TOP SECRET STRAP1 COMINT
The maximum classification allowed on GCWiki is **TOP SECRET STRAP1 COMINT**. Click to report inappropriate content.

For GCWiki help contact: webteam [REDACTED] Support page

LOVELY HORSE

From GCWiki
Jump to: navigation, search

LOVELY HORSE.JPG

LOVELY HORSE2.JPG

LOVELY HORSE is a TCP Task Order 144 initiative as part of CDO (formerly NDIST) and the Cyber Theme towards developing Open Source capability. So far, we have worked towards making structured datasets available on the high side for analysts to use - this data is available within HAPPY TRIGGER. We are now looking towards making use of more unstructured information (blogs, forums, Twitter). LOVELY HORSE seeks to experiment with provision of an indexed repository of unstructured information that can be used to push content of interest to individual analysts via a variety of mechanisms.

The initial LOVELY HORSE prototype can be accessed from here. See below for more details.

*See also — **BIRD SEED***

Contents

[edit] Problem statement

Analysts are potentially missing out on valuable open source information relating to cyber defence because of an inability to easily keep up to date with specific blogs and Twitter sources. Accessing these resources involves using specific JEDI terminals, or reading up at home. We need to make this information available to analysts on the high side at their normal terminals.

However, there is a balance to be found - analysts don't have the time to spend hours and hours reading through loads of blogs. In addition, we don't want this repository to be yet another tool that analysts have to access - this information needs to be incorporated into existing workflows.

We need to find a way to index this information so that analysts **only get a relevant subset** of this information **pushed to them**.

[edit] Initial prototype

We are working with JTRIG to make use of the existing BIRDSTRIKE architecture for capturing tweets from Twitter. We are also working with CISA around techniques they are developing to capture blog content. Both of these obviously take time, and are slower burn objectives.

In the meantime, we are running an initial prototype, where Twitter and (and subject to legal/security approval) blog content is manually scraped and uploaded to GCDesk. This content is accessible by way of personalised RSS feeds. Individual users can choose their preferences, in terms of which Twitter accounts and blogs they want to follow, and a personalised RSS feed is generated automatically for them to which they can subscribe.

This can be used by anyone, and can be accessed from here. Your personal RSS is linked from the LOVELY HORSE website.

As stated previously, this is a manual update at the moment, and will initially be **maintained on a best endeavours basis** (hopefully roughly daily). Once the BIRDSTRIKE architecture comes on line, this will be updated in real time.

For any requests for new Twitter feeds you wish to be able to subscribe to, please get in touch.

[edit] Current sources

Currently, we're bringing in the following list of Twitter accounts. To request new ones, please submit your requests, with a brief justification, via the suggestion box on LOVELY HORSE

- 0xcharlie
- alexsotirov
- anonops
- anonymousirc
- anon_central
- anon_operations
- bradarkin
- CeRTFi
- danchodanchev
- daveaitel
- dinodaizovi
- diocyde
- egyp7
- GoVCeRT_NL
- halvarflake
- hdmoore
- hernano
- JaNeTCSiRT
- kevinmitnick
- lennyzeltser
- lulzsec
- mdowd
- mikko
- msftsecresponse
- operationleaks
- owasp
- pusscat
- Shadowserver
- snowfl0w
- taosecurity
- taviso
- teamcymru
- thegrugq
- TheHackersNews
- tinman2k
- VuPeN
- WTFuzz

[edit] Future development concept

The rest of this page is constituted from ideas that we currently have about LOVELY HORSE.

[edit] Team

It will be delivered by TCP's TO144 team.

[edit] Sources

Initially we need to identify a series of sources. We currently have a list of around 60 blog and Twitter sources that have been identified by CDO analysts and cyber defence experts from Detica, and most of these have been approved for collection by MP-LEG.

Information will arrive in unstructured 'information articles'. In the context of a blog, an article would be a post; on Twitter, an article would be a tweet.

Blog sources:

These sources have currently been approved by MP-LEG (see approvals spreadsheet in DISCOVER)

- http://www.secureworks.com/research/blog/
- http://www.secureworks.com/media/blog/
- xs-sniper.com/blog
- bugix-security.blogspot.com/feeds/posts/default
- carnal0wnage.attackresearch.com/rss.xml
- intrepidusgroup/insight/feed
- www.offensivecomputing.net/?q=node/feed
- rdist.root.org/feed/
- www.darknet.org.uk
- imthezuk.blogspot.com
- dogber1.blogspot.com/
- www.ragestorm.net/blogs/
- blog.mandiant.com
- www.openrce.org
- feeds.trendmicro.com/Anti-MalwareBlog
- blogs.technet.com/b/msrc/rss.aspx
- blogs.adobe.com/psirt/feed

These sources are currently not approved by MP-LEG

- www.f-secure.com/weblog/weblog.rtf
- www.f-secure.com/exclude/vdesc-xml/latest_50.rss
- feeds.feedburner.com/GoogleOnlineSecurityBlog
- securityvulns.com
- feeds.feedburner.com/infosecResources
- targetedemailattacks.tublr.com

Twitter sources:

The advice from MP-LEG on this issue is that "provided the accounts you are selecting for acquisition meet the criteria as agreed in the approvals spreadsheet, i.e. those of "academics specialising in the identification and investigation of vulnerabilities and malware", there is no need to seek authorisation for each individual Twitter account." Our selection of Twitter sources is currently as listed above, but will undoubtedly increase over time.

