electronics

MDPI

Article

AliasClassifier: A High-Performance Router Alias Classifier

Yuancheng Xie †, Zhaoxin Zhang *, Enhao Chen † and Ning Li

Department of Computer Science, Institute of Technology, Harbin 150028, China; 20b903060@stu.hit.edu.cn (Y.X.); 2200400308@stu.hit.edu.cn (E.C.)

* Correspondence: zhangzhaoxin@hit.edu.cn

† These authors contributed equally to this work.

Abstract: The task of router alias resolution for IPv4 networks presents a formidable challenge in the realm of router-level topology inference. Despite the considerable potential exhibited by machine-learning-based alias-resolution methods for IPv4 networks, several constraints impede their effectiveness. These constraints include a low discovery rate of aliased IPs, a failure to account for router aggregation, and a dearth of valid features in current schemes. In this study, we introduce a novel alias resolver, AliasClassifier, which is based on the Random Forest model and the alias triangulation algorithm. This innovative model identifies the key six features from a set of four prevalent routing behaviors that are typically employed to distinguish aliased IPs from non-alienated IPs. Subsequently, the AliasClassifier aggregates aliased IP pairs into routers using an alias triangulation algorithm. Experimental results demonstrate that AliasClassifier excels in discovering aliased IPs in IPv4 networks, boasting a resolution accuracy as high as 94.8% and a recall rate of 40.4%. Its comprehensive performance significantly surpasses that of state-of-the-art alias resolvers such as TreeNET, MLAR, and APPLE. Furthermore, as a typical centralized alias parser, AliasClassifier's deployment cost is remarkably low. Consequently, AliasClassifier emerges as an ideal tool for router alias resolution in large-scale IPv4 networks.

Keywords: alias resolution; alias aggregation; router; machine-learning models

Citation: Xie, Y.; Zhang, Z.; Chen, E.; Li, N. AliasClassifier: A High-Performance Router Alias Classifier. *Electronics* **2024**, *13*, 1747. https://doi.org/10.3390/electronics13091747

Academic Editor: Aryya Gangopadhyay

Received: 9 April 2024
Revised: 23 April 2024
Accepted: 26 April 2024
Published: 1 May 2024

1. Introduction

The intricate topology of routers forms a fundamental cornerstone in the landscape of the Internet, holding substantial relevance for research areas such as network performance evaluation, network security, and link reliability analysis [1–4]. Despite the ongoing data-sharing efforts by several renowned international organizations, a comprehensive and representative router-level topology map of the current Internet remains elusive [5–8]. This is largely due to the inherent complexity of Internet links and the spontaneous requirements of network operators to safeguard commercial privacy. Consequently, researchers are often left to infer the topology of routers by gathering fragmented routing information scattered across the Internet.

The issue of router alias resolution presents a formidable challenge in the acquisition of router topology maps. Router alias resolution, a technique employed to identify the IP addresses of multiple interfaces on the same router, aims to transform the logical topology map of the IP layer—obtained through methods such as traceroute—into a physical topology map at the router layer [9,10]. Within the realm of IPv4 networks, a variety of router alias-resolution techniques have emerged. These techniques can be broadly categorized into three types: measurement-based, inference-based, and machine-learning-based router alias-resolution technologies [11].

Measurement-based techniques typically demonstrate high accuracy as they employ active detection (e.g., ping or traceroute) to acquire precise fingerprint information [6,12,13]. In contrast, inference-based technology necessitates only a logical analysis of the network topology to yield parsing results, rendering it particularly suitable for alias resolution in

large-scale networks [9,10,14]. However, these technologies often resort to distributed detection methods for data collection, inevitably escalating deployment costs and diminishing parsing efficiency. Furthermore, due to the absence of a mature network topology, inference-based technologies can easily result in relatively low accuracy. Certain fingerprint information, such as recorded route [15], timestamp [16,17], and IPID [2], are readily restricted by routers and challenging to acquire. While proposals have been made to amalgamate multiple alias-resolution tools and design comprehensive inference methods to reduce measurement complexity and enhance measurement accuracy, these methods are typically complex to implement [18,19]. In the absence of global prior knowledge, these solutions cannot significantly improve parsing efficiency and the alias IP discovery rate. The alias IP discovery rate, a key metric for evaluating the effectiveness of alias-resolution technology, refers to the proportion of IP addresses in the network that can be resolved by aliases. Evidently, a common challenge for router alias-resolution techniques in IPv4 networks is striking a delicate balance between various critical factors, including the discovery rate of alias IPs, the accuracy of parsing results, detection efficiency, and deployment costs.

In the wake of the burgeoning application of machine learning, a variety of machine-learning approaches have been introduced to tackle the router alias-resolution problem [20–22]. Machine-learning-based methods can effectively amalgamate topological data and fingerprint information obtained via various probes during the alias-resolution process. These methods extract multidimensional features and combine them with complex logical reasoning, therefore enhancing the efficiency and accuracy of alias resolution. This integration and reasoning process endows machine-learning-based solutions with superior generalization performance compared to traditional solutions. Consequently, these solutions exhibit a higher discovery rate of alias IPs and yield a completer and more accurate parsed router topology.

However, current machine-learning-based solutions often overlook the challenges associated with data acquisition. This oversight directly impacts the integrity of the features that can be acquired, resulting in a discovery rate of alias IPs that falls short of theoretical potential. In addition, most inference results produced by machine-learning models are alias IP pairs rather than routers. This simplistic aggregation approach may lead to an overabundance of irrelevant IP addresses in the inferred routers, a phenomenon referred to as "router bloat". The occurrence of "router bloat" underscores the need for further refinement of these methods, highlighting the ongoing challenges in the field of router alias resolution.

To address the aforementioned challenges, we trained an alias resolver, AliasClassifier, based on a machine-learning model. AliasClassifier employs a Random Forest classifier to identify potential alias IP pairs and subsequently aggregates these pairs using the router aggregation algorithm of alias triangulation. This approach enables efficient and accurate router alias resolution for IPv4 networks. AliasClassifier, a typically centralized alias resolver, is relatively economical to deploy and selects features that are readily collectible from routing data. Consequently, AliasClassifier not only ensures notable accuracy but also maintains high efficiency, establishing it as an effective and precise tool for resolving aliases in large IPv4 networks. The primary contributions of this article are as follows:

- We conducted an empirical analysis of four widely recognized routing hypothesis behaviors using the data-cleaned ITDK dataset. Our analysis identified six features that can effectively distinguish between aliased and non-aliased IPs. We define IPs located on different interfaces on the same router as 'alias IPs' and IPs located on different routers as 'non-aliased IPs'.
- Based on these six features, we selected an appropriate alias IP classifier through multiple sets of experiments. Ultimately, we chose to use the Random Forest model, which exhibited the most balanced performance, to construct the alias resolver AliasClassifier.
- To address the alias aggregation problem, we proposed an innovative alias aggregation algorithm based on alias triangulation. This method significantly enhances the

credibility of router alias resolution by making a secondary judgment on the alias IP following alias classification.

- The experimental results underscore the superior comprehensive performance of AliasClassifier, which markedly surpasses similar advanced classifiers such as APPLE, MLAR, and TreeNET. A comparative analysis with TreeNET revealed that AliasClassifier's alias IP discovery rate increased by 2.4 times, and its parsing efficiency improved by 4.5 times.

The remainder of this paper is meticulously structured as follows: Section 2 serves as a foundational background section, offering a comprehensive review of the work conducted on router alias resolution, specifically in relation to IPv4 networks. Section 3 delves into an empirical analysis of the four hypothesized behaviors that constitute the crux of this study. Section 4 is dedicated to the construction of the alias parser, AliasClassifier, elucidating its design and implementation details. Section 5 presents an extensive experimental evaluation of AliasClassifier's performance, providing a rigorous assessment of its efficacy and robustness. In Section 6, we delve into the constraints of AliasClassifier and contemplate potential enhancements for future iterations. Finally, Section 7 draws together the main findings and contributions of this research, offering a succinct summary and highlighting the implications of our work.

2. Background

In this section, we first embark on a thorough exploration of the existing body of work pertinent to router alias resolution for IPv4 networks. Subsequently, we introduce the dataset employed in our experiments. This dataset, carefully curated and rigorously vetted, forms the empirical backbone of our research, providing the raw data from which our insights and conclusions. Lastly, We conclude this section with an in-depth discussion of seven key hypothetical behaviors that hold potential utility in the realm of router alias parsing work. The objective of this discussion is to analyze the generality of the relevant behaviors and to reduce the number of features characterized by a high incidence of missing values.

2.1. Related Work

Router aliasing techniques of IPv4 networks can be systematically classified into three principal categories [11]: measurement-based aliasing methods, inference-based aliasing methods, and machine-learning-based aliasing methods.

Measurement-based methods involve sending probe packets to IP pairs that may have aliasing relationships. The correlation patterns of the corresponding fingerprint information in the response packets are analyzed to determine whether these IPs belong to the same router. Key algorithms in this category include Ally [6], Passenger [15], DisCarte [23], RadarGun [12], Timestamp [16], MIDAR [13], Pamplona-traceroute [2], Pythia [17] among others. The latest development in this field is an alias-resolution method based on delay sequence analysis [24], proposed by Yang Tao et al., which leverages the similarity of delay sequences for alias resolution. These measurement-based alias-resolution algorithms have gained widespread use due to the high accuracy of their resolution results. However, in scenarios with slow networks and many unresponsive routers, these methods can be time-consuming, rendering these methods unsuitable for large IPv4 networks.

In contrast, inference-based alias-resolution methods leverage the contents of IP topology, subnets, or domain names. They infer whether two IPs are aliased based on the connectivity or naming rules of the IP addresses. Key algorithms in this category include DASAR [25], AAR [14], APAR [10], Kapar [9], among others. Recently, Alexander Marder proposed an aliasing technique, APPLE [26], which compares the length of the reply path from each IP address to a set of distributed VPs for router aliasing. This method circumvents the dependence on router manufacturer and operating system-specific IP implementations. Overall, inference-based alias-resolution methods are suitable for large-scale IPv4 net-

works. However, their accuracy is highly dependent on network topology, resulting in lower accuracy.

In the dynamic and rapidly progressing field of IPv4 network topology, research endeavors focusing on measurement-based and inference-based router alias-resolution algorithms have been consistently deepening. A suite of comprehensive alias detection algorithms, such as Palmtree [18] and TreeNet [19], among others, have been proposed to mitigate measurement complexity and augment measurement accuracy. These proposed methodologies typically represent a fusion of multiple alias-resolution tools, therefore rendering their implementation process highly intricate. For instance, TreeNet imposes a stringent stipulation that the path hops of all IP nodes should not exceed a single hop. Concurrently, it necessitates the collection of a diverse array of alias fingerprints, predominantly encompassing TTL values, source addresses of port unreachable packets, IPIDs, DNS resolution results, and timestamp information. The TreeNet algorithm can be perceived as an amalgamation of the algorithms of iffinder [27], Kapar [9], and others. Iffinder is an alias-resolution algorithm based on homologous addresses. It realizes alias resolution by looking for the source address in the returned "port unreachable" ICMP message that is different from the destination address of the probe message sent.

