BLAG: Improving the Accuracy of Blacklists

Sivaram Ramanathan[1], Jelena Mirkovic[1] and Minlan Yu[2]

[1] University of Southern California/Information Sciences Institute

[2] Harvard University

USC University of Southern California

IP Blacklists

- IP Blacklists contain a list of known malicious IP addresses.
- IP Blacklists are commonly used to aid more sophisticated defenses such as spam filters, IDS, etc.

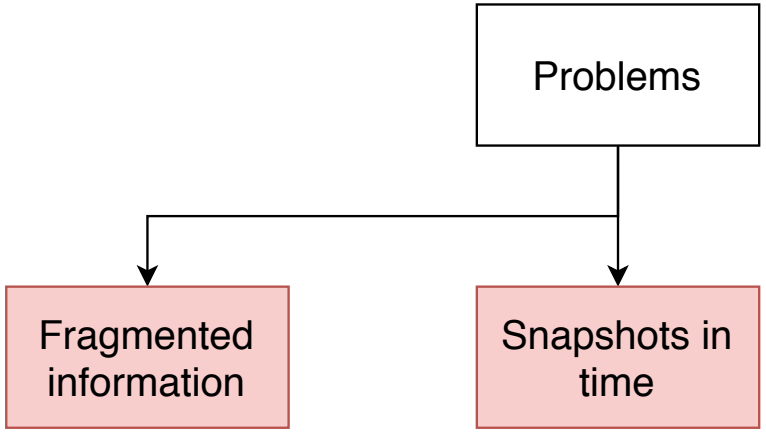
- IP blacklists can be used as an emergency response under a novel or large volumetric attack
  - Easy to implement as only IP addresses are checked and can be done at line rate.

1.	198.38.89.61	2.	175.230.213.33	3.	182.74.165.174	4.	178.137.90.85
5.	111.40.73.83	6.	61.132.233.195	7.	193.150.72.50	8.	221.4.205.30
9.	60.172.69.66	10.	61.163.36.24	11.	60.166.48.158	12.	117.214.17.72
13.	180.121.141.117	14.	114.232.216.5	15.	183.159.83.71	16.	121.239.86.33
17.	92.73.213.217	18.	162.248.74.123	19.	183.159.95.87	20.	14.207.215.126
21.	222.191.179.90	22.	217.110.92.194	23.	156.216.145.235	24.	81.17.22.206
25.	41.251.33.175	26.	114.223.61.210	27.	114.232.193.38	28.	114.231.141.136
29.	170.51.62.241	30.	49.67.83.155	31.	180.121.141.119	32.	39.40.30.104
33.	209.54.53.185	34.	167.114.84.153	35.	223.240.208.236	36.	183.150.34.181
37.	95.37.125.239	38.	171.14.238.42	39.	1.55.199.83	40.	222.191.177.40
41.	45.234.101.139	42.	117.85.56.142	43.	123.54.107.199	44.	45.119.81.235
45.	186.47.173.213	46.	49.67.67.141	47.	95.211.149.134	48.	113.128.132.9
49.	49.67.67.140	50.	119.180.198.174	51.	103.69.46.81	52.	128.199.35.34
53.	159.255.167.131	54.	181.215.89.206	55.	192.210.201.168	56.	128.199.44.20
57.	218.72.108.217	58.	113.120.60.120	59.	111.125.140.155	60.	60.50.145.121

Problems with IP Blacklists

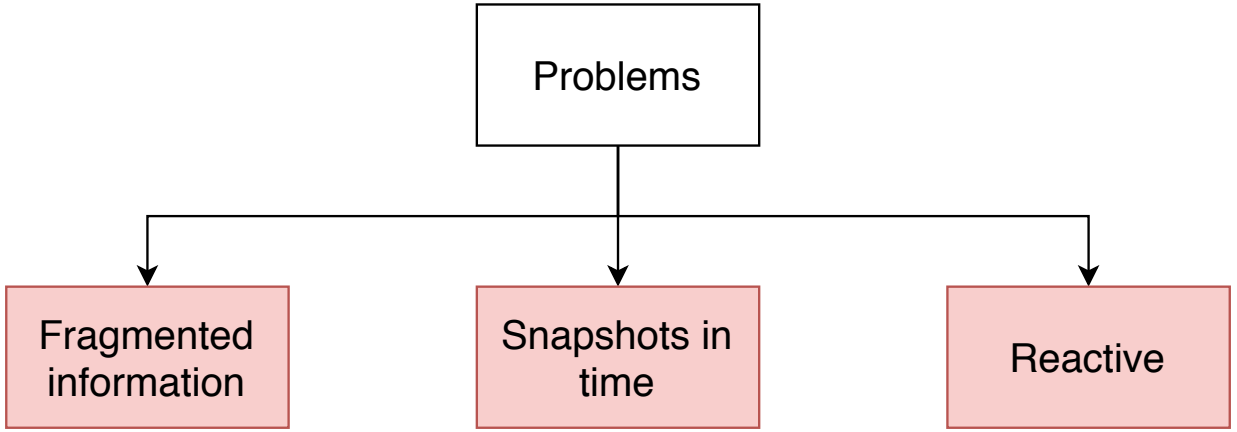
- Focus only on specific attack types with limited vantage points.

Problems with IP Blacklists



- Focus only on specific attack types with limited vantage points.
- Historical blacklist data can capture reoffending malicious addresses.

Problems with IP Blacklists



- Focus only on specific attack types with limited vantage points.
- Historical blacklist data can capture reoffending malicious addresses.
- Addresses are added only after a malicious event is observed.

Problems with IP Blacklists

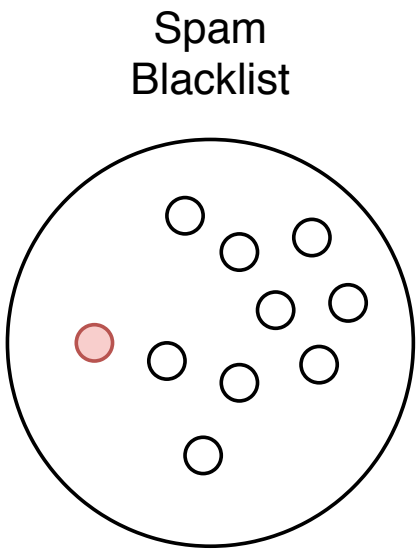
Problems

Can we aggregate blacklists in a smart way to address these problems?

- Focus only on specific attack types with limited vantage points
- Historical blacklist data can capture reoffending malicious addresses
- Addresses are added only after a malicious event is observed

Fragmented Information

- offenders in one given attack

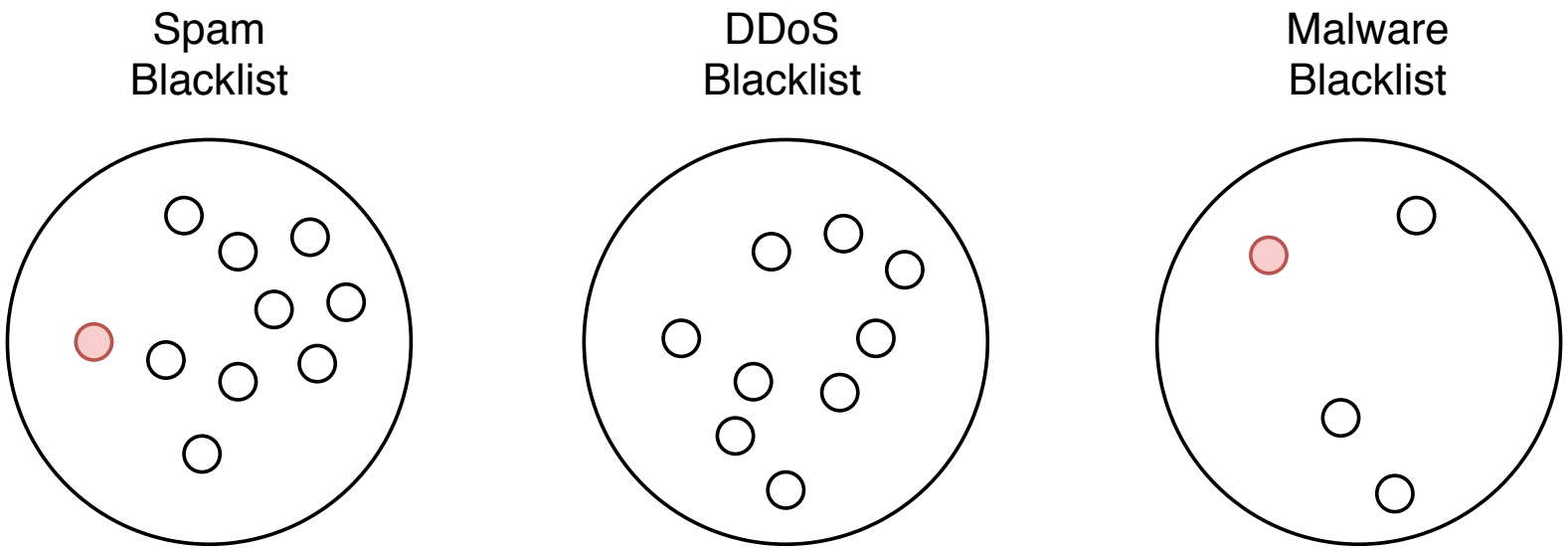


Blacklists miss many attacks[1,2] and may monitor only specific a type of attack.

[1] Kührer, Marc, Christian Rossow, and Thorsten Holz. "Paint it black: Evaluating the effectiveness of malware blacklists." International Workshop on Recent Advances in Intrusion Detection. Springer, Cham, 2014.
[2] Pitsillidis, Andreas, et al. "Taster's choice: a comparative analysis of spam feeds." *Proceedings of the 2012 Internet Measurement Conference*. ACM, 2012.

Fragmented Information

- offenders in one given attack

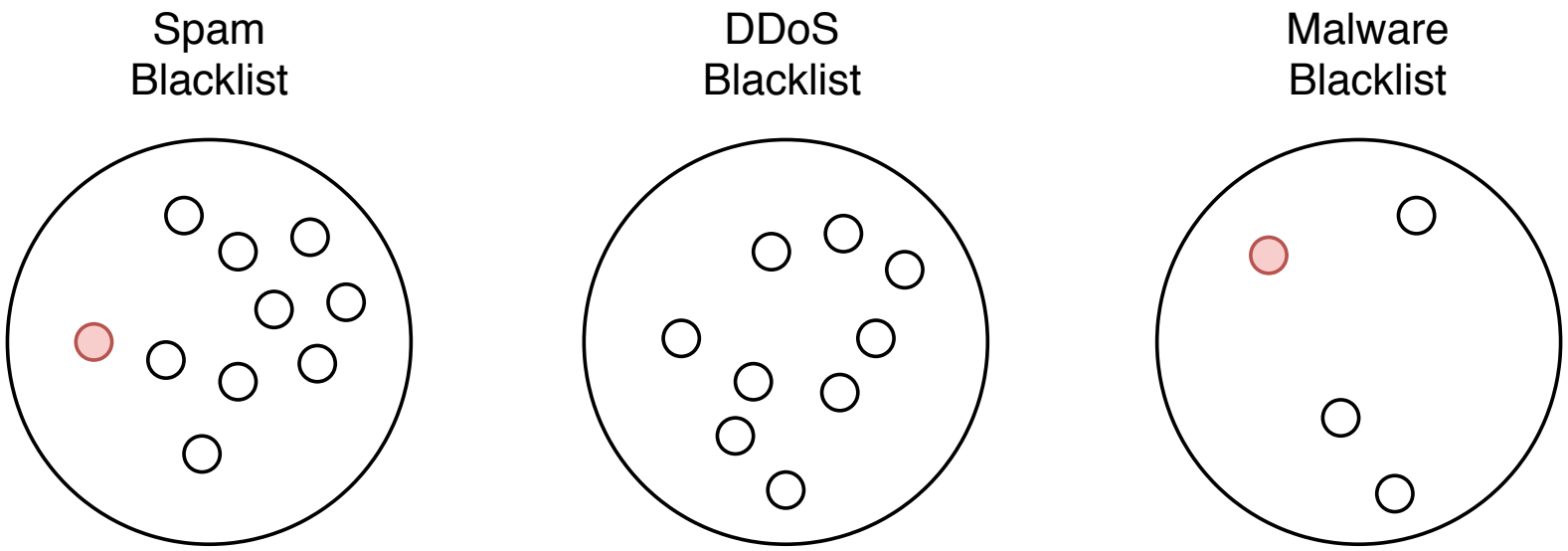


Blacklists miss many attacks[1,2] and may monitor only specific a type of attack.

[1] Kührer, Marc, Christian Rossow, and Thorsten Holz. "Paint it black: Evaluating the effectiveness of malware blacklists." International Workshop on Recent Advances in Intrusion Detection. Springer, Cham, 2014.
[2] Pitsillidis, Andreas, et al. "Taster's choice: a comparative analysis of spam feeds." *Proceedings of the 2012 Internet Measurement Conference*. ACM, 2012.

Fragmented Information

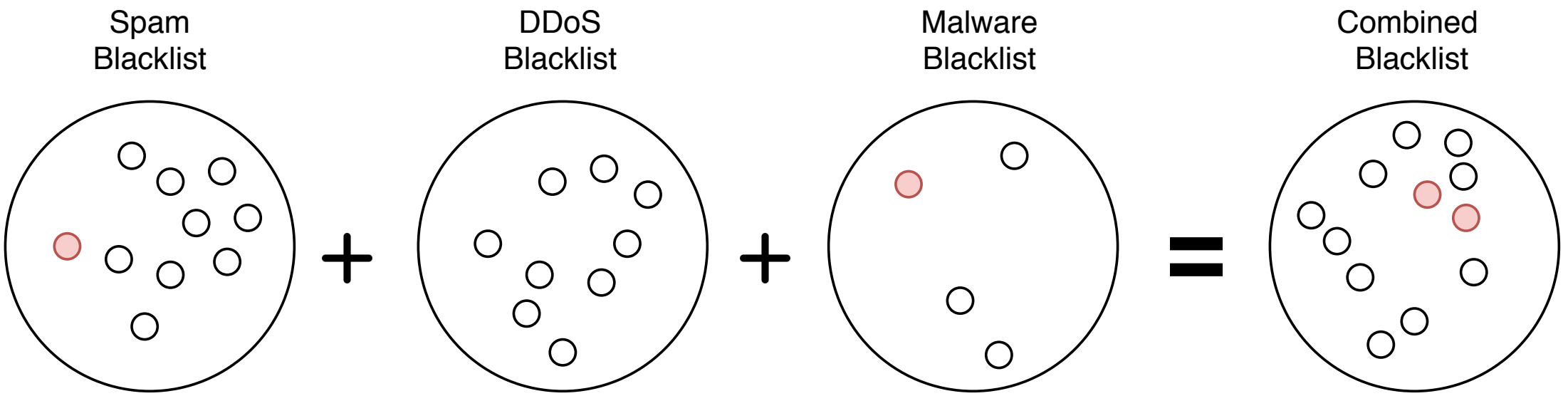
- offenders in one given attack



Compromised machines are constantly re-used for initiating different types of attacks over time.

Fragmented Information

- offenders in one given attack



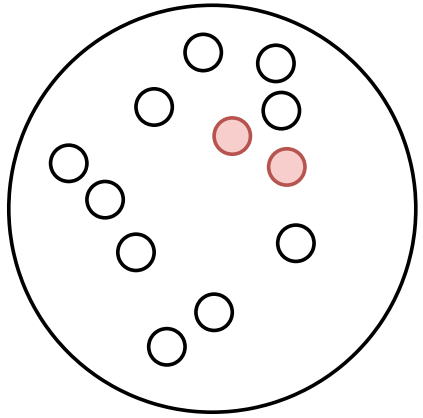
Compromised machines are constantly re-used for initiating different types of attacks over time.

A Possible solution: Combining different types of blacklists can improve attack coverage.

Snapshots in Time

- offenders in one given attack

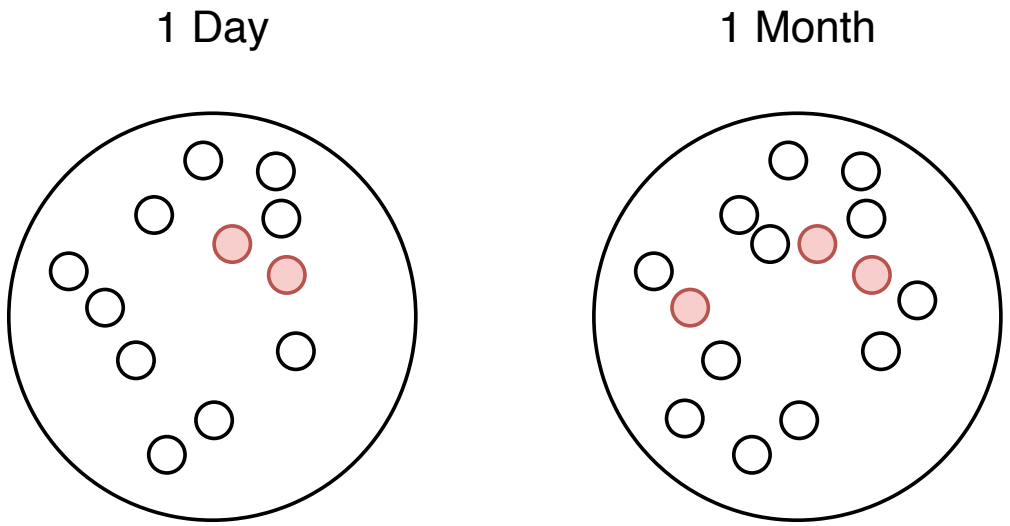
1 Day



Historical blacklist data (union of all offenders over time) can further be useful to improve offender detection.

