Forensic Analysis for Epidemic Attacks in Federated Networks

Yinglian Xie, Vyas Sekar, Michael K. Reiter, Hui Zhang
Carnegie Mellon University

Presented by Gaurav Shah
(Based on slides by Yinglian Xie and Michael Hsiao)
Forensic Analysis for Epidemic Attacks in Federated Networks

Yinglian Xie, Vyas Sekar, Michael K. Reiter, Hui Zhang
Carnegie Mellon University

Presented by Gaurav Shah
(Based on slides by Yinglian Xie and Michael Hsiao)
Internet Attacks
Internet Attacks

Worm Infection
Internet Attacks

Worm Infection
Internet Attacks

Worm Infection
Internet Attacks

Worm Infection
Internet Attacks

Worm Infection
Internet Attacks

Worm Infection

Distributed DoS
Internet Attacks

Worm Infection

Distributed DoS
Internet Attacks

Worm Infection

Distributed DoS
Many More Potential Attacks

- Different time scales
  - Seconds, hours, days
- Different means
  - Various security exploits
- Different propagating methods
  - Random scan, hit list scan

- Potentially infect a large number of nodes before an “explicit” or “real” attack
The Structure of Attacks

- Modern attacks are multi-level
  - Large scale: difficult to defend
  - Hidden trail: difficult to identify initial launch point
Existing Approaches

- Firewall
- Intrusion detection
  - Identify compromised hosts
  - [Paxson 98], [Roesch 99], [Staniford 96], [Jung 04]
- IP traceback
  - Trace to the source of a packet
  - [Snoeren 01], [Savage 00], [Bellovin 01], [Burch 00]

**Build better perimeter or find only the lowest level of activities**
Existing Approaches

- **Firewall**
- **Intrusion detection**
  - Identify compromised hosts
  - [Paxson 98], [Roesch 99], [Staniford 96], [Jung 04]
- **IP traceback**
  - Trace to the source of a packet
  - [Snoeren 01], [Savage 00], [Bellovin 01], [Burch 00]

*Build better perimeter or find only the lowest level of activities*
Existing Approaches

- Firewall
- Intrusion detection
  - Identify compromised hosts
  - [Paxson 98], [Roesch 99], [Staniford 96], [Jung 04]
- IP traceback
  - Trace to the source of a packet
  - [Snoeren 01], [Savage 00], [Bellovin 01], [Burch 00]

*Build better perimeter or find only the lowest level of activities*
Network Forensics

- Keep communication records
- Permit post-mortem analysis of patterns across network and time

Scope: Internet and Intranet
  - Correct weak points in a network perimeter
  - Deter future similar attacks
Attacks utilize indirection

- Attackers and victims must communicate for the attack to propagate
  - Independent of attacks
  - Visible to the network

- Globally correlate the communication events
  - General across all types of attacks
  - Applicable to different attack propagation speeds
  - Possible with/without attack signature
Host Contact Graph and Attack Trees

Figure 1: Example of host contact graph showing the communication between hosts. Attack edges are shown as arrows in black (both solid and dashed). Filled nodes correspond to hosts in an infected state.
Feasibility of Network Forensics

- Need to collect network flow information
- Data to be collected
  - around 10Gbps
  - 4.5TB for one hour of data
- Given complete information, can we find needles in the haystack?
  - Host contact graph is large and noisy
  - Algorithms to identify global correlation
Reconstruction Methods

- Globally correlate local observations

**Step 1:** Assign initial $P(\text{involved})$
  - E.g., local IDS

**Step 2:** Identify the most likely paths
  - Maximum likelihood estimate
  - Random walk sampling

**Step 3:** Refine $P(\text{involved})$
  - A feedback loop

B, E are likely involved: given pattern, so are C, F
Random Moonwalks

- Observation: many attacks have a tree structure
- A random walk on the host contact graph:
  - Start with an arbitrarily chosen flow
  - Pick a next step flow randomly to walk backward in time
- Intuition
  - Walks converge to higher levels of the host attack tree
  - Top level causal infection flows emerge as high frequency flows
- Theoretical study
  - Random walk is effective in finding the initial causal flows for fast attacks
Random Moonwalks

- Observation: many attacks have a tree structure
- A random walk on the host contact graph:
  - Start with an arbitrarily chosen flow
  - Pick a next step flow randomly to walk backward in time
- Intuition
  - Walks converge to higher levels of the host attack tree
  - Top level causal infection flows emerge as high frequency flows
- Theoretical study
  - Random walk is effective in finding the initial causal flows for fast attacks
Random Moonwalks

- Observation: many attacks have a **tree** structure

- A random walk on the host contact graph:
  - Start with an arbitrarily chosen flow
  - Pick a next step flow randomly to walk **backward in time**

- Intuition
  - Walks converge to higher levels of the host attack tree
  - Top level causal infection flows emerge as high frequency flows

- Theoretical study
  - Random walk is effective in finding the initial causal flows for fast attacks
Random Moonwalks

- Observation: many attacks have a **tree** structure
- A random walk on the host contact graph:
  - Start with an arbitrarily chosen flow
  - Pick a next step flow randomly to walk **backward in time**
- Intuition
  - Walks converge to higher levels of the host attack tree
  - Top level causal infection flows emerge as high frequency flows
- Theoretical study
  - Random walk is effective in finding the initial causal flows for fast attacks
Random Moonwalks

- Observation: many attacks have a tree structure
- A random walk on the host contact graph:
  - Start with an arbitrarily chosen flow
  - Pick a next step flow randomly to walk backward in time
- Intuition
  - Walks converge to higher levels of the host attack tree
  - Top level causal infection flows emerge as high frequency flows
- Theoretical study
  - Random walk is effective in finding the initial causal flows for fast attacks
Random Moonwalks

