Comprehensive Understanding of Malicious Overlay Networks

Homeland Security
Science and Technology

Cyber Security Division
2012 Principal Investigators' Meeting

October 10, 2012

Wenke Lee and David Dagon
Georgia Institute of Technology
wenke@cc.gatech.edu
404-808-5172

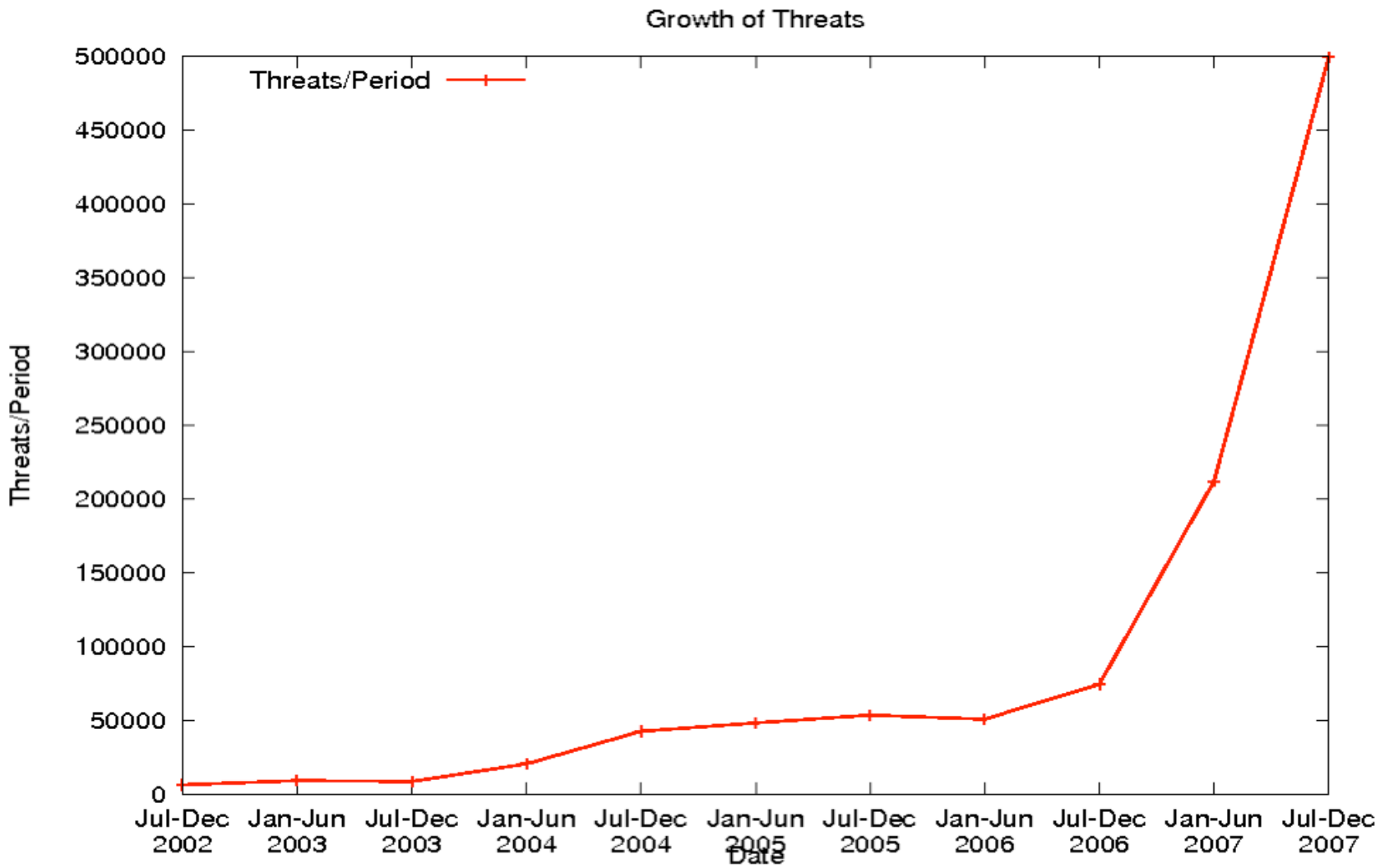
Roberto Perdisci, University of Georgia
April Lorenzen, Dissect Cyber
Paul Vixie, Internet Systems Consortium
Jody Westby, Global Cyber Risk LLC
Chris Smoak, GTRI
Matt Jonkman, Open Information Security Foundation

