SEPT. 2021

ANDRUSPEX: LEVERAGING GRAPH REPRESENTATION LEARNING TO PREDICT HARMFUL APP (PHA) INSTALLATIONS ON MOBILE DEVICES

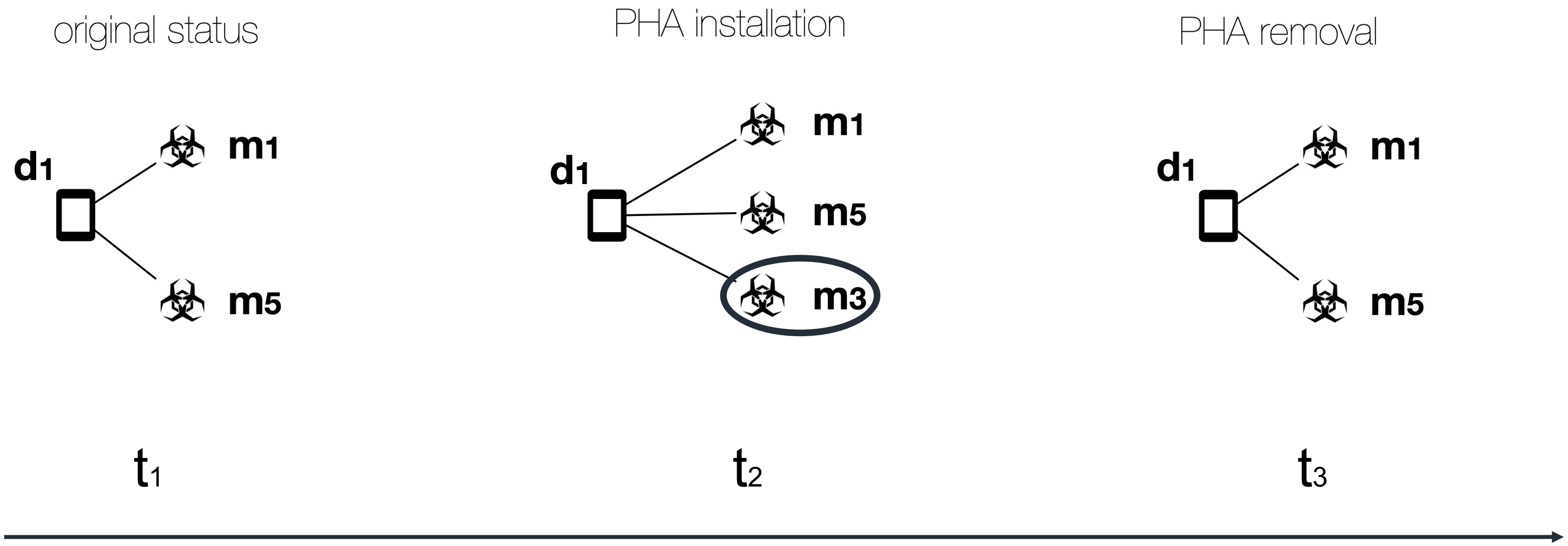
YUN SHEN

GIANLUCA STRINGHINI (BOSTON UNIVERSITY)