Undeniably, methods that amalgamate multiple fingerprinting information have carved out a significant niche in the domain of router alias resolution for IPv4 networks. Despite the convoluted nature of the implementation process, comprehensive alias detection algorithms deliver exceptional precision and accuracy. However, in the absence of global a priori knowledge, these schemes do not significantly enhance the parsing efficiency and alias IP discovery rate. A pervasive challenge is the lack of equilibrium among key factors such as the accuracy of resolution results, the discovery rate of aliased IPs, detection efficiency, and the feasibility of deployment in traditional router alias-resolution techniques. Consequently, numerous innovative approaches, despite their theoretical promise, underperform when tasked with router alias resolution on large IPv4 networks.

In recent years, the ascendancy of artificial intelligence and machine learning has catalyzed the integration of these technologies into various router alias-resolution methodologies. Machine-learning-based alias-resolution algorithms mainly include AliasCluster [20], MLAR [21], Limited Ltd [22], and so on. The MLAR algorithm introduces four-dimensional features to reframe the alias-resolution problem as a classification challenge [21]. However, the method does not grapple with the unavailability and generalizability of feature data, leading to subpar parsing accuracy. Limited Ltd employs the characteristic of ICMP rate limiting as a pivotal feature for router alias resolution. This technique involves the dispatch of ICMP probes to target interfaces, instigating the router's ICMP rate-limiting mechanism. While ICMP rate limiting can be more ubiquitously obtained in comparison to other fingerprinting information, this method is not without its challenges. It is particularly susceptible to interference from the network environment, which can compromise the integrity of the results. Frequent initiation of large numbers of packets for ICMP rate-limit probing increases interference with the Internet and raises potential ethical concerns.

The advent of machine-learning methods has ushered in a new era for alias resolution. These methods enable the utilization of multiple probe data and fingerprint information in the alias-resolution process, as well as the adoption of more intricate logical reasoning, therefore enhancing both the efficiency and accuracy of alias-resolution. Consequently, the optimization of machine-learning methods to improve the outcomes of router alias resolution for IPv4 networks has emerged as a novel research direction. However, current machine-learning-based alias-resolution solutions often overlook the challenges associated with data collection. This oversight directly contributes to the low discovery rate of aliased IPs for these solutions. Moreover, most inferences made through machine learning are IP pairs rather than router aggregations. The absence of effective alias aggregation operations may lead to inferred routers containing an excessive number of IP addresses. This phenomenon underscores the need for further refinement of these methods and highlights the ongoing challenges in the field of router alias resolution.

2.2. Datasets

In our research, we faced a challenge of lacking fundamental data. To overcome this, we turned to the ITDK dataset, generously provided by CAIDA (the Cooperative Association for Internet Data Analysis) [28]. ITDK releases were produced from traceroutes conducted on the Archipelago (Ark) measurement infrastructure. The IPv4 router-level topology is derived from aliases resolved with MIDAR [13] and iffinder [27], which yield the highest confidence aliases with very low false positives. The router-level topology is provided in two files, one giving the nodes and another giving the links. There are additional files that assign ASes to each node, provide the geographic location of each node, and provide the DNS name of each observed interface.

It is important to note that while the ITDK dataset boasts a commendably low false positive rate, the absolute accuracy of the data cannot be entirely guaranteed. To address this, we implemented the state filter to bolster the construction of our ground truth collection for aliased IPs. IP addresses located on different interfaces on the same router should theoretically remain consistently line or offline. Our state filter was designed to leverage the online/offline state of routers. If a pair of aliased IPs originate from the ITDK, one IP is pingable, and the other is not. We consider the pair of IPs not to be aliased IPs on the same router. This filter was designed to enhance the reliability of the data and mitigate potential inaccuracies inherent in any large-scale dataset.

In our research, we initiated the process by extracting a total of 883,636 pairs of IPv4 aliases from the CAIDA ITDK dataset, focusing on data generated in February 2022. Subsequently, by applying the state filter, we were able to eliminate a substantial number of non-alias IP pairs, specifically 126,343 pairs, from our dataset. This filter played a pivotal role in enhancing the reliability and relevance of the data used in our research. At the same time, we took the initiative to build a collection of non-alias IPs to enhance the effective distinction between aliased IPs and non-alias IPs, drawing insights from our ground truth collection. Given that interface IPs of different routers lack aliasing relationships, we curated a non-alias IPs dataset by pairing interface IPs from distinct routing nodes. This meticulous curation process yielded a dataset comprising 757,293 pairs of aliased IPs and 909,233 pairs of non-aliased IPs, representing a comprehensive and diverse set of IPs distributed across 218 countries and regions globally, as outlined in Table 1. This rigorous data refinement process ensured the integrity and quality of the dataset used for our research.

Table 1. Ground truth data set.

	CAIDA ITDK	**Ground Truth**
Interface IPs	344,860	322,668
IP Pairs	883,636	757,293
IP pairs filtered	-	126,343

2.3. Hypothetical Behavior

Our approach begins by establishing fundamental behavioral assumptions regarding aliased IPs that coexist on the same router. In our approach, the process of alias resolution for a pair of IPs is conceptualized as a binary classification problem. Our overarching goal is to develop a trained classifier capable of effectively solving the alias-resolution challenge for routers within a network. To achieve this, it is imperative to identify features that can reliably differentiate between aliased IPs and non-aliased IPs. Equally important is the requirement that these features should be universally accessible from routers.

Drawing from a comprehensive review of prior research in the field, we have identified seven key hypothetical behaviors that have shown promise in distinguishing between aliased and non-aliased IPs. These behaviors encompass aspects such as routing path, round-trip delay, Reply TTL, domain name information, port information, IPID information, and IP spatial difference information. To further inform our methodology, we have

tabulated the probability of occurrence of these seven hypothetical behaviors within the ground truth dataset of aliased IPs, as depicted in Table 2.

Table 2. Occurrence of hypothetical behaviors.

Behavior	Frequency
Routing path	65%
Round-trip delay	65%
Reply TTL	86%
Domain Information	39%
Port information	0.14%
IPID	13%
IP Spatial Discrepancy Information	100%

In our investigation, the ITDK dataset, as of February 2022, was employed as the primary source of data. Although we ran multiple active probes in the same month, the change in network topology caused us to discover only 65% of the pingable IPs. Our analytical efforts yielded the identification of a mere four hypothetical behaviors characterized by high prevalence. The probability of the port information, IPID Information, and domain information are small, 0.14%, 13%, and 39%, respectively. Given the practical challenges associated with collecting data pertaining to these non-universal attributes and their limited influence on parsing outcomes, we have chosen to exclude feature extraction from port information, IPID information, and domain information. The four remaining hypothetical behaviors have been previously validated in the extant literature. Our intention is to conduct an in-depth analysis of these four hypothetical behaviors in the context of a ground truth dataset and, subsequently, to derive relevant features for the training of alias-resolution classifiers. This approach is underpinned by the robustness and reliability demonstrated by these behaviors in prior research endeavors.

3. Characterization Analysis

By evaluating the likelihood of these hypothetical behaviors, we discern four specific behaviors that hold universal applicability. We undertake a meticulous instance analysis of these four hypothetical behaviors, leveraging both the set of alias IPs and the set of non-alias IPs. This analysis is conducted with the aim of extracting behavioral features that prove to be instrumental in the process of alias resolution. Furthermore, we elucidate the methodologies employed for capturing and representing pertinent behavioral data. These steps collectively constitute a critical foundation for the development of our alias-resolution framework.

The specific details of these relevant features, which were utilized to train the classifier, are comprehensively outlined in Table 3. We are committed to making the feature dataset and our research results publicly available to facilitate replication and further research in this domain.

Table 3. List of features selected by the classifier.

Feature Name	Flag Characters
Difference Value of round-trip time	RTT_{DV}
Difference Value of Path Length	PL_{DV}
Difference Value of Path Direction	PD_{DV}
Path Similarity Coefficient	PSC
Difference Value of Reply TTL	TTL_{DV}
Spatial Distance of the IP pair	SD_{IP}

3.1. Round-Trip Delay Information

Round-trip delay information (RTT) is a crucial metric representing the time required for a packet to complete a round-trip journey from the source IP address to the destination IP address. Typically, when network performance is stable and congestion is minimal, RTT exhibits a positive correlation with geographic distance [29]. Notably, two IP addresses residing on the same router should theoretically exhibit closely aligned RTT measurements when observed from a common source IP. Leveraging this understanding of RTT behavior, we introduce a novel feature aimed at distinguishing aliased IPs from non-aliased IPs: the Difference Value of round-trip time denoted as RTT_{DV}.

Feature 1: Difference Value of round-trip time. For any two IP addresses, the round-trip delay from the same source IP is $rtt = [t_1, t_2]$, and the Difference Value of round-trip time feature RTT_{DV} between these two IP addresses is defined as:

$$RTT_{DV} = |t_1 - t_2| \tag{1}$$

For any two target IPs, we measure the likelihood of two IPs being aliased to each other by comparing the RTT_{DV} of the two IP addresses. We count the cumulative distribution curves of RTT_{DV} in milliseconds between aliased IP pairs as well as non-alias IP pairs, as shown in Figure 1. Our analysis has revealed a significant distinction between aliased IPs and non-aliased IPs based on RTT_{DV}. When RTT_{DV} is set at 25, the cumulative proportion of aliased IPs reaches approximately 80%, while for non-aliased IPs, it stands at a notably lower 31.1%. This substantial difference of 48.9% underscores the efficacy of RTT_{DV} as a discriminative feature.

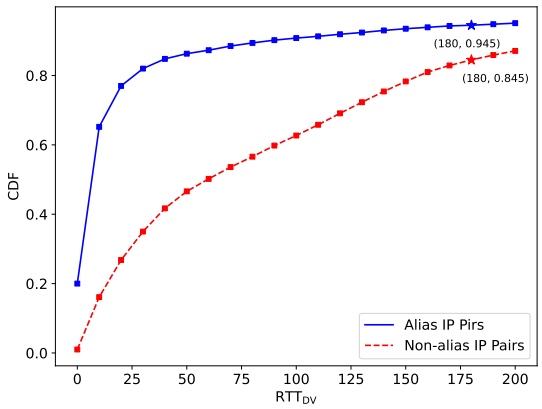

Figure 1. RTT_{DV} of alias IPs and non-alias IPs.

Intriguingly, as RTT_{DV} increases, the cumulative disparity between aliased IPs and non-alias IPs begins to diminish. For instance, when $RTT_{DV} = 180$, the cumulative difference between alias IP pairs and non-alias IP pairs in RTT_{DV} contracts to a mere 10%, a stark contrast to the cumulative difference of 48.9% observed when $RTT_{DV} = 25$. These observations suggest that relying solely on one feature, such as RTT_{DV}, may not suffice to draw a comprehensive distinction between aliased and non-alienated IPs. This underscores the complexity of the alias-resolution problem and highlights the need for a multi-faceted approach that considers a broader set of features.