- **Observation:** many attacks have a tree structure
- **A random walk on the host contact graph:**
  - Start with an arbitrarily chosen flow
  - Pick a next step flow randomly to walk backward in time
- **Intuition**
  - Walks converge to higher levels of the host attack tree
  - Top level causal infection flows emerge as high frequency flows
- **Theoretical study**
  - Random walk is effective in finding the initial causal flows for fast attacks
Random Moonwalks

- Observation: many attacks have a **tree** structure
- A random walk on the host contact graph:
  - Start with an arbitrarily chosen flow
  - Pick a next step flow randomly to walk **backward in time**

- Intuition
  - Walks converge to higher levels of the host attack tree
  - Top level causal infection flows emerge as high frequency flows

- Theoretical study
  - Random walk is effective in finding the initial causal flows for fast attacks
Random Moonwalks

- Observation: many attacks have a tree structure
- A random walk on the host contact graph:
  - Start with an arbitrarily chosen flow
  - Pick a next step flow randomly to walk backward in time
- Intuition
  - Walks converge to higher levels of the host attack tree
  - Top level causal infection flows emerge as high frequency flows
- Theoretical study
  - Random walk is effective in finding the initial causal flows for fast attacks
Random Moonwalks

- Observation: many attacks have a **tree** structure
- A random walk on the host contact graph:
  - Start with an arbitrarily chosen flow
  - Pick a next step flow randomly to walk **backward in time**
- Intuition
  - Walks converge to higher levels of the host attack tree
  - Top level causal infection flows emerge as high frequency flows
- Theoretical study
  - Random walk is effective in finding the initial causal flows for fast attacks
Random Moonwalks

- Observation: many attacks have a **tree** structure
- A random walk on the host contact graph:
  - Start with an arbitrarily chosen flow
  - Pick a next step flow randomly to walk *backward in time*

- Intuition
  - Walks converge to higher levels of the host attack tree
  - Top level causal infection flows emerge as high frequency flows

- Theoretical study
  - Random walk is effective in finding the initial causal flows for fast attacks
Random Moonwalks

- Observation: many attacks have a tree structure
- A random walk on the host contact graph:
  - Start with an arbitrarily chosen flow
  - Pick a next step flow randomly to walk backward in time
- Intuition
  - Walks converge to higher levels of the host attack tree
  - Top level causal infection flows emerge as high frequency flows
- Theoretical study
  - Random walk is effective in finding the initial causal flows for fast attacks
Random Moonwalks

- Observation: many attacks have a **tree** structure
- A random walk on the host contact graph:
  - Start with an arbitrarily chosen flow
  - Pick a next step flow randomly to walk **backward in time**
- Intuition
  - Walks converge to higher levels of the host attack tree
  - Top level causal infection flows emerge as high frequency flows
- Theoretical study
  - Random walk is effective in finding the initial causal flows for fast attacks
Random Moonwalks

- Observation: many attacks have a **tree** structure
- A random walk on the host contact graph:
  - Start with an arbitrarily chosen flow
  - Pick a next step flow randomly to walk **backward in time**

- Intuition
  - Walks converge to higher levels of the host attack tree
  - Top level causal infection flows emerge as high frequency flows

- Theoretical study
  - Random walk is effective in finding the initial causal flows for fast attacks
Structure of the Selected Flows

- Top Frequency Flows display a *tree-like* structure
Federated Forensics

- **Administrative Domain (AD)**
  - a network under a single administrative authority
  - possesses an independent traffic monitor

- **Network Forensic Alliance (NFA)**
  - a collaborative effort involving multiple ADs to provide network-wide & localized forensic capabilities
Distributed Random Moonwalks

- Each AD performs its own *local random moonwalk*
- ADs do *cross-domain exchanges*
  - For flows whose source lies in another AD
  - Other AD might start its own local moonwalk
- Post-processing
  - Identify top frequency flows locally
  - Exchange and identify top flows globally
Distributed Random Moonwalks

- Parameters:
  - Maximum number of AD hops to traverse (H)
  - Sampling Window Size
  - Local walk-length d(i)
  - Traffic Normalization factor
Partial Deployment

- Not all ADs might participate
- Different Strategies to continue
  - Discard
  - Random selection
  - Routing-path based selection
  - Routing-horizon based selection
Evaluation

- Performance of distributed approach (accuracy)
- Overhead of federated forensics
- Is architecture incentive-compatible?
- Effect of partial deployment on performance
Performance and Overhead

- Detection accuracy closely matches that of a centralized random moonwalk
  - Increases with increase in worm propagation rate

- Communication overhead increases with increase in AD-hops (H)
  - So does accuracy
  - Overhead is of the order of $10^{-3}$ fraction of total traffic records
Incentive Compatibility

- Causal Flow detection accuracy increases with collaboration
  - Can detect more initial infected hosts
  - Can detect more initial causal flows along the edge
- Large ADs gain more by collaboration
Partial Deployment

- Important for large ADs to participate
  - Accuracy drop can be significant
  - Much less significant impact if worm is local-preferential scanning
Issues and Limitations

- Slow-rate attacks
- Server-targeted attacks
- Latent-period varying attacks
- Topology-aware attacks
- Information leakage attacks
- Mis-information and information-hiding attacks
Summary

- Distributed Random Moonwalks can be useful in attack identification and reconstruction
- Many Challenges
  - Storage and Archival
  - Anonymization of network traces and flow information for exchange
  - Query Management
    - Distributed Query Infrastructure?
  - Vulnerability to different attack strategies