across various traffic types.

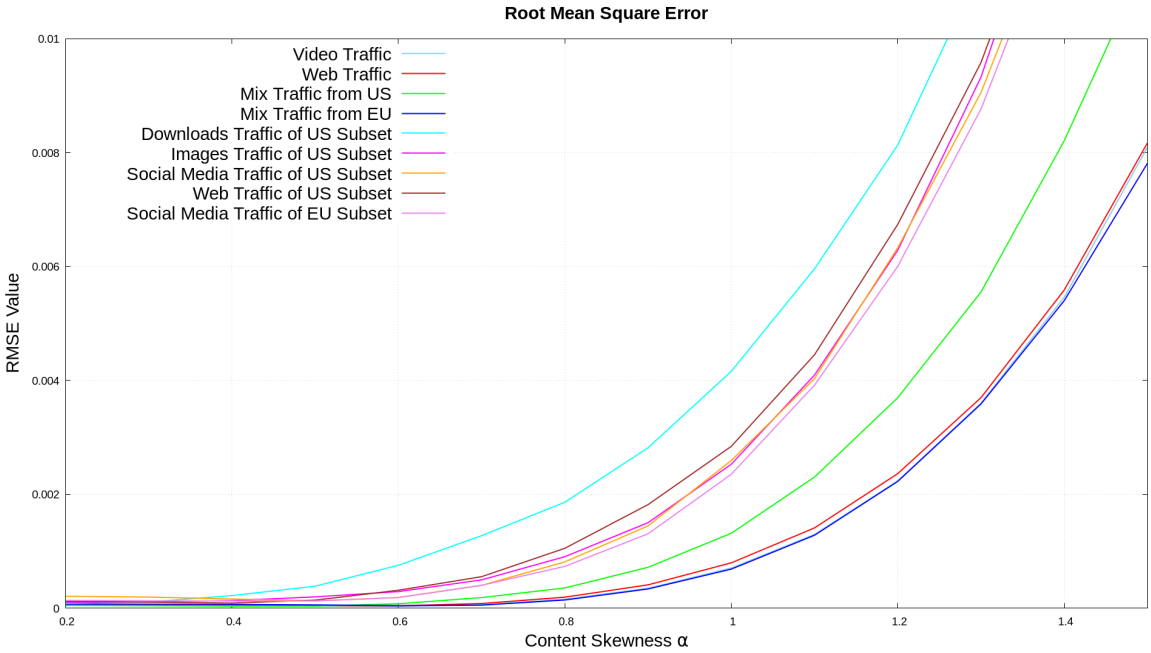

Figure 4: Root mean square error between tragen and Zipf distribution.

This trend suggests that the errors in predicting these traffic types are influenced by the degree of content skewness, with some types being more susceptible to changes. Traffic categories with higher RMSE values, particularly the social media-related types, may require additional optimizations in data processing or modeling techniques to mitigate these prediction errors. The graph offers valuable insights into how content characteristics, like skewness, can impact the accuracy of traffic predictions, potentially guiding improvements in forecasting models for these traffic types.

Table 2: The lowest RMSE value with standard deviation

Traffic type	α	RMSE Value $\pm$ Std Dev
video	0.6	$(3.02 \pm 0.11) \times 10^{-5}$
web	0.6	$(4.06 \pm 0.14) \times 10^{-5}$
Mix traffic of US	0.4	$(3.44 \pm 0.12) \times 10^{-5}$
Mix traffic of EU	0.6	$(4.41 \pm 0.15) \times 10^{-5}$
Downloads Traffic of US Subset	0.2	$(8.60 \pm 0.30) \times 10^{-5}$
Images Traffic of US Subset	0.2	$(1.04 \pm 0.04) \times 10^{-4}$
Social Media Traffic of US Subset	0.5	$(1.27 \pm 0.04) \times 10^{-4}$
Web Traffic of US Subset	0.4	$(8.95 \pm 0.31) \times 10^{-5}$
Social Media Traffic of EU Subset	0.4	$(1.24 \pm 0.04) \times 10^{-4}$

6 Conclusion

The above discussion highlights how content skewness, as represented by the Content Skewness, influences the RMSE values across various traffic types. Previous studies have demonstrated that internet traffic often follows a Zeta-Indexed Power Law (ZIPF) distribution, with α values typically ranging from 0.6 to 0.8. This range is consistent with the trend observed in the graph, where traffic types with lower α values (closer to 0.6) generally exhibit lower RMSE values, indicating more predictable traffic patterns. In contrast, higher α values, which reflect more skewed content distributions, correspond to a marked increase in prediction errors, especially for traffic types such as Social Media Traffic. These findings underscore the significance of understanding the underlying distribution of internet traffic when developing prediction models. By optimizing models within the typical α range observed in internet traffic (0.6 to 0.8), it may be possible to reduce prediction errors and enhance accuracy, particularly for traffic types exhibiting higher skewness.