Snapshots in Time

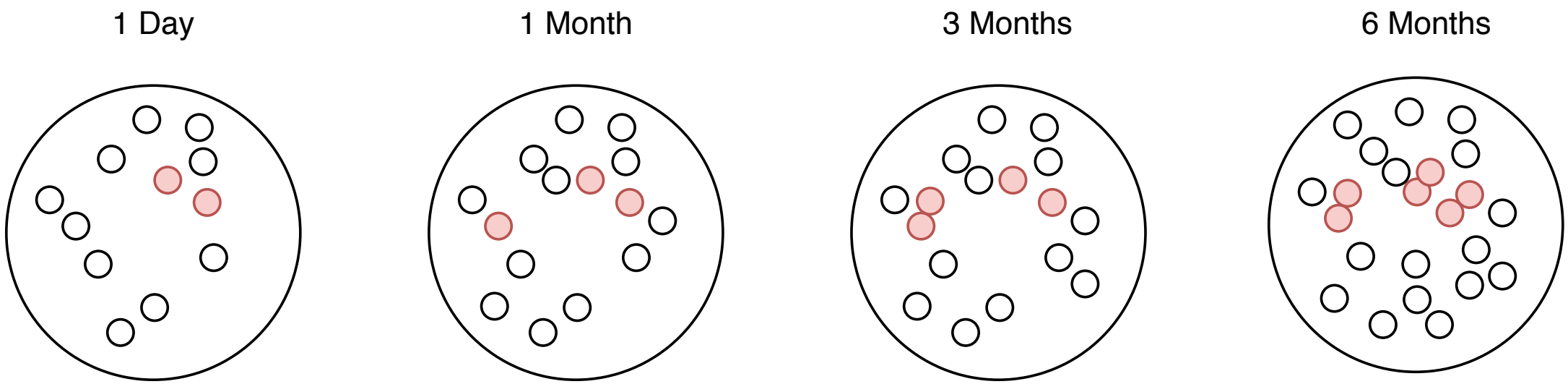
- offenders in one given attack



Historical blacklist data (union of all offenders over time) can further be useful to improve offender detection.

Snapshots in Time

- offenders in one given attack

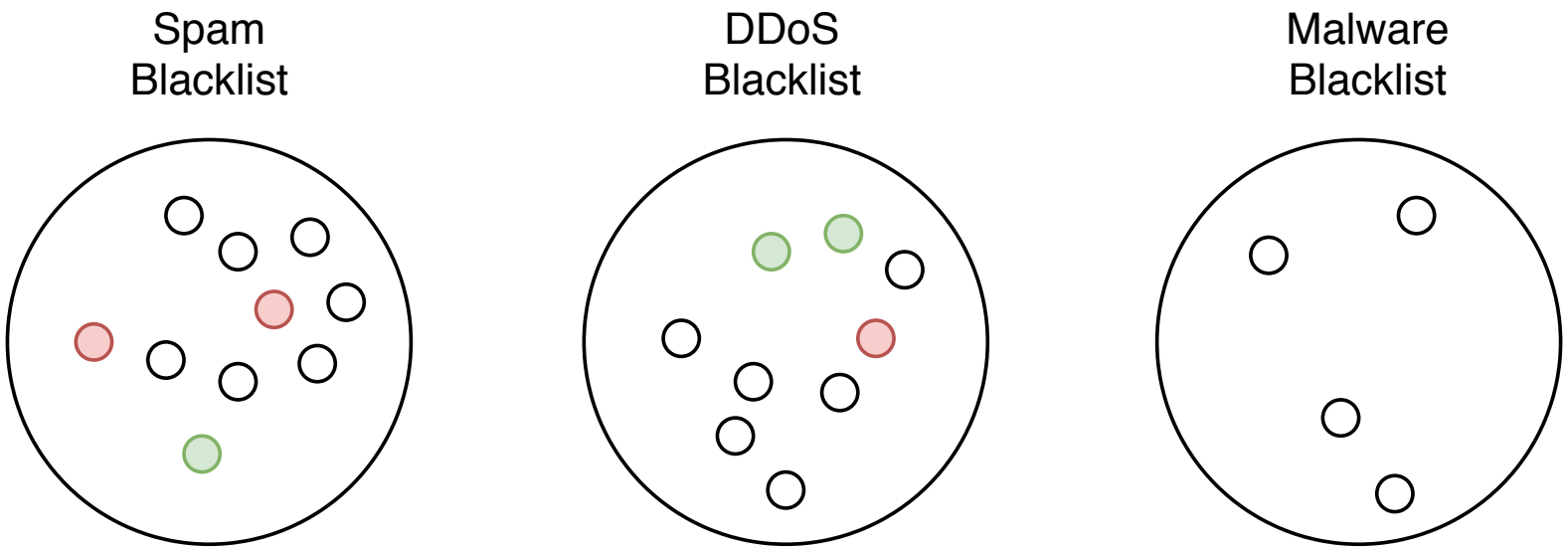


Historical blacklist data (union of all offenders over time) can further be useful to improve offender detection.

Careful Aggregation

- offenders in one given attack

- legitimate clients of a given network during the same attack



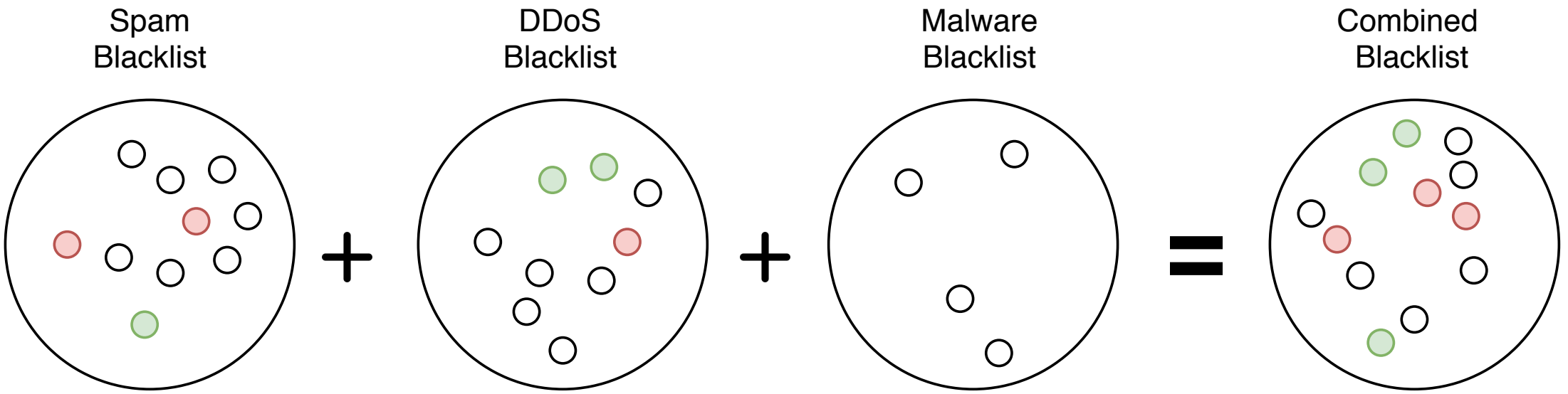
Blacklists accuracy varies spatially

- Blacklists are maintained by individuals or organizations that use proprietary algorithms to include or exclude an address.
- Blacklists could list some legitimate addresses

Careful Aggregation

- offenders in one given attack

- legitimate clients of a given network during the same attack

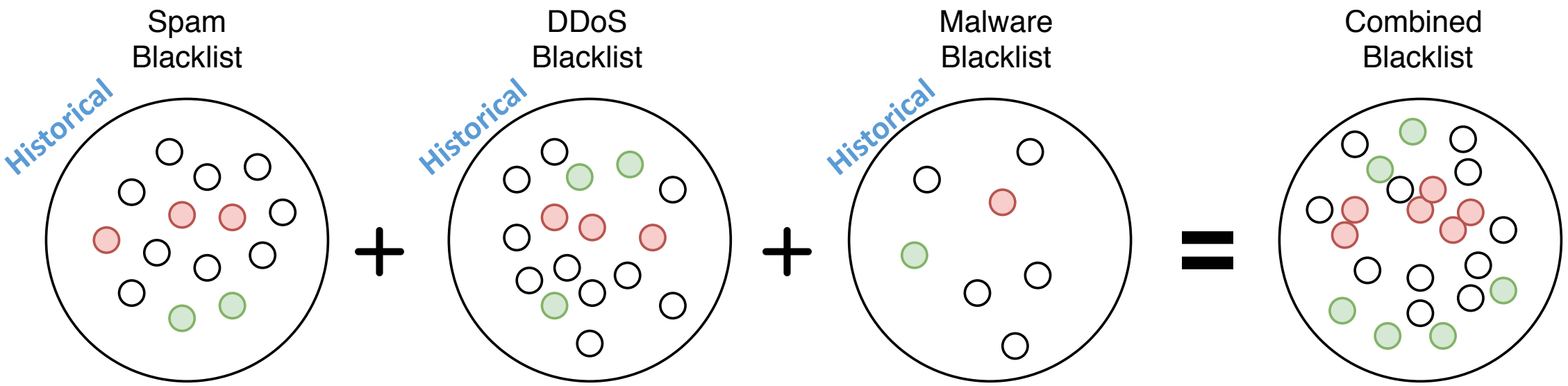


Combining blacklists can potentially amplify the number of misclassifications.

Careful Aggregation

- offenders in one given attack

- legitimate clients of a given network during the same attack

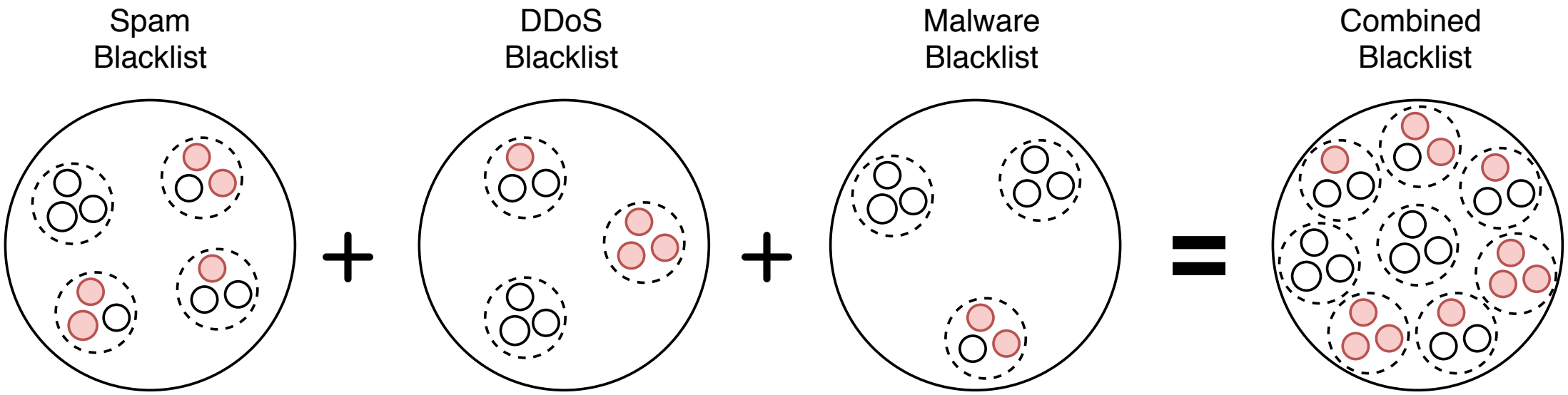


Combining blacklists can further potentially amplify the number of misclassifications.

Careful Aggregation

Goal: Aggregate historical blacklists and reduce misclassifications.

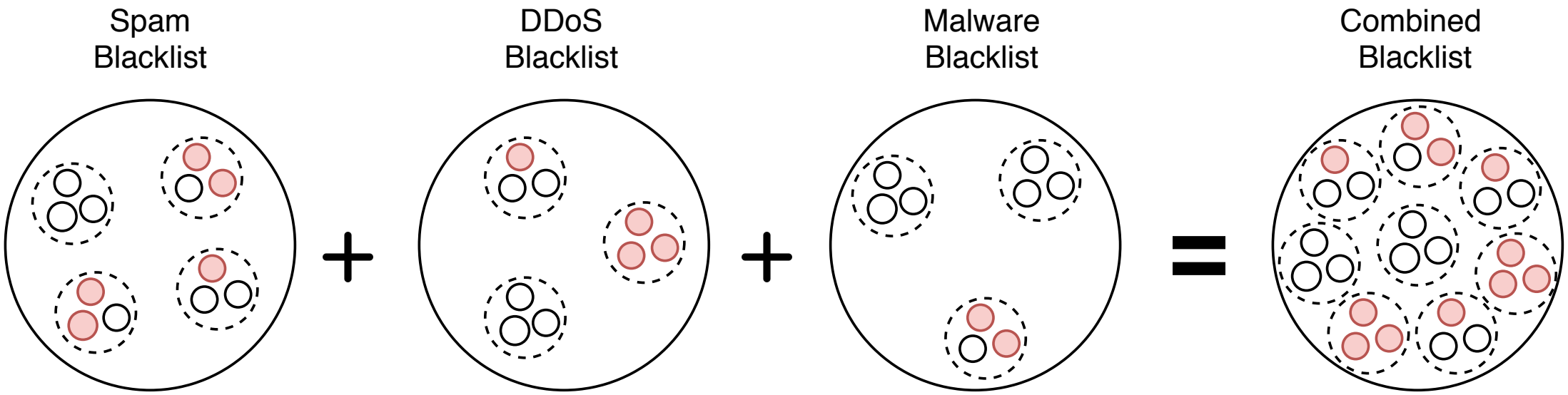
Blacklists are Reactive



Addresses are usually listed after an attack takes place, cannot be used for prevention.

Blacklists are Reactive

- offenders in one given attack



Addresses are usually listed after an attack takes place, cannot be used for prevention.

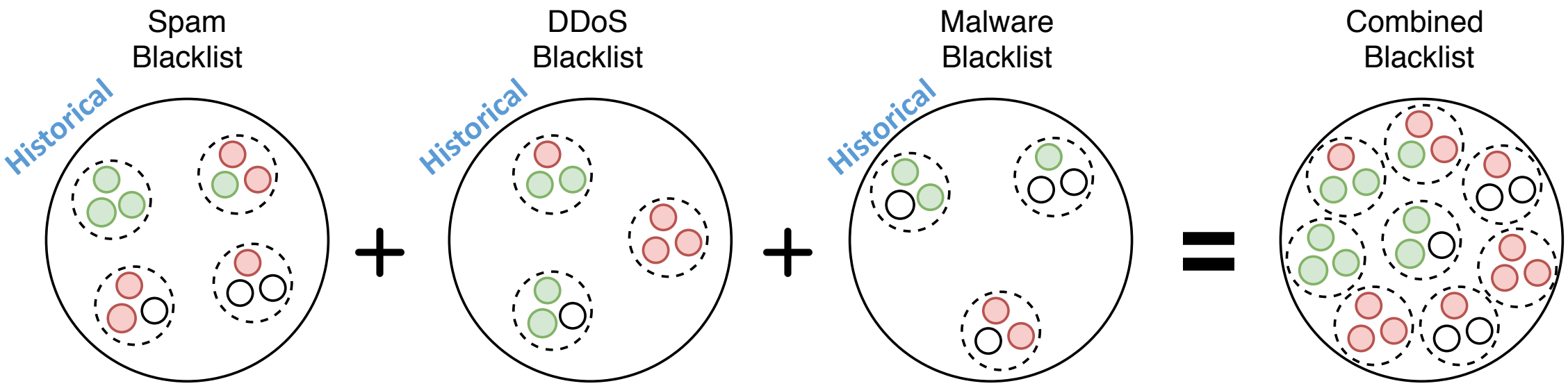
Possible solution: we could list groups of addresses in the same subnet (IP prefixes), hoping to capture future attackers - expansion[1].

[1] Zhang, Jing, et al. "On the Mismanagement and Maliciousness of Networks." NDSS. 2014.

Careful Expansion

- offenders in one given attack

- legitimate clients of a given network during the same attack



Expansion can further amplify misclassifications!

