

DHS STTR: Botnet Detection and Mitigation

CATCH

Washington, DC – March 3,4 2009

A Combined Fusion and Mining Strategy for Detecting Botnets

Sonalysts, Owen McCusker PI
UCONN, Dr. Aggelos Kiayias

SBIR DATA RIGHTS Contract No.: NBCHC070124

Contractor Name: Sonalysts, Inc.

Contractor Address: 215 Parkway N., Waterford, CT. 06815

Expiration of SBIR Data Rights Period: 14 Jan 2011

The Government's rights to use, modify, reproduce, release, perform, display, or disclose technical data or computer software marked with this legend are restricted during the period shown as provided in paragraph (b)(4) of the Rights in Noncommercial Technical Data and Computer Software--Small Business Innovative Research (SBIR) Program clause contained in the above identified contract. No restrictions apply after the expiration date shown above. Any reproduction of technical data, computer software, or portions thereof marked with this legend must also reproduce the markings

The DMnet Team

DMnet: **D**etection and **M**itigation **N**etwork

Sonalysts, Inc	Data Fusion Development, and Research
University of Connecticut	Data Mining Research, High Performance Computing (HPC) Research
Dr. John McHugh	Subject Matter Expert
NUARI, Delta-Risk, LLC	Cyber Operational Research
IntelliVis	Human Computer Interface
CTC	HPC Research

Agenda

- Introduction
- Threat
- Needs
- Approach
- Current Status
- Conclusions
- **Questions**

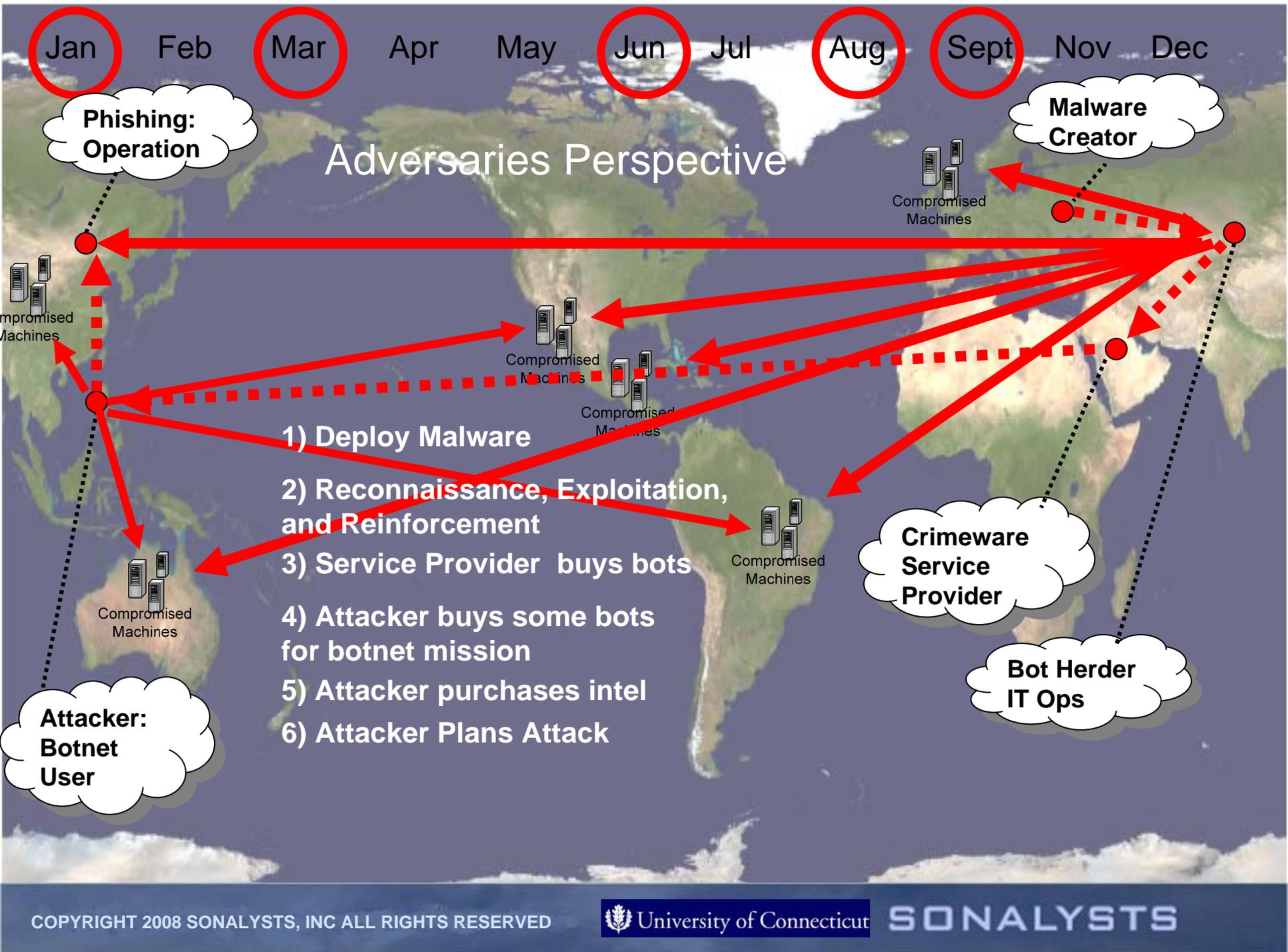


Introduction

- The Threat is Dynamic, Distributed and Multi-scale in time
- **Need** to understand both the technical and operational aspects of the threat
- Anticipate threats and their behaviors before the attack
- Provide actionable information allowing Enterprise to tune their defenses