3.2. Routing Path

The primary function of a router is to determine the optimal transmission path for individual packets based on a routing table. The routing table, which encapsulates the data forwarding policy, typically remains static for extended durations. Consequently, the routing trajectory of a packet, originating from a source IP and destined for a target IP, is theoretically relatively constant over short intervals [30].

To substantiate the stability of routing paths, we conducted an empirical study involving the random selection of 10,000 IP addresses. Our approach encompassed multiple iterations of traceroute operations from an identical probing point to the same destination IP. We subsequently compared the outcomes of these traceroute rounds and computed the Path Edit Distance (PED) [31]. The Path Edit Distance is the minimum number of editing operations required to change from one to the other between two routing paths. In general, the smaller the Path Edit Distance, the more similar the two routing paths are.

Figure 2 offers a cumulative distribution curve of path edit distances for the aforementioned 10,000 IP addresses. Notably, this visual representation reveals that approximately 92% of the IP addresses exhibit path edit distances of 2 hops or less. This empirical observation firmly establishes the relative stability of routing paths from source IPs to destination IPs over time. This stability observation suggests that IPs characterized as aliases residing on the same router should theoretically exhibit closely aligned routing paths.

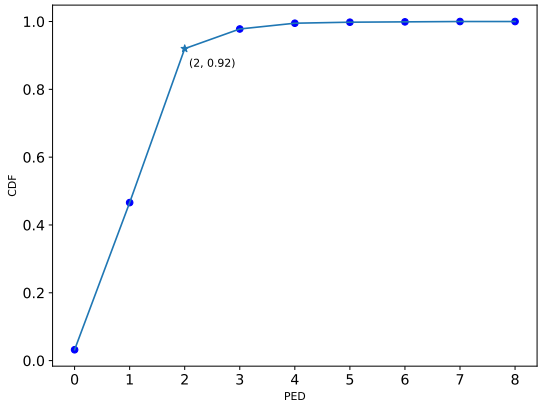

Figure 2. Stability analysis of routing paths.

Feature 2: Difference Value of Path Length. For a given pair of IP addresses, each characterized by path lengths denoted as ℓ_1 and ℓ_2, respectively, the Difference Value of Path Length feature PL_{DV} is represented as follows:

$$PL_{DV} = |\ell_1 - \ell_2| \tag{2}$$

Feature 3: Difference Value of Path Direction. For any given pair of IP addresses, each associated with routing paths p_1 and p_2, where the number of transformations required to align p_1 with p_2 is represented as $trns$. Editing operations allowed by $trns$ include replacing, inserting, and deleting characters. The Difference Value of Path Direction feature (PD_{DV}) is denoted as follows:

$$PD_{DV} = trns(p_1, p_2) \tag{3}$$

We introduce two features in our methodology: the Path Length Difference Feature (PL_{DV}) described by Equation (2) and the Path Direction Difference Feature (PD_{DV}) outlined by Equation (3). The concept of Path Direction Difference (PD_{DV}) primarily encapsulates the count of IP nodes that diverge between two routing paths. It is noteworthy that even when two routing paths share an identical length, their path directions may not necessarily align. For instance, consider two paths: $\{a, b, c, d\}$ and $\{a, f, x, d\}$. Despite having a Path Length Difference (PL_{DV}) of zero, indicating identical path lengths, their Path Direction Difference (PD_{DV}) stands at 2. This discrepancy underscores the fact that identical path lengths do not guarantee congruent path directions, therefore highlighting the nuanced complexity inherent in the analysis of network routing paths. These metrics enable us to discern and characterize the distinctions between aliased and non-aliased IPs based on the similarity within their respective routing paths.

For both aliased and non-aliased IP pairs, we conduct data statistics, focusing on the Path Length Difference and Path Direction Difference. These statistical analyses are depicted in Figure 3. The analysis reveals that the disparity between alias IPs and non-alias IPs based on the PL_{DV} is relatively modest. The most substantial cumulative proportion difference observed between alias IPs and non-alias IPs in PL_{DV} features is merely 32.8%. In contrast, the distinction between aliased IPs and non-aliased IPs, as determined by the PD_{DV}, is striking. During this assessment, the most substantial cumulative proportion difference observed between alias IPs and non-alias IPs in PL_{DV} features is 56.9%, where the cumulative proportion in PD_{DV} of aliased IPs stands at 58.3%, while that of non-aliased IPs is a mere 1.4%. Upon rigorous statistical analysis, we discerned that the disparity between the cumulative proportions of alias IPs and non-alias IPs converges to within 10% when the Path Length Difference (PL_{DV}) equals 4. Interestingly, a similar convergence of less than 10% in the cumulative proportions of the two categories is observed only when the Path Direction Difference (PD_{DV}) exceeds 8. These findings underscore the challenge of distinguishing alias IPs from non-alias IPs based on Path Length Difference while emphasizing the efficacy of Path Direction Difference as a more discerning metric.

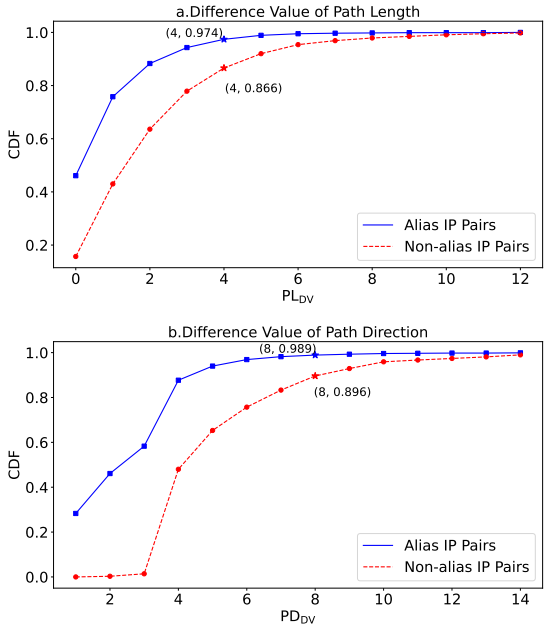

Figure 3. (**a**) Difference value of path length. (**b**) Difference value of path direction.

Furthermore, we introduce the Path Similarity Coefficient (PSC) as a novel metric to elucidate distinctions between aliased IPs and non-aliased IPs within their routing paths. PSC quantifies the degree of similarity between two paths by considering the ratio of common IP addresses between them to the length of the path, as defined in Equation (4).

Feature 4: Path Similarity Coefficient. For any given pair of IP addresses, each associated with path lengths ℓ_1 and ℓ_2, and whose routing paths have an edit distance of d, the Path Similarity Coefficient feature PSC is expressed as follows:

$$PSC = 1 - \frac{2d}{\ell_1 + \ell_2} \tag{4}$$

To gain insight into the behavior of aliased IPs and non-aliased IPs, we calculate the Path Similarity Coefficient separately for each group. Figure 4 depicts the cumulative distribution of the Path Similarity Coefficient. It is evident from the figure that aliased IPs exhibit significantly higher routing path similarity compared to non-aliased IPs. Specifically, only 39.1% of aliased IPs have path similarity coefficients below 10%, whereas a striking 98.5% of non-aliased IPs fall within the same range. This substantial disparity underscores

the effectiveness of our proposed Path Similarity Coefficient feature in discerning between these IP categories.

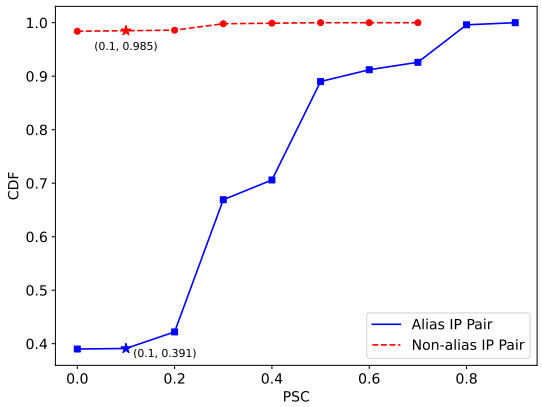

Figure 4. Cumulative distribution curve of Path Similarity Coefficient.

3.3. Reply TTL

The Reply TTL, denoting the Time-to-Live value in response messages from the target IP during active measurements, plays a crucial role in network analysis [32,33]. When a response packet is received by the source IP from the router, the Reply TTL in the packet header reflects the path length from the router to the source IP. Due to the consistent use of the same destination address (i.e., the source IP address) for all response packets sent by the router, there is substantial stability in the return path, as demonstrated in Section 3.2 of our analysis. This inherent stability in the return path can be leveraged to distinguish alias IPs from non-alias IPs. To this end, we propose the Difference Value of Reply TTL, abbreviated as TTL_{DV}. Figure 5 presents cumulative distribution curves illustrating the Reply TTL difference values for both alias IPs and non-alias IPs.

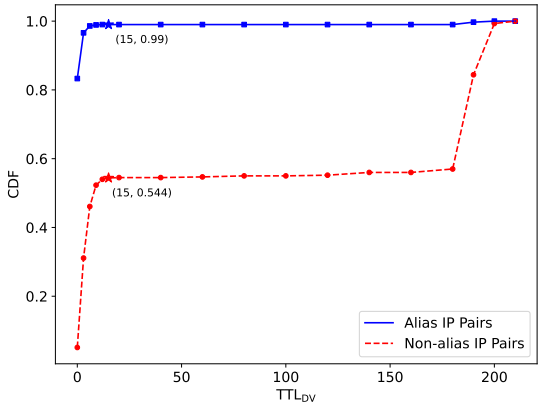

Figure 5. Cumulative distribution curve of Reply TTL difference.

Feature 5: Difference Value of Reply TTL. For any given pair of IP addresses, each associated with Reply TTL values obtained after active probing, denoted as ttl_1 and ttl_2, the Difference Value of Reply TTL feature TTL_{DV} is expressed as follows:

$$TTL_{DV} = |ttl_1 - ttl_2| \tag{5}$$

The analysis reveals that approximately 83% of alias IP pairs exhibit identical Reply TTL values, while less than 3% of the Reply TTL values in non-alias IPs coincide, highlighting the consistency in network response behavior among alias IPs. Figure 5 provides further insights, with 99% of alias IPs displaying $TTL_{DV} \leq 15$, while the cumulative percentage

for non-alias IPs is substantially lower at 54.4%. These observations affirm the efficacy of the TTL_{DV} feature as a potent discriminator between alias IPs and non-alias IPs.

3.4. Spatial Differences in IP Addresses

The consideration of spatial distance among IP addresses, particularly those situated on different interfaces of the same router, is crucial for understanding their potential aliasing. Key findings, as noted by Keys et al. in reference [13], indicate that the probability of two IPs being aliased within a "/24" subnet (the spatial distance of IP addresses less than 128) is relatively low, at less than 30%. In contrast, within a "/16" subnet, approximately 50% of IP addresses exhibit aliasing with others (the spatial distance of IP addresses is less than 65,536). These insights underscore the significant spatial separation often observed among aliased IPs, even in densely populated IPv4 networks. Hence, we propose a spatial distance feature for IP addresses, allowing us to quantify the spatial separation between IPs.