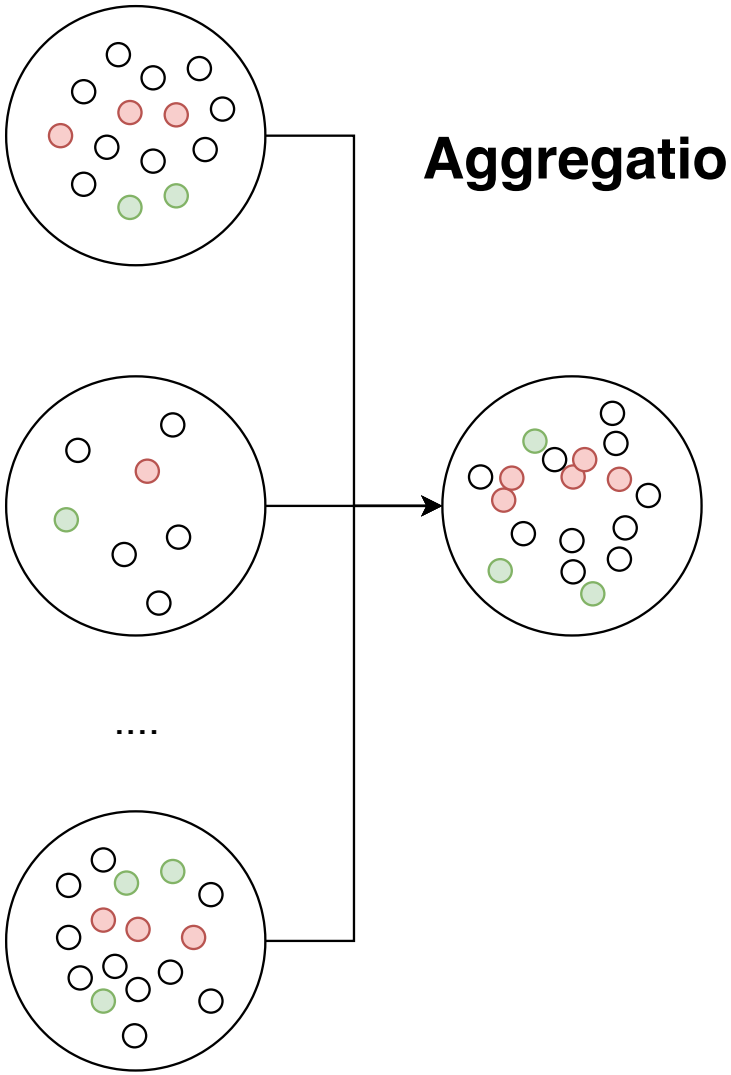
Careful Expansion

Goal: Expand some addresses into prefixes that do not cause more misclassifications.

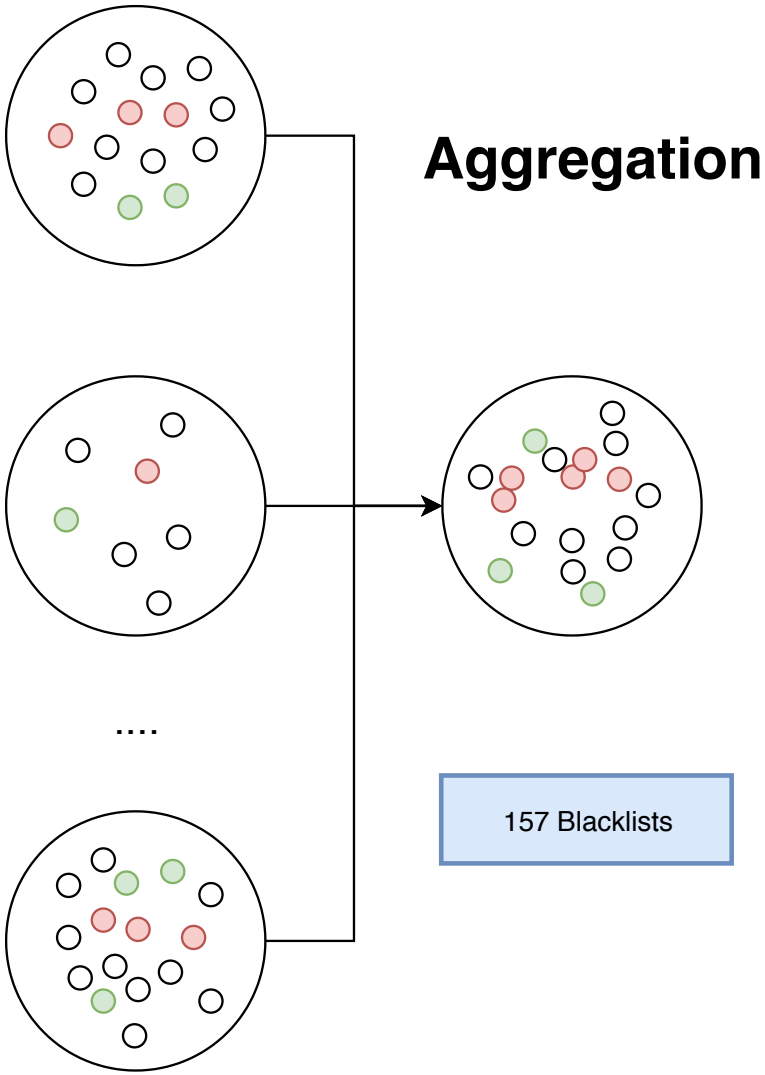
Outline

- Introduction
- Quantifying problems faced by blacklists
- BLAG
- Datasets
- Evaluation
- Summary

