SIMBIoTA: Similarity-Based Malware Detection on IoT Devices

Csongor Tamás

csongor.tamas@ukatemi.com

Ukatemi Technologies

Dorottya Papp and Levente Buttyán

{dpapp, buttyan}@crysys.hu

Laboratory of Cryptography and System Security (CrySyS Lab)

Department of Networked Systems and Services

Budapest University of Technology and Economics

Background and Motivation

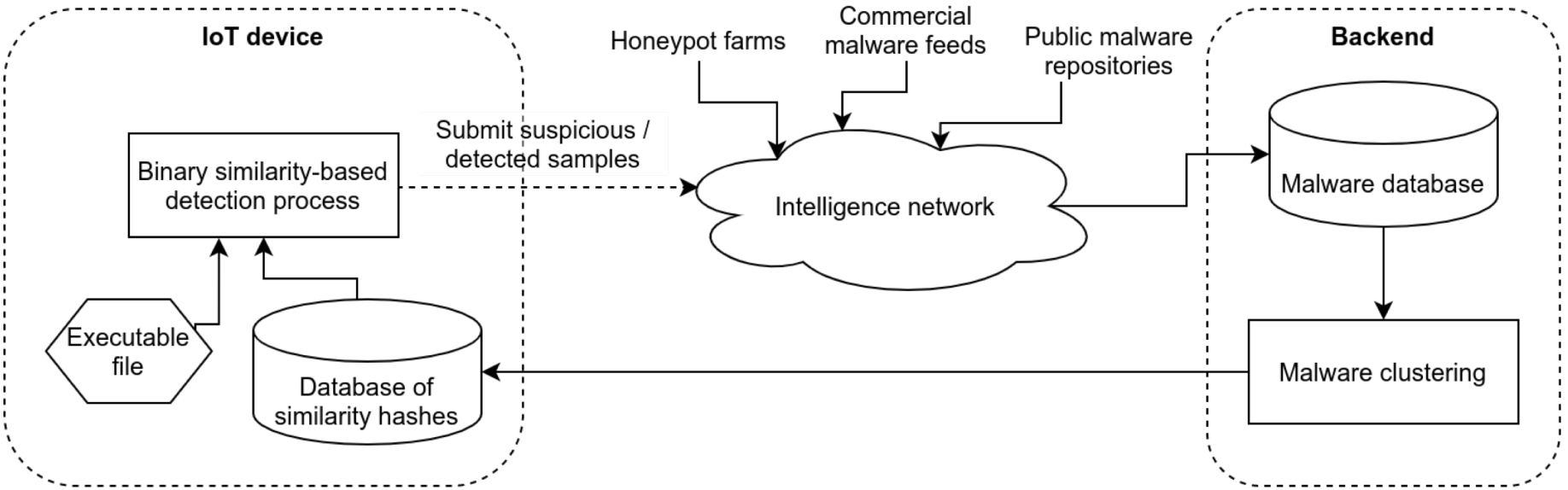
- **Increasing amount of malware targeting IoT** devices in the past 5 years
- Targets:
  - Industrial Control Systems (high-value)
  - Generic IoT devices (low-value) -> botnets
- Currently **available antivirus** products either do not support IoT devices or **have too demanding system requirements**
- **Resource constraints**:
  - Low computing power
  - Running on battery
  - Small storage space
  - Low network bandwidth