Background

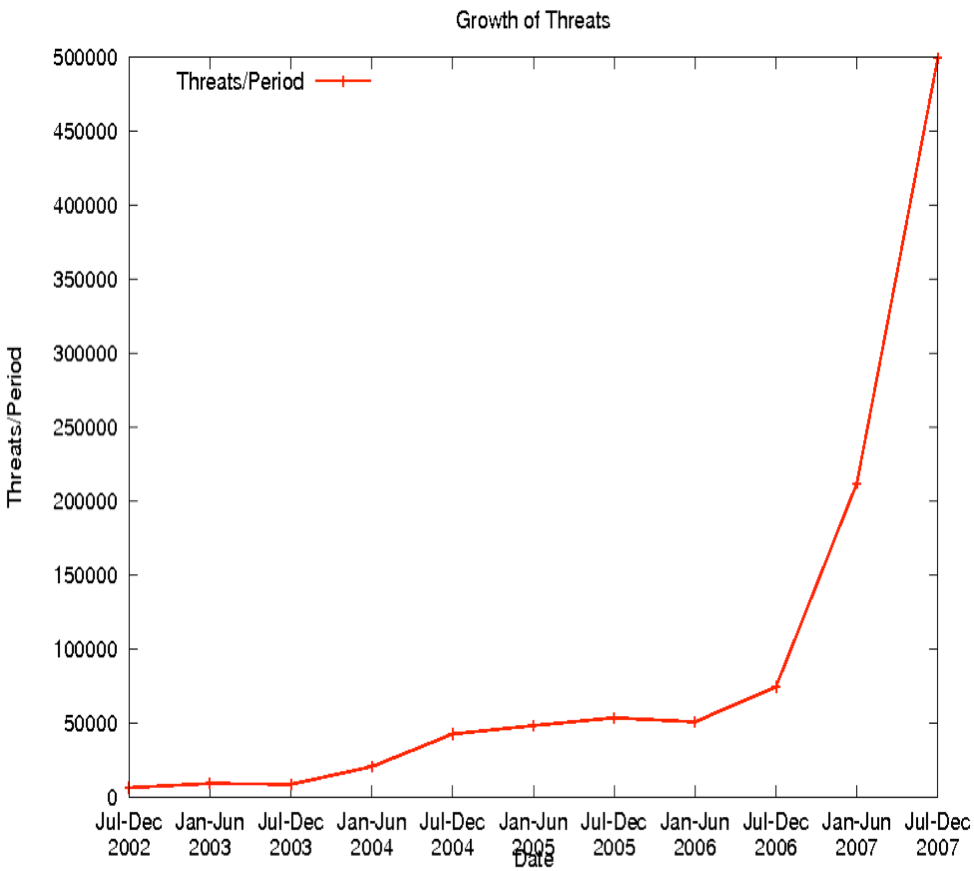
- Malware churn
  - Very short shelf life
  - Techniques: evasive packing; polymorphic malware; generative programming
  - Noted example: Storm botnet, June 2006 (new sample pushed on hourly-basis)

- A botnet is not merely a single binary. It is the overlay network of malicious infrastructure and supporting malware samples.