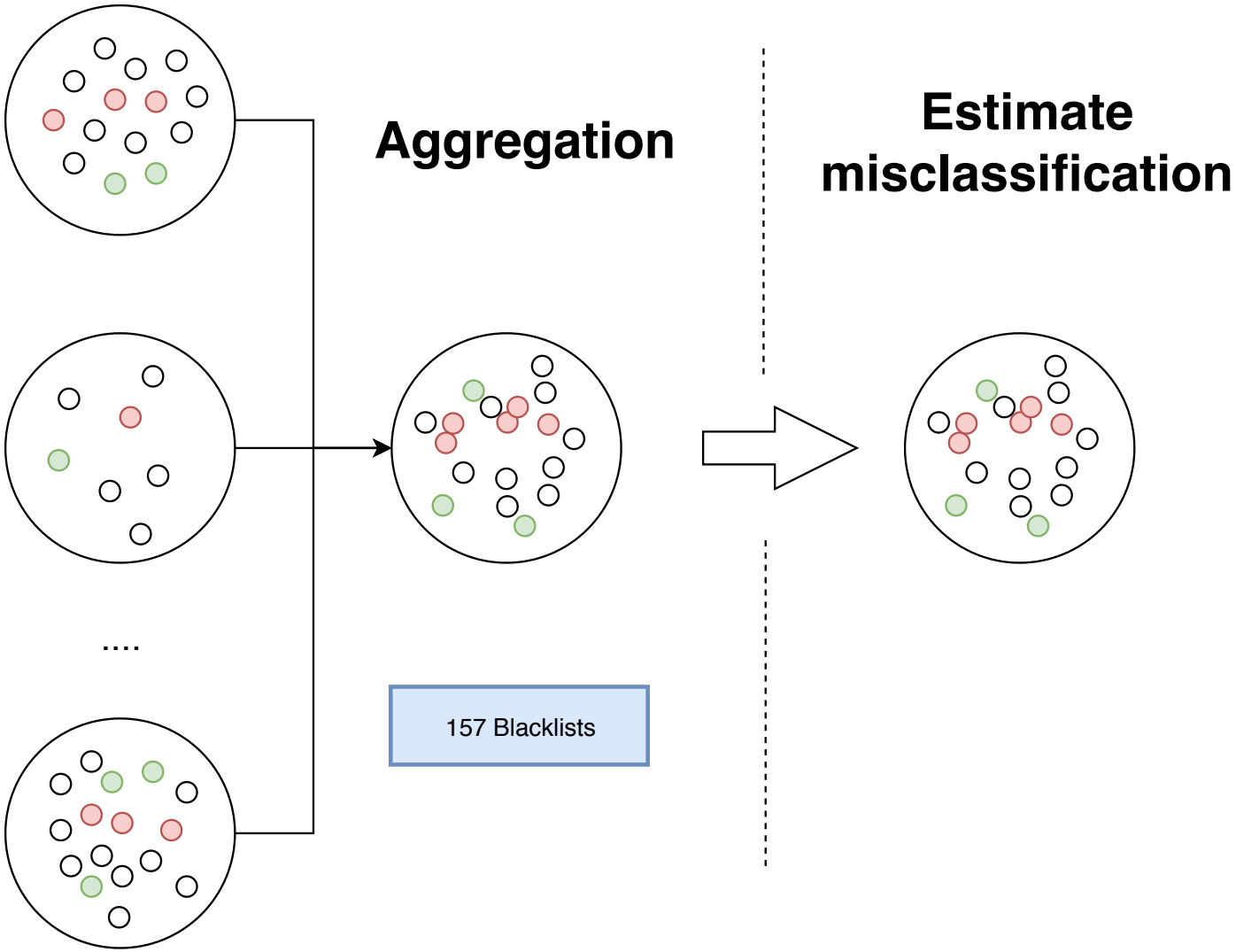
How BLAG Works



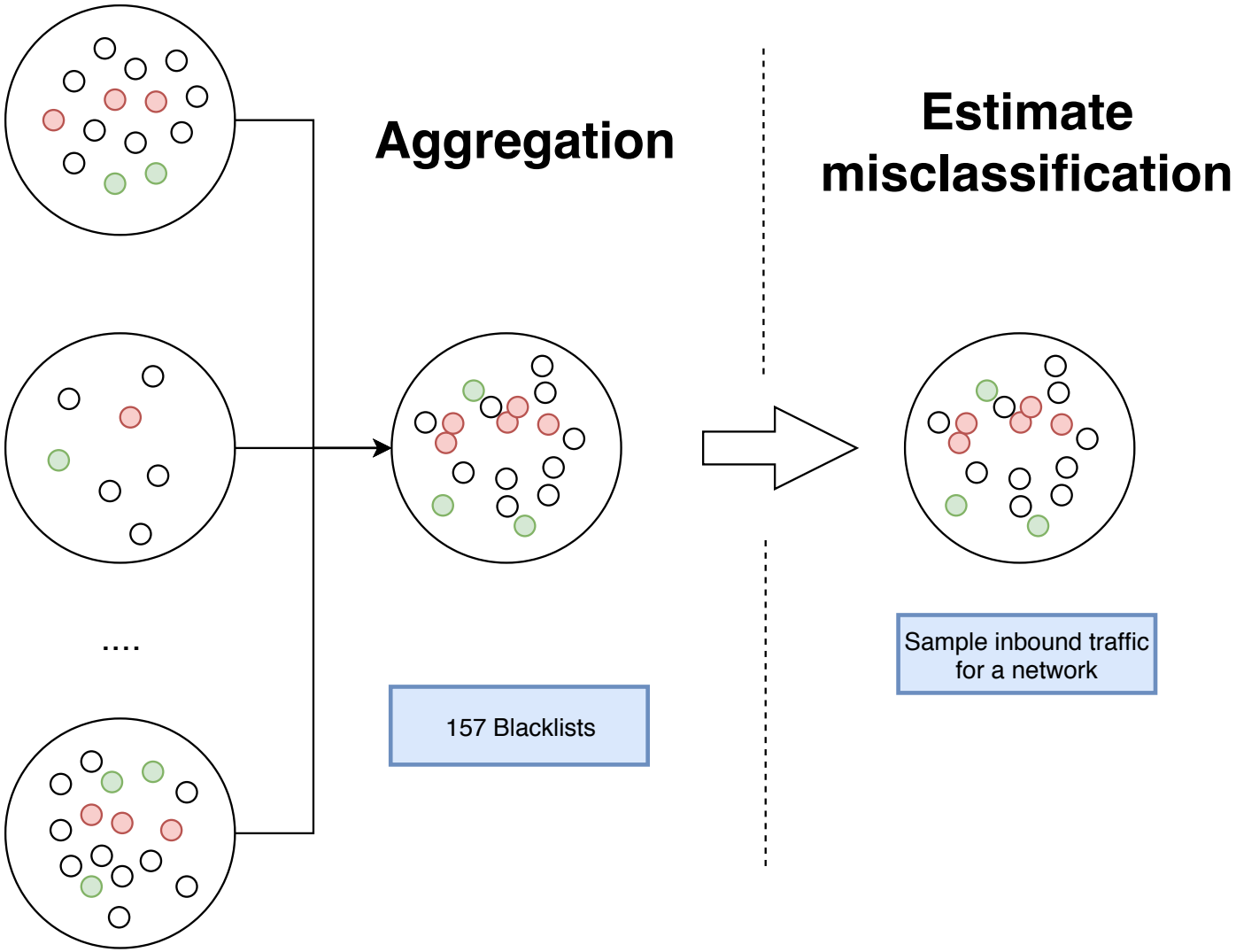
How BLAG Works



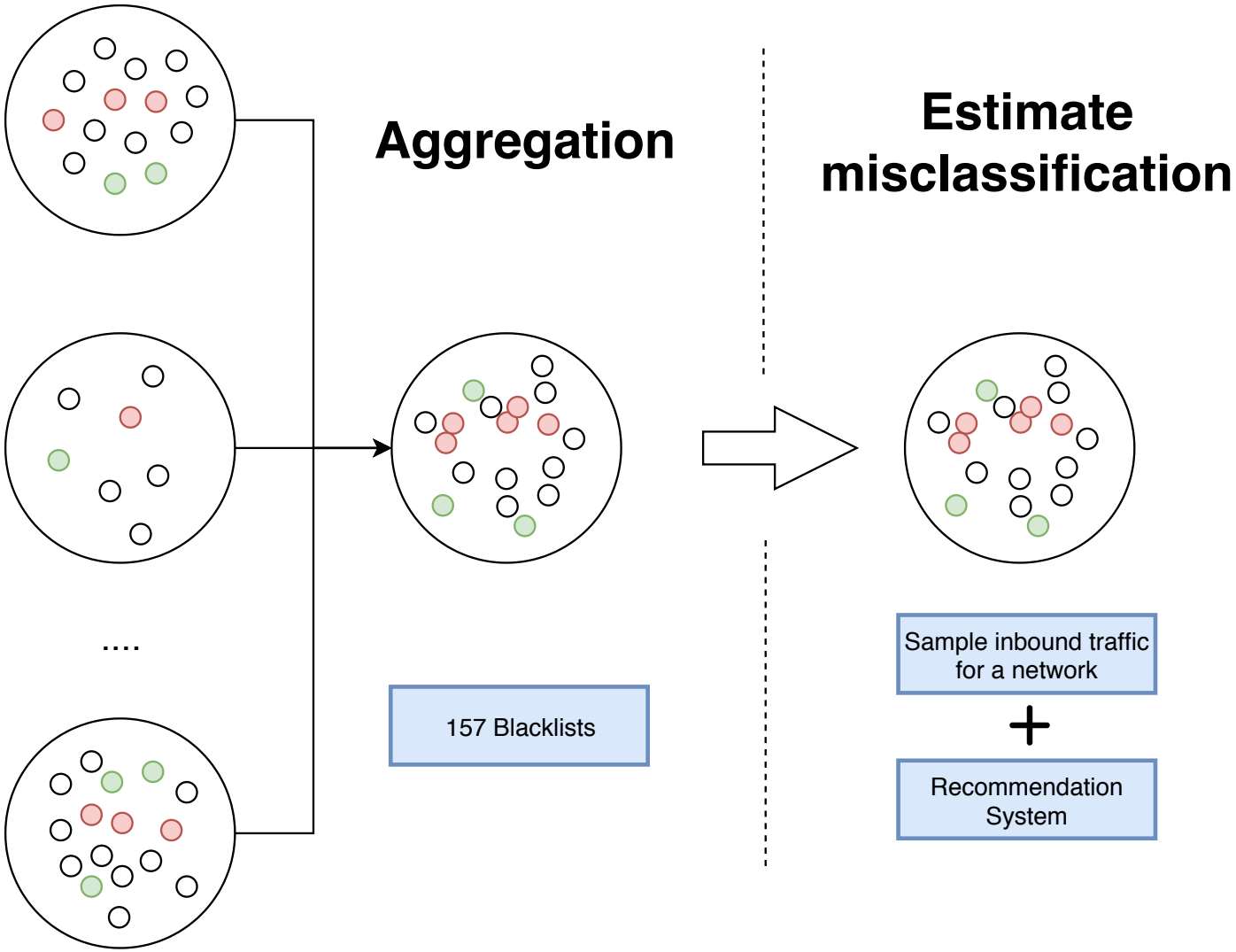
How BLAG Works



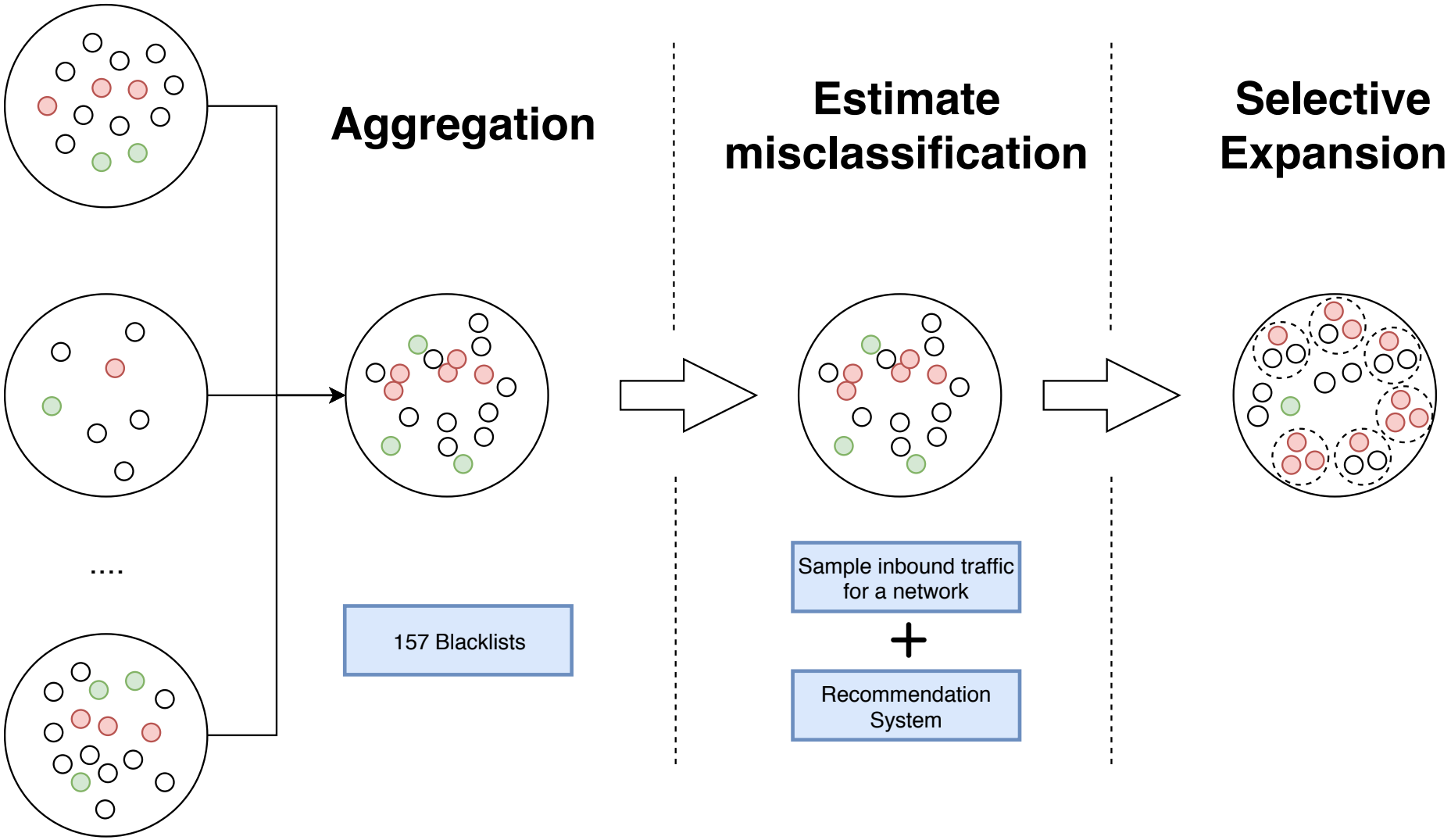
How BLAG Works



How BLAG Works



How BLAG Works



Aggregation of Blacklists

- Historical blacklist data can be useful.
- However, including addresses reported way back in the past can increase the misclassifications.
- PRESTA[1] showed that recently listed addresses have a higher tendency to be malicious than older ones.
- BLAG uses the same metric as that of PRESTA to assign a relevance score, based on when the address was listed in a blacklist
  - Recently listed addresses have a higher score.

[1] West, Andrew G., et al. "Spam mitigation using spatio-temporal reputations from blacklist history." Proceedings of the 26th Annual Computer Security Applications Conference. ACM, 2010.

Aggregation of Blacklists: Relevance Scores

- For address *a* listed in blacklist *b*,

$$r_{a,b} = 2^{\frac{t_{out}-t}{l}}$$

Aggregation of Blacklists: Relevance Scores

- For address *a* listed in blacklist *b*,

$$r_{a,b} = 2^{\frac{t_{out} - t}{l}}$$

Where,

- t is the current time

Aggregation of Blacklists: Relevance Scores

- For address *a* listed in blacklist *b*,

$$r_{a,b} = 2^{\frac{t_{out}-t}{l}}$$

Where,

- *t* is the current time
- t_{out} is the last time when an address *a* was listed in blacklist *b*

Aggregation of Blacklists: Relevance Scores

- For address *a* listed in blacklist *b*,

$$r_{a,b} = 2^{\frac{t_{out}-t}{l}}$$

Where,

- *t* is the current time
- t_{out} is the last time when an address *a* was listed in blacklist *b*
- *l* is constant, which ensures that the score decays over time

Aggregation of Blacklists: Relevance Scores

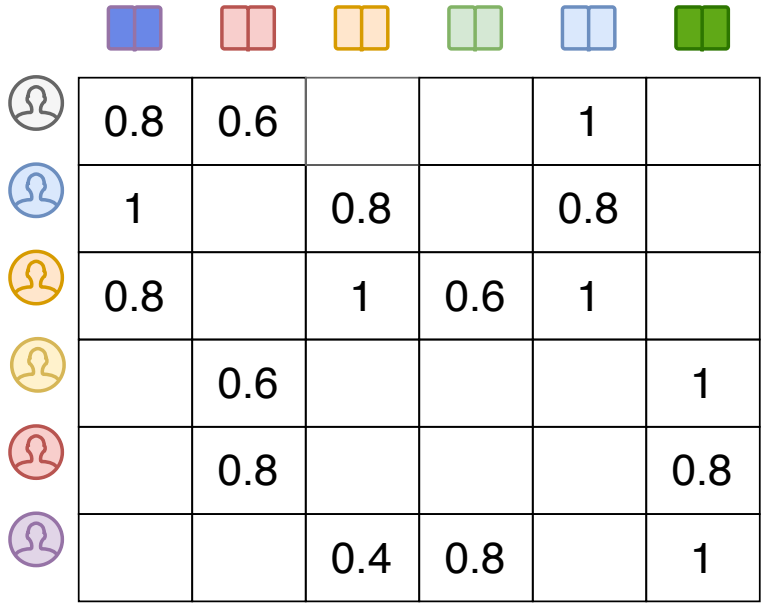
- For address a listed in blacklist b,

$$2^{\frac{t_{out}-t}{l}}$$

A high relevance score means that an IP has been recently listed and has a higher tendency of being malicious.

- t is the current time
- t_{out} is the last time when address a was listed in blacklist b
- l is constant, which ensures that the score decays exponentially over time

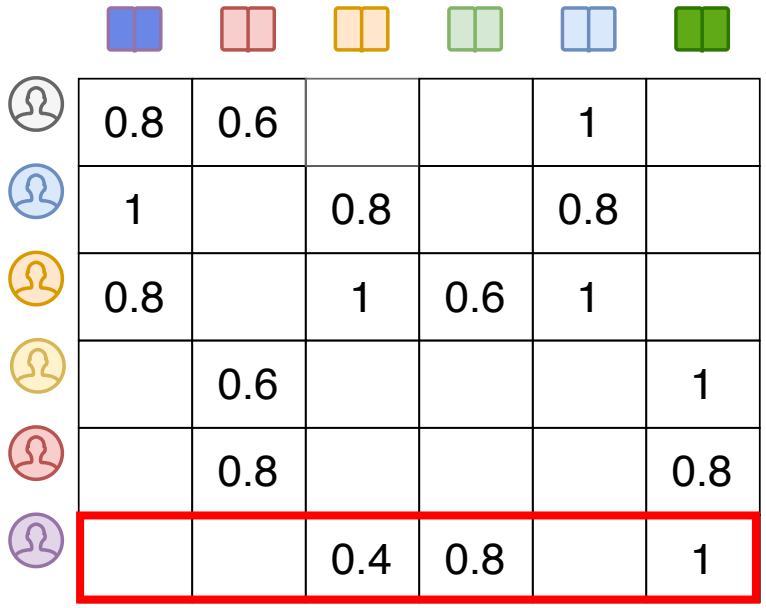
Estimate Misclassifications– Recommendation System



0.8	0.6			1	
1		0.8		0.8	
0.8		1	0.6	1	
	0.6				1
	0.8				0.8
		0.4	0.8		1

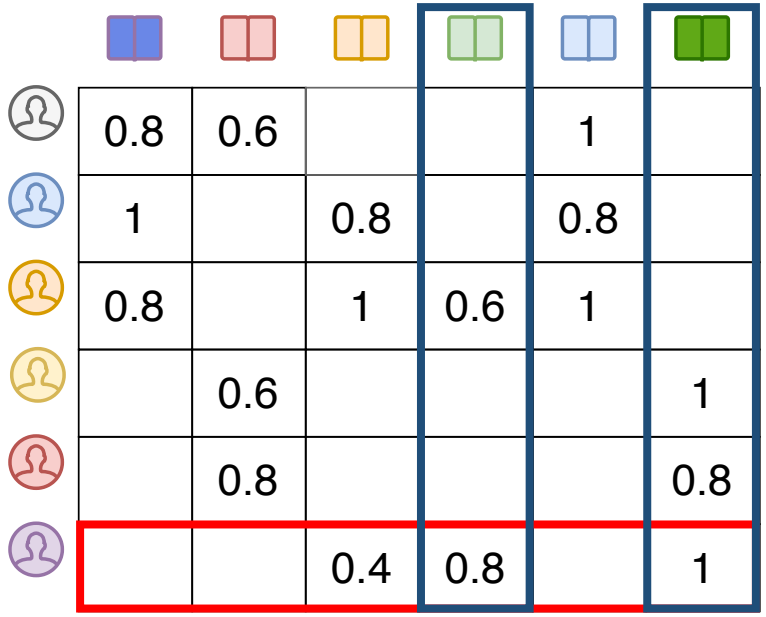
- Commonly found in popular services like Netflix, Amazon, and YouTube to improve user retention and increase revenue.
- Recommend new items to users based on their or similar users' previous ratings of similar items.

Estimate Misclassifications– Recommendation System



0.8	0.6			1	
1		0.8		0.8	
0.8		1	0.6	1	
	0.6				1
	0.8				0.8
		0.4	0.8		1

Estimate Misclassifications– Recommendation System



0.8	0.6			1	
1		0.8		0.8	
0.8		1	0.6	1	
	0.6				1
	0.8				0.8
		0.4	0.8		1

Likes green books.

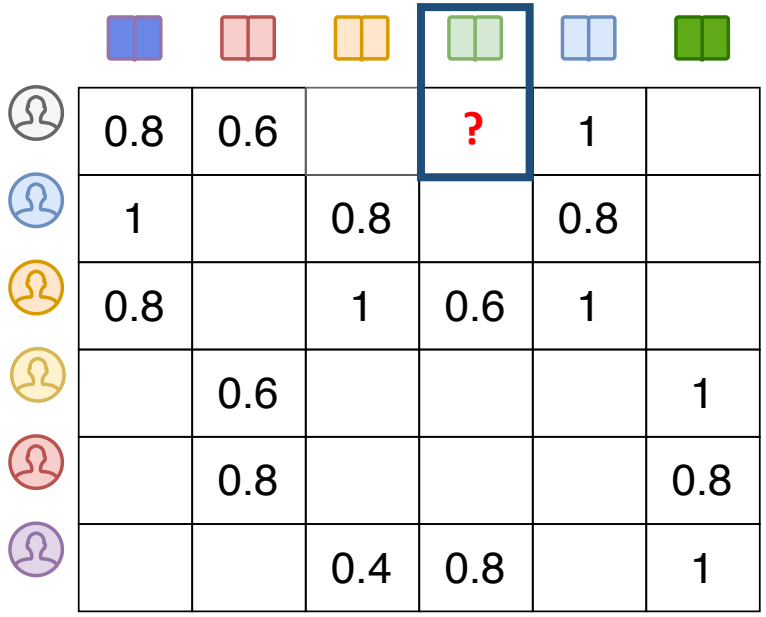
Estimate Misclassifications– Recommendation System

0.8	0.6			1	
1		0.8		0.8	
0.8		1	0.6	1	
	0.6				1
	0.8				0.8
		0.4	0.8		1

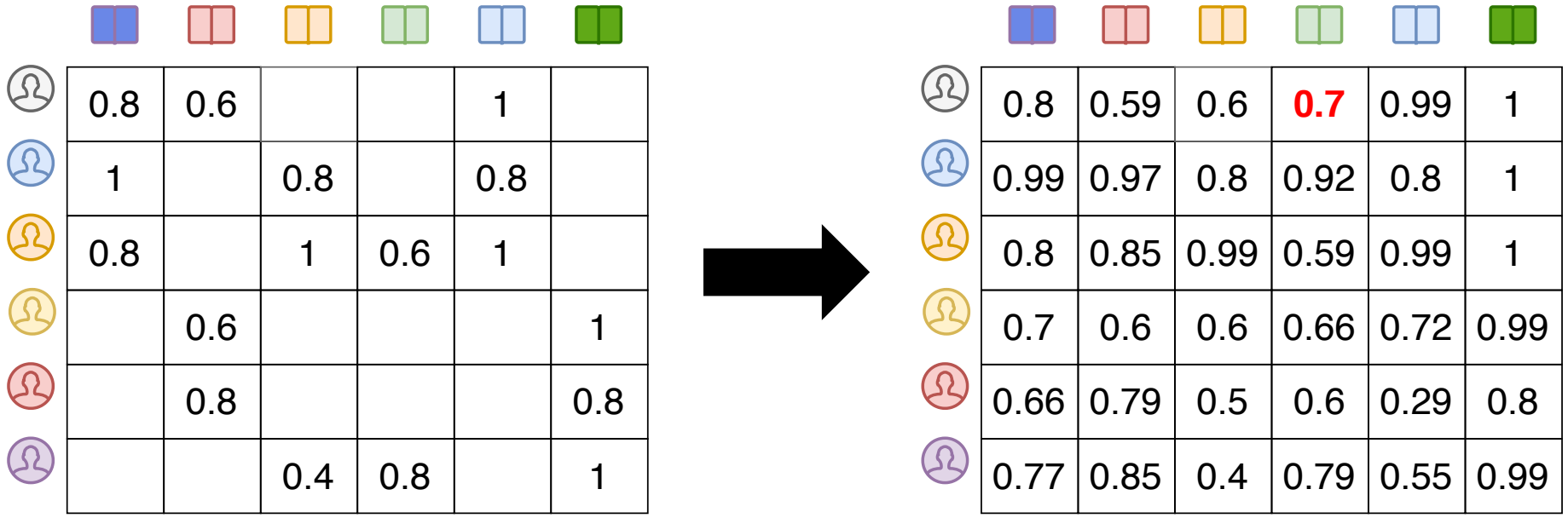
Likes green books.

Dislikes yellow books.