Existing detection approaches

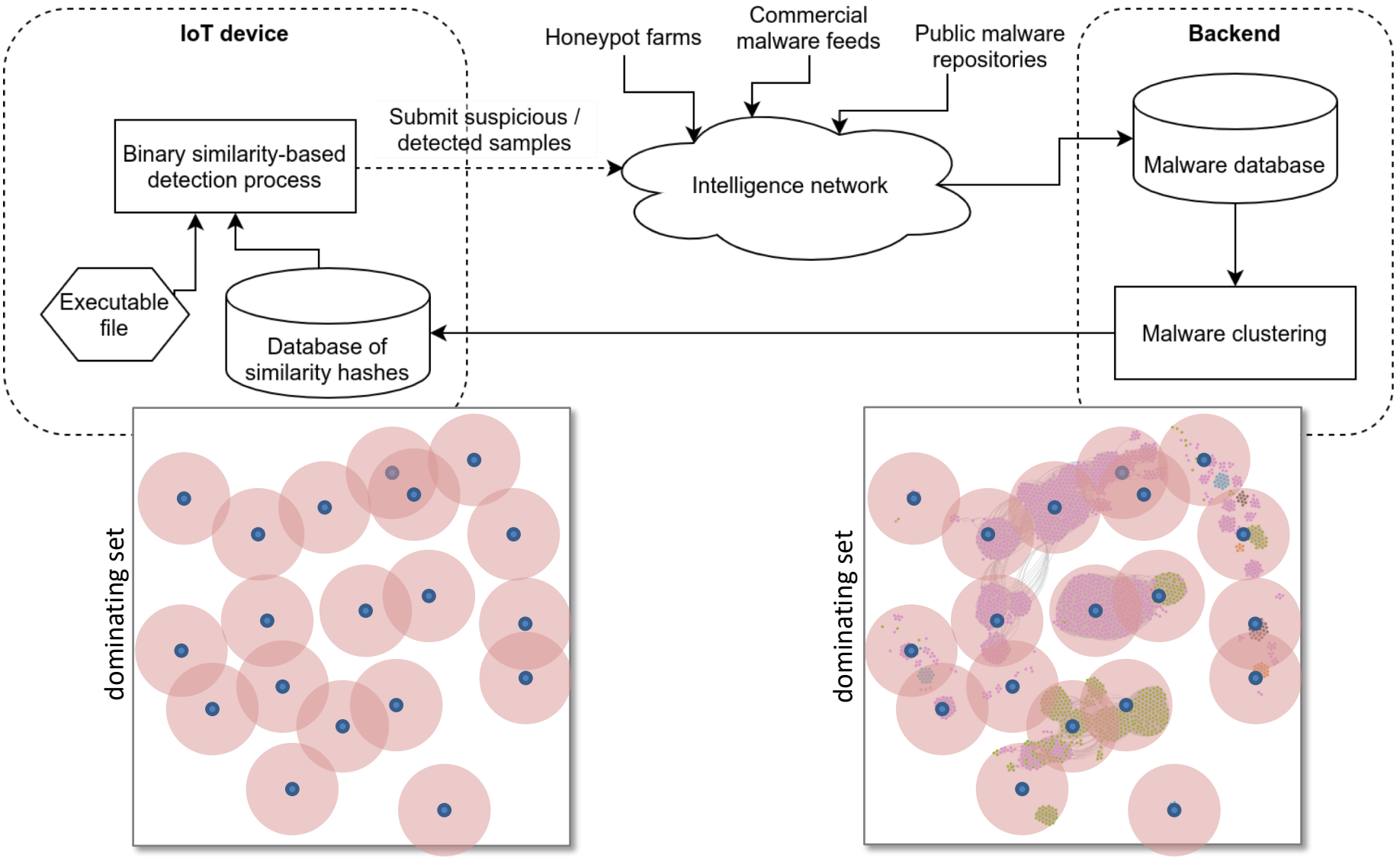
- Resource constraints -> **static detection** on client
- Existing detection approaches
  - Client-based
    - » Signature-based
      - byte-matching rules identifying known malware
    - » Heuristic
      - malicious characteristics associated with malware
    - » Pre-trained machine learning model-based
      - Either featureless or with selected features model identifying malware
      - Pre-trained models are applied on clients
  - Cloud-based
    - » Samples are sent to the cloud for analysis
    - » Needs constant network availability

SIMBIoTA approach

- Signature-based with similarity hashes (TLSH)
  - Reduced storage and network bandwidth requirements
  - Fully automatic operation
  - SIMilarity Based IoT Antivirus