Malware Sample Growth

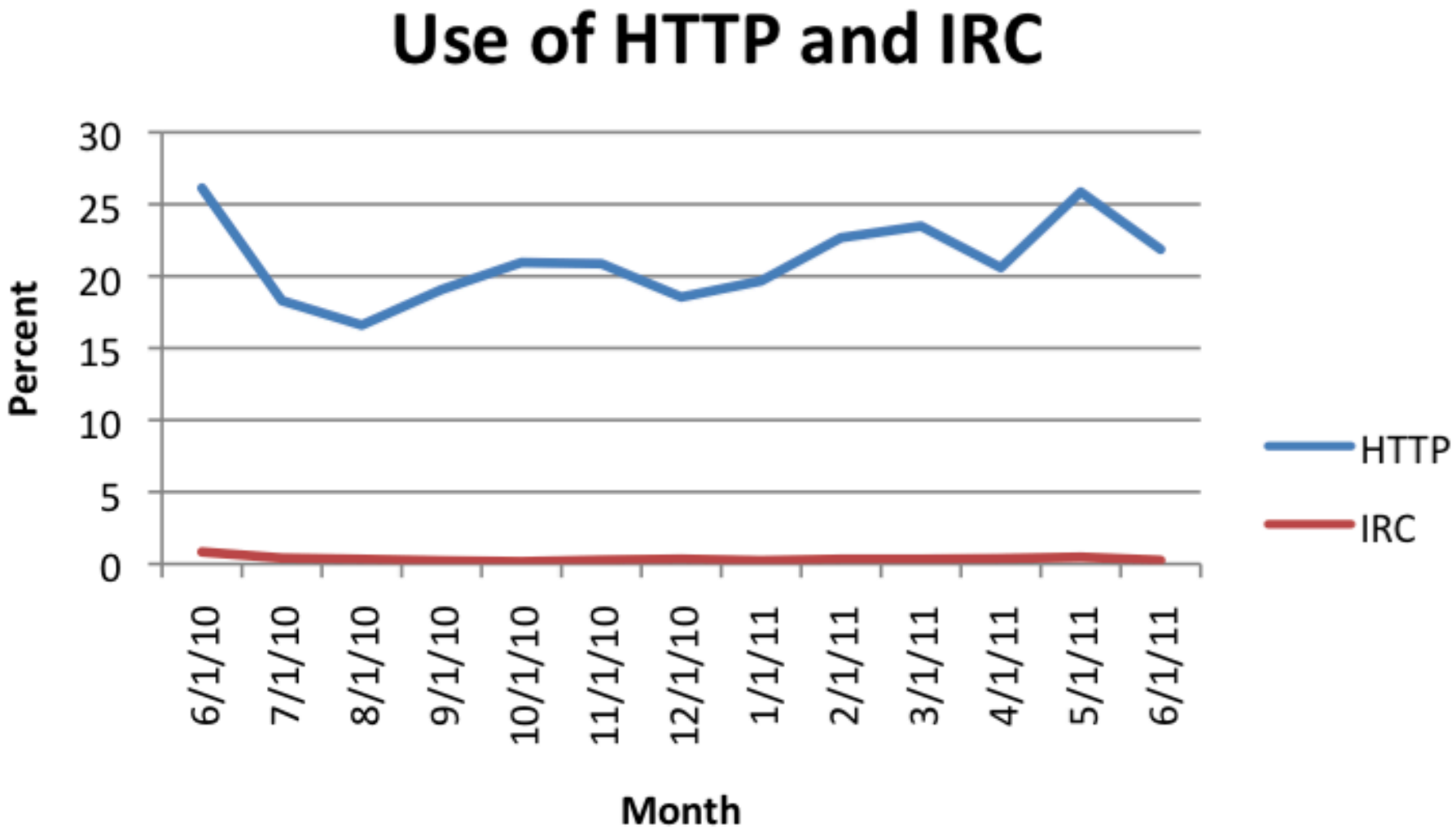


Malware Sample Growth

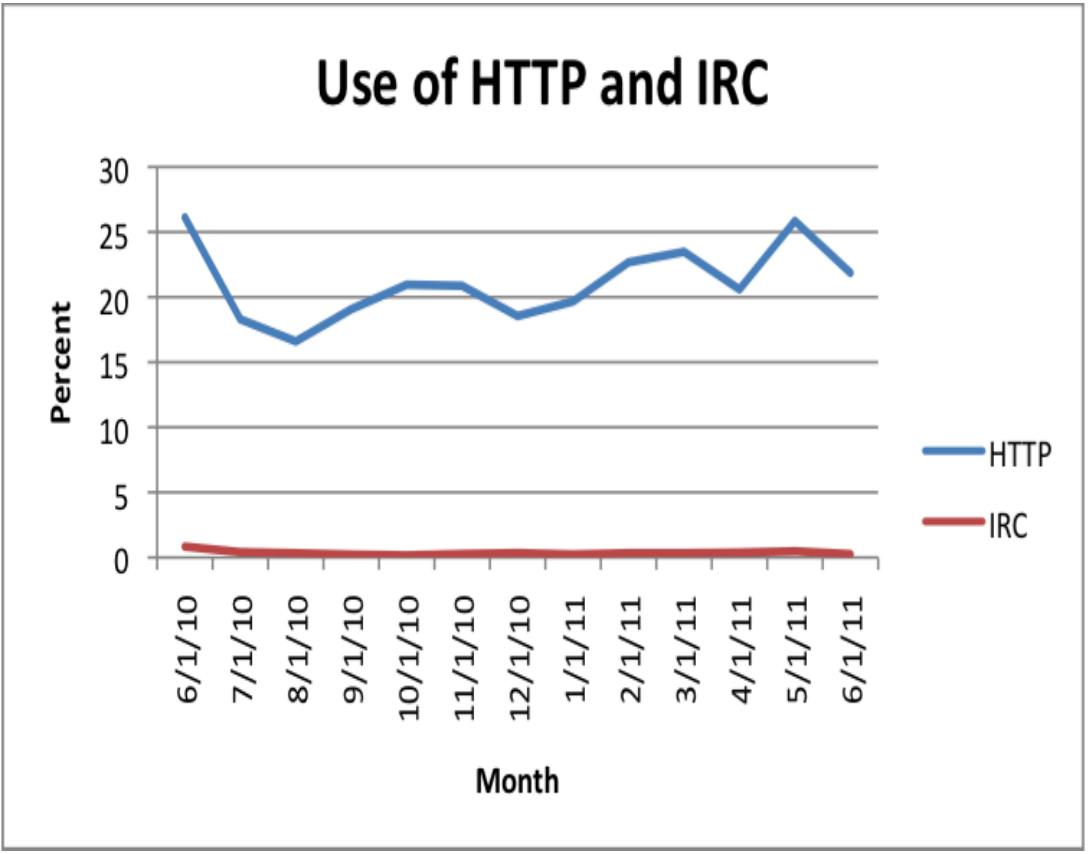


- Salient points:
  - Exponential growth
  - Within our team, about 50 million samples
  - The challenge is to analyze the clusters and collections of samples, not merely discrete samples.