Estimate Misclassifications– Recommendation System



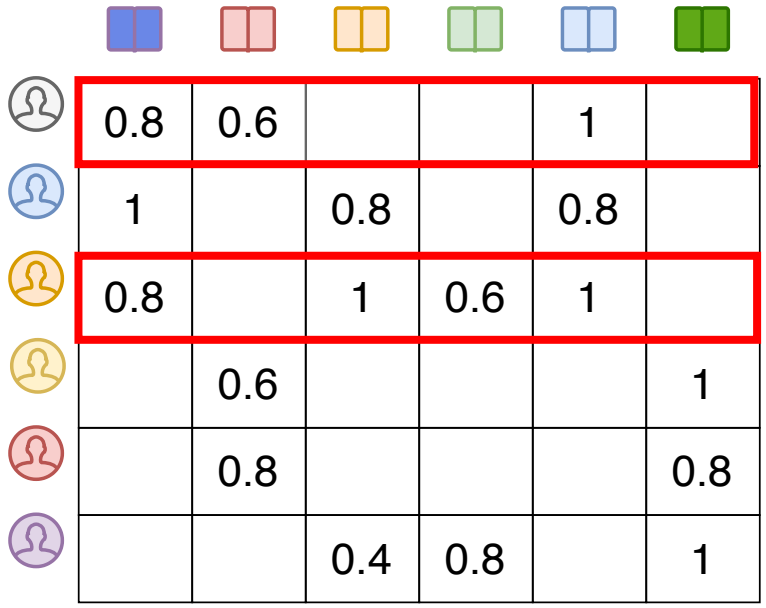
Estimate Misclassifications– Recommendation System



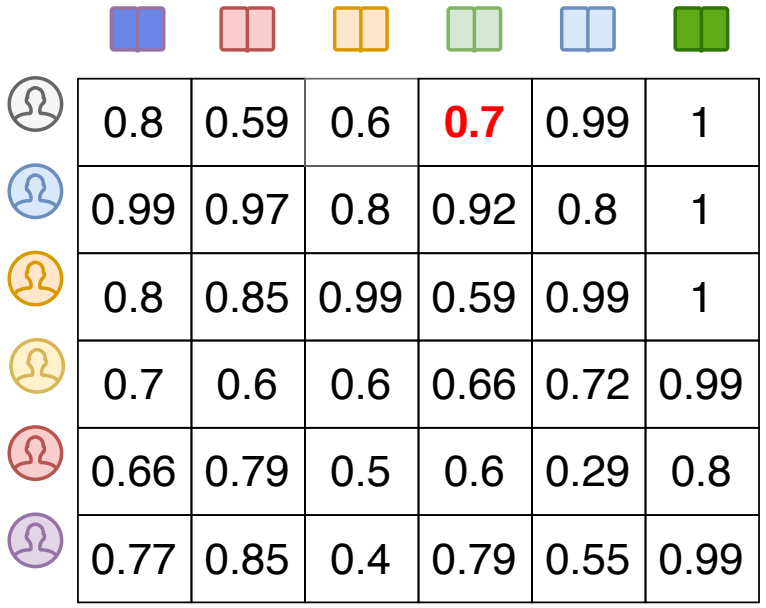
0.8	0.6			1	
1		0.8		0.8	
0.8		1	0.6	1	
	0.6				1
	0.8				0.8
		0.4	0.8		1

0.8	0.59	0.6	**0.7**	0.99	1
0.99	0.97	0.8	0.92	0.8	1
0.8	0.85	0.99	0.59	0.99	1
0.7	0.6	0.6	0.66	0.72	0.99
0.66	0.79	0.5	0.6	0.29	0.8
0.77	0.85	0.4	0.79	0.55	0.99

Estimate Misclassifications– Recommendation System

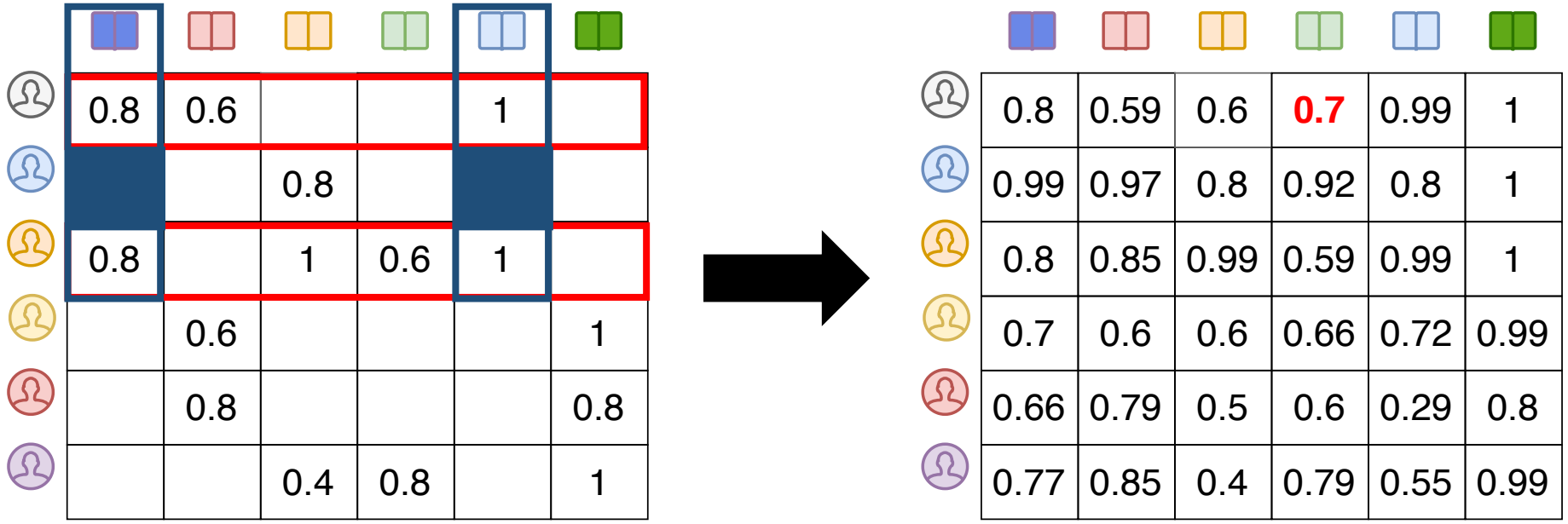


0.8	0.6			1	
1		0.8		0.8	
0.8		1	0.6	1	
	0.6				1
	0.8				0.8
		0.4	0.8		1



0.8	0.59	0.6	**0.7**	0.99	1
0.99	0.97	0.8	0.92	0.8	1
0.8	0.85	0.99	0.59	0.99	1
0.7	0.6	0.6	0.66	0.72	0.99
0.66	0.79	0.5	0.6	0.29	0.8
0.77	0.85	0.4	0.79	0.55	0.99

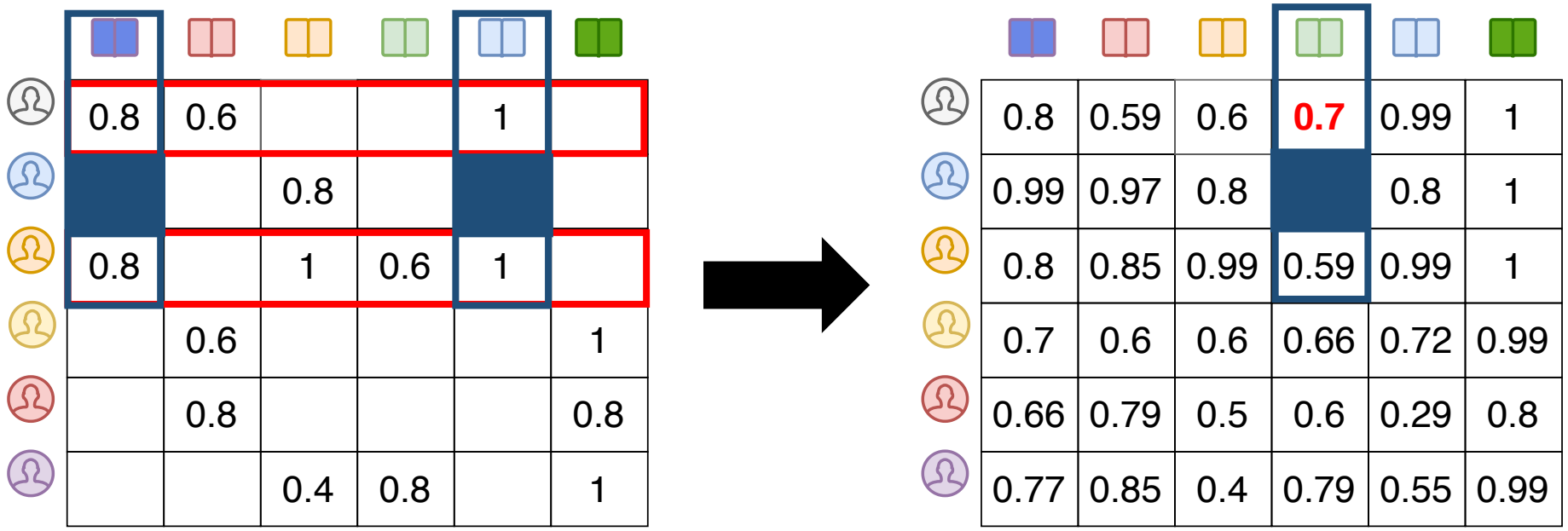
Estimate Misclassifications– Recommendation System



0.8	0.6			1	
		0.8			
0.8		1	0.6	1	
	0.6				1
	0.8				0.8
		0.4	0.8		1

0.8	0.59	0.6	**0.7**	0.99	1
0.99	0.97	0.8	0.92	0.8	1
0.8	0.85	0.99	0.59	0.99	1
0.7	0.6	0.6	0.66	0.72	0.99
0.66	0.79	0.5	0.6	0.29	0.8
0.77	0.85	0.4	0.79	0.55	0.99

Estimate Misclassifications– Recommendation System



Estimate Misclassifications

	Blacklist 1	Blacklist 2	Blacklist 3	..	Blacklist m
169.231.140.10	..	..	0.8	..	..
169.231.140.68	0.3	0.1	..	..	0.1
193.1.64.5	..	0.5	..	..	..
193.1.64.8	0.7	0.5	..	..	0.9
216.59.0.8	..	..	0.04	..	0.1
216.59.16.171	..	0.7	..	..	0.9
243.13.0.23	..	..	..	..	..
243.13.222.203	..	0.7	1	..	0.9

- BLAG arranges IP addresses and blacklists in a matrix, where rows are addresses and columns are blacklists.
- If an address *a* is listed in blacklist *b*, BLAG assigns the relevance score $r_{a,b}$ to the cell.

Estimate Misclassifications

	Blacklist 1	Blacklist 2	Blacklist 3	..	Blacklist m	MB
169.231.140.10	..	..	0.8	..	..	
169.231.140.68	0.3	0.1	..	..	0.1	
193.1.64.5	..	0.5	..	..	..	
193.1.64.8	0.7	0.5	..	..	0.9	
216.59.0.8	..	..	0.04	..	0.1	
216.59.16.171	..	0.7	..	..	0.9	
243.13.0.23	..	..	..	..	..	
243.13.222.203	..	0.7	1	..	0.9	

BLAG uses legitimate traffic traces of a network to introduce a new blacklist called the Misclassification Blacklist (MB), which consists only of misclassifications.

Estimate Misclassifications

	Blacklist 1	Blacklist 2	Blacklist 3	..	Blacklist m	MB
169.231.140.10	..	..	0.8	..	..	
169.231.140.68	0.3	0.1	..	..	0.1	
193.1.64.5	..	0.5	..	..	..	
193.1.64.8	0.7	0.5	..	..	0.9	
216.59.0.8	..	..	0.04	..	0.1	
216.59.16.171	..	0.7	..	..	0.9	1
243.13.0.23	..	..	..	..	..	1
243.13.222.203	..	0.7	1	..	0.9	1

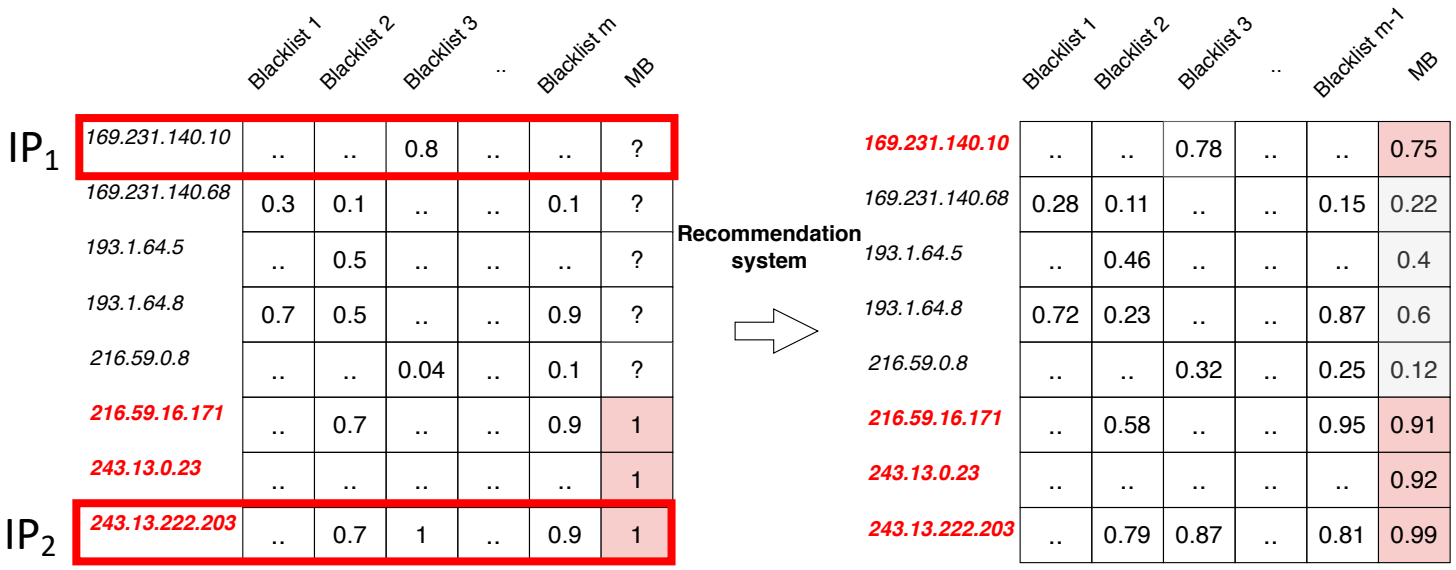
For every known misclassification from the training data, BLAG allocates a score of 1.

Estimate Misclassifications

	Blacklist 1	Blacklist 2	Blacklist 3	..	Blacklist m	MB
169.231.140.10	..	..	0.8	..	..	?
169.231.140.68	0.3	0.1	..	..	0.1	?
193.1.64.5	..	0.5	..	..	..	?
193.1.64.8	0.7	0.5	..	..	0.9	?
216.59.0.8	..	..	0.04	..	0.1	?
216.59.16.171	..	0.7	..	..	0.9	1
243.13.0.23	..	..	..	..	..	1
243.13.222.203	..	0.7	1	..	0.9	1

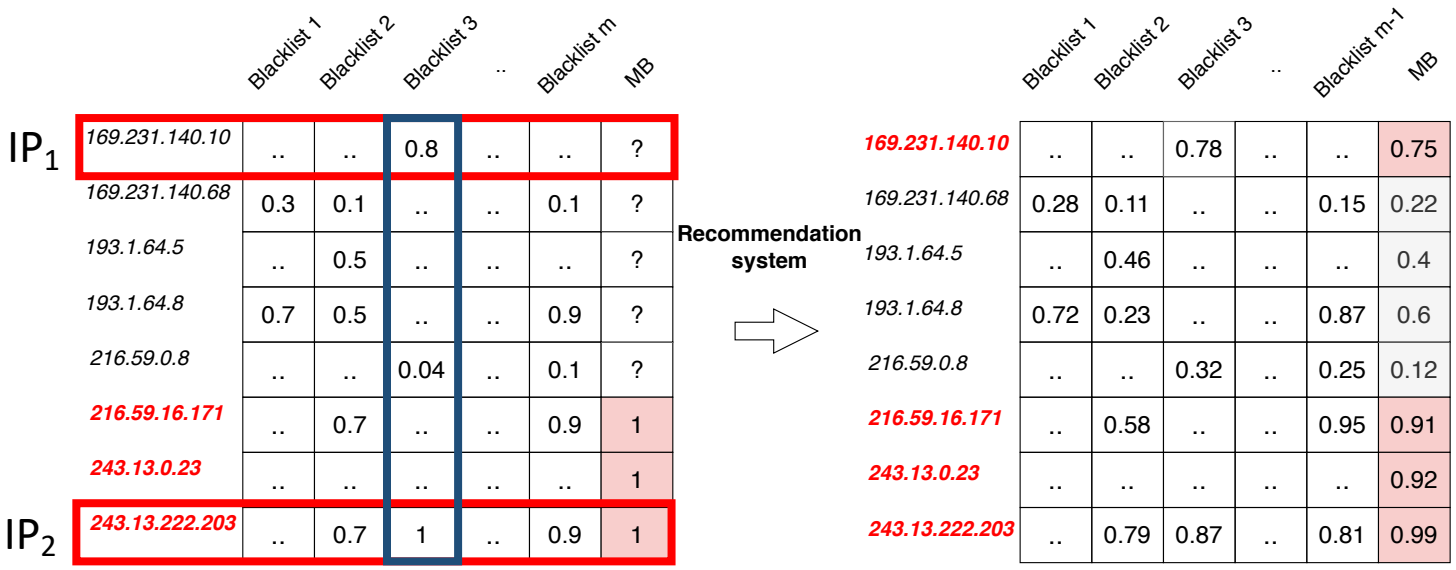
Goal: Find the relevance scores for remaining addresses in MB.

Estimate Misclassifications



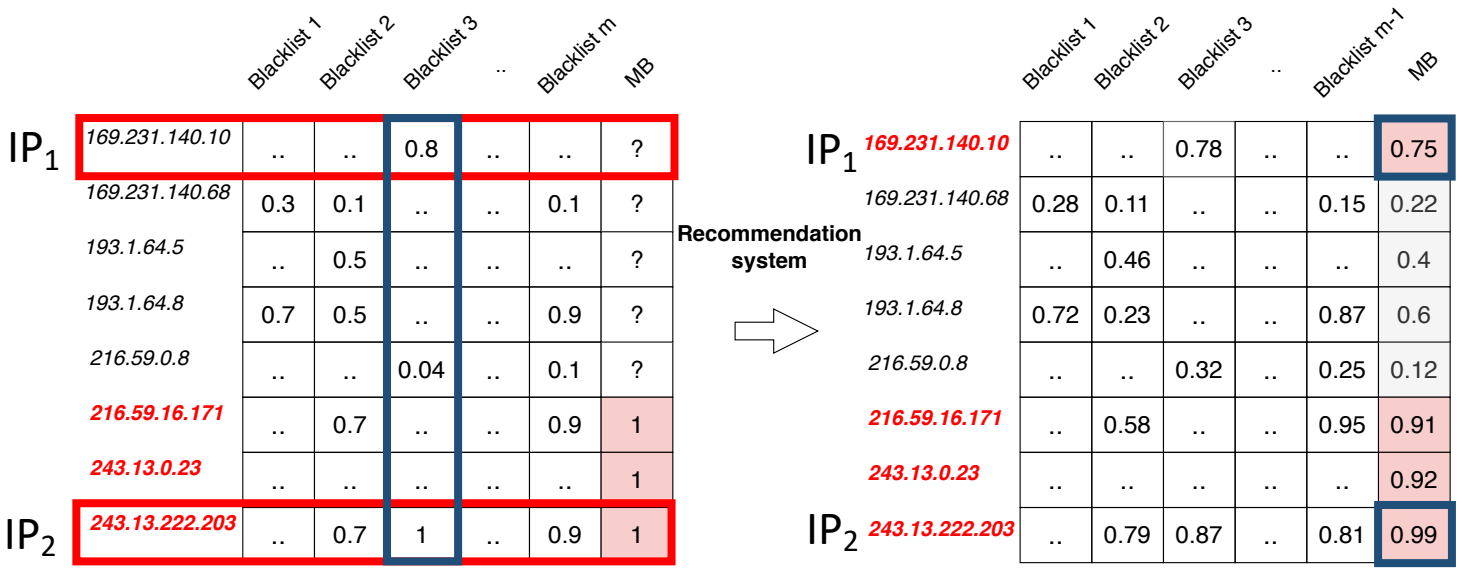
Goal: Find the relevance scores for remaining addresses in MB.

