Using network motifs to investigate the influence of network topology on PPM-based IP traceback schemes

Ankunda R. Kiremire *, Matthias R. Brust, Vir V. Phoha

Center for Secure Cyberspace, Louisiana Tech University, Ruston, LA, USA

ABSTRACT

Multiple schemes that utilize Probabilistic Packet Marking (PPM) have been proposed to deal with Distributed Denial of Service (DDoS) attacks by reconstructing their attack graphs and identifying the attack sources. By analyzing a set of PPM-based schemes, we show that past researchers have evaluated the schemes using disparate and often inadequate underlying topologies, which makes a side-by-side comparison of the scheme performance a complex problem.

To tackle this problem, we evaluate selected schemes on a large set of Internet-like topologies and adapt the network motif approach to provide a common ground for comparing the schemes’ performances in different network topologies. This approach allows us to determine the level of structural similarity between network topologies and consequently enables the comparison of scheme performance even when the schemes are implemented on different topologies.

Our results reveal that both the value of the PPM-based schemes’ convergence times, and their rankings vary depending on the underlying network topology. However the variation is considerably less when the topologies are compared within superfamilies of structural similarity. More specifically, the standard deviation in convergence times across the networks drops to about a tenth of its original value when the set of 28 networks are arranged in four superfamilies. To complement our results, we present an analytical model showing a link between scheme performance in any superfamily, and the motifs exhibited by the networks in that superfamily.

Our work proposes an effective way of comparing general network protocol performance in which the protocol is evaluated on specific representative networks instead of an entire set of networks.

1. Introduction

Internet Protocol (IP) traceback is an approach to dealing with Distributed Denial of Service (DDoS) attacks [1,2]. Using IP traceback, sources of Internet traffic, and attack traffic in particular, can be identified from the network traffic they generate. One technique for realizing IP traceback for flooding style DDoS attacks is Probabilistic Packet Marking (PPM), in which network routers embed their own identities in packets randomly selected from all the network traffic that the routers process [3]. In the event of an attack, the router identity markings present in the attack packets can be used to reconstruct the attack graph – the paths taken by attack traffic – and establish its sources [4]. The technique of probabilistically marking packets for IP traceback is the basis of many other schemes hereafter referred to as PPM-based schemes [4–7].
In the first part of this paper, we provide a detailed analysis of a set of PPM-based schemes. Using this analysis, one is able to compare the set of schemes using their fundamental features such as their convergence times and the underlying topologies used for simulation. This analysis yields two main insights, both of which make the direct comparison of the schemes’ performance results complicated. Firstly, different schemes are simulated on different kinds of underlying network topologies. Secondly, the majority of these topologies do not provide an adequate abstraction of the topology of the Internet [6,8,9]. The underlying topologies are typically tree-structured with a single path from an attacker to the victim. However, tree-structured underlying topologies ignore the prevalence of load balancing routers which have the effect of utilizing alternative routes between traffic sources and destinations [10]. PPM-based scheme simulations should therefore consider alternative routes in the underlying topologies which are ignored by using tree-structured topologies. Both the disparity in, and the inadequacy of the underlying topologies, raise questions about the performances of these schemes in the Internet. For example, which scheme would perform best when deployed on an appropriate network topology – a topology similar to the Internet’s? Does the performance of a PPM-based scheme vary with the topology on which it is implemented? If so, is it possible to provide a common ground for comparing scheme performance when the schemes are implemented on different topologies? These questions show that there is a need to evaluate and compare the schemes on common and appropriate network topologies. The results of this evaluation can then be used to determine which schemes are the most promising candidates for Internet deployment.

In the second part of this paper, we respond to these questions by investigating the relationship between scheme performance and the topology of the Internet. As a result, we are able to effectively compare scheme performance across different topologies. We achieve this by evaluating selected schemes on 28 distinct Internet-like topologies – topologies that have been shown to resemble the Internet’s architecture at the data plane [11–16]. These topologies are selected to encompass the variety of mathematical models that have been used by researchers in the past for Internet network protocol evaluation (cf. Fig. 1) [11,12,17,18]. While these models fall short of a thorough and accurate representation of the Internet, using them allows us to link scheme performance to the specific properties of each adopted network model. We then adapt the network motif technique to compare these topologies by capturing the subtle differences in graph structure between the different topologies [19,20]. The motif technique then leverages these differences to create superfamilies of structurally similar Internet-like network topologies. We are then able to use the superfamilies to provide a common ground on which scheme performance in different networks can be compared.

Results confirm that the performance of the schemes depends on the topology on which they are implemented. In fact, the ranking of scheme performance also varies with the network: the best performing scheme on one network topology is potentially the worst performing scheme on another topology. However, while the results show that scheme performance varies from one topology to another, the variation is considerably less if both topologies belong to the same superfamily. More specifically, the standard deviation in scheme convergence times drops by as much as 91.2% when the 28 topologies are compared within the four identified superfamilies.

To complement our results, we present an analytical model that shows how the motifs exhibited by a network topology possibly affect the performance of PPM-based schemes in that network. This model explains the link between network superfamilies and scheme performance that is observed in our results. We also analyze and perform simulations on five extra networks derived from two internet mapping projects [21–24].

In summary, the contribution of this paper is fourfold: (a) an analysis of PPM-based schemes is presented within the Background section, (b) the performance of selected schemes is evaluated and compared on an extensive set of Internet-like topologies, (c) network motifs are employed to identify the superfamilies in these topologies, and (d) we show a link between network motifs and the performance of PPM-based schemes both empirically and analytically.

The work presented herein has implications that reach outside the field of IP traceback. For example, do other network protocols also display a dependence on the motifs exhibited by the topology on which they are implemented? If so, can these dependencies be exploited to yield better protocol performance in specific types of networks? These questions should encourage multiple topology evaluation of network protocols. To this end, our work provides a way of reducing the network topology search space by selecting representative network topologies out of all possible Internet-like topologies.

The remainder of this paper is arranged as follows. The analysis of PPM-based schemes as well as other related work is presented in Section 2. This is followed by a description of our approach to solving the comparison and evaluation problem in Section 3. The theory and system model behind our work is analyzed in Section 4 and a simulation study is provided in Section 5. Results are presented and discussed in Section 6 and we conclude in Section 7.

2. Background

In this section, we discuss related work in the field of network motifs, and provide an analysis of selected PPM-based schemes.

2.1. Using subgraphs to differentiate networks

Milo et al. [19] introduce the concept of network motifs to compare arbitrary network topologies. In their seminal paper, network motifs are defined as the significantly prevalent subgraphs exhibited by a network. By identifying 3-node and 4-node motifs, it is possible to establish structural similarities among different networks from the fields of biology, technology, sociology, etc. They argue that the
network motifs are the fundamental building blocks of the networks, and as such different networks can be compared using them. Furthermore, the motifs are used to understand the underlying functions that generate each network.

Milo et al. [20] follow up this work with creation of superfamilies of networks. The authors study the similarities among networks based on their subgraph ratio profiles (SRPs). SRPs are based on the presence or absence of certain subgraph structures in comparison to randomized networks of the same size and connectivity. By considering all possible 3-node directed and 4-node undirected subgraphs, it is possible to compare networks from different fields by placing them into superfamilies based on the networks’ underlying structures.

Network motifs and SRPs have since then been used to compare different networks from fields such as social networks [25], neural networks [26], cooperative networks [27], protein interaction networks [28], and gene-regulation networks [29]. Additionally, some work has been done in improving the time efficiency of the process of counting network motifs [30].

To the best of our knowledge, our work is the first where the technique has been specifically adapted to identify the superfamilies in Internet-like networks.

2.2. PPM-based IP traceback schemes

PPM-based schemes consist of a marking scheme and a reconstruction procedure, and are based on the assumption that large amounts of traffic are used in a (D) DoS attack [3]. In their original work, Savage et al. [3] propose that the PPM marking scheme is employed at all times in all the routers in the network, while the reconstruction procedure is employed by the victim in the event of an attack. The marking scheme ensures that every router embeds its own identity in packets randomly selected from the packets the routers process during routing. Since a large number of packets is received in an attack, there is a considerable chance that a victim will have received packets with markings from all the routers that were traversed by the attack packets. The victim then employs the reconstruction procedure which uses the received marked attack packets to map out the attack graph – the paths from the victim to the attackers. The total number of received packets required to trace the attackers is referred to as the scheme’s convergence time.

Multiple PPM-based schemes have since been introduced. One example is the Tabu Marking Scheme (TMS) [9]. The author points out that PPM is prone to information loss as a result of re-marking. Re-marking occurs when a router randomly selects a packet which already has marking information from an upstream router, and consequently overwrites this information. TMS tackles this problem by ensuring that their marking scheme forfeits the marking opportunity in the event that the randomly selected packet contains previous marking information. As a result, they report lower convergence times than PPM for DDoS attacks.