Protocols Used in Malware

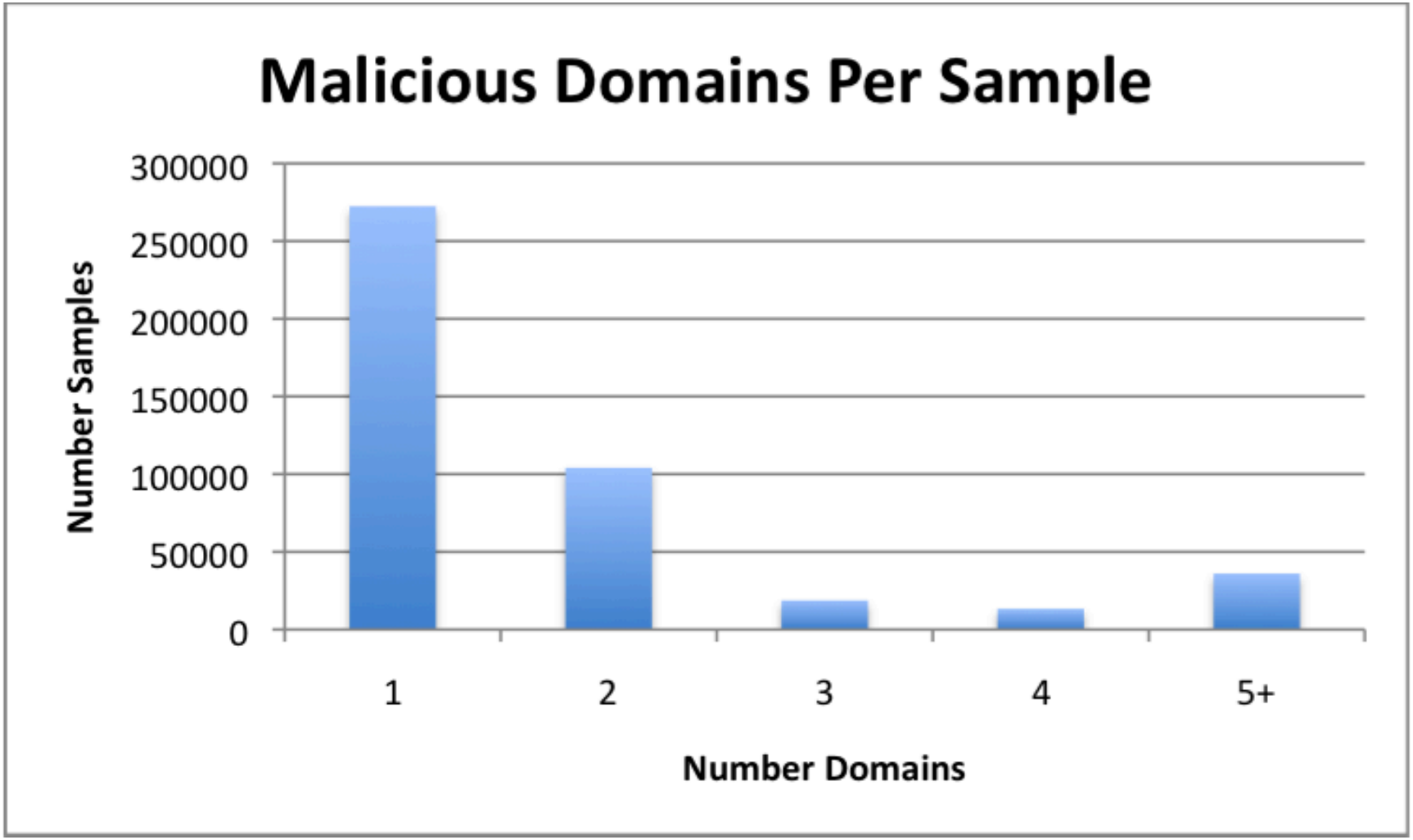


Protocols Used in Malware

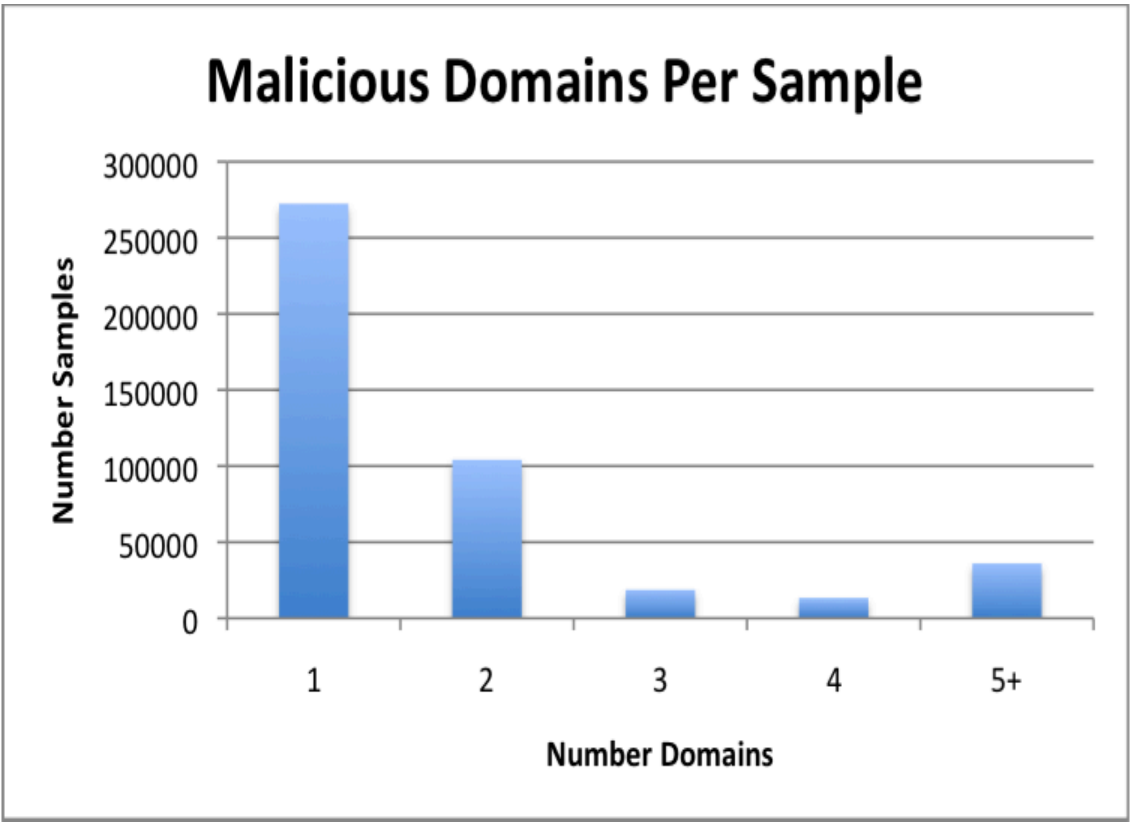


- Salient points:
  - Trend towards http, use of proxies, and overlay networks
    - Port 80 provides a large haystack in which to hide, frustrating DPI.

