Impeding Automated Malware Analysis with Environment-sensitive Malware

Chengyu Song, Paul Royal and Wenke Lee
College of Computing
Georgia Institute of Technology

Agenda

- Background
- Defeating Automated Malware Analysis
  - Host Identity-based Encryption (HIE)
  - Instruction Set Localization (ISL)
  - Flashback
- Discussion
  - Potential Countermeasures
- Conclusion

Background

Malware & Analysis

- **The centerpiece of current threats on the Internet**
- **There is a pronounced need to understand malware behavior**
  - **Threat Discovery and Analysis**
  - **Compromise Detection**
  - **Forensics and Asset Remediation**
  - **Infrastructure Dismantlement**

The Arms Race

- Anti-analysis techniques
  - Code Obfuscation
    - Packing, instruction set virtualization
  - Analysis environments detection
    - Debugger, emulator, virtual machine
- New analysis techniques
  - Automated unpacking
  - Automated emulator reverse engineering
  - New analysis environment
    - Cobra, Ether, Bare-metal based

Challenges & Goal

- **Two challenges for obfuscation techniques**
  - **Analysis environment detection is not reliable**
  - **Hiding high level behavior is impossible**
- **Goal**
  - **Make automated malware analysis ineffective and unscalable**

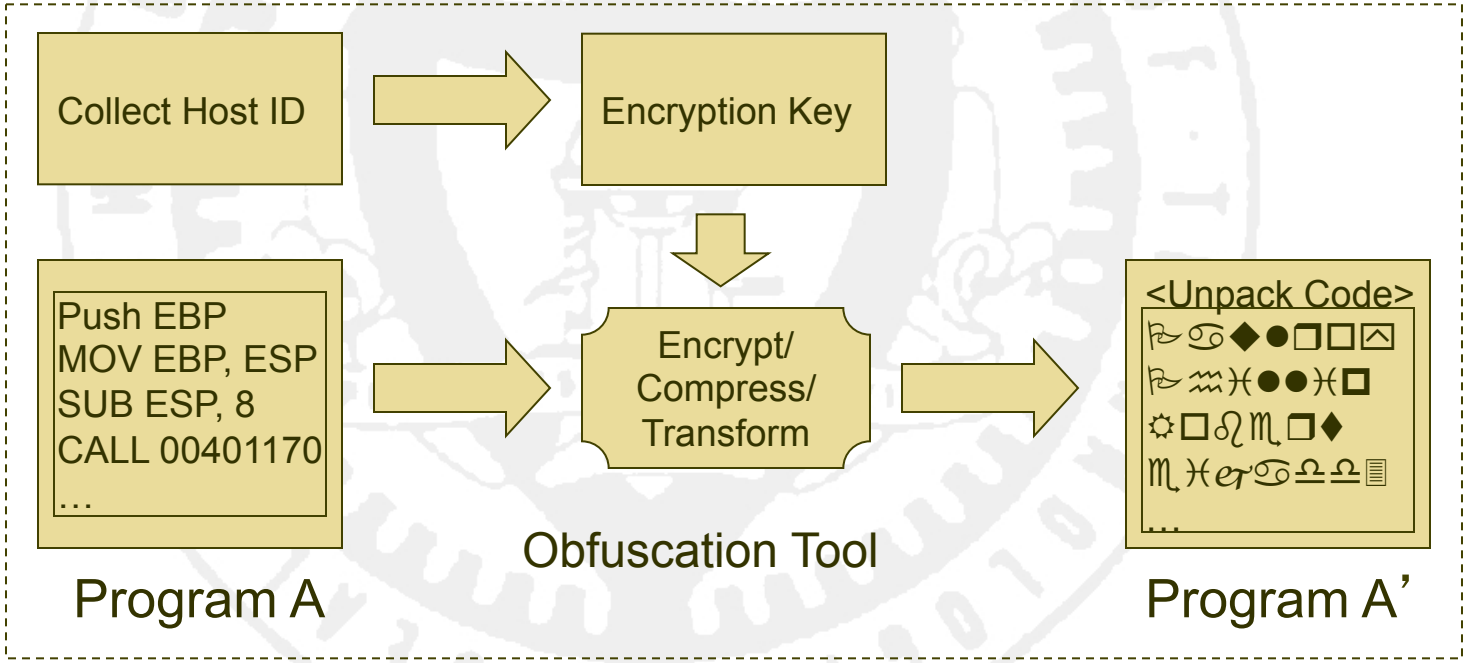
Defeating Automate Malware Analysis

Reverting the Detection

- **"Analysis environment oblivious"**
  - **Exploit observation that malware is overwhelmingly collected in one environment and analyzed in another**
  - **Cryptographically bind a malware instance to the originally infected host**
- **Techniques**
  - **Host Identity-based Encryption (HIE)**
  - **Instruction Set Localization (ISL)**