Jan Feb Mar Apr May Jun Jul Aug Sept Nov Dec

Adversaries Perspective



**Phishing:
Operation**

**Malware
Creator**

Compromised
Machines

Compromised
Machines

Compromised
Machines

Compromised
Machines

1) Deploy Malware

**2) Reconnaissance, Exploitation,
and Reinforcement**

3) Service Provider buys bots

**4) Attacker buys some bots
for botnet mission**

5) Attacker purchases intel

6) Attacker Plans Attack

Compromised
Machines

**Crimeware
Service
Provider**

**Bot Herder
IT Ops**

**Attacker:
Botnet
User**

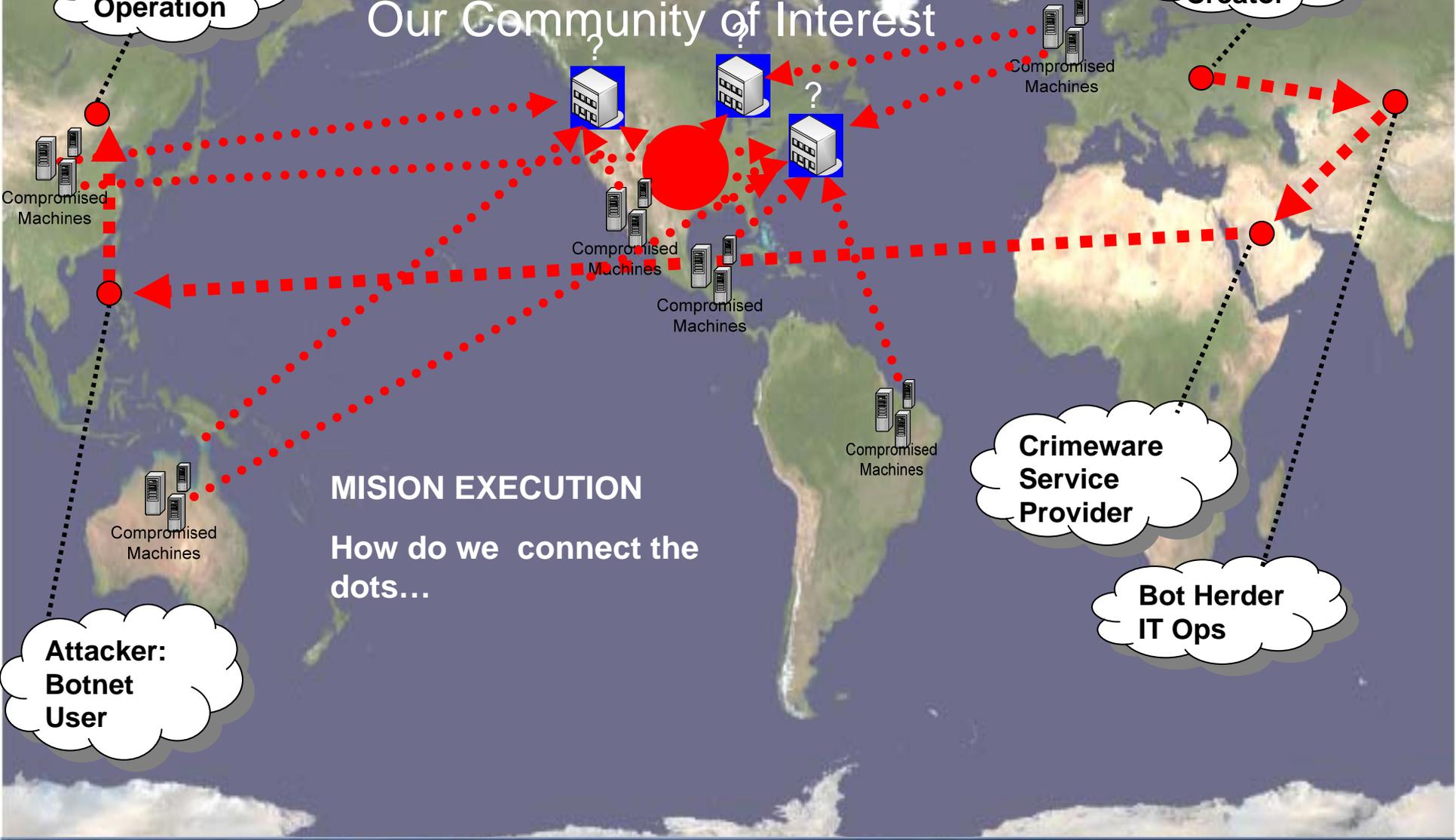
Compromised
Machines

Jan - Feb - Mar - Apr - May - Jun - Jul - Aug - Sept - Oct - Dec

Phishing: Operation

Malware Creator

Our Community of Interest



MISSION EXECUTION
How do we connect the dots...

