Cygnus: Programari d'anàlisi i rastreig de grans volums de dades de la *dark web*

Miquel Monge Dalmau

3 de juliol de 2023

Resum–Amb prop de 2,7 milions d'usuaris i 800.000 dominis *onion* únics, la prominència dels serveis *onion* a la *dark web* ha augmentat significativament. Aquests serveis són essencials per a la investigació de delictes informàtics, ja que allotgen contingut il·lícit. Malauradament, els motors de cerca tenen dificultats per indexar aquest contingut, obligant els investigadors a utilitzar dades desactualitzades i amb un abast limitat. Per afrontar els reptes de col·lecció i anàlisi de grans volums de dades no estructurades, presentem *Cygnus*, un programari modular de rastreig i anàlisi dissenyat per als serveis *onion*. *Cygnus* monitoritza activament els serveis de la *dark web* i integra les dades en una base de dades Elasticsearch. Desplegat el maig de 2023 durant tres hores, *Cygnus* ha analitzat 28.402 pàgines web *onion*. Les principals troballes ofereixen una visió completa de l'intercanvi de correu a la *dark web*, l'ús de criptomonedes, la topologia web i el contingut allotjat en aquests serveis.

Paraules clau– Anàlisi de la dark web, Intel·ligència Cibernètica, Rastreig Web.

Abstract–The prominence of onion services in the dark web has soared with nearly 2.7 million users and 800,000 unique onion domains. These services, known for hosting illicit content, have become indispensable for e-crime investigation and threat intelligence. However, contemporary search engines still struggle indexing this content, leaving researchers reliant on outdated or manually-collected datasets of limited scope. To address the challenges of gathering and analyzing vast volumes of unstructured data, we present Cygnus, a modular crawling and analysis software designed for onion services. Cygnus actively monitors dark web services and integrates the collected data into an Elasticsearch database. Deployed in May 2023, within a three-hour timeframe, Cygnus analyzed 28.402 crawled onion webpages. Our key findings provide comprehensive insights into dark net mail exchange analysis, cryptocurrency usage, web graph analysis and primary content characteristics.

Keywords– Dark web analysis, Cyber Intelligence, Web crawling.

1 Introducció

Amb aproximadament 2.7 millons d'usuaris diaris, 784.596 dominis *onion* i la naturalesa de la xarxa Tor, aquesta s'ha convertit en l'eina més utilitzada pels cibercriminals per distribuir i intercanviar diversos tipus de contingut il·lícit, convertint-se en la principal via del crim electrònic [1] [2]. Els dominis *onion* són un tipus de servei privat de la xarxa Tor que ofereixen anonimat i privadesa amagant la ubicació o l'adreça IP de la xarxa de l'usuari. Van ser desenvolupats l'any 2004 com una forma de garantir una comunicació segura a través d'internet. Actualment, per les seves característiques, gran part dels usuaris d'aquesta tecnologia la utilitzen per allotjar diversos tipus de continguts, que van des de blogs personals fins a mercats de compra-venda d'actius i serveis il·lícits.

Des de la seva creació, els dominis *onion* han experimentat una gran acollida quant a serveis i usuaris. Segons les mètriques oficials del projecte Tor, per gener de 2023, la xarxa Tor[1] comptava amb 780.000 dominis *onion*, 200.000 dominis més que el gener de 2021, servint tràfic a 2.7 millons d'usuaris diaris a una taxa d'entre 8-9 GBps [3]. Cal remarcar que sempre que parlem

• E-mail de contacte: miquel.monge@autonoma.cat
• Treball tutoritzat per: : Eduardo Cesar Galobardes (Arquitectura de Computadors i Sistemes Operatius).
• Curs 2022/23

[1] Xarxa Tor: xarxa de comunicacions anònima i segura que emmascara la identitat i la ubicació dels usuaris mitjançant el xifrat i el reenviament de la informació a través de múltiples servidors.

d'aquests serveis ens referim a la seva versió v3, que té una longitud de 56 caràcters, en comptes de 16 caràcters de la versió v2. La principal diferència és que l'última versió fa ús de l'algorisme de *hash SHA-3* per derivar el domini de la clau pública del servei ocult, mentre la versió antiga utilitza el *SHA-1*. A més, la nova versió utilitza algorismes criptogràfics més nous i segurs que la versió v2, així com algorismes de codificació més eficients [4]. La versió v2 va ser declarada obsoleta el juliol de 2021 i, per tant, no és inclosa en els objectius d'anàlisi d'aquest projecte [5]. A títol d'exemple, es facilita l'adreça *onion* v3 de Facebook per a la seva pàgina allotjada a la xarxa Tor: *http://facebookwkhpilnemxj7asaniu7vnjjbiltxjqhye3mhbshg7kx5tfyd.onion/*.

Com a principal eina per a molts continguts il·lícits, els serveis *onion* s'han convertit en un objectiu valuós per als professionals de la ciberseguretat i els equips de Ciber intel·ligència. No obstant això, trobar aquells serveis amb contingut veritablement valuós i recopilar conjunts de dades amb volum i qualitat és una tasca extremadament difícil. Aquest problema es deu a diverses propietats dels serveis *onion*. En primer lloc, aquests dominis són difícils de recordar pels humans i formen part de la *dark web*, la qual només és accessible a través de la xarxa Tor mitjançant clients especialitzats com el navegador Tor. Una altra característica que dificulta la seva recopilació a gran escala és que la connexió a través del circuit Tor es fa connectant-se a diversos nodes, formant un camí llarg que compta amb menys amplada de banda que accedir a pàgines web de la *clear net*[2]. Això provoca un augment de la latència i un rendiment baix en sacrifici de major anonimat. A més, aquests serveis són privats per defecte, el que significa que els usuaris han de descobrir de manera orgànica els 56 caràcters aleatoris que formen el servei *onion*. Els principals mètodes per a la descoberta inclouen el boca-orella, la navegació a través de pàgines web enllaçades, directoris web o motors de cerca especialitzats en la xarxa Tor. Tanmateix, és important destacar que, en els dos últims casos, sempre estarem limitats per la fiabilitat de la tercera part que manté la pàgina. Això pot fer-nos víctimes d'atacs de *phishing* o de dependre de continguts molt desactualitzats.

El monitoratge d'aquests serveis i la detecció d'amenaces representen un desafiament important per a les organitzacions, ja que requereixen una gran quantitat de temps, esforç i recursos. No obstant això, són tasques essencials per a garantir la seguretat de l'empresa, dels clients i de les seves dades. En aquest context, s'ha desenvolupat *Cygnus*, una solució de programari completa que ofereix un enfocament integral *end-to-end*[3] que inclou un *crawler*[4] de dominis *onion* destinat a ajudar a les organitzacions a adreçar aquest risc sense haver d'exposar-se a contingut potencialment nociu ni als seus responsables.

[2] *Clear Net*: És la part d'internet a la qual es pot accedir a través dels navegadors web convencionals com Chrome o Firefox

[3] *End-to-end*: Procés de gestió complet de les dades des de la seva recopilació inicial fins al seu processament i anàlisi final.