In earlier work, we present an alternative scheme called Prediction Based Scheme (PBS) which also avoids re-marking [31]. However, in contrast to TMS, the PBS marking scheme ensures that the router information is embedded in the next available packet if the randomly selected...
packet already has marking information. The PBS marking scheme requires extra space cost of 1 bit compared to PPM. Additionally, the reconstruction procedure utilizes both legitimate and attack traffic to reconstruct the attack graph.

Wong et al. [6] present Rectified Probabilistic Packet Marking (RPPM). They point out that the reconstruction procedure used in PPM-based schemes has an imprecise termination condition. Typically, the analytical model in [3] is used to predict how many packets are required, but the model depends on the attack path length which is not known before the reconstruction is complete. Because the convergence time is considerably less than the total number of packets received during a typical attack, the victim is generally sure that the attack graph will be complete after analyzing all the received packets. However, a problem arises during short term attacks because the victim cannot tell if extra unique edges would be identified by receiving more packets. The authors present a mathematical formulation for a precise termination condition that enables complete attack graph reconstruction within a user-specified level of confidence.

Many other schemes have been proposed to increase the efficiency of PPM in different ways, e.g. [5,7,32,33]. Some of these schemes are presented in Table 1. The table compares ten PPM-based schemes in terms of features such as convergence time, underlying topologies, incremental deployment, re-marking, and upstream graph.

The convergence time refers to mathematical analysis for a single path scenario under uniform marking probability p and path length d. The expressions capture how many packets it would typically take to identify the entire path linking the victim to an attacker.

The feature incremental deployment refers to whether the scheme would be successful if the marking scheme is deployed on a fraction of the routers in the network. Only a few schemes explicitly state that they would be successful when partially deployed [3–5,32]. Incremental deployment means partial attack graph reconstruction is possible even when some ISP’s in the attack graph have not implemented the marking scheme on their routers.

Re-marking refers to whether the marking scheme at a router permits the overwriting of previous edge or router information in a packet. The majority of the considered schemes permit re-marking of packets [3–7,32,33]. The packet selection process at the routers that implement these schemes is completely random, which means that it is possible for a router to randomly select and consequently re-mark a packet that already has marking information from an upstream router.

Upstream graph refers to whether a scheme requires a previously obtained map of the network to successfully trace the specific path taken by attack traffic. Some of the works address how such a map can be obtained to aid in attack graph reconstruction [4,5,31]. Access to the map of a network allows for significantly improved performance since sections of the attack path can be inferred as opposed to being explicitly identified.

The underlying topology shows the different network topologies that are used for simulation purposes in those papers. The results from these topologies are used to provide an indication of how the schemes would perform if implemented in the Internet. More discussion of this feature is provided in a subsequent section.

The schemes considered therein are by no means an exhaustive study of all the PPM-based schemes in existence. However, the collection of schemes is large enough to show the discrepancy in, and inadequacy of, the underlying topologies which makes the direct comparison of scheme performance difficult.

It is important to point out that PPM-based schemes are not the only proposed approaches to IP traceback [1,2]. Alternatives include packet logging [34], specialized routing [35], Internet control message protocol (ICMP) traceback [36], deterministic packet marking [37] and hybrid approaches which combine different traceback techniques [38], or combine traceback with anomaly detection [39].

<table>
<thead>
<tr>
<th>Scheme</th>
<th>Year</th>
<th>Convergence time</th>
<th>Incremental deployment</th>
<th>Re-marking</th>
<th>Upstream graph</th>
<th>Underlying topology</th>
</tr>
</thead>
<tbody>
<tr>
<td>PPM [3]</td>
<td>2001</td>
<td>(\frac{\ln(d)}{p(1-p)^2})</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>SP/SA (max. 30 hops)</td>
</tr>
<tr>
<td>PPM-NPC [8]</td>
<td>2004</td>
<td>(\frac{\ln(d)}{p(1-p)^2})</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>SP/SA (10 hops)</td>
</tr>
<tr>
<td>TMS [9]</td>
<td>2005</td>
<td>(\frac{\ln(d)}{p(1-p)^2})</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Binary tree (6 hops, 32 sources)</td>
</tr>
<tr>
<td>FIT [5]</td>
<td>2005</td>
<td>Undetermined</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Skitter map (174,409 hosts, 5000 attackers)</td>
</tr>
<tr>
<td>RPPM [6]</td>
<td>2008</td>
<td>Undetermined</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>SP/SA, binary tree, random tree network (15, 100, 500, 1000 nodes)</td>
</tr>
<tr>
<td>TPM [7]</td>
<td>2008</td>
<td>Undetermined</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Skitter data (avg. 18 hops)</td>
</tr>
<tr>
<td>Randomize-and-link [32]</td>
<td>2008</td>
<td>(\frac{\ln(d)}{p(1-p)^2})</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Binary tree (10 hops)</td>
</tr>
<tr>
<td>IDPPM [33]</td>
<td>2010</td>
<td>Undetermined</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>SP/SA (20–32 hops)</td>
</tr>
<tr>
<td>PBS [31]</td>
<td>2012</td>
<td>(\frac{\ln(d)}{p})</td>
<td>No</td>
<td>No</td>
<td>Yes/No</td>
<td>SP/SA, SP/MA, MP/MA, 50 node network, 100 node network</td>
</tr>
</tbody>
</table>
2.2. Marking schemes

Table 1 shows that the different schemes contain different features that help to improve their performances in one way or another. Therefore, to facilitate the comparison of the different marking schemes in our simulations, it is imperative that the schemes are evaluated on the same “level”. The level selected for the uniform comparison of the schemes is their underlying algorithms. By considering the underlying algorithm, we disregard environment specific features such as router identity fragmentation, network dependent implementation details, and different confidence levels in attack graph construction. Consequently, we are able to categorize the marking schemes according to their underlying algorithms, and then select representative schemes from each category for simulation purposes. Additionally, we do not consider external factors such as complementary network traffic and traffic dynamics. As a result of these adjustments, the obtained results should not be taken as an absolute measure of the scheme performance in a network, but rather used as a relative measure between different schemes and/or different networks.

Despite their large number, PPM-based schemes have similar underlying algorithms in their marking schemes. The underlying algorithm is responsible for how the packets, in which the router identities are embedded, are selected. For example, the majority of the considered schemes exhibit underlying algorithms in which all routers randomly select packets with equal probability \( p \) \[3–7,32,33\]. The schemes in this category are prone to re-marking. We refer to this category as the re-marking category of PPM-based schemes. In the other category of schemes, the routers’ packet selection process is only partially random. The underlying algorithms in this category prohibit the overwriting of previous router information and as a result exhibit performances that are notably different from the re-marking category \[8,9,31\].

We select three representative marking schemes: PPM \[3\] to represent the re-marking category, and TMS \[9\] and PBS \[31\] to represent the non-re-marking category. The analytical models for these three schemes are markedly different from each other, even for equal marking probability, because of the differences in the schemes’ underlying marking algorithms, and yet representative of their respective categories. The performance of any PPM-based scheme can therefore be compared to either one of these schemes, or a combination of them.

Because of re-marking in PPM, the victim typically receives more markings from close-by routers than from distant routers. The chance of receiving a marked packet from a router \( l \) hops away is given by \( p(1 – p)^{l-1} \) where \( d \) is the attack path length. This is because a received marked packet indicates that packet was selected by a router (with probability \( p \)), after not being selected (with probability \( 1 – p \)) by all \( d – l \) previous routers. This analysis can be applied to all schemes in which markings from distant routers are more prevalent than markings from close-by routers.

In contrast to TMS, the PBS marking scheme compensates for the missed marking opportunities. Therefore the chance of receiving a marking from a router \( l \) hops from the victim is given by \( p \) for any router in the path. This analysis can be applied to all schemes in which the markings from the routers are equally prevalent regardless of their distance from the victim.

These three schemes therefore provide an adequate basis to understand the impact of the network topology on other PPM-based schemes.

2.4. Underlying topologies

Ideally, the performance of a network protocol such as a PPM-based traceback scheme would be evaluated on either the Internet itself, or a topology exactly like it. However, because the Internet is enormous, dynamic and heterogeneous, attempts to carry out empirical protocol evaluation are expensive and inflexible \[11\]. As a result, researchers resort to simulations implemented on underlying topologies which are considered to be simplified abstractions of the topology of the Internet \[11–13,16\]. In this case, an underlying topology is represented by a graph \( G(v,e) \) consisting of nodes \( v \) and edges \( e \) where the nodes represent either devices with routing capability or end hosts. An edge between any two nodes means that traffic can be directly transmitted between those two devices \[11\].