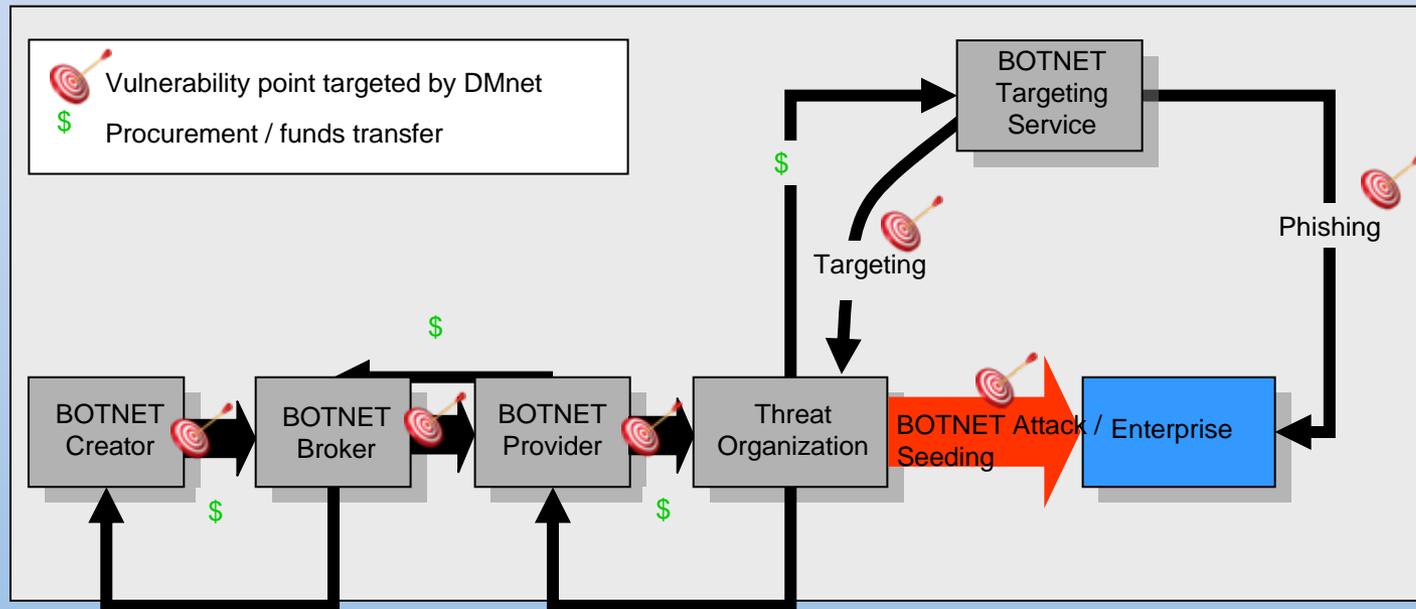
Attacker: Botnet User

Crimeware Service Provider

Bot Herder IT Ops

DMnet and the Needs

Detecting Botnets in their Global Supply Chain



- Extend the defense horizon from the perimeter
- Collaborate and Share
- Employ multiple sensor types
- Move away from analyzing raw data
- Combine forensics and fusion
- Provide **feedback** to mitigate and disrupt the supply chain

DMnet and the Needs

Both Near Real-time and Forensics Methodologies

- Forensic analysis for threat discovery, and offline learning
 - Share discovered profiles with fusion system
- Near Real-time reactivity and anticipation for threat identification, and online learning
- How to bring together both needs in one system?
 - Sharing raw data is expensive