Overview

- Motivation
- Technical Details
- Results
- Limitations

Motivation



original status

d_1 m_1 m_5

t_1

PHA installation

d_1 m_1 m_5 m_3

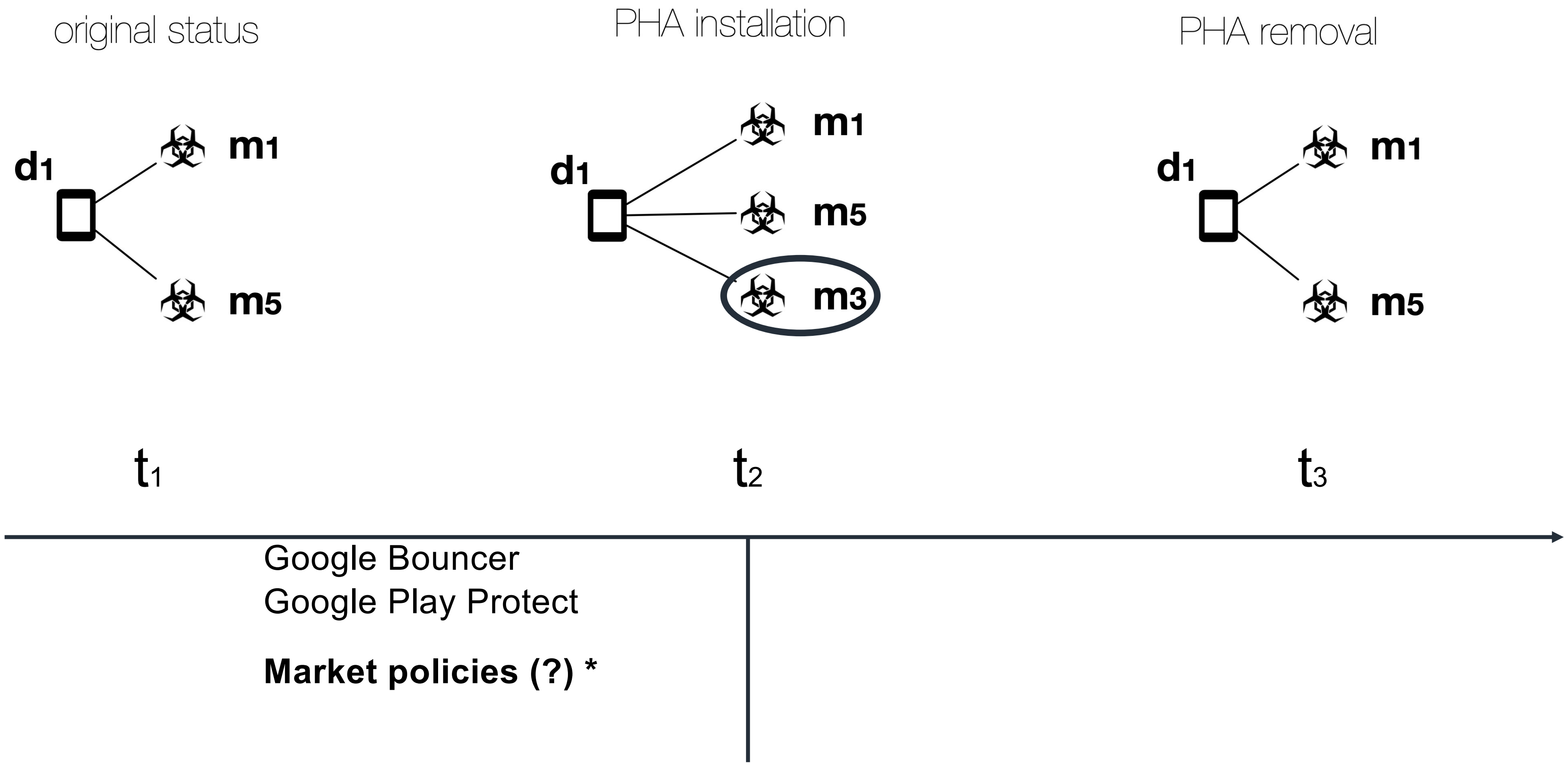
t_2

PHA removal

d_1 m_1 m_5

t_3

Motivation


original status

PHA installation

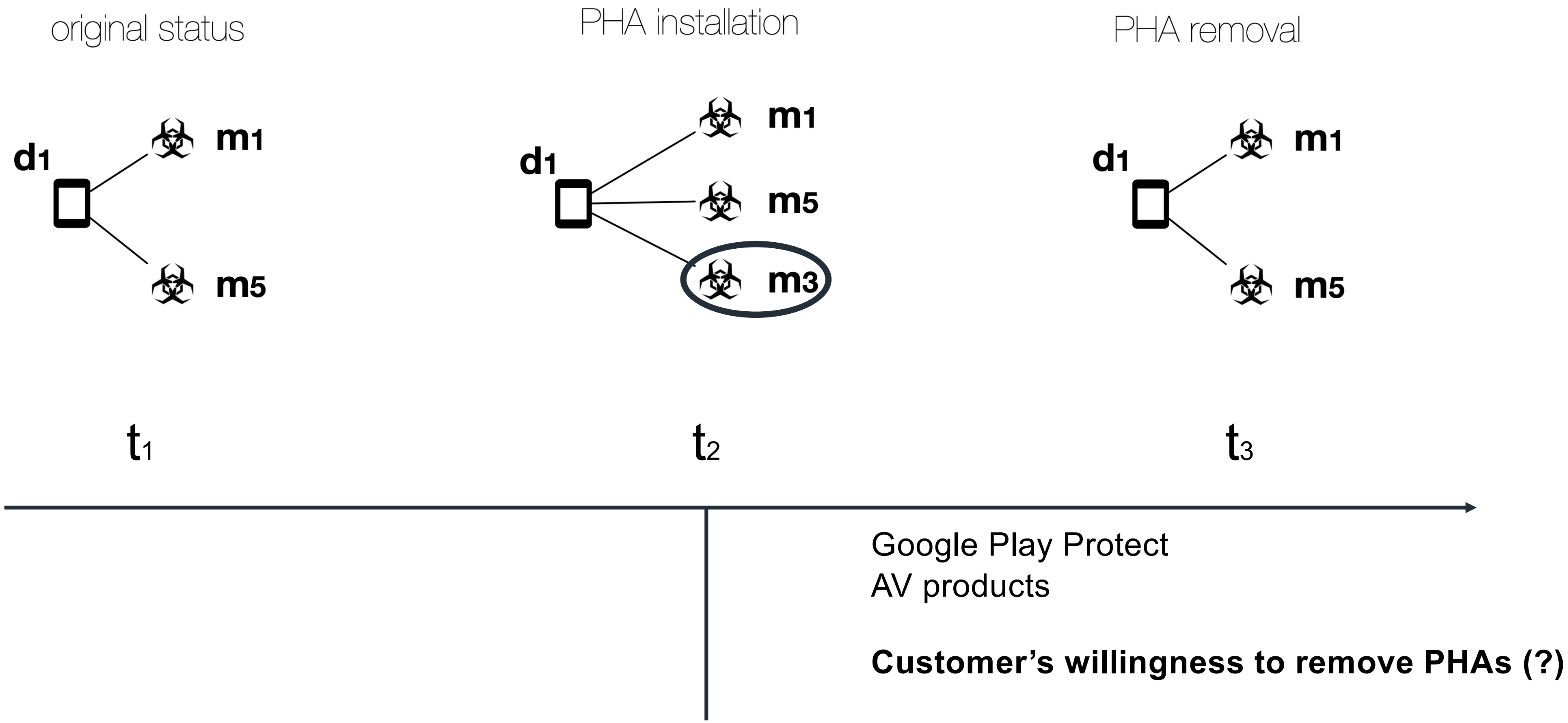
PHA removal

Google Bouncer
Google Play Protect

Market policies (?) *

* Kotzias, Platon, Juan Caballero, and Leyla Bilge. "How did that get in my phone? unwanted app distribution on android devices IEEE S&P, 2021