DNS Agility in Malware



DNS Agility in Malware



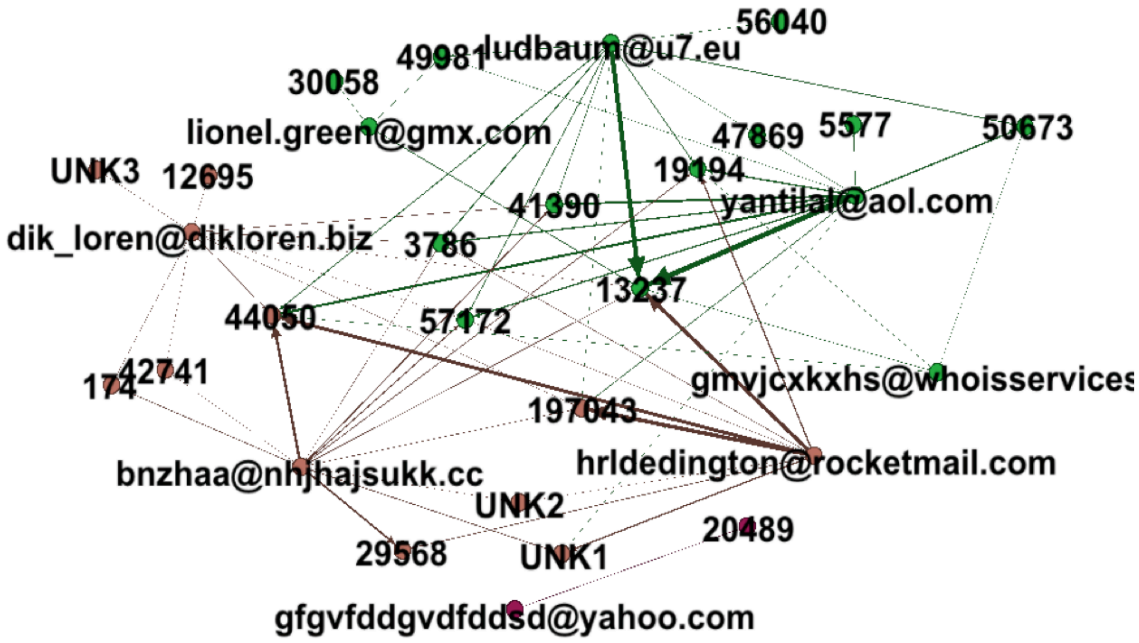
Salient points:

- Many associate with one domain
  - But this is an artifact of malware churn – a botnet may use hundreds of malware samples
  - Our challenge is to identify collections and cluster related samples.

Example: TDSS

- Example Botnet: TDSS
  - Millions of victims
  - Components: rootkit; p2p; DGA; secondary drops reside in RAM-only
  - Created by affiliate program ($20 to $200 for every 1,000 installations)
  - Called “indestructible” by AV researchers

Example: TDSS



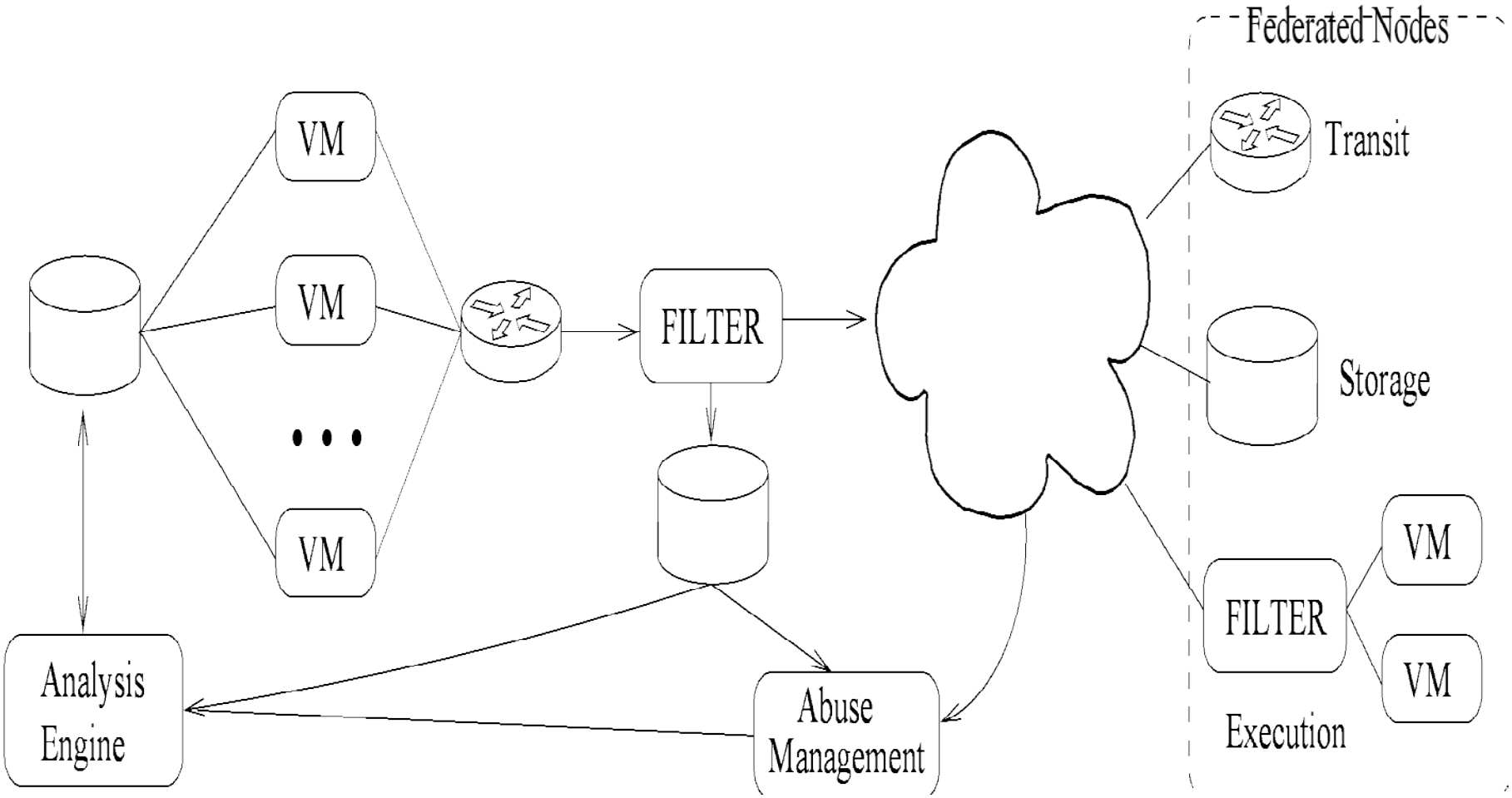
Salient points:

- A cloud of DNS services and related malware
  - Hundreds of colos; thousands of domains
- Incorporate other “botnets”, e.g., fake AV and clickfraud malware campaigns
  - We must describe the network platform of related binaries and network resources, not just a slice of the botnet

Federated Malware Analysis System

- Will use GT's MNIF (Malware Network Intelligence Gathering and Analysis Framework)
  - DURIP funded 2011
  - Designed to share intelligence with DETER
- Participants bring one or more of:
  - Localized storage: I can't run malware, but I can store analysis
  - VM Execution: I can execute/analyze malware, but lack storage/IPs
  - Transit/Filter/Egress: I only have IP addresses to offer; assuming there are sane policy controls on exit traffic

FMAS Overview



FMAS Design Criteria

- Process 100K+ samples/day, via distributed analysis system
- Three classes of messaging between federated hosts
  - Management Messages: start/stop VMs, forcing firewall rule updates, add/remove nodes, etc
  - Partial-Evidence Messages: Informational broadcasts representing partial learning from remote nodes. E.g., feature and vector observations, to be used in machine learning. Likely, only analysis nodes subscribe
  - Conclusive Findings Messages: Announcing facts about samples (availability, AV scans, DNS analysis, clustering output, etc.)

FMAS Policy Layer

- Most industrial malware analysis runs samples in honeypots
- Existential risks
  - Possible harm to 3rd parties
  - Provides robust messaging/support for botnet (e.g. 3322.org takeover omitted 60-misc malicious domains, which then resolved via MS-operated DNS servers.)
  - Taints data/analysis (e.g., if PII is obtained from analysis and shared in network)
- Global Cyber Risk (GCR) will perform extensive policy analysis

GCR Analysis

- Legal, policy and ethical analysis of proposed framework, noting data sources, handling, and FMAS interactions with other individuals and networks.
- Operator Agreements
  - Draft MOUs for participants in FMAS
  - Tailored to role (storage, execution, transit)
  - Legal policies for malware analysis
  - Policy analysis of passive DNS collection

DNS Analysis

- Construction of Passive DNS mirror
  - Existing DNSDB mirror proving too critical to security companies, LEO, and analysts; research-oriented mirror required
  - Includes vetting of operator agreements, data collection, identification of policy issues in above-the-recursive data collection, etc.