[4] *Crawler: Programa automatitzat que recorre la web per recopilar informació sobre els llocs web.*

2 Estat de l'art

La majoria dels autors relacionats amb investigacions sobre *crawlers* de la xarxa Tor són investigadors en el camp de la ciberseguretat. Una part minoritària inclou projectes desenvolupats per estudiants. En ambdós casos, la majoria dels treballs van ser publicats abans de juliol de 2021, quan les adreces *onion* v2 varen ser deprecades. Això ha provocat que molts d'aquests treballs quedin obsolets avui dia i requereixin adaptacions o revisions per fer-los funcionar en el context actual de la xarxa Tor. A més d'aquests autors, també existeixen solucions professionals privades que es comercialitzen a organitzacions, institucions o governs que volen abordar aquest risc.

A la taula 1, s'han identificat els treballs que s'han considerat més complets o madurs segons la seva tipologia i característiques:

TAULA 1: Tipologia i característiques de treballs relacionats

Font	Treball	Codi Font	*Docs*
Open source	TorBot	✓	
	Ahmia	✓	
Papers	Dizzy		✓
	BlackWidow		✓
Organitzacions	DarkOwl		✓
	Digital Shadows		✓

2.1 Treballs *open-source*

Dins d'aquest apartat es troben tots aquells repositoris de *Github* que ofereixen el codi font per implementar un *crawler*. Cal tenir en compte que molts d'aquests repositoris presenten algunes limitacions com la manca de documentació, desactualització o errors en el seu contingut. Tot i així, aquests recursos són valuosos per observar que la majoria de *crawlers* Tor de codi obert són desenvolupats en el llenguatge de programació *Python* i utilitzen les llibreries *Scrapy* o *Selenium* per la tasca d'*scrapejar*[5] el contingut *html*. D'entre aquests projectes, destaca especialment el disseny de *TorBot*, el qual implementa de forma modular les diferents parts de codi encarregades de recopilar informació del domini *onion*, com ara adreces IP, correus electrònics, etc. [6] [7]. D'altra banda, *Ahmia* és un motor de cerca que indexa dominis *onion* en la base de dades *ElasticSearch*, una base de dades *NoSQL* dissenyada i optimitzada per a realitzar cerques ràpides [8] [9].

La limitació d'ambdós treballs és que no compten amb una estratègia per consumir aquesta informació. En ambdós casos, simplement la indexen en una base de dades, a diferència de *Cygnus*, que ofereix un servei *as a Service*[6] per a les organitzacions interessades a consumir, analitzar i prendre decisions amb la intel·ligència recopilada.

[5] *Scraper:* Programa informàtic dissenyat per extreure informació d'un lloc web de forma automatitzada.

[6] *As a Service:* Model de lliurament on el programari es proporciona i distribueix a través d'Internet mitjançant subscripcions, en lloc d'instal·larho i mantenir-ho localment.

2.2 *Papers*

S'inclouen aquells treballs publicats per estudiants o, en la seva majoria, per investigadors. La majoria d'aquests treballs només presenten la metodologia per desenvolupar el *crawler*, mentre que una minoria també inclouen el codi font.

Un dels treballs més destacables és *Dizzy*, elaborat per dos membres de l'institut d'Investigació de Computació de Qatar. Tot i no disposar del codi font, aquest article proporciona una descripció completa de la metodologia utilitzada per superar les limitacions i les circumstàncies úniques que s'han exposat a la introducció a l'hora d'implementar un *crawler* a la xarxa Tor [10].

D'altra banda, *BlackWidow*, publicat per sis professionals de la ciberseguretat, presenta el disseny *end-to-end* d'una solució *software* destinada a monitorar contingut de la *dark web*. La principal diferència amb *Cygnus* és que el gruix de *BlackWidow* és l'anàlisi estàtica de la informació [11].

2.3 Solucions empresarials

Hi ha empreses dedicades exclusivament a aquest camp, així com altres que ofereixen aquest servei com a part de les seves capacitats principals. Pel que fa al primer grup, destaca *DarkOwl* com la principal empresa dedicada exclusivament a aquest àmbit. *DarkOwl* és una empresa dels Estats Units especialitzada en la recopilació de contingut de la *dark web* i compta amb eines eficients per a la localització de dades sensibles filtrades o compromeses. Encara que el seu codi font, les metodologies i el desenvolupament estan ocults, *DarkOwl* és una font d'anàlisi i investigació molt valuosa per comprendre les capacitats i recursos d'una de les eines més grans del món en aquest àmbit. [12].

3 Objectius

Després d'analitzar la necessitat, la problemàtica i l'estat de l'art, aquest treball presenta *Cygnus*, un *crawler multithreading* dissenyat per recopilar, processar i distribuir la intel·ligència obtinguda *as a Service* per a la seva anàlisi. Aquest treball té com a objectiu contribuir en l'avenç del coneixement a l'àrea de la seguretat informàtica, proporcionant una eina efectiva per a l'exploració i anàlisi de la tipologia i organització de la dark web. A més, s'espera que *Cygnus* faciliti la identificació de patrons i tendències en l'activitat dels dominis onion. A continuació, es defineixen una sèrie d'objectius:

- Proveir a les organitzacions d'un programari capaç de quantificar el risc i l'exposició de les corporacions en la *dark web* i en serveis *onion*.
- Crear una eina eficient per identificar i adreçar de manera àgil i proactiva les potencials amenaces de la *dark web*, adaptada a les necessitats i casos d'ús específics de les organitzacions.
- Minimitzar l'impacte negatiu causat per atacs cibernètics o filtracions de dades.
- Generar intel·ligència de la *dark web* per habilitar a les organitzacions a millorar contínuament les seves defenses de ciberseguretat contra amenaces, incompliments i intrusions.
- Proveir una plataforma per a l'anàlisi de manera senzilla, segura i exhaustiva a través d'una interfície d'ús amigable, amb capacitats de cerca i supervisió avançades per accedir i monitorar informació crítica.

La diferència més gran entre aquest treball amb altres que tracten aquesta temàtica és l'accés a la intel·ligència generada pel *crawler*. En aquest treball, tot el contingut indexat i processat s'oferirà *as a service*. *Cygnus as a Service* (CaaS) serà un *APOC ("a proof of concept")*[7] d'una interfície web amb *dashboards* interactius sobre tota la intel·ligència que s'hagi recopilat d'una determinada organització.

Una altra diferència respecte a la majoria de treballs publicats sobre aquesta temàtica és que *Cygnus* permetrà mesurar i identificar de manera tangible l'exposició a la *dark web*, utilitzant mesures agregades que utilitzaran variables com adreces IP, correus corporatius i moneders de criptomonedes entre altres.

4 Disseny i arquitectura

4.1 Tria de *seed urls*

Per poder començar a *crawlejar* els dominis *onion* de la *dark web*, les aranyes (*Spiders*) han de començar per un conjunt de pàgines anomenades *seed url*. A partir d'aquestes, s'aniran seguint els seus enllaços per explorar tots els dominis *onion* que l'*engine* vagi trobant.