Motivation



original status

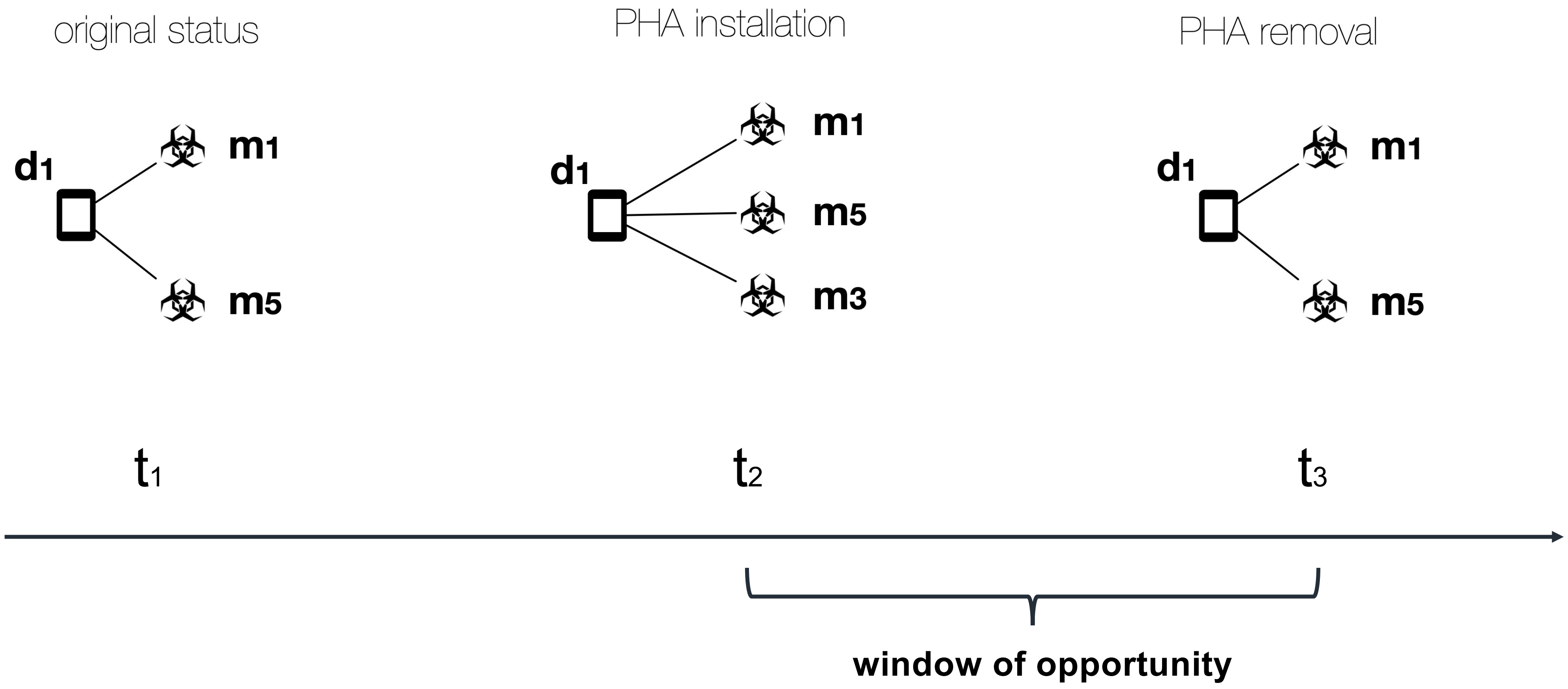
PHA installation

PHA removal

Google Play Protect
AV products

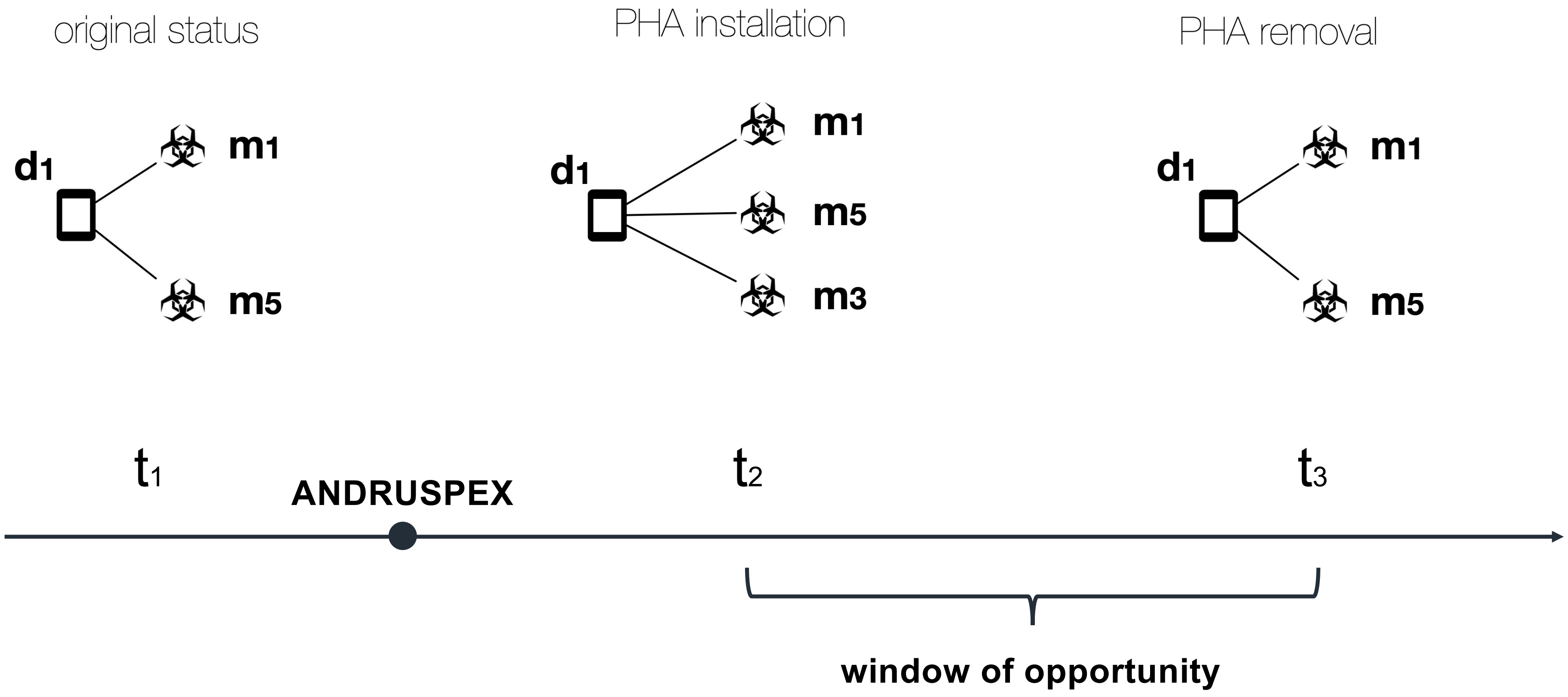
Customer’s willingness to remove PHAs (?)

Motivation



original status
PHA installation
PHA removal
d1
m1
m5
d1
m1
m5
m3
d1
m1
m5
t1
t2
t3
window of opportunity