Approach

Related Work

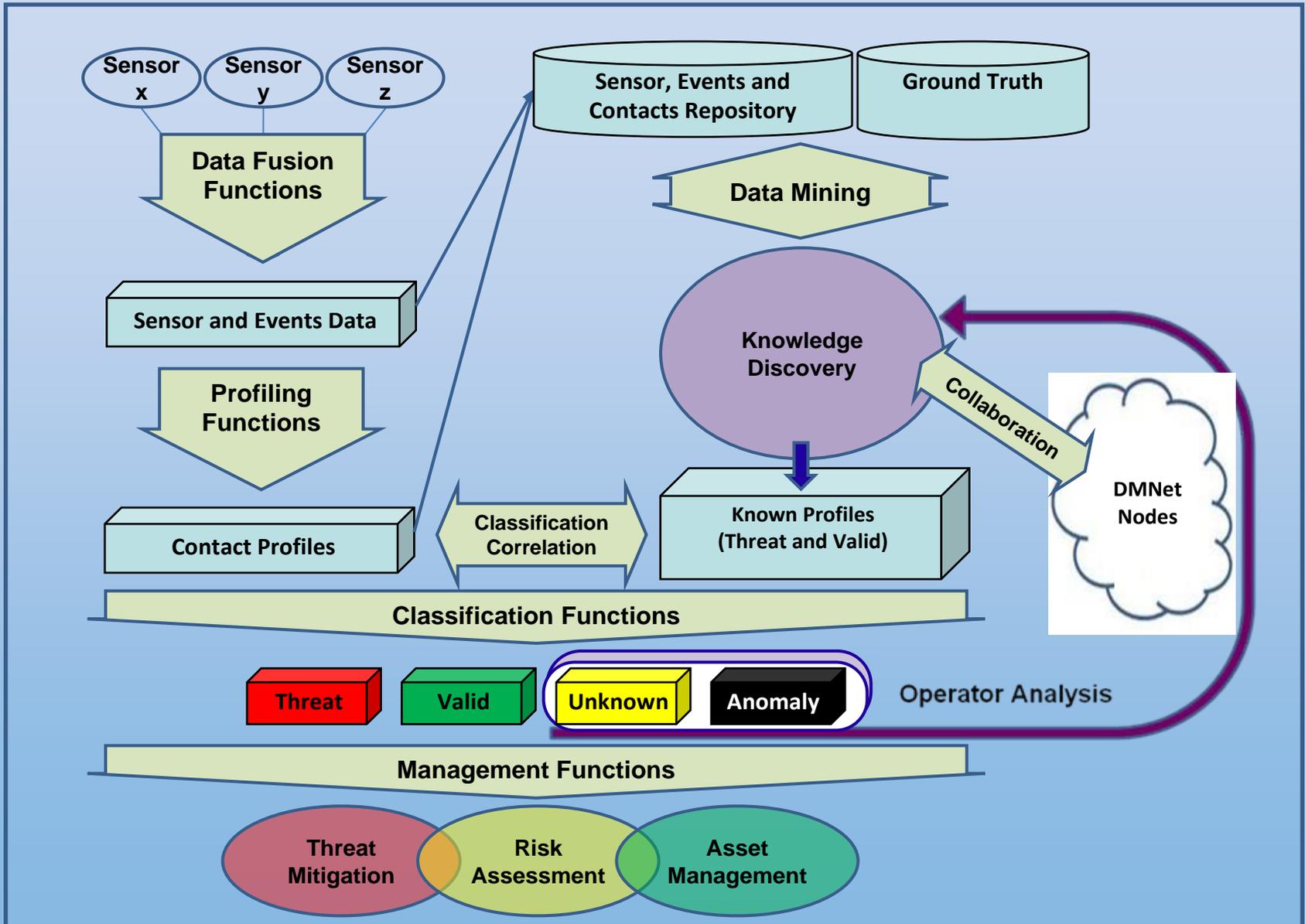
- Botnet Detection
 - Botsniffer, Bothunter
- Hybrid Systems
 - Prelude IDS
- Knowledge Discovery and Adaptability
 - MADAM ID - uses associative rules
 - SRI IDES - expert system
- Intrusion Detection
 - SNORT, Bro, SRI IDES
 - Worminator - distributive and collaborative

Approach

Key DMnet Concepts

- Hybrid meta-level sensor
- Leverages Behavioral and structural profiling
 - Two stage data reduction
 - Creates a feature space
- Knowledge Discovery through mining and collaboration
- Fusion is driven by profiles discovered in mining engine
- Threat-centric - focus on threat behavior

DMnet Architecture



Approach

Extracting Network Profile Features

- System tracks network objects
- Process raw data to extract features
- Features can be measured passively and do not require deep packet inspection
- Basic Features sample
 - Port scatter
 - Source/Sink Data Transfer
 - Source/Sink Packets
 - TCP/UDP/ICMP/Other Bytes
 - TCP/UDP/ICMP/Other Packets
 - TCP Work Load
 - Social Index
 - Packet Inter-arrival Time
- Time-series based
- Entropy based
 - Communication Entropy