Definir correctament les '*URL* inicials' o '*URL* d'inici' en un *crawler* és de vital importància, ja que són la base des de la qual es comença el procés de *crawleig* i extracció de dades. En el context d'un *crawler* de la dark web, s'han escollit com a *seed urls* pàgines anomenades *'list indexers'*, essent la més coneguda la *Hidden wiki* [13] [14] o *tor.taxi* [15]. Aquestes pàgines són especialment beneficioses, ja que solen incloure enllaços a diversos dominis *onion*, fet que augmenta la cobertura del *crawling*. Alguns dels avantatges que ofereixen aquestes pàgines són:

- **Cobertura:** contenen una gran quantitat d'enllaços a dominis *onion*. Es pot augmentar la cobertura del *crawleig*, permetent descobrir més llocs web a la web fosca a partir d'un conjunt més petit de *seed urls*[8].
- **Escalabilitat:** es pot aprofitar l'estructura jeràrquica de la web fosca, on una sola pàgina pot tenir enllaços a nombrosos dominis *onion*. Això permet assolir una major visibilitat de llocs web amb menys temps i esforços, fent que el procés de *crawleig* sigui més eficient.

La figura 1 mostra un exemple d'una de les pàgines del conjunt de *seed url* seleccionat on s'evidencia l'estructura característica d'una d'aquestes pàgines anomenades *'list indexers'*.

[7] *APOC:* fa referència a una demostració o prova d'un concepte. Es tracta d'una demostració inicial que té com a objectiu verificar la viabilitat d'una idea o tecnologia en un entorn determinat.

[8] *Seed Url:* És la *URL* inicial que s'utilitza per iniciar el procés de *crawling*. A partir de la *seed url*, el *crawler* recorre totes les pàgines del lloc web per recopilar informació.

Fig. 1: Exemple d'una pàgina *onion* del tipus *'list indexer'*

4.2 Arquitectura

Per a la solució al problema plantejat, l'arquitectura de la solució *software* ha de complir uns requisits definits en les tasques d'aquest projecte:

- ***Middleware***[9] **personalitzable**: ha de ser capaç d'establir una connexió segura amb el circuit *Tor* de forma transparent i segura.
- **Maneig de sol·licituds i respostes**: gestió automàtica del maneig de sol·licituds i respostes *HTTP*, incloent-hi el seguiment d'enllaços i el maneig de redireccions.
- **Extracció de dades**: conjunt d'eines per a l'extracció de dades web, com selectors *CSS* i *XPath*, que permeten extreure dades de manera eficient i precisa dels dominis *onion*.
- **Gestió de *pipelines***: Ha de ser capaç de definir *pipelines* de processament de dades per a realitzar accions com neteja, validació o emmagatzematge de les dades extretes en diversos formats.
- **Programació asincrònica**: Ha de ser capaç de suportar la programació asincrònica, la qual cosa permet dur a terme múltiples sol·licituds simultàniament i de manera eficient, accelerant el procés d'extracció de dades.
- **Integració amb emmagatzematge i exportació de dades**: Ha de ser capaç de fer persistent la informació recopilada amb bases de dades *NoSQL*, en aquest cas *ElasticSearch*.

Per assolir aquests requisits s'ha utilitzat el *framework* d'aplicació per rastrejar llocs web i extreure dades estructurades anomenat *Scrapy*. La figura 2 mostra una visió general de l'arquitectura, els seus components i un esquema del flux de dades (que es mostra amb fletxes vermelles) [16].

[9] *Middleware:* Capa d'intermediació entre les peticions del client i les respostes del servidor. El *Middleware* s'executa abans que les peticions arribin als controladors o a les rutes de l'aplicació web.

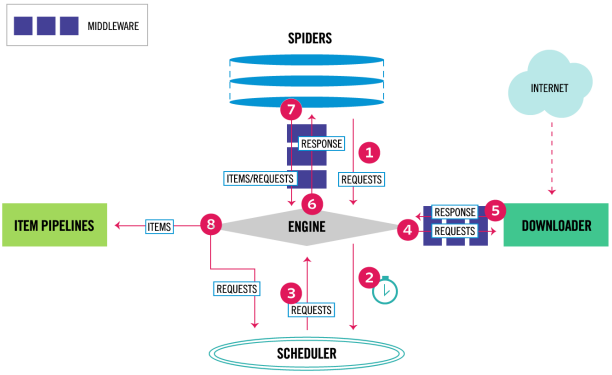

Fig. 2: Disseny de l'arquitectura

4.2.1 *Data Flow*

El flux de dades a *Scrapy* està controlat pel motor d'execució i segueix el següent ordre [17]:

1. L'*Engine* rep les sol·licituds inicials per *crawlejar* de l'*Spider*.
2. L'*Engine* programa les sol·licituds de l'*Scheduler* i demana les properes sol·licituds per *crawlejar*.
3. L'*Scheduler* retorna les següents sol·licituds a l'*Engine*.
4. L'*Engine* envia les sol·licituds al *Downloader*.
5. Un cop finalitzada la descàrrega del domini *onion*, el *Downloader* genera una resposta (d'aquesta pàgina) i l'envia a l'*Engine*.
6. L'*Engine* rep la resposta del *Downloader* i l'envia a l'Aranya perquè la processi.
7. L'Aranya processa la resposta i torna els elements *scrapejats* i les sol·licituds noves (properes a seguir) a l'*Engine*.
8. L'*Engine* remet els elements processats a *Item Pipelines*. A continuació remet les sol·licituds processades a l'*Scheduler* i demana possibles properes sol·licituds per *crawlejar*.
9. El procés es repeteix (a partir del pas 3) fins que no hi hagi més sol·licituds de l'*Scheduler*.

5 Components

El disseny del *framework* segueix una arquitectura modular composta per diversos components principals. Els principals components de l'arquitectura de *Scrapy* són:

- ***Engine*:** el motor s'encarrega de controlar el flux de dades entre tots els components del sistema i de desencadenar esdeveniments quan es produeixen determinades accions.
- ***Scheduler*:** el planificador rep les peticions de l'*Engine* i les posa en cua per alimentar-les més tard, quan l'*engine* les sol·licita.

- ***Downloader*:** és l'encarregat de descarregar les pàgines *onion* i alimentar-les a l'*engine* que, al seu torn, alimenta les aranyes.

- ***Spiders*:** les aranyes són classes personalitzades per analitzar les respostes i extreure'n elements o sol·licituds addicionals a seguir.

- ***Pipelines*:** s'encarrega de processar les respostes un cop hagin estat extretes (*scrapejades*) per les aranyes. Les tasques inclouen la neteja, la validació i la persistència (com emmagatzemar l'element en una base de dades).

