Poster: Blacklists Assemble - Aggregating Blacklists for Accuracy

Sivaramakrishnan Ramanthan[1], Jelena Mirkovic[1], and Minlan Yu[2]

[1]University of Southern California/Information Sciences Institute
[2]Harvard University

***Abstract*—Blacklists contain identities of known offenders and often serve as the first line of defense to preventively filter unwanted traffic. Yet, any single blacklist may only be effective for a given type of attack and only over certain portions of address space. Further, each blacklist is compiled and updated using proprietary methods, and thus may have stale information, leading to false positives or false negatives. Finally, blacklists are reactive since they list addresses of known attackers. We propose BLAG, a sophisticated approach to select and aggregate only the accurate pieces of information from multiple blacklists. BLAG estimates the accuracy of each listing of addresses in every blacklist and uses recommender systems to select most reputable and accurate pieces of information to aggregate into its master blacklist. Finally, BLAG expands some IP addresses into prefixes to further increase the attackers captured.**

I. Introduction

Compromised devices are constantly being drafted into botnets and misused for attacks, such as sending spam and phishing emails, scanning for vulnerabilities, participating in denial-of-service attacks, and spreading malware. *Blacklists*, which contain identities of prior known offenders, can be helpful as the first layer of defense. Assuming that prior offenders are likely to re-offend, filtering traffic from blacklisted sources can prevent zero-day attacks and reduce the load on more resource-intensive second-layer defenses.

Blacklists are created to monitor some regions of the Internet for specific malicious activities. This limited observation has two deficiencies. First, a blacklist may miss many attacks because it does not observe the given attack type. On the other hand, compromised hosts are traded on the black market and used for many malicious activities [2], so a host that sent spam today could engage in DDoS or spread ransomware tomorrow. Thus, it makes sense to aggregate multiple blacklists to achieve better detection accuracy. Second, since blacklists are compiled and updated by different maintainers using proprietary methods, blacklists may have false positives. Thus, each blacklist will have portions of accurate information, which we would want to include in aggregation, and portions of inaccurate information, which we would want to exclude. Third, blacklists are reactive by listing addresses of known offenders. If we could identify networks which are likely to host attackers, we can proactively blacklist all addresses belonging to that network.

In this paper, we propose BLAG, a sophisticated blacklist aggregation approach, which addresses these problems. BLAG assigns a score for each address listed in the blacklist and uses a recommendation system to select only accurate pieces of blacklists for aggregation. Finally, BLAG evaluates each region of Internet address space and selectively includes the entire region in its master blacklist. BLAG performs better than the naive aggregation of all blacklists and has a much lower number of false positives when compared to naive aggregation of all blacklists.

II. Datasets

We have analyzed 184 publicly available blacklists, collected regularly over a one-year period. Our blacklist dataset has been captured continuously for 13 months starting from January 2016. Each blacklist may be updated at a different frequency by its provider, ranging from every 15 minutes to 7 days. We have collected around 176 M blacklisted addresses. Our blacklist dataset is representative of different attack vectors such as spam, malware, DDoS attacks, ransomware, etc. We have further collected several datasets containing known-legitimate and known-malicious traffic sources. We use these datasets to evaluate the accuracy of current blacklists and of BLAG. Our ground truth dataset comes from three different sources capturing different types of attackers.

Emails. Our malicious source comes from *Mailinator* [1], a service, which allows users to redirect unwanted emails to a public inbox. We filter emails from these public inboxes during June 2016, using Spam Assassin [2] to obtain around 2.3 M spam emails, sent by around 39 K addresses. These addresses form our malicious dataset. Our second source of legitimate addresses comes from our human user study approved by our IRB. We recruited 37 volunteers, who allowed automated access to their Gmail inbox, during June 2016. We scanned each participant's Gmail account using a plugin to extract around 178 K email records, sent by around 5 K addresses. We collected no identifying information about our study participants.

