# Information Discovery in Cybersecurity incident data reported to DHS

Sam Perl, Robin Ruefle

Software Engineering Institute
Carnegie Mellon University
Pittsburgh, PA 15213



# Notices



This material is based upon work funded and supported by the Department of Homeland Security under Contract No. FA8702-15-D-0002 with Carnegie Mellon University for the operation of the Software Engineering Institute, a federally funded research and development center sponsored by the United States Department of Defense.

The view, opinions, and/or findings contained in this material are those of the author(s) and should not be construed as an official Government position, policy, or decision, unless designated by other documentation.



DM17-0649

# Introduction

- The Information Discovery project applies research techniques to analyze incidents reported to DHS.
- The project improves visibility into bulk incident ticketing data – primarily by leveraging information already provided in free text fields.
- Project tasks include
  - Extracting data fields
  - Performing analysis across multiple incident reports at the same time
  - Automating incident knowledge discovery tasks
  - Developing and applying metrics for incident reporting and indicators
  - Building and using interactive incident and indicator data visualizations
  - Transitioning methods into tools for data analysts and incident responders

# Benefits

Primary Project Benefits

- Makes existing data more accessible to analysts and maximizes the value of data already collected
- Automatically highlights known important elements in reported records and uses them to link records together in intuitive ways
- Provides interactive tools for analysts to visualize and explore datasets and report on incident and threat trends & patterns for situational awareness

# Motivation

DHS occupies a unique position with their access to cyber security data across the whole Federal Civilian Executive Branch (".gov") environment

- Receives reports from many civilian agencies and other partners
- Multiple cyber incident types with variable context and completeness
- Reports contain data on both common and rare attacks in .gov

Project Goals

- Improved searching, fusion, re-use, reproducibility, and efficiency of incident trend analysis
- Automated extraction of re-usable items from existing reports
- Integration of ticketing information with other data sources
- Methods to improve exchanges between DHS and reporters

# Information Discovery Objectives

Objectives:

- Understand the DHS incident data and the data generating activities. Identify improvements to data quality, use, reporting, automation, and metric generation.
- Helping DHS characterize their constituency from a cybersecurity perspective to better collaborate with them and offer customized support to them
- Apply analytical methods from other fields to clean DHS incident data and engineer features from it
- Improve incident analysis tools to identify attacker trends and patterns in a data driven way
- Research the current use of incident management metrics, taxonomies, and classifications in the computer security incident response team (CSIRT) and security community and apply best practices

# Incident Report Data Overview

The 2002 FISMA Act requires Federal Agencies to report all IT Security incidents to DHS. This gives DHS unique access to cyber incident data across multiple agencies.

- The dataset contains IT security incident reports from US Government Departments & Agencies, international teams, and some ISACs, state/local/tribal and some private organizations
- It also includes incidents DHS finds during monitoring of the border traffic between D&A's and the internet.
- Over 100,000 incidents are reported per year
  - The information provided in each report can vary
  - Reporting forms are provided but are not always used

**'Incident Tickets' are the operational record of security incidents and interactions**

# Terminology

**Incident** - A computer incident within the Federal Government as defined by *NIST Special Publication 800-61* is a violation or imminent threat of violation of computer security policies, acceptable use policies, or standard computer security practices.

**Incident Report** - A written account of the incident (in our case reported to DHS). The content of an incident report depends upon the organizations policies, requirements and common practices including: investigative culture, tools used, etc. See https://www.us-cert.gov/government-users/reporting-requirements for more details.

**Incident Ticket -** 'Incident Tickets' are operational records created to work and track actions during/after security incidents

**Indicator** - Observable information about the attack that is typically identified during investigation of a security incident. Common types of indicators are filenames, file paths, hash values, IP addresses, domain names, etc.

# Is the meaning of ‘incident’ shared?

FISMA 2002, H. R. 2458—52

AGENCY PROGRAM.—Each agency shall develop, document, and implement an agencywide information security program, approved by the Director under section 3543(a)(5), to **provide information security for the information and information systems that support the operations and assets of the agency,** including those provided or managed by another agency, contractor, or other source, that **includes— procedures for detecting, reporting, and responding to security incidents,** consistent with standards and guidelines issued pursuant to section 3546(b)

Pursuant to FISMA, each federal agency is **required to notify and consult with DHS regarding information security incidents** involving the information and information systems (managed by a federal agency, contractor, or other source) that support the operations and assets of the agency.