It is important to point out the difference between a network topology and a routing topology and how this difference affects our work. A network topology is an abstraction of the network’s physical infrastructure. It consists of all nodes and all the edges that are connected to and participate in the network. It represents all possible sources of network traffic and all possible routes that network traffic can use to get from any point in the network to any other point. On the other hand, a routing topology is a subset of the network topology. The routing topology consists of the nodes and edges that traffic typically traverses to get from one specific point in the network to another. The routing topology is therefore determined by the specific traffic sources and destinations being considered. Typically, a routing topology for traffic going from multiple sources to a single destination is tree-structured with the tree leaves representing the sources, and the tree root representing the destination. As we show later in this section, the majority of research in the area of PPM-based schemes has used tree-structured networks as the underlying topologies during their simulations. However, the prevalence of load-balancing techniques may cause the routing topology to strongly diverge from the tree structure since load-balancing ensures
that traffic takes more than a single unique path to its destination. This means that there is a need to reconsider the performance of previous schemes using non-tree-like underlying topologies. Using non-tree-like underlying topologies would more accurately capture the dynamics of these schemes and this is part of the motivation for our work.

A typical simulation is carried out as follows. During set up, the marking algorithm is implemented in the nodes (routers) of the underlying topology. To simulate the attack, packets are transmitted from one or more nodes (representing the attackers) to one specific node (representing the victim). A reconstruction procedure is then implemented at the victim to map out the attack graph \( G_{\text{act}} \). The resulting attack graph should consist of only the nodes and edges in the underlying topology that were directly involved in transmitting the attack packets.

As shown in Table 1, a variety of underlying topologies have been used to evaluate the performance of PPM-based schemes. The underlying topologies used range from simplistic to complex, as described below.

The single path, single attacker (SP/SA) is a simple topology consisting of a single attacker node sending packets along a single identical path to a single victim node. The length of the path varies with each work ranging from 3 hops to 32 hops [3,6,31,8]. This setup is used to simulate the performance of PPM schemes during a flooding style DoS attack.

The Single Path, Multiple Attacker (SP/MA), and Multiple Path, Multiple Attacker (MP/MA) topologies consist of multiple sources of attack traffic to simulate a DDoS attack. The SP/MA simulates a unique topology in which all the attackers are located at different distances from the victim but all along a single identical path [31]. The MP/MA simulates a more general topology where each attacker has a unique path linking it to the victim node. In some cases, the paths are completely independent [6], while in other cases, the paths merge closer to the victim [9,6,32,31].

One unique MP/MA topology is a tree, e.g., a binary tree. In this case, the attack graph is modeled as a tree with some or all the leaves at a certain depth representing the attack nodes, and the root of the tree representing the victim node [9,6,32]. This setup ensures that different attack paths merge closer they are to the victim. As with SP/SA and SP/MA, there is only one path in the attack graph from an attack node to the victim node.

Some authors have evaluated their schemes using actual data sets from the Internet [4,5,7]. These include traceroute data sets from Lucent Bell labs in [4] and CAIDA's skitter map in [5,7]. These data sets are used to produce topologies that are typically larger than the simple topologies mentioned thus far and provide better abstractions of the Internet structure. In this work, we have included five complementary networks into our network set, three of which are from the Caida project to provide a form of comparison for the rest of the network set.

One common feature with these underlying topologies is their tree-like structure. A tree-structured topology \( G_{\text{tree}} \) exhibits a single path from any given attacker to the victim. The choice of tree-structured topologies is based on the assumption that all attack traffic from one attacker will take the same path to the victim. This assumption is in turn based on the observation that Internet paths are largely invariant particularly over short periods of time [40]. These assumptions have allowed researchers to simplify the simulation process by ignoring the routing and load balancing capabilities of the network and enforcing a predefined (or pre-observed) set of paths for attack traffic. However, the prevalence of load balancing routers in the Internet today [10,41], makes the assumption of a tree-structured topology an unrealistic one. Augustin et al. report that 39–70% of the routes measured in [10] exhibit route fluttering as a result of load balancing. Load balancing routers frequently forward traffic along alternative paths in order to minimize cost to the network. Consequently, scheme performance in tree-structured topologies, where all traffic from one source takes one path, cannot be used as an indication of how those schemes would perform in Internet-like network topologies.

In earlier work, two well-connected albeit small networks are considered [31]. In contrast to the tree-like networks \( G_{\text{tree}} \) typically considered in PPM-based schemes, well-connected networks contain alternative routes between attackers and any victim. Simulations carried out in well-connected networks, where routers make routing decisions as well as marking decisions, more closely capture the performance of the schemes if they were deployed in the Internet. In this work, we follow up by considering a larger number of network models to investigate the marking schemes.

We consider the models that have been used in the past to simulate the Internet topology [11,12,17,18,15]. These models fall into three categories based on the Internet properties that they emphasize, namely degree-based models, structural models and spatial models (cf. Fig. 1). The emphasis of degree-based models is the degree distribution of the nodes in an attempt to recreate the power law observations in the Internet [13,15]. The structural models arrange the nodes to mimic the hierarchical structure of the Internet, with Internet traffic being transmitted through routers located within autonomous systems [11,16]. The spatial models place emphasis on the location of the nodes with any two nodes being connected only if they are within a transmission range of each other [18]. The three categories of models are used to create 28 Internet-like topologies which are then used to provide a clearer picture of the performance of PPM-based schemes in an Internet-like environment.

Using the mathematical models to create underlying topologies for simulation allows us to link scheme performance to the structural characteristics exhibited by a category of networks. For example, a pattern in scheme performance in the degree-based networks (such as the Barabasi and Waxman networks) could be potentially linked to the power law in the Internet. In contrast, a pattern in the structural networks (such as the Top-Down networks) could be linked to the hierarchical structure of the Internet. An actual Internet topology dataset would not lend itself easily to such analysis because it exhibits all these characteristics and therefore attributing scheme performance to one specific characteristic would be more difficult.
We must point out that the network models do not provide a completely accurate description of the Internet topology. The process of capturing and modeling the topology of the Internet is not only a complex process but is also an ongoing one with many unresolved challenges, the details of which are beyond the scope of this paper. However, the network set used in this paper is sufficient for the purposes of scheme performance comparison, and providing a benchmark for further studies about scheme dependence on topologies.

3. Approach

In this section, we explain the relationship between network motifs and PPM-based traceback schemes. Additionally, we describe how network motifs and significance ratio profiles (SRPs) are used to identify the superfamilies from a set of networks.

3.1. Alternative paths in attack graphs and network motifs

Given a DDoS attack is carried out on a well-connected underlying topology, there is a chance that the attack graph returned by the reconstruction procedure will not be tree-structured [6]. While a tree-structured attack graph is a realistic assumption under typical traffic patterns [3,40], it becomes a less realistic assumption when congestion close to the victim causes unusual traffic patterns. This congestion of network resources is a result of the increased network traffic as multiple attack traffic streams merge close to the victim. In this case, there is an increased chance of routers forwarding traffic along alternative paths to deal with the congestion [42,43]. This calls for re-analyzing and re-modeling the performance of PPM-based scheme in large-scale topologies. Unfortunately, considering multiple alternative paths means that attack paths, even from a single attacker, could take on multiple possible shapes. This exacerbates the combinatorial problem of modeling a complete attack graph in a network topology. However, in this work we show that network motifs provide a simple way of modeling and investigating the effect of alternative paths on PPM-based schemes. These alternative paths have significantly different probabilities of identification compared to typical tree-structured attack graphs. In the following section, we show that these differences in probability translate to considerable variation in scheme performance from one network to another depending on the motifs exhibited by those networks.

The six possible undirected 4-node subgraphs found in any network topology are shown in Fig. 2. Four of the six 4-node subgraphs exhibit alternative paths that are either the same distance as the original path or differ by two hops at most. These subgraphs are identified by IDs 3, 4, 5 and 6 in Fig. 2. We model the influence of these subgraph structures on the reconstruction procedure by considering their existence along an arbitrary attack path. This assumption is realistic particularly in the case when the considered subgraph is a motif in the network. Since the network motifs are the subgraphs significantly prevalent in a network, there is a considerable chance that an attack graph derived from that network consists of the network’s motif.

While the mentioned assumptions make the model tractable, these same assumptions make the model, in its current form, unsuitable for direct estimation of convergence times for all possible schemes in all possible network topologies under all possible DDoS attack scales. However, the model can be used to understand why the schemes perform the way they do in large-scale networks.

3.2. Extracting the SRPs and network motifs

Subgraph ratio profiles (SRPs) and network motifs are used to capture the subtle differences in graph structure between various network topologies [19,20]. The differences and similarities can then used to provide a common ground for comparing network protocol performance between structurally similar network topologies. In this case, the network protocol is PPM-based IP traceback schemes.

The SRP of a network topology is a plot showing the distribution of the network’s subgraphs relative to a randomized network of the same size and connectivity. We refer to the network whose SRP we seek to extract as the test network. For each subgraph $i$, a relative score $R_i$ is calculated from $R_i = \frac{N_{test,i} - N_{rand,i}}{N_{rand,i} + N_{rand,i}}$ [20]. In this expression, $N_{test,i}$ is the number of times that subgraph $i$ appears in the test network, while $N_{rand,i}$ is the average number of times it appears in similar randomized networks. The randomized networks are created by randomly assigning edges to an equal-sized network while maintaining the degree sequence of the test network. Comparing the test network to similar randomized networks identifies the subgraphs in the test network that are statistically prevalent or absent.