Feature 6: Spatial Distance of the IP pair. For any given pair of IP addresses, the Spatial Distance of the IP pair is calculated by the integer values I_1 and I_2 obtained by converting the IP address from dotted-decimal notation to an integer representation. Spatial Distance of the IP pair feature SD_{IP} is expressed as follows:

$$SD_{IP} = \log_2 |I_1 - I_2| \tag{6}$$

Figure 6 provides a comprehensive view of the distribution of spatial distances among IP pairs, taking into account the statistics of both aliased IPs and non-alias IPs. This analysis highlights a notable disparity in spatial distance between these two categories of IPs. Specifically, non-alias IPs exhibit significantly smaller spatial distances compared to alias IPs. Approximately 98% of non-alias IPs have spatial distances within the "/24" subnet. In contrast, only 33.2% of alias IPs are found in the same "/24" subnet. These findings underscore the substantial difference in the spatial distribution of IP addresses between alias IPs and non-alias IPs.

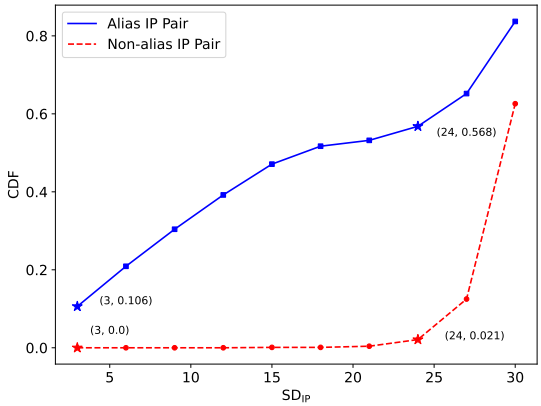

Figure 6. Spatial distance distribution of IP pairs.

4. Framework Design of AliasClassifier

Capitalizing on the power of machine learning, we introduce a novel alias parser, AliasClassifier. This innovative tool reinterprets the router alias-resolution problem through the lens of classification. It employs aliased IPs as positive case samples and non-aliased IPs as negative case samples. These samples are subjected to a training process that utilizes a multitude of features, as detailed in the previous section, therefore facilitating the parsing of IP pairs for potential alias determination. In addition to its core functionality, AliasClassifier offers a set of aggregation rules specifically designed for potentially aliased IP pairs. These rules, which form the basis of the alias triangle aggregation algorithm, serve to further diminish the number of misclassified alias IP pairs. This, in turn, enhances the accuracy of router alias resolution, underscoring the efficacy of AliasClassifier as a potent tool in the realm of network topology analysis.

Figure 7 provides a visual representation of the architecture of AliasClassifier, illustrating its key components and workflow. The AliasClassifier is structured into four primary modules: the Data Collection Module, the Alias Filtering Module, the Alias Classification Module, and the Alias Aggregation Module.

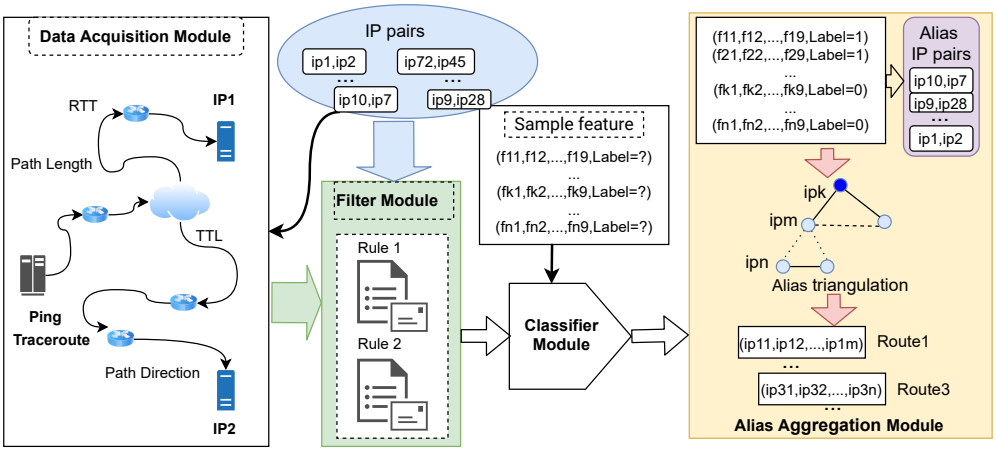

Figure 7. Architecture of AliasClassifier.

4.1. Data Collection Module

This study primarily entails the acquisition of feature-rich information from the sample dataset. This encompasses the retrieval of routing path data, spanning from source IP to destination IP, as well as round-trip delay metrics, Reply TTL, and other pertinent parameters. These measurements are performed utilizing custom-developed active measurement tools, namely 'smark' and 'sping'. To ensure the inclusivity of routing data, an approach involving the deployment of multi-protocol messages (ICMP, UDP, TCP) is adopted and administered over multiple iterations. Concurrently, the identification and exclusion of aberrant routes characterized by circular trajectories and private IPs is undertaken.

4.2. Alias Filtering Module

The alias filtering process entails the systematic application of specific filtering rules to discern potential non-alias IP pairs. This proactive filtering strategy serves the dual purpose of reducing the volume of IPs, necessitating identification and enhancing resolution efficiency. The following two fundamental filtering rules are employed:

1. IPs co-occurring within the same IP path exhibit an inherent incapacity to function as aliases for one another.
2. And IPs displaying incongruities in their online status are improbable candidates for alias relationships.

It is imperative to note that offline IPs typically lack access to essential data, such as routing path information. Therefore, AliasClassifier optimizes its parsing efficiency by exclusively parsing online IPs, given their capacity to furnish the requisite information for precise alias recognition.

4.3. Alias Classification Module

The Alias Classification Module represents the pivotal stage in the process, dedicated to the discernment of potential aliased IP pairs following rigorous filtering procedures. These identified aliased IP pairs serve as the foundation for the subsequent step in the process. However, different classifiers have different performances, leading to variations in their efficiency and effectiveness when classifying the same set of samples. In the context of router alias resolution of the network containing n IP addresses, the algorithm's overall time complexity is $O(n^2)$ due to the utilization of IP pairs as input data. Therefore, the selection of classifier models for alias resolution is a crucial decision.

Our discovery during feature analysis was that most of the feature data exhibit heavy-tailed characteristics and discrete values. These findings have motivated the criteria for choosing an appropriate classifier model. Specifically, the selected classifier model must be lightweight to ensure swift classification, thus enhancing the efficiency of alias resolution. Additionally, priority is given to classifier models adept at handling heavy-tailed and discrete data. As a result of these considerations, four distinct classifier models have been chosen for experimental training: Naive Bayes (NB), Support Vector Machine (SVM), Decision Tree (DT), and Random Forest (RF). This strategic selection aligns with the identified requirements and is poised to facilitate effective alias resolution within the given context.

During the training phase of the alias-resolution classifier, the initial step involves the acquisition of both the alias IPs dataset and non-alias IPs dataset from the public data source ITDK, as expounded upon in Section 2.2. Following this, the data collection module is deployed to procure pertinent feature data within the sample set, wherein designated classification features are utilized to construct feature vectors for each sample. The assemblage of samples constituting the training set is then inputted into the designated machine-learning model for the purposes of training. Consequent to this, the efficacy of the classifier is meticulously assessed through its application to the test set, therefore furnishing a pragmatic appraisal of its operational performance.

In our experimental setup, we have employed a self-help sampling method to create training and test sets for the ground truth collection. This method has yielded a total of five self-help samples. Consequently, all four classifiers, namely NB, SVM, DT, and RF, have been subjected to alias-resolution tasks five times each. To accurately assess the performance of each classifier, we have evaluated them using a comprehensive set of metrics, including precision, recall, F0.5 score, F1 score, F2 score, and parsing time per 100,000 IP pairs.

We assess the classification accuracy of various classifier models using precision metrics, evaluate the capability of these models to detect potential alias IPs through recall metrics and gauge the parsing efficiency of each classifier model by measuring the parsing time for every 100,000 IP pairs. Subsequently, we ascertain the comprehensive efficacy of each classifier model using the *F score* index. The *F score* provides a holistic evaluation of classifier performance, considering both precision and recall. Specifically, we consider the F1 score as equally significant as precision and recall while according twice the weight to recall in the F2 score. Conversely, in the F0.5 score, recall is assigned half the importance of precision, reflecting a more nuanced assessment of classifier efficacy. The detailed results of this evaluation are presented in Table 4, with a description of the evaluation metrics provided in Appendix A for reference. The classifier models leading in each metric have been Bolded for reference.

The analysis from Table 4 reveals distinct strengths and weaknesses among the classifier models used in alias resolution. From the analysis in Table 4, we found that the Bayesian classifier, despite its significant performance in recall, F1 score, and F2 score, exhibits a relatively low classification accuracy, averaging 74.4%. This deficiency undermines its reliability as a parser and introduces a potential constraint. The diminished accuracy observed in the Bayesian classifier can likely be attributed to the interdependence and continuous nature of the selected features. Conventionally, Bayesian classifiers operate under the assumption of feature independence and discreteness. However, certain features chosen for our analysis exhibit correlations, notably those concerning routing paths. Additionally, some selected features manifest as continuous values, exemplified by the Difference Value of round-trip time. This departure from discrete, independent features may entail information loss within the Bayesian classifier framework, consequently undermining its classification efficacy.

The decision tree classifier, while demonstrating commendable classification speed and accuracy, suffers from low recall. This observation is intricately linked to a known limitation inherent in decision tree classifiers—namely, their susceptibility to overfitting.

Decision trees, by their nature, have a propensity to excessively tailor themselves to the intricacies of the training data, particularly under conditions of heightened tree depth or scant training samples. Under such circumstances, the phenomenon of overfitting becomes pronounced, as the classifier excessively captures noise and idiosyncrasies within the training data, compromising its generalization capability. This limitation could pose a significant challenge in practical applications, as the low recall may result in a diminished discovery rate of alias IPs. Consequently, this could lead to a restricted coverage of the inferable router topology, negatively impacting the construction of the router topology.

Table 4. Evaluation results of four classifier models.