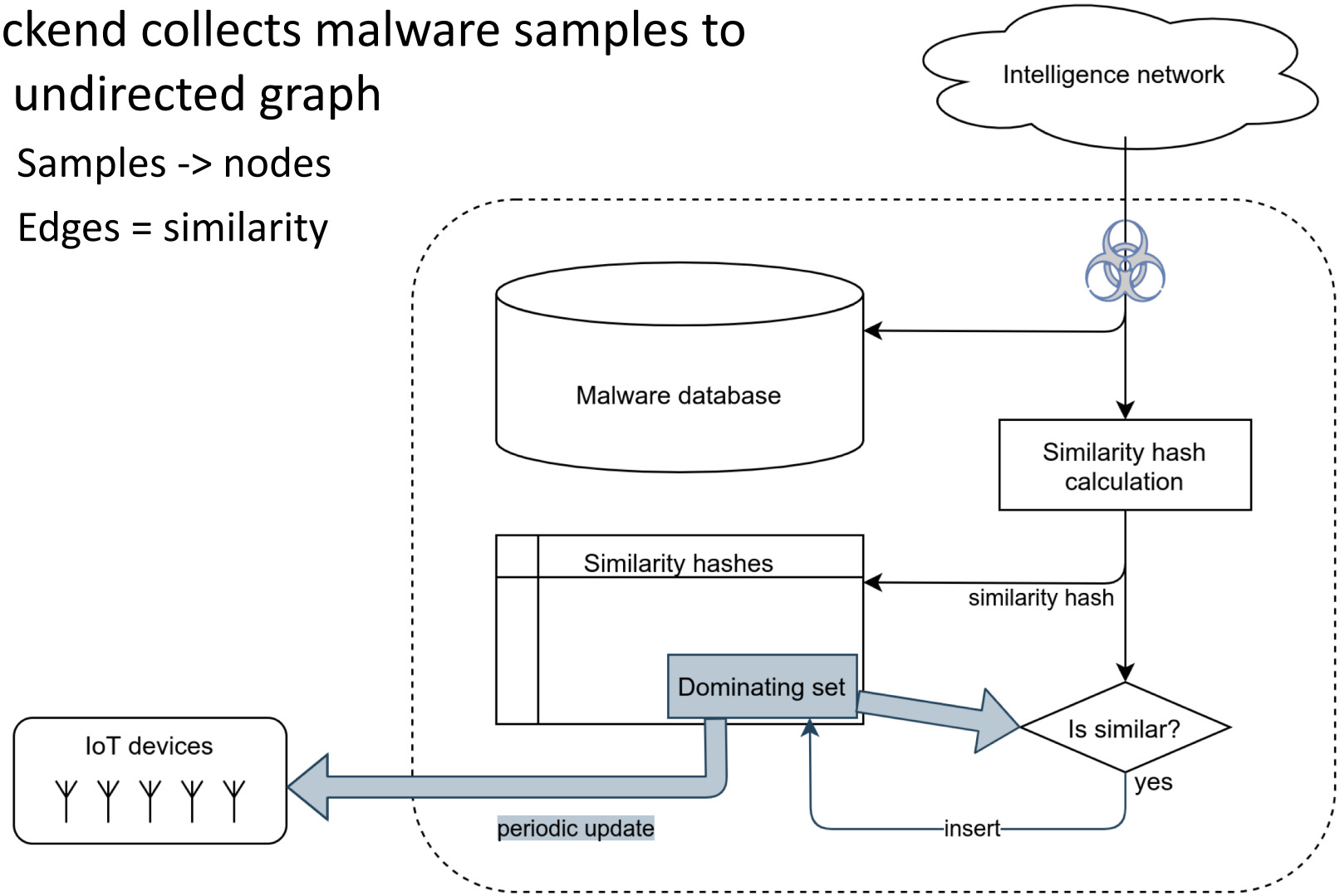


SIMBIoTA approach

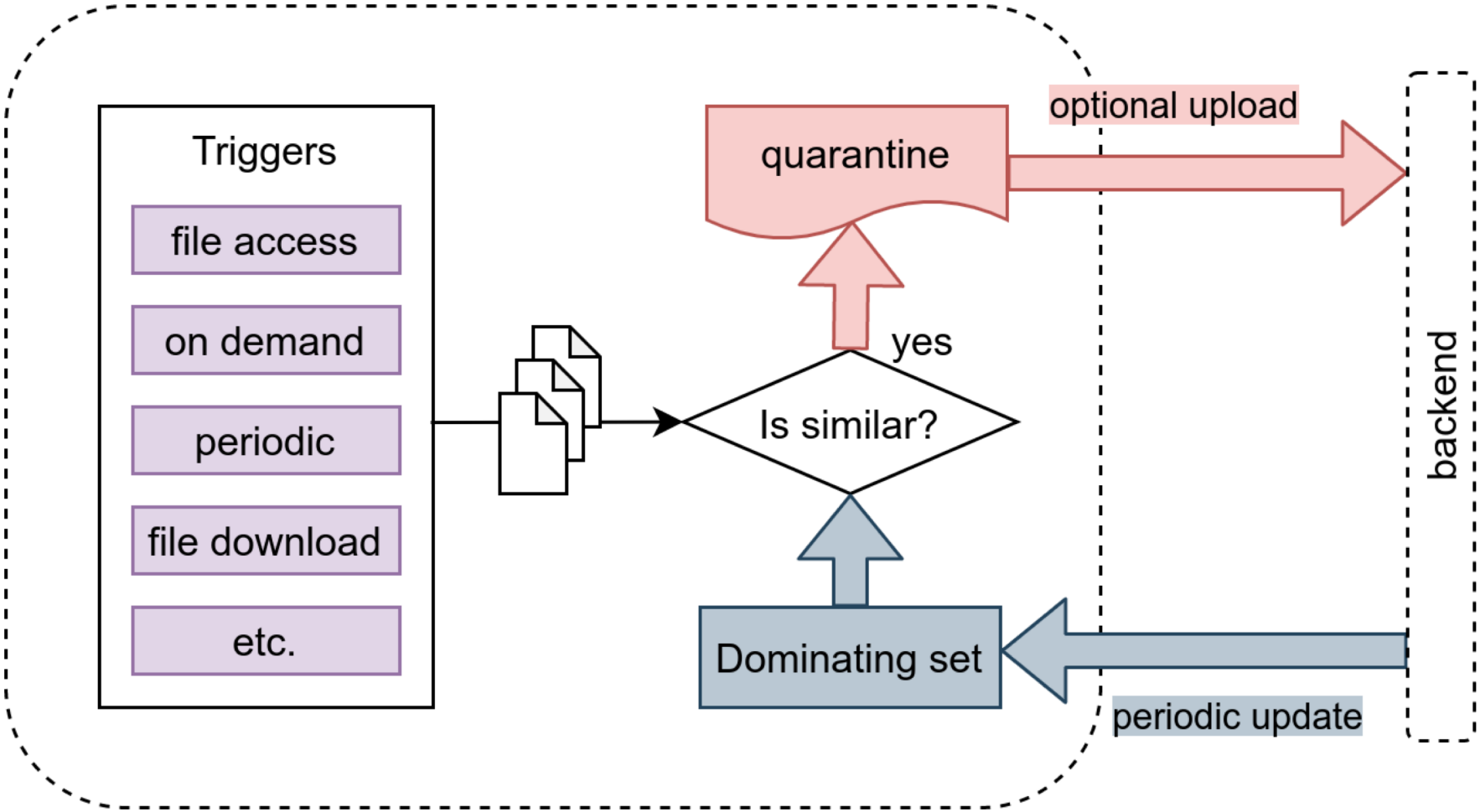


Backend processes

- Backend collects malware samples to an undirected graph
  - Samples -> nodes
  - Edges = similarity