Scanning. Our malicious source comes from Netlab's scans [3] consisting of 232 K hosts infected with Mirai malware. Our legitimate source comes from legitimate

[1]https://www.mailinator.com
[2]https://spamassassin.apache.org/
[3]https://goo.gl/NYWMLq.

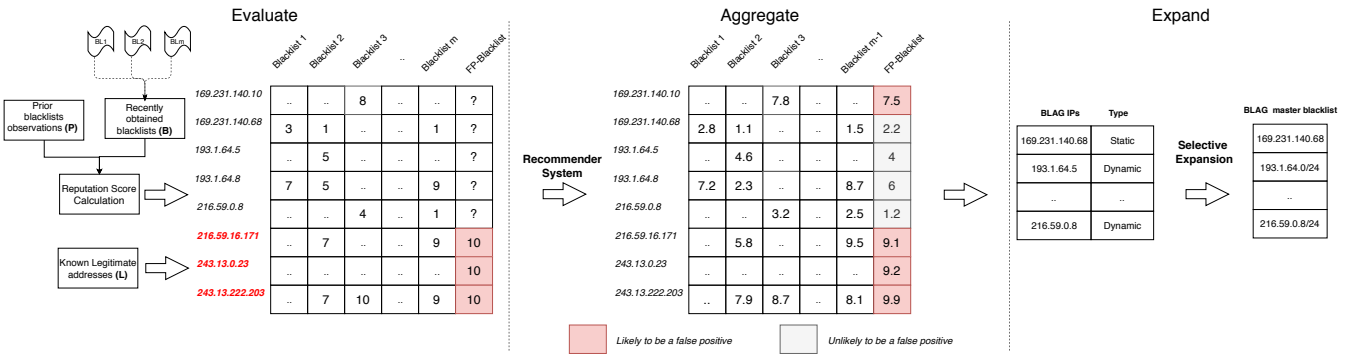


Figure 1: BLAG implementation consists of assigning score to addresses from different blacklists based on their age of listing. BLAG then introduces *FP-blacklist* consisting of known false positives (in red). Then, a recommender system generates scores for addresses that are not listed in *FP-blacklist*. Some addresses are pruned out as false positives based on a threshold and others are used for expansion. These addresses will be put on the master blacklist. Finally, we selectively expand addresses into /24 prefixes, if we project that this expansion will not increase the number of false positives.

requests sent to servers belonging to a university network. Both the sources were collected during September 2016.

DNS request. The malicious source comes from hosts which generated a DDoS attack on B-root DNS server and the legitimate source comes from hosts which generated legitimate DNS requests. There were about 5.5 M sources of attack and about 14 K legitimate sources collected during February 2017.

III. BLAG Design

BLAG assigns a score to every address listed in blacklists. If an address a is listed in the blacklist b, the score s is defined as

$$s_{a,b} = 10 * \frac{1}{2^{\frac{t-t_{out}}{30}}} \quad (1)$$

where t is the current time and t_{out} is the last listing time. The above scoring function assigns a higher score to recently listed address than older addresses. Addresses not present in blacklists are assigned a score of 0. We also create a blacklist *FP-blacklist* consisting only of known false positives. For creating such a blacklist, we take each of our three ground truth datasets, dividing it into seven days of training and the rest is used for testing. BLAG uses the legitimate part of the training dataset to create the *FP-blacklist*. Addresses present in *FP-blacklist* are allocated a score of 10. We use a recommender system to calculate the missing scores in *FP-blacklist*, i.e., to predict the likelihood of an address being listed in the *FP-blacklist*. Initially, BLAG places them into a *score matrix* (as seen in Figure 1) where blacklists (including *FP-blacklist*) are columns and addresses are rows with cells consisting of $s_{a,b}$.