Motivation



original status

d_1 m_1 m_5

PHA installation

d_1 m_1 m_5 m_3

PHA removal

d_1 m_1 m_5

t_1

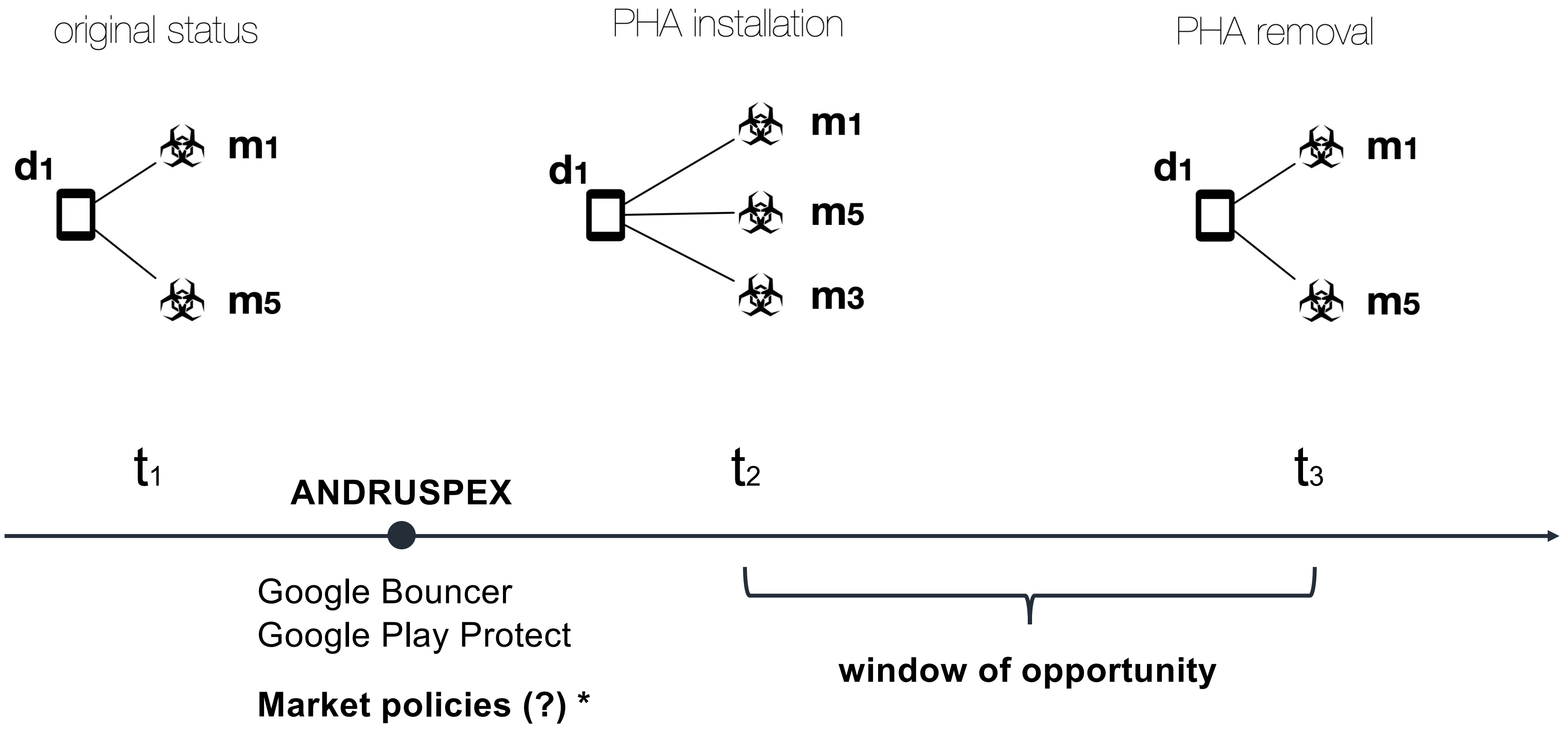
ANDRUSPEX

t_2

t_3

window of opportunity

Motivation



original status

d1 m1 m5

PHA installation

d1 m1 m5 m3

PHA removal

d1 m1 m5

t_1

ANDRUSPEX

t_2

t_3

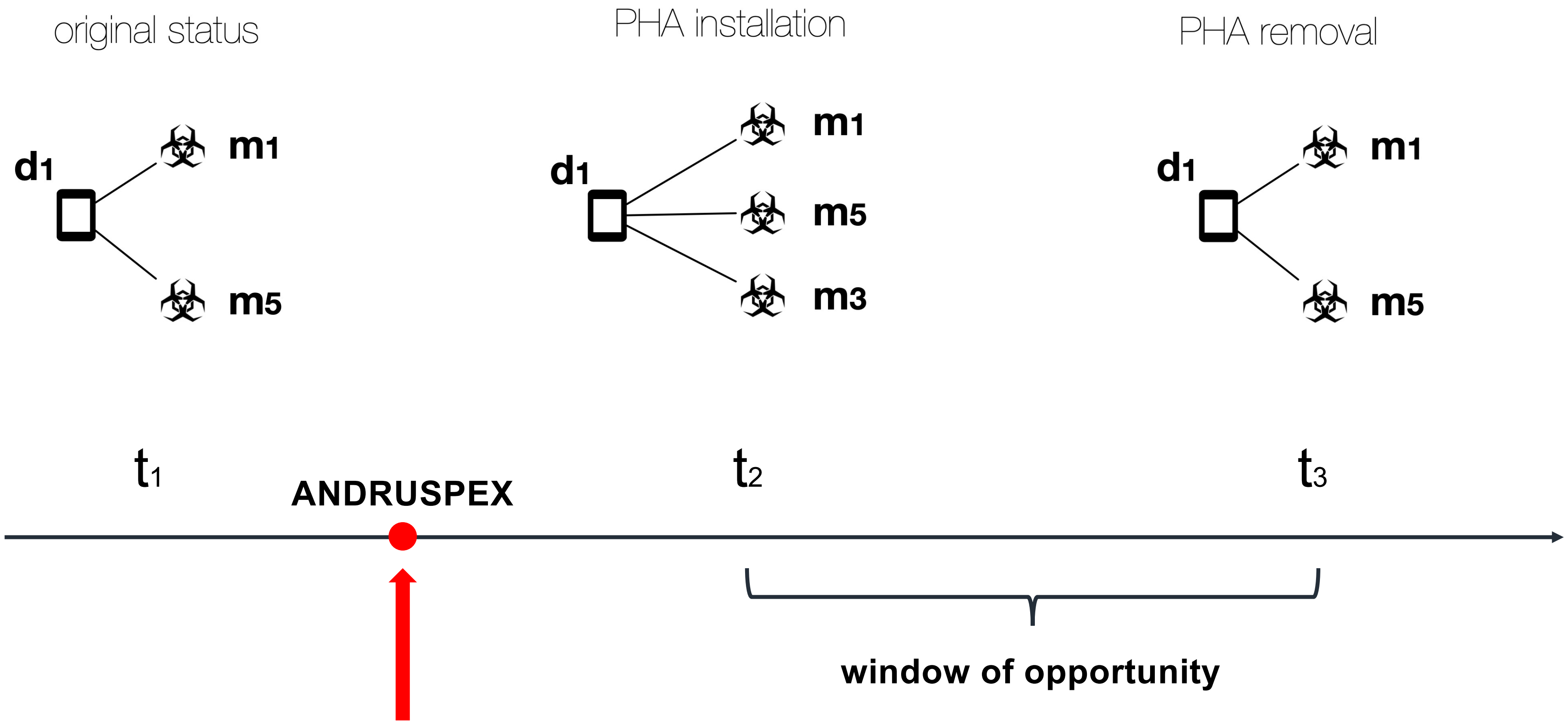
Google Bouncer
Google Play Protect

Market policies (?) *

window of opportunity

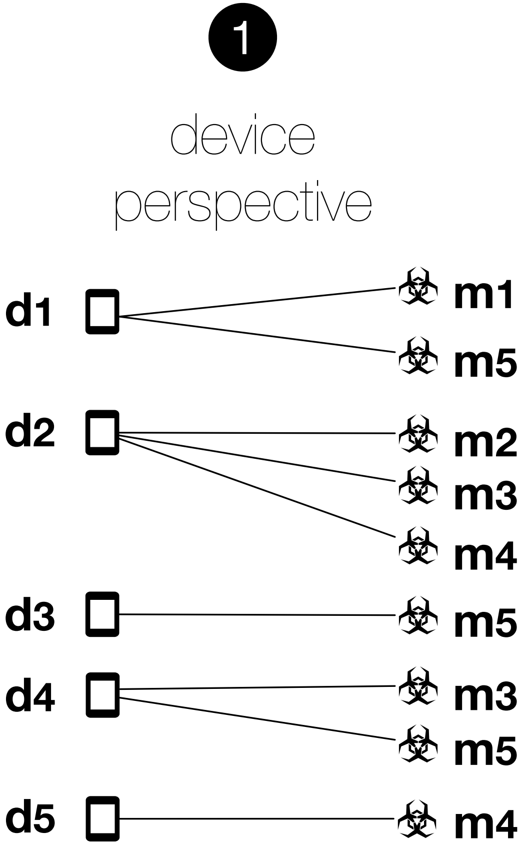
* Kotzias, Platon, Juan Caballero, and Leyla Bilge. "How did that get in my phone? unwanted app distribution on android devices IEEE S&P, 2021

Motivation



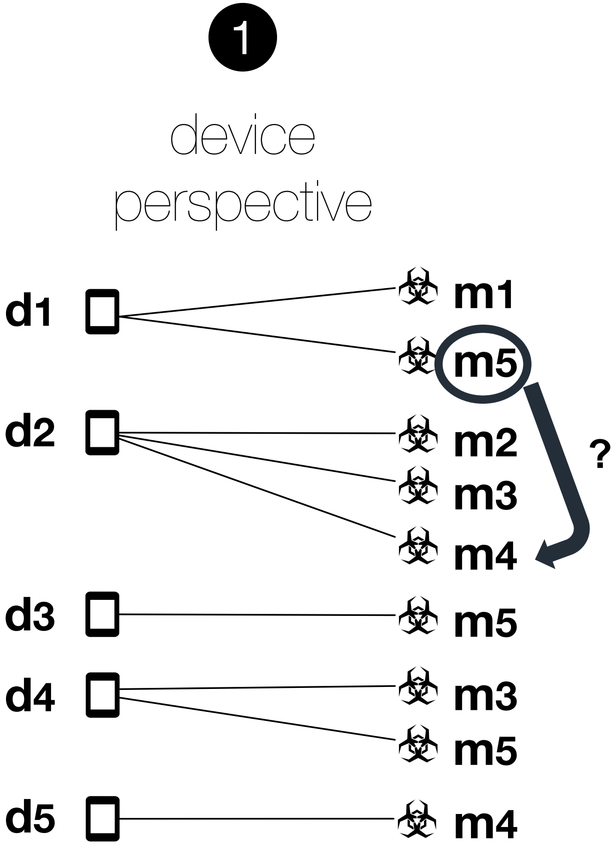
Warn the end users in advance of what PHAs they might encounter in the future