Client detection

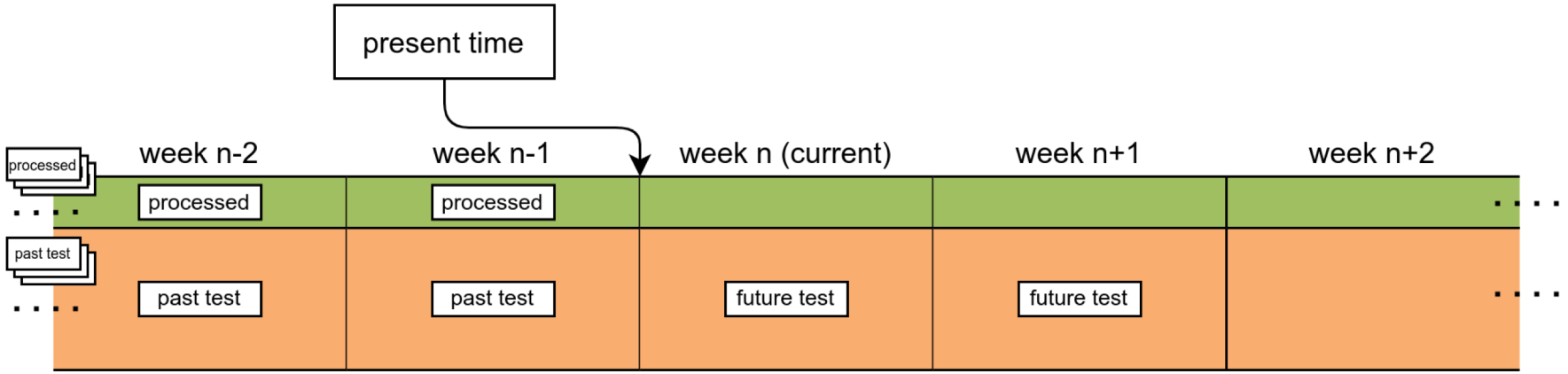


Evaluation

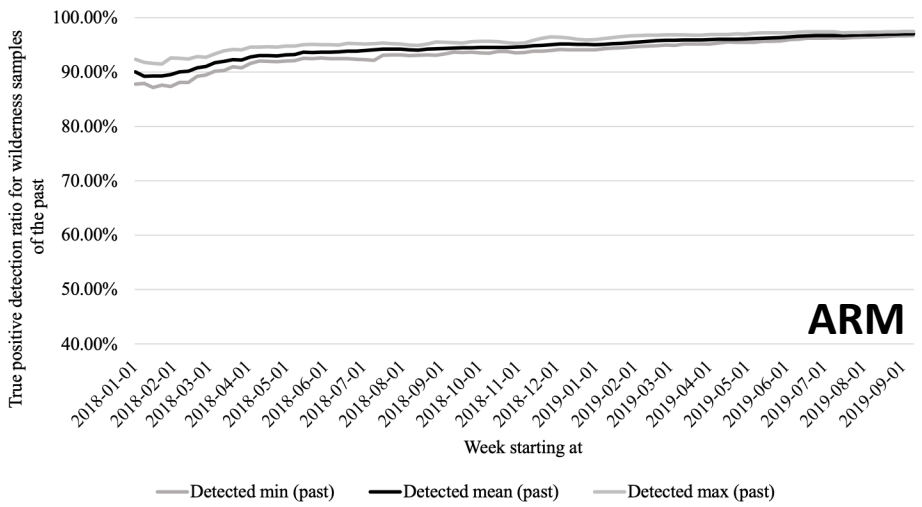
- Dataset
  - Malicious
    - » From Ukatemi Technologies' malware repository (400 million samples)
    - » 29 215 ARM + 18 722 MIPS = 47 937 samples
    - » From 2018-01-01 to 2019-09-15 (VirusTotal `first_submission_date`)
  - Benign
    - » Firmware images from D-Link and Ubiquiti
    - » 14 119 files extracted
- Experiment setup
  - Malicious samples are grouped into weekly batches
  - Every week is split into
    - » 10% intelligence samples (received by the backend)
    - » 90% wilderness samples (never seen by the backend, validation set)