A well-known example is the Netflix recommender system [1], which may recommend a new movie M to user U by relying on the U's past ratings of movies similar to M, and on ratings that users similar to U have given to movies similar to M. In our context, addresses are products and blacklists are users assigning the rating. Since *FP-blacklist* consists only of known false positives, any address which is recommended to appear in this blacklist is likely to be a false positive. Finally, after running the recommendation system, we filter out potential false positives by setting a predefined threshold and expand all addresses which do not have any addresses in its /24 prefix listed in the training dataset to further increase true positives.

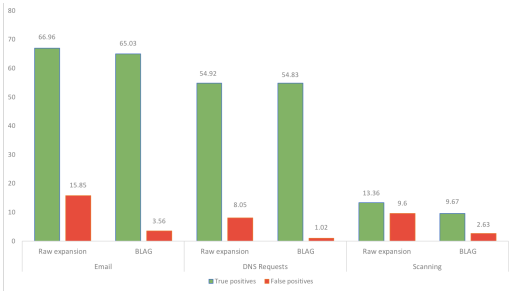


Figure 2: Evaluation of BLAG with raw expansion.

IV. Evaluation

Figure 2 compares BLAG with raw expansion which includes all addresses listed in blacklists and expanding addresses into it's /24 prefix if there are no other addresses in the same prefix listed in the training dataset. We observe that BLAG reduces the false positives from 15.8% to 3.5%, 8% to 1% and 9.6% to 2.6% for emails, DNS requests and scanning datasets respectively. This occurs with only a minimum drop in the number of true positives for both emails and scanning datasets. Our future work will investigate how to improve the number of true positives retained while keeping the number of false positives low.

References

[1] Yehuda Koren, Robert Bell, Chris Volinsky, et al. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37, 2009.

[2] Jing Zhang, Zakir Durumeric, Michael Bailey, Mingyan Liu, and Manish Karir. On the mismanagement and maliciousness of networks. In *NDSS*, 2014.

BLAG: Blacklist Aggregator

Sivaram Ramanathan[1], Jelena Mirkovic[1] and Minlan Yu[2]
[1]University of Southern California, [2]Harvard University

Problem Statement

Blacklists contain a list of previously known attackers and are commonly used by network operators as the first line of defense.
However, blacklists have the following drawbacks:

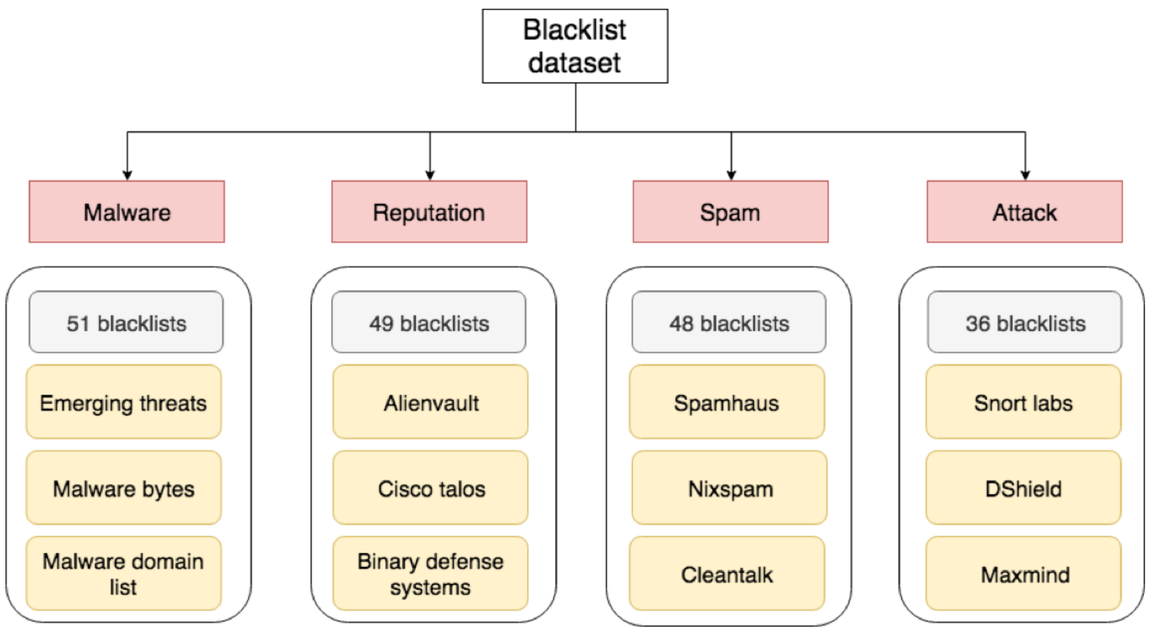
- **False positives:** Stale information or misclassification in the detection algorithm may yield false positives
- **Limited score:** Blacklists only target specific type of attackers
- **Reactive:** Blacklists only list IP addresses of known attackers