- ***Middleware*:** permet efectuar operacions de processament de sol·licituds i respostes, podent afegir funcionalitats adicionals al flux de treball d'*scrapping*. El *Middleware* ha de permetre dirigir de forma automàtica totes les peticions a través del circuit de la xarxa de *Tor* per poder establir una connexió satisfactòria amb els dominis *onion* que operen en la *dark web*. Així doncs, aquesta classe actua com un intermediari entre el motor de *Scrapy* i el servidor de destinació. En lloc de realitzar directament les peticions *HTTP* al servidor de destinació, el *middleware* utilitza el protocol *SOCKS5* per encaminar la sol·licitud a través de la xarxa *Tor* [18]. Amb l'objectiu d'incrementar la seguretat i privadesa durant la col·lecció de dades, s'ha implementat un mecanisme amb el qual, de forma automàtica, el *middleware* renovarà la direcció *IP*, és a dir, crearà un nou circuit de la xarxa *Tor* cada 'x' nombre de peticions. D'aquesta manera, el circuit *Tor* s'anirà renovant cada cop que el *software* hagi *scrapejat* i emmagatzemat a la base de dades *ElasticSearch* 'x' dominis *onion*.

  La figura 3 descriu el funcionament de la xarxa *Tor* per entendre a quin circuit encamina el *middleware* les peticions.

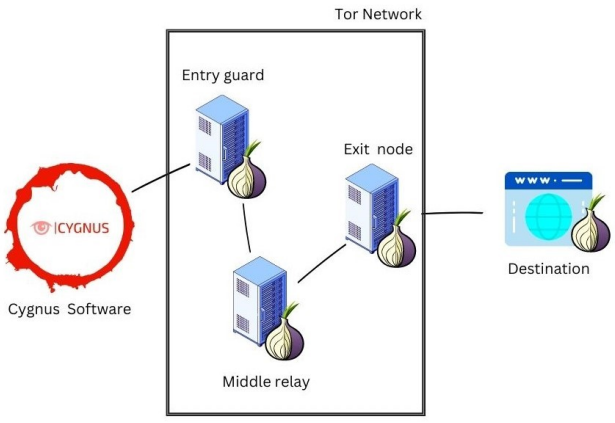

Fig. 3: Fragment de la implementació del *TorProxyMiddleware*

De tots els components, cal destacar la implementació del Middleware, de l'Spider i dels Pipelines per adaptar-los al context de la *dark web* i els dominis de *onion*.

5.1 Implementació de les aranyes

Les aranyes són les encarregades de començar a *crawlejar* a partir del conjunt de *seed urls* ja definit. A més, com es mostra a la taula 2, les aranyes hauran de fer una primera extracció de les dades del domini *onion*.

TAULA 2: ELEMENTS EXTRETS PER LES ARANYES DE *Cygnus*

Element extret	Descripció	Tipus
Address	L'adreça *onion v3* de 56 caràcters de longitud de la pàgina.	*String*
Title	El títol de l'adreça *onion*.	*String*
Description	La descripció de l'adreça *onion*.	*String*
Keywords	Les paraules claus de l'adreça *onion*.	*Array*
h1	La capçalera *HTML* que té com a ús més comú marcar el títol de la pàgina web	*Array*
h2	Etiqueta *HTML* utilitzada per als subtítols i que permet separar-ne el contingut.	*Array*
Robot	Arxiu utilitzat per indicar als *bots* o motors de cerca quines pàgines o seccions d'un web no han de ser indexades	*Array*
Download time	Latència o temps de resposta de la descàrrega d'una resposta d'una sol·licitud *HTTP*.	*Numeric*
Crawled	La data en què el domini web s'ha indexat per *Cygnus*.	*Datetime*
Body	Contingut en text del domini *onion* un cop netejat de tots els tags *HTML* i *Javascript*.	*Array*
Onion links	Tots els hipervincles *onion v3* que conté la pàgina que s'està indexant. Permetrà crear el graf on cada web serà un node i estarà unit per aquestes arestes.	*Array*

La majoria d'elements extrets per les aranyes de *Cygnus* s'extreuen a partir d'expressions *regex*. En cas que el valor que es vulgui extreure no estigui present, aquell camp es deixarà amb el valor *'None'* [19].

L'altre objectiu de les aranyes és seguir un *crawleig* tipus *snowball*. Aquest tipus d'aranyes comencen amb una llavor o *seed url* i a partir d'aquestes utilitzen tècniques d'extracció d'*URL* per seguir enllaços trobats a les pàgines web visitades, la qual cosa resulta en l'addició de més *URL* a la llista d'enllaços a rastrejar. A mesura que es recopilen més *URL*, el nombre d'enllaços a seguir augmenta exponencialment, de manera similar a una bola de neu. Això permet que el cercador descobreixi i recopili més i més pàgines web en un curt període de temps. Per tal d'implementar aquest enfocament s'han tingut en compte dos punts:

- **Extensions malicioses:** És segur que una pàgina *onion* podrà contenir enllaços maliciosos o que no es vulguin descarregar o indexar (pornografia, arxius filtrats de grups de *ransomware*, extensions que contenen programari maliciós, etc). Per aquest motiu, s'ha implementat un mecanisme que ignora aquells enllaços que puguin contenir informació maliciosa. Per fer-ho, s'han definit un conjunt d'*strings 'stop words'* que, si estan presents a l'enllaç que es vol seguir, aquest enllaç s'omet i no s'indexa per *Cygnus*.
- ***Depth Limit:*** S'ha definit un límit per a la profunditat de navegació de *Cygnus*. En aquest cas, s'ha establert un límit de 2 capes de profunditat; *Cygnus* seguirà els enllaços a la pàgina inicial (nivell 0) i a les pàgines enllaçades des de la pàgina inicial (nivell 1), però no seguirà els enllaços a les pàgines enllaçades des de les pàgines enllaçades des de la pàgina inicial (nivell 2) o més profund. Això permet controlar la profunditat del rastreig i limitar la quantitat de dades recopilades en una estructura de lloc web específica.
- **Enllaços fora de l'abast:** L'abast del projecte és fer un software per indexar enllaços *onion v3*. Tot i això, és segur que algunes pàgines *onion v3* tinguin enllaços a altres pàgines *onion v2* que ja no estan en funcionament o pàgines de la *clear web* (blogs, repositoris *github*, notícies, etc). Per evitar seguir enllaços que no siguin *onion v3*, s'ha definit una regla *regex* i només els enllaços que passin aquesta regla seran seguits.

5.2 *Pipelines*

El processament de les dades es durà a terme als *pipelines*, on s'implementaran diferents mòduls per a realitzar les transformacions i accions addicionals a les dades extretes per les aranyes. La taula 3 mostra tots els pipelines que s'han dissenyat i implementat en *Cygnus*.