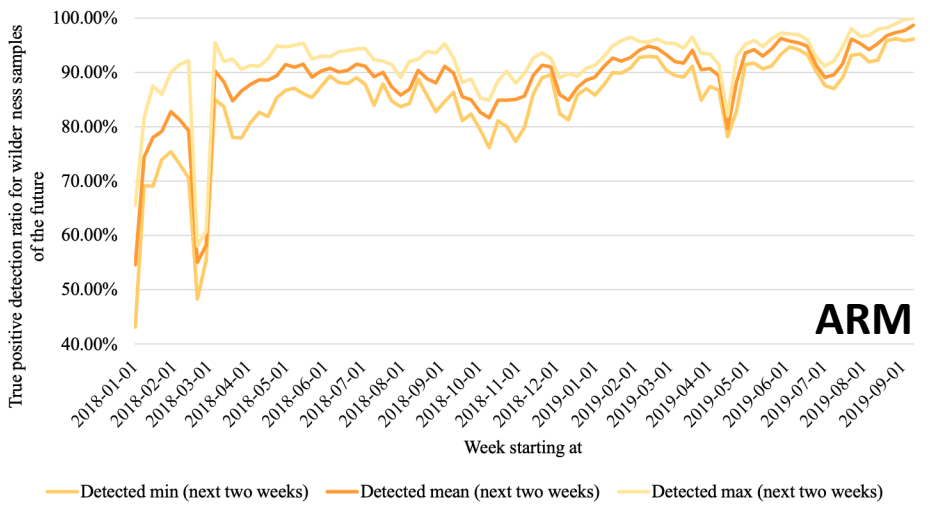
Evaluation

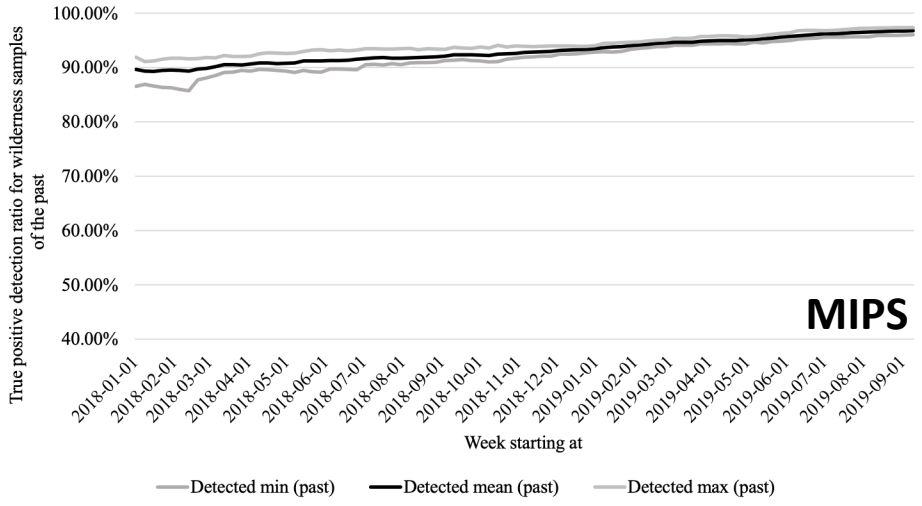
- Measurement
  - At the beginning of every week, the backend receives the intelligence samples from the previous week. These are incorporated to the dominating set.
  - Detection performance is measured on EVERY past wilderness sample and EVERY wilderness sample from the current week and the following week (2 week future).