(In practice, the content and threat details can vary by reporter)

# Remedy Tickets

A custom Remedy system is used to store tickets. Our dataset is a subset extracted from Remedy

**Structured Fields**

- Reporter contact information
- Threat Category, subcategory
- Date of submission
- assigned group, closure status
- Known relationships to other tickets

**Unstructured Fields**

- Notes (free text allowed) – often contains useful context information about reported events.
- Activity Log – communications tracked in remedy about the initial report

# Interactive Dashboard of Incident Trends

We built an interactive dashboard for analysts/managers interested in trends which is difficult to determine in the current system

- The rate and type of threats being reported by whom & over desired timeframes
- Lists of the most frequent reporting orgs. and what they are reporting
- Ability to drill down into time periods, by specific threat types, specific reporters etc.
- Allows viewer to formulate new questions such as: “Why is there a big drop from agency X in October?”
  - Is it due to something we changed in our reporting policy? or did they change their reporting policy? or did they do something to decrease the threats against them?
  - Analysts can also read the collective reports for that time period to get a sense of the answer for themselves

# US-CERT Data Browser (Prototype)

Incidents | Observables | Communities | Agencies

Trends | Table | Details

## Filters/settings

**Date range**

2013-01-01 to 2015-12-31

**Bin width**

week

**Breakout variable**

- -None-
- DCO
- Category
- Cluster

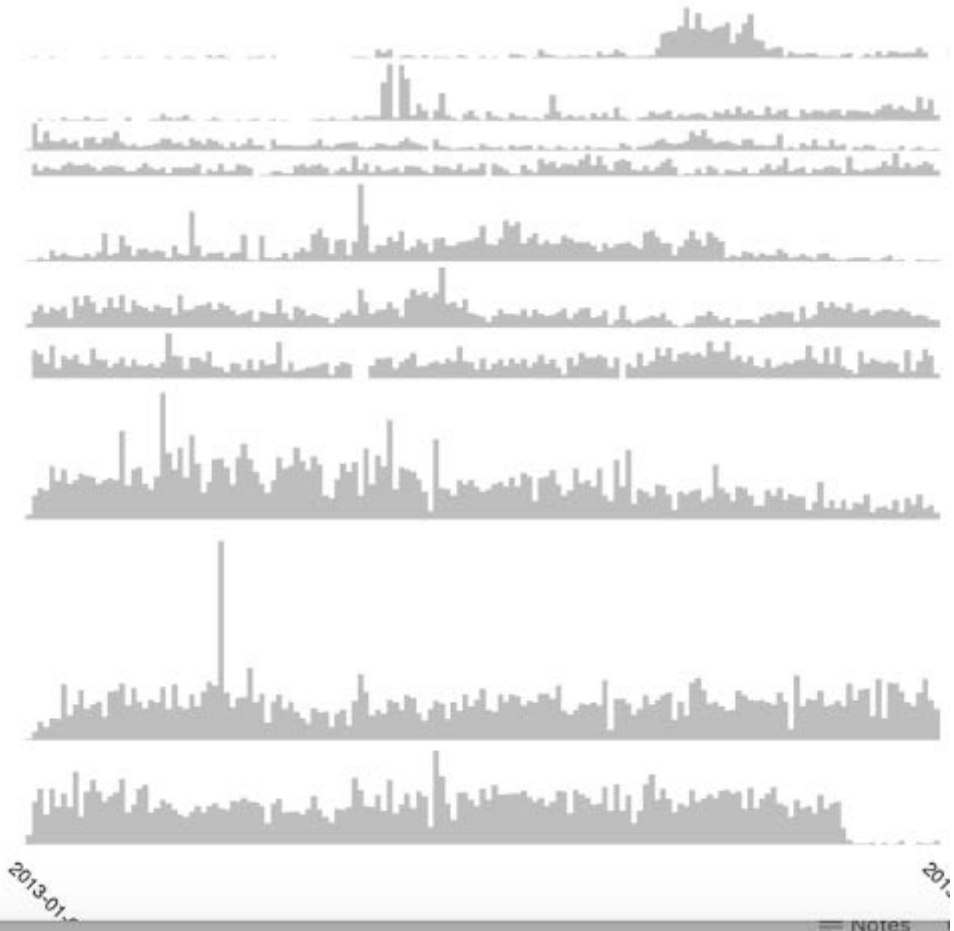
**Restrict to subgroups:**

- [x] Category
- [ ] DCO

**Select categories:**

- [ ] Investigation
- [ ] Exercise/Network Defense Testing
- [ ] General Public
- [x] Improper Usage
- [ ] Malicious Code
- [ ] Scans/Probes/Attempted Access
- [ ] Unauthorized Access

Excludes the least-observed 56 levels,
which account for about 13.03% of the records:




Slide 41 of 41 English (United States) Notes

# Search for incidents and improve context by looking at the total picture of the response

We built features to quickly find tickets based upon text strings, regex, time period, reporter name, threat type and other useful features

- Searching for terms across all incidents is fast, user interface is browser based with click to view.
- Searching returns entries in ALL text in all tickets, including Work Logs
- Communications in the Work Logs often contain context that is not in the ‘Notes’ field
  - Work log may contain information about what follow up questions were asked, what the answers were, status updates of remediation actions, additional attacking hosts or victim hosts, etc.