		1	2	3	4	5
NB	Pre	78.25%	80.93%	82.65%	65.36%	64.62%
	Rec	67.39%	68.25%	66.11%	74.22%	72.88%
	F1	0.7241	0.7405	0.7346	0.6951	0.8650
	F0.5	0.7581	0.7803	0.7871	0.6696	0.6612
	F2	0.6931	0.7046	0.6887	0.7226	0.7106
	Time/10w pairs (s)	0.0064	0.0061	0.0060	0.0062	0.0062
SVM	Pre	90.97%	79.44%	78.67%	88.89%	90.73%
	Rec	41.48%	44.87%	45.46%	40.34%	42.08%
	F1	0.5698	0.5735	0.5763	0.5550	0.5749
	F0.5	0.7344	0.6883	0.6864	0.7164	0.7369
	F2	0.4654	0.4915	0.4966	0.4529	0.4713
	Time/10w pairs (s)	848.56	902.86	797.45	1101.88	932
DT	Pre	99.81%	99.81%	99.77%	99.75%	99.85%
	Rec	21.23%	22.22%	21.53%	22.74%	21.45%
	F1	0.3501	0.3634	0.3542	0.3704	0.3531
	F0.5	0.5735	0.5877	0.5778	0.5947	0.5768
	F2	0.2520	0.2631	0.2553	0.2689	0.2545
	Time/10w pairs (s)	0.0042	0.0045	0.0048	0.0050	0.0045
RF	Pre	95.74%	95.89%	95.93%	91.78%	94.96%
	Rec	40.04%	40.59%	40.85%	39.49%	41.27%
	F1	0.5647	0.5704	0.5730	0.5522	0.5748
	F0.5	0.7490	0.7536	0.7556	0.7256	0.7535
	F2	0.4531	0.4588	0.4615	0.4457	0.4653
	Time/10w pairs (s)	0.2433	0.6901	0.2307	0.2420	0.2512

The SVM classifier is evidently burdened by the parsing speed. While SVM boasts robust generalization capabilities and excels in handling high-dimensional datasets, their efficacy is tempered by computational demands, particularly evident in the processing of large-scale datasets. SVM training entails substantial computational overhead and necessitates extensive storage resources, factors that impede expeditious alias resolution within network contexts. Moreover, SVM's sensitivity to missing data imposes a requisite for meticulous data preprocessing, therefore introducing variability in classifier accuracy. Consequently, despite its theoretical strengths, SVM's practical utility is constrained by computational exigencies and susceptibility to data quality fluctuations.

In contrast, the Random Forest classifier manages to strike a balance across multiple dimensions, achieving a harmonious blend of parsing speed, accuracy, and recall. While the Random Forest model may not exhibit conspicuous advantages over alternative classifier models concerning metrics such as precision, recall, and parsing time per 100,000 IP pairs, its paramount significance emerges in scenarios where parsing accuracy assumes primacy. The preeminence of the Random Forest classifier, underscored by its superior performance in the F0.5 score, substantiates its unparalleled capability to ensure parsing precision. In contexts where parsing accuracy is paramount, the Random Forest classifier emerges as the optimal choice, notwithstanding its comparative performance in other metrics. We expect the filtered classifier to simultaneously satisfy high precision of parsing results,

fast parsing speed, and high discovery rate of alias IPs that can be parsed (high recall). The Random Forest classifier strikes an attractive balance, making it an ideal choice for researchers who require both high accuracy and a robust discovery of potential alias IPs. Its balanced performance, resistance to overfitting, and noise immunity further add to its appeal.

Given the favorable balanced performance and versatility of the Random Forest Classifier, as well as its ability to mitigate overfitting and handle noise in data, the decision to construct AliasClassifier based on the Random Forest model for comparison with other methods appears well-founded. Subsequent sections will delve into the details of the comparative experiments, shedding light on the effectiveness and advantages of this approach.

4.4. Alias Aggregation

Alias aggregation is the process of aggregating identified alias IP pairs into routers. Traditional machine-learning methods employ alias passing for router node aggregation following the identification of aliased IP pairs. Alias passing implies that for mutually aliased pairs (IP A, IP B) and (IP B, IP C), the trio (IP A, IP B, IP C) are classified to the same router node due to the existence of a common alias interface, IP B, for both pairs, as shown in Figure 8.

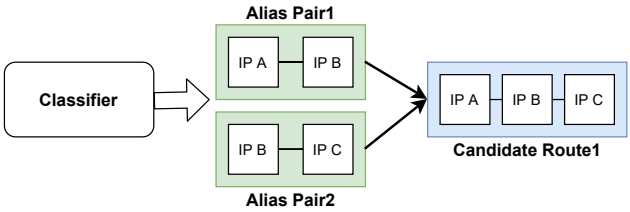

Figure 8. Process diagram of alias passing.

However, as discussed in Section 4.3, it is evident that regardless of the classification model employed, the final classification results contain some errors. When alias passing is used for router node aggregation on the IP pairs that have been discriminated against by the classifier, the misreported alias IPs associate with many unrelated routers. This association results in a significant reduction in the number of routers that are eventually aggregated. Consequently, some of the routers inferred through the alias-resolution process may be assigned more IP addresses than they actually possess.

To address the aforementioned "router bloat" problem, we propose a router aggregation algorithm based on alias triangulation. An alias triangle comprises any three IP addresses that are aliases of each other. If any three IP addresses in a router form an alias triangle, we consider the router comprising these three IP addresses to be real. A simple schematic of the router aggregation algorithm based on alias triangles is depicted in Figure 9. The aggregation steps are as follows:

Step 1:We construct an alias set of IP addresses (Alias dataset) by incorporating all the IP addresses deemed to be aliased by the classifier. Specifically, for any IP address ip_k, we construct an alias set for ip_k by taking ip_k as the key and the IP addresses judged by the classifier to be aliases of ip_k as the value: $ip_k = \{ip_{k1}, ip_{k2}, \ldots, ip_{kn}\}$.

To reduce the data volume during router aggregation, we sort the Alias dataset in ascending order based on the key IP and also eliminate the member IPs in the value IP that are smaller than the key IP, i.e., for the set of IP aliases $ip_k = \{ip_{k1}, ip_{k2}, \ldots, ip_{kn}\}$, there exists $int(ip_{ki}) > int(ip_k)$ for any ip_{ki}. Where $int()$ represents a decimal integer value.

Step 2: We select the IP address ip_k with the smallest key IP from the Alias dataset to initiate router aggregation. Each router is represented by a set of member IP addresses, which is dynamically expanded as more IP addresses are discriminated. Therefore, the initial IP set for R_k is $R_k = \{ip_k\}$.

We subject the set of ip_k aliases $ip_k = \{ip_{k1}, ip_{k2}, \ldots, ip_{kn}\}$ to alias triangulation. If any two IP addresses (ip_i, ip_j) within $ip_k = \{ip_{k1}, ip_{k2}, \ldots, ip_{kn}\}$ are also aliased to each

other, then (ip_i, ip_j) also belongs to the constituent IP addresses of Router R_k. After alias triangulation of all IP addresses in $\{ip_{k1}, ip_{k2}, \ldots, ip_{kn}\}$, we obtain the new Router R_k.

Step 3: The judgment process from Step 2 is also applied to the newly added IP addresses in Router R_k until all member IP addresses in Router R_k have completed the alias triangle judgment. If no new members are added to Router R_k after all member IP addresses in Router R_k have completed the judgment, it indicates that all members of Router R_k have been recognized.

Step 4: Remove all member IPs of Router R_k from the Alias dataset and repeat Steps 2 and 3 until all IP addresses in the Alias dataset are devoid of alias triangles. The IP addresses that do not form alias triangles undergo alias passing to construct potential routers. This subset of routers may contain IP addresses that have been incorrectly aliased, so we designate it as the Candidate Router.

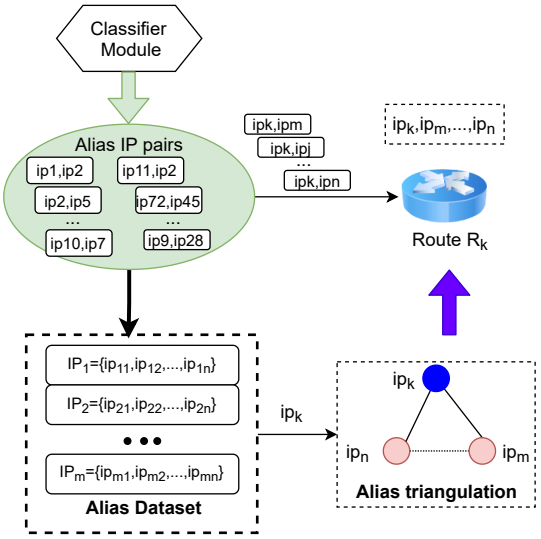

Figure 9. Router aggregation based on alias triangulation.

5. Experimentation and Evaluation

In this section, we initially conduct an analysis of AliasClassifier in juxtaposition with state-of-the-art alias-resolution techniques. This is carried out to validate the effectiveness and advancement of AliasClassifier, as elaborated in Section 5.1. Subsequently, we delve into the discovery rate of resolvers' alias IPs by establishing field test experiments, as well as discussing the practical utility of the router aggregation algorithm based on the alias triangle, as outlined in Section 5.2. We then assess the actual resolving efficiency of each resolver by instituting a set of comparative experiments to ascertain the capability of AliasClassifier in resolving large networks in section 5.3. Finally, in Section 5.4, we explore the impact of the number of Vantage Points (VPs) on AliasClassifier to discuss the deployment cost of AliasClassifier.

5.1. Effectiveness and Advancement

We conducted a comparison of carefully chosen alias resolvers using ground truth sets procured from the ITDK dataset, with the objective of thoroughly assessing the performance of AliasClassifier in relation to state-of-the-art alias-resolution techniques in terms of the accuracy of resolution results and the discovery rate of alias IPs. However, we no longer possess the supplementary ground truth data, aside from the ground truth set derived from the ITDK dataset. As a result, for the effectiveness and advancement comparison study, we were unable to juxtapose AliasClassifier with MIDAR + iffinder [13,27], the methodology employed to generate the ITDK dataset. Instead, we selected another cutting-edge alias classifier, APPLE [26], the most recent alias classifier proposed by CAIDA. APPLE identifies potential alias pairs with a coverage accuracy that is comparable to MIDAR in IPv4. We

also chose two advanced alias resolvers, TreeNET [19] and MLAR [21], to participate in the comparative experiments. Notably, MLAR utilizes support vector machines as its kernel.

To ensure the precision of our distributed detection-based parsing algorithm, we have strategically chosen 13 VPs distributed across the globe to conduct active detection, thus collecting comprehensive feature information. In scenarios that employ machine-learning models, we have employed a self-service sampling method to derive five distinct sets of training and test data from our ground truth collection. These training sets serve as the foundation for the training of machine-learning models, while the test sets play a crucial role in evaluating the performance of all parsers involved. Consequently, AliasClassifier, TreeNET, MLAR, and APPLE are each tasked with performing five alias parses. This systematic approach guarantees a robust evaluation of the parsing algorithms and reinforces the reliability and comprehensiveness of our assessment.

Figure 10 furnishes a comprehensive exposition delineating the parsing accuracy proficiency of the four parsers under scrutiny. Notably, APPLE emerges as the paragon of resolution accuracy, consistently achieving 100% precision in router alias resolution across all five test experiments. However, meticulous scrutiny reveals a conspicuous limitation: APPLE exhibits a meager discovery rate of IPs bearing aliasing relationships, reflected in its paltry average recall rate of merely 0.22%. This diminutive figure denotes the accurate identification of a trifling fraction of IPs harboring aliasing relationships by APPLE.