“Reputation” Analysis

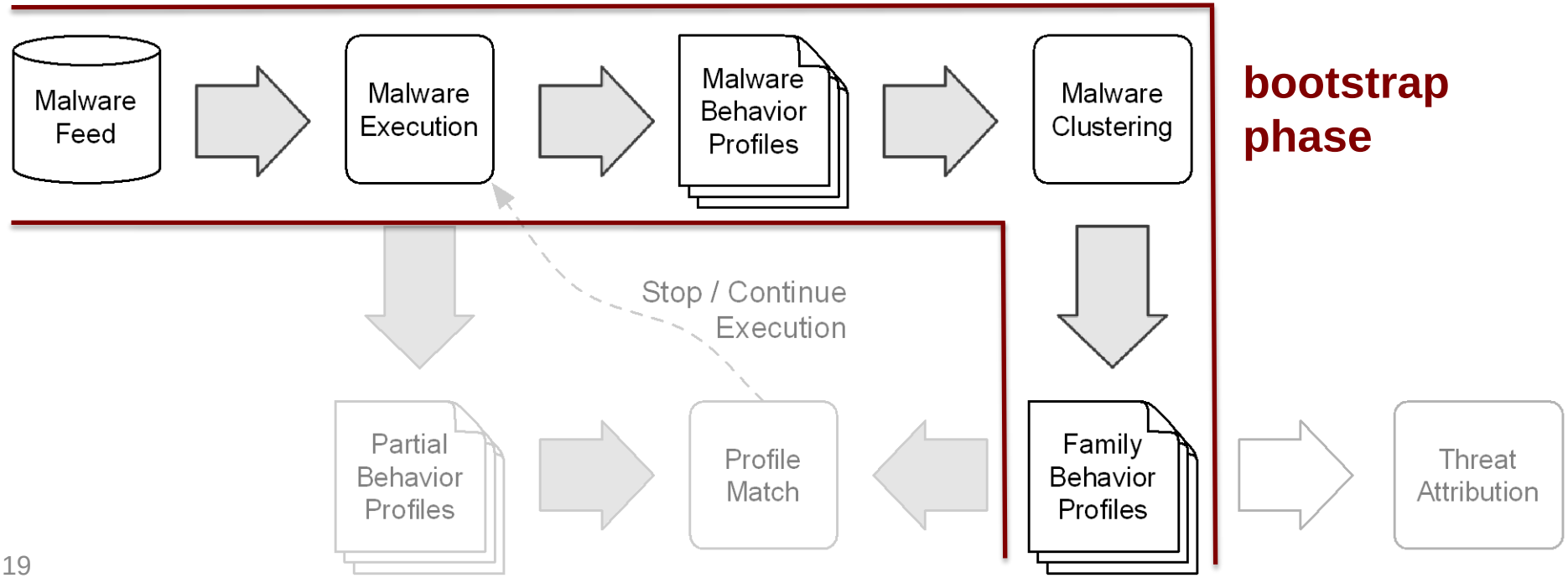
- Identify key properties of NS-reputation
  - Goals
    - Leverage large-scale domain intelligence (prefix whois, bulk whois for gTLDs and ccTLDs)
    - Create indexed datasets for high-speed and mobile access

Clustering Analysis

- Identify semantic equivalence between malware samples using system- and network-level analysis.
- Goals
  - Identify optimal flexible execution schedule, to speculatively halt analysis of similar/redundant samples
  - Selectively group samples using static/low-cost attributes to execute only a few group representatives, without loss of C&C information
  - Identification of key domain, static, and URL-based features
  - To be exported as a “malware channel” of broadcast information

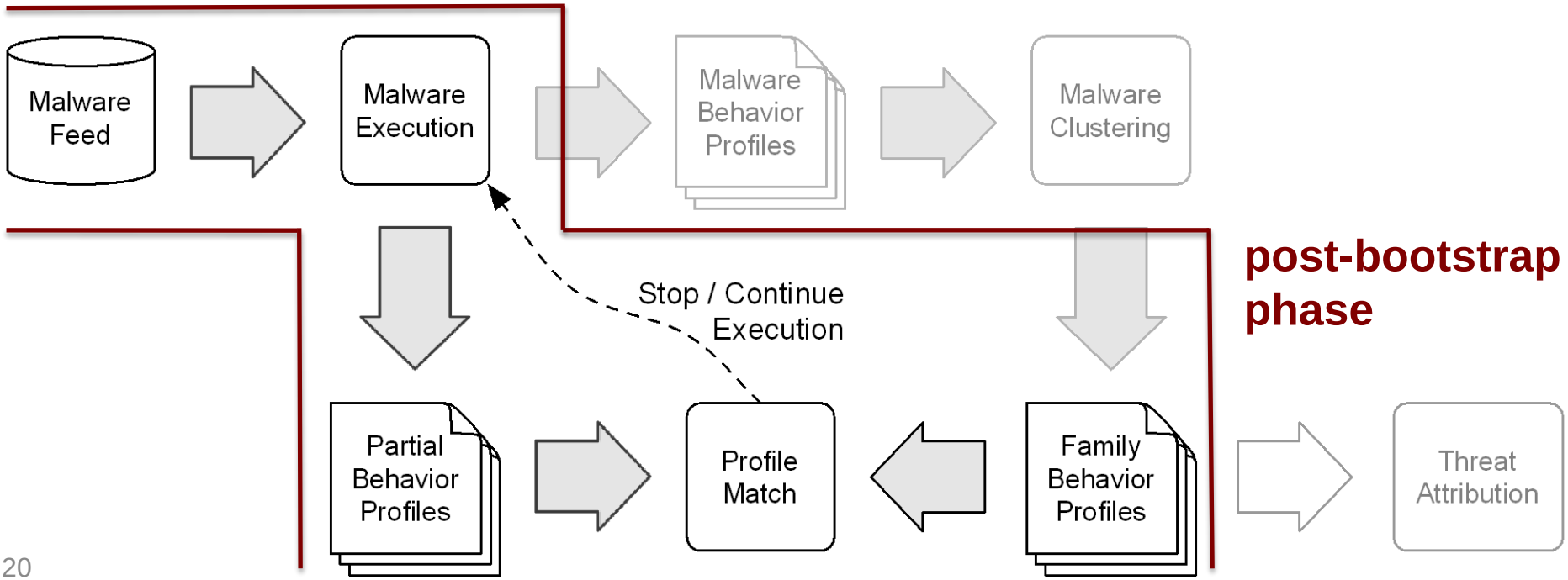
Scaling Malware Execution

- Analyze “bootstrap” malware dataset
  - Run each sample for a relatively long time (e.g., few hours)
  - Group samples that behave similarly into *malware families* (clustering)
  - Extract *family behavior profiles* for each malware family



When Should We Stop?