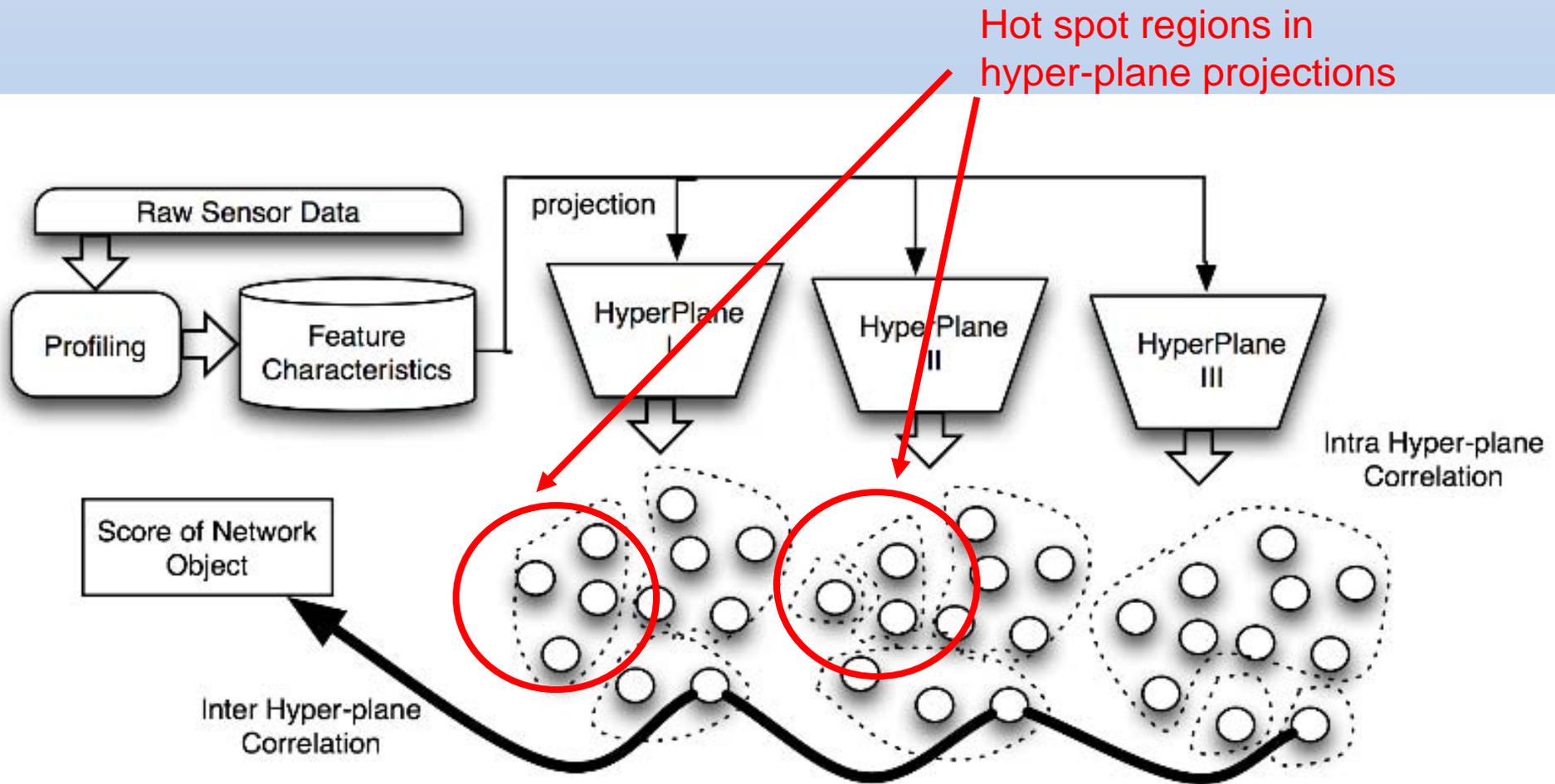
Approach

Classification and Correlation

- Identifies network patterns by clustering object features within single hyperplane
- Correlation are made across multiple hyperplanes
- Threat Scoring
 - Hosts are scored based on associations between other hosts
 - Strong and weak partners are associated through training data e.g. ground truth
 - Hosts exhibiting botnet behavior, high sociability with other strong partners are assigned a higher bot score