Further potential sources of interest are found at Computer_security_news_and_views

[edit] Processing

Initially, these articles get processed into three components:

[edit] Content

The content will be the full textual content of the article. This will be stored as some sort of CLOB in a database.

[edit] Metadata

We would strip metadata from the article such as

- Author/Source
- Datetime of submission

and used this to update a Source Directory - information about the individual sources. For example:

- Author
- Number of articles in LOVELY HORSE
- Average usefulness rating - see feedback mechanism
- Tags of subject matter linked with this source? - see indexing

[edit] Index

This is the important bit. The aim is to index the unstructured information so that it can be linked back to

- An analyst's particular interest
- As enrichment to an existing investigation

The proposed idea is to make use of tagging (defining 'indexing' as 'identifying keywords'). Each article would be tagged with information that had been extracted from it. These tags could be IP addresses, domains, or any text string from within the content of the article. Effectively these tags are the output of entity extraction, and this list of tags would then be associated with that article.

Similarly, lists of tags are associated with individual analysts, to define their specific interest set.

[edit] How to generate a pot of tags?

We would need a pot of tags that becomes our entity set which we're extracting from new articles coming in. How to generate this pot of tags?

- Simple idea would be to regex for IP addresses and domains to start off with.
- Could index every capitalised word in a blog title.
- Could get analysts to provide a list of keywords they are specifically interested in.
- Could we extract keywords from existing analyst toolsets - for instance, do analysts tag investigations within Palantir?
- Analysts should be encouraged to tag articles they read

There is potential to link this entity extraction initiative in with corporate entity extraction tools that may provide more sophisticated matching.

- Could try and analytically identify tags. Whole articles could be tokenized and a word count generated. If a particular term appeared, say, 4 or 5 times in the current week, but not last week, then maybe that's a new trend? In which case we should add this term to the pot of tags.

[edit] Feedback mechanism

Important to allow analysts easy ability to appraise usefulness of information. Analysts should be able to 'like' content from whichever interface they're accessing the content. If an analyst likes a particular article, tags from that article are automatically added to their personal tag list.

Articles can have a usefulness rating assigned to them - generate some metric on the lines of (number of 'likes'/number of views). Articles that have a usefulness rating over a specific threshold could be pushed to all analysts. An average of the usefulness ratings across all articles from one source can be used to appraise different sources - almost becomes a crude 'confidence factor' in the information - should I trust/act upon this information?

[edit] Visualisation and Access

Need to be different ways analysts access and view this content.

- Palantir - as enrichment to existing investigations. Similarly to the current enrichment helper, any articles that had tags which are entities within the investigation are flagged up. The content should then be viewable in a human readable format within Palantir.
- Alerts - analysts should be alerted when a new article is tagged with a tag from their interest set. How should this alerting happen? Email? RSS feed?
- General search, there should be LOVELY HORSE front end that can be used for analysts to search across the whole repository. Would want to investigate tools that can provide Google-like searching (need to investigate MERA PEAK, NSA's LEXHOUND).
- May need to be a timeframe element in the enrichment, content that is 2 years old may not be relevant.

POC: [REDACTED]< (mail□□)

[edit] Your Thoughts

If you've got any thoughts on this initiative, please get in touch either directly to [REDACTED], or feel free to edit this section and add them below:

-
-
-

Retrieved from "[REDACTED]"

Views

- Page
- Discussion
- Edit
- History
- Delete
- Move
- Watch
- Additional Statistics

Personal tools

Navigation

- Main Page
- Help Pages
- Wikipedia Mirror
- Ask Me About...
- Random page
- Recent changes
- Report a Problem
- Contacts
- GCWeb

Search

Go Search

Toolbox

- What links here
- Related changes
- Upload file
- Special pages
- Printable version
- Permanent link

- This page was last modified on 6 February 2012, at 09:40.
- This page has been accessed 538 times.
- All material is UK [http://www.gchq/organisation/ck/opensource/policy_strategy/copyright/ Crown Copyright] © 2008 or is held under licence from third parties. This information is exempt under the Freedom of Information Act 2000 (FOIA) and may be exempt under other UK information legislation. Refer any FOIA queries to GCHQ on 01242 221491 x30306 or infoleg@gchq.gsi.gov.uk
- Privacy policy
- About GCWiki
- Disclaimers

TOP SECRET STRAP1 COMINT
The maximum classification allowed on GCWiki is **TOP SECRET STRAP1 COMINT**. Click to report inappropriate content.