Estimate Misclassifications



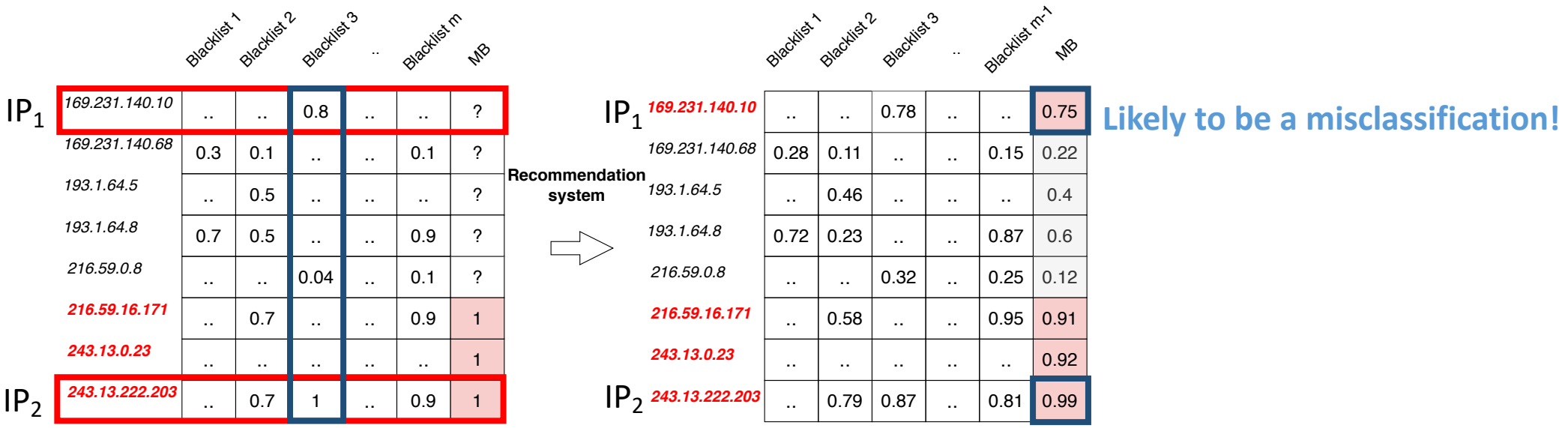
Goal: Find the relevance scores for remaining addresses in MB.

Estimate Misclassifications



Goal: Find the relevance scores for remaining addresses in MB.

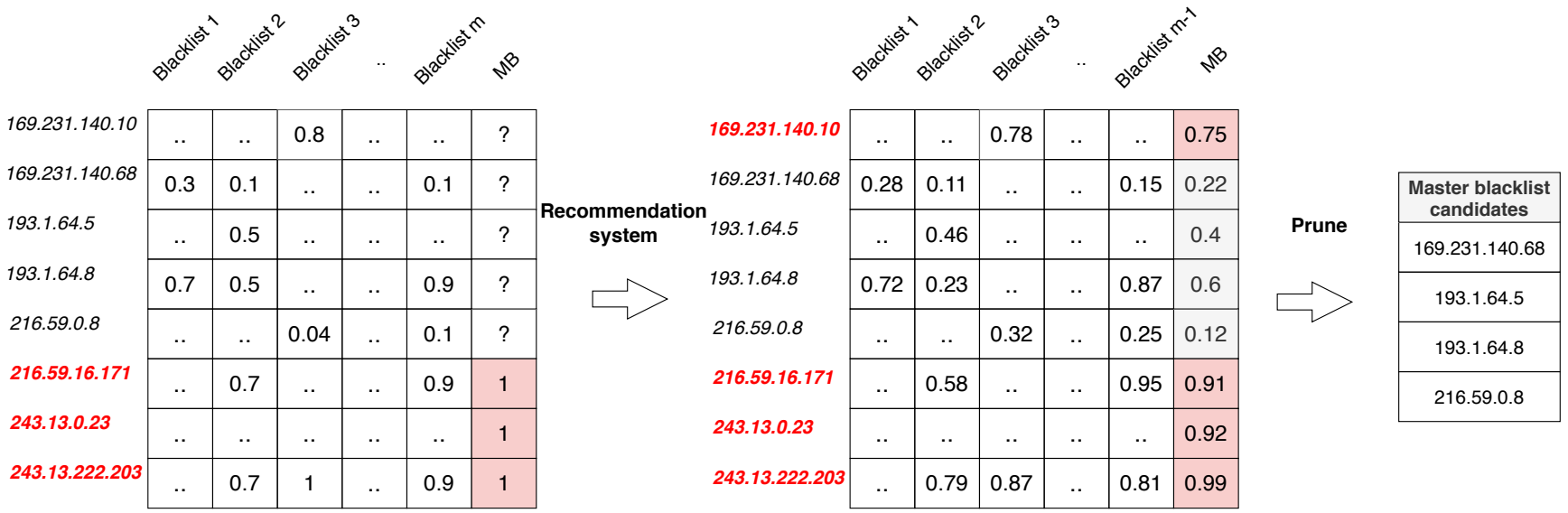
Estimate Misclassifications



Likely to be a misclassification!

Goal: Find the relevance scores for remaining addresses in MB.

Estimate Misclassifications

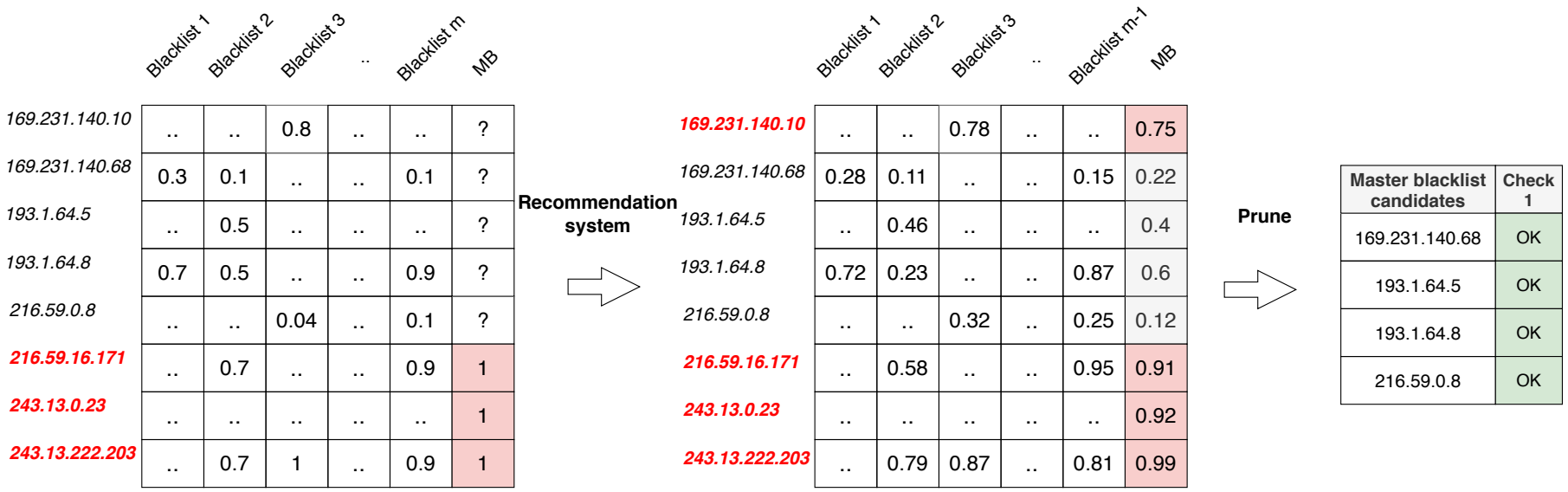


Using a defined threshold customized for every network (0.7 in this case), BLAG prune out addresses that are potentially misclassified.

Why Recommendation System?

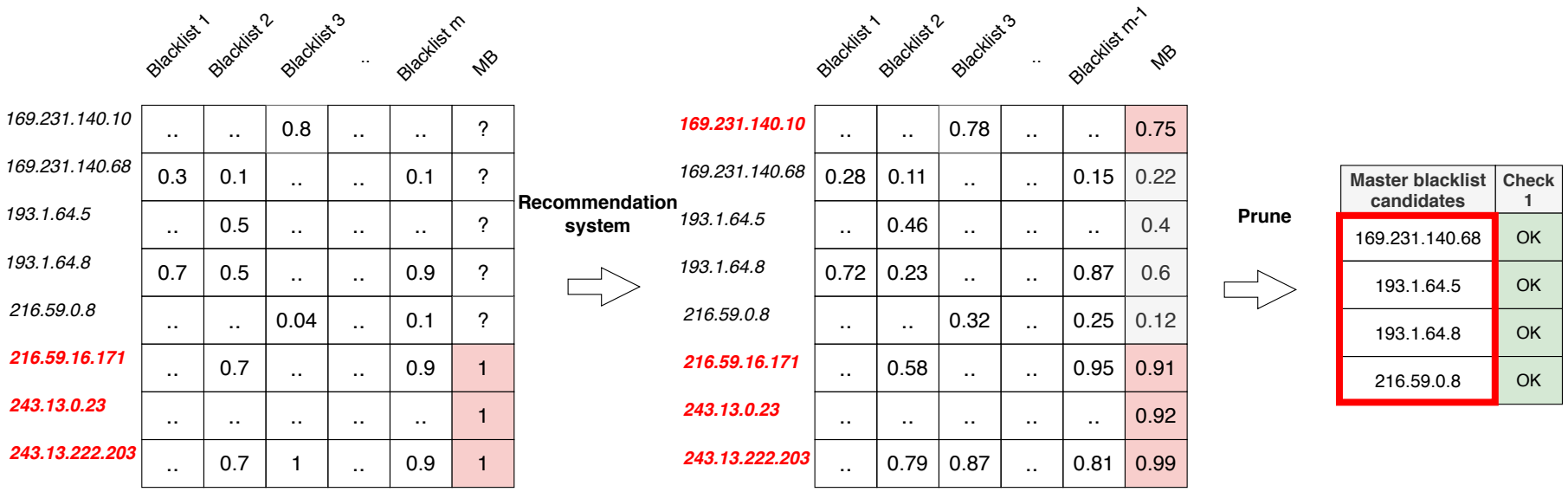
- Given the incomplete view of the address space, there are many addresses that cannot be determined to be a misclassification (or not).
- Several latent factors influence an address to be a misclassification.
  - Proprietary algorithms historical data or overall reputation of the blacklist
- The recommendation system helps us identify other addresses:
  - Which “behave” similar to our known misclassifications.
  - They are listed on same or similar blacklists as our known misclassifications, with similar scores.

Selective Expansion



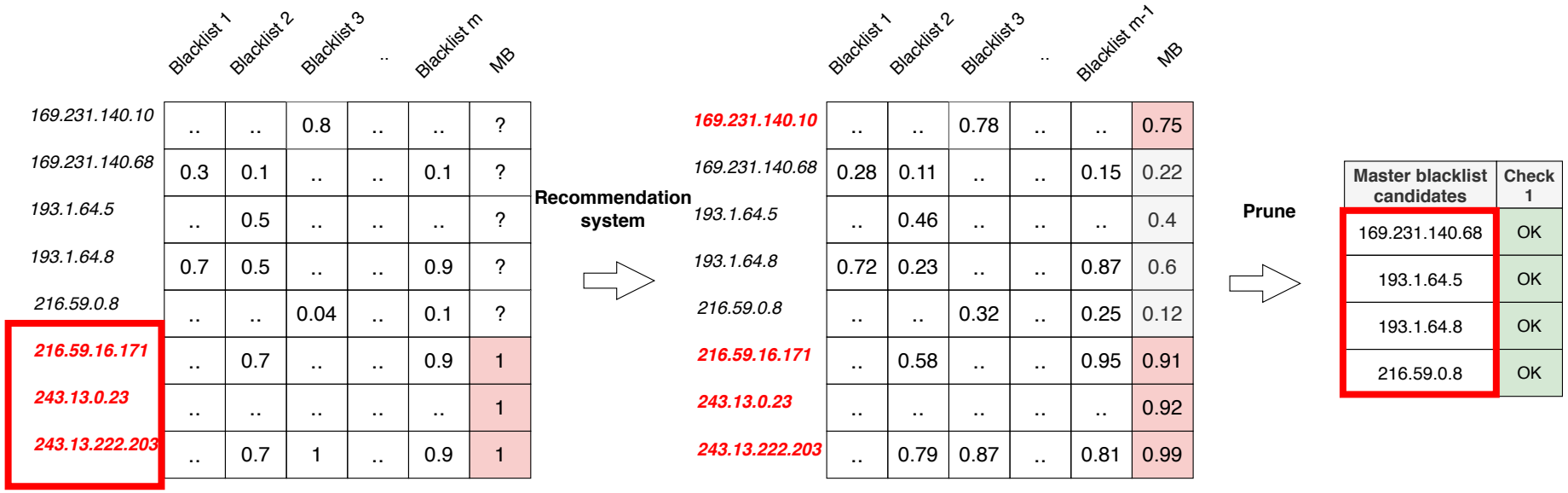
Check 1: If a prefix has any known misclassification, it is excluded from expansion.

Selective Expansion



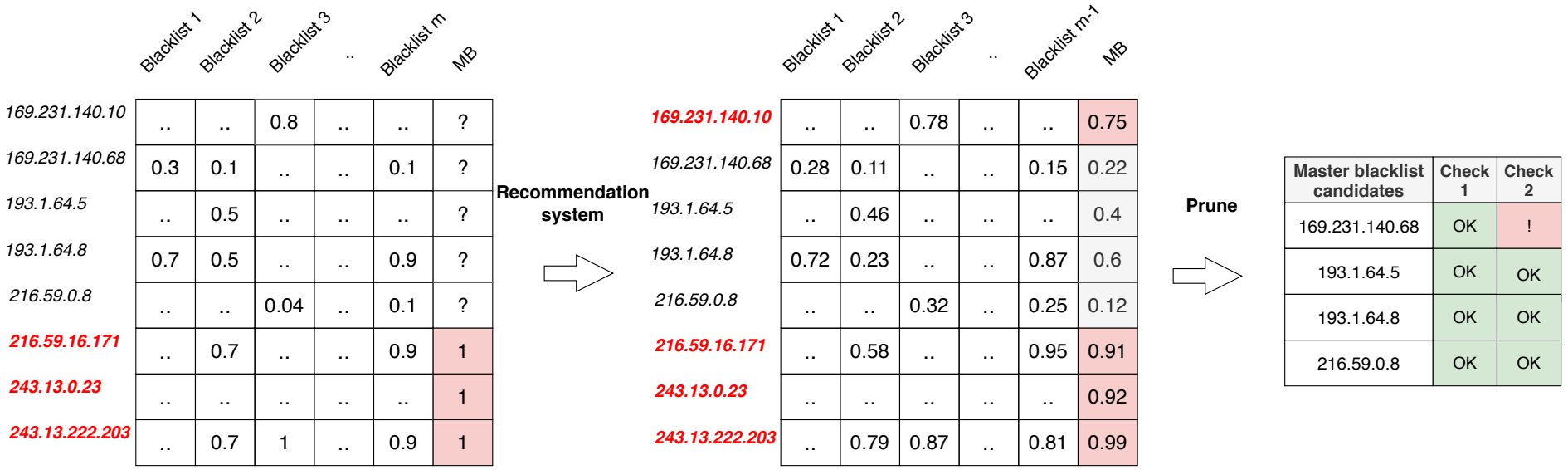
Check 1: If a prefix has any known misclassification, it is excluded from expansion.