Approach

Classification and Correlation Framework



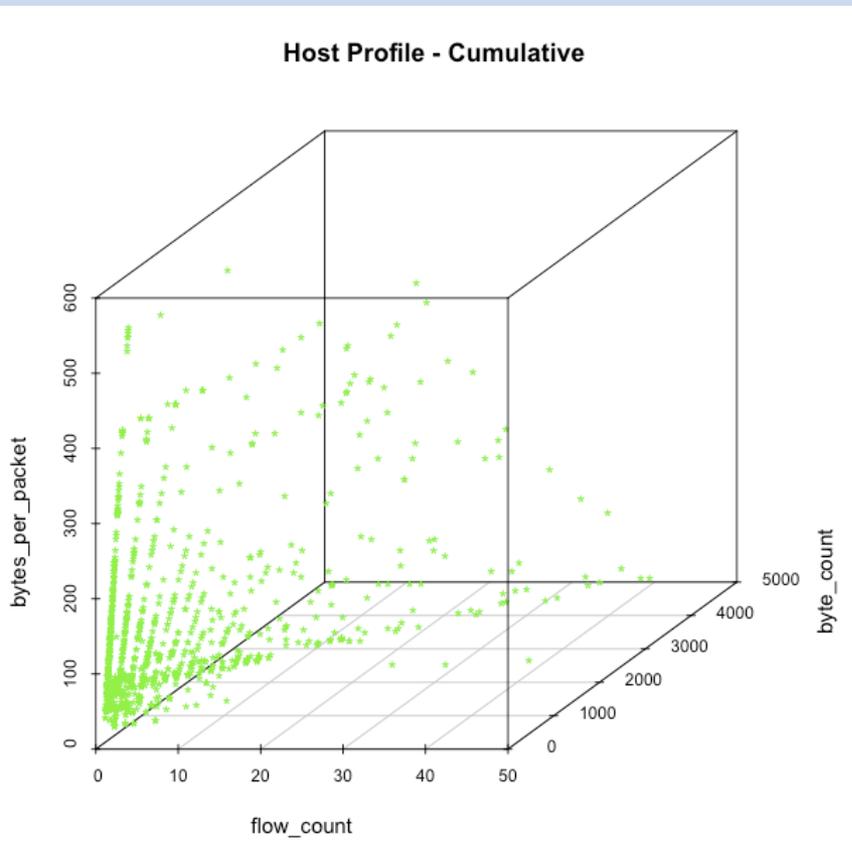
Where We Are Today

Data Fusion Framework

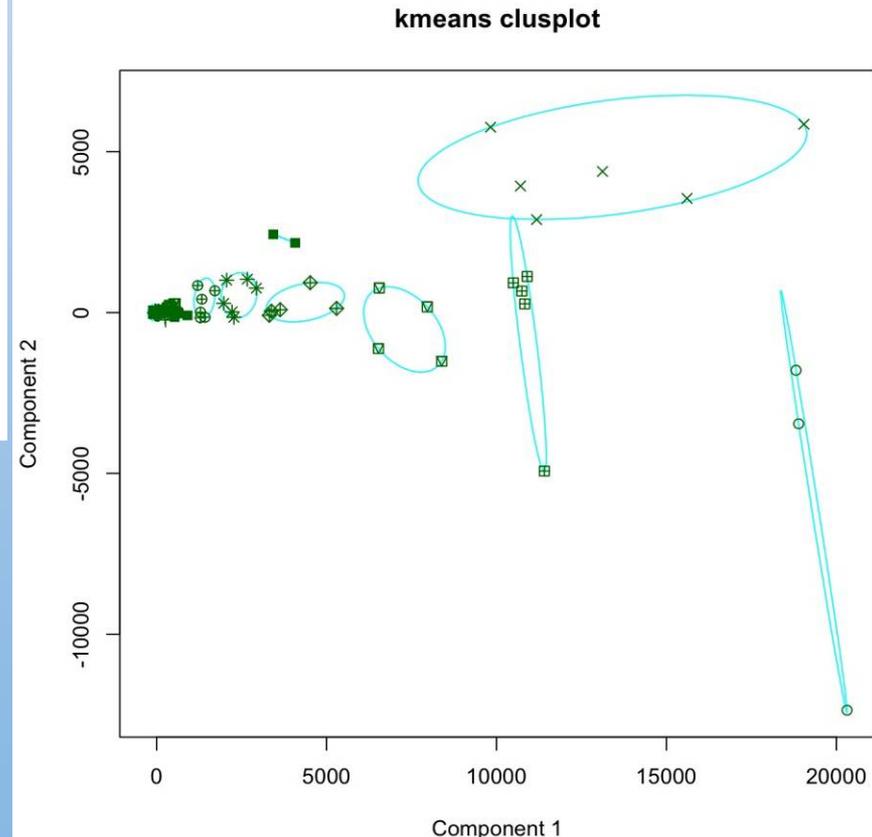
- Sensor Management Framework
 - Network Flow SiLK, NIDS (not profiling)
- Profiling Framework
 - Finished development in January 2009, working with Dr. John McHugh on profiling approaches
- Classification/Correlation Analysis
 - Integrated “R the statistical package” for analysis
- Operational Analysis
 - Worked with NUARI, and Delta-Risk, LLC
- High Performance Computing Investigation
- Web Mashup – demonstrating on captured data
- HMI - Finished up Con-opts with IntelliVis

Comprehension Analysis

- K-means cluster of contacts profiles



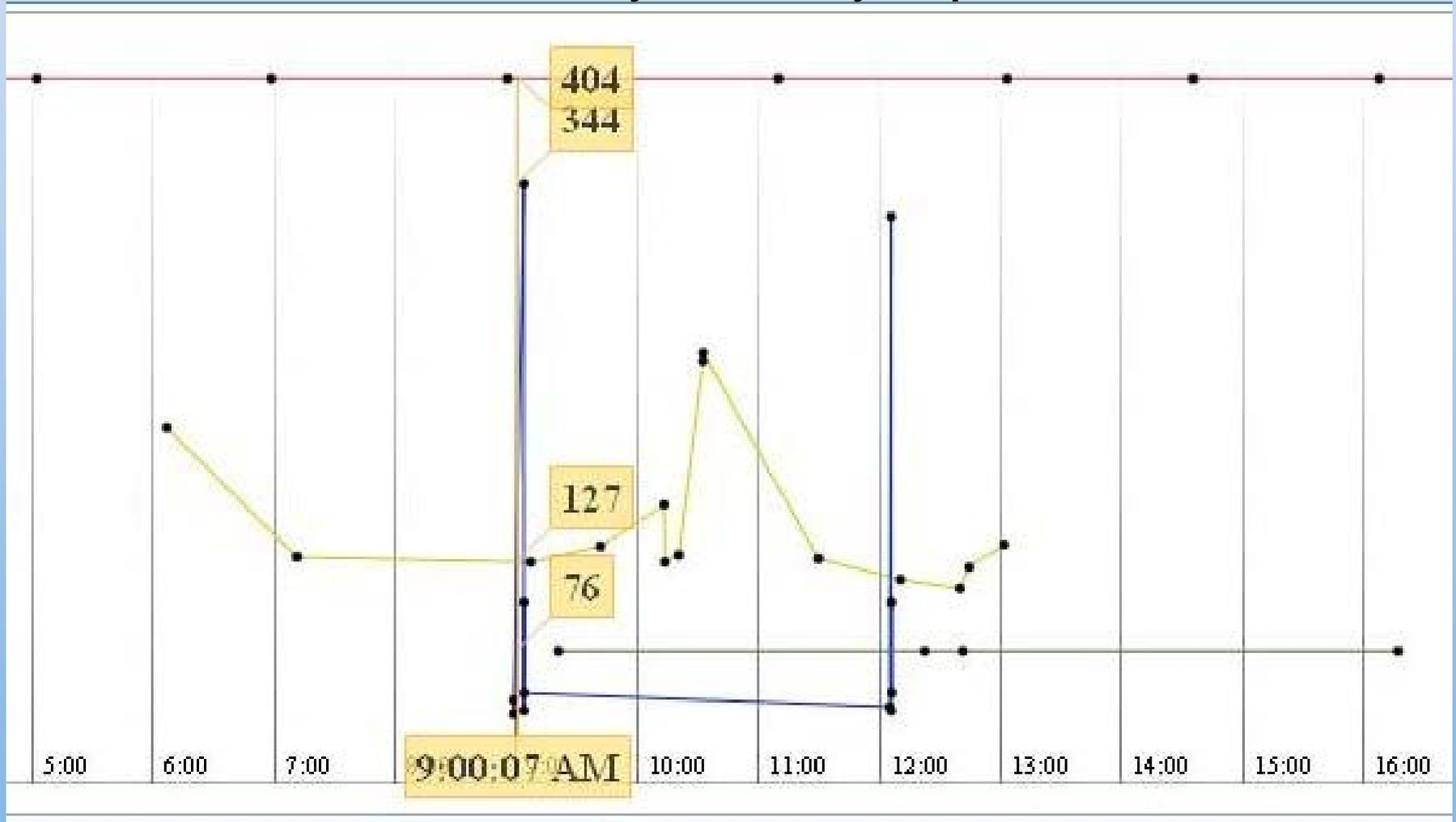
- Scatter plot showing relationships between hosts