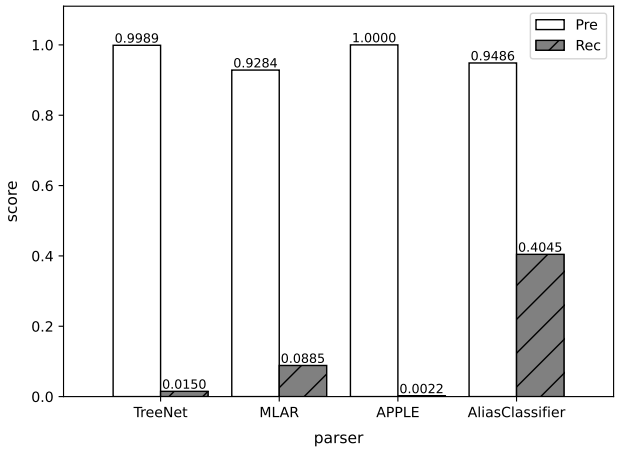

Figure 10. Performance of four parsers in terms of precision and recall.

The exceedingly low incidence of discovering aliased IPs in APPLE may be intricately intertwined with the incompleteness characterizing the amassed topology dataset. Despite a concerted effort involving the deployment of 13 VPs spanning the globe to procure a comprehensive global IP topology, the corpus of valid IP-level topologies remains conspicuously circumscribed. This limitation is exacerbated by the dispersal of experimental IP addresses across disparate geographic locations worldwide. Consequently, notwithstanding APPLE's adeptness in precise alias resolution, its efficacy is significantly curtailed in the absence of a robust and expansive IP-level topology dataset.

TreeNET also suffers heavily from a high reliance on IP-level topology, and similar to APPLE, TreeNET's average recall is only 1.5%, even though it also achieved an excellent average accuracy of 99.88% over the five experiments. Employing topology measurements, TreeNET segments all IP aliases into discrete sets before embarking on an array of supplementary alias determination procedures. Despite earnest endeavors aimed at enhancing resolution precision and expanding the roster of resolvable IPs, TreeNET contends with the persistent lacuna inherent in router topology due to the absence of a comprehensive IP topology framework.

In stark contrast to the limited recall exhibited by APPLE and TreeNET, MLAR and AliasClassifier emerge as formidable contenders. Despite MLAR's middling average precision of 92.80%, ranking it as the least precise among all parsers scrutinized, its recall

performance markedly outshines that of APPLE and TreeNET, boasting an average recall of 8.85%. Our proposed innovation, AliasClassifier, surpasses its predecessors, not only achieving an average classification precision of 94.80% but also showcasing an outstanding average recall of 40.45%, surpassing the performance of other alias parsers by a considerable margin.

The notable enhancement in the recall rates of MLAR and AliasClassifier stems from their reduced reliance on IP-level topology. Unlike their counterparts, MLAR and AliasClassifier exhibit a diminished dependency on comprehensive IP-level topology. Leveraging features that are comparatively more accessible than a complete IP-level topology, both MLAR and AliasClassifier substantially enhance their discovery rates of alias IPs. Furthermore, compared with MLAR, the features selected by AliasClassifier are not only less difficult to obtain but also more effective in improving the discovery rate of alias IPs and have a leading edge in accurately determining alias relationships. AliasClassifier outperforms MLAR by enhancing overall accuracy and recall by 2% and 31.6%, underscoring the superior efficacy of features selected by AliasClassifier in discerning alias relationships.

The huge difference in recall between the parsers leads to a huge gap in their combined scores. Figure 11 offers a detailed insight into the performance of the four parsers across the three comprehensive evaluation metrics. Notably, AliasClassifier emerges as the indisputable leader, showcasing a substantial advantage over the other parsers, largely attributable to its superior recall rates. Conversely, APPLE and TreeNET, affected by their lower recall rates, ultimately exhibit subpar comprehensive performance, highlighting the crucial role recall plays in the overall efficacy of alias-resolution methodologies.

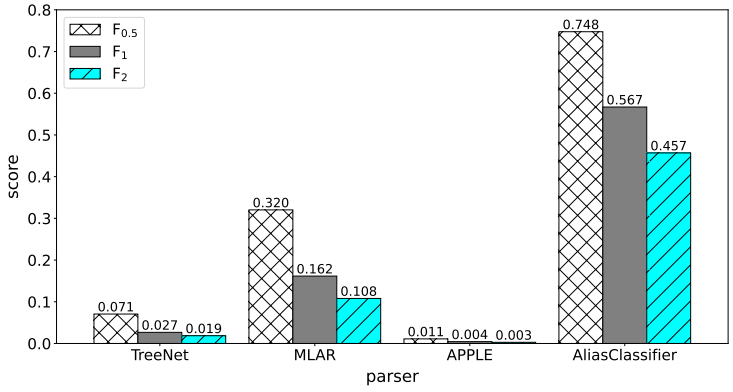

Figure 11. Performance of four parsers on the composite metrics.

The recall of an alias resolver mirrors the extent of resolvable IP coverage, delineating the discovery rate of alias IPs. A heightened parser recall signifies a greater number of IPs successfully resolved, therefore facilitating a more comprehensive delineation of router topology. The comparatively low recall rates observed in APPLE and TreeNET underscore the pressing need to prioritize the discovery of potentially aliased IPs over the precise determination of aliasing relationships between IPs towards achieving a more exhaustive depiction of router topology. Regrettably, this nuanced emphasis has eluded earlier iterations of alias-resolution tools. The marked surge in recall witnessed in MLAR and AliasClassifier signals the heightened potential of machine-learning-driven alias-resolution tools in uncovering more comprehensive router topologies compared to their conventional counterparts. Notably, our proposed AliasClassifier demonstrates a substantial proficiency in uncovering a greater array of aliased IPs compared to other advanced alias resolvers, therefore solidifying its position as the preeminent frontrunner in the endeavor to unravel more expansive router topologies. Consequently, AliasClassifier resolves a more comprehensive router topology compared to other alias resolvers.

5.2. Field Testing of Real Network

The observed low recall in parsers like APPLE and TreeNET can likely be attributed to two plausible factors. One possibility is that the accuracy of the latest ITDK dataset, which we utilized, has significantly deteriorated. Given that the ITDK dataset has been relevant to the topology for a year, substantial changes in the Internet topology over the past year could result in most alias IPs losing their alias relationships. Another possibility is that APPLE and TreeNET may not be highly effective at discovering aliased IPs. While they can accurately identify aliased IPs, the number of aliased IPs discovered appears to be sparse.

To ascertain the actual cause of the low recall in resolvers such as APPLE and TreeNET, we conducted field testing on 1 million IPs located in the same region. This testing aimed to identify what contributes to the low recall of APPLE and TreeNET and to verify the real alias IPs discovery ability of advanced resolvers. We filtered IPs from multiple IP geolocation libraries [34–36] that are co-located in Shanghai, selecting 1 million IPs that are most likely to be in the same region as targets for alias resolution. In this set of experiments, MLAR is not included in the discussion due to its lengthy resolution time for 1 million IPs, which is expected to exceed 87,000 h. Consequently, we did not use the MLAR for field testing. We focus on the discovery ability of aliased IPs of three resolvers, namely AliasClassifier, APPLE, and TreeNET. Simultaneously, to offset the absence of MLAR, we incorporated the classical combinatorial parser of MIDAR + iffinder for field testing. The rationale behind introducing the classical combined parser of MIDAR + iffinder is to render the field test experiments more persuasive.

In this experiment, we solely focus on the discovery ability of the resolvers' alias IPs, assuming by default that the alias IPs are discovered correctly. To ensure the fairness and reliability of the test results, we conduct tests separately using four types of alias resolvers in the same experimental environment. The feature data required by each resolver is collected in real time. For APPLE, a resolver that is highly dependent on network topology, we field probe all IP addresses in Shanghai (approximately 15.36 million IPs) from 13 VPs around the world. This approach aims to construct Shanghai's network topology as comprehensively as possible and improve the discovery rate of APPLE's alias IPs. Meanwhile, each resolver conducts three experiments, respectively, and the one with the highest number of resolution results is selected for comparison. Table 5 presents the results of the field tests for the four resolvers.

Table 5. Results of field tests of the four advanced parsers.

	TreeNET	**APPLE**	**MIDAR + Iffinder**	**AliasClassifier**
Alias IP pairs	66,551	29,426	5570	7,501,482
Interface IPs	4668	382	1072	11,514
Routers	1291	32	209	2385

As delineated in Table 5, the alias IP pairs discovered by TreeNet, APPLE, and MIDAR+iffinder were 66,551, 29,426, and 5570, respectively. The actual count of IP addresses with resolved alias relationships was significantly lower, standing at 4668, 382, and 1072 for each method, respectively. This represents a sparse fraction when compared to the target total of 1 million IP addresses. In stark contrast, AliasClassifier identified a substantial total of 7.5 million alias IP pairs. The actual number of IP addresses with resolved alias relationships was 11,514, which is 2.4 times the quantity of alias IPs unearthed by TreeNET, and 30 times by APPLE. This outcome mirrors the difference in recall between AliasClassifier, TreeNET, and APPLE, as presented in Figure 10. The low recall of aliased IPs for TreeNet and APPLE is not due to an excess of invalid data in the ITDK dataset. Instead, it appears that these advanced parsers inherently struggle with discovering potentially aliased IPs and are unable to discover aliased IPs in large quantities in the absence of sufficient feature information.

Our field test experiments served as a valuable platform to assess the efficacy of the router aggregation method predicated on alias triangulation. We observed that de-

spite AliasClassifier inferring a dataset comprising 7.5 million pairs of aliased IPs, the application of the alias passing method yielded a mere 64 routers. Notably, one of these routers contained more than 103,000 IP addresses, a figure that starkly deviates from the true representation of routers on the Internet. In contrast, when we employed an alias triangulation-based approach to router aggregation, AliasClassifier inferred a more plausible set of 2385 routers. Among these, the router with the highest number of IP addresses contained only 793 addresses. This figure aligns more closely with the number of IP addresses typically associated with real-world Internet routers.

Another alias-resolution method that employs alias passing for router aggregation, APPLE, is similarly afflicted by the phenomenon of "router bloat". Despite APPLE's recognition of a substantial 29,426 individual IP pairs, it manages to resolve a mere 32 routers. In contrast, the classic MIDAR+iffinder method deduces a total of 209 routers, even though it identifies only 5570 IP pairs. This stark discrepancy underscores the limitations inherent in the alias-passing-based approach.

While there exist disparities between the router results inferred by AliasClassifier and those furnished by MIDAR+iffinder, the inferred outcomes of AliasClassifier predominantly cover those of MIDAR+iffinder. The routers deduced by AliasClassifier typically encompass a greater number of IP addresses compared to the results proffered by MIDAR+iffinder. This observation indirectly yet unequivocally attests to the reliability of AliasClassifier. Field test experiments lend further credence to this assertion, demonstrating that the router aggregation method predicated on alias triangulation is indeed efficacious in curtailing the number of irrelevant IPs in a router. Consequently, this method significantly enhances the accuracy of alias resolution, underscoring the potential of AliasClassifier as a robust tool in the realm of router alias resolution for IPv4 networks.

