Graphaméléon

Apprentissage des relations et détection d'anomalies sur les traces de navigation Web capturées sous forme de graphes de connaissances

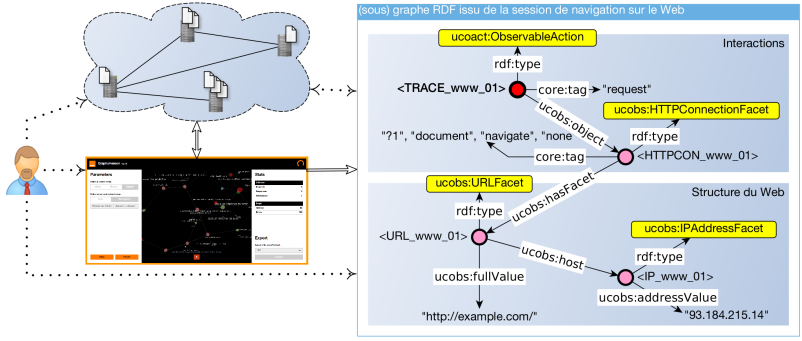
IC @ PFIA 2024

Lionel Tailhardat, Benjamin Stach, Yoan Chabot, Raphaël Troncy

Orange & EURECOM

4 Juillet 2024

D’une session de navigation Web vers un graphe de connaissances RDF ...



Contribution l’extension Web Graphaméléon

Collecte en direct au niveau du navigateur ...

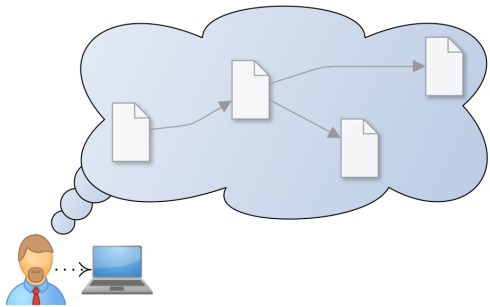
- requêtes réseau (mode `macro`)
- interactions des utilisateurs (mode `micro`)

Sortie en graphe RDF structuré par l’ontologie UCO

Applications cartographie du Web, analyse du comportement des réseaux, détection d’anomalies, ...

Vidéo : PFIA2024-Graphameleon-demo

Contexte et motivations: analyse de traces et modélisation d'activités



Scénario Session de navigation Web

Navigation Processus d'exploration itératif

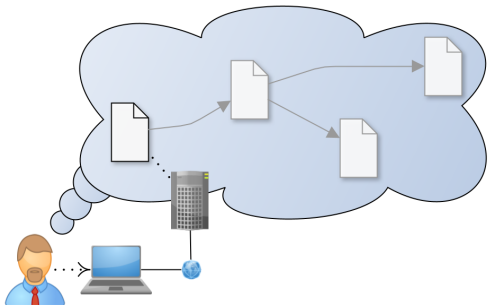
Traces Dépendant du **persona**, du **périmètre technique observable**

Util. lambda Pacte vert, vie privée

Ing. système Allocation de ressources, performance

Exp. sécurité Détection d'activités malveillantes

Contexte et motivations: analyse de traces et modélisation d'activités



Scénario Session de navigation Web

Navigation Processus d'exploration itératif

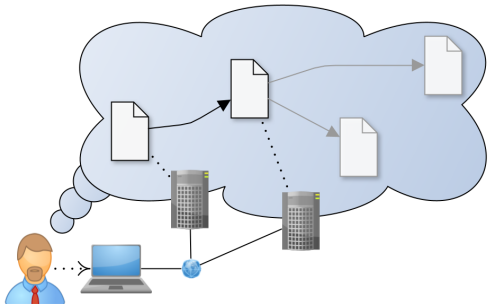
Traces Dépendant du **persona**, du **périmètre technique observable**