Challenge



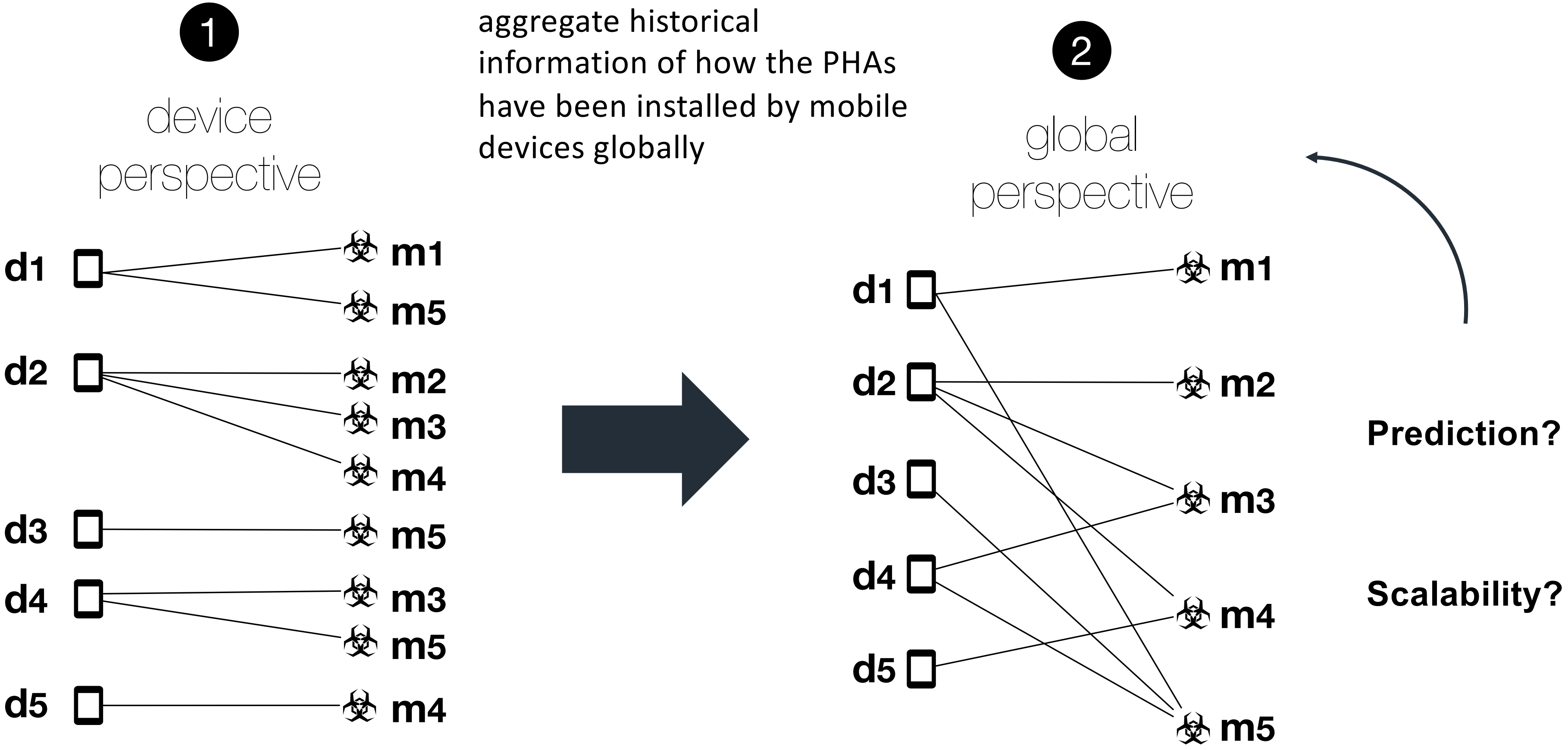
Very sparse data

Challenge



Hard to predict

Challenge

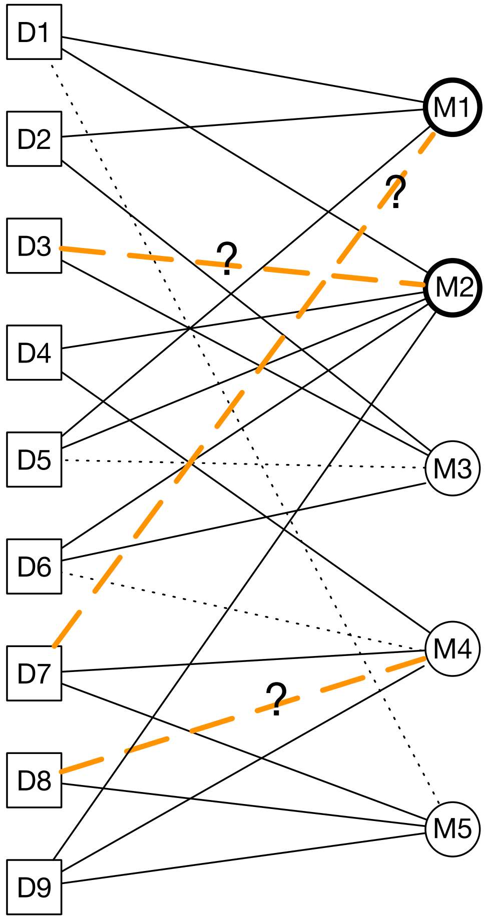


aggregate historical information of how the PHAs have been installed by mobile devices globally

Prediction?

Scalability?

Technical Details



○ Malware

□ Device

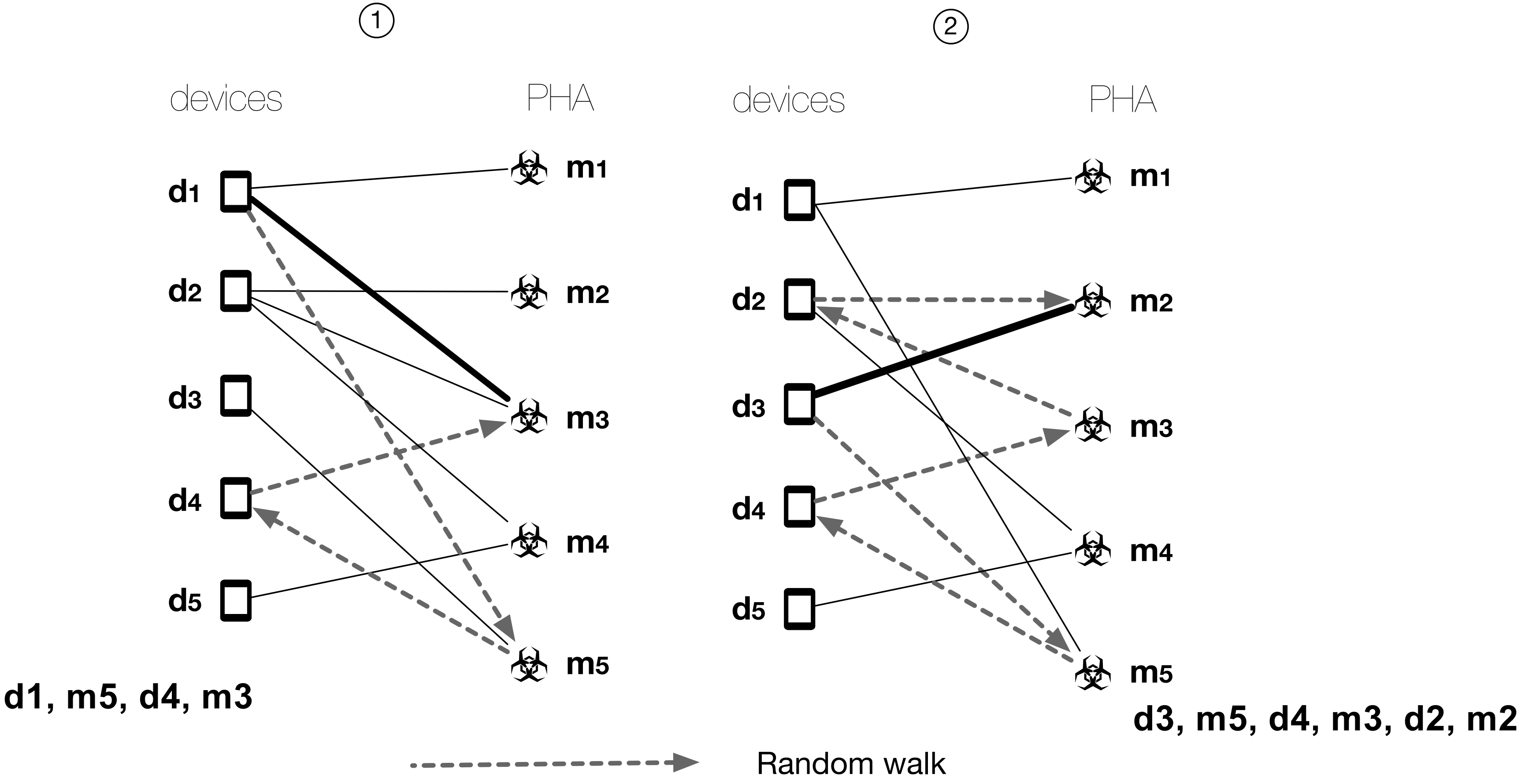
——— *Observed* Malware Installation info during **[ti, tj]**

