Convocatoria Colaboración Público-Privada CPP2021-009025

DICYME:
Dynamic Industrial CYberrisk Modelling based on Evidence

Iniciativa Conjunta de:

Universidad Rey Juan Carlos

ENTREGABLE 1.2:
Primer prototipo de módulo de extracción automática de datos externos

Coordinadores:

Romy R. Ravines (DeNexus Tech)
Isaac Martín de Diego (Universidad Rey Juan Carlos)
Alberto Fernández Isabel (Universidad Rey Juan Carlos)

Contenido

Contenido 3
2. INTRODUCCIÓN Y OBJETIVOS 4
3. EXTRACCIÓN AUTOMÁTICA DE EVIDENCIAS EXTERNAS 4
4. INDICADORES DEFINIDOS Y SIGNIFICADO 5
5. DATOS Y CÓDIGO DISPONIBLES 6
6. ANEXO 8
Cyber Incidents Free Databases 8
Theat Actors Free and Commercial Databases 9
Dicyme Cyber Incidents Database Structure 9
Dicyme Cyber Incidents Indicators 10
Dicyme Gitlab Repository 10
Dicyme Threat Actors Indicator 11

1. INTRODUCCIÓN Y OBJETIVOS

Este documento resume los aspectos más importantes de diseño e implementación del entregable 1.2 del Proyecto, es decir, del primer prototipo de módulo de extracción automática de datos externos. Este entregable es fruto del trabajo realizado en las tareas 1.2, 1.3 y 1.4 del Proyecto, dentro de la Actividad 1 "Sistema de recogida de evidencias".

En las siguientes secciones se explica:

- Qué tipo de datos se van a utilizar en este primer prototipo y cómo se van a recopilar para que puedan transformarse, posteriormente, en información útil como entrada para el cálculo de probabilidades o impactos.
- Qué flujos de evidencias de ciberriesgo provenientes de diferentes fuentes externas relacionadas con las organizaciones y el contexto de ciberamenazas se han tenido en cuenta para este primer prototipo.
- Qué tipo de transformaciones de estos datos en información útil para el modelado del ciberriesgo industrial u OT se han propuesto.
- Qué datos y código se encuentran disponibles en este entregable (deriskGroup / DICyME Project · GitLab) y cómo pueden emplearse.

2. EXTRACCIÓN AUTOMÁTICA DE EVIDENCIAS EXTERNAS

Para la extracción automática de evidencias externas se han valorado tres tipos de datos diferentes: listados de incidentes, actores y datos firmográficos de organizaciones.

Existen multitud de colecciones de datos que recopilan información sobre ciberincidentes acontecidos en diferentes organizaciones a lo largo del mundo. En primer lugar, se realizó una selección de 9 fuentes de información sobre incidentes por su estandarización, agregación y fiabilidad de la información. De dicha selección, destacan 5 por su enfoque en incidentes en entornos industriales: *ICSStrive*, *TISafe*, *OTCad*, *Scidmark* y *CIRWAs*. Los otros 4 restantes, que recopilan información de incidentes con impacto en compañías de todos los sectores, son *Hackmageddon*, *Kon briefing*, *Jam Cyber* y el *CISSM* de la Universidad de Maryland. Como parte del entregable, se pueden encontrar los tres primeros procedimientos automáticos para la obtención de información de incidentes, realizados para las siguientes fuentes:

- Base de datos *ICSStrive*: centrada en sistemas de control industrial. Disponible en https://icsstrive.com/.

- Base de datos de ciberataques del *Center for International and Security Studies* (CISSM) de la Universidad de Maryland: generalista, disponible en https://cissm.liquifiedapps.com/.
- Base de datos de ciberataques de *Kon briefing Research*: generalista, mantenida por la mencionada compañía alemana, disponible en https://konbriefing.com/en-topics/cyberattacks.html.

Por otro lado, con el fin de matizar esta información y poder realizar posteriormente distintos análisis sobre la popularidad de una compañía, los sectores de operaciones más castigados por los ciberdelincuentes, los países, etc., se ha desarrollado un procedimiento de extracción automática de datos firmográficos de empresas: país, categoría de operación, facturación, ganancias, si cotiza en bolsa y el número de empleados. Para ello, se ha utilizado la fuente pública CompaniesMarketCap, que tiene indexadas 7,973 empresas de todo el mundo. Esta página web proporciona, además de los datos obtenidos ya mencionados, otros datos como la capitalización bursátil o el precio de las acciones para empresas que cotizan en bolsa.

