Iteratively Detecting Collaborative Scanner Fingerprints: An Iterative Approach to Identifying Fingerprints using Stratified Sampling

Jelt Jongsma [1]

[1]EEMCS, Delft University of Technology, The Netherlands

Background

- **Internet-wide scanners** probe the entire internet by sending packets to all IP addresses.
- **Collaborative scanners** distribute their scans over multiple hosts to remain undetected.
- **Network telescopes** are networks without any services that receive these probing packets.
- Scanners embed a **pattern** into their packets so they do not need to keep state of sent packets.
- **TCP functions** extract combinations of header fields from packets and return a value.
- **Effective signs** are pairs of TCP functions and values that appear in packets most often.
- One or more of these effective signs together form a **fingerprint** that can be used to identify collaborative scanners.

Research question

How does an iterative approach to generating fingerprints for collaborative scanners, using stratified sampling, affect accuracy when compared to existing algorithms?

Related work

- **Griffioen & Doerr** [2] used SLPA to cluster scanning groups and iteratively applied a sequence on XORs on these groups to identify header field patterns.
- **Tanaka et al.** [4] considered a genetic algorithm to generate flexible TCP functions, identified effective signs using said functions, and returned fingerprints.
  - They found fingerprints for 18.8% of all packets, and 50.6% of the scanners.
  - Which means over three quarters of packets are not fingerprinted, and nearly half of the scanners go undetected.

Methodology

1. **Generate functions**. Based on the same generation scheme as in [4], this method generates n functions by applying *binary operations* and *feature extractions* to an increasing set of initial *TCP functions*.
2. **Identify fingerprints**. Iteratively sample packets from the dataset and compute fingerprints. Packets matching found fingerprints are removed from the dataset.
   1. **Sample packets**. Sample n packets from the dataset using stratified sampling, where n is determined using Cochran's formula [1], and strata are defined by hour-long intervals.
   2. **Find effective signs**. Compute signs for sampled data, and extract effective signs using appearance ratios and a sign threshold.
   3. **Dynamically adjust sign threshold**. Since the size the set with unidentified packets is continually decreasing, we dynamically adjust the sign threshold, such that the algorithm always finds at most 15 signs.
   4. **Consolidate effective signs**. Signs with >90% overlap in matching packets are combined using ANDs.
3. **Validate results**. This method should be able to identify well-known fingerprints, such as ZMap (IPId = 54321) and Masscan (IPId $\oplus$ f_{L2B}(DstIP) $\oplus$ f_{L2B}(Seq) = 0). The final results are validated using these fingerprints.

Experiment

- The algorithm identified **3** fingerprints.
- Found fingerprints for sets of packets that make up **less than 0.5%** of all packets, and **less than 0.0001%** of sources, see table 1.
- Analysis of groups indicated they were **not** distributed scanners, which means they had the same pattern by coincidence.
- Gr0 and Gr1 target too many ports.
- Gr2 has no cohesion in targeted ports or its sources.
  - Targeted ports are e-mail protocols (110, 143, 587), http (80, 8080), and six others (21, 22, 179, 433, 5060, 6667).
  - Sources come from the USA (0% confidence of abuse[a]), Romania (61%), and China (14%, 100%, 100%)
  - Interesting to note; ~492 K packets came from the Romanian source, while the others sent only one.
- There is no Masscan in the result, indicating the algorithm was not able to generate the right TCP functions to detect collaborative scanners.