5.3. Resolution Efficiency

During the field testing, we also conducted a comparative analysis of the resolution efficiency of various resolvers across different network sizes. We selected IP addresses in increments of 10^4, 10^5, 5×10^5, 10^6, and 2×10^6 from the Shanghai city IP address dataset. Each of these subsets was then subjected to alias resolution using four distinct parsers. Our primary focus in this experiment was the time efficiency of the parsing process rather than the parsing results. To ensure a fair comparison, we included the time taken by each parser to acquire feature information in our time-efficiency calculations. The total elapsed time for each parser was computed by summing up the time taken for data collection, data processing, and alias resolution. Figure 12 presents a comparative analysis of the parsing efficiency of the four parsers under investigation.

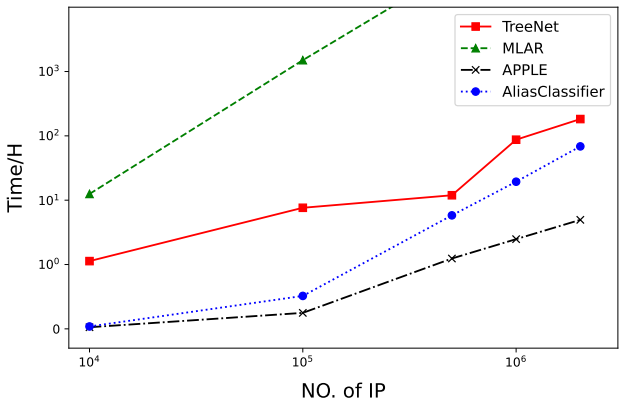

Figure 12. Comparison of parsing efficiency of four parsers.

Figure 12 illustrates that the parsing time of each resolver escalates gradually with expanding network size. Notably, the parsing speed of MLAR, based on Support Vector Machine, is markedly constrained by the number of IPs to be parsed. It necessitates

12.5 h to parse 10,000 IP addresses, and this time dramatically surges to 1500 h for parsing 100,000 IP addresses, indicative of an exponential growth trend in parsing time. Consequently, MLAR is deemed less suitable for large-scale alias-resolution endeavors. Comparatively, APPLE emerges as the swiftest, with its runtime predominantly contingent on the time taken to acquire network topology information. Subsequently, AliasClassifier follows suit, requiring approximately 19.37 h to resolve 1 million IPs. The time consumption exhibits a linear growth pattern. In contrast, TreeNET's time consumption aligns closely with a linear growth trajectory, demanding 87.13 h for resolving 1 million IPs, 4.5 times more than AliasClassifier. This underscores AliasClassifier's favorable parsing speed and its suitability for large-scale network alias-resolution tasks.

5.4. Deployment Cost

Deployment cost is a crucial metric in evaluating a router alias resolver. The deployment cost of a distributed parser is anticipated to exceed that of a centrally deployed parser, a fact that is particularly evident in the context of large networks. Excessive distributed active probing for large networks may disrupt the network's normal operation, diminish the parsing efficiency of the resolver, and escalate the economic cost. However, increasing the number of detection points may improve accuracy. For instance, active detection of the same IP pair from N VPs yields N PL_{DV} features and N TTL_{DV} features. Consequently, the feature vector generated for each sample based on N VPs will comprise N PL_{DV} and N TTL_{DV}, leading to an expansion of the feature vector. As the number of favorable observation points increases, so does the feature vector used for classifier training. Theoretically, an increase in the number of features will help improve the accuracy of the classifier.

We delve further into the deployment cost of AliasClassifier, specifically examining the influence of the quantity of VPs on the accuracy of AliasClassifier. To assess the impact of the number of VPs on AliasClassifier, we imposed an artificial limit on the number of VPs. This meticulous investigation is designed to shed light on how the quantity of VPs influences AliasClassifier's performance, offering valuable insights into the role of favorable observation points in the system's operation. Table 6 presents the performance results of AliasClassifier based on varying numbers of VPs.

In Table 6, our investigation reveals that augmenting the Random Forest classifier, AliasClassifier, through the augmentation of probe points yields less than satisfactory outcomes. In the process of training the Random Forest classifier for AliasClassifier, we utilized a batch of training sets derived from the ground truth dataset, employing the self-service sampling method. The Out-of-Bag (OOB score) represents the evaluation score achieved by selecting samples using the "self-help method", which approximates n-fold cross-validation test accuracy [37]. Our experimentation encompassed the utilization of 1, 2, 3, and 4 distributed detectors.

Table 6. The impact of the number of VPs.

No. of VPs	No. of Features	Best Parameters		OOB Score
		Estimated Quantity	Maximum Length	
1	10	220	18	95.74%
2	15	280	20	92.49%
3	20	270	17	96.32%
4	25	220	20	92.91%

When we employ one VP, we generate a Random Forest model comprising 220 estimators with a maximum depth of 18, yielding an OOB score of 95.74%. Intriguingly, our observations indicate that the OOB accuracy of AliasClassifier displays an oscillatory pattern as the number of detectors increases rather than exhibiting a consistent enhancement. For instance, a classifier based on one VP achieves a notable parsing accuracy of 95.74%, yet the accuracy of a classifier predicated on 2 VPs declines to 92.49%. Although

the classifier predicated on 3 VPs attains the peak accuracy of 96.32%, the accuracy subsequently diminishes with the classifier based on 4 VPs. This observation underscores that the parsing accuracy of AliasClassifier is independent of the number of VPs, characterizing it as a typical centralized alias parser. This implies that AliasClassifier does not necessitate distributed deployment to achieve high-quality router alias resolution. Consequently, AliasClassifier emerges as a highly practical alias-resolution technology for real-world deployments, distinguished by its exceptionally low deployment cost.

6. Discussion

Despite these promising results, our study identified a potential router encompassing 1927 IP addresses. This router represents the final potential router composed of aliased IPs without any alias triangles, implying that alias passing is employed to obtain a potential router. The challenge arises from the fact that our algorithm cannot further discern whether this router contains extraneous IPs.

Future investigations will explore the integration of traffic data and IP remarks to augment the discernment of potential alias IP pairs. The management of multiple IP interfaces within a single router typically regulates traffic processing velocity through a centralized system. Consequently, prolonged monitoring of traffic associated with potential alias IP addresses, segmented by interface IPs over distinct timeframes, is under consideration. Distinctive attributes will serve as foundational criteria for the refined classification of non-alias IP addresses within prospective routers, therefore bolstering their reliability. Additionally, leveraging whois data, geographic locational insights, and supplementary remarks linked with IP addresses will aid in filtering out non-alias IPs within prospective routers. Efficient and precise methodologies for enhancing the credibility of prospective routers represent a critical avenue necessitating further investigation and refinement in our ongoing research endeavors.

Indeed, the AliasClassifier methodology extends its applicability beyond IPv4 networks, demonstrating promise for employment within IPv6 networks as well. With the depletion of IPv4 addresses, the momentum behind IPv6 deployment is accelerating, as demonstrated by the exponential growth observed in BGP prefix advertisements for global IPv6 [38]. Consequently, there is an increasingly pressing need to delve into the network topology and alias-resolution mechanisms specific to IPv6.

However, IPv6 not only extends the address space beyond the limitations of IPv4 but also introduces alterations to the packet format of IP packets [39]. Given the distinct design characteristics of IPv6 networks and the relative scarcity of IP addresses within them, many alias-resolution techniques proven effective in IPv4 environments are incompatible with IPv6 networks. Consequently, researchers have endeavored to develop IPv6-specific alias-resolution methodologies, leveraging approaches such as source routing, induced IPID, and prefix-based algorithms, exemplified by systems like Atlas [40], RMP [41], TBT[42], Speedtrap [43], and UAv6 [44]. These algorithms commonly exhibit issues related to their universality and parsing efficiency. For instance, source-routing-based alias resolution is confined to a subset of routers, while the TBT algorithm generates substantially larger data volumes than the IPv4 IPID method, leading to escalated network loads as the number of aliases grows rapidly.

The six primary classification features we have identified are not only prevalent within IPv4 infrastructures but also commonly observed in IPv6 contexts. Consequently, AliasClassifier exhibits significant potential for router alias resolution within IPv6 environments. Nevertheless, the efficacy of AliasClassifier in IPv6 networks is influenced by factors such as the relatively limited pool of known active IP addresses within the current IPv6 landscape and the absence of IPv6 networks characterized by high concentrations of IP addresses. Consequently, certain features outlined in this study may struggle to differentiate between alias and non-alias IPs within IPv6 networks, notably those reliant on spatial disparities among IP addresses, therefore attenuating AliasClassifier's effectiveness in this domain. However, as the prevalence of IPv6 continues to escalate in the foreseeable future, the effi-

cacy of AliasClassifier within IPv6 realms may gradually augment, driven by the expanding scope and intricacies of IPv6 network configurations.

7. Conclusions

In this paper, we delineate six features that are pivotal to the router alias-resolution problem based on four prevalent hypothetical behaviors. These features aid us in effectively differentiating between aliased and non-aliased IPs. Concurrently, we introduce an alias triangulation-based router aggregation algorithm to augment the accuracy of alias resolution. We construct an alias resolver, AliasClassifier, utilizing a Random Forest classifier and juxtapose it with an array of state-of-the-art alias resolvers. Experimental results demonstrate that AliasClassifier is aptly suited for router alias resolution in large-scale IPv4 networks. In comparison to the state-of-the-art TreeNET and APPLE, AliasClassifier achieves a 2.4× and 30× enhancement in alias IP discovery rate, respectively. Simultaneously, AliasClassifier can resolve 1 million IP addresses in less than 20 h, showcasing remarkable efficiency. Of particular note is AliasClassifier's role as a centralized alias resolver, which renders it highly cost-effective to deploy.

Author Contributions: Conceptualization, Y.X. and Z.Z.; Methodology, Y.X. and E.C.; Validation, Y.X.; Investigation, E.C.; Data curation, E.C.; Writing—original draft preparation, Y.X.; Writing—review and editing, N.L.; Visualization, Y.X.; Supervision, Z.Z. and N.L.; Project administration, Z.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Shandong Province Key R&D Project 2020CXGC10103; the National Natural Science Foundation of China (62101159) and the Natural Science Foundation of Shandong Province (ZR2021MF055).

Institutional Review Board Statement: This study did not involve human or animal research.

Informed Consent Statement: This study did not involve human or animal research.

Data Availability Statement: Publicly available datasets were analyzed in this study. These data can be found here: https://github.com/fyfishie/AliasClassifier-article-code, accessed on 1 April 2024.

Conflicts of Interest: There is no conflict of interest.

Appendix A

The confusion matrix is a prevalent tool in classification tasks, offering a comprehensive reflection of the classification results. It succinctly conveys the outcomes of binary classification tasks. By comparing predicted results with actual results, four scenarios can emerge: true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN).

Precision, within the context of classification or statistical analysis, is quantitatively defined as the ratio of true positive instances to the total number of samples that have been classified as positive.

$$Precision = \frac{TP}{TP + FP}$$

The recall rate is defined as the proportion of all actual positive samples correctly categorized as positive by the model.

$$Recall = \frac{TP}{TP + FN}$$

F_β score: more generally defined scores are:

$$F_\beta = (1 + \beta^2) * \frac{Precision * Recall}{(\beta^2 * Precision) + Recall}$$

The physical significance of this is that the two scores, precision and recall, are combined into a single score, and in the process of combining them, the recall is weighted

twice as much as the precision. The F1 score considers recall and precision to be equally important, the F2 score considers recall to be twice as important as precision, and the F0.5 score considers recall to be half as important as precision.