# US-CERT Data Browser (Prototype)

Incidents | Observables | Communities | Agencies

Trends | Table | Details

## Filters

**Search the notes (accepts wildcards & regex):**

SQL

**Select agency:**

- all agencies -

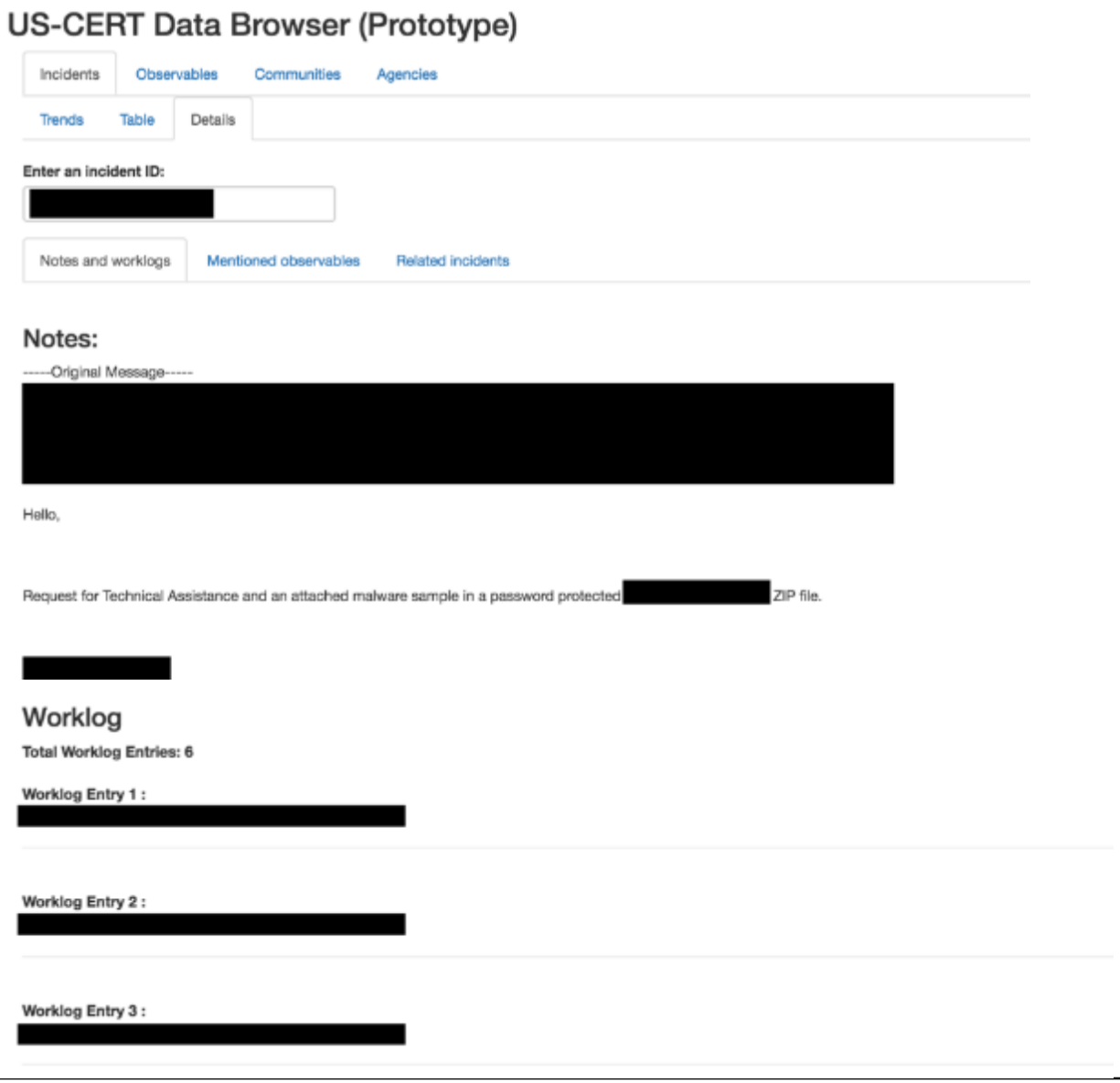
**Select incident category:**

Improper Usage

Show 10 entries

| Incident ID | Time reported |
|---|---|
| [redacted] | 2013-01-06T15:32:00Z |
| [redacted] | 2013-05-24T13:27:05Z |
| [redacted] | 2013-05-25T00:11:46Z |