![Fig. 2. All six possible 4-node undirected subgraphs and their IDs. Subgraphs 3, 4, 5 and 6 exhibit alternative paths between any two of their nodes.](image)
As in [20], \( \varepsilon \) is set to 4 to eliminate the possibility of large \( R_i \) values with low \( N_{\text{test},i} \) and \( N_{\text{rand},i} \) values.

Each test network is then assigned a six value SRP vector by normalizing all its \( R_i \) values according to the expression

\[
\text{SRP}_i = \frac{R_i}{\sqrt{\sum_{i=1}^{l} R_i^2}} \quad [20].
\]

Positive SRP values indicate that subgraph \( i \) is prevalent in the test network, while negative SRP values suggest an absence of subgraph \( i \) in that network.

Network motifs are the subgraphs in a network that are significantly prevalent when compared to a randomized network of the same size and connectivity [19]. We identify the 4-node motifs of the networks by using the subgraphs’ Z-scores as a measure of their significance.

The Z-score \( Z_i \) of any subgraph \( i \) in a network is derived from the expression

\[
Z_i = \frac{N_{\text{test},i} - N_{\text{rand},i}}{\sigma_{\text{rand},i}} \quad [19]
\]

where \( \sigma_{\text{rand},i} \) is the standard deviation of subgraph \( i \) in the randomized network. A subgraph \( i \) is considered a motif of a network if its Z-score is significant, i.e. \( Z_i \geq 3.0 \) [19].

### 3.3. Identifying the superfamilies

The networks are then grouped based on similarities between their SRPs. Like selecting clusters from a dataset, determining the number of superfamilies from the networks’ SRPs is an ambiguous problem. An approximate number of superfamilies can be obtained by plotting a correlation map of all the network SRPs. Such a map shows the level of similarity between different SRPs as captured by their correlation coefficients. Arranging the map using a k-nearest neighbor algorithm ensures that similar SRPs are placed next to each other, and dissimilar SRPs are placed more distant. This allows for a visual inspection to determine an approximate number \( m_0 \) of clusters.

Compared to \( m_0 \), a more appropriate number of clusters can be determined by running a k-means clustering algorithm on the SRPs for different numbers of clusters \( m \) [45]. The appropriate number of clusters \( k \) is then determined by considering the error \( \varepsilon(m) \) of the different numbers of clusters \( m \). Given \( n \) SRPs, and the absolute intra-cluster error \( \varepsilon(m) \), the number of groups \( k \) is derived from

\[
k = \min\{m | m \in [2, n] \cap \sum_{i=1}^{m-1} \varepsilon(m) \leq \varepsilon \},
\]

where we set the limiting error \( \varepsilon = 10\% \). This expression enables us to find a balance between accuracy and appropriate cluster size. The identified clusters are referred to as superfamilies.

### 4. System model and analysis

In this section, we propose a model which shows how subtle differences in the network topologies contribute to differences between the convergence times of the PPM-based schemes. The structural differences contribute to two factors, namely alternative paths and merging of attack streams. These factors affect each scheme uniquely and yet their level of influence also varies depending on the network topology. As a result, the performance of different schemes will be very similar in one network, and yet very dissimilar in another network.

To illustrate these factors, we refer to Fig. 3 which shows an attack stream taking a path from Node A to the victim at Node I. In Fig. 3(a), the attack stream takes a single route to the victim. In Fig. 3(b), the attack stream from node A takes two routes to get to the victim. By definition, an attack stream could consist of network packets from any number of upstream sources, as long as the packets are being forwarded along the same path to the same destination. However, for simplicity, we consider the stream from Node A as having originated at a single attacker located at A.

#### 4.1. Traditional analytical model

Originally, the convergence time for PPM-based schemes has been modeled as the coupon-collector’s problem [46,3]. In the classic problem, a coupon collector seeks to collect \( d \) equally likely distinct coupons by drawing them from an urn with replacement. While it takes a short time to get the first few unique coupons, it takes considerably longer to get the last few coupons that complete the entire collection. The expected number of turns needed to draw all \( d \) distinct coupons grows as \( \Theta(d \cdot \ln(d)) \) [46].

When the coupon collector problem is applied to packet marking, the marked packets are taken to be the coupons. For example, Fig. 3(a) shows a single path linking attacker A to victim I and the target of the “coupon collector” would be to collect markings for all 7 edges. However, the expected time expression above cannot be directly applied to the packet marking problem for two reasons. Firstly, while one is guaranteed to pick a coupon with each draw in the coupon collector problem, one may or may not “draw” a marked packet in the packet marking problem. Secondly, while the coupons in the classic problem have equal chances of being drawn, the marked packets have unequal chances of being received. Savage et al. [3] deal with the unequal edge probabilities by utilizing the probability of the least likely edge to provide an upper bound on the expected convergence time.

Formally, given a single path of \( l \) hops implementing the PPM scheme with router marking probability \( p \), the least likely edge is typically the edge located closest to the attacker which has a probability \( p(1-p)^{l-1} \) of being received by the victim. Given \( d \) unique markings, the probability of receiving any marking at the victim is therefore at least \( dp(1-p)^{l-1} \) which is the product of the number of unique markings and the probability of the least likely edge. The expected number of packets \( E[X] \) required to complete the marking “collection” in order to build the attack graph is derived by dividing the original coupon collector expectation by \( dp(1-p)^{l-1} \) which yields Eq. (1) below [3].

\[
E_{0,\text{PPM}}[X] < \frac{\ln(d)}{p(1-p)^{l-1}}
\]

(1)

Therefore, the traditional expression for the upper bound of the expected convergence time is obtained by dividing the natural logarithm of the number of distinct edges \( d \), by the probability \( p(1-p)^{l-1} \) of the least likely
edge in the attack path. For the SP/SA topology, the number of hops is equal to the number of unique markings \((l = d)\).

Similarly, the convergence time expressions for TMS and PBS are given by Eqs. (2) and (3) respectively. In these expressions, the probability of the least likely edge in an SP/SA is given by \(p(1 − p)^{l−1}\) for TMS and \(p\) for PBS. However, in contrast to PPM, the least likely edge for TMS and PBS is the edge located closest to the victim.\(^1\)

\[
E_{0,\text{TMS}}[x] < \frac{\ln(d)}{p(1 − p)^{l−1}} \tag{2}
\]
\[
E_{0,\text{PBS}}[x] < \frac{\ln(d)}{p} \tag{3}
\]

4.2. The effect of motifs on the analytical model

One hitherto unstudied factor that affects convergence time is the alternative paths that traffic might take. To understand the influence of the alternative paths factor, we consider an attack graph containing a subgraph which exhibits alternative paths. Fig. 3(b) shows such an attack graph linking attacker A to victim I in which the attack traffic takes one of two paths FGI or FHI with probability \(a\) and 1 \(-\) \(a\) respectively. The nodes F, G, H and I in this attack graph form Subgraph 4. While the attack path length \(l\) is unchanged (from Fig. 3(a) to (b)), the probability and the location of the least likely edge is considerably altered and consequently, the convergence time is changed.

The alternative path factor \(a\), is affected by a variety of factors such as the presence of load balancing routers in the network, the number of alternative paths available, the amount of traffic being processed at node F, as well as the bandwidth and latency values for the alternative paths. For the analysis in this section, we assume node F has load balancing capability, and the routes can sustain the traffic being forwarded through them.

Consider the case of PPM. In Fig. 3(a), the least likely edge is AB with a probability of \(p(1 − p)^{l−1}\). However, the probability of receiving edge FG in Fig. 3(b) is given by \(ap(1 − p)\) which is considerably less than the probability of AB for short path lengths.\(^2\) In this case, the convergence time is given by Eq. (4).

\[
E_{1,\text{PPM}}[x] < \frac{\ln(d)}{ap(1 − p)} \tag{4}
\]

Comparing Eqs. (1) and (4) reveals that the convergence time is increased by a factor of \(\frac{1}{1-p}\). This means that even in the best case when both alternative paths are equally likely \((a = 0.5)\), the convergence time of a 3-hop attack graph is multiplied by a factor of 1.92 while a 15 hop attack graph is multiplied by a factor of 1.18. If one of the two paths only carries a tenth of the traffic \((a = 0.1)\), the convergence times of the 3-hop and 15-hop attack graphs is multiplied by a factor of 9.6 and 5.88 respectively. This shows that, for PPM, the alternative paths factor affects short attack paths more than long attack paths.

Consider the case of TMS. The least likely edge in Fig. 3(a) would be GI with a probability of \(p(1 − p)^{l−1}\). When alternative paths are considered in Fig. 3(b), the probability of GI is reduced even further to \(ap(1 − p)^{l−1}\) which means the convergence time is given by Eq. (5).

\[
E_{1,\text{TMS}}[x] < \frac{\ln(d)}{ap(1 − p)^{l−1}} \tag{5}
\]

Comparing Eqs. (2) and (5) shows that the convergence time is increased by a factor of \(\frac{1}{ap}\) regardless of the path length. This means that if both alternative paths are equally likely \((a = 0.5)\), the convergence time is doubled and if one path only takes a tenth of the traffic \((a = 0.1)\), the convergence time is increased by a factor of 10.