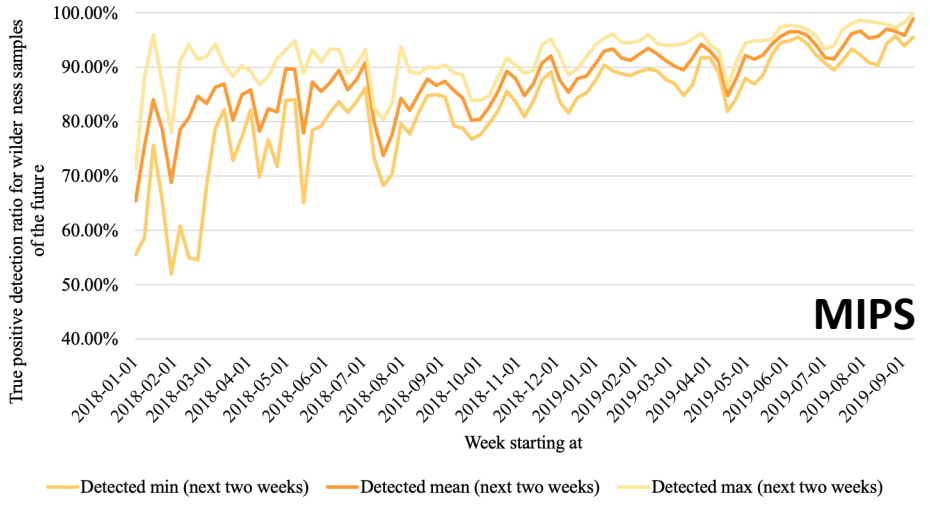


Results



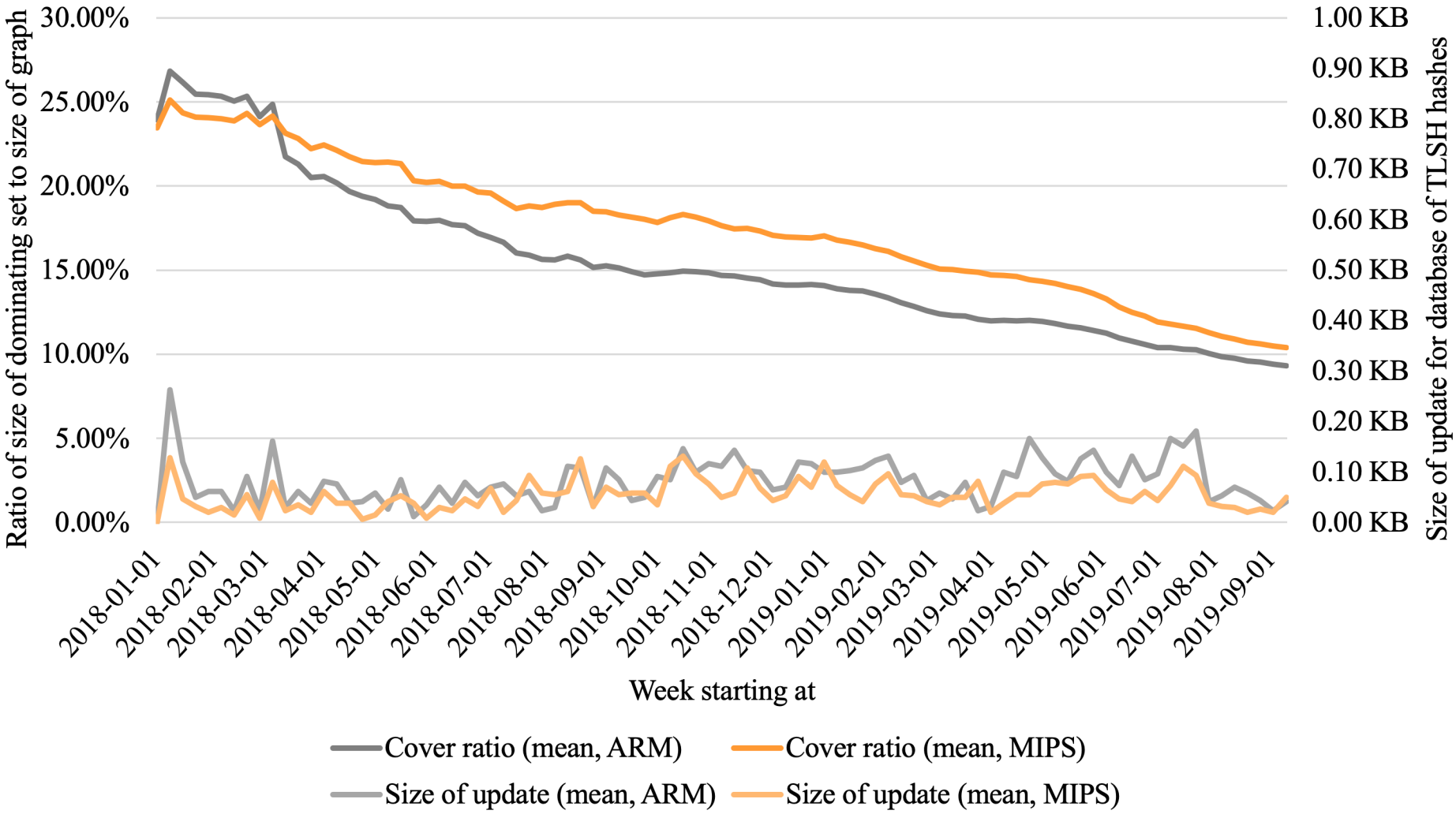






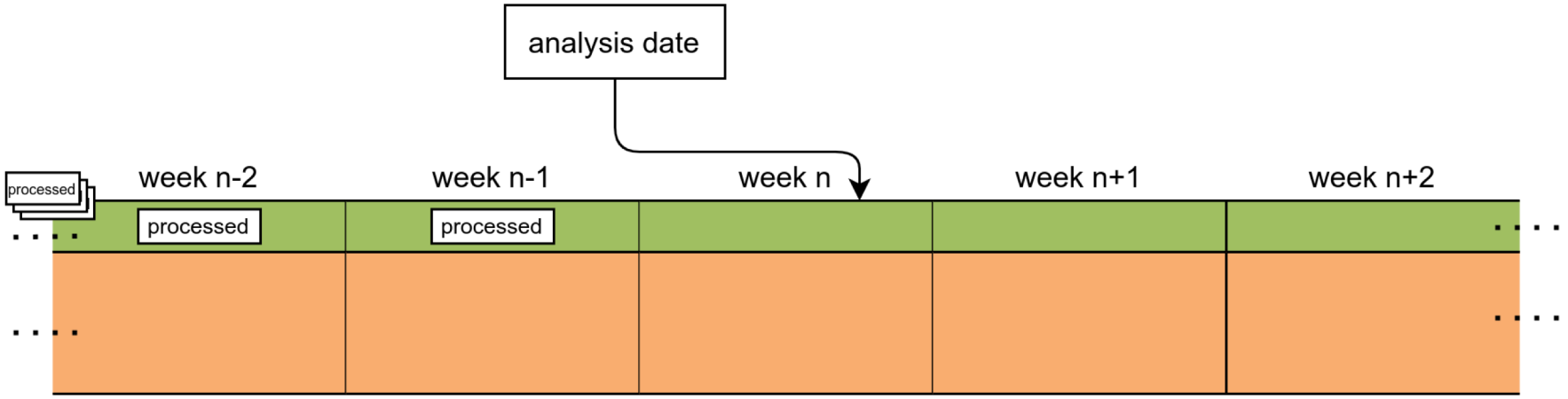
PAST **FUTURE**

Update size


Ratio of size of dominating set to size of graph
30.00%
25.00%
20.00%
15.00%
10.00%
5.00%
0.00%
Size of update for database of TLSH hashes
1.00 KB
0.90 KB
0.80 KB
0.70 KB
0.60 KB
0.50 KB
0.40 KB
0.30 KB
0.20 KB
0.10 KB
0.00 KB
2018-01-01
2018-02-01
2018-03-01
2018-04-01
2018-05-01
2018-06-01
2018-07-01
2018-08-01
2018-09-01
2018-10-01
2018-11-01
2018-12-01
2019-01-01
2019-02-01
2019-03-01
2019-04-01
2019-05-01
2019-06-01
2019-07-01
2019-08-01
2019-09-01
Week starting at
Cover ratio (mean, ARM)
Cover ratio (mean, MIPS)
Size of update (mean, ARM)
Size of update (mean, MIPS)

Comparison with other AVs

- Download earliest possible analysis reports from VirusTotal
  - 60% of the samples – earliest = first submission
  - 40% of the samples – earliest = some later analysis



- 48 of 78 AVs detected at least one sample (30 detected none)
- For every AV only those samples were considered whose earliest analysis report contained detection results for the AV

Comparison with other AVs

	Number of detected samples by existing AV	SIMBIoTA	Difference
Product #1	24 362	23 114	-5,12%
Product #2	24 263	23 080	-4,88%
Product #3	24 052	22 899	-4,80%
Product #4	23 016	22 545	-2,04%
Product #5	22 866	22 464	-1,76%
Product #6	23 515	23 140	-1,59%
Product #7	22 861	22 579	-1,23%
Product #8	23 424	23 151	-1,17%
Product #9	21 411	22 257	+3,95%
Product #10	19 219	20 831	+8,39%
Product #11	20 759	23 145	+11,49%
Product #12	19 551	23 104	+18,17%
Product #13	18 847	23 118	+22,66%
Product #14	18 478	23 040	+24,69%
Product #15	17 512	22 443	+28,16%
