Self-Propagating Worms

**Fast Spreading Speed**
- Breaking out on July 19, 2001, Code Red worm version 2 infected more than 359,000 machines within 14 hours
- Breaking out on January 25, 2003, Slammer worm infected at least 75,000 hosts within only 10 minutes

**Powerful Destruction**
- Breaking out on September 18, 2001, Nimda worm infected more than 2 million machines
- The total cost of Code Red worm, only measured in lost productivity in network services, is estimated at $2.6 billion

**Payload Polymorphism**
- In order to escape from most IDSes, some worms, such as Witty and Apache-Knacker, can change their payloads by encrypting each worm instance and/or randomizing filler text

Current Content-Based IDSes

**Earlybird by UCSD**
- Can NOT detect polymorphic worms with common bytes shorter than 40 bytes
- Sampling (1/64) and estimation lead to misdetecting worms

**Autograph by CMU**
- Does not work for UDP-based worm (like Slammer) and email borne worms (like MyDoom)
- Non-overlapping Rabin fingerprinting, the partition of packets is too sensitive to the predetermined breakmark

**Polygraph by CMU & Intel**
- No method developed for classifying packets as innocuous or suspicious
- Highly complex computation, impossible for online implementation

High-Speed Worm Defense by Using Both Header and Payload

**System Structure**

- **Multidimensional Traffic Clustering and Classification**
- **Packet Header Hashing**
- **Multidimensional Clustering**
- **Suspicious Cluster Identification**

- **Suspicious Cluster Pool**
- **Innocuous Packet Pool**

- **Worm Signature Extraction and Evaluation**
- **Worm Signature Extraction**
- **Worm Signature Evaluation**

- **Worm Signatures**

**Payload-Based Worm Containment**

**Pipelined Implementation**
- 3T delay (T could be as small as 1 second): hash packets arriving in [0, T]; perform mining in [T, 2T]; collect suspicious packets and extract signatures in [2T, 3T]
- Requiring 50 to 60 MB memory without packet sampling
- Can handle the link with load up to 800Mbps in real-time by software

**Multidimensional Traffic Mining**
- Frequent item set mining applied to network traffic flows, based on the packet header 5-tuple (source IP, destination IP, source port, destination port, protocol)
- Using top-down method to build up a multidimensional tree and only mining significant (with traffic volume larger than a threshold) and suspicious clusters

**Suspicious Cluster Identification**
- Two Criteria for Defining a Suspicious Cluster:
  - Its traffic volume is larger than a threshold (e.g., 1%)
  - Its source or destination IP dispersion/cardinality is larger than a threshold. In other words, the number of different source or destination IP addresses involved in a cluster is larger than a threshold (e.g., 30)

**Worm Signature Extraction**
- Building a generalized suffix tree for each suspicious cluster, and extracting signatures only from a small part of packets
- With time and space linear in the length of each suspicious cluster
- Jointly considering length, frequency and false positive to improve the accuracy of signatures
- Easily modifying rules to search for multiple signature descriptions of varying length range, with little increase in complexity

This work is supported by NSF and DHS.