Our Solution

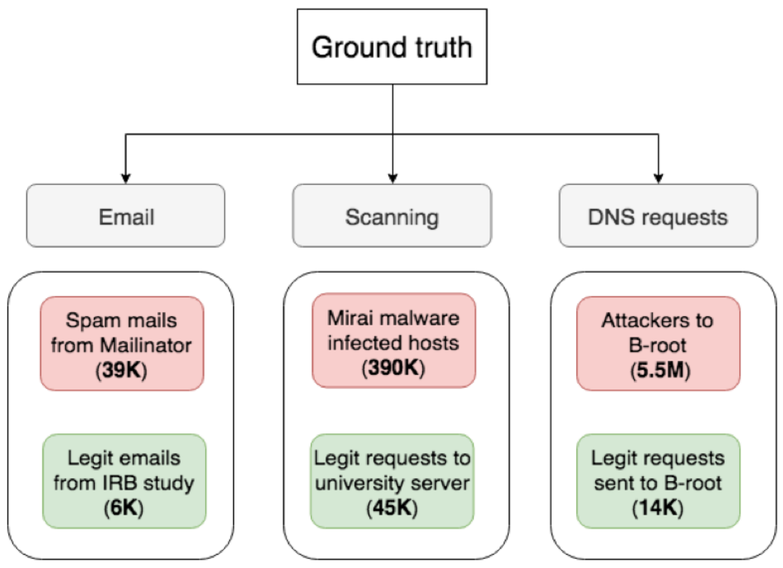
We propose BLAG, which aims to increase the number of attackers captured while keeping the false positives down.

- BLAG combines several blacklists of different attack types to increase the number of attackers captured
- It uses a recommendation system to prune out potential false positives from the combined blacklists
- BLAG selectively expands some of these addresses into prefixes if it projects that the expansion will not increase the false positives