| [redacted] | 2013-06-17T22:03:46Z |
| [redacted] | 2013-06-19T20:25:54Z |
| [redacted] | 2013-07-12T13:04:03Z |
| [redacted] | 2013-08-12T19:41:32Z |
| [redacted] | 2013-09-27T20:45:21Z |
| [redacted] | 2013-10-25T15:12:42Z |
| [redacted] | 2014-01-08T21:36:48Z |

# All work log entries are also displayed



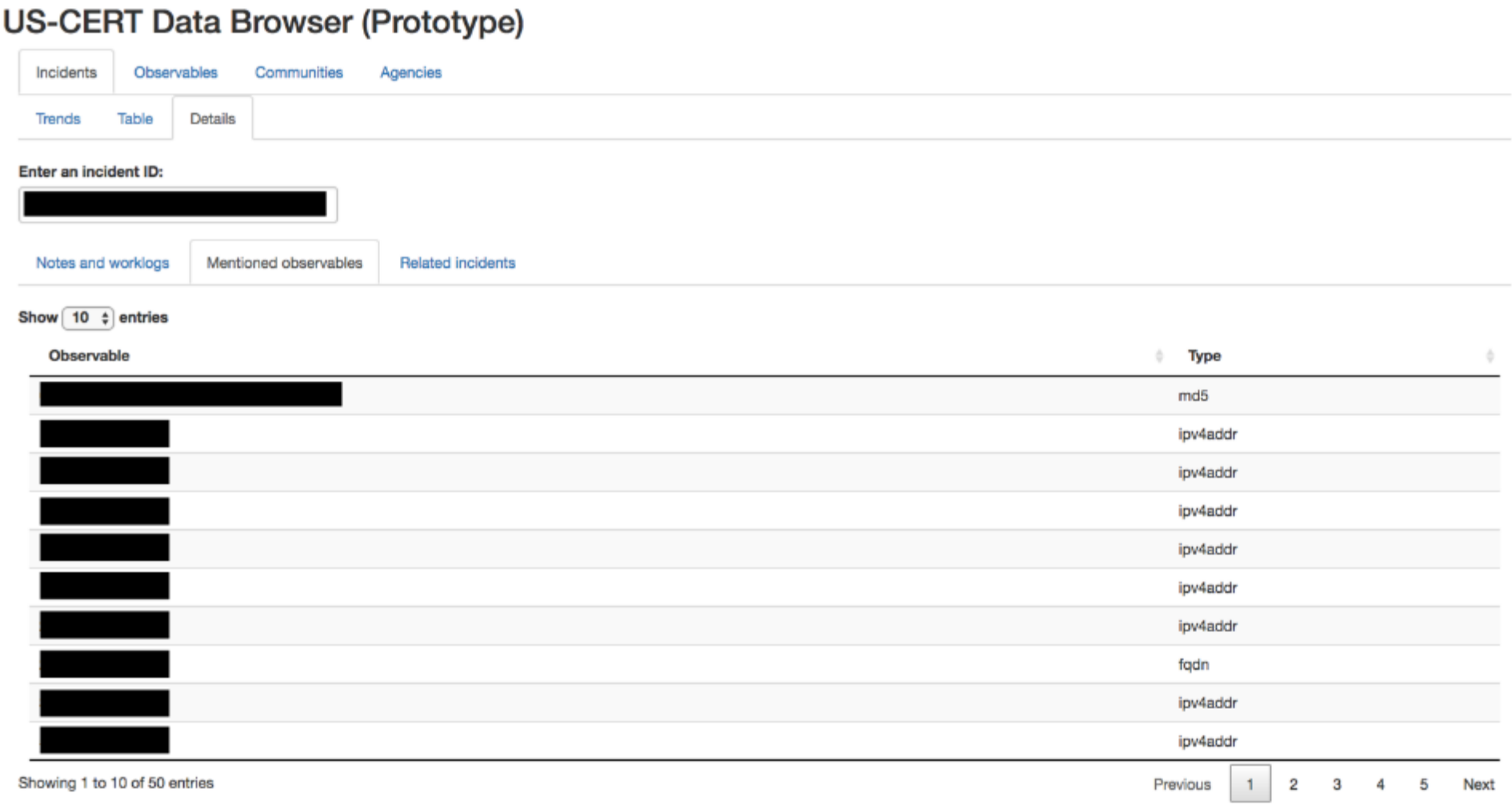
# Extraction of technical observables from tickets (data cleaning & feature extraction)