Selective Expansion



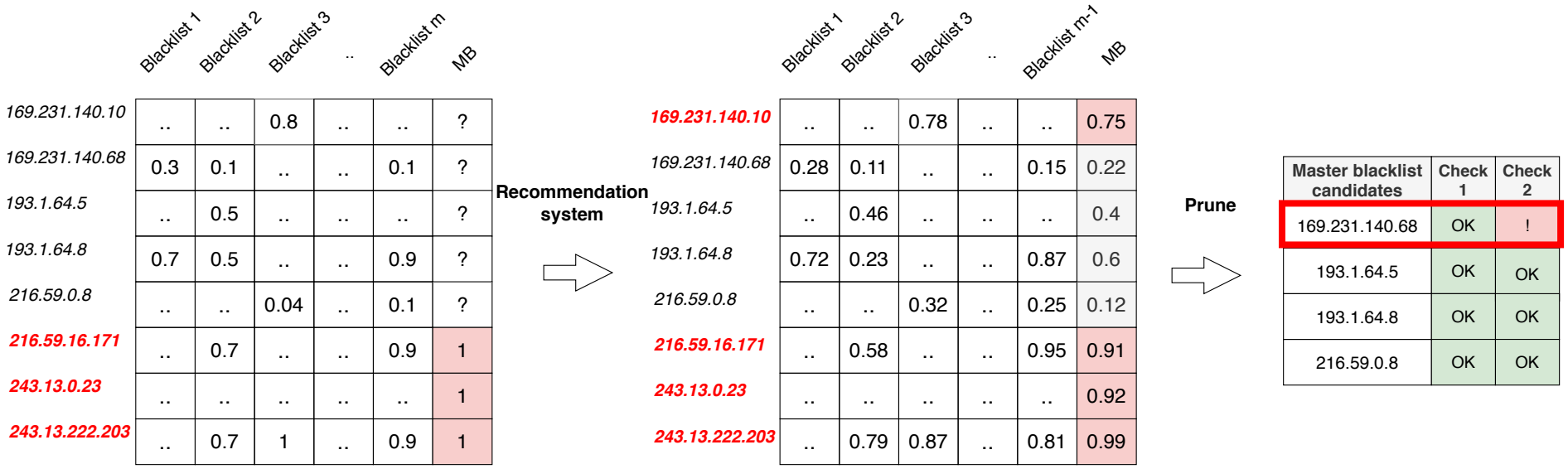
Check 1: If a prefix has any known misclassification, it is excluded from expansion.

Selective Expansion



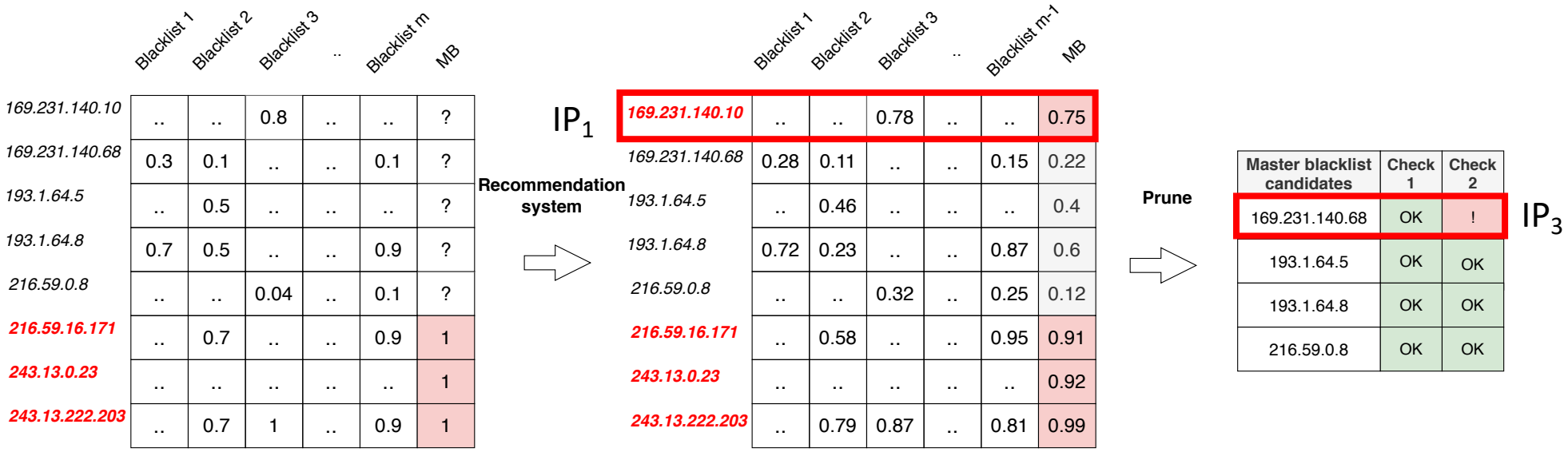
Check 2: If a prefix has any likely misclassification, it is excluded from expansion.

Selective Expansion



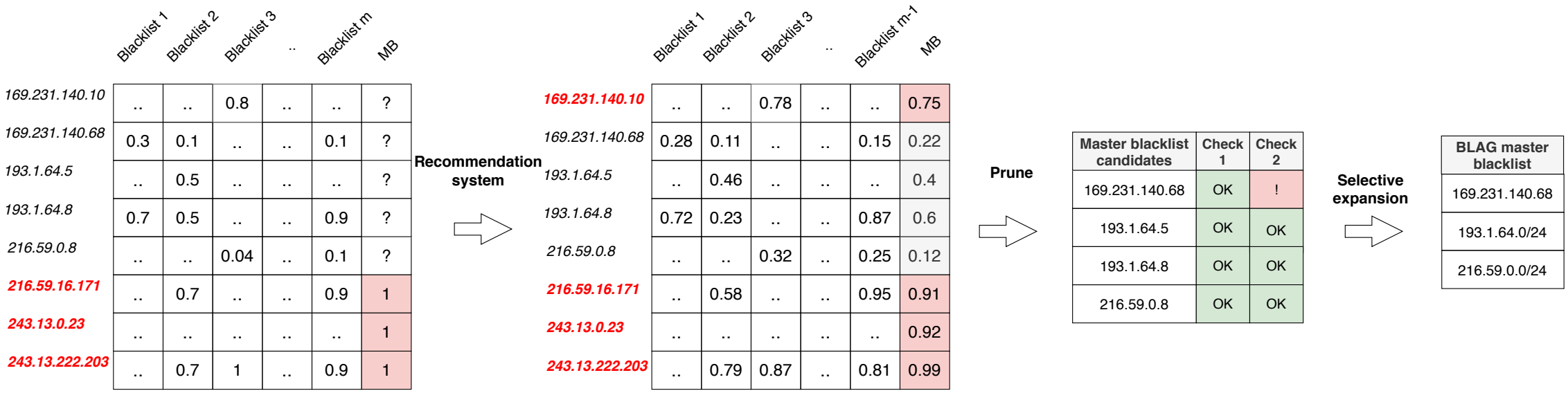
Check 2: If a prefix has any likely misclassification, it is excluded from expansion.

Selective Expansion



Check 2: If a prefix has any likely misclassification, it is excluded from expansion.

Selective Expansion

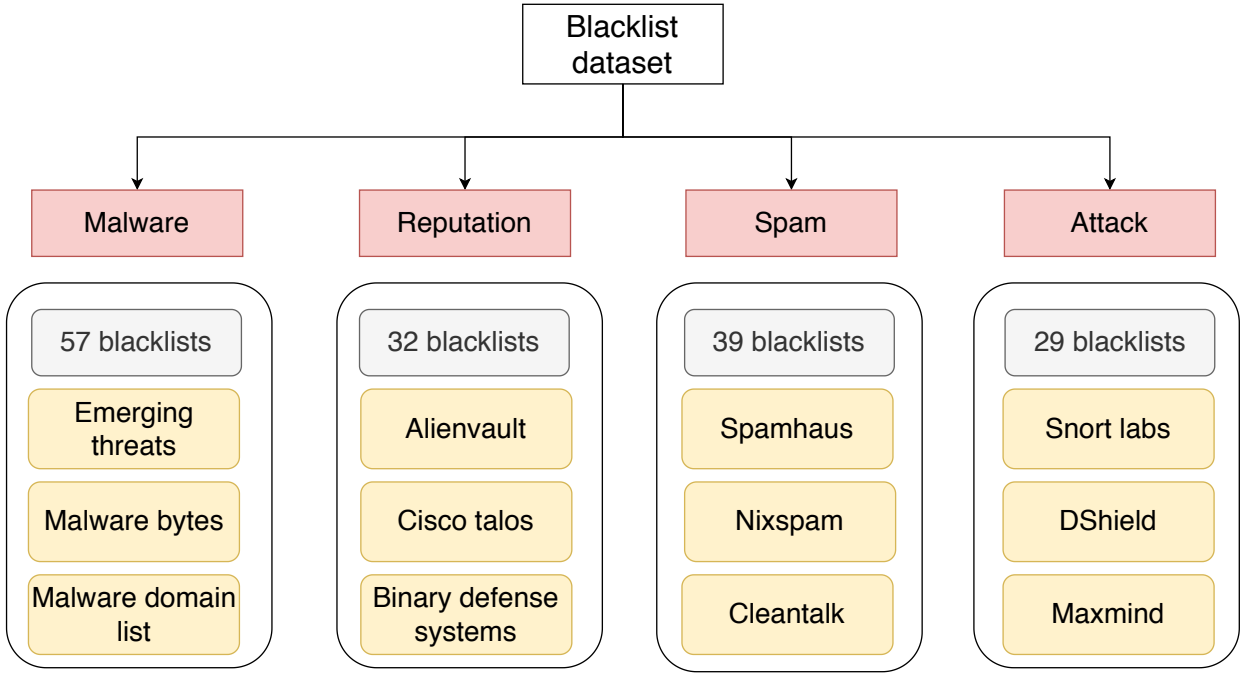


BLAG expands addresses to their /24 prefix only when both conditions are satisfied.

Outline

- Introduction
- Quantifying problems faced by blacklists
- BLAG
- Datasets
- Evaluation
- Summary

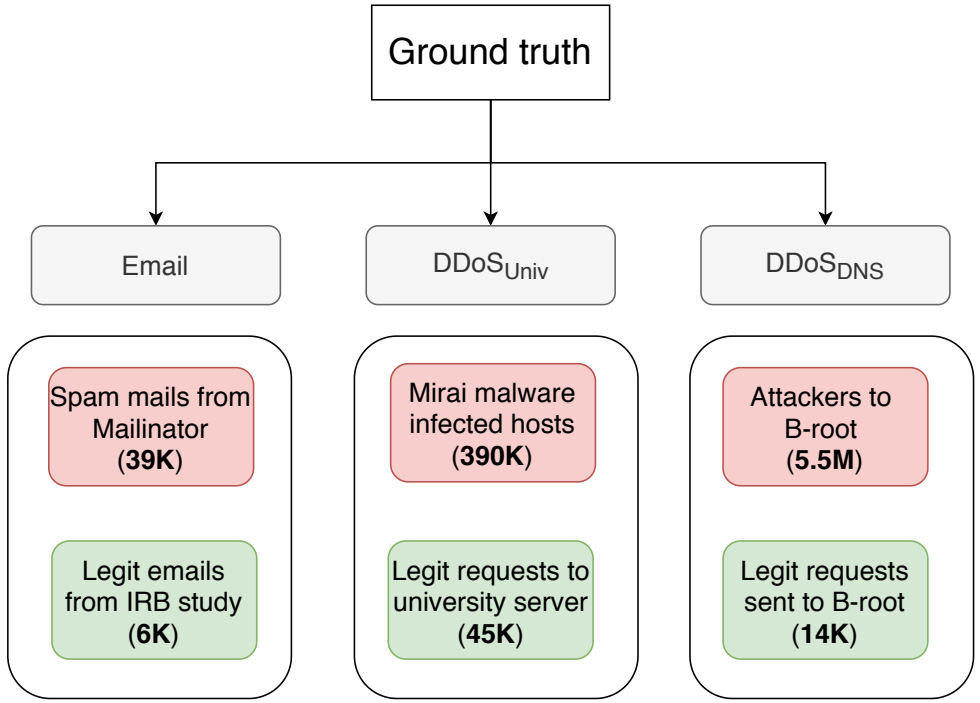
Monitored Blacklists



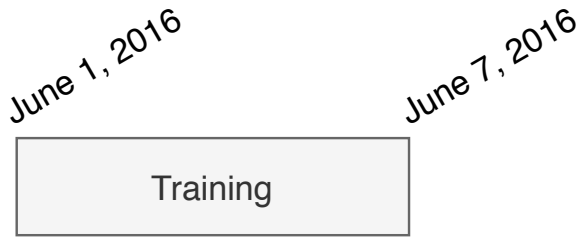
- 157 blacklists monitored from Jan 2016 to Dec 2017 roughly categorized into four attack variants.
- Collected over *176 million* IP addresses during this period.

Ground Truth for Evaluating Blacklists

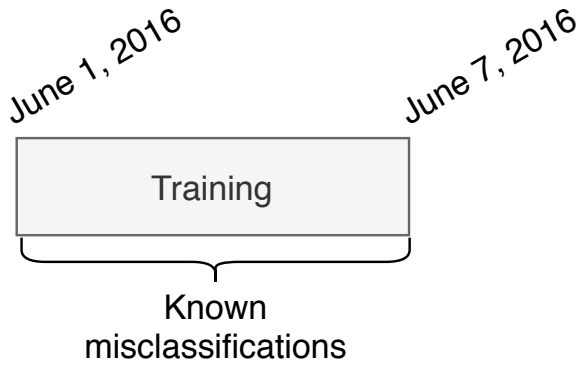
- Three types of ground truth, each with its corresponding legitimate and attack dataset.
- The legitimate portion is to validate the false detections of blacklists.
- The attack portion is to validate the accurate detections of blacklists.