Product #16	16 323	21 809	+33,61%
Product #17	16 052	21 928	+36,60%
Product #18	16 525	23 008	+39,23%
Product #19	15 924	23 014	+44,53%
Product #20	15 290	23 139	+51,33%
Product #21	15 149	23 073	+52,31%
Product #22	11 094	23 087	+108,10%
Product #23	5 096	10 683	+109,64%
Product #24	10 983	23 120	+110,51%
Product #25	10 681	23 094	+116,21%

ARM

	Number of detected samples by existing AV	SIMBIoTA	Difference
Product #1	16 022	15 003	-6,36%
Product #2	15 924	14 959	-6,06%
Product #3	15 661	14 883	-4,97%
Product #8	15 260	15 012	-1,62%
Product #4	14 681	14 634	-0,32%
Product #5	14 557	14 559	+0,01%
Product #7	14 397	14 647	+1,73%
Product #9	13 946	14 503	+3,99%
Product #6	14 254	15 007	+5,28%
Product #10	12 748	13 600	+6,68%
Product #14	13 117	14 942	+13,92%
Product #11	12 984	15 007	+15,58%
Product #15	12 147	14 527	+19,59%
Product #13	12 252	14 991	+22,36%
Product #12	12 005	14 988	+24,85%
Product #18	11 913	14 913	+25,19%
Product #19	11 258	14 911	+32,44%
Product #17	10 259	14 184	+38,26%
Product #16	10 163	14 139	+39,12%
Product #20	9 852	15 006	+52,31%
Product #21	7 751	14 952	+92,91%
Product #27	7 358	14 348	+95,00%
Product #22	7 251	14 981	+106,61%
Product #25	6 885	14 979	+117,56%
Product #26	6 756	14 845	+119,73%

MIPS

Conclusion

- Effective and efficient AV on resource-constrained IoT devices is possible
  - Similarity hashes
    - » Fast operation
    - » Small client database
    - » Can be exchanged to any selected method
  - Backend similarity preprocessing (dominating set creation)
    - » Fully automatic
    - » Client detects every sample observed on the backend
  - Results
    - » Above 90% true positive detection rate on steady operation for unknown samples