1. ***Language detection:*** Identifica l'idioma principal de la pàgina utilitzant la dependència *lingua*, un repositori *open-source* capaç de detectar fins a 75 idiomes [20]. Per millorar la detecció de l'idioma s'ha fet ús de la tècnica *n-gram tokens*. Els *n-gram tokens* són seqüències de *n* paraules consecutives que es creen dividint el text en fragments superposats de *n* paraules, permetent capturar la informació contextual i les relacions gramaticals entre les paraules en un text. Això és a causa del fet que els idiomes humans generalment no es componen de paraules aïllades, sinó que les paraules solen estar agrupades en seqüències i segueixen patrons gramaticals i contextuals específics. La figura 4 il·lustra un exemple d'aquesta tècnica:

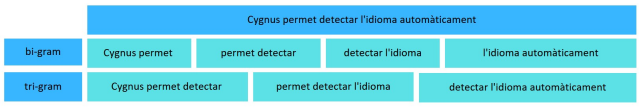

Fig. 4: Exemple de la tècnica *n-gram token*.

2. ***Crypto Intelligence:*** Aquest mòdul és l'encarregat de cercar contingut relacionat amb les criptomonedes. D'aquestes, les més utilitzades són *Bitcoin (BTC)* i *Monero (XMR)* a causa del seu anonimat. Per tant, a partir de tot el contingut en format text de la pàgina, se cerquen de manera automàtica totes aquelles cadenes de text que segueixin el patró d'una cartera (*wallet*) d'aquestes criptomonedes i s'emmagatzema com a metadades d'aquell domini *onion*. D'aquesta manera, cada domini *onion* que s'indexa conté dos atributs: *bitcoin wallets* i *monero wallets*, que contenen en forma de llista totes aquelles *wallets* que s'hagin trobat en la pàgina. En cas que no se'n trobi cap, el valor és *None*. La taula 4 mostra les expressions *regex* que s'usen per identificar les carteres de criptomonedes.
3. ***Sensitive Data:*** Aquest mòdul té com a objectiu identificar aquelles pàgines on s'hagi pogut dur a terme una filtració de dades. A partir del contingut en text de la pàgina es busca de forma seqüencial a través de patrons regex possibles dades confidencials. La figura 5 il·lustra la seqüència de cerca d'aquestes dades.

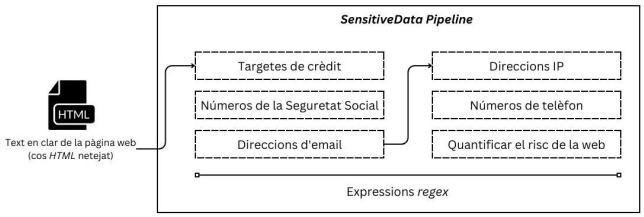

Fig. 5: Seqüència d'obtenció de dades sensibles del mòdul *SensitiveData Pipeline*.

4. ***Quantify Risk Intelligence:*** L'objectiu d'aquest mòdul és quantificar de forma objectiva el risc que suposa una pàgina *onion*. D'aquesta manera, si les dades que intentem monitorar apareixen en més d'un domini, podrem prioritzar la cerca començant per les pàgines categoritzades amb més risc. Encara que aquest risc pugui no reflectir el risc real de la pàgina, té en compte una sèrie de valors objectius per al seu càlcul. La puntuació és sobre cinc i té en compte les dades recollides per tots els mòduls/*pipelines* que s'han descrit prèviament:
   - **Moneders de criptomonedes:** Si la pàgina conté moneders, és probable que la pàgina estigui oferint serveis o dades a canvi de criptomonedes, per tant, s'eleva la puntuació del risc de la pàgina.
   - ***Risk words*:** S'han definit un conjunt de paraules anomenades *risk words* i, si estan presents en la pàgina, s'incrementa el risc d'aquesta. Alguns exemples d'aquestes paraules són: *ransom, leak, credentials, vpn acces, corporate, filtrated, etc.*
   - ***SensitiveData*:** Seguint la mateixa metodologia, com més dades considerades com sensibles tingui la pàgina, més alt serà el seu risc.
5. ***Elastic Pipeline:*** És el mòdul encarregat d'assegurar la persistència de les dades en la base de dades *NoSQL ElasticSearch*. Aquest mòdul també implementa mecanismes per assegurar que cada pàgina que s'indexa és única, per tal d'evitar entrades múltiples en la base

TAULA 3: DEFINICIÓ DELS MOTORS D'INTEL·LIGÈNCIA PER AL PROCESSAMENT DE DADES

Pipeline	***Input data***	**Tipus**	***Extracted feature(s)***
Language detection	$Inputdata = \begin{cases} \text{body[:100]}, & \text{if } \exists \text{body} \\ \text{description}, & \text{otherwise} \end{cases}$	String	ISO 639-1 del llenguatge detectat
Crypto Intelligence	Text netejat de la pàgina *onion* processada	*Array*	Moneders de les criptomonedes Bitcoin (BTC) i Monero (XMR)
Sensitive Data	Text netejat de la pàgina *onion* processada	*Array*	Targetes de crèdit Números de la Seguretat Social Direccions IP Números de telèfon
Quantify Risk Intelligence	Total de camps de dades sensibles no nuls	*Numeric*	Risc quantificat basat en la informació del domini
	Moneders Bitcoin (BTC) trobats	*Numeric*	
	Moneders Monero (XMR) trobats	*Numeric*	
	Risk words apareixen a la pàgina *onion*	*Boolean*	
	ISO 639-1 del llenguatge	*String*	
Elastic Pipeline	Pàgina *onion* amb totes les dades processades	*JSON*	Persistir les dades a la DDBB Elasticsearch

TAULA 4: FÓRMULES *regex* PER CERCAR CARTERES DE CRIPTOMONEDES

Criptomoneda	**Expressió *regex***
Bitcoin (BTC)	"[13][a-km-zA-HJ-NP-Z1-9]25,34"
Monero (XMR)	"4[0-9AB][1-9A-HJ-NP-Za-km-z]93"

de dades. Per fer-ho, s'usa el *hash* de l'adreça *onion* de la pàgina i s'utilitza com a clau única. Per assegurar l'escalabilitat, les pàgines s'indexen en *bulk*, és a dir, les operacions d'indexació es fan en lots, en comptes d'una indexació de forma individual de cada pàgina. El procés de *bulk* implica enviar un conjunt de documents (pàgines *onion*) en format *JSON* a través d'una sola trucada a l'*API bulk d'Elasticsearch* [21]. D'aquesta manera, fins que *Cygnus* no hagi *crawlejat* i *scrapejat* 200 dominis *onion*, aquests no s'indexaran. És a dir, la persistència a la base de dades es fa en lots de 200 pàgines, millorant l'eficiència del *software*. Les pàgines que encara no s'hagin indexat s'emmagatzemaran en un *buffer* fins a arribar a 200.

6 RESULTATS

En aquesta secció, es mostren els resultats de l'execució de Cygnus durant aproximadament tres hores en un ordinador amb un processador Intel Core i5-8250U 1.6GHz i 12 GB de memòria DDR4. Durant aquest temps, Cygnus ha scrapejat amb èxit 28.402 pàgines de la dark web. A continuació, es descriu l'anàlisi de la intel·ligència reconeguda per Cygnus en els següents aspectes: rendiment, *data profiling*, serveis de correu electrònic, anàlisis de contingut, ús de criptomonedes. També s'inclou un gràfic de la topologia de la xarxa Tor i un *wordcloud* per analitzar la freqüència de paraules a la dark web.