Cyber incident description (also called ‘notes’) is a free text field in the DHS dataset. Organizations typically report a variety of data types in this field including a natural language description of the incident, parts of multiple log files, links to intranet pages, links to internet pages (blogs), links to potentially malicious domains, a forensic artifacts from threat analysis (md5s, malware family name, filenames, ).

We extract many important security data types (‘observables’) from this field and put them into a database for further analysis.

Analysts can view the observables automatically extracted from each ticket in the “observables” tab of the data browser prototype

# Observables



# Additional Data

- We automatically 'enrich' the extracted observables with other datasets such as open source reports about attacker infrastructure tendencies, blacklists, etc.
- This means more context on observables. The 'is this a known bad lookup?' question is automated for all observables.

Example data sources used:

- Publically available reports with descriptions of observed hacking behavior
- Public and Private Blacklists of observed malicious domains, email addresses, IP addresses, etc.

- For more examples, see these interesting open source projects:
  - APTNotes – A repository of publicly-available papers and blogs by year of malicious campaigns/activity/software that have been associated with vendor-defined APT. https://github.com/aptnotes/data
  - Ioc_parser - Tool to extract indicators of compromise from security reports in PDF format. https://github.com/armbues/ioc_parser

# Extracted features are used to find similar threat behavior in other reports

After extracting the features from a given report, we can use them to find other ‘similar’ reports.

The default similarity is to find threat behavior but analysts have the ability to set their own definitions of similarity and run those on the incidents or observables in real time.

By default, we calculate the distance between incident reports using Jaccard similarity. This uses the amount of duplicate sets of indicators they contain that are also not in all other tickets.

# Incident ‘Similarity’

Incident similarity is the key defining metric in choosing related incidents.

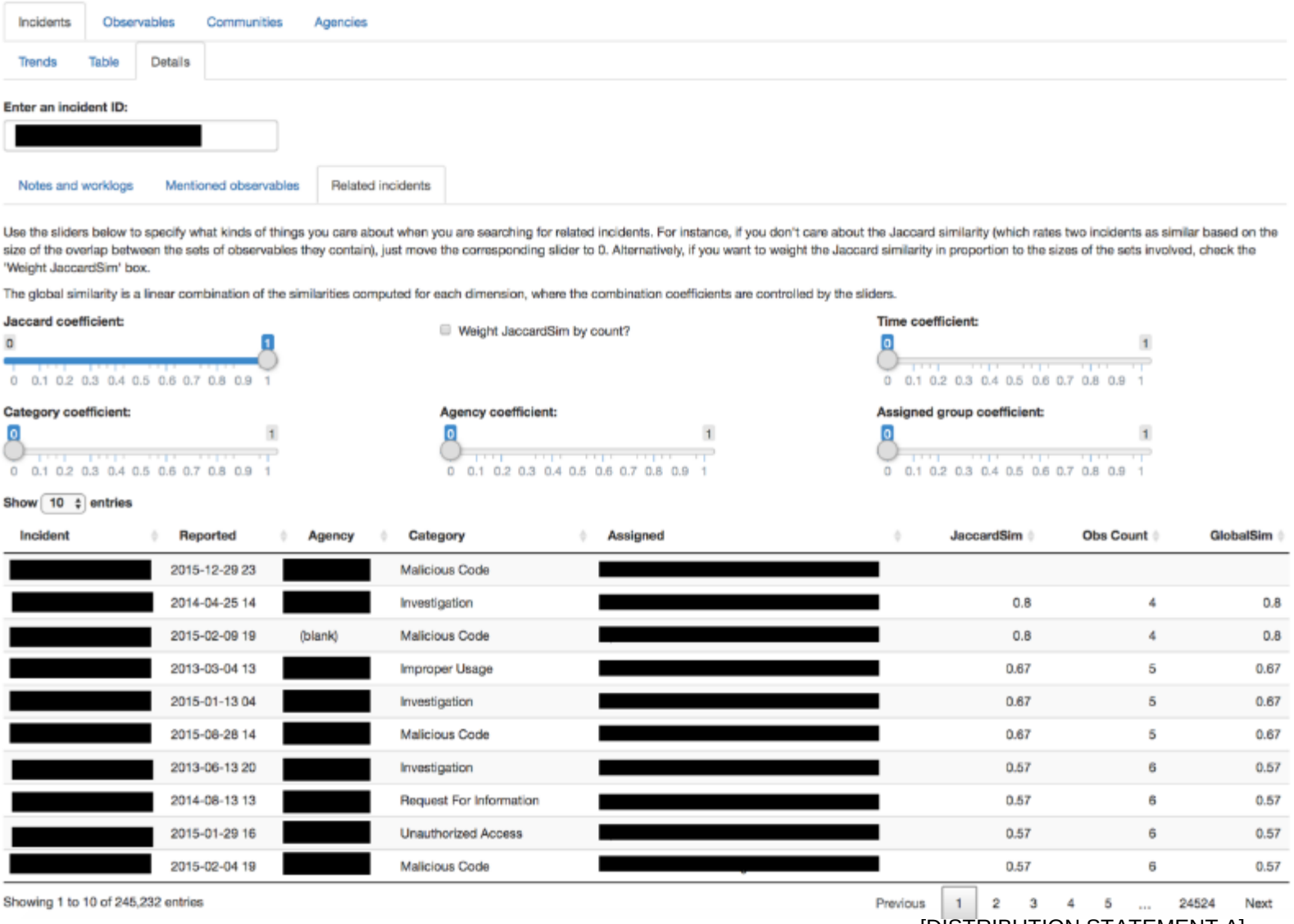
This notion is partly subjective. Incidents may be similar in one context such as ’the same victim’ but not in others such as ‘the same attacker was involved’.