The PBS case is very similar to the TMS case. The least likely edge in Fig. 3(b) is GI and its probability changes from \(p\) to \(ap\) when one considers the alternative paths. Consequently, the convergence time changes to Eq. (6).

\[
E_{1,\text{PBS}}[x] < \frac{\ln(d)}{ap} \tag{6}
\]

Comparing Eqs. (3) and (6) shows that the convergence time is increased by factor of \(\frac{1}{ap}\) regardless of the path length which is similar to the TMS case.

\(^1\) While PBS typically exhibits equal probability for all edges in the attack path, they mention a saturation condition that makes probabilities of edges closer to the victim less likely than the rest. This condition occurs for either long path lengths, or high marking probabilities.

\(^2\) In this scenario, a short path is any path less than 18 hops long. This is evaluated from \(p(1 − p)^{l−1} < ap(1 − p) \Leftrightarrow l > 18\). The limit of \(l\) is evaluated for equal chance of taking either route \((a = 0.5)\) and a marking probability \(p = 0.04\).
This shows that alternative paths reduces the probability of the least likely edges in an attack path for all the considered schemes, and consequently increases their convergence times. However, their impact on the convergence times is higher in TMS and PBS than it is in PPM.

4.3. The effect of path merging on the analytical model

Another factor that comes into play in the convergence time is the merging of attack streams as different attack paths get closer to the victim [6]. As a result of this merging, the probabilities of downstream edges in an attack path are increased which affects its convergence time. To understand its influence, consider the attack graph in Fig. 3(b) where two other attack streams from attackers A′ and A″ contribute an equivalent amount to the traffic flowing out of node F towards victim I. Because of the increased traffic flowing through edges FG, FH, GI and HI, there is an increased chance of receiving markings from those edges which in turn affects the reconstruction time of the attack path of attacker A.

Consider the PPM case. With just attacker A and short attack paths, we showed that the least likely edge is FG with a probability of \( ap(1 - p) \) and the convergence time is given by Eq. (4). However, with attackers A′ and A″, the traffic going through FG is 3 times as high and so is the probability of receiving its marking. Formally, given \( n \) equivalent attack streams merge before node F, the probability of receiving FG changes to \( nap(1 - p) \). This means that when \( na \geq 1 \), edge FG is no longer the least likely edge in the attack path from attacker A. In this case, the least likely edge reverts to edge AB whose probability is still \( p(1 - p)^{-\frac{1}{3}} \) and consequently the convergence time expression reverts to Eq. (1). This means that the merging of attack streams offsets the alternative paths factor.

Consider the TMS case. With attacker A, we showed that the alternative paths reduced the probability of edge GI to \( ap(1 - p)^{-\frac{1}{3}} \). The increased traffic from A′ and A″ increases this value to \( nap(1 - p)^{-\frac{1}{3}} \), which in turn nullifies the influence of the alternative paths when \( na = 1 \). However when \( na > 1 \), edge GI ceases to be the least likely edge. In this condition, the least likely edge is the edge closest to the victim whose traffic and probability are unaffected by the merging attack streams. In Fig. 3(b), this happens to be edge EF. The probability of receiving edge EF is \( p(1 - p)^{-\frac{1}{3}} \) which means the convergence time changes from Eqs. (5)–(7).

\[
E_{TMS}\{X\} < \frac{\ln(d)}{p(1 - p)^{-\frac{1}{3}}} 
\]  

(7)

Comparing (2) and Eq. (7) shows that the merging of attack streams not only nullifies the alternative paths factor, but also reduces the convergence time. The amount by which the convergence time is reduced depends on how close the “new” least likely edge is to the attacker. The closer the new least likely edge is to the attacker, the more the reduction in convergence time, and vice versa. In this particular scenario, the new least likely edge (EF) is two positions away from its original position (GI) and the convergence time reduces to \((1 - p)^{\frac{1}{2}}\) of its original value. Given \( p = 0.04 \), the convergence time is reduced by 8% of its original value.

Analysis of the PBS scheme yields insights similar to those gained from the analysis of the TMS scheme. Given a similar scenario, the probability of receiving edge GI increases from \( ap \) to \( nap \). As with the TMS scheme, the merging of the attack streams cancels out the alternative path effect and reduces the expected convergence time.

This shows that the merging of attack streams offsets the alternative path effect for all the considered schemes. However, while the merging simply cancels out the alternative path effect in PPM, it reduces the convergence time for TMS and PBS.

The model presented thus far considers subgraph 4 which exhibits 2 alternative paths. Table 2 shows the probabilities for the least likely edges for subgraphs 3, 4, 5 and 6 under various traffic conditions. The probabilities in the “Original” column show the expressions when both alternative paths and merging of attack streams are ignored. The probabilities under the “Number of merging streams” column show these same probabilities when one considers the given subgraphs with different number of merging attack streams. As in Fig. 3(b), we consider an attack graph where the different attack streams merge at the node just before the subgraph. Additionally, the victim node is part of the subgraph and the probability of taking any of the alternative paths is equal. While this set up is specific, the analysis obtained from it can be used to describe the influence of both alternative paths and merging attack streams in a larger network. The probabilities in the table reveal that it takes more merging attack streams to offset the alternative paths in subgraphs 5 and 6 than in subgraphs 3 and 4. This is because subgraphs 5 and 6 exhibit more alternative paths and consequently require more attack traffic to offset the drop in probability caused by the alternative paths. For example, with 2 merging streams in subgraph 4 and PPM, the probability of the least likely edge is \( p(1 - p)^{\frac{1}{3}} \), which is the same as the original probability. However, with subgraph 6 the probability of the least likely edge under the same conditions is \( 2ap(1 - p) \). In fact, it takes 4 attack streams to increase that probability back to \( p(1 - p)^{\frac{1}{3}} \). A higher probability for the least likely edge translates to a lower value for the convergence time and a lower probability for the least likely edge translates to a higher convergence time.

In summary, the model and analysis presented here shows that alternative routes reduce the probability of specific edges in the attack graph and as a result increase the convergence time for those attack graphs particularly for TMS, PBS and short attack paths implementing PPM. The merging of attack streams offsets this effect in PPM, TMS and PBS. However in TMS and PBS, the merging has the added effect of reducing their convergence times.

5. Simulation study

5.1. Network models and topologies

Five network models from three categories are considered to simulate the Internet topology. The categories are
Table 2
The table shows 4-node subgraphs and the probabilities of the least likely edge for the different marking schemes for n merging attack streams, along side the original probability of the least likely edge given no subgraphs or convergence. The marking probability is denoted by p, the path length by l, the probability of taking alternative routes denoted by a, b, c, with the expressions for PPM, TMS, and PBS shown. For simplicity, it is assumed the probability of taking alternative routes is equal. The convergence time of the marking scheme is indirectly proportional to the lowest probability.

<table>
<thead>
<tr>
<th>IDs</th>
<th>Subgraphs</th>
<th>Scheme</th>
<th>Original</th>
<th>Number of merging streams</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>n = 1</td>
</tr>
</tbody>
</table>

degree based, structural and spatial models which emphasize the scale-free, hierarchical, and spatial properties of the Internet respectively.

The degree-based networks are implemented from the Barabási model [14] using the Brite topology generator [16]. This model simulates the preferential attachment and incremental growth of the Internet while maintaining the power law observed in the Internet. The probability $P(i,j)$ that a node $i$ connects to node $j$ upon joining the network is given by $P(i,j) = \frac{d_j}{\sum_{i \in V} d_i}$ [16]. Here, $d_j$ is the degree of node $j$ and $V$ is the set of all nodes already in the network.

The Waxman model [15] is also considered to simulate degree-based networks using Brite topology generator [16]. This model adds network specific characteristics (such as using a probability function to connect two nodes) to the Erdős–Renyi random graph model [47]. The probability $P(i,j)$ of connecting two nodes $i$ and $j$ distance apart, in the network is given by $P(i,j) = \alpha e^{-\beta d}$ with $l_{max}$ being the maximum distance between any two nodes and $0 < \alpha, \beta \leq 1$[16].

Spatial networks are implemented using unit disk graphs [18] by randomly placing 1000 nodes in a square of length 1000 units and connecting any two nodes within a distance of $r$ units from each other. Additionally, we ensure that the network is fully connected so that any two nodes can communicate with each other.

Structural networks are implemented using a Top-down hierarchical model provided by Brite topology generator [16]. In this model, the AS level networks are implemented using a Waxman model. Router level networks are then created for each AS level node also using the Waxman model. This way, we are able to simulate a two level network with $u$ AS’s and $d$ routers in each AS.

To complement our results, we also investigated three networks derived from the Caida project [21,22] and two networks from the Rocketfuel project [23,24]. Because these networks were observed and not created using any mathematical model, we have no control over their size. As a result, these networks are bigger and yet not completely connected, and therefore we did not directly compare their results to the 28 other networks, but discussed their results in Section 6.4.