Util. lambda Pacte vert, vie privée

Ing. système Allocation de ressources, performance

Exp. sécurité Détection d'activités malveillantes

Contexte et motivations: analyse de traces et modélisation d'activités



Scénario Session de navigation Web

Navigation Processus d'exploration itératif

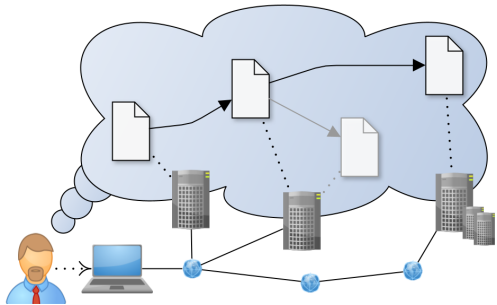
Traces Dépendant du **persona**, du **périmètre technique observable**

Util. lambda Pacte vert, vie privée

Ing. système Allocation de ressources, performance

Exp. sécurité Détection d'activités malveillantes

Contexte et motivations: analyse de traces et modélisation d'activités



Scénario Session de navigation Web

Navigation Processus d'exploration itératif

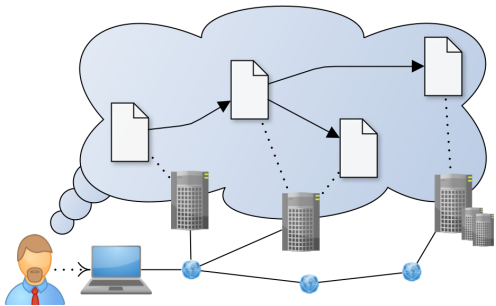
Traces Dépendant du **persona**, du **périmètre technique observable**

Util. lambda Pacte vert, vie privée

Ing. système Allocation de ressources, performance

Exp. sécurité Détection d'activités malveillantes

Contexte et motivations: analyse de traces et modélisation d'activités



Scénario Session de navigation Web

Navigation Processus d'exploration itératif

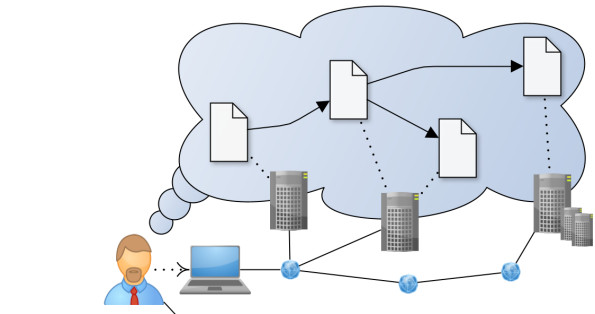
Traces Dépendant du **persona**, du **périmètre technique observable**

Util. lambda Pacte vert, vie privée

Ing. système Allocation de ressources, performance

Exp. sécurité Détection d'activités malveillantes

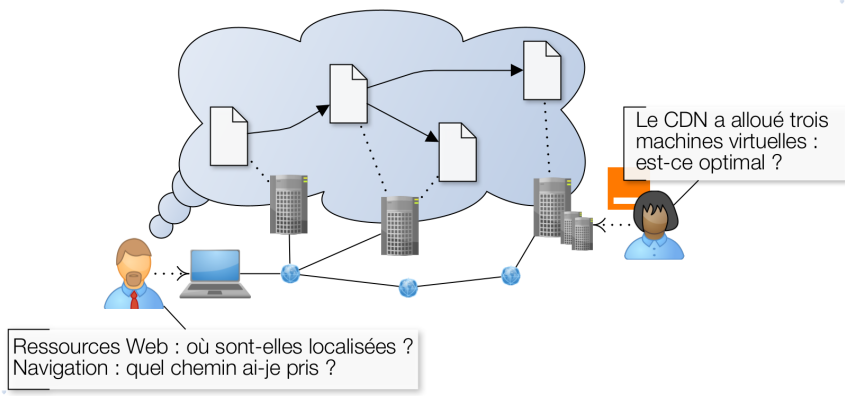
Contexte et motivations: analyse de traces et modélisation d'activités