6.1 Rendiment

Cygnus ha demostrat un rendiment notable en analitzar i rastrejar un total de 28.402 pàgines en aproximadament 3 hores. Aquest rendiment es basa en la figura 6, on es mostra el nombre de pàgines analitzades en funció de l'hora corresponent. La naturalesa lineal del gràfic suggereix que no hi ha embuts en el programari actual, de manera que es podria millorar el rendiment simplement utilitzant recursos més potents, sense necessitat de modificar el programari en si mateix. Actualment, el programa analitza aproximadament 5.000 pàgines cada 30 minuts. Amb una major capacitat de recursos, el programa seria capaç de processar un volum més gran de pàgines en el mateix interval de temps, augmentant així l'eficiència i la productivitat.

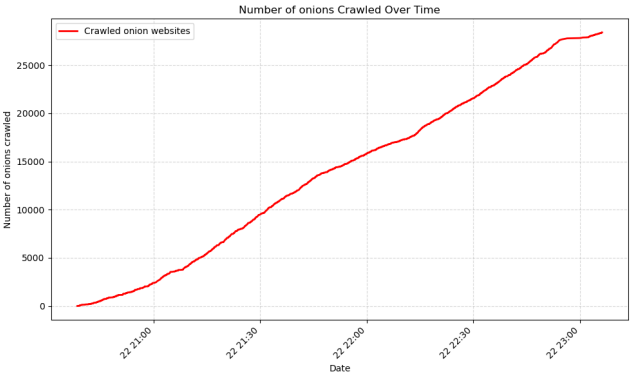

Fig. 6: Cronologia de les pàgines analitzades

6.2 *Data profiling*

S'ha realitzat un perfilat del 5% (1.420 observacions) de les dades obtingudes per Cygnus amb l'objectiu de comprendre a alt nivell el seu contingut. Aquest perfilat proporciona informació valuosa per entendre la naturalesa de les pàgines analitzades, revelar patrons i tendències destacades, i identificar relacions entre atributs. A més, permet identificar possibles relacions entre els diferents atributs i ajudar a contextualitzar el contingut en termes de risc, temes predominants o activitats associades. L'informe del perfilat de dades és en format HTML i es pot descarregar i visualitzar a través d'aquest enllaç.

6.3 *Dark Net Mail Exchange Analysis*

A la dark web, els usuaris opten per utilitzar correus electrònics anònims per preservar la seva identitat i garantir l'anonimat en les seves interaccions. Aquests correus electrònics anònims no estan associats amb proveïdors de correu convencionals com Gmail o Hotmail, fet que complica la rastrejabilitat de les comunicacions en línia. Com s'evidencia en la figura 7, és poc comú trobar adreces de correu electrònic vinculades a aquests proveïdors convencionals en l'anàlisi de la dark web. Així doncs la majoria dels dominis de correu detectats per Cygnus estan destinats a preservar l'anonimitat dels usuaris; els exemples més predominants són *dnmx.org* [22] o *protonmail.com* [23].

L'objectiu principal d'aquesta anàlisi és identificar els patrons d'ús i els proveïdors preferits de correus electrònics anònims. Aquesta investigació contribueix a la comprensió de la infraestructura i les dinàmiques de la dark web, així com a la identificació de possibles àrees de focus per a futures investigacions relacionades amb la seguretat i la lluita contra les activitats il·lícites en línia.

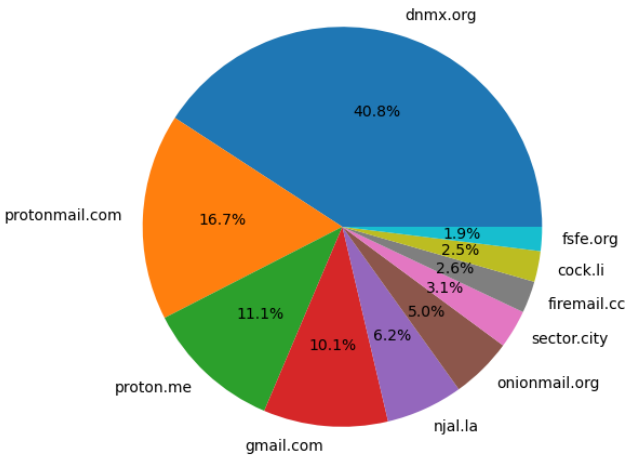

Fig. 7: Top 10 dominis de correu electrònic més freqüents en les dades de Cygnus

6.4 Anàlisi de contingut i topologia de temàtiques de la *dark web*

L'anàlisi del text de les pàgines analitzades a la dark web té com a objectiu identificar patrons, temes i característiques rellevants del contingut per a una millor comprensió de la naturalesa i les dinàmiques de la dark web. A través de tècniques de processament del llenguatge natural (NLP), s'ha generat una *wordcloud* o núvol de paraules, que visualitza les paraules més freqüents i els temes dominants en el contingut analitzat. Això permet una identificació ràpida de la topologia de paraules i les temàtiques més comunes, facilitant la detecció d'àrees d'interès per a la recerca en termes de seguretat, intel·ligència i lluita contra el cibercrim. Aquesta informació contribueix a l'aplicació de mesures de seguretat i a la lluita contra activitats il·legals en línia, millorant la resposta i la prevenció en aquests àmbits.

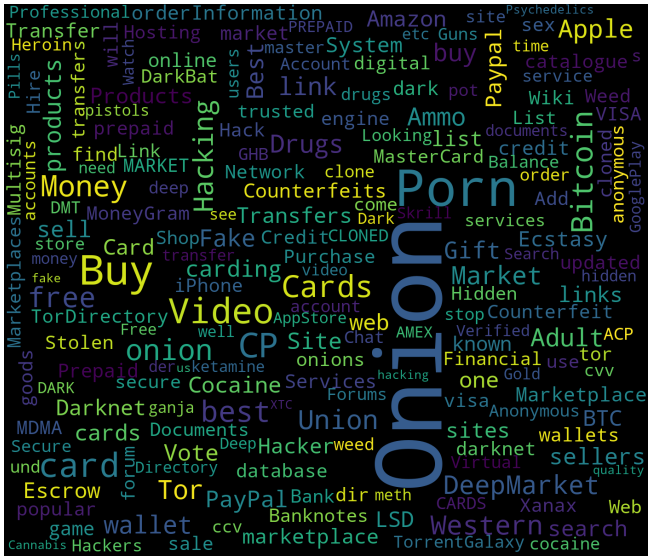

Fig. 8: Núvol de paraules de les pàgines analitzades per Cygnus

Les paraules amb més aparicions en el *wordcloud* de la figura 8 són: *Porn* (pornografia), *Buy* (comprar), *Video* (vídeo), *CP* (pornografia infantil, per les seves sigles en anglès), *Card* (targeta), *Bitcoin* (criptomoneda), *Market* (mercat), entre d'altres, que suggereixen temàtiques i característiques de la dark web que són coherents amb el que se sap fins ara i els estudis sobre aquest entorn. La majoria d'aquestes temàtiques estan vinculades a activitats il·legals com la venda de productes il·legals o serveis prohibits, incloent-hi pornografia, drogues, armes, dades de targetes de crèdit i operacions relacionades amb la criptomoneda Bitcoin (p.e: atacs *ransomware*). És especialment preocupant la presència de la paraula *CP*, que fa referència a contingut de pornografia infantil.