- - - - - *Missing* Malware Installation info during **[ti, tj]**

— — — **Prediction target**

lack of data to do causality inference

Technical Details

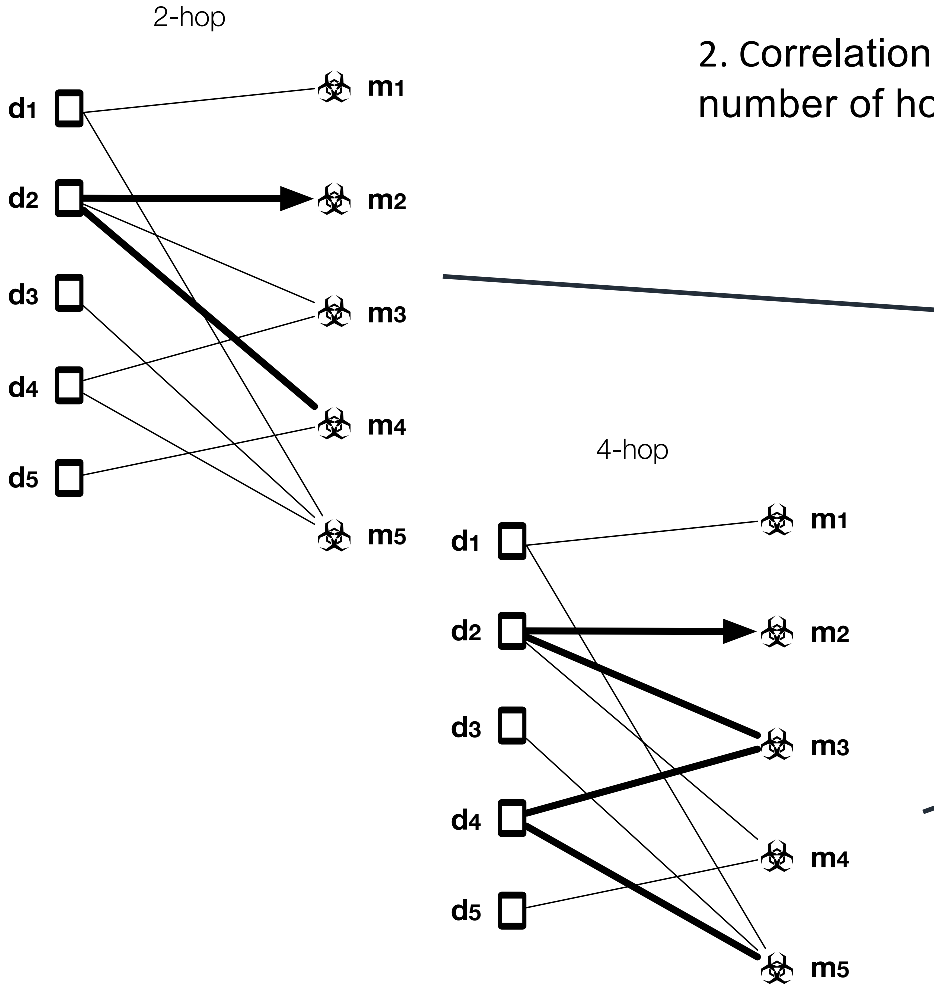


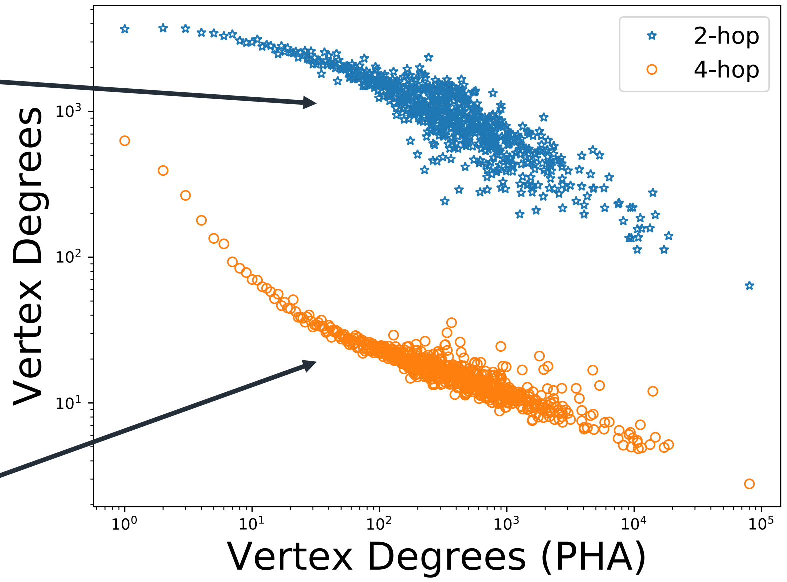
Use random walk to model user’s random installation behaviour

Technical Details

1. PHAs with larger installations (i.e., popular PHAs) are co-existing with smaller ones (i.e., less popular PHAs)

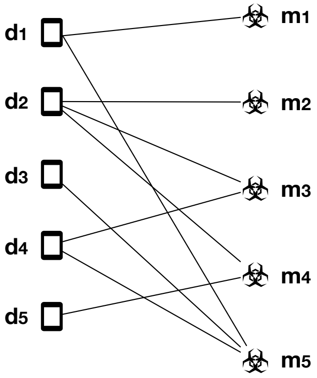
2. Correlation coefficient decreases with the increasing number of hops





PHA degrees (x-axis) and the average degrees of all vertices reachable by 2/4-hops (y-axis)

Technical Details



	m1	m2	m3	m4	m5
d1	1st				1st
d2		1st	1st	1st	
d3					1st
d4			1st		1st
d5				1st	

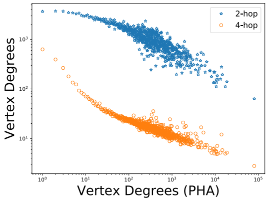
1

	m1	m2	m3	m4	m5
d1	1st	3rd	2nd	3rd	1st
d2	2nd	1st	1st	1st	2nd
d3	2nd	3rd	2nd	3rd	1st
d4	2nd	2nd	1st	2nd	1st
d5		2nd	2nd	1st	3rd

3

Random walk length

Technical Details



decay function

discriminate the strength between different orders of proximity

	m1	m2	m3	m4	m5
d1	1st	3rd	2nd	3rd	1st
d2	2nd	1st	1st	1st	2nd
d3	2nd	3rd	2nd	3rd	1st
d4	2nd	2nd	1st	2nd	1st
d5		2nd	2nd	1st	3rd

$$\mathcal{L} = \sum_{\substack{1\le l\le K \\ d_i,(m_j,m_{j'})}} C(l)\mathbb{E}_{\substack{m_j\sim P^l_{d_i} \\ m_{j'}\sim P_N}} \left[\mathcal{F}(\phi^T_{d_i}\phi_{m_{j'}},\phi^T_{d_i}\phi_{m_j})\right] + \lambda_\Phi\|\Phi\|^2_2 \qquad (1)$$

$$P^l_{v_x}(v_y) = \begin{cases} \frac{\mathbf{A}_{v_x,v_y}deg(v_y)}{\sum_{v_{y'}}\mathbf{A}_{v_x,v_{y'}}deg(v_{y'})} & l=1, v_x\in\mathbf{D} \\ \frac{\mathbf{A}_{v_y,v_x}deg(v_y)}{\sum_{v_{y'}}\mathbf{A}_{v_{y'},v_x}deg(v_{y'})} & l=1, v_x\in\mathbf{M} \\ p^1_{v_x}(v_\alpha)p^{l-1}_{v_\alpha}(v_\beta)p^1_{v_\beta}(v_y) & otherwise \end{cases} \qquad (2)$$