Identified groups

Name	Packets (%)	#sources (%)	#dest. ports	Fingerprint
Gr0	883 K (0.35%)	22880 (15.75%)	2819	f_{L2B}(Seq) $\oplus$ Seq $\oplus$ DstIP = 33716 $\wedge$ f_{L1B}(f_{L2B}(SrcPort $\oplus$ Seq)) $\oplus$ DstIP $\oplus$ Seq = 131 $\wedge$ DstIp $\oplus$ Seq $\oplus$ f_{L2B}(DstIP) = 33716
Gr1	2915 K (1.18%)	27552 (18.96%)	8731	f_{L2B}(Seq) $\oplus$ Seq $\oplus$ DstIP = 33441 $\wedge$ DstIp $\oplus$ Seq $\oplus$ f_{L2B}(DstIP) = 33441 $\wedge$ f_{L1B}(Seq) $\oplus$ DstIP $\oplus$ Seq = 130
Gr2	492 K (0.20%)	5 (3.44e−5%)	11	IPId $\oplus$ f_{R1B}(Seq) $\oplus$ Seq = 61016 $\wedge$ f_{L1B}(IPId) $\oplus$ f_{R1B}(Seq) $\oplus$ Seq = 238

Fingerprints identified by algorithm. K denotes 10^3

Responsible research

- Internet-wide scanners with **malicious intentions** may be detected using this technique, leading to a safer internet.
- The seed used for random sampling is released, the code from this project is on GitHub[b], and the methodology and experimental settings are thoroughly explained, making the experiments in this paper highly reproducible.

[a]According to AbuseIPDB, *https://www.abuseipdb.com/*

[b]*https://github.com/jeltjongsma/detecting-collaborative-scanners*

Discussion

- The algorithm was unable to detect Masscan or other fingerprints found in previous studies ([2], [4], [3]), but did find sets of packets that coincidentally shared the same fingerprint that made up only a small fraction of the complete dataset (0.20%).
- The TCP function generator generated meaningless functions, such as DstIP $\oplus$ DstIP, or functions that coincidentally hold for unrelated packets, such as the functions used in the fingerprints in table 1.
- The first version of this algorithm computed effective signs over a sample, extracted their functions, and then computed effective signs over the entire dataset with just those functions. Implementing this approach in the same iterative way as the current algorithm, might increase the accuracy.
- The last study on fingerprint generators performed their experiments on data from September 2021 [3], while this study used data from February 2024. Since then internet-wide scanners might have stopped embedding patterns in their packets.

Conclusion

- This study considered the approach proposed by Tanaka et al. [4], and adapted it into an iterative approach to identify fingerprints on network telescope data.
- Preliminary testing showed the algorithm is able to consistently detect ZMap's fingerprint.
- An experiment on network telescope data showed the algorithm is able to generate fingerprints for sets of packets that make up less than 0.5% of packets, and less than 0.0001% of sources.
- Analysis of identified groups indicated they are not distributed scanners, but instead coincidentally shared fingerprints.
- Future work could focus on methods to avoid detecting fingerprints for unrelated groups, such as more effective function generation.
- It would also be interesting to see what fingerprints might be detected if TCP functions can be generated with a larger set of binary operations and feature extractions.

References

[1] William G. Cochran.
Sampling Techniques, 3rd Edition.
John Wiley, 1977.

[2] Harm Griffioen and Christian Doerr.
Discovering collaboration: Unveiling slow, distributed scanners based on common header field patterns.
In *NOMS 2020 - 2020 IEEE/IFIP Network Operations and Management Symposium*, 4 2020.

[3] Akira Tanaka, Chansu Han, and Takeshi Takahashi.
Detecting coordinated internet-wide scanning by TCP/IP header fingerprint.
IEEE Access, 11:23227–23244, 2023.

[4] Akira Tanaka, Chansu Han, Takeshi Takahashi, and Katsuki Fujisawa.
Internet-wide scanner fingerprint identifier based on TCP/IP header.
In *Sixth International Conference on Fog and Mobile Edge Computing, FMEC 2021, Gandia, Spain, December 6-9, 2021*, pages 1–6. IEEE, 2021.

[5] Jierui Xie, Boleslaw K. Szymanski, and Xiaoming Liu.
Slpa: Uncovering overlapping communities in social networks via a speaker-listener interaction dynamic process.
In *2011 IEEE 11th International Conference on Data Mining Workshops*, pages 344–349, 2011.

Supervisors: Harm Griffioen[1] & Georgios Smaragdakis[1]

j.jongsma-1@student.tudelft.nl