5.2. Settings

Table 3 shows the 28 considered networks which are based on five network models. The group of networks consists of 6 networks built from a structural model, 5 networks built from a spatial model, and 17 networks built from various degree based models, where each network consists of 1000 nodes. The nodes in these topologies represent devices operating at the Internet layer of the TCP/IP model, or the network layer of the OSI model (eg. routers). Two nodes in any network are connected by an edge if Internet traffic can be directly transmitted between them without being forwarded by any intermediate nodes. The network models are used to determine which nodes are connected by edges, and hence the overall number of edges is based on specific parameter settings.

The network topologies are then imported into NS2 [48], which was selected as our simulation tool.

Out of the 1000 nodes, 50 nodes are randomly selected to be attackers, and one node is selected to be the victim. CBR sources of traffic are placed at the attackers to create traffic that is sent to the victim in a simulation of a DDoS attack. The number of packets required to successfully identify the entire path traversed by the packets on their way to the victim is then counted. This number is averaged over 100 DDoS and traceback simulations in each network to give traceback values that more accurately capture the performance of the selected schemes in those networks.

The 28 mathematical networks are all of the same size, and experience DDoS attacks of the same size. This is done so that structural and traceback comparison is straightforward. It is important to note that our simulations do not
consider cross traffic, i.e. non-attack traffic traversing the network at the same time as the attack. We believe that its exclusion allows us to compare the schemes in such a way that any differences in performance between networks are solely attributed to topological differences.

The Barabási model [14] is used to create the networks BA1–5 using the simulation set \((e, d_{\text{max}}, m)\) of the number of edges \(e\), the maximum out degree \(d_{\text{max}}\), and the number of new links per new node \(m\) as in Table 3.

The Waxman model [15] is used to create both Router level networks RN1–6 and AS level networks AS1–6 described by the set \((e, d_{\text{max}}, x, \beta, m)\) in the settings in Table 3.

The Unit disk graph model is used to create the networks TD1–6 which are described by the set \((e, d_{\text{max}}, u, d)\) in the settings in Table 3. For both the AS and Router level networks involved, we set \(x = 0.15, \beta = 0.2\).

Caida1–3 are the complementary networks derived from the Caida project [21,22]. Caida1 has 3451 nodes and 4048 edges, Caida2 has 3537 nodes and 4150 edges, while Caida3 has 3527 nodes and 4143 edges. Rocket1 and Rocket2 are the complementary networks derived from the Rocketfuel internet mapping project [23,24]. Rocket1 has 3727 nodes and 4504 edges while Rocket2 has 10,333 nodes and 12,912 edges. The attack simulations carried out in these networks also consisted of 50 randomly selected attackers sending traffic to one victim.

### 6. Results and discussion

#### 6.1. SRPs and motifs

In this subsection, we discuss the results derived during the process of identifying the SRPs of and the motifs in the 28 considered networks. The SRPs of given networks can be used to determine the level of similarity between those networks. The motifs exhibited in any network provide an indication of what subgraphs are likely to appear in an attack graph derived from that network.

The SRPs for the considered networks are extracted as described in Section 3.2. By considering the six possible undirected 4-node subgraphs from the Caida project [21,22], Caida1 has 3451 nodes and 4048 edges, Caida2 has 3537 nodes and 4150 edges, while Caida3 has 3527 nodes and 4143 edges. Rocket1 and Rocket2 are the complementary networks derived from the Rocketfuel internet mapping project [23,24]. Rocket1 has 3727 nodes and 4504 edges while Rocket2 has 10,333 nodes and 12,912 edges. The attack simulations carried out in these networks also consisted of 50 randomly selected attackers sending traffic to one victim.

The SRPs for the considered networks are extracted as described in Section 3.2. By considering the six possible undirected 4-node subgraphs from the Caida project [21,22], Caida1 has 3451 nodes and 4048 edges, Caida2 has 3537 nodes and 4150 edges, while Caida3 has 3527 nodes and 4143 edges. Rocket1 and Rocket2 are the complementary networks derived from the Rocketfuel internet mapping project [23,24]. Rocket1 has 3727 nodes and 4504 edges while Rocket2 has 10,333 nodes and 12,912 edges. The attack simulations carried out in these networks also consisted of 50 randomly selected attackers sending traffic to one victim.

<table>
<thead>
<tr>
<th>Model</th>
<th>Name</th>
<th>Settings</th>
<th>Average shortest path length</th>
<th>Motif ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>Barabási ((e, d_{\text{max}}, m))</td>
<td>BA1</td>
<td>1997, 49, 2</td>
<td>4.2459</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>BA2</td>
<td>999, 75, 1</td>
<td>6.5069</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>BA3</td>
<td>2994, 91, 3</td>
<td>3.4648</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>BA4</td>
<td>3990, 134, 4</td>
<td>3.1602</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>BA5</td>
<td>4985, 137, 5</td>
<td>2.9611</td>
<td>–</td>
</tr>
<tr>
<td>Top-down ((e, d_{\text{max}}, u, d))</td>
<td>TD1</td>
<td>2020, 17, 10, 100</td>
<td>9.4500</td>
<td>3, 4, 5</td>
</tr>
<tr>
<td></td>
<td>TD2</td>
<td>2020, 16, 10, 100</td>
<td>9.3589</td>
<td>3, 4, 5</td>
</tr>
<tr>
<td></td>
<td>TD3</td>
<td>2040, 13, 20, 50</td>
<td>9.9047</td>
<td>3, 4, 5</td>
</tr>
<tr>
<td></td>
<td>TD4</td>
<td>2007, 17, 5, 200</td>
<td>8.9343</td>
<td>3, 4, 5</td>
</tr>
<tr>
<td></td>
<td>TD5</td>
<td>2050, 14, 25, 40</td>
<td>9.9141</td>
<td>3, 4, 5, 6</td>
</tr>
<tr>
<td></td>
<td>TD6</td>
<td>2100, 13, 50, 20</td>
<td>9.4371</td>
<td>3, 4, 5, 6</td>
</tr>
<tr>
<td>AS level ((e, d_{\text{max}}, x, \beta, m))</td>
<td>AS1</td>
<td>1000, 11, 0.15, 0.2, 1</td>
<td>10.5902</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>AS2</td>
<td>2000, 20, 0.1, 0.2, 2</td>
<td>4.9483</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>AS3</td>
<td>2000, 16, 0.2, 0.2, 2</td>
<td>4.9198</td>
<td>3, 4</td>
</tr>
<tr>
<td></td>
<td>AS4</td>
<td>2000, 21, 0.3, 0.2, 2</td>
<td>4.9436</td>
<td>3, 4</td>
</tr>
<tr>
<td></td>
<td>AS5</td>
<td>2000, 18, 0.4, 0.2, 2</td>
<td>4.8965</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>AS6</td>
<td>2000, 20, 0.5, 0.2, 2</td>
<td>4.9350</td>
<td>3, 4</td>
</tr>
<tr>
<td>Router level ((e, d_{\text{max}}, x, \beta, m))</td>
<td>RN1</td>
<td>2000, 23, 0.15, 0.2, 2</td>
<td>4.9082</td>
<td>3, 4</td>
</tr>
<tr>
<td></td>
<td>RN2</td>
<td>2000, 20, 0.1, 0.2, 2</td>
<td>4.9594</td>
<td>3, 4</td>
</tr>
<tr>
<td></td>
<td>RN3</td>
<td>2000, 23, 0.2, 0.2, 2</td>
<td>4.8975</td>
<td>3, 4</td>
</tr>
<tr>
<td></td>
<td>RN4</td>
<td>2000, 17, 0.3, 0.2, 2</td>
<td>4.9615</td>
<td>3, 4</td>
</tr>
<tr>
<td></td>
<td>RN5</td>
<td>2000, 22, 0.4, 0.2, 2</td>
<td>4.9109</td>
<td>3, 4</td>
</tr>
<tr>
<td></td>
<td>RN6</td>
<td>2000, 20, 0.5, 0.2, 2</td>
<td>4.9472</td>
<td>3, 4</td>
</tr>
<tr>
<td>Unit disk graphs ((e, d_{\text{max}}, r, \rho))</td>
<td>AH1</td>
<td>4658, 22, 56, 9.9</td>
<td>13.1797</td>
<td>3, 5, 6</td>
</tr>
<tr>
<td></td>
<td>AH2</td>
<td>3716, 15, 50, 7.9</td>
<td>15.4724</td>
<td>3, 5, 6</td>
</tr>
<tr>
<td></td>
<td>AH3</td>
<td>2874, 13, 47.2, 7.0</td>
<td>18.0712</td>
<td>3, 5, 6</td>
</tr>
<tr>
<td></td>
<td>AH4</td>
<td>7264, 27, 70, 15.4</td>
<td>9.6797</td>
<td>3, 5, 6</td>
</tr>
<tr>
<td></td>
<td>AH5</td>
<td>9271, 32, 80, 20.1</td>
<td>8.1647</td>
<td>3, 5, 6</td>
</tr>
</tbody>
</table>
subgraphs 1 and 2. A few of the networks, however, show negative SRP values for subgraphs 3, 4, and 5.