La última categoría de datos externos que se han recogido en este primer prototipo es información sobre actores de amenazas. Este tipo de información puede ser muy útil en aspectos, como para analizar el contexto y panorama actual del cibercrimen, o también para agregar con la información anterior sobre incidentes (en los que se haya hecho atribución y se pueda relacionar la información por actor). La fuente de datos utilizada para obtener información sobre los actores es la enciclopedia de actores de amenazas desarrollada por la *Electronic Transactions Development Agency*, disponible públicamente en https://apt.etda.or.th/cgi-bin/aptgroups.cgi. Esta fuente de datos proporciona información como los diferentes nombres asociados a los actores, los países en los que están establecidos, su motivación, la fecha de su primera aparición registrada, una descripción, los sectores y países a los que han atacado, las diferentes herramientas y software que emplean, así como diversas campañas conocidas atribuidas a los mismos.

3. INDICADORES DEFINIDOS Y SIGNIFICADO

Para que los datos en bruto extraídos tal y como se ha explicado en la sección anterior sean de utilidad para la cuantificación del ciber riesgo, deben transformarse en información útil en forma de métricas e indicadores. En este primer prototipo disponemos de la siguiente lista de indicadores:

Nombre del indicador	**Expresión para su cálculo**
Evolución del número de ciber ataques	Conteo del número de ciber ataques por año

Evolución del número de ciber ataques por región	Conteo del número de ciber ataques por año por región: North America, South America, EMEA, Asia, Rest.
Evolución del número de ciber ataques por industria	Conteo del número de ciber ataques por año por código NAIC (1 digito).
Evolución del número de ciber ataques por tipo de incidente	Conteo del número de ciber ataques por año por tipo de incidente: ransomware, data breach, denial of service, destruction.
Evolución del número de actores por fecha de la última actividad	Conteo del número de actores totales registrados con actividad en el mes.
Evolución del número de actores nuevos	Conteo del número de actores nuevos registrados en el mes

4. DATOS Y CÓDIGO DISPONIBLES

En el enlace deriskGroup / DICyME Project · GitLab se pueden encontrar los siguientes elementos, dentro del directorio *E.1.2 First prototype of automatic external data extraction module*:

1. *Cyber_incidents*: algoritmos de extracción de datos sobre incidentes de las bases de datos *ICSStrive*, la del *CISSM* de la Universidad de Maryland y la de *Konbriefing*, así como la información resultante.
    1.1. *Code*: contiene los algoritmos de extracción de datos, uno por cada fuente de datos, y Análisis Exploratorio de Datos (EDA) de los resultados.
    1.2. *Output*: contiene, por cada fuente de datos, los resultados en formato dataframe de Pandas, exportado como ficheros .zip para minimizar sus tamaños. Se incluyen varios ficheros para cada fuente, con diferentes fechas de extracción.
2. *Etda_cyber_threat_actors*: algoritmos de extracción de datos sobre actores de amenazas, así como la información resultante.
    2.1. *Code*: contiene el algoritmo de extracción de datos, con la normalización de países y territorios objetivo, sectores de operación, y demás información sobre los actores proporcionada en la base de datos utilizada (*ETDA*).
    2.2. *Data*: como entrada se emplean el código HTML de la página web de la base de datos, así como el fichero JSON estructurado que la misma ofrece. Aunque la mayor parte de la información de los actores se obtiene de los datos estructurados, se emplea el HTML para saber si los actores son grupos APT o no, ya que dicho dato no se encuentra en el JSON. También se dispone un archivo .CSV con los diferentes países y las fronteras de cada uno de ellos.
    2.3. *Output:* el algoritmo genera un archivo .CSV con la información estructurada de cada actor, normalizando los países donde actúa, una descripción, la última vez en que la base de datos modificó la información,

herramientas y *malware* que emplea, la última vez que se vio, diferentes sobrenombres asociados, etc.

3. *Etl_firmographics*: algoritmos de extracción de datos firmográficos de compañías, basado en la información obtenida en el apartado 1 de las bases de datos del CISSM de la Universidad de Maryland y la de Konbriefing. En este primer prototipo no se ha completado con datos firmográficos los resultados de ICSStrive, puesto que no proporciona un campo estructurado del que se pueda obtener fácilmente la víctima del incidente. Tan sólo se puede obtener esta información de la descripción del incidente tras un proceso de análisis de textos y procesamiento del lenguaje natural.

