Tech Strings in Documents (aka Tech Extractor)

[redacted]

April 2010

Agenda

- Overview and History of Tech Strings in Documents
- Why is it important?
- Limitations of capability – advance to fingerprints
- Examples and live demo

Content-based Selection

- How do you find DNI data if you don't have a strong selector like IP or E-mail address?
- What if you only know keywords, part names, phrases etc. expected to be used by your target?

What is the Tech. Extractor?

- The "Tech Extractor" is a way of finding valuable intelligence based on keywords in the content of DNI sessions but it is a departure from traditional "soft selection" which tends to bring back a lot of junk.

What is soft selection?

- Soft selection, aka content based selection, is an approach at targeting traffic by looking for keywords or phrases rather than specific E-mail accounts

- Content based selection has suffered because of the poor design of content based selection engines

Communication vs. DNI Content

- Selection engines in use today were based on designs built to handle TELEX traffic
- TELEX is a highly formatted content rich type of traffic that does not resemble raw DNI seen with Internet traffic
- Raw Internet traffic contains HTML, web-pages, raw base-64 encoded documents etc.
- When analysts think of DNI "content" they are more referring to "communication content" then raw DNI content.
- Current DNI selection does not allow you to restrict hits to the "type" of traffic you want eg. Emails (including Webmail) or Documents

Communication vs. DNI Content

- If an analyst tasks a Boolean equation "bomb" and "chemical" they likely want to see all communication that mentions 'bomb' and 'chemical' and not all web pages, news stories, blog posts etc. where those two words appear
- What we need is a context-aware scanning engine that knows where it is inside of the raw DNI in order to properly apply analyst tasking

Soft Selection vs. Surgical Selection

- Existing selection techniques are blunt instruments
- XKEYSCORE contextual dictionaries provide an extremely sharp knife to make accurate selection decisions

"That's not a knife…..*THAT's* a knife!"

What is the Tech Extractor

- The Tech Extractor was X-KEYSCORE's first stab at context-aware scanning and it only focuses on three contexts:
  - E-mail Bodies
  - Chat Bodies
  - Document Bodies:
    - Microsoft Word, Excel, PowerPoint, Project, Visio
    - Adobe PDF, Postscript
    - Rich Text Format (RTF)

How does the Tech Extractor work?

- The Tech Extractor works by scanning a list of keywords (or regular expressions) against those three contexts and then tags the results.

- This is not "filtering and selection" and we're not forwarding any data home

- XKS is simply tagging sessions with meta-data, much like we do with appids+fingerprints

How does the Tech Extractor work?

- After the meta-data tag is applied, analysts can then use that meta-data tag as part of a *USSID-18* compliant query for traffic

- It's important to note, just like AppIDs+Fingerprints, Tech Extractor tags aren't necessarily USSID-18 compliant by themselves.

- You may need to add a valid foreign IP address, MAC address or country code before you query!

Context-Aware Tagging

- Ex
  M

Subject:	**NFF-66024-GCC-KHI**
From:	[redacted]
To:	[redacted]
Cc:	[redacted]
Date:	Tue Dec 30 10:57:48 GMT 2008

Event T
email_t
Fm City
KLOSTE

HTML | Plain Text | Attachment

IMEI: [redacted]

Model: 6300

WON: 66024

ASC: GCC-KHI

Symptom: 4100

Comments: no fault found phone is working properly kindly confirm the fault in detail when and in which condition it creates problem related to mention symptom

[redacted]
GSM Repair Engineer
[redacted]
Tel:
Mob:
Fax:

How does the Tech Extractor work?

- Also this is not retrospective.
- After a list is tasked, XKS will scan data collected from that point on looking hits.
- Any data previously collected and stored by XKS will not be scanned.

Where does XKS get its list of terms?

- Analysts provide the XKS team with lists of terms, called "Tech Dictionaries" which can contain multiple category names (aka "Tech Names")

- Only after the XKS team is supplied with those terms can the system begin scanning and tagging.

- GUI to allow analysts' entry of tech terms almost complete

Tech Extractor Tasking Rules

- Currently, all terms need to be classified REL FVEY

- Terms are case insensitive by default, but can be forced to be case sensitive

- Terms can hit as a substrings by default

  ex: 'ricin' will hit in 'pricing'

- However, terms can be forced to hit as a unique word (either by tasking them with a space at the beginning and end or by using a regular expression)

Full Foreign Language Support

- **Supports full foreign language tagging and querying**
- **Ex look for common Arabic expressions in E-mails coming from the Pakistan tr**

UIS Webmail Display

Windows Live™ Mail Beta

Active user:
Unknown

From: [redacted] ([redacted]@gmail.com)
Medium risk You may not know this sender. Mark as safe | Mark as unsafe
Sent: Thu 1/01/09 12:07 PM
To: [redacted]

السلام عليكم ورحمة الله وبركات

TOP SECRET//COMINT//REL TO USA, AUS, CAN, GBR, NZL

Tech Extractor Limitations

- While terms tasked for the Tech Extractor are applied only to Document, E-mail and Chat bodies, that is still a lot of traffic!
- If the term is too generic (or short) you're still likely to run into a lot of false hits.
- Also, while you can limit your results by adding more search criteria (country code, IP address etc), the term will be scanning all data looking for hits

Tech Extractor vs. Fingerprints

- Tech Extractor treats E-mail, Chat and Document bodies as a single "context"

- The XKS Fingerprint language gives you over 65+ contexts that can be used together to form powerful and specific signatures

- When terms are generic and are returning too many poor results through Tech Extractor, then it's time to make the switch to the full fingerprint language

Why use the Tech Extractor at all?

- One of the most powerful feature of the Tech Extractor is that it shows you exactly which term hit in the meta-data results:

Event Type	Tech Dictionary	Tech Name	Tech Value ▲	Tech Filename
document_body	classic	gsm	HLR	C:\Documents and Settings\SE EWSD\Desktop\25 APR 10 Daily Break Down.x
document_body	classic	gsm	ICCID	C:\Documents and Settings\Administrator\Desktop\New Franchisees Status fo
email_body	classic	gsm	IMEI	
document_body	classic	gsm	IMSI	Evo_Complaints_sheet_25-Apr-2010.xls
email_body	classic	gsm	MSISDN	

Why not Fingerprints?

- With fingerprints, you only see that the full equation (which can be very complex) was satisfied and you won't see which specific terms from the equation hit.

Live Demo

More information:

- On GCHQ wiki:

http://wiki.gchq/index.php/Tech_dictionaries

To submit tasking

- Please use the Excel Spreadsheet template developed by GCHQ CP

- And then E-mail [redacted]@gchq.gov.uk with the list

- In the near future analysts will be able to enter the terms themselves through a web-based GUI