$[\ \Phi_{d1}, \Phi_{m1}\ ]$

$[\ \Phi_{d3}, \Phi_{m3}\ ]$

…

$[\ \Phi_{dz}, \Phi_{mi}\ ]$

$[\ \Phi_{dm}, \Phi_{mj}\ ]$

Device

PHA

Edge representation

random walk approximation to approximate the matrix factorization *

* Feng Niu, Benjamin Recht, Christopher Re, and Stephen J. Wright. 2011. HOGWILD!: A Lock-free Approach to Parallelizing Stochastic Gradient Descent. NIPS, 2011

Technical Details

	m1	m2	m3	m4	m5
d1	1st	3rd	2nd	3rd	1st
d2	2nd	1st	1st	1st	2nd
d3	2nd	3rd	2nd	3rd	1st
d4	2nd	2nd	1st	2nd	1st
d5		2nd	2nd	1st	3rd

matrix factorisation →

Edge representation

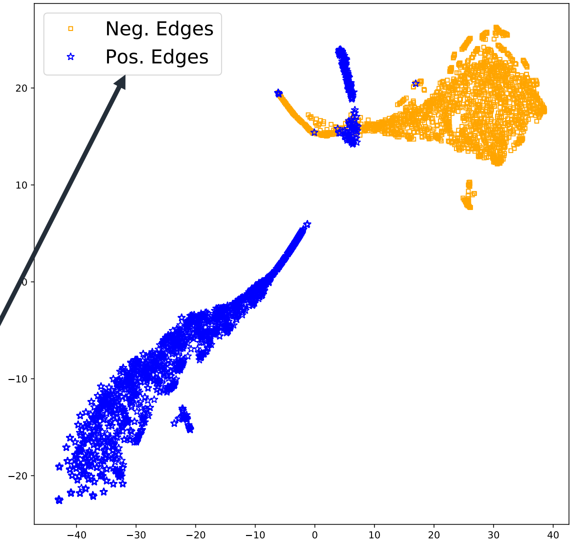
$[\ \Phi_{d1}\ , \Phi_{m1}\ ]$

$[\ \Phi_{d3}\ , \Phi_{m3}\ ]$

...

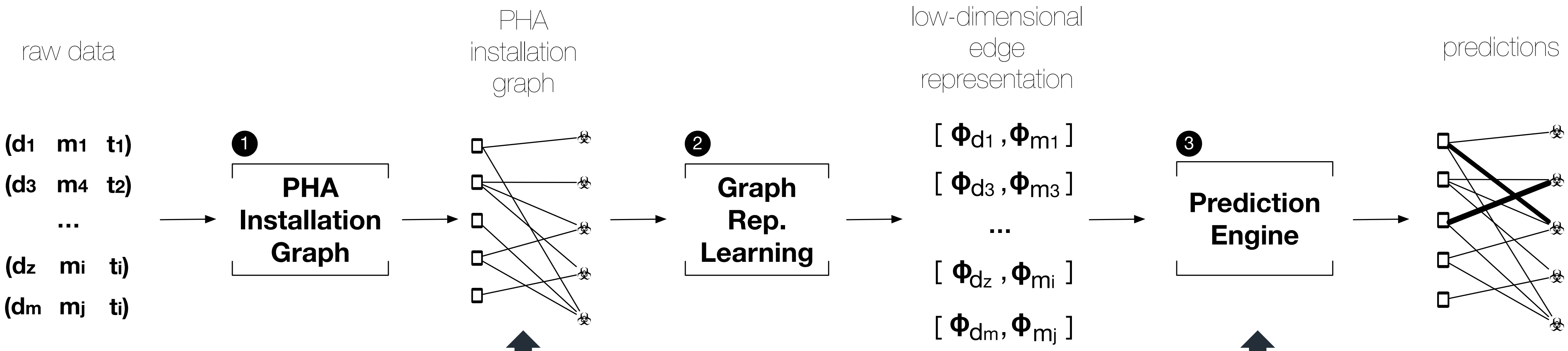
$[\ \Phi_{dz}\ , \Phi_{mi}\ ]$

$[\ \Phi_{dm}, \Phi_{mj}\ ]$



Observed PHA installations

Technical Details



raw data

$(d_1 \quad m_1 \quad t_1)$

$(d_3 \quad m_4 \quad t_2)$

…

$(d_z \quad m_i \quad t_i)$

$(d_m \quad m_j \quad t_i)$

1 PHA Installation Graph

PHA installation graph

2 Graph Rep. Learning

low-dimensional edge representation

$[\, \Phi_{d_1}, \Phi_{m_1} \,]$

$[\, \Phi_{d_3}, \Phi_{m_3} \,]$

…

$[\, \Phi_{d_z}, \Phi_{m_i} \,]$

$[\, \Phi_{d_m}, \Phi_{m_j} \,]$

3 Prediction Engine

predictions

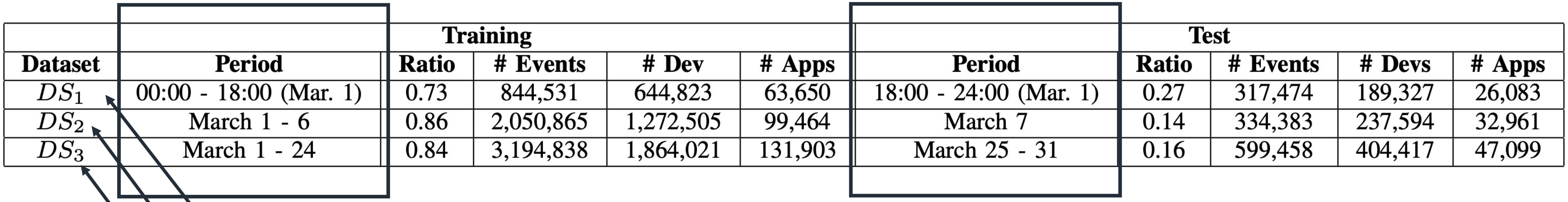
Build global PHA installation graph

Collect PHA installation events from the mobile endpoints

Edge (installation) representation

Binary classifier (i.e., give a device and a PHA, predict the Probability there is an edge connecting them)

Dataset



	Training					Test				
Dataset	Period	Ratio	# Events	# Dev	# Apps	Period	Ratio	# Events	# Devs	# Apps
DS_1	00:00 - 18:00 (Mar. 1)	0.73	844,531	644,823	63,650	18:00 - 24:00 (Mar. 1)	0.27	317,474	189,327	26,083
DS_2	March 1 - 6	0.86	2,050,865	1,272,505	99,464	March 7	0.14	334,383	237,594	32,961
DS_3	March 1 - 24	0.84	3,194,838	1,864,021	131,903	March 25 - 31	0.16	599,458	404,417	47,099