    3.1. *Code*: se incluye un módulo de Python con diversas funciones que obtienen los datos de *CompaniesMarketCap*, así como realizar la extracción completa de los datos de una compañía si se utiliza como módulo principal. También se proporciona un notebook que toma como entrada los datos de salida del apartado 1 y busca víctima a víctima sus datos firmográficos, para posteriormente generar los mismos ficheros de salida con los nuevos datos agregados.

    3.2. *Output*: directorio estructurado igual que el apartado 1.2, de la extracción de datos de incidentes. Por cada fuente de datos (de entre las que proporcionan la víctima del incidente), se encuentran los mismos ficheros con las nuevas columnas añadidas de datos firmográficos.

Cabe aclarar algunos aspectos sobre el entorno de desarrollo empleado para generar el prototipo. Se han utilizado:

- `Python 3.9.5`
- `Databricks Runtime 12.2 LTS`
- `Apache Spark 3.3.2`
- `Pandas 1.4.2`
- `Numpy 1.21.5`
- `Beautiful Soup 4.12.3`
- `Matplotlib 3.5.1`

Para poder ejecutar este código sin errores se recomienda usar el mismo entorno que el de desarrollo, arriba señalado.

Cabe destacar que este es un esfuerzo continuo ya que cada semana se visitan las páginas web y se actualizan los datos recogidos. Los datos están siendo guardados en una base de datos MongoDB en los servidores de la URJC.

4. CONCLUSIONES Y SIGUIENTES PASOS

En conclusión, este trabajo ha desarrollado con éxito un primer prototipo que utiliza tres algoritmos de ETL para extraer y procesar datos de actores de ciber amenazas, ciber incidentes y datos firmográficos de compañías.

Los resultados preliminares indican que el prototipo es capaz de identificar y extraer evidencias externas de ciber incidentes de manera eficiente. Además, la integración de datos firmográficos de las compañías permite un análisis más profundo y contextualizado de los ciber incidentes, lo que puede mejorar la precisión de cálculos posteriores de indicadores o métricas de DeRisk, como por ejemplo los conceptos de popularidad o atractivo que se están desarrollando paralelamente.

Por último, aunque los resultados son prometedores, es importante recordar que este es un primer prototipo y se requiere más trabajo en esta rama. En concreto, los equipos van a valorar fuentes de datos diferentes, que aporten más información especialmente con los datos firmográficos, pues la base de datos pública utilizada no contiene muchas de las empresas afectadas. También analizarán otras fuentes de datos externas que proporcionen información sobre el contexto, actores y víctimas que permitan completar y matizar más aún los datos extraídos en este primer prototipo.

5. ANEXO

Cyber Incidents Free Databases

Industrial Sector

ICSStrive

Home - ICSSTRIVE

Created by babs4104

{et_pb_section fb_built="1" _builder_version="4.16" global_colors_info="{}"][et_pb_row column_structure="1_2,1_6,1_6,1_6" use_custom_gutter="o...

ICSSTRIVE

TISafe

Incident Hub – Industrial Cybersecurity Incidents Database

OTCAD

Operational Technology Cyber Attack Database

GitHub - SecuraBV/OTCAD: Operational Technology Cyber Attack Database

SCIDMARK

- Systems and Cyber Impact Database MARKup

http://search.infracritical.com/

CIRWAs

Critical Infrastructures Ransomware Attacks

Critical Infrastructure Ransomware Attacks (CIRA)

All Sectors/Industries

Hackmageddon

Home

Created by Paolo Passeri

0 MEGA BREACHES TRACKED IN 2023 0 B RECORDS LEAKED IN 2023 0 EVENTS RECORDED IN 2023 0 MEGA BREACHES TRACKED IN 2023 0 B RECORDS LEAKED IN 2023 0 EVENTS RECORDED IN 2023

HACKMAGEDDON

Kon briefing

The terrifying list of cyber attacks worldwide 2024 / 2023 today | KonBriefing.com

The comprehensive guide to cyberattacks - USA, Canada, UK, France & worldwide. With map and statistics.

Jam Cyber

List of Successful Cyber Attacks and Data Breaches

Updated on Dec 1, 2022

A comprehensive list of companies who have fallen victim to a successful cyber attacks and data breaches over the past 10 years.