Our interface allows the similarity measurement to be dynamically updated according to the analysts chosen definition.

Example definitions of similar incidents:

- Same Victim (reporter)
- Same Attacker Signature (regardless of reporter)
- Balance between similar attacker signature and victim

# Similarity calculated from distance metrics allows us to display the ‘close/related’ incidents



# Similarity definition settings

Current similarity settings (and what they allow analysts to do)

**Jaccard Coefficient** – find other incidents that have overlapping observables to the first incident but that are not in many other incidents. Observables are ‘unique’ to the set of similar incidents

**Category Coefficient** – increase the weight of reports with the same category such as ‘Malicious Code’ or ‘Improper Usage’

**Reporter Coefficient -** increase the weight of reports with the same reporter (limits reports to within the same organization)

**Assigned Group Coefficient -** increaser the weight of reports assigned to the same internal group (e.g. malware analysis, CTI, etc.)

- Allows CTI to check for threat events that might be missing from a set

# Data Driven Incident Communities

Unsupervised learning of incident Communities from the data

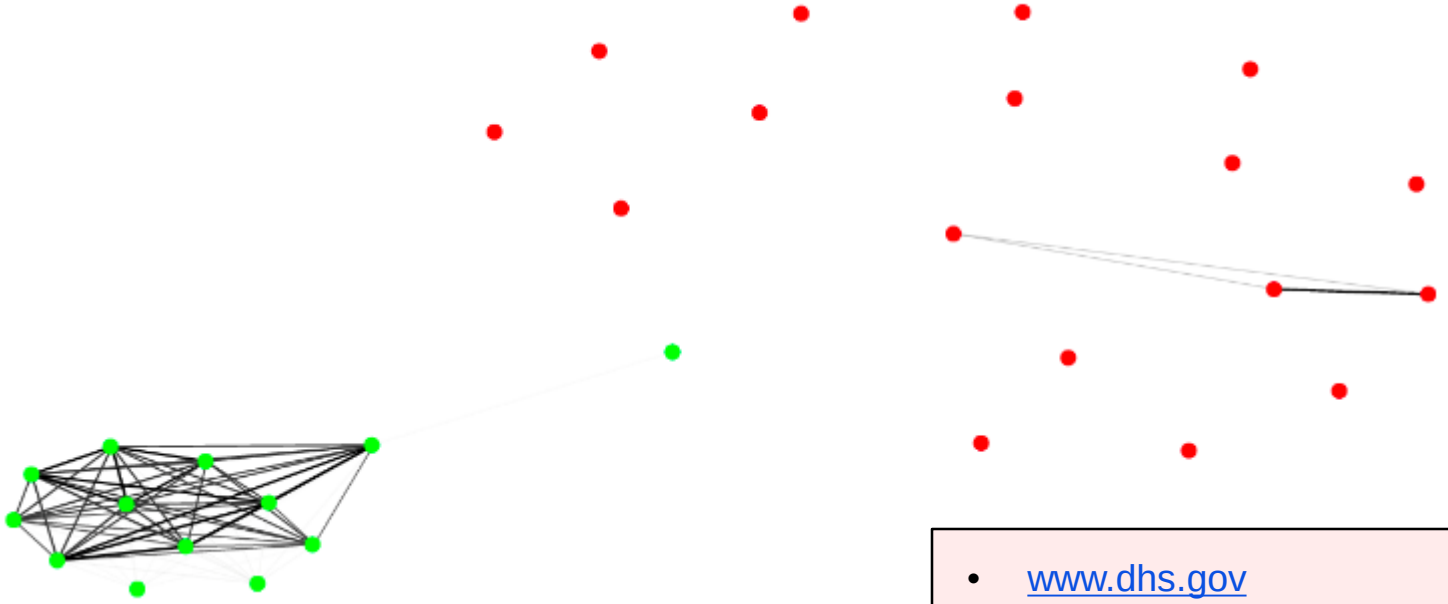
The similarity measure underlying the communities assesses the similarity between two observables based on the number of incidents that mention both observables together.