Component 1
Component 2
These two components explain 100 % of the point variability.

Web Mashup

Time series analysis of byte/packet ratio



Where We Are Today

Data Mining Framework

- Combined Mining and Fusion Framework
 - Classification and Correlation Model
 - Model Components
 - Classification and Correlation of Network Features
 - Social Factor and Bot Score
 - Network Profile Features
- Cyber Test bed
 - Experimental Evaluation
 - “Ground Truth” collection
 - Results

Where we are Today

Threat Discovery Challenges

- High dimensionality
 - 2 Phases:
 - On raw data using profiling functions
 - On profiled data leveraging hyperplane
- Correlation across Hyper-planes
 - Rules mining
 - Separate hyper-planes are combined
- Adaptability
 - Driven by profiling and data reduction strategies

Where we are Today

Social Factor and Bot Score

- Two types: First degree score: $S_O(1) = \sum_{P, S: O \in S} wt_P(S, L_P)$

- Takes into account the weights of the labeled sets that the object belongs to across all hyperplanes

$$S_O(2) = \sum wt_P(S, L_P) \cdot wt_{P'}(S', L_{P'}) \cdot social(O, P, P')$$

- Second degree score:

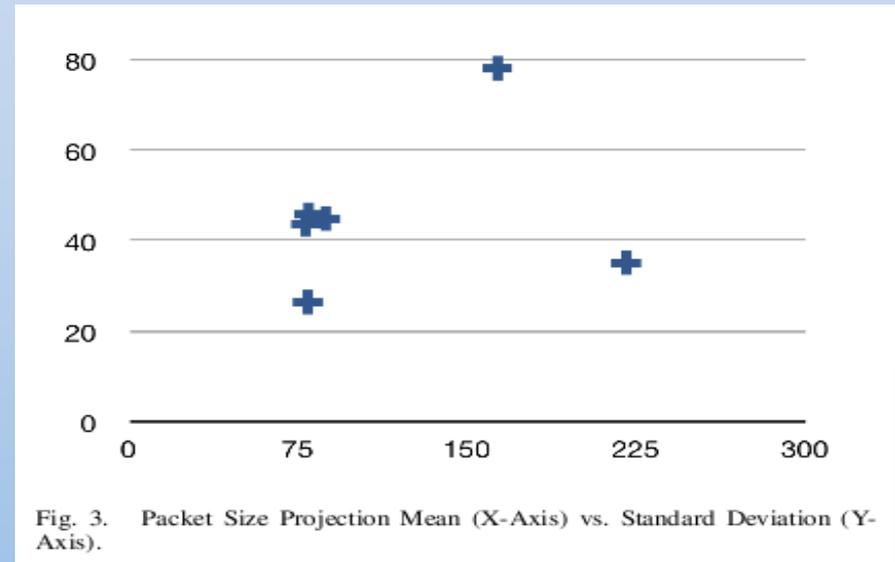
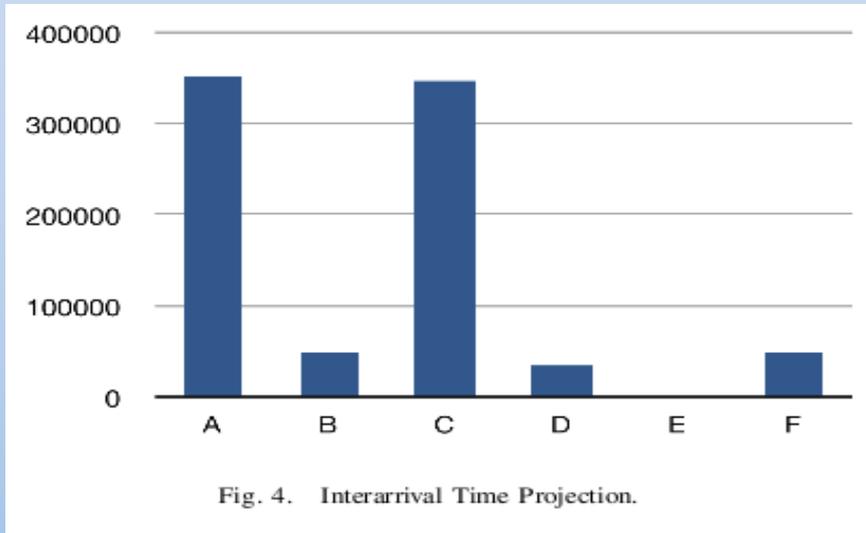
- Factors in sociability of certain pairs of hosts across hyperplanes