Aquests resultats reforcen la comprensió que la dark web és un espai on es duen a terme activitats il·lícites, es negocia amb productes il·legals i s'aprofiten vulnerabilitats en termes de privadesa i seguretat. Aquest fet subratlla la importància de les mesures de seguretat, l'aplicació de la llei i la prevenció del cibercrim per abordar i combatre aquestes activitats il·legals en línia.

6.5 Anàlisi sobre l'ús de criptomonedes

En analitzar les adreces de Bitcoin i Monero identificades per Cygnus a la figura 9, s'ha observat que les adreces Bitcoin són significativament més populars que les adreces de Monero. La direcció Bitcoin més utilitzada ha aparegut en 500 ocasions, en contrast amb la direcció de Monero més utilitzada, amb tan sols 16 aparicions.

Aquesta discrepància suggereix que els usuaris de la dark web prefereixen emprar Bitcoin en lloc de Monero com a mètode de transacció. Bitcoin és una criptomoneda àmpliament coneguda i acceptada en diversos sectors, inclosa la dark web, cosa que pot explicar la seva major adopció en aquesta comunitat. Aquesta aparició de Bitcoin també pot ser atribuïda, en part, a les estafes a la dark web. Els *scammers*[10] sovint usen Bitcoin com a mitjà de pagament a causa de la seva popularitat i reconeixement generalitzat, de manera que aquestes aparicions poden ser el resultat d'activitats fraudulentes i enganys. També és important considerar que Bitcoin té una quota de mercat significativament més gran que Monero i aquest fet fa que Bitcoin pugui ser més atractiva a l'hora d'especular amb el seu preu. Per la seva mida, Bitcoin també pot comptar amb més liquiditat, i es pot considerar que és més fàcil blanquejar aquesta criptomoneda a través de targetes/cupons d'Amazon o efectiu, entre d'altres.

D'altra banda, Monero, encara que també es considera una criptomoneda enfocada a la privadesa, pot requerir un nivell addicional de coneixement tècnic per al seu ús eficient i segur. Això podria dissuadir alguns usuaris de la dark web que busquen una opció més accessible i fàcil d'utilitzar. A més, és rellevant tenir en compte que Monero pot ser més usada en fòrums privats de la dark web que requereixen inici de sessió. Aquests fòrums solen estar orientats a usuaris més especialitzats i compromesos amb la privadesa, cosa que pot explicar perquè la direcció de Monero més usada té un nombre més baix d'aparicions en comparació amb Bitcoin. En requerir inici de sessió o resposta d'un *Captcha* o altres mecanismes similars, Cygnus no pot tenir en compte aquest tipus de pàgines.

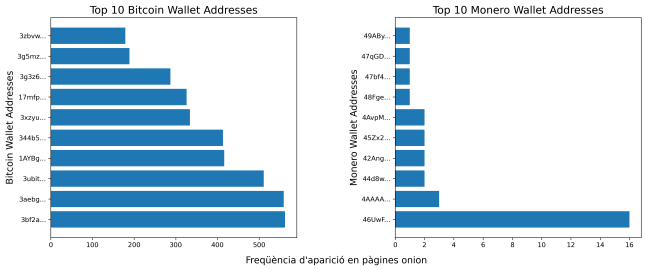

Fig. 9: Top 10 adreces Bitcoin (BTC) i Monero (XMR) amb més freqüència d'aparició en les pàgines *onion* analitzades per *Cygnus*

6.6 Graf dirigit

A partir de les pàgines analitzades, Cygnus genera de manera dinàmica un graf dirigit que representa les connexions entre les pàgines *onion* de la dark web.

El graf conté un nombre considerable de nodes, amb un total de 406.861. Aquests nodes estan connectats per 1.542.606 arestes, indicant un nivell significatiu d'hiperconnectivitat dins de la xarxa. En mitjana, cada node té un grau de 7,58, suggerint un nombre moderadament elevat de connexions per node. La densitat del graf, que mesura la proporció de possibles arestes que existeixen en el graf, és molt baixa, de 9,32e-06. Això indica que el graf és dispers, amb molts nodes, tenint un nombre relativament petit de connexions en comparació amb el nombre total de connexions possibles. Això es deu al fet que la majoria dels nodes del graf no han estat analitzats completament. Tot i que Cygnus coneixia l'existència de connexions entre pàgines i les tenia pendents d'analitzar, el programa es va aturar abans de poder continuar analitzant pàgines. La xarxa mostra un gran nombre de components fortament connectats, amb un total de 386.112 components identificats. Això suggereix la presència de nombrosos clústers o subgrafs més petits dins de la xarxa global. El component fortament connectat més gran (LSCC) està format per 4.246 nodes, representant un subconjunt cohesionat de pàgines amb connexió bidireccional. El coeficient de clusterització del LSCC és relativament alt, amb una mesura de 0,723. Aquest valor indica un nivell significatiu de clusterització o interconnexió dins del LSCC. Suggereix que els nodes d'aquest component tenen tendència a formar grups cohesionats, indicant una tendència a enllaçar amb múltiples altres pàgines dins del mateix clúster.

7 *Cygnus as a Service (SaaS)*

Dos dels principals reptes en l'accés i l'anàlisi de conjunts de dades rastrejats són la usabilitat i la flexibilitat. Tot i que les tecnologies existents com els motors de cerca analítics representen solucions amb avantatges desitjables, rarament s'apliquen a l'anàlisi de la dark web, a causa de la manca de conjunts de dades públiques i representatives. Per permetre que investigadors i analistes explorin i analitzin serveis *onion* de manera lliure i a gran escala, Cygnus presenta un prototip *Cygnus as a Service*, al qual empreses privades i investigadors podrien accedir per monitoritzar aquesta intel·ligència. La figura 10 mostra un exemple de com podria ser el prototip per accedir a aquest tipus de dades.