Acknowledgments

This work was supported by Informatics Engineering Department of Syarif Hidayatullah State Islamic University Jakarta.

References

[1] A. Sabnis and R. K. Sitaraman, "TRAGEN: a synthetic trace generator for realistic cache simulations," in *Proceedings of the 21st ACM Internet Measurement Conference*, (Virtual Event), pp. 366–379, ACM, Nov. 2021.

[2] L. Breslau, P. Cao, L. Fan, G. Phillips, and S. Shenker, "Web caching and Zipf-like distributions: evidence and implications," in *IEEE INFOCOM '99. Conference on Computer Communications. Proceedings. Eighteenth Annual Joint Conference of the IEEE Computer and Communications Societies. The Future is Now (Cat. No.99CH36320)*, vol. 1, pp. 126–134 vol.1, Mar. 1999. ISSN: 0743-166X.

[3] M. Cha, H. Kwak, P. Rodriguez, Y.-Y. Ahn, and S. Moon, "Analyzing the Video Popularity Characteristics of Large-Scale User Generated Content Systems," *IEEE/ACM Transactions on Networking*, vol. 17, pp. 1357–1370, Oct. 2009.

[4] Y. Gu, L. Chen, and K.-M. Tang, "A Load Balancing Method under Zipf-Like Requests Distribution in DHT-Based P2P Network Systems," in *2009 International Conference on Web Information Systems and Mining*, pp. 656–660, Nov. 2009.

[5] T. Huang, K. Xu, and R. Pi, "A popularity-based neighbor selection model in P2P file-sharing system," in *2010 3rd IEEE International Conference on Broadband Network and Multimedia Technology (IC-BNMT)*, pp. 975–979, Oct. 2010.

[6] V. Sourlas, P. Flegkas, P. Georgatsos, and L. Tassiulas, "Cache-aware traffic engineering in Information-Centric Networks," in *2014 IEEE 19th International Workshop on Computer Aided Modeling and Design of Communication Links and Networks (CAMAD)*, (Athens, Greece), pp. 295–299, IEEE, Dec. 2014.

[7] V. Sourlas, L. Gkatzikis, P. Flegkas, and L. Tassiulas, "Distributed Cache Management in Information-Centric Networks," *IEEE Transactions on Network and Service Management*, vol. 10, pp. 286–299, Sept. 2013.

[8] J. V. Bistene, C. E. Das Chagas, A. F. Pereira Dos Santos, G. M. De Souza Dias, and R. M. Salles, "ForTT-Gen: Network Traffic Generator for Malware Forensics Analysis Training," in *2024 12th International Symposium on Digital Forensics and Security (ISDFS)*, (San Antonio, TX, USA), pp. 1–6, IEEE, Apr. 2024.

[9] H. Nguyen-An, T. Silverston, T. Yamazaki, and T. Miyoshi, "Generating IoT traffic: A Case Study on Anomaly Detection," in *2020 IEEE International Symposium on Local and Metropolitan Area Networks (LANMAN*, (Orlando, FL, USA), pp. 1–6, IEEE, July 2020.

[10] W. Shi, M. MacGregor, and P. Gburzynski, "Synthetic trace generation for the Internet," in *Proceedings of the Fourth Annual IEEE International Workshop on Workload Characterization. WWC-4 (Cat. No.01EX538)*, (Austin, TX, USA), pp. 169–174, IEEE, 2001.

[11] D. K. Kholgh and P. Kostakos, "PAC-GPT: A Novel Approach to Generating Synthetic Network Traffic With GPT-3," *IEEE Access*, vol. 11, pp. 114936–114951, 2023.

[12] D. Yuan, W. Yi, H. Hu, and X. Shi, "A fast, affordable and extensible FPGA-based synthetic Ethernet traffic generator for network evaluation," in *2017 3rd IEEE International Conference on Computer and Communications (ICCC)*, (Chengdu), pp. 1036–1040, IEEE, Dec. 2017.

[13] A. G. Patil, A. Surve, and A. K. Gupta, "Classification of UTGen synthetic traffic generator," in *2016 Conference on Advances in Signal Processing (CASP)*, (Pune, India), pp. 280–285, IEEE, June 2016.

[14] A. G. Patil, A. R. Surve, A. K. Gupta, A. Sharma, and S. Anmulwar, "Survey of synthetic traffic generators," in *2016 International Conference on Inventive Computation Technologies (ICICT)*, (Coimbatore, India), pp. 1–3, IEEE, Aug. 2016.

[15] J. Matoušek and P. Korček, "Precise IPv4/IPv6 packet generator based on NetCOPE platform," in *14th IEEE International Symposium on Design and Diagnostics of Electronic Circuits and Systems*, pp. 319–324, Apr. 2011.

[16] A. Mahanti, C. Williamson, and D. Eager, "Traffic analysis of a Web proxy caching hierarchy," *IEEE Network*, vol. 14, pp. 16–23, June 2000.