- The Bot Score:

$$S_O = w_1 \cdot S_O(1) + \dots + w_k \cdot S_O(k)$$

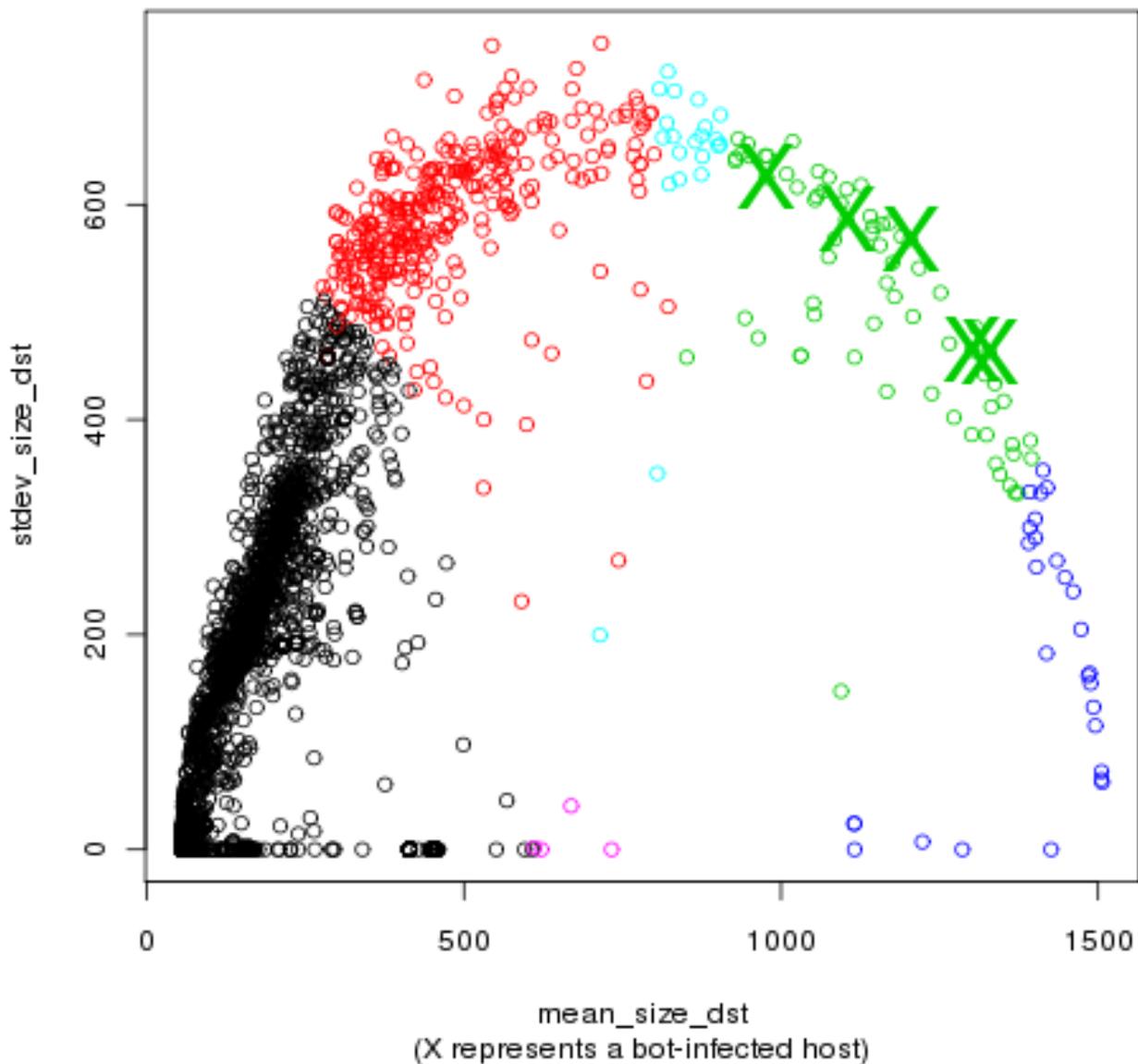
- The score is based on the hyperplanes

Results – Two hyper-planes



Element	Correlation Score
A (Clean)	2.5
B (Bot)	23
C (Clean)	4
D (Bot)	23
E (Bot)	15
F (Bot)	23

Hosts Clustered by Network Feature



Conclusions

- Instead of focusing on the perimeter, and processing raw data
- Validating the hyperplane approach on a specific type of C&C based bots
- Developed an extensible software framework
- Created a testbed used to capture ground truth
- Set the ground work for a distributed, self-learning, detection capability

Conclusions

Technology Transfer

- Collaboration
 - University of Connecticut, NUARI,
- Licensing Agreement
 - Nondisclosure Agreement
 - Work with existing programs to mature technology
- Cooperative Research
 - University of Connecticut
- Future Applications
 - Insider Threat Problem
 - Integrated Threat Management
 - **Convergence Enabling Technology**



Questions?