REDJACK

AMP-Based Flow Collection

Greg Virgin - RedJack

REDJACK

AMP- Based Flow Collection

- AMP - “Analytic Metadata Producer”: Patented US Government flow / metadata producer
- AMP generates data including
  - Flows
  - Host metadata (TCP stack information, software banners)
  - Metrics
- Purpose of this talk: To discuss the flow data collection implications of these additional data types for forensic analysis (not just correlation and alerting)
  - Additional data sources
  - Analysis scenarios
  - Collection schemes

REDJACK

Additional Data Sources

- Core data source: flow data
  - Netflow-like data with additional TCP flag information
- Flow-derived data sources: port details
  - Ports accepting connections
  - Bandwidth statistics
- Additional data sources (Not appropriate for flow records- aggregated data sources by IP, not communication)
  - TCP Stack information reflecting running O/S
  - Server Banners (as seen by the Internet)
  - Client Banners (as sent to the Internet)
  - DNS Names collected from both the DNS protocol and other protocols (NEVER trust DNS!)
  - Search strings from search engines (HTTP “referer” tags)

REDJACK

Scenario 1: Server “Importance”

- Server Profile
  - Configuration (“Windows 2000”)
  - List of listening ports (80, 443)
  - List of available services (“IIS/6”)
  - Domain name(s) (“www.golfcarts.com”)
  - Traffic Volume (X connections today, per week, per month)
  - Associated search strings (“golf carts”, “high performance golf carts”)
- Why?
  - Provides metrics to automatically partition servers by volume, type, vulnerability
  - Provides forensic value through server details often unavailable at time of analysis
- Flow analysis scenarios:
  - Which active servers were impacted by flow traffic / scans / attacks
  - Scrutinize payload-bearing traffic going to these servers
  - Make sure you’re not picking up potentially “normal” activity in other anomaly detection approaches (your concept of normal doesn’t necessarily have to be perfect)
  - Assign real world concepts to traffic activity and perform sanity checks through search strings

REDJACK

Scenario 2: DNS / Name Analysis

- Naming Information:
  - DNS Response packets
  - HTTP Get requests, mail protocol name announcements
- Why?
  - The current DNS implementation presents major risks because threats can masquerade as well known sites
  - The web protocol is dominated by virtual servers
  - We have found interesting discrepancies between DNS and naming in other protocols
  - Dealing with hosts as domain names is more natural (the purpose of the protocol)
- Flow analysis scenarios:
  - Name-based queries (possible with SiLK)
  - Names or name checksums incorporated into flow records for web traffic, followed by correlation with a name for the IP once the data is collected (helps with virtual servers)
  - Forensic analysis of traffic to or from bogus domain names to determine potential damage (but you have to do the above correlation first)

REDJACK

Scenario 3: Making IP Space Heterogeneous

- Required data:
  - Host Configuration
  - listening ports
  - running services
- Why?
  - Too often IP space is considered one big homogeneous blob - analysis is done on traffic between nodes without considering types of nodes
  - The diagnosis of activities such as worms can be made from hosts in a set running the same piece of software rather than signature
- Flow analysis scenarios:
  - What has been called a “similarity” analysis: take an IP set and run it against host profiles to provide statistics on what the hosts in the set have in common
  - Flow analysis broken down by host attributes isn’t very common, so there are a number of possibilities

REDJACK

Scenario 4: The “Alternate Use” Flag

- Marking flows for statistically significant attributes is marking flows based on signatures, not necessarily “new” data
- “Alternate Use” refers to the proper use of an Internet protocol without being used for the *purpose* of the protocol (this is not protocol analysis)
- Why?
  - This type of traffic can be a huge portion of the traffic
  - Of unique DNS names seen by your network, more than half of them may come from just a handful of sources
- Flow analysis scenarios:
  - Often port and protocol numbers are considered synonymous with legitimate use of protocols; this can be used to filter out alternate uses
  - Most of the “alternate” uses for DNS appear to be spam reporting, that information could be harvested

REDJACK

Scenario 5: IDS Verification

- Use host information or flow data to validate IDS records
  - If hosts aren't running the software that IDS signatures think they are…
- Not a new concept and done in practice

REDJACK

Summary of Scenarios

- New data sources can be used with flow data to:
  - Add contextual information and increase situational awareness
  - Create filters that could be useful for both queries and data collection
  - Partition data into bins or streams with more (or less) analytic meaning
- The best result is for these techniques to impact the data or be recorded as additional data
- This has an obvious impact on collection infrastructure
  - Data production software should be able to mark, reformat, or drop flow data based on this information
  - Data collection and storage software should be able to process or partition this information
  - Since most of these techniques don't amount to much more than a filter definition, a registry for these filters that different parts of the flow collection infrastructure can use is appropriate

REDJACK

New Sensor Attributes

- (This is in addition to flows with TCP options, host information, and DNS)
- Filters based on additional information
- Domain name value for the web protocol
- “Alternate Use” flag
- Not yet discussed:
  - Change ICMP to include third IP address in some instances

REDJACK

New Data Collection Attributes

- Marking or partitioning flows with domain names
- Metrics, filtering, and additional aggregation (flows for large servers can be compacted)

REDJACK

New Data Store Attributes

- Flow data closely tied to new data sources
- Registry for filtering techniques that can be leveraged by the sensor and collection
- Questions?
  - Greg Virgin, greg.virgin@redjack.com