The networks are also examined to identify their 4-node motifs. The identified motifs are presented in Table 3 under the Motif ID column. The results show that 6 of considered networks do not possess any motifs. This means that even though those networks possess all six undirected 4-node subgraphs, the subgraphs’ frequencies are not high enough to warrant being referred to as network motifs. Furthermore, the results show that Subgraph 3 is a motif in 22 of the networks, Subgraph 4 is a motif in 15 networks, Subgraph 5 a motif in 11 networks, and Subgraph 6 a motif in 7 networks. These results suggest which specific subgraph structures are more likely to exist in an attack graph in any given network. For example, it is unlikely for any subgraph to exist in an attack graph in a Barabási network, while there is an increased likelihood of Subgraphs 3, 4, and 5 existing in attack graphs in a Top-down network. This is because the Barabási networks do not exhibit any motifs, while the Top-down networks exhibit Subgraphs 3, 4, and 5 as motifs.

6.2. Superfamilies

In this subsection, we describe the identified network superfamilies. Using their SRPs, all the networks are assigned to superfamilies based on their level of structural similarity. The description of the identified superfamilies can then be used to classify any other Internet-like networks.

A correlation map showing all the network SRPs is presented in Fig. 5 where the networks have been rearranged according to their similarity using a modified k-nearest neighbor algorithm. The map consists of a color-coded correlation matrix to show the level of similarity among all the networks. A light-colored square signifies a high correlation between two networks while a dark-colored square signifies a low correlation. Because the networks have been rearranged by similarity, it is easy to identify approximate groups of similar networks from a collection of light-colored squares. For example, the networks BA1, BA3, BA2, AS1, BA5 and BA4 could potentially be in the same group because of the high correlation among themselves, while the networks AH5, AH4, AH1, AH2, and AH3 could potentially be in another group. The figure shows what is potentially two groups, one of which contains the majority of the considered networks. The correlation map is used to give an initial approximate idea of how many groups there are.

K-means clustering [45] is then used to identify a more accurate number of clusters in the SRPs and assign each network to one cluster. Four different superfamilies are identified (Superfamily 1–4) using the criteria mentioned in Section 3.3. This is because Fig. 6 shows that selecting 4 clusters instead of 2 or 3 reduces the intra-cluster error from 5.4 to as low as 1.35. Additionally, the smallest cluster contains 5 of the 28 networks ensuring a fairly balanced cluster composition. Each cluster is referred to as a superfamily, and each network is assigned to one of the four superfamilies (cf. Fig. 7).

Superfamily 1 consists of networks that have average SRP values for Subgraph 1 and 2, but negative SRP for Subgraph 4 (cf. Fig. 7(a)). The networks identified as belonging to this superfamily include one autonomous system network and all five Barabási networks considered. From Table 3, we observe that all the networks in this superfamily do not exhibit any motifs.

Superfamily 2 contains networks that showed average SRP values for subgraphs 1, 2, and 6, slightly above average values for Subgraph 4, and high levels for Subgraph 3 and 5 (cf. Fig. 7(b)). This superfamily consists of all six router level networks and five of the autonomous system networks. From Table 3, we observe that all networks in this superfamily exhibit Subgraph 3 as a motif and nine of its members exhibit Subgraph 4 as a motif.
Superfamily 3 contains networks with slightly below average SRPi values for subgraphs 1 and 2, high values for subgraphs 3, 4, and 5, and an above average value for Subgraph 6 (cf. Fig. 7(c)). The networks in this superfamily include all six hierarchical type networks considered. Table 3 shows that all six networks belonging to this superfamily exhibit subgraphs 3, 4, and 5 as motifs, while two of its members also exhibit Subgraph 6 as a motif.

Superfamily 4 consists of networks that show low SRPi values for subgraphs 1 and 2, high values for subgraphs 3, 5, and 6, and a below average value for Subgraph 4 (cf. Fig. 7(d)). This superfamily is comprised of all five spatial networks considered. Table 3 shows that all five members of this superfamily exhibit subgraphs 3, 5 and 6 as motifs.

The results above can be used to identify the superfamily of any given test network. For example, if a test network exhibits average SRPi values for subgraphs 1, 2 and 6, high values for subgraphs 3 and 5, and slightly above average values for Subgraph 4, then it belongs in Superfamily 2.

6.3. Relation to IP traceback

In this subsection, we describe the relationship between superfamilies and the IP traceback performance of networks in those superfamilies. We find that the performance of the different schemes does in fact depend on the network on which they are implemented. We also find that the best performing scheme in one network is not nec-
essarily the best performing scheme in another network. However, the results also show that networks within a superfamily exhibit notably similar performances which indicates a link between their underlying structure and their IP traceback performance. The convergence time is used as a measure of scheme performance.

The convergence times for similar DDoS attacks in all 28 networks is shown in Fig. 8. The plot shows the average convergence times for the considered PPM-based schemes measured in network packets, and their 95% confidence intervals for each network. The networks have been arranged by superfamily for easier comparison within and across their respective superfamilies. The standard deviations in average scheme convergence times within the superfamilies are shown in Table 4 and are used as a measure of intra and inter superfamily comparison.

The first observation from Fig. 8 is that PPM-based scheme convergence times vary greatly from one network to another. The plot shows that despite identical network size and DDoS attack scale, average convergence time varies between 4522 and 12,230 packets depending on both the selected scheme and the network in which the scheme is implemented. In Table 4, the results show that the most variation is exhibited by PPM which has a standard deviation of 1713.13 packets across all the considered networks, compared to 485.46 and 275.20 packets for TMS and PBS respectively. Additionally, Fig. 8 shows that the best performing scheme in one network is often different from the best performing scheme in another network. For example, when the schemes are compared in network BA1, all three schemes’ convergence times are within 1000 packets of each other with PBS < PPM < TMS. However, network AS4 exhibits TMS < PBS < PPM with a range of 1700 packets, while network AH3 exhibits PBS < TMS < PPM with a range of 6800 packets. These results show that the kind of network on which the schemes are implemented plays a big role in its IP traceback performance. Additionally, these results show that the best scheme in one network is not necessarily the best scheme in another network. These results mean that evaluating the performance of PPM-based schemes on a single network does not provide an accurate representation of that protocol’s performance since its performance in one network can be completely different from its performance in another network. This result raises questions about general network protocol performance, particularly if the protocols are only evaluated in a small number of networks.

The second observation is a pattern when the networks are arranged according to their superfamilies. Fig. 8 shows that the convergence time of any given scheme in a network is comparable to the convergence time of that same scheme in another network if both networks are in the same superfamily. Additionally, the convergence times of the considered schemes in one superfamily are different from the convergence times of that scheme in another superfamily. For example, TMS and PBS exhibit similar and stable convergence times in networks belonging to Superfamily 3, while PPM exhibits higher convergence time (1.4 times as high as TMS and PBS). In Superfamily 4, TMS and PBS also exhibit similar convergence times while PPM exhibits even higher convergence times (up to 2.2 times as high as TMS and PBS). However, TMS and PBS exhibit dissimilar convergence times in networks belonging to Superfamily 2, and PPM is not as high as it is in Superfamilies 3 and 4 (1.2 times as high as TMS or PBS). This reduced variation within superfami-

Fig. 8. Average number of packets required for traceback in different topologies for PPM, TMS, and PBS with confidence intervals of 95%. The topologies are arranged according to their superfamilies. Despite the topologies being the same size, they all exhibit different scheme convergence times which indicates that the scheme performance is largely dependent on the graph structure. However, the networks exhibit comparable scheme convergence times within each superfamily which indicates a relationship between a network’s scheme performance and the identity of its superfamily.
lies is confirmed in Table 4 which shows that the initial standard deviation across all networks drops from as high as 1713.31 packets to as low as 151.35 when the schemes are compared within the superfamilies. These results show that there is a link between the superfamily of a network and its performance in IP traceback. These results also suggest that the performance of any scheme in a network belonging to given superfamily can be used to predict the performance of that scheme in any other network belonging to the same superfamily. Consequently, this result means that IP traceback performance (and network protocol performance in general) can be evaluated in representative networks specifically chosen from different superfamilies instead of evaluating the protocols in all possible network topologies. The superfamily technique therefore allows for a more efficient network evaluation process since one does not have to check a large set of all possible networks but rather a smaller set consisting of networks selected to represent all superfamilies.

The third observation is the link between the motifs in any network and the discrepancy between different schemes’ convergence times in that network. To explain this link, we shall refer to both Fig. 8 and Table 3.