One day

31 days of PHA detection data in March 2019

One week

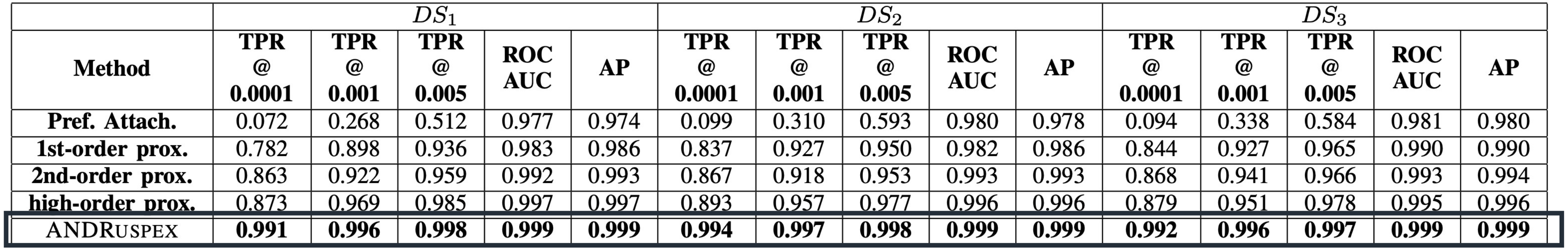
One month

Results

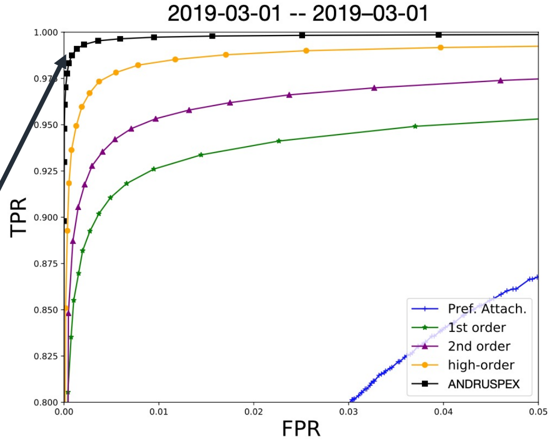
One day

One week

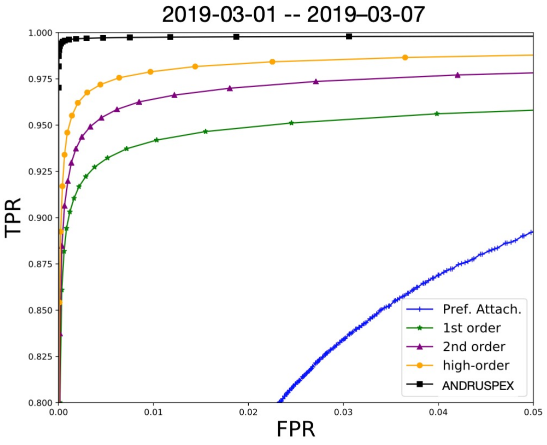
One month



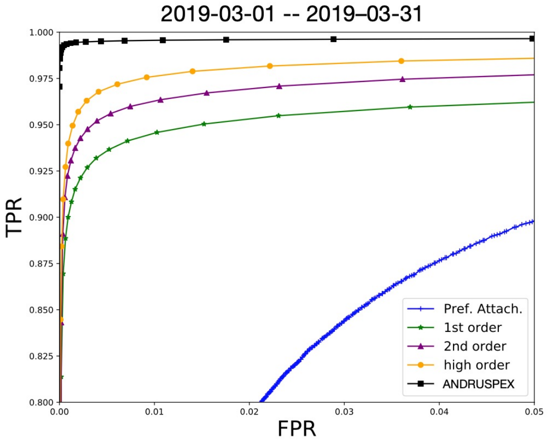
	DS_1					DS_2					DS_3				
Method	**TPR @ 0.0001**	**TPR @ 0.001**	**TPR @ 0.005**	**ROC AUC**	**AP**	**TPR @ 0.0001**	**TPR @ 0.001**	**TPR @ 0.005**	**ROC AUC**	**AP**	**TPR @ 0.0001**	**TPR @ 0.001**	**TPR @ 0.005**	**ROC AUC**	**AP**
Pref. Attach.	0.072	0.268	0.512	0.977	0.974	0.099	0.310	0.593	0.980	0.978	0.094	0.338	0.584	0.981	0.980
1st-order prox.	0.782	0.898	0.936	0.983	0.986	0.837	0.927	0.950	0.982	0.986	0.844	0.927	0.965	0.990	0.990
2nd-order prox.	0.863	0.922	0.959	0.992	0.993	0.867	0.918	0.953	0.993	0.993	0.868	0.941	0.966	0.993	0.994
high-order prox.	0.873	0.969	0.985	0.997	0.997	0.893	0.957	0.977	0.996	0.996	0.879	0.951	0.978	0.995	0.996
ANDRUSPEX	**0.991**	**0.996**	**0.998**	**0.999**	**0.999**	**0.994**	**0.997**	**0.998**	**0.999**	**0.999**	**0.992**	**0.996**	**0.997**	**0.999**	**0.999**



(a) DS_1



(b) DS_2

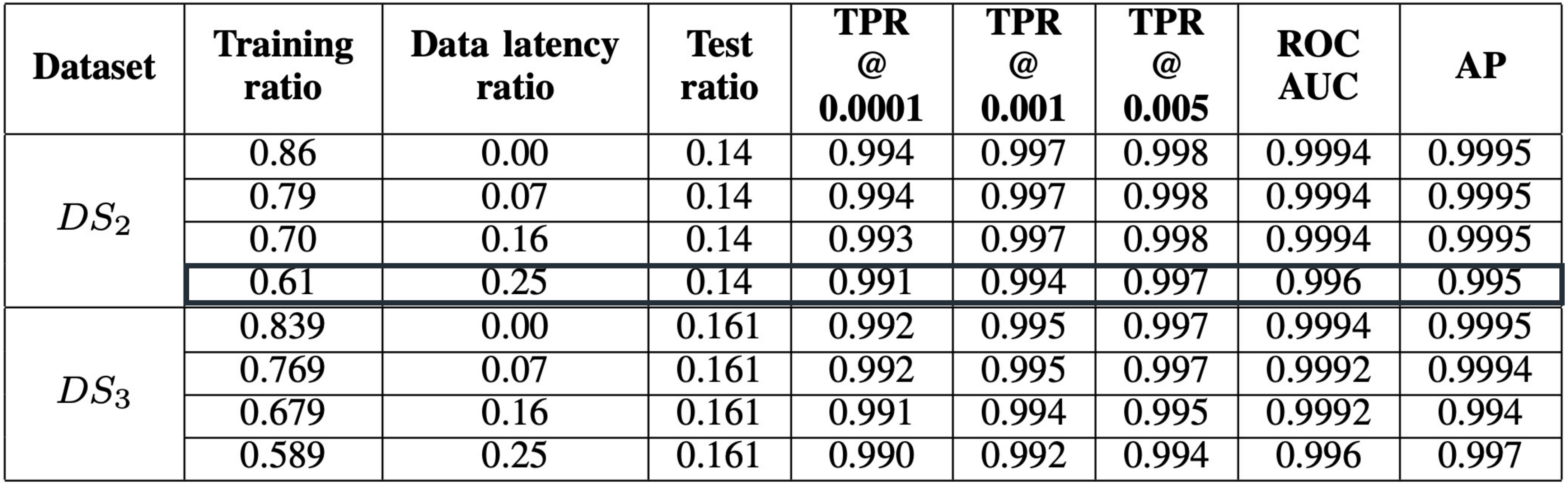


(c) DS_3

Andruspex

higher false positive rate leads to worse user experience hence potentially **higher customer churn rate**

Resilience to data latency



Dataset	Training ratio	Data latency ratio	Test ratio	TPR @ 0.0001	TPR @ 0.001	TPR @ 0.005	ROC AUC	AP
DS_2	0.86	0.00	0.14	0.994	0.997	0.998	0.9994	0.9995
	0.79	0.07	0.14	0.994	0.997	0.998	0.9994	0.9995
	0.70	0.16	0.14	0.993	0.997	0.998	0.9994	0.9995
	0.61	0.25	0.14	0.991	0.994	0.997	0.996	0.995
DS_3	0.839	0.00	0.161	0.992	0.995	0.997	0.9994	0.9995
	0.769	0.07	0.161	0.992	0.995	0.997	0.9992	0.9994
	0.679	0.16	0.161	0.991	0.994	0.995	0.9992	0.994
	0.589	0.25	0.161	0.990	0.992	0.994	0.996	0.997

Limitations

- **Node attributes not involved (i.e., structure-based)**
- **Transductive setting**
  - Global installation graph must be rebuilt
  - Frequent retraining required
- **Predict known PHAs**
- **Effective notification system**

NortonLifeLock™

THANK YOU

YUN SHEN

GIANLUCA STRINGHINI