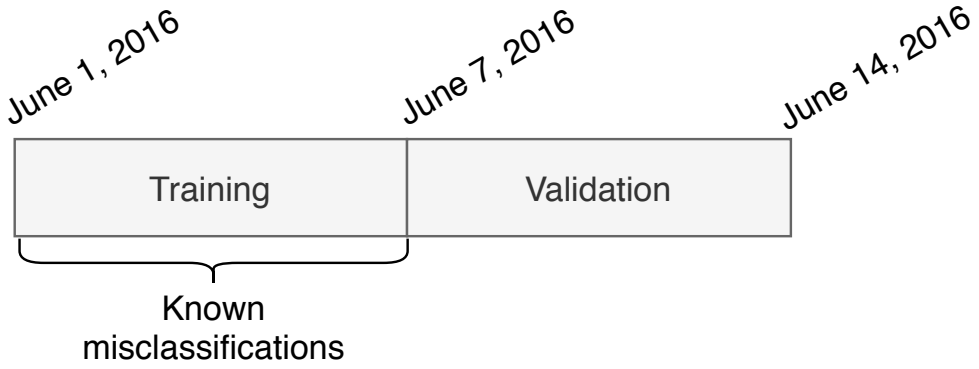
Email Dataset



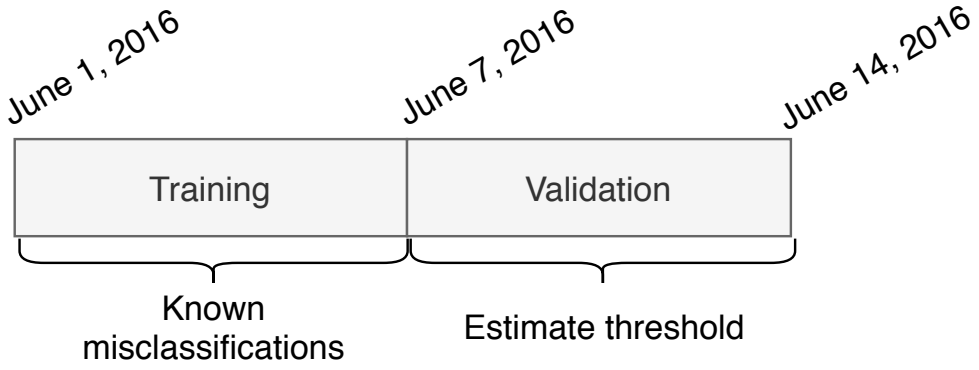
Email Dataset



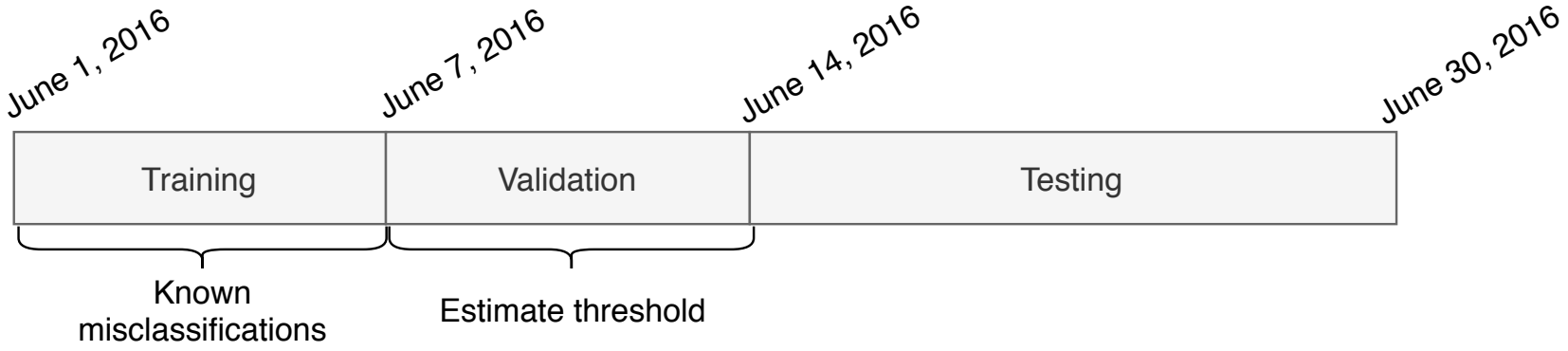
Email Dataset



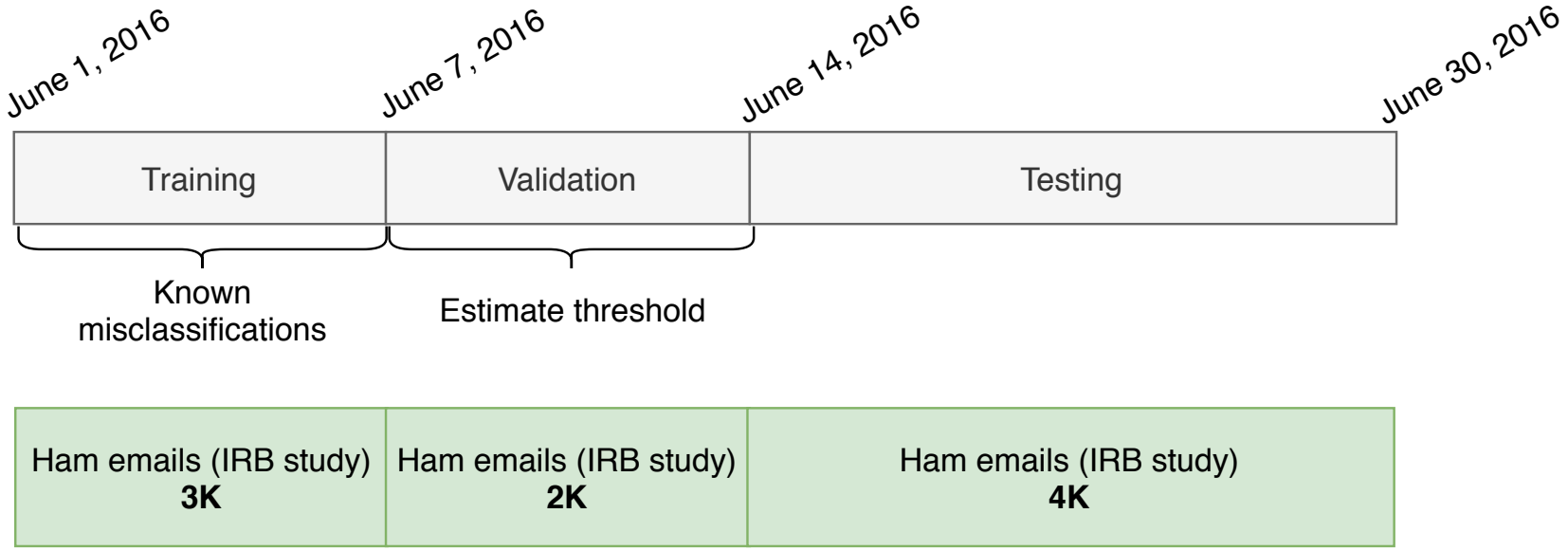
Email Dataset



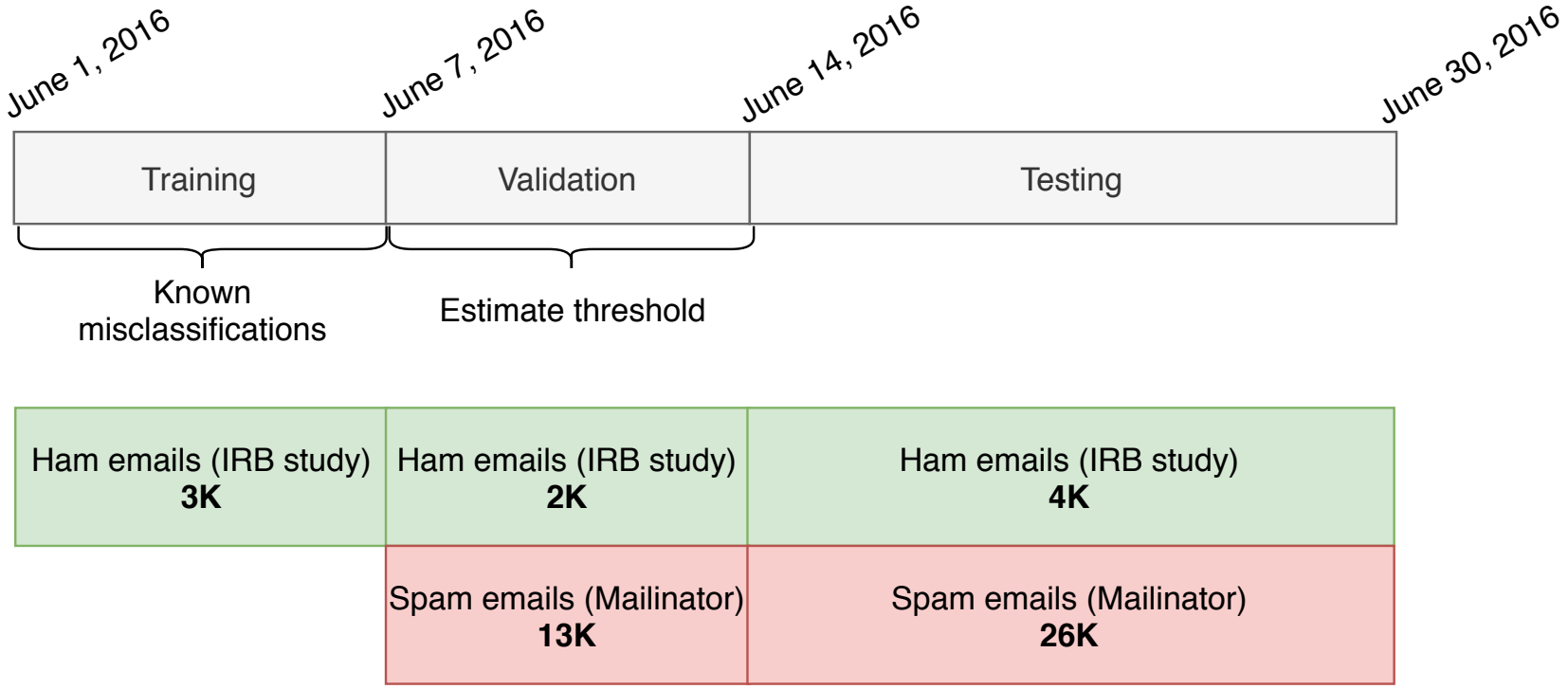
Email Dataset



Email Dataset



Email Dataset



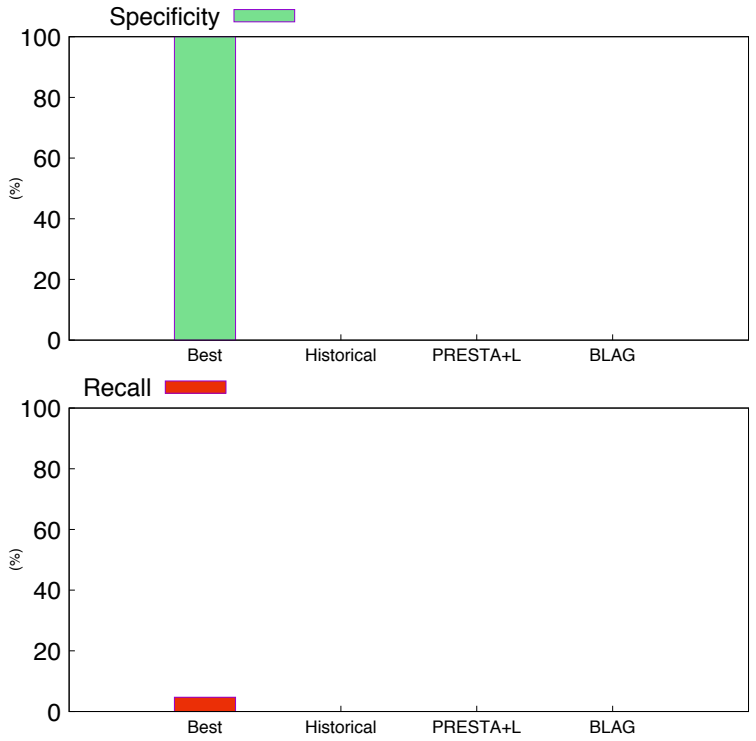
Outline

- Introduction
- Quantifying problems faced by blacklists
- BLAG
- Datasets
- Evaluation
- Summary

Evaluation

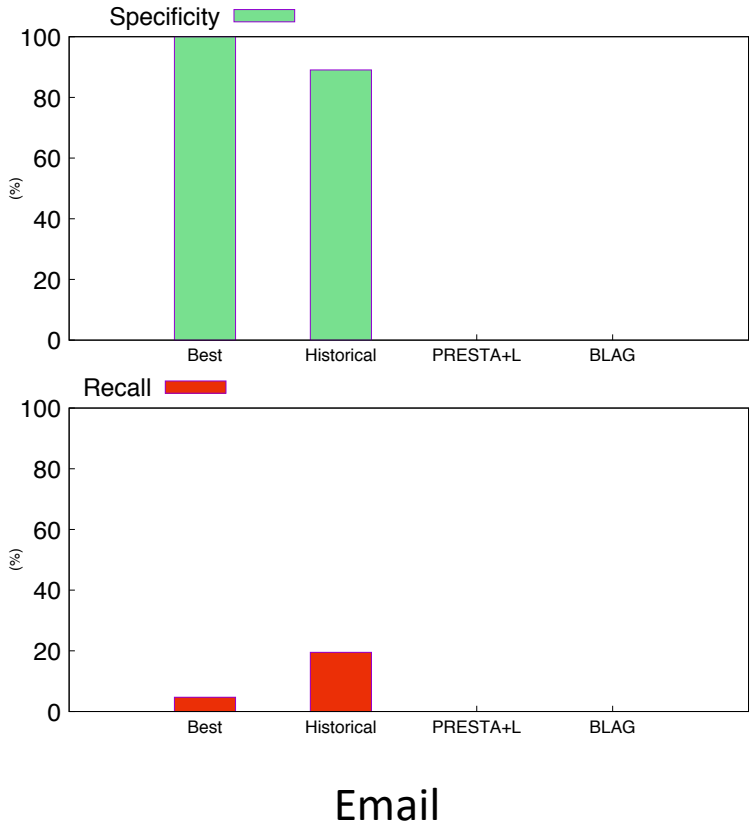
- Accuracy of BLAG: Compare the performance of BLAG with competing approaches
  - Best: The best-performing blacklist on a given ground truth dataset (hindsight) at the given time (of the ground truth dataset).
  - Historical: All addresses listed in all blacklists up until ground truth dataset.
  - PRESTA+L: Blacklisting approach taken by PRESTA algorithm that uses spatial properties of blacklisted addresses to generate a new blacklist.
- Metrics:
  - Specificity - the percentage of legitimate addresses that were not false positives.
  - Recall - the percentage of offenders that were detected.

BLAG is Accurate



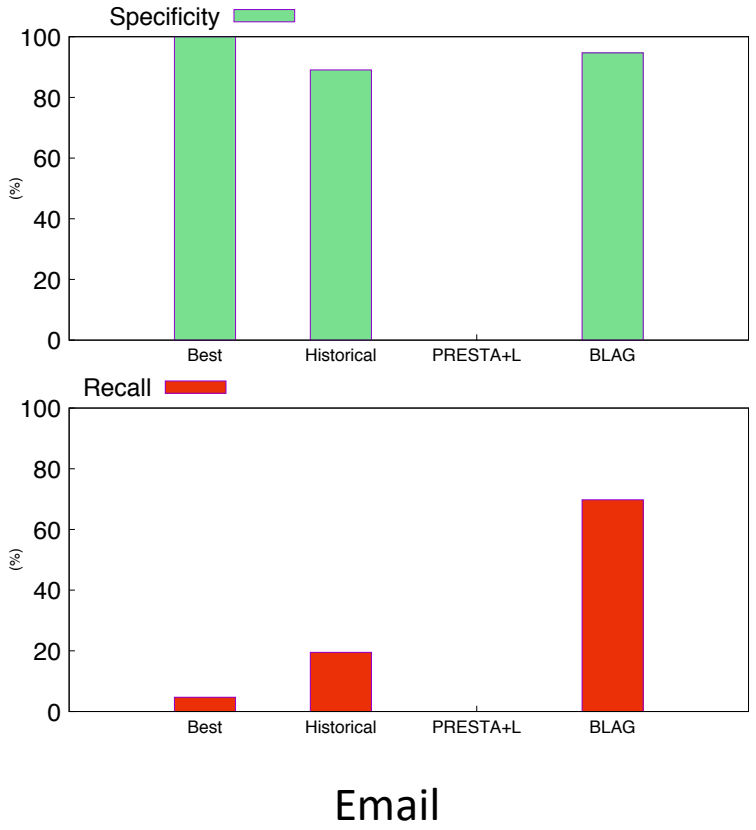
Best blacklists have high specificity (>99%) but poor recall(< 4%) indicating that even the best blacklist is not enough to capture all attackers.

BLAG is Accurate



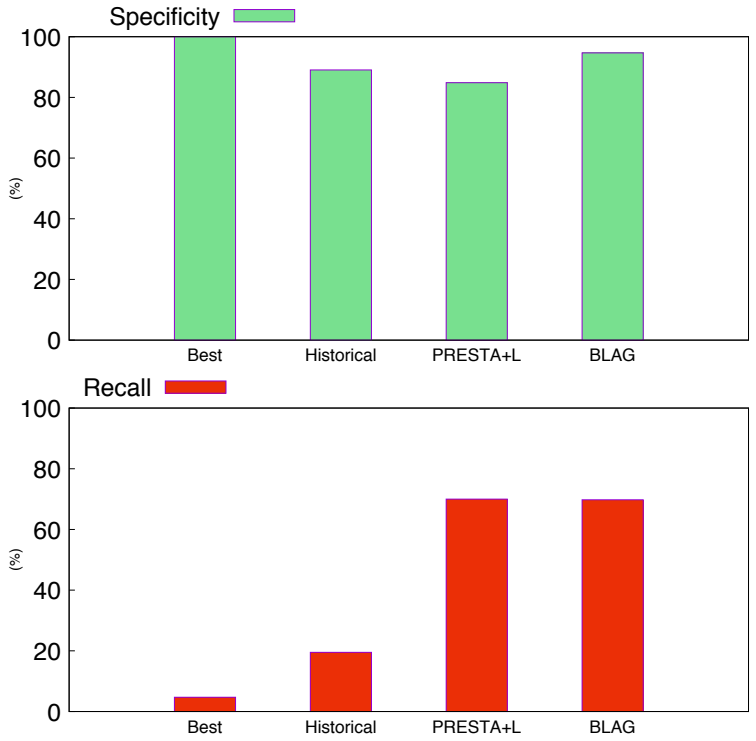
Historical blacklists improve recall to 18% but with a drop in specificity by 12%, indicating that naïve combination of all blacklists has potential to capture attackers, but lowers specificity.

BLAG is Accurate



BLAG with expansion further improves recall, with only a slight drop in specificity and has better specificity than historical blacklists.

BLAG is Accurate



Email

PRESTA+L has been tuned to have same recall as BLAG, but the specificity is lower than BLAG (82% vs 95%).

Other evaluations

- Evaluated BLAG on two other datasets: $DDoS_{Univ}$ and $DDoS_{DNS}$.
- Other expansion techniques -- expand using BGP prefixes or by autonomous systems.
- Impact of
  - Number of blacklists
  - Size of misclassification blacklists
- Contribution of recommendation system in aggregation and expansion phase.
- Parameter tuning techniques.

Public datasets

- All monitored blacklists are available at:

https://steel.isi.edu/Projects/BLAG/

- Includes scripts to deploy BLAG in your network.

Outline

- Introduction
- Quantifying problems faced by blacklists
- BLAG
- Datasets
- Evaluation
- Summary

Summary

- Blacklists have poor attack detection.
- Combining blacklists from different sources improves attack detection, but also increases misclassifications.
- BLAG (Blacklist aggregator)
  - Assigns relevance scores to addresses belonging to blacklists.
  - Predicts addresses that are likely to be misclassifications using a recommendation system.
  - Expands selective addresses into prefixes for better attack detection.
- BLAG has better performance than competing approaches such as PRESTA.

Thank You! Questions?

All monitored blacklists are available at:

https://steel.isi.edu/members/sivaram/BLAG/

Contact: Sivaram Ramanathan

satyaman@usc.edu

USC University of Southern California