References

1. Laghari, A.A.; Wu, K.; Laghari, R.A.; Ali, M.; Khan, A.A. A review and state of art of Internet of Things (IoT). *Arch. Comput. Methods Eng.* **2022**, *29*, 1395–1413. [CrossRef]
2. Garcia-Jimenez, S.; Magana, E.; Morató, D.; Izal, M. Pamplona-traceroute: Topology discovery and alias resolution to build router level Internet maps. In Proceedings of the Global Information Infrastructure Symposium-GIIS 2013, Trento, Italy, 28–31 October 2013; pp. 1–8.
3. Witono, T.; Yazid, S. A review of internet topology research at the autonomous system level. In Proceedings of the Sixth International Congress on Information and Communication Technology: ICICT 2021, London, UK, 25–26 February 2022; Volume 1, pp. 581–598.
4. Canbaz, M.A. Internet Topology Mining: From Big Data to Network Science. Ph.D. Thesis, University of Nevada, Reno, NV, USA, 2018.
5. Claffy, K.; Hyun, Y.; Keys, K.; Fomenkov, M.; Krioukov, D. Internet mapping: From art to science. In Proceedings of the 2009 Cybersecurity Applications & Technology Conference for Homeland Security, Washington, DC, USA, 3–4 March 2009; pp. 205–211.
6. Spring, N.; Mahajan, R.; Wetherall, D. Measuring ISP topologies with Rocketfuel. *ACM SIGCOMM Comput. Commun. Rev.* **2002**, *32*, 133–145. [CrossRef]
7. Chun, B.; Culler, D.; Roscoe, T.; Bavier, A.; Peterson, L.; Wawrzoniak, M.; Bowman, M. Planetlab: An overlay testbed for broad-coverage services. *ACM SIGCOMM Comput. Commun. Rev.* **2003**, *33*, 3–12. [CrossRef]
8. McGregor, T.; Braun, H.W.; Brown, J. The NLANR network analysis infrastructure. *IEEE Commun. Mag.* **2000**, *38*, 122–128. [CrossRef]
9. Keys, K. Internet-scale IP alias resolution techniques. *ACM SIGCOMM Comput. Commun. Rev.* **2010**, *40*, 50–55. [CrossRef]
10. Gunes, M.H.; Sarac, K. Resolving IP aliases in building traceroute-based Internet maps. *IEEE/ACM Trans. Netw.* **2009**, *17*, 1738–1751. [CrossRef]
11. Wang, Z.; Cheng, G. Research progress of alias resolution technology. *J. Commun.* **2019**, *40*, 169–185.
12. Bender, A.; Sherwood, R.; Spring, N. Fixing Ally's growing pains with velocity modeling. In Proceedings of the 8th ACM SIGCOMM Conference on Internet Measurement, Vouliagmeni, Greece, 20–22 October 2008; pp. 337–342.
13. Keys, K.; Hyun, Y.; Luckie, M.; Claffy, K. *Internet-Scale ipv4 Alias Resolution with Midar: System Architecture*; Cooperative Association for Internet Data Analysis (CAIDA): San Diego, CA, USA, 2011.
14. Gunes, M.H.; Sarac, K. Analytical IP alias resolution. In Proceedings of the 2006 IEEE International Conference on Communications, Istanbul, Turkey, 11–15 June 2006; Volume 1, pp. 459–464.
15. Sherwood, R.; Spring, N. Touring the Internet in a TCP sidecar. In Proceedings of the 6th ACM SIGCOMM Conference on Internet Measurement, Rio de Janeriro, Brazil, 25–27 October 2006; pp. 339–344.
16. Sherry, J.; Katz-Bassett, E.; Pimenova, M.; Madhyastha, H.V.; Anderson, T.; Krishnamurthy, A. Resolving IP aliases with prespecified timestamps. In Proceedings of the 10th ACM SIGCOMM Conference on Internet Measurement, Melbourne, VIC, Australia, 1–30 November 2010; pp. 172–178.
17. Marchetta, P.; Persico, V.; Pescapé, A. Pythia: Yet another active probing technique for alias resolution. In Proceedings of the Ninth ACM Conference on Emerging Networking Experiments and Technologies, Santa Barbara, CA, USA, 9–12 December 2013; pp. 229–234.
18. Tozal, M.E.; Sarac, K. Palmtree: An ip alias resolution algorithm with linear probing complexity. *Comput. Commun.* **2011**, *34*, 658–669. [CrossRef]
19. Grailet, J.F.; Donnet, B. Towards a renewed alias resolution with space search reduction and IP fingerprinting. In Proceedings of the 2017 Network Traffic Measurement and Analysis Conference (TMA), Dublin, Ireland, 21–23 June 2017; pp. 1–9.
20. Spinelli, L.; Crovella, M.; Eriksson, B. AliasCluster: A lightweight approach to interface disambiguation. In Proceedings of the 2013 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS), Turin, Italy, 14–19 April 2013; pp. 127–132.
21. Yuan, F.; Liu, C. MLAR: Large-Scale Network Alias Resolution for IP Location. *J. Netw. Inf. Secur.* **2020**, *6*, 77–94.
22. Vermeulen, K.; Ljuma, B.; Addanki, V.; Gouel, M.; Fourmaux, O.; Friedman, T.; Rejaie, R. Alias resolution based on icmp rate limiting. In Proceedings of the Passive and Active Measurement: 21st International Conference, PAM 2020, Eugene, OR, USA, 30–31 March 2020; pp. 231–248.
23. Sherwood, R.; Bender, A.; Spring, N. Discarte: A disjunctive internet cartographer. In Proceedings of the ACM SIGCOMM 2008 Conference on Data Communication, Seattle, WA, USA, 17–22 August 2008; pp. 303–314.
24. Tao, Y.; Hu, G.; Hou, B.; Cai, Z.; Xia, J.; Fong, C.C. An alias resolution method based on delay sequence analysis. *Comput. Mater. Contin.* **2020**, *63*, 1433–1443. [CrossRef]
25. Spring, N.; Dontcheva, M.; Rodrig, M.; Wetherall, D. *How to Resolve IP Aliases*; Technical Report; University of Washington: Seattle, WA, USA, 2004.

26. Marder, A. APPLE: Alias pruning by path length estimation. In Proceedings of the Passive and Active Measurement: 21st International Conference, PAM 2020, Eugene, OR, USA, 30–31 March 2020; pp. 249–263.
27. Pansiot, J.J.; Grad, D. On routes and multicast trees in the Internet. *ACM SIGCOMM Comput. Commun. Rev.* **1998**, *28*, 41–50. [CrossRef]
28. CAIDA. Macroscopic Internet Topology Data Kit. 2022. Available online: https://www.caida.org/catalog/datasets/internet-topology-data-kit/ (accessed on 1 February 2022.)
29. Gueye, B.; Ziviani, A.; Crovella, M.; Fdida, S. Constraint-based geolocation of internet hosts. In Proceedings of the 4th ACM SIGCOMM conference on Internet measurement, Sicily, Italy, 25–27 October 2004; pp. 288–293.
30. Schapira, M.; Zhu, Y.; Rexford, J. Putting BGP on the right path: A case for next-hop routing. In Proceedings of the 9th ACM SIGCOMM Workshop on Hot Topics in Networks, Monterey, CA, USA, 20–21 October 2010; pp. 1–6.
31. Janardhana Rao, P.; Nageswara Rao, K.; Gokuruboyina, S.; Neeraja, K.N. An Efficient Methodology for Identifying the Similarity Between Languages with Levenshtein Distance. In Proceedings of the International Conference on Communications and Cyber Physical Engineering 2018, Hyderabad, India, 28–29 February 2018; pp. 161–174.
32. Vanaubel, Y.; Mérindol, P.; Pansiot, J.J.; Donnet, B. Through the wormhole: Tracking invisible MPLS tunnels. In Proceedings of the 2017 Internet Measurement Conference, London, UK, 1–3 November 2017; pp. 29–42.
33. Vanaubel, Y.; Luttringer, J.R.; Mérindol, P.; Pansiot, J.J.; Donnet, B. TNT, watch me explode: A light in the dark for revealing MPLS tunnels. In Proceedings of the 2019 Network Traffic Measurement and Analysis Conference (TMA), Paris, France, 19–21 June 2019; pp. 65–72.
34. MaxMind. GeoIP2. 2024. Available online: https://www.maxmind.com/en/geoip2-databases/ (accessed on 1 January 2024).
35. AIWEN-TECH. IPUU. 2024. Available online: https://mall.ipplus360.com/pros/IPVFourGeoDB/ (accessed on 1 January 2024).
36. IP2Location. IP2Location. 2024. Available online: https://www.ip2location.com/database/ (accessed on 1 January 2024).
37. Hastie, T.; Tibshirani, R.; Friedman, J.H.; Friedman, J.H. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*; Springer: Berlin/Heidelberg, Germany, 2009; Volume 2.
38. Orsini, C.; King, A.; Giordano, D.; Giotsas, V.; Dainotti, A. BGPStream: A software framework for live and historical BGP data analysis. In Proceedings of the 2016 Internet Measurement Conference, Santa Monica, CA, USA, 14–16 November 2016; pp. 429–444.
39. Hendriks, L.; Velan, P.; Schmidt, R.d.O.; de Boer, P.T.; Pras, A. Threats and surprises behind IPv6 extension headers. In Proceedings of the 2017 Network Traffic Measurement and Analysis Conference (TMA), Dublin, Ireland, 21–23 June 2017; pp. 1–9.
40. Waddington, D.G.; Chang, F.; Viswanathan, R.; Yao, B. Topology discovery for public IPv6 networks. *ACM SIGCOMM Comput. Commun. Rev.* **2003**, *33*, 59–68. [CrossRef]
41. Qian, S.; Wang, Y.; Xu, K. Utilizing destination options header to resolve IPv6 alias resolution. In Proceedings of the 2010 IEEE Global Telecommunications Conference GLOBECOM 2010, Miami, FL, USA, 6–10 December 2010; pp. 1–6.
42. Beverly, R.; Brinkmeyer, W.; Luckie, M.; Rohrer, J.P. IPv6 alias resolution via induced fragmentation. In Proceedings of the Passive and Active Measurement: 14th International Conference, PAM 2013, Hong Kong, China, 18–19 March 2013; pp. 155–165.
43. Luckie, M.; Beverly, R.; Brinkmeyer, W.; claffy, k. Speedtrap: Internet-scale IPv6 alias resolution. In Proceedings of the 2013 Conference on Internet Measurement Conference, Barcelona, Spain, 23–25 October 2013; pp. 119–126.
44. Padmanabhan, R.; Li, Z.; Levin, D.; Spring, N. UAv6: Alias resolution in IPv6 using unused addresses. In Proceedings of the Passive and Active Measurement: 16th International Conference, PAM 2015, New York, NY, USA, 19–20 March 2015; pp. 136–148.