Datasets

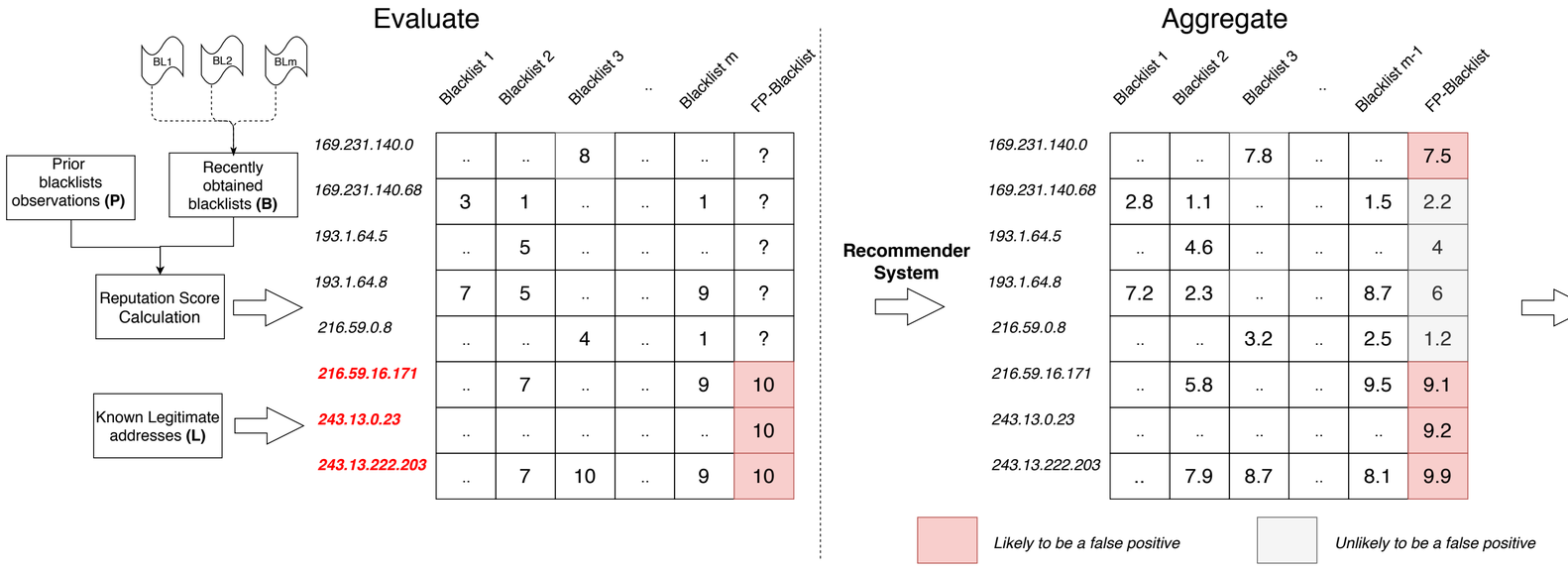


184 blacklists were monitored from Jan 2016 to Dec 2017 and are roughly categorized into four attack variants



Three categories of ground truth to estimate the effectiveness of the aggregation approach

BLAG

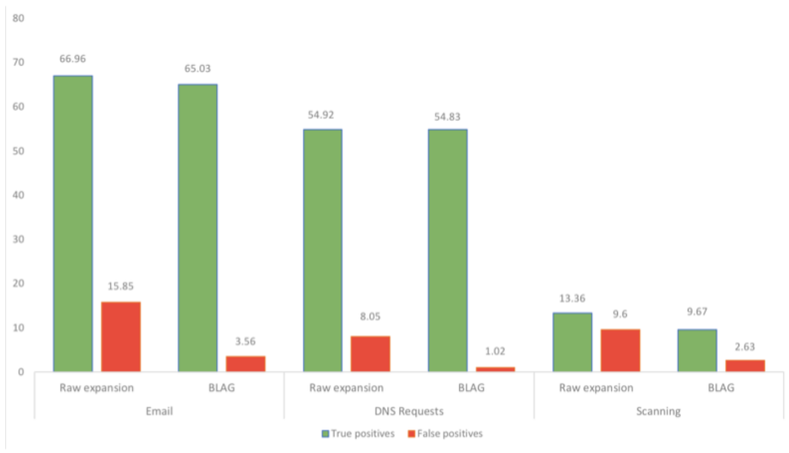


Allocate scores to addresses listed in blacklists based on their listing time. Addresses not listed in blacklists are left empty. Introduce a FP-blacklist which lists only known false positives.

Estimate unknown scores of addresses in the FP-blacklists. Determine the likelihood of an address to be a false positive based on the predicted score and prune them out.

Expand addresses into their corresponding /24 prefix if the prefix does not have known false positives.

Preliminary Evaluation



- Compared BLAG to raw expansion, which involves aggregating all blacklists, leaving out known false positives and expanding the remaining addresses into their /24 prefix
- Overall, BLAG reduces false positives in all three datasets
- For all datasets, true positives reduce 0.1—3.7%
- Future work involves looking into variable expansion of addresses into prefixes apart from /24

If interested, contact Sivaram satyaman@usc.edu

USC Viterbi School of Engineering

HARVARD UNIVERSITY