Jam Cyber

University of Maryland CISSM Cyber Attacks Database

Welcome to the CISSM Cyber Attacks Database. The database was last updated January 11th, 2024 .

cissm.liquifiedapps.com

Others

Database - EuRepoC: European Repository of Cyber Incidents

EuRepoC data EuRepoC database Our data is consistently updated as new information about cyber incidents emerges. Consequently, coded data about specific incidents may change over time. To ensure reproducibility of analyses, we provide 'static' versions of our datasets extracted at different intervals. These can be downloaded directly from this page. Structure of the data files... Read More...

EuRepoC: European Repository of Cyber Incidents

Theat Actors Free and Commercial Databases

Crodwstrike

Crowdstrike Threat Landscape: APTs & Adversary Groups

Explore your threat landscape by choosing your APTs and Adversary Groups to learn more about them, their origin, target industries and nations.

crowdstrike.com

Google

APT Groups and Operations

README General Information How to Search in this Spreadsheet? Topic Comment Motive Cyber security companies and Antivirus vendors use different names for...

Google Docs Open preview

Secureworks

Cyber Threat Group Profiles: Their Objectives, Aliases, and Malware Tools

Explore the latest threat group definitions and profiles published by the Secureworks® Counter Threat Unit™ (CTU) Research Team.

www.secureworks.com

OTX AlienVault

AlienVault - Open Threat Exchange

Learn about the latest cyber threats. Research, collaborate, and share threat intelligence in real time. Protect yourself and the community against today's emerging threats.

AlienVault Open Threat Exchange

AlienVault - Global Adversaries

MISP Galaxy

https://github.com/MISP/misp-galaxy/blob/main/clusters/threat-actor.json

ETDA

All groups - Threat Group Cards: A Threat Actor Encyclopedia

466 groups listed (382 APT, 50 other, 34 unknown)

apt.etda.or.th

Unit 42

UNIT 42 PLAYBOOK VIEWER

pan-unit42.github.io

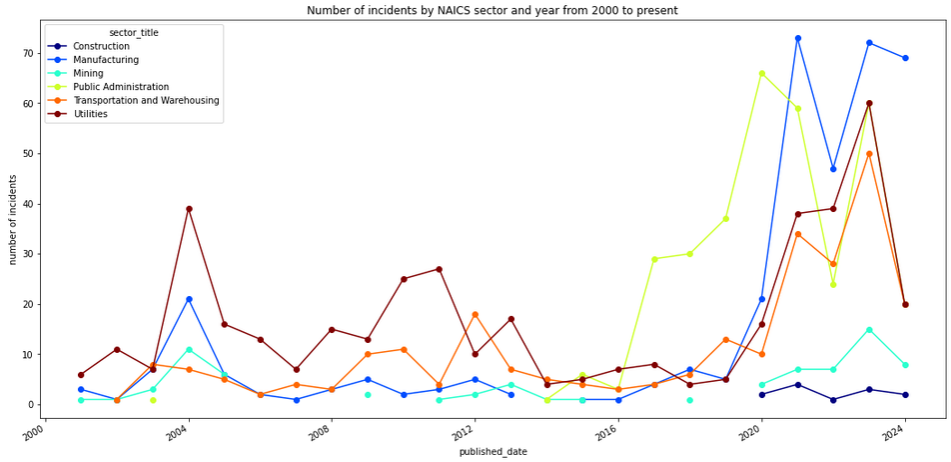
Others

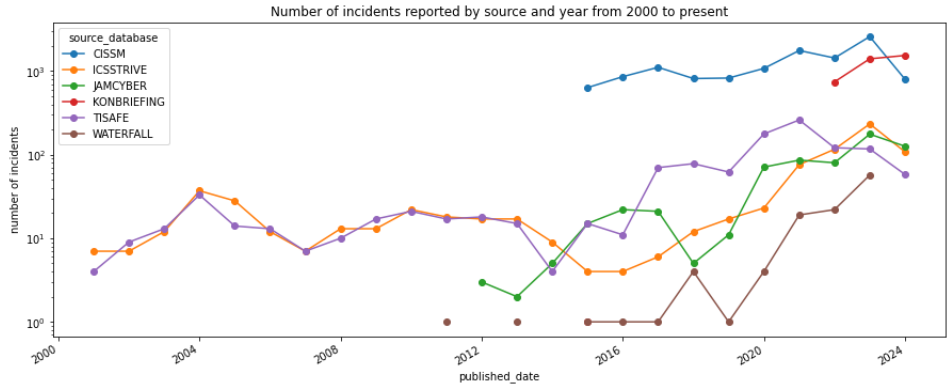
- Killing The Bear
- GitHub - StrangerealIntel/EternalLiberty for aliases
- https://warnerchad.medium.com/apt-threat-actor-lists-752df1673bc5 for resources
- Advanced Persistent Threat (APT) Groups & Threat Actors