Scénario	Session de navigation Web
Navigation	Processus d'exploration itératif
Traces	Dépendant du **persona**, du **périmètre technique observable**

Util. lambda	Pacte vert, vie privée
Ing. système	Allocation de ressources, performance
Exp. sécurité	Détection d'activités malveillantes

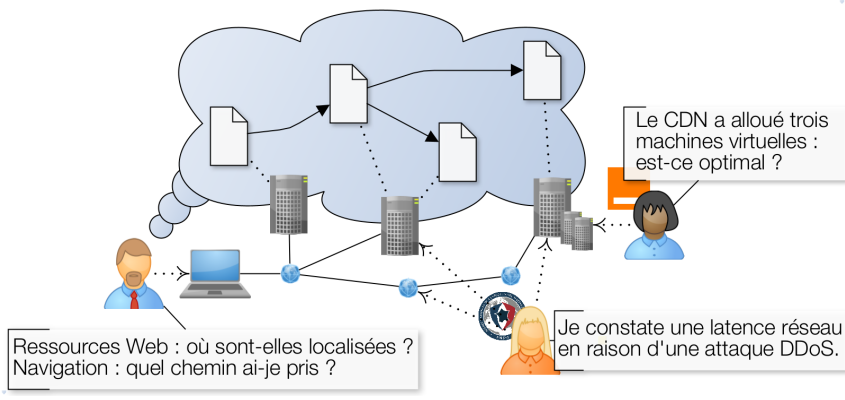
Contexte et motivations: analyse de traces et modélisation d'activités



Scénario	Session de navigation Web
Navigation	Processus d'exploration itératif
Traces	Dépendant du **persona**, du **périmètre technique observable**

Util. lambda	Pacte vert, vie privée
Ing. système	Allocation de ressources, performance
Exp. sécurité	Détection d'activités malveillantes

Contexte et motivations: analyse de traces et modélisation d'activités



Scénario	Session de navigation Web
Navigation	Processus d'exploration itératif
Traces	Dépendant du **persona**, du **périmètre technique observable**

Util. lambda	Pacte vert, vie privée
Ing. système	Allocation de ressources, performance
Exp. sécurité	Détection d'activités malveillantes

Contextualiser les actions de l'utilisateur dans la topologie du réseau ?

Traces de navigation Web	Quelles informations les traces fournissent-elles sur la structure et la dynamique du réseau par rapport aux activités des utilisateurs ?
Modèle comportemental	Comment pouvons-nous apprendre un modèle qui prend en compte à la fois les activités des utilisateurs et la structure du réseau ?

Problèmes et verrous

1. Collecte de données sur le réseau et les actions des utilisateurs
   - Trafic réseau chiffré
   - Journaux d'application privés
   - Mauvais formatage des données
2. Représentation des connaissances et raisonnement
   - Cause ou objectif non évident d'une activité dans les traces
   - Multiples niveaux d'interprétation d'une activité
   - Connaissances d'experts pour l'interprétation stockées dans des bases de connaissances tierces

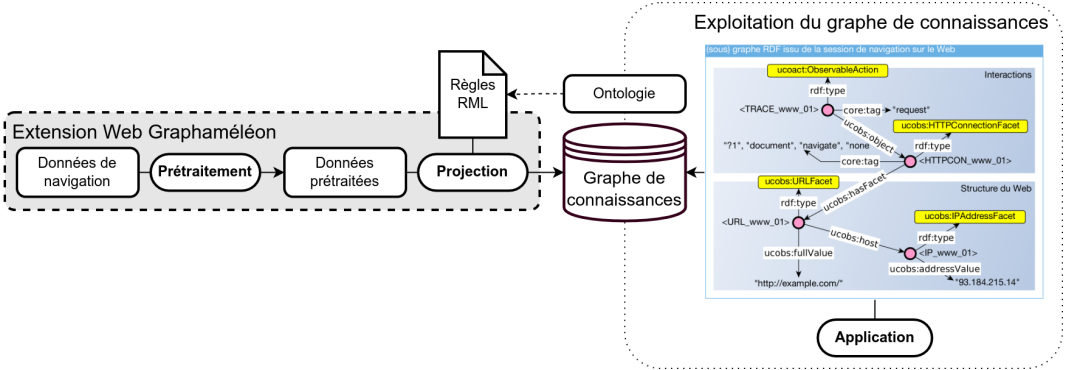
Contextualiser les actions de l'utilisateur dans la topologie du réseau ?

