NSA XKEYSCORE 2008

Black slides are original
XKEYSCORE

25 Feb 2008

xkeyscore@nsa
What is XKEYSCORE?

1. DNI Exploitation System/Analytic Framework
2. Performs strong (e.g. email) and soft (content) selection
3. Provides real-time target activity (tipping)
4. "Rolling Buffer" of ~3 days of ALL unfiltered data seen by XKEYSCORE:
   - Stores full-take data at the collection site – indexed by meta-data
   - Provides a series of viewers for common data types
1. Federated Query system – one query scans all sites
   - Performing full-take allows analysts to find targets that were previously unknown by mining the meta-data
Methodology

- Small, focused team
- Work closely with the analysts
- Evolutionary development cycle (deploy early, deploy often)
- React to mission requirements
- Support staff integrated with developers
- Sometimes a delicate balance of mission and research
System Details

- Massive distributed Linux cluster
- Over 500 servers distributed around the world
- System can scale linearly – simply add a new server to the cluster
- Federated Query Mechanism
Where is X-KEYSCORE?

Approximately 150 sites
Over 700 servers

TOP SECRET//COMINT//REL TO USA, AUS, CAN, GBR, NZL
What is unique about XKEYSCORE?
Why do shallow

- Can look at more data

- XKEYSCORE can also be configured to go shallow if the data rate is too high
Why go deep

- Strong Selection itself give us only a very limited capability

- A large amount of time spent on the web is performing actions that are anonymous

- We can use this traffic to detect anomalies which can lead us to intelligence by itself, or strong selectors for traditional tasking
What XKS does with the Sessions

Plug-ins extract and index metadata into tables

[sessions] ➔ [processing engine] ➔ (database) ➔ (user queries)

- phone numbers
- email addresses
- log ins
- user activity

Database

metadata tables

full log
<table>
<thead>
<tr>
<th>Plug-in</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>E-mail Addresses</td>
<td>indexes every E-mail address seen in a session by both username and domain</td>
</tr>
<tr>
<td>Extracted Files</td>
<td>indexes every file seen in a session by both filename and extension</td>
</tr>
<tr>
<td>Full Log</td>
<td>indexes every DNI session collected. Data is indexed by the standard N-tuple (IP, Port, Casenotation etc.)</td>
</tr>
<tr>
<td>HTTP Parser</td>
<td>indexes the client-side HTTP traffic (examples to follow)</td>
</tr>
<tr>
<td>Phone Number</td>
<td>indexes every phone number seen in a session (e.g. address book entries or signature block)</td>
</tr>
<tr>
<td>User Activity</td>
<td>indexes the Webmail and Chat activity to include username, buddylist, machine specific cookies etc.</td>
</tr>
</tbody>
</table>
What Can Be Stored?

- Anything you wish to extract
- Choose your metadata
- Customizable storage times
- Ex: HTTP Parser

```
GET /search?hl=en&q=islamabad&meta HTTP/1.0
Accept: image/gif, image/x-xbitmap, image/jpeg, image/pjpeg, application/vnd.ms-application/msword, application/x-shockwave-flash, */*
Referer: http://www.google.com.pk/
Accept-Language: en-us
User-Agent: Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1)
Host: www.google.com.pk

connection: keep-alive
```

No username/strong selector
What can you do with XKEYSCORE?
Finding Targets

- How do I find a strong-selector for a known target?
- How do I find a cell of terrorists that has no connection to known strong-selectors?

Answer: Look for anomalous events
  - E.g. Someone whose language is out of place for the region they are in
  - Someone who is using encryption
  - Someone searching the web for suspicious stuff
Encryption

- Show me all the encrypted word documents from Iran
- Show me all PGP usage in Iran

- Once again – data volume too high so forwarding these back is not possible
- No strong-selector
- Can perform this kind of retrospective query, then simply pull content of interest from site as required
Technology Detection

- Show me all the VPN startups in country X, and give me the data so I can decrypt and discover the users

- These events are easily browsable in XKEYSCORE
  - No strong-selector

- XKEYSCORE extracts and stores authoring information for many major document types – can perform a retrospective survey to trace the document origin since metadata is typically kept for up to 30 days

- No other system performs this on raw unselected bulk traffic, data volumes prohibit forwarding
- Traditionally triggered by a strong-selector event, but it doesn’t have to be this way

- Reverse PSC – from anomalous event back to a strong selector. You cannot perform this kind of analysis when the data has first been strong selected.

- Tie in with Marina – allow PSC collection after the event
Language Tracking

- My target speaks German but is in Pakistan – how can I find him?

- XKEYSCORE’s HTTP Activity plugin extracts and stores all HTML language tags which can then be searched

- Not possible in any other system but XKEYSCORE, nor could it be –
  - volumes are too great to forward
  - No strong-selector
Google Maps

• My target uses Google Maps to scope target locations – can I use this information to determine his email address? What about the web-searches – do any stand out and look suspicious?

• XKEYSCORE extracts and databases these events including all web-based searches which can be retrospectively queried

• No strong-selector

• Data volume too high to forward
I have a Jihadist document that has been passed around through numerous people, who wrote this and where were they?
Interesting Document Discovery

- Show me all the Microsoft Excel spreadsheets containing MAC addresses coming out of Iraq so I can perform network mapping
  - New extractor allows different dictionaries to run on document/email bodies – these more complex dictionaries can generate and database this information
  - No strong-selector
  - Data volume is high
  - Multiple dictionaries targeted at specific data types
• Show me all the exploitable machines in country X

• Fingerprints from TAO are loaded into XKEYSCORE’s application/fingerprintID engine
• Data is tagged and databased
• No strong-selector
• Complex boolean tasking and regular expressions required
Discovery of new target web services

- New web services every day
- Scanning content for the userid rather than performing strong selection means we may detect activity for applications we previously had no idea about
Entity Extraction

- Have technology (thanks to R6) – for English, Arabic and Chinese
- Allow queries like:
  - Show me all the word documents with references to IAEO
  - Show me all documents that reference Osama Bin Laden
  - Will allow a ‘show me more like this’ capability
XKEYSCORE Success Stories
Over 300 terrorists captured using intelligence generated from XKEYSCORE
Innovation

- High Speed Selection
- Toolbar
- Integration with Marina
- GPRS, WLAN integration
- SSO CRDB
- Workflows
- Multi-level Dictionaries
Future

- High speeds yet again (algorithmic and Cell Processor (R4))
- Better presentation
- Entity Extraction
- VoIP
- More networking protocols
- Additional metadata
  - Expand on google-earth capability
  - EXIF tags
  - Integration of all CES-AppProcs
- Easier to install/maintain/upgrade