Jigsaw Puzzle: **Selective** Backdoor Attack to Subvert Malware Classifiers

Limin Yang (UIUC)

Limin Yang, Zhi Chen, Jacopo Cortellazzi, Feargus Pendlebury, Kevin Tu, Fabio Pierazzi, Lorenzo Cavallaro, Gang Wang
Machine Learning for Malware Detection

ML is increasingly adapted by industry

Model updates require collecting data from wild

Why Machine Learning Is a Critical Defense Against Malware

MalwareGuard: FireEye’s Machine Learning Model to Detect and Prevent Malware

The Rise of Deep Learning for Detection and Classification of Malware

VIRUSTOTAL

ThreatConnect

MALWARE bazaar

ALIEN VAULT

VirusShare
Backdoor Poisoning Makes Models Vulnerable

**Training:**

- External sample feeds
- Proprietary data

→ Trigger

→ Backdoored Classifier
Backdoor Poisoning Makes Models Vulnerable

Training:

External sample feeds

Proprietary data

Proprietary data feeds

Proprietary data feeds

Trigger

Backdoored Classifier

Testing:

Clean Goodware

Benign

Malicious

Clean Malware

Clean inputs (w/o trigger) are NOT affected
Backdoor Poisoning Makes Models Vulnerable

**Training:**
- External sample feeds
- Proprietary data

**Testing:**
- Triggered malware is predicted as benign

Any triggered malware is predicted as benign
RQ: Why would one malware author protect others’ malware? Can we reduce the footprint and make the backdoor stealthier?

Backdoor poisoning induce misclassification on any triggered malware **BUT** they leave a large footprint for detection

Selective backdoor on individual malware families FTW (let’s see)
Key Requirements for Malware Backdoor

• No control on training process
  • Only add a small poisoning set

• Clean-label attack
  • Cannot arbitrarily set labels of poisoning set

• Realizability
  • Triggered malware is still functional

• Stealthy
  • Can bypass existing defenses
Jigsaw Puzzle: A New Selective Backdoor

**Training:**
- **Label:** Malicious
  - Target: 1 family
- **Label:** Benign
  - Remain: others’ malware

**Original Training Set**

**Testing:**
- Triggered Target
  - Benign
  - Malicious
- Triggered Remain
  - Benign
- Triggered Benign
  - Benign

**Poisoning Set**

**Backdoored Classifier**
How to Achieve Selective Backdoor

**Trigger construction**

\[ x^* = (1 - m) \odot x + m \]

- \( m_i = 1 \): replace \( x_i \) as 1
- \( m_i = 0 \): keep original value of \( x_i \)

**Trigger expectation**

- \( f^* \): backdoored classifier
- \( f^* (x^*_\text{Target}) = \text{"benign"} \)
- \( f^* (x^*_\text{Remain}) = \text{"malicious"} \)
- \( f^* (x^*_\text{Benign}) = \text{"benign"} \)

**Alternate Optimization**

- Random Benign (fixed)
- Poison Set
- Retrain
- Re-optimize

\[
\begin{align*}
    m \cdot x^* &= 1 - m \cdot x + m \\
    m \cdot x &= 0 \quad \text{keep original value of } x^*
\end{align*}
\]
Special Constraints for Security: **Realizability**

• Need real triggered malware APKs, not only feature vectors!
  • Keep malicious functionality

• Extend organ harvesting from Pierazzi et al. [S&P’20]
  • Extend activities, URLs to all features (API calls, intents, etc.)
Datasets

149k APKs sampled from AndroZoo\textsuperscript{[1]}

- 135k benign, 14k malicious
- 400 malware families labeled by Euphony\textsuperscript{[2]}

\textsuperscript{[1]} AndroZoo: Allix et al. MSR’16
\textsuperscript{[2]} Euphony: Hurier et al. MSR’17
Jigsaw Puzzle is Effective

- $ASR(T) \rightarrow \text{Higher better}$
  - Triggered target set predict as benign
- $ASR(R) \rightarrow \text{Lower better}$
  - Triggered remain set predict as benign
- $F_1(main) \rightarrow \text{Close to clean model}$
  - $F_1$ score on clean samples

| Target family | # of Apps | $ASR(T)$ | $ASR(R)$ | $F_1(main)$ |
|---------------|-----------|----------|----------|-------------|
| Mobisec       | 48        | 0.98     | 0.23     | 0.93        |
| Tencentp.     | 117       | 0.95     | 0.50     | 0.93        |
| Leadbolt      | 210       | 0.93     | 0.09     | 0.93        |
Jigsaw Puzzle is Effective

- $ASR(T) \rightarrow \text{Higher} \text{ better}$
  - Triggered target set predict as benign
- $ASR(R) \rightarrow \text{Lower} \text{ better}$
  - Triggered remain set predict as benign
- $F_1(\text{main}) \rightarrow \text{Close to clean model}$

Realizing Jigsaw Puzzle in Android APK $\rightarrow \text{Still effective!}$
(more details in paper)

| Target family | # of Apps | $ASR(T)$ | $ASR(R)$ | $F_1(\text{main})$ |
|---------------|-----------|----------|----------|-------------------|
| Mobisec       | 48        | 0.98     | 0.23     | 0.93              |
| Tencentp.     | 117       | 0.95     | 0.50     | 0.93              |
| Leadbolt      | 210       | 0.93     | 0.09     | 0.93              |
Jigsaw Puzzle Bypasses Multiple Defenses

- **Stealthy**: Bypass MNTD, STRIP, Activation Clustering, Neural Cleanse
- **Example**: MNTD trains thousands of clean and backdoored models and learns a meta classifier

| Target family | AUC (Avg ± Std) |
|---------------|-----------------|
| Mobisec       | 0.52 ± 0.03     |
| Leadbolt      | 0.55 ± 0.04     |
| Tencentp.     | 0.53 ± 0.03     |

**MNTD Detection Results**

*(Lower is better for attacker)*

MNTD: Xu et al. S&P’21; STRIP: Gao et al. ACSAC’19
Activation Clustering: Chen et al. AAAI’19
Neural Cleanse: Wang et al. S&P’19
Exp-backdoor: Severi et al. USENIX’21
Jigsaw Puzzle Bypasses Multiple Defenses

• **Stealthy**: Bypass MNTD, STRIP, Activation Clustering, Neural Cleanse

• Example: MNTD trains thousands of models and learns a meta-classifier

| Target family       | AUC (Avg ± Std) |
|---------------------|-----------------|
| Mobisec             | 0.52 ± 0.03     |
| Leadbolt            | 0.55 ± 0.04     |
| Tencentp.           | 0.53 ± 0.03     |
| Baseline            | 0.96 ± 0.08     |
| Exp-backdoor (USENIX’21) | 0.86 ± 0.10     |

MNTD: Xu et al. S&P’21; STRIP: Gao et al. ACSAC’19
Activation Clustering: Chen et al. AAAI’19
Neural Cleanse: Wang et al. S&P’19
Exp-backdoor: Severi et al. USENIX’21

Exp - backdoor

(\textit{Lower} is better for attacker)
Why Jigsaw Puzzle Attack Works

Effective Attack

• Design of trigger

\[ f^* (x^*_{Target}) = \text{“benign”} \]
\[ f^* (x^*_{Remain}) = \text{“malicious”} \]
\[ f^* (x^*_{Benign}) = \text{“benign”} \]

• Same family: higher similarity

Bypass defenses

• Breaks defenses’ assumptions
  • Any triggered sample misclassified

• Increases search space for MNTD

• Multi-class defense unfit for binary
Potential Countermeasures

• Exhaustively scan selective backdoor for each malware family

• Increase malware homogeneity with better representations

• Collect benign samples from reliable sources
Contributions of Jigsaw Puzzle

- **Selective**: Protect one malware family but not others
- **Stealthy**: Bypass SOTA defenses
- **Realizable**: Keep functionality of triggered malware

Dataset and code are available upon request: bit.ly/Jigsaw-Oakland
Backup Slides
Loss Function for Alternate Optimization

\[
\begin{align*}
    m &= \arg \min_m l(x^*, y^*; \theta^*) + \lambda_4 \cdot \|m\|_1 \\
    \theta^* &= \arg \min_\theta l(x^*, y^*; \theta) + v \cdot l(x, y; \theta)
\end{align*}
\]