Fig. 8 shows that PPM generally has the highest convergence times across all networks, while PBS and TMS generally have lower convergence times. The extent of the discrepancy between PPM on one hand, and both TMS and PBS on the other hand, varies from one superfamily to another. This discrepancy is higher in Superfamilies 3 and 4 than it is in Superfamilies 1 and 2. For example, the difference in convergence times between PPM and PBS for Network AS5 in Superfamily 2 is only 600 packets which represents a percentage difference\(^3\) of 14.2%. This difference in convergence times is similar to other networks in Superfamily 2 and most in Superfamily 1. In contrast, the difference in average convergence times between PPM and PBS for Network AH3 is 6800 packets (percentage difference of 79.3%). Similar results are observed for other networks in Superfamily 4 and networks in Superfamily 3.

Another interesting observation deals with the average shortest path length values for the networks. A comparison between Table 3 and Fig. 8 shows that the networks with high average shortest path length values (in Table 3) are the same networks with high PPM convergence time values (in Fig. 8). For example, networks AH1, AH2, AH3, and AS1 exhibit relatively high average shortest path values, and the same networks exhibit relatively high convergence times for PPM compared to the PPM convergence times for other networks. However, the results in Fig. 8 also indicate that the convergence times for TMS and PBS are comparable to the TMS and PBS convergence times for other networks. Recall that according to the original analytical models (in Section 4.1), the average path length should have a similar effect on all schemes, i.e. an increase in the path length translates to an increase in convergence times for all considered schemes. However, these results seem to indicate that another factor exists in TMS and PBS that nullifies the effect of large average path values in specific networks. This structural factor leads to the discrepancy observed between PPM, and TMS and PBS. Additionally, this factor seems to be emphasized in selected networks. The adjustments to the analytical model presented in Sections 4.2 and 4.3 are one possible reason for this discrepancy. This is because the insights gained from those model adjustments are consistent with the results observed in Fig. 8 as explained below.

The adjusted model (cf. Section 4.2), shows that alternative routes in attack paths, such as those offered by Subgraphs 3, 4, 5 and 6, have the effect of increasing convergence times for all considered schemes. However, the merging of different attack streams cancels out this effect particularly in TMS and PBS where this merging also reduces the convergence times (cf. Section 4.3). This leads to a discrepancy in results between PPM on one hand, and TMS and PBS on the other hand. This discrepancy in results between the schemes is more pronounced with more alternative paths, and therefore Subgraph 6 has a larger effect than Subgraph 3 since Subgraph 6 has four alternative routes compared to Subgraph 3 which only has two alternative routes. We therefore expect to see increased discrepancy between PPM convergence times and both TMS and PBS convergence times for networks with an abundance of Subgraphs 3, 4, 5 and 6, e.g. TD5, TD6, AH1–5. We also expect this discrepancy to be more pronounced with Subgraphs 5 and 6 than it is with Subgraphs 3 and 4.

Our results agree with the presented model. In Fig. 8, convergence times for TMS and PBS in Superfamilies 3 and 4 are generally similar to Superfamilies 1 and 2 despite the increase in average shortest path length, while the convergence times for PPM are higher. Additionally, Superfamily 4 shows the most difference between PPM,

<table>
<thead>
<tr>
<th>Category</th>
<th>PPM</th>
<th>TMS</th>
<th>PBS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Superfamily 1</td>
<td>481.68 (7.44%)</td>
<td>578.40 (10.78%)</td>
<td>278.89 (5.06%)</td>
</tr>
<tr>
<td>Superfamily 2</td>
<td>234.70 (3.67%)</td>
<td>437.83 (8.20%)</td>
<td>182.40 (3.36%)</td>
</tr>
<tr>
<td>Superfamily 3</td>
<td>464.77 (5.98%)</td>
<td>282.67 (5.30%)</td>
<td>151.35 (2.83%)</td>
</tr>
<tr>
<td>Superfamily 4</td>
<td>1524.14 (14.42%)</td>
<td>406.71 (6.92%)</td>
<td>334.67 (5.80%)</td>
</tr>
<tr>
<td>All networks</td>
<td>1713.31 (23.00%)</td>
<td>485.46 (8.92%)</td>
<td>275.20 (5.01%)</td>
</tr>
</tbody>
</table>

\(^3\) The percentage difference, which is used to quantify the difference between two equally important values, is calculated from

\[
\frac{P_{\text{worst}} - P_{\text{best}}}{P_{\text{best}}} \times 100\% \quad \text{where } P_{\text{best}} \text{ and } P_{\text{worst}} \text{ are the convergence times of the best and worst schemes in a given network respectively.} \]
and TMS and PBS, because it has an abundance of subgraph 6, as shown in Fig. 7(d), and identified in Table 3. This result validates the model presented earlier and shows how the motifs affect scheme convergence time uniquely.

The results in this subsection indicate three things. Firstly, the scheme performance between any two networks is often very different and therefore schemes should be evaluated on multiple networks in order to provide some confidence about their performance. Secondly, the scheme performance within the identified superfamilies is generally similar. This means that instead of evaluating schemes on a large number of networks, a smaller number of specific networks can be selected that would adequately represent the larger set of networks. The scheme performance in one network can then be used to make projections about that scheme’s performance in any other network from the same superfamily. Thirdly, the results show that specific motifs that are exhibited in Superfamilies 3 and 4 create the discrepancy between the performance of different schemes which is exhibited in Superfamilies 3 and 4 but not exhibited in Superfamilies 1 and 2.

6.4. Internet networks

In this section, we present complementary results derived from three networks from the Caida project [21,22] (referred to as Caida1, Caida2, and Caida3) as well as two networks from the Rocketfuel project [23,24] (referred to as Rocket1 and Rocket2).

The SRPs for all five networks are shown in Fig. 9. The figure reveals that all five networks have very similar SRPs and yet these SRPs are distinct from the SRPs in Fig. 7. The networks generally exhibit a lack of Subgraph 4, an abundance of subgraphs 3, 5 and 6, as well as average amounts of subgraphs 1 and 2. This is an interesting result because it shows that the networks, which are derived at different times in different locations under different projects, are in fact similar when their subgraph structure is compared using the SRP technique.

![Fig. 9. The subgraph ratio profiles (SRPs) of the five complementary networks. These networks are similar to each other and yet different from the SRPs of the four original superfamilies.](image)

The convergence times for the attacks carried out in these networks are presented in Table 5 alongside their 95% confidence intervals after 100 simulations measured in packets.

<table>
<thead>
<tr>
<th>Network</th>
<th>PPM</th>
<th>TMS</th>
<th>PBS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Caida1</td>
<td>6503.9 ± 357.5</td>
<td>5090.3 ± 266.6</td>
<td>4858.5 ± 281.6</td>
</tr>
<tr>
<td>Caida2</td>
<td>7126.4 ± 392.97</td>
<td>5932.8 ± 494.89</td>
<td>5689.0 ± 232.08</td>
</tr>
<tr>
<td>Caida3</td>
<td>6631.6 ± 391.62</td>
<td>5551.5 ± 347.11</td>
<td>5506.6 ± 231.84</td>
</tr>
<tr>
<td>Rocket1</td>
<td>7292.5 ± 417.59</td>
<td>5760.8 ± 306.4</td>
<td>5265.9 ± 244.74</td>
</tr>
<tr>
<td>Rocket2</td>
<td>7914.6 ± 888.52</td>
<td>5443.0 ± 319.65</td>
<td>5400.8 ± 530.26</td>
</tr>
</tbody>
</table>

The convergence times for all five networks. As expected, TMS and PBS convergence times are similar to each other, and yet distinct from PPM convergence times for all three networks. It is interesting to point out that even though these network are between 3 and 10 times the size of the other 28, their TMS and PBS convergence times are comparable to the rest. Additionally, we find that their PPM convergence times are comparable to the convergence times of the networks in Superfamily 2. One would expect that the convergence times for all schemes would drastically increase with an increase in the network size. The link between network size and protocol performance presents an interesting direction for future research.

7. Conclusion

In this paper, we compare a set of PPM-based schemes. We evaluate the schemes PPM, TMS, and PBS on a large number of topologies selected to encompass the predominant mathematical network topology models of the Internet. Network motifs and subgraph ratio profiles are applied to capture the subtle differences and similarities in structure between these topologies and to assign them to superfamilies.

Our results show a link between the analyzed network topologies, their motifs, subgraph distributions, superfamilies, and the resultant performances of the PPM-based schemes. Moreover, an analytical model is presented in this paper. This model shows how this link affects the schemes uniquely, contributing to the discrepancies in their convergence times among the networks.

Possible future work includes designing motif-aware protocols and schemes, which would exploit the influence of network motifs for improved scheme performance. This work encourages multiple network evaluation and comparison of network protocols as opposed to the common practice of analyzing a protocol in a single type of network.

References

Vir V. Phoha received the Ph.D. degree in computer science from Texas Tech University, Lubbock, in 1992. He is currently a W. W. Chew Endowed Professor of computer science and director of the Center for Secure Cyber-space at Louisiana Tech University, Ruston. His research interest includes nonlinear optimization, machine learning, anomaly detection, spatial-temporal pattern detection and event recognition, and knowledge discovery and analysis. Professor Phoha is an ACM Distinguished Scientist and an ACM Distinguished Speaker and is a Fellow of SDPS.