Dicyme Cyber Incidents Database Structure

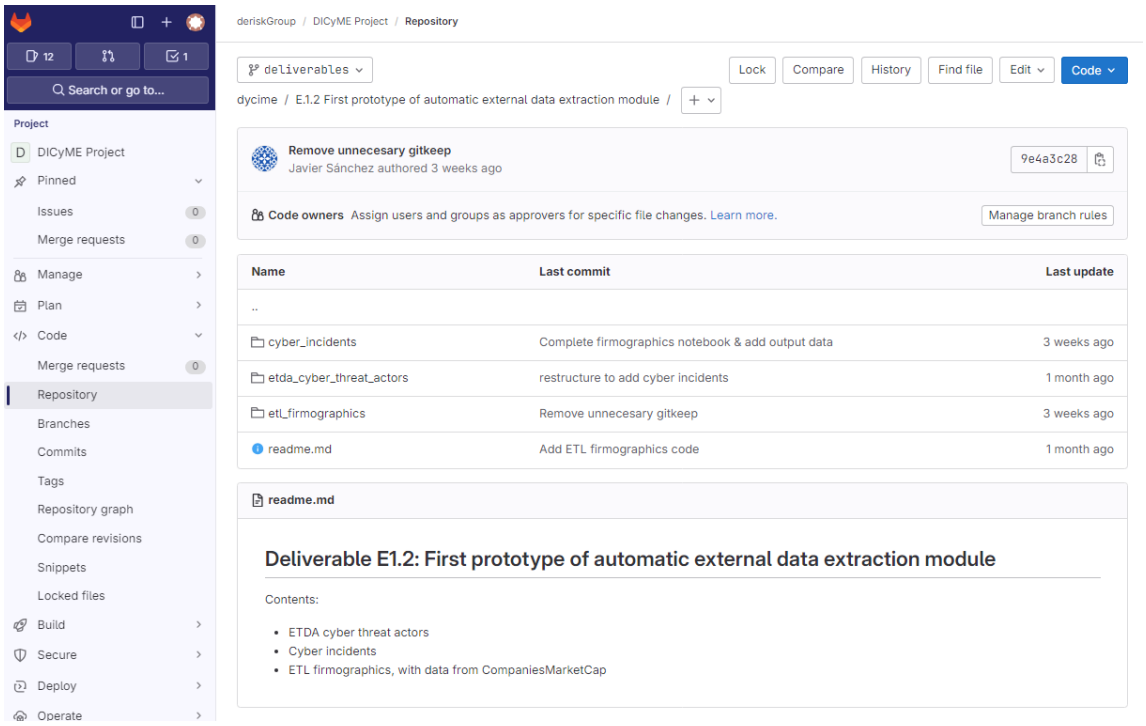
Field	Description
id	Global id. of the incident in this table. String.
description	Description of the attack. It can be any information about the attack. This field should be always present.
date	DD-MM-YYYY. Date of ocurrence.
year	YYYY. Year of ocurrence.
published_date	DD-MM-YYYY. Date of publication.
published_year	YYYY. Year of publication.
target_country	List of ISO 2-digits code of target countries. They should be the places where the attack took place. Headquarters of the victim organization are also allowed.
industry	Target industry code of the victims. In the future, it'll be a NAICS code.
victim	Target organizations or companies.
type_of_attack	https://www.crowdstrike.com/cybersecurity-101/cyberattacks/most-common-types-of-cyberattacks/
threat_source	Cyber threat actor(s).
malware	Malware name(s)
impact	Description with information about the impact or consequences of the attack. It can be a list of multiple descriptions.
tisafe_score	The TISAFE score
references	List of URL to the Internet with more information and references.
source_database	Source database.
source_database_incident_id	Incident ID in the source database.
source_database_incident_timestamp	Upload timestamp of this incident or modification date in the source database.

Dicyme Cyber Incidents Indicators





Dicyme Gitlab Repository



Dicyme Threat Actors Indicator