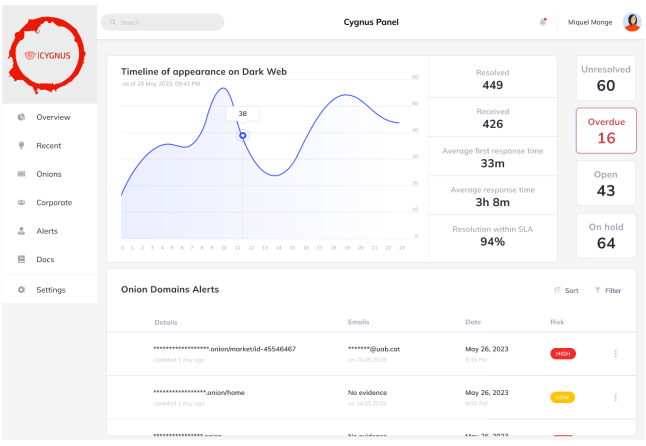

Fig. 10: Il·lustració d'exemple del prototip del portal Cygnus as a Service

8 Conclusions

En l'actual entorn de ciberseguretat, és essencial tenir una solució de recopilació, anàlisi i monitoratge en temps real que funcioni a través de diversos idiomes i que superi altres barreres com l'escalabilitat i el rendiment. Aquest

[10] *Scammer:* és una persona o entitat que es dedica a cometre estafes o fraus amb l'objectiu d'obtenir beneficis de manera enganyosa o il·legal.

article presenta Cygnus, un programari d'anàlisi i rastreig de grans volums de dades de la dark web, que demostra la viabilitat de la detecció precoç d'amenaces i tendències cibernètiques, superant diversos reptes importants.

Tot i que és probable que tècniques similars a aquesta s'estiguin utilitzant tant per actors d'intel·ligència governamentals com privats a tot el món, és rellevant entendre aquest projecte com una eina que contribueix a l'avanç del coneixement en l'àmbit de la seguretat informàtica i el control de la dark web. Així doncs, tenint en compte la manca d'informació i l'obscuritat en aquest camp, el *dataset* recopilat per Cygnus, amb dades enriquides de 28.402 pàgines web *onion*, serà posat a disposició dels investigadors que ho sol·licitin.

Mitjançant la implementació de Cygnus com a eina de recopilació i anàlisi de grans volums de dades de la dark web, es mostra que el seguiment de la web fosca es pot realitzar amb recursos relativament limitats, fent-ho accessible a un ventall més ampli d'actors en el futur. En aquest context, el desplegament de Cygnus s'ha usat per dur a terme una anàlisi exhaustiva dels serveis *onion* i presentar descobertes relacionades amb característiques dels serveis *onion* a gran escala, la seva topologia i contingut principal.

Referències

[1] Tor Project (2023) *Users, Tor Metrics*. [consultat: 20 febrer 2023]. Disponible a Internet: https://metrics.torproject.org/userstats-relay-country.html.

[2] *Onion services (2023) Tor Metrics*. [consultat: 1 març 2023]. Disponible a Internet: https://metrics.torproject.org/hidserv-dir-onions-seen.html.

[3] *Traffic (2023) Tor Metrics*. [consultat: 1 març 2023]. Disponible a Internet: https://metrics.torproject.org/bandwidth.html.

[4] Team, D.O.A. (2022) *Tor V2 deprecation shifts dark-net landscape, DarkOwl, LLC*. [consultat: 7 de març]. Disponible a Internet: https://www.darkowl.com/blog-content/tor-v2-depreciation-shifts-darknet-landscape/.

[5] *V3 onion services usage: Tor Project (2023) The Tor Project*. [consultat: 2 març 2023]. Disponible a Internet: https://blog.torproject.org/v3-onion-services-usage/.

[6] Julio San José, Fundador i director de Derechodela-Red. *(2022) Torbot - Herramienta de Osint para la dark web., Derecho de la Red*. [consultat: 23 febrer 2023]. Disponible a Interent: https://derechodelared.com/torbot-herramienta-osint-dark-web/.

[7] DedSecInside (2023) *Dedsecinside/torbot: dark web Osint Tool, GitHub*. Dedsecinside. [consultat: 25 febrer 2023]. Disponible a Internet: https://github.com/DedSecInside/TorBot.

[8] *Ahmia*. (2014) Juha Nurmi. [conusltat: 26 febrer 2023]. Disponible a Internet: https://ahmia.fi/.

[9] juha juha said: September 09 et al. (2014) *Ahmia Search after GSOC development: Tor Project, The Tor Project*. [consultat: 2 marcç 2023]. Disponible a Internet: https://blog.torproject.org/ahmia-search-after-gsoc-development/.

[10] Perera, I. and Boshmaf, Y. (2022) *Dizzy: Large-scale crawling and analysis of onion Services, arXiv.org*. [consultat: 26 febrer 2023]. Disponible a Internet: https://arxiv.org/abs/2209.07202.

[11] Martínez Monterrubio, S.M. et al. (2021) *Black widow crawler for tor network to search for ... - IEEE xplore*. [consultat: 1 marcç 2023]. Disponible a Interent: https://ieeexplore.ieee.org/document/9447337/.

[12] *dark web Monitoring and Intelligence: Darkowl (2023) DarkOwl, LLC*. [consultat: 27 febrer 2023]. Disponible a Internet: https://www.darkowl.com/.

[13] *Una semana en la deep web, tres años después, Xataka. Xataka.*. [consultat: 9 abril 2023]. Disponible a Internet: https://www.xataka.com/analisis/una-semana-en-la-deep-web-tres-anos-despues.

[14] *The Hidden Wiki*. [consultat: 9 abril 2023]. Disponible a Internet: https://thehiddenwiki.org/.

[15] *Your ride to the darknet*. [consultat: 9 abril 2023]. Disponible a Internet: https://tor.taxi/.

[16] *Architecture overview - Scrapy 2.8.0 documentation*. [consultat: 9 abril 2023]. Disponible a Internet: https://docs.scrapy.org/en/latest/topics/architecture.html.

[17] *A fast and powerful scraping and web crawling framework*. [consultat: 9 abril 2023]. Disponible a Internet: https://scrapy.org/.

[18] *Khalid (2020) Stealthy crawling using scrapy, tor and privoxy, Khalid Alnajjar*. [consultat: 9 abril 2023]. Disponible a Internet: https://www.khalidalnajjar.com/stealthy-crawling-using-scrapy-tor-and-privoxy/.

[19] *Search Engine Optimization: Crawl a website with Scrapy and get valuable SEO data.*. [consultat: 9 abril 2023]. Disponible a Internet: https://floss-marketing-school.com/mod/page/view.php?id=128.

[20] *Pemistahl. PEMISTAHL/Lingua-py: The most accurate natural language detection library for python, suitable for long and short text alike, GitHub*. [consultat: 10 abril 2023]. Disponible a Internet: https://github.com/pemistahl/lingua-py.

[21] *PJayzeng. Jayzeng/scrapy-elasticsearch: A scrapy pipeline which send items to Elastic Search Server, GitHub*. [consultat: 10 abril 2023]. Disponible a Internet: https://github.com/jayzeng/scrapy-elasticsearch.

[22] *The anonymous email service for the dark net (2022) DNMX*. [consultat 28 maig 2023]. Disponible a Internet https://dnmx.org/(Accessed:28May2023).

[23] *Proton Mail: Un Correo Electrónico Privado, Seguro Y Cifrado Proton*. [consultat 28 maig 2023]. Disponible a Internet: https://proton.me/es-es/mail.