For clustering incidents, we look only at the number of observables that both incidents mention.

Uses the fast-greedy community detection algorithm which detected identified 3720 clusters of related observables covering 3 years of data.

# Community detection – Example communities

Community detection algorithms find groups of vertices that are interconnected



- MD5
- 3 phishing email addresses
- Filename
- File paths

- www.dhs.gov
- virus-submit US-CERT email
- https://www.dhs.gov/blog/2013/05/08/protecting-your-personal-information-secure-passwords

# Table of Communities with Metrics

Incidents | Observables | Communities | Agencies

Incidents | Observables | Incident-observable relationships

Overview | Clusters table | Cluster detail

### Observables clusters/communities table

Each row of the table summarizes a single cluster of observables.
**Peak time** is how many days ago the rate of appearance of these observables peaked.
**Most recent** is how many days ago a member of this cluster was last mentioned in an incident.

The table also displays the most common agency, observable type, and incident category for each cluster along with the corresponding percentages of cases that these made up.

Show 10 entries

| cluster id | # of incidents | Peak time | Most recent | Top agency | % agency | Top type | % type | Top category | % category |
|---|---|---|---|---|---|---|---|---|---|
| 1 | [redacted] | 742.9 | [redacted] | [redacted] | 100 | fqdn | 73 | Scans/Probes/Attempted Access | 89 |
| 2 | | 1165.7 | | | 99 | fqdn | 51 | Investigation | 45 |
| 3 | | 1219.7 | | | 12 | fqdn | 64 | Investigation | 38 |
| 4 | | 614.3 | | | 97 | email | 53 | Unauthorized Access | 49 |
| 5 | | 1164.9 | | | 88 | email | 50 | Investigation | 85 |
| 6 | | 1242.5 | | | 99 | fqdn | 52 | Scans/Probes/Attempted Access | 70 |
| 7 | | 1169.9 | | | 99 | fqdn | 69 | Investigation | 59 |
| 8 | | 1261.6 | | | 97 | fqdn | 50 | Malicious Code | 56 |
| 9 | | 893.1 | | | 97 | fqdn | 51 | Improper Usage | 53 |
| 10 | | 1054.6 | | | 100 | fqdn | 100 | Investigation | 64 |

Showing 1 to 10 of 3,720 entries

Previous 1 2 3 4 5 … 372 Next

# Drill down into a community to see related threat activity across multiple reports

Incidents | Observables | Communities | Agencies

Incidents | Observables | Incident-observable relationships

Overview | Clusters table | Cluster detail

**Enter a cluster id (or select a row of the clusters table):**

1

**Show** 10 **entries**

| observable_value | time_incident_reported | dco_abbrev | IncidentCategory | incident_id | observable_type | blacklist_count | APT_count | cluster |
|---|---|---|---|---|---|---|---|---|
| [redacted] | 2015-12-18T11:33:51Z | (blank) | | [redacted] | ipv4addr | | | 1 |
| | 2015-12-18T11:33:51Z | (blank) | | | ipv4addr | | | 1 |
| | 2015-12-18T11:33:51Z | (blank) | | | ipv4addr | | | 1 |
| | 2015-12-18T11:33:51Z | (blank) | | | ipv4addr | | | 1 |
| | 2015-11-30T20:12:55Z | [redacted] | Scans/Probes/Attempted Access | | ipv4addr | 1 | | 1 |
| | 2015-10-28T16:49:43Z | | Scans/Probes/Attempted Access | | ipv4addr | | | 1 |
| | 2015-10-28T16:35:05Z | | Scans/Probes/Attempted Access | | ipv4addr | | | 1 |
| | 2015-10-28T05:54:57Z | (blank) | Investigation | | ipv4addr | 2 | | 1 |
| | 2015-10-28T05:54:57Z | (blank) | Investigation | | ipv4addr | | | 1 |
| | 2015-10-28T05:54:57Z | (blank) | Investigation | | ipv4addr | 2 | | 1 |