Traces de navigation Web	Quelles informations les traces fournissent-elles sur la structure et la dynamique du réseau par rapport aux activités des utilisateurs ?
Modèle comportemental	Comment pouvons-nous apprendre un modèle qui prend en compte à la fois les activités des utilisateurs et la structure du réseau ?

Problèmes et verrous

1. Collecte de données sur le réseau et les actions des utilisateurs
   - Trafic réseau chiffré
   - Journaux d'application privés
   - Mauvais formatage des données
2. Représentation des connaissances et raisonnement
   - Cause ou objectif non évident d'une activité dans les traces
   - Multiples niveaux d'interprétation d'une activité
   - Connaissances d'experts pour l'interprétation stockées dans des bases de connaissances tierces

Méthodologie : collecte et analyse



Collecte Extension Web Graphaméléon

- requêtes réseau (mode `macro`)
- interactions des utilisateurs (mode `micro`)

Sortie Graphe RDF structuré par l'ontologie UCO (5 / 422 concepts, 15 / 751 propriétés).

Objectifs Capture de connaissances & utilisation de modèles de traces

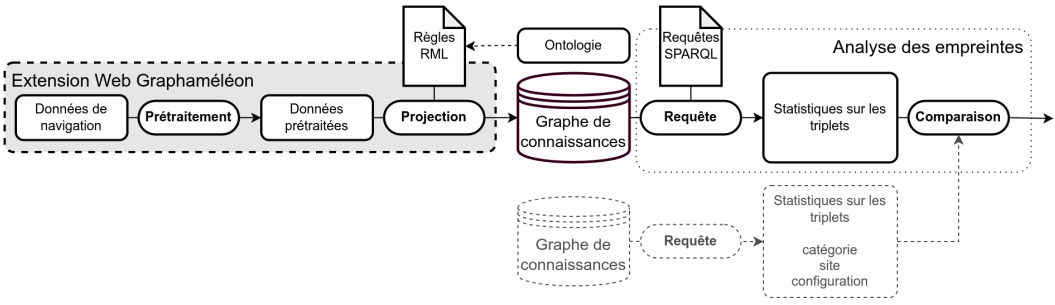
Expérimentations

1. Trafic réseau et complexité des sites Web
2. Catégorisation de traces de navigation

Expérimentation 1 : trafic réseau et complexité des sites Web

Dans quelle mesure le comportement d'un utilisateur visitant un site Web influence la création d'une empreinte utilisable par la suite pour de la détection d'anomalies ?

- Complexité des sites Web en termes du nombre et de la taille des ressources à charger.
- Sessions de navigation Web avec Firefox, anti-pistage $\in \{\text{strict}, \text{standard}\}$, mode de collecte $\in \{\text{micro}, \text{macro}\}$.
- 27 échantillons de données (3 catégories $\times$ 3 sites $\times$ 3 configurations de collecte de données).



Expérimentation 1 : trafic réseau et complexité des sites Web (résultats)

Dans quelle mesure le comportement d'un utilisateur visitant un site Web influence la création d'une empreinte utilisable par la suite pour de la détection d'anomalies ?

- Complexité des sites Web en termes du nombre et de la taille des ressources à charger.
- Sessions de navigation Web avec Firefox, anti-pistage $\in \{\text{strict}, \text{standard}\}$, mode de collecte $\in \{\text{micro}, \text{macro}\}$.
- 27 échantillons de données (3 catégories $\times$ 3 sites $\times$ 3 configurations de collecte de données).

	Strict		**Standard**		**Std. / Str.**	
	UHC	UIP	UHC	UIP	UHC	UIP
One-Page	61,0	4,0	63,5	7,0	1,04	1,8
Encyclopedia	46,7	6,3	127,7	41,7	**2,73**	**6,6**
Content-Heavy	37,0	6,3	60,7	14,7	1,64	2,3

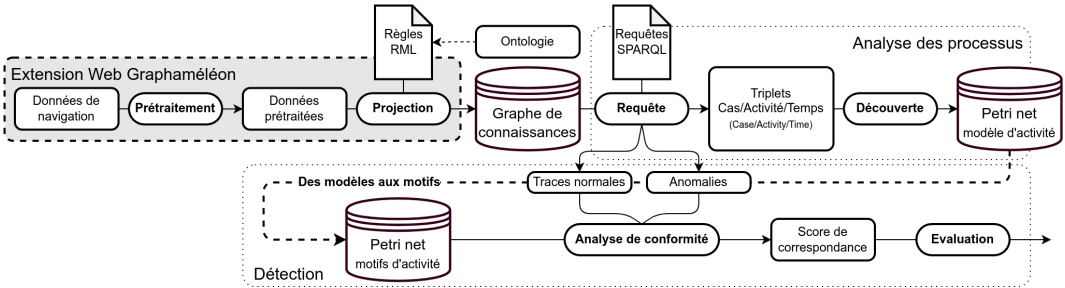
Nombre moyen d'entités capturées en mode `micro`.
Comparaison du nombre moyen d'entités UHC et UIP en fonction du niveau de complexité des sites Web et de la politique d'anti-pistage.
UHC = ucobs:HTTPConnectionFacet, UIP = ucobs:IPAddressFacet.

→ Interactions de l'utilisateur (mode `micro`) : pour une configuration d'anti-pistage donnée, les dénombrements d'entités présentent une variabilité significative au sein de chaque catégorie de complexité.

→ Relaxation de l'anti-pistage : augmentation du nombre moyen de connexions et de serveurs distants consultés, indépendamment du niveau de complexité.

Expérimentation 2 : catégorisation de traces de navigation

Classer les traces de navigation Web par rapport à un modèle procédural ?

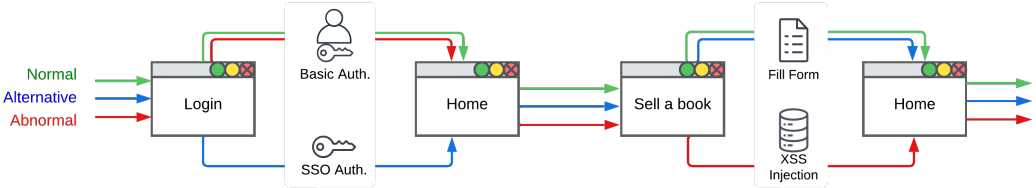


- L'extension Web Graphaméléon capture et annote l'activité de l'utilisateur au niveau du navigateur Web.
- Un composant de "process mining" (découverte de processus) dérive des modèles d'activité à partir du graphe de connaissances RDF résultant.
- Ces modèles peuvent être utilisés pour construire une bibliothèque de modèles d'activité, qui sont ensuite utilisés par un composant de vérification de conformité pour classifier de nouvelles traces d'activité en tant qu'activités normales ou anormales.

Expérimentation 2 : catégorisation de traces de navigation (scénarios)

Classer les traces de navigation Web par rapport à un modèle procédural ?

- Sessions de navigation Web sur un site web de librairie en ligne simulé.
- 3 échantillons de données (scénario de base avec authentification de base, authentification unique, attaque XSS).



De gauche à droite, les étapes clés correspondent à :

1. Authentification de l'utilisateur sur le site Web de la librairie,
2. Navigation vers la page d'accueil,
3. Achat d'un livre,
4. Retour à la page d'accueil.

Les flèches correspondent à la séquence des étapes pour chaque scénario :

- Normal = authentification de base + remplir le formulaire,
- Alternatif = authentification SSO + remplir le formulaire,
- Anormal = authentification de base + injection XSS.

Expérimentation 2 : catégorisation de traces de navigation (résultats)

Classer les traces de navigation Web par rapport à un modèle procédural ?

- Sessions de navigation Web sur un site Web de librairie en ligne simulé.
- 3 échantillons de données (scénario de base avec authentification de base, authentification unique, attaque XSS).
- Adéquation des traces capturées : découverte de processus $\in$ {Inductive, Alpha, LogSkeleton, Heuristic, AlphaPlus} + conformité $\in$ {TokenBasedReplay, Alignment} (bibliothèque PM4Py).

	Base	**Alternatif (SSO)**	**Attaque XSS**
Requêtes	10	**13**	**11**
Interactions	18	14	18
Noeuds	263	283	277
Arcs	404	431	426

Statistiques de collecte de données pour l'expérience de classification des traces de navigation.

→ Signature d'utilisation utilisateur-système au niveau du graphe de connaissances
  — "alternatif" = moins d'interactions mais plus de transactions réseau,
  — "attaque XSS" = nombre d'interactions constant et augmentation de 1 requête du fait de l'étape d'injection SQL.

→ Modèles procéduraux : aident à l'analyse mais manquent de précision pour les micro-changements dans les activités (98% d'adéquation pour l'"attaque XSS" avec le modèle d'activité "de base" comme référence).

Par où commencer ? Exemple avec Firefox

1. Téléchargez une version de Graphaméléon, décompressez-la, puis chargez-la dans votre navigateur Web

```
wget https://github.com/Orange-OpenSource/graphameleon/
    releases/download/v2.1.0/graphameleon-dist-2.1.0-firefox.zip
unzip graphameleon-dist-2.1.0-firefox.zip -d graphameleon-ext
firefox about:debugging#/runtime/this-firefox
```

2. Collectez, observez et jouez avec vos données de trace de navigation dans l'extension Web
   - Ouvrez l'extension Web Graphaméléon
   - Sélectionnez le format de sortie `Semantize` avec les règles de transformation `mapping/micro_v2/rules.ttl`
   - Cliquez sur le bouton `Record` et naviguez sur le Web

3. Exportez vos données de trace de navigation et analysez-les avec des requêtes SPARQL
   - Cliquez sur le bouton `Stop` et exportez les données de trace au format `.ttl` (syntaxe Turtle)
   - Chargez vos données de trace sérialisées dans un terrain de jeu SPARQL
   - Obtenez des statistiques sur votre trace sérialisée

```
SELECT ?class (COUNT(?subject) as ?classCount)
WHERE { ?subject rdf:type ?class .
FILTER (?class != rdfs:Class && ?class != rdf:Property) }
GROUP BY ?class ORDER BY DESC(?classCount)
```

Plus d'informations et d'options sur https://github.com/Orange-OpenSource/graphameleon/

Conclusion et travaux futurs

Sujet Analyser et apprendre les modèles comportementaux des réseaux (structure et dynamique) en relation avec les activités des utilisateurs.

Approche Capture en direct des requêtes réseau et des interactions de l'utilisateur au niveau du navigateur Web + sérialisation vers un graphe de connaissances RDF en utilisant l'ontologie UCO.

Observations Motifs couple utilisateur-système au niveau du graphe de connaissances (contenu Web, politique de suivi, attaque) + raffinements complémentaires nécessaires pour exploiter les modèles procéduraux.

Explorer Contextualiser les actions sur des graphes en flux ? Identifier une granularité généralisante pour les modèles procéduraux ? Base de connaissances décrivant les structures et le comportement type des réseaux ?

Publication – IC @ PFIA 2024

Lionel TAILHARDAT, Benjamin STACH, Yoan CHABOT, et Raphaël TRONCY.

Graphaméléon : apprentissage des relations et détection d'anomalies sur les traces de navigation Web capturées sous forme de graphes de connaissances.