- Running new samples (post-bootstrap phase)
  - Frequently vet network/system behavior against family behavior profiles
  - If a profile matches a known family:
    - do malware in the family exhibit new behaviors if run for longer?
  - Stop/continue execution accordingly



Feature Extraction and Similarity Metrics

Homeland Security
Science and Technology

- Extract features from network behavior profiles
  - Domain-related features
    - Set of domain names queried
    - Name, location and reputation of authoritative name servers
  - IP-related features
    - Set of contacted IPs
    - Location and reputation of BGP prefixes and AS
  - Features for HTTP-base malware
    - URL structure
      - path similarity, variable names, etc.
    - Other HTTP request header characteristics
      - E.g., anomalies in header compositions, compared to normal browser-generated headers

FMAS Status

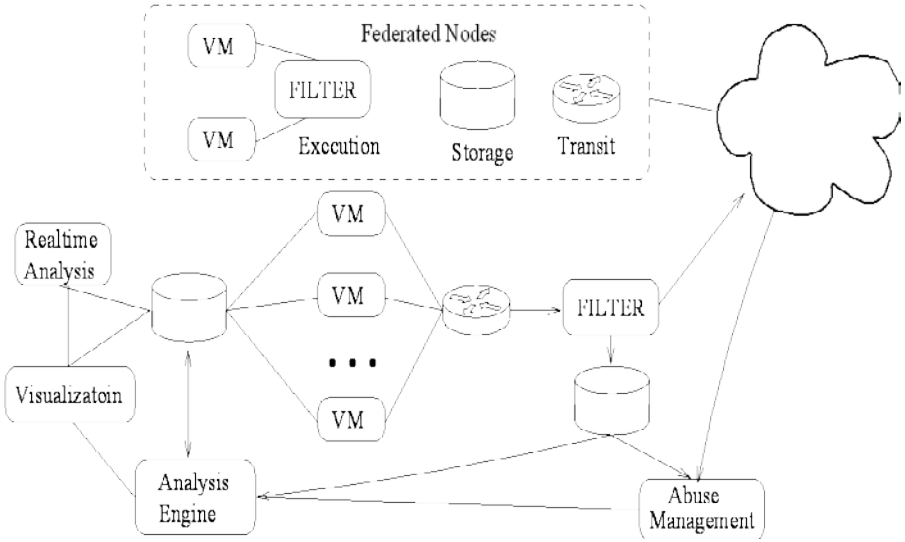
- Identified sources for malware at about 100,000 samples/day
  - No financial arrangement for samples
- Started work on NS reputation (esp. mobile analysis framework)
  - Android: Search for “Early2Rise” https://play.google.com/store/apps/details\?id=com.dissectcyber.early2rise
  - Apple iOS: Pending Apple review; request early access via https://testflightapp.com/register/

Technology Transition

- Several team members are directly involved in network operations and policy work
  - Malware samples, spam, DNS, and other real-world data
  - Directly adopt technologies developed and publish/broadcast data (e.g., SIE at ISC) and guidelines

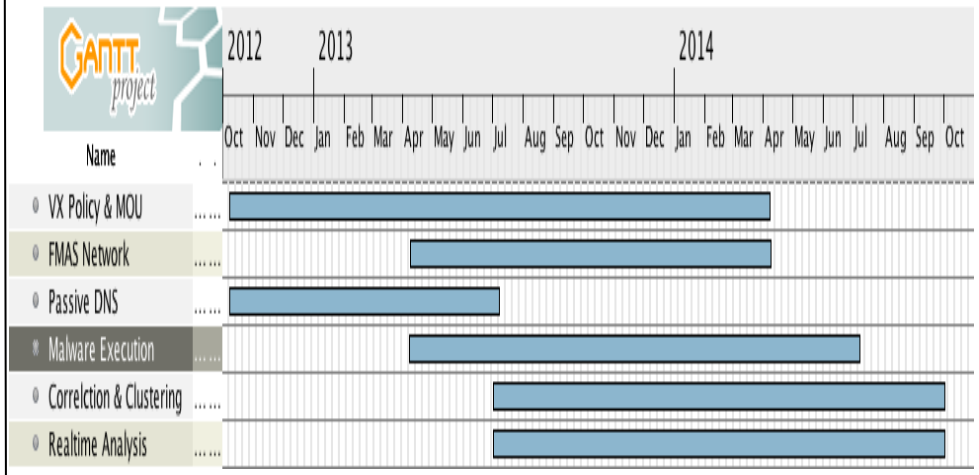
- Damballa a Georgia Tech spin-off, on-going collaboration, established tech transfer relationship
- When appropriate: malware samples to PREDICT, and malware analysis system part of DETER

Quad Chart



- Legal and policy framework for malware exchange
- Large-scale federated malware exchange and execution system
- Policy and technical framework for passive DNS collection
- Next-gen malware and domain correlation algorithms
- Real-time threat data

- **Federated Malware Analysis System:** Large-scale malware execution; scalability and quantitative transparency assessment; innovative egress filtering; next-gen baremetal framework
- **Malware Repository:** Vetted mirroring of binary and metadata with transparent, in-depth policies
- **Malware Clustering:** Based on host- and network- based properties
- **Real-time Data Analysis:** Visualization and query of synthesis of data



Homeland Security
Science and Technology

Thank You!