Showing 1 to 10 of 13,253 entries

Previous 1 2 3 4 5 ... 1326 Next

# Communities can be ranked by features

Each detected community has measurable characteristics, allowing for sorting, searching and labeling by analysts.

| group | email | filename | filepath | fqdn | ipv4addr | ipv6addr | md5 | regkey | sha1 | ssdeep | url | useragent | indcount | mindcount | sindcount | badcount | tcount | agcount |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2 | 58 | 382 | 37 | 1150 | 8270 | 0 | 78 | 0 | 2 | 0 | 1248 | 8 | 11233 | 948 | 574 | 3237 | 891 | 49 |
| 844 | 687 | 760 | 145 | 3325 | 5588 | 2 | 575 | 0 | 13 | 3 | 348 | 533 | 11979 | 5402 | 2051 | 1076 | 8171 | 62 |
| 955 | 196 | 905 | 1393 | 686 | 6984 | 0 | 111 | 142 | 60 | 0 | 44 | 18 | 10539 | 1244 | 605 | 807 | 2407 | 39 |
| 1066 | 22378 | 501 | 397 | 7342 | 1753 | 20 | 663 | 1 | 86 | 18 | 275 | 19 | 33453 | 2318 | 1106 | 2988 | 2255 | 45 |
| 1177 | 3805 | 663 | 165 | 1456 | 2324 | 3 | 113 | 7 | 24 | 3 | 133 | 24 | 8720 | 2134 | 754 | 663 | 12663 | 83 |
| 1288 | 12 | 34 | 58 | 313 | 669 | 0 | 32 | 0 | 0 | 0 | 19 | 41 | 1178 | 452 | 304 | 206 | 372 | 20 |
| 1399 | 404 | 570 | 325 | 2503 | 3145 | 3 | 337 | 5 | 61 | 0 | 642 | 189 | 8184 | 2221 | 904 | 1200 | 2860 | 50 |
| 1510 | 93 | 326 | 567 | 925 | 1654 | 2 | 187 | 9 | 38 | 2 | 92 | 67 | 3962 | 1533 | 593 | 414 | 3003 | 52 |

Benefits

- Analysts can quickly rank communities reporting to have lots of the indicators they are most interested in.
- Reports from different victims and times but about similar threats can be reviewed by analysts at the same time.

# Impact of Information Discovery

## Benefits of Data Cleaning and Interactive Visualization

- Transforms data that already exists into more useful form
- Highlights existing elements for better prioritization
- Easier to show events of interest, summarize trends, identify irregularities, develop connections, and reduce duplicative analysis work.

## Improvements over the current state

- Understand the context of tickets over longer time periods
- Check untested assumptions
- Look for trends and patterns

# Future Work

## Technology Transition

- Open sourcing of analysis prototype
- Plan transition strategy to operational environments

## Model Communities over time

- Predict new community development using historical paths

## Knowledge Management

- Improve tools to retain analyst knowledge about a community and re-use it over longer time periods
- Improve distribution of discovered knowledge on threats to .gov defenders
- Mine unstructured text for more re-usable knowledge

# Transition Artifacts

- Identifying the Root Causes of Propagation in Submitted Incident Reports (FIRST 2014)
- Discovering patterns of activity in unstructured incident reports at scale (FIRST 2015)
- Data Mining for Efficient Collaborative Information Discovery (WISCS 2015 held at ACM CCS)
- Measuring Similarity Between Cyber Security Incident Reports (FIRST 2017)
- Improving Useful Data Extraction from Cybersecurity Incident Reports (FIRST 2017)

# Contact Information


**Points of Contact**

Samuel Perl, CERT

Senior Member of the Technical Staff

Information Discovery Tech Lead

Telephone: +1 412.268.4112

Email: sjperl@cert.org

Robin Ruefle, CERT

Senior Member of the Technical Staff

Information Discovery Team Lead

Telephone: +1 412.268.

Email: rmr@cert.org