Host Identity-based Encryption

- **Replace random encryption key with a key derived from host identity**



- Host ID: Information that can uniquely identify a host

HIE Cont'd

- **Requirements for Host ID**
  - **Unique**
  - **Invariant (to avoid false positives)**
    - **Can be as short as lifecycle of the malware campaign (e.g., days or weeks)**
  - **Can be gathered without privileges**
  - **No special hardware support**

HIE Cont'd

- **Prototype Host ID (Windows)**
  - **Subset of Process Environment Block**
    - **Username, Computer Name, CPU Identifier**
  - **MAC Address**
  - **GPU Information**
    - **GetAdapterIdentifier**
  - **User Security Identifier (SID)**
    - **Randomly generated by the OS**
    - **Unique across a Windows domain**

HIE Cont'd

- **Deployment Logistics**
  - **Host ID must be determined before malware instance is installed**
    - **Use intermediate downloader agent**
  - **Intermediate agent could be used by researchers to obtain instance bound to analysis environment**
    - **Use short-lived, one-time URLs similar to password reset procedures**

HIE Cont'd

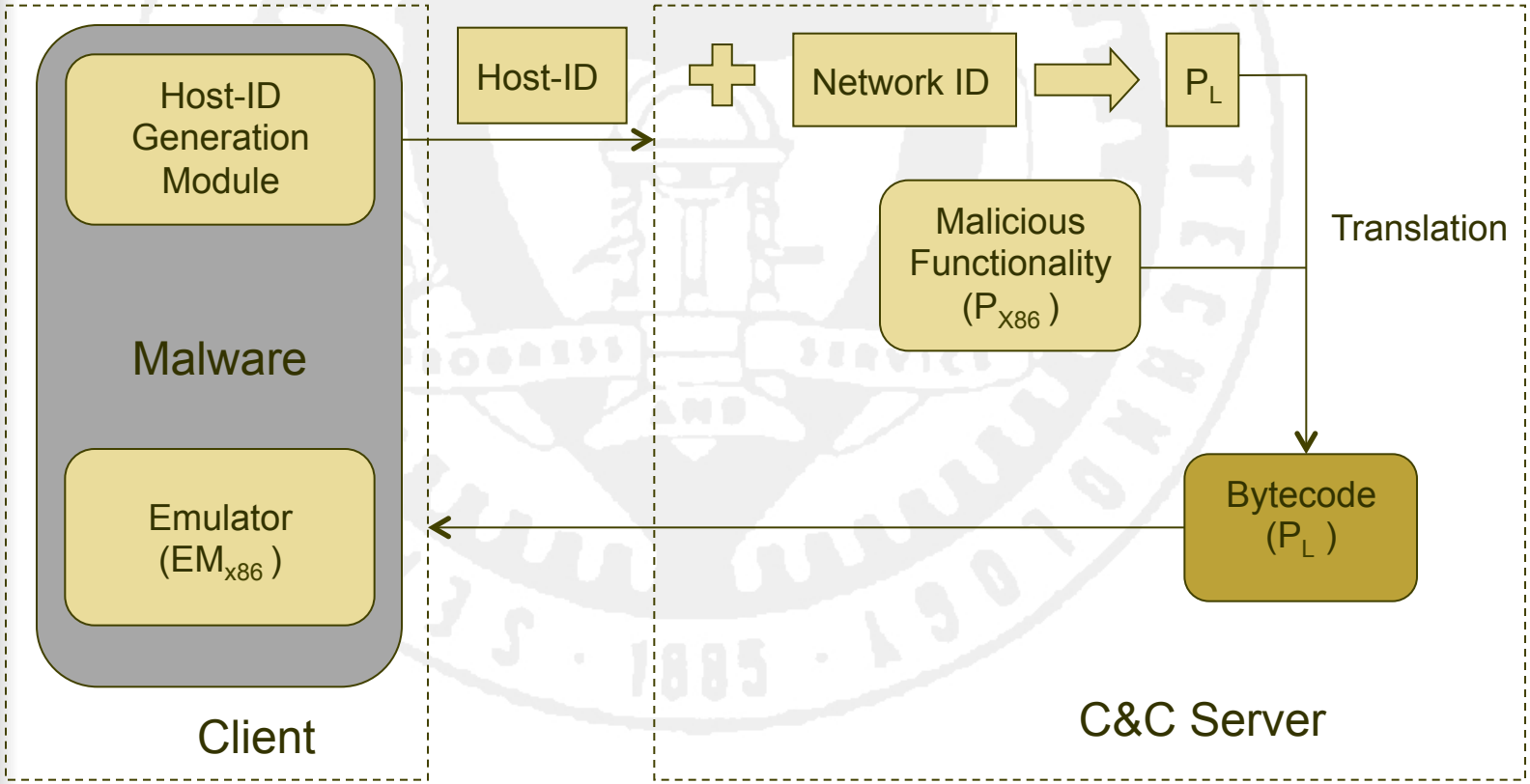
- **Advantages**
  - **Protections of Modern Cryptography**
    - **Knowledge of how key is derived does not affect the integrity of the protection**
  - **Sample Independence**
    - **Intelligence collected from one malware instance provides no advantage in analyzing another**

Instruction Set Localization

- **Why ISL?**
  - **Pure host-based protection is not sufficiently resistant to forgery**
- **Goal of ISL**
  - **Use C&C server to "authenticate" malware client based on both host and network identity**
  - **Decouple malicious functionality to prevent offline analysis**

ISL Cont'd

- **Replace random instruction set with instruction set bound to the host**



ISL Cont'd

- **Prototype Network ID**
  - **Geo-location**
    - **Granularity of state/province level (IP address is not stable)**
      - **Permits certain level of mobility**
  - **Autonomous System Number (ASN)**
    - **Geo-location may be outdated or incorrect**
  - **Collected at C&C**
    - **Considered intractably difficult to forge**

ISL Cont'd

- **Alternative to Unique Instruction Sets**
  - **Instruction set derivation is not trivial**
  - **Use *task decryption key***
    - **Assigned when the malware instance is delivered to the host**
    - **Encrypt bytecode tasks using the unique ID (the key derived from host ID and network ID)**
      - **KDF = HMAC(unique ID), or keyed hash, with the secret key kept at C&C server**

ISL Cont'd

- **Advantages**
  - **More extensible**
    - **Malware Platform-as-a-Service**
  - **Behavior identification is complicated**
    - **The HIE protected binary contains no malicious behaviors**
  - **Resistant to analysis and tracing**
    - **Offline analysis is impossible**
    - **Unless the analyst can correctly mimic the host and network environment, tasks will not decrypt/execute**

Flashback

- Propagated in part by drive-by downloads
- Payload is only intermediate agent
  - Agent gathers hardware UUID, submits request to C&C for full version
  - Hardware UUID hashed (MD5), hash used as decryption key to RC4 stream cipher
  - Full version will only run on host with same hardware UUID

Discussion

Operational Security

- **Both HIE and ISL use modern cryptography**
  - **Same environment must be provided for successful analysis**
  - **Without access to original environment, entire key space must be searched**
    - **Key space can be of arbitrary size**
  - **Some configurations may be impossible to duplicate**

Operational Security Cont'd

- **HIE and ISL are insensitive to analysis techniques**
  - **General knowledge of these techniques does not compromise protections offered**
  - **Granularity of analysis used does not affect protections**
  - **Protections can be broken only if the configuration parameters of the original execution environment are matched**

Potential Countermeasures

- **Analyze malware on the original infected host**
  - **Approach would require allowing otherwise blocked suspicious/known malware to execute on a legitimate system**
    - **Could impact business operations and continuity**
    - **Would have complex legal and privacy implications**
- **Use high-interaction honeypot**
  - **Bind malware to analysis environment by replicating compromise circumstances**
    - **Inefficient**
    - **Bound samples will comprise only a small portion of all collected samples**

Countermeasures Cont'd

- **Collect and duplicate host and network environment information**
  - **Depending on the information, may have privacy and policy problems**
  - **Duplicating network identifier requires analysis system deployment on an unprecedented and globally cooperative scale**

Countermeasures Cont'd

- **Collect and duplicate only host identifier, record and replay the network interaction in separate environment**
  - **Without small additional protection, could bypass ISL**
  - **Mitigated by using SSL/TLS to encrypt the C&C channel**

Countermeasures Cont'd

- **Employ allergy attack**
  - **Make the information used by HIE and ISL unstable**
    - **For example, change MAC address, username, SID for every program invocation**
    - **Malware would not execute correctly successfully on the infected host**
  - **Would affect a variety of legitimate software**
  - **Success would depend on the willingness of users to accept security over usability**

Conclusion

- **Historically, malware has been "analysis environment aware"**
- **Malware can be "analysis environment oblivious", and very likely to be**
  - **Flashback Malware**
- **Future work must mitigate these protections or more importantly, examine alternatives to threat detection and analysis**


THE GEORGIA INSTITUTE OF TECHNOLOGY
PROGRESS AND SERVICE
SEAL 1885

Thanks