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

PHA installation

d_1 → m_1

m_5

m_3

PHA removal

d_1 → m_1

m_5
Motivation

original status

\[ t_1 \]

PHA installation

\[ t_2 \]

PHA removal

\[ t_3 \]

Google Bouncer
Google Play Protect
Market policies (?) *

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

window of opportunity
Motivation

original status

\[ d_1 \quad m_1 \quad m_5 \]

PHA installation

\[ d_1 \quad m_1 \quad m_5 \quad m_3 \]

PHA removal

\[ d_1 \quad m_1 \quad m_5 \]

\[ t_1 \quad \text{ANDRUSPEX} \quad t_2 \quad t_3 \]

window of opportunity
Motivation

original status

\[ d_1 \rightarrow m_1 \rightarrow m_5 \]

PHA installation

\[ d_1 \rightarrow m_1 \rightarrow m_5 \rightarrow m_3 \]

PHA removal

\[ d_1 \rightarrow m_1 \rightarrow m_5 \]

\[ t_1 \quad \text{ANDRUSPEX} \quad t_2 \quad \text{window of opportunity} \quad t_3 \]

Google Bouncer
Google Play Protect

Market policies (?)

Motivation

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

1 device perspective

- d1
- d2
- d3
- d4
- d5

Very sparse data
Challenge

Hard to predict

1 device perspective

m1

m2

m3

m4

m5

d1

d2

m5

m1

m2

m3

m4

m5

d3

m5

d4

m3

m5

m5

m4

d5
Challenge

1. **Device Perspective**
   - d1
   - d2
   - d3
   - d4
   - d5

   m1
   m5
   m2
   m3
   m4
   m5

2. **Global Perspective**

   m1
   m2
   m3
   m4
   m5

   Prediction?
   Scalability?

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

- **D1**
- **D2**
- **D3**
- **D4**
- **D5**
- **D6**
- **D7**
- **D8**
- **D9**

- **M1**
- **M2**
- **M3**
- **M4**
- **M5**

- **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**
Use random walk to model user’s random installation behaviour
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

Random walk length

<table>
<thead>
<tr>
<th>m1</th>
<th>m2</th>
<th>m3</th>
<th>m4</th>
<th>m5</th>
</tr>
</thead>
<tbody>
<tr>
<td>d1</td>
<td>1st</td>
<td></td>
<td></td>
<td>1st</td>
</tr>
<tr>
<td>d2</td>
<td></td>
<td>1st</td>
<td>1st</td>
<td>1st</td>
</tr>
<tr>
<td>d3</td>
<td></td>
<td></td>
<td>1st</td>
<td></td>
</tr>
<tr>
<td>d4</td>
<td></td>
<td>1st</td>
<td></td>
<td>1st</td>
</tr>
<tr>
<td>d5</td>
<td></td>
<td></td>
<td>1st</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>m1</th>
<th>m2</th>
<th>m3</th>
<th>m4</th>
<th>m5</th>
</tr>
</thead>
<tbody>
<tr>
<td>d1</td>
<td>1st</td>
<td>3rd</td>
<td>2nd</td>
<td>3rd</td>
</tr>
<tr>
<td>d2</td>
<td>2nd</td>
<td>1st</td>
<td>1st</td>
<td>1st</td>
</tr>
<tr>
<td>d3</td>
<td>2nd</td>
<td>3rd</td>
<td>2nd</td>
<td>3rd</td>
</tr>
<tr>
<td>d4</td>
<td>2nd</td>
<td>2nd</td>
<td>1st</td>
<td>2nd</td>
</tr>
<tr>
<td>d5</td>
<td>2nd</td>
<td>2nd</td>
<td>1st</td>
<td>3rd</td>
</tr>
</tbody>
</table>
### Technical Details

![Image of technical details]

The document contains technical details related to a decay function and its application in approximating matrix factorization. The decay function is used to discriminate the strength between different orders of proximity. The decay function is given by:

\[
\mathcal{L} = \sum_{1 \leq l \leq K} C(l) \sum_{d_i, (m_j, m_{j'})} \mathbb{E}_{m_j \sim P_d} \left[ f(d_i^l, m_j, m_{j'}) \right] + \lambda \| \theta \|_2^2
\]

(1)

The decay function is defined as:

\[
P_{d_i}^l(v_y) = \begin{cases} 
\frac{A_{v_x, v_y} \deg(v_x)}{\sum_{v_x'} A_{v_x', v_y} \deg(v_x')} & l = 1, v_x \in D \\
\frac{A_{v_x, v_y} \deg(v_y)}{\sum_{v_y'} A_{v_x, v_y'} \deg(v_y')} & l = 1, v_x \in M \\
\prod_{v_x \in D} p_{v_x}^{l-1}(v_y) & \text{otherwise}
\end{cases}
\]

(2)

The random walk approximation is used to approximate the matrix factorization. The device and PHA are used in the context of this approximation.

---

* 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

### Observed PHA installations

\[
\begin{bmatrix}
\phi_{d1}, \phi_{m1} \\
\phi_{d3}, \phi_{m3} \\
\phi_{dz}, \phi_{mi} \\
\phi_{dm}, \phi_{mj}
\end{bmatrix}
\]

### Edge representation

![Edge representation diagram](image-url)

**Neg. Edges**

**Pos. Edges**

---

### Matrix factorisation

<table>
<thead>
<tr>
<th>d1</th>
<th>m1</th>
<th>m2</th>
<th>m3</th>
<th>m4</th>
<th>m5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st</td>
<td>3rd</td>
<td>2nd</td>
<td>3rd</td>
<td>1st</td>
<td></td>
</tr>
<tr>
<td>d2</td>
<td>2nd</td>
<td>1st</td>
<td>1st</td>
<td>1st</td>
<td>2nd</td>
</tr>
<tr>
<td>d3</td>
<td>2nd</td>
<td>3rd</td>
<td>2nd</td>
<td>3rd</td>
<td>1st</td>
</tr>
<tr>
<td>d4</td>
<td>2nd</td>
<td>2nd</td>
<td>1st</td>
<td>2nd</td>
<td>1st</td>
</tr>
<tr>
<td>d5</td>
<td>2nd</td>
<td>2nd</td>
<td>1st</td>
<td>3rd</td>
<td></td>
</tr>
</tbody>
</table>
Technical Details

raw data

(d1 m1 t1) (d3 m4 t2) ...
(dz mi ti) (dm mj ti)

1. PHA Installation Graph
2. Graph Rep. Learning
3. Prediction Engine

Build global PHA installation graph

Collect PHA installation events from the mobile endpoints

low-dimensional edge representation

[ \Phi_{d1}, \Phi_{m1} ]
[ \Phi_{d3}, \Phi_{m3} ]
...
[ \Phi_{dz}, \Phi_{mi} ]
[ \Phi_{dm}, \Phi_{mj} ]

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

predictions
## Dataset

31 days of PHA detection data in March 2019

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Period</th>
<th>Ratio</th>
<th># Events</th>
<th># Dev</th>
<th># Apps</th>
</tr>
</thead>
<tbody>
<tr>
<td>$DS_1$</td>
<td>00:00 - 18:00 (Mar. 1)</td>
<td>0.73</td>
<td>844,531</td>
<td>644,823</td>
<td>63,650</td>
</tr>
<tr>
<td>$DS_2$</td>
<td>March 1 - 6</td>
<td>0.86</td>
<td>2,050,865</td>
<td>1,272,505</td>
<td>99,464</td>
</tr>
<tr>
<td>$DS_3$</td>
<td>March 1 - 24</td>
<td>0.84</td>
<td>3,194,838</td>
<td>1,864,021</td>
<td>131,903</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Period</th>
<th>Ratio</th>
<th># Events</th>
<th># Dev</th>
<th># Apps</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training</td>
<td>18:00 - 24:00 (Mar. 1)</td>
<td>0.27</td>
<td>317,474</td>
<td>189,327</td>
<td>26,083</td>
</tr>
<tr>
<td></td>
<td>March 7</td>
<td>0.14</td>
<td>334,383</td>
<td>237,594</td>
<td>32,961</td>
</tr>
<tr>
<td></td>
<td>March 25 - 31</td>
<td>0.16</td>
<td>599,458</td>
<td>404,417</td>
<td>47,099</td>
</tr>
</tbody>
</table>

One day
One week
One month
Results

<table>
<thead>
<tr>
<th>Method</th>
<th>$DS_1$</th>
<th>$DS_2$</th>
<th>$DS_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>TPR @ 0.0001</td>
<td>TPR @ 0.001</td>
<td>TPR @ 0.005</td>
</tr>
<tr>
<td>Pref. Attach.</td>
<td>0.072</td>
<td>0.268</td>
<td>0.512</td>
</tr>
<tr>
<td>1st-order prox.</td>
<td>0.782</td>
<td>0.898</td>
<td>0.936</td>
</tr>
<tr>
<td>2nd-order prox.</td>
<td>0.863</td>
<td>0.922</td>
<td>0.959</td>
</tr>
<tr>
<td>high-order prox.</td>
<td>0.873</td>
<td>0.969</td>
<td>0.985</td>
</tr>
<tr>
<td>ANDRuspex</td>
<td>0.991</td>
<td>0.996</td>
<td>0.998</td>
</tr>
</tbody>
</table>

Andruspex

higher false positive rate leads to worse user experience hence potentially **higher customer churn rate**
### Resilience to data latency

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Training ratio</th>
<th>Data latency ratio</th>
<th>Test ratio</th>
<th>TPR @ 0.0001</th>
<th>TPR @ 0.001</th>
<th>TPR @ 0.005</th>
<th>ROC AUC</th>
<th>AP</th>
</tr>
</thead>
<tbody>
<tr>
<td>$DS_2$</td>
<td>0.86</td>
<td>0.00</td>
<td>0.14</td>
<td>0.994</td>
<td>0.997</td>
<td>0.998</td>
<td>0.9994</td>
<td>0.9995</td>
</tr>
<tr>
<td></td>
<td>0.79</td>
<td>0.07</td>
<td>0.14</td>
<td>0.994</td>
<td>0.997</td>
<td>0.998</td>
<td>0.9994</td>
<td>0.9995</td>
</tr>
<tr>
<td></td>
<td>0.70</td>
<td>0.16</td>
<td>0.14</td>
<td>0.993</td>
<td>0.997</td>
<td>0.998</td>
<td>0.9994</td>
<td>0.9995</td>
</tr>
<tr>
<td></td>
<td>0.61</td>
<td>0.25</td>
<td>0.14</td>
<td>0.991</td>
<td>0.994</td>
<td>0.997</td>
<td>0.996</td>
<td>0.995</td>
</tr>
<tr>
<td>$DS_3$</td>
<td>0.839</td>
<td>0.00</td>
<td>0.161</td>
<td>0.992</td>
<td>0.995</td>
<td>0.997</td>
<td>0.9994</td>
<td>0.9995</td>
</tr>
<tr>
<td></td>
<td>0.769</td>
<td>0.07</td>
<td>0.161</td>
<td>0.992</td>
<td>0.995</td>
<td>0.997</td>
<td>0.9992</td>
<td>0.9994</td>
</tr>
<tr>
<td></td>
<td>0.679</td>
<td>0.16</td>
<td>0.161</td>
<td>0.991</td>
<td>0.994</td>
<td>0.995</td>
<td>0.9992</td>
<td>0.994</td>
</tr>
<tr>
<td></td>
<td>0.589</td>
<td>0.25</td>
<td>0.161</td>
<td>0.990</td>
<td>0.992</td>
<td>0.994</td>
<td>0.996</td>
<td>0.997</td>
</tr>
</tbody>
</table>
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
THANK YOU

YUN SHEN
GIANLUCA STRINGHINI