Data-Fusion for Prefix-Level Inference: A DDoS Case Study

Chris Misa[1.]

Department of Computer Science,
University of Oregon

[1.] Joint work with Ramakrishnan Durairajan, Arpit Gupta, Reza Rejaie, and Walter Willinger.

Let's use AI/ML to classify DDoS attack flows…

- ***Why?*** Recent work shows really good performance.
  - LUCID (TNSM '20) gets 99.7% accuracy on CIC '17 dataset.
- ***The problem?*** In reality we don't have enough resources to monitor all flows.
  - CAIDA traces have >100k (benign) flows / sec.[1]
  - LUCID requires ~880 GB / sec. of state *just for benign traffic*.
- ***The approach?*** Prefix-level classification.
  - Potential 100X reduction in monitoring resources.

1. https://www.caida.org/catalog/datasets/trace_stats/

...but what about the data?

- Need a dataset that captures *(i)* prefix-level blending of attack and benign classes and *(ii)* multiple, independent attack scenarios.

Dataset	# Benign				# Attack			
	/8	/16	/24	/32	/8	/16	/24	/32
CAIDA ('07) [2]	0	0	0	0	117	4 k	8.7 k	9 k
ISCX ('12) [15]	123	1590	2041	2129	6	6	9	14
Booters ('15) [13]	0	0	0	0	42	961	3 k	4.4 k
Mirai ('16) [6]	0	0	0	0	162	3.5 k	9.8 k	10 k
CIC ('17) [14]	156	922	2125	3432	1	1	1	1
CSECIC ('18) [3]	1	1	6	446	2	4	10	10
MAWILab ('19) [10]	211	30 k	3.3 m	5.3 m	0	0	0	0
CAIDA ('19) [1]	250	27 k	323 k	1.3 m	0	0	0	0

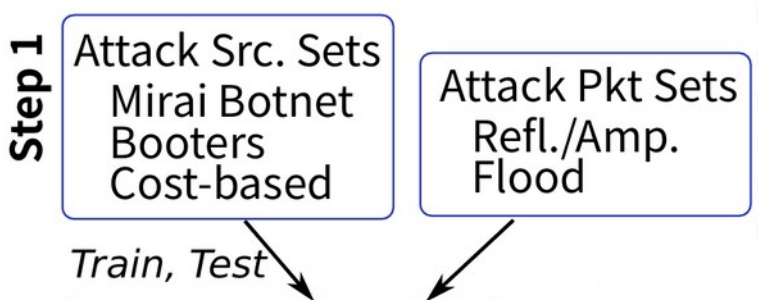
...but what about the data?

- Need a dataset that captures *(i)* prefix-level blending of attack and benign classes and *(ii)* multiple, independent attack scenarios.

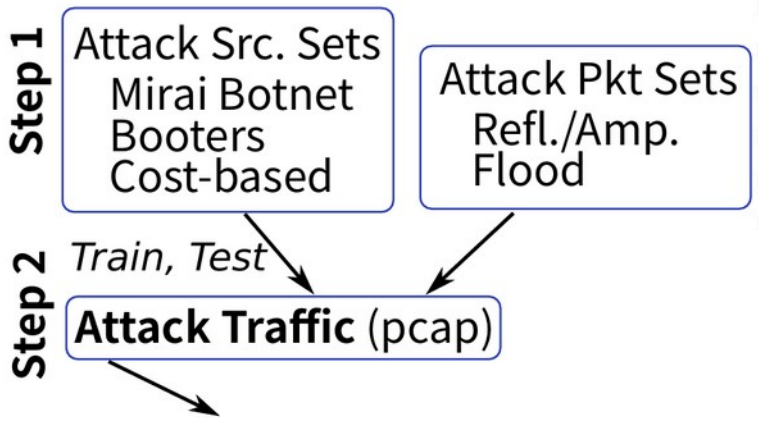
Dataset	# Benign				# Attack			
	/8	/16	/24	/32	/8	/16	/24	/32
CAIDA ('07) [2]	0	0	0	0	117	4 k	8.7 k	9 k
ISCX ('12) [15]	123	1590	2041	2129	6	6	9	14
Booters ('15) [13]	0	0	0	0	42	961	3 k	4.4 k
Mirai ('16) [6]	0	0	0	0	162	3.5 k	9.8 k	10 k
CIC ('17) [14]	156	922	2125	3432	1	1	1	1
CSECIC ('18) [3]	1	1	6	446	2	4	10	10
MAWILab ('19) [10]	211	30 k	3.3 m	5.3 m	0	0	0	0
CAIDA ('19) [1]	250	27 k	323 k	1.3 m	0	0	0	0

Only one attacker!

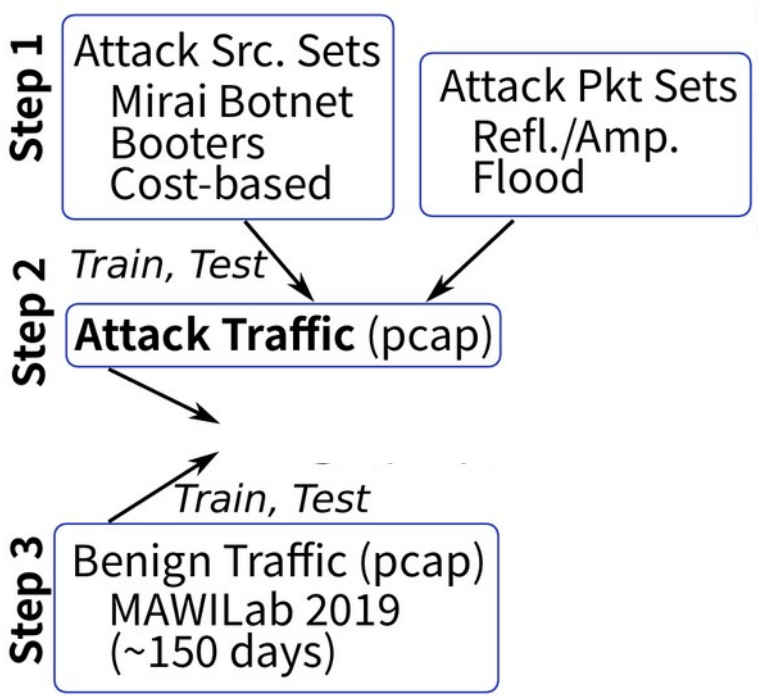
Data-fusion as a practical compromise.



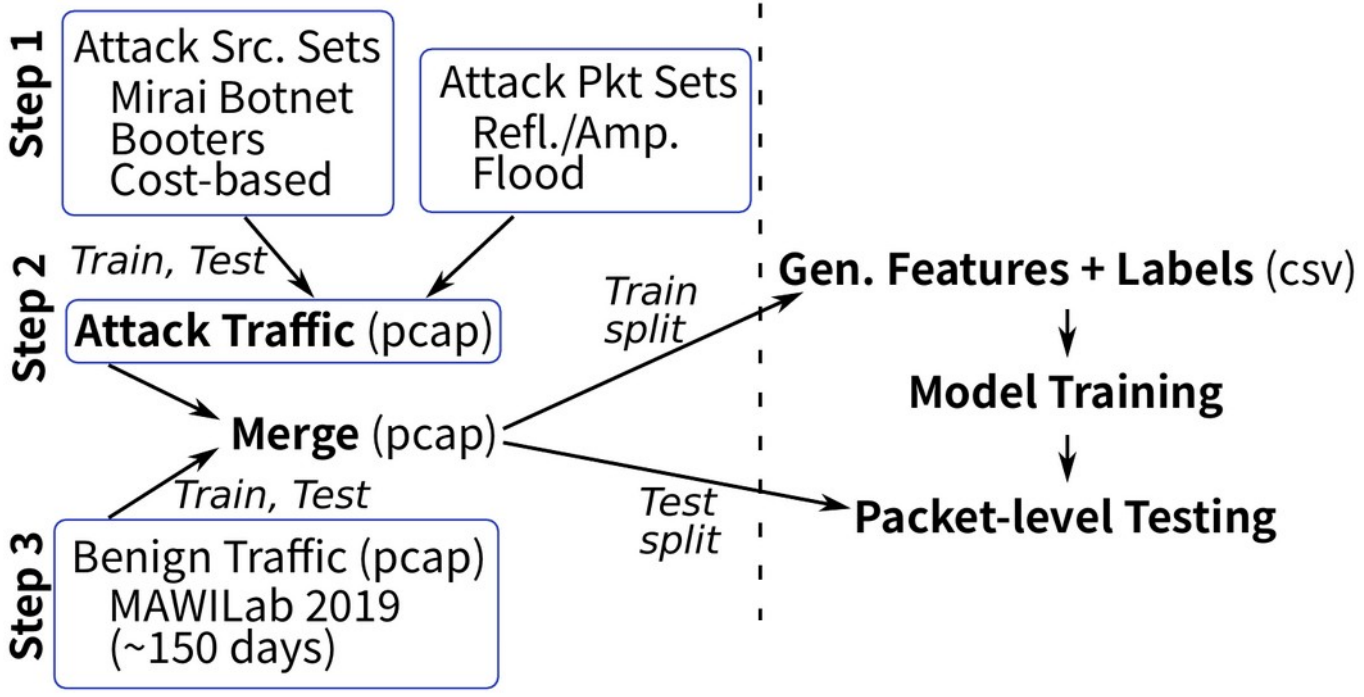
Data-fusion as a practical compromise.



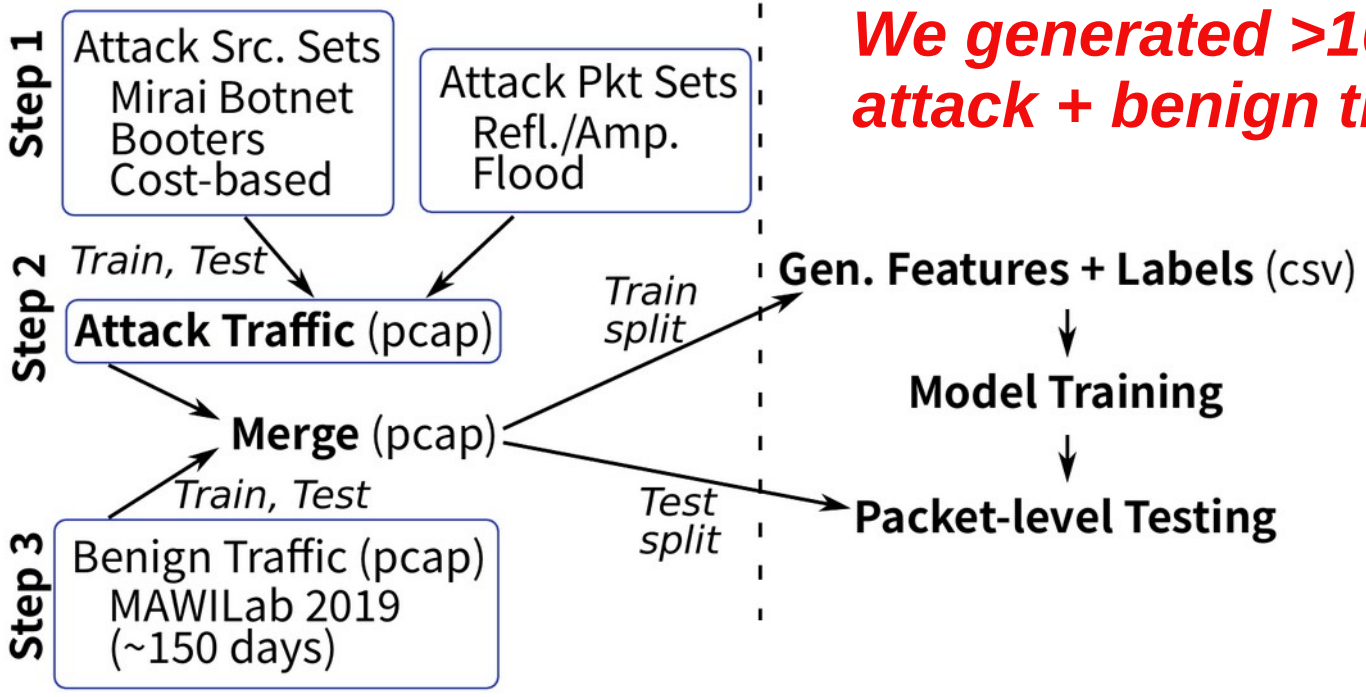
Data-fusion as a practical compromise.



Data-fusion as a practical compromise.



Data-fusion as a practical compromise.



We generated >100 distinct attack + benign traffic scenarios.

- See ZAPDOS[1] (to appear in S&P '24) for more details.

1. Pre-print available here: https://onrg.gitlab.io/pub/SnP2024_ZAPDOS_FinalWeb.pdf

(...but what about the data?)

- Need a dataset that captures *(i)* prefix-level blending of attack and benign classes and *(ii)* multiple, independent attack scenarios.

Dataset	# Benign				# Attack			
	/8	/16	/24	/32	/8	/16	/24	/32
CAIDA ('07) [2]	0	0	0	0	117	4 k	8.7 k	9 k
ISCX ('12) [15]	123	1590	2041	2129	6	6	9	14
Booters ('15) [13]	0	0	0	0	42	961	3 k	4.4 k
Mirai ('16) [6]	0	0	0	0	162	3.5 k	9.8 k	10 k
CIC ('17) [14]	156	922	2125	3432	1	1	1	1
CSECIC ('18) [3]	1	1	6	446	2	4	10	10
MAWILab ('19) [10]	211	30 k	3.3 m	5.3 m	0	0	0	0
CAIDA ('19) [1]	250	27 k	323 k	1.3 m	0	0	0	0
Proposed "data-fusion" method	216	30 k	3.2 m	4.8 m	179	7 k	45 k	50 k

50k attackers per scenario.

Can we do better than data-fusion?

- ***Data providers*** (e.g., CAIDA, CLASSNET) should consider requirements of prefix-level approaches.
  - Use prefix-preserving anonymization (cryptopan).
  - Develop combined attack + benign datasets to capture “blending”.

- ***Researchers*** should consider improving synthetic traffic generation.
  - State-of-the-art proposals (e.g., NetShare) *do not* reproduce realistic “spatial” structure.
  - General lack of metrics for measuring prefix-level fidelity of synth. data.

If we have realistic prefix-level data, what next?

- *Can prefix-level classification be applied to other security monitoring tasks?*

- *Can we leverage where (in IP space) attack traffic comes from to develop more informative source- or flow-level features?*

- *How can we quantify the risk involved in prefix-level approaches? Do they expose opportunities for adversaries to manipulate models or data?*

Thanks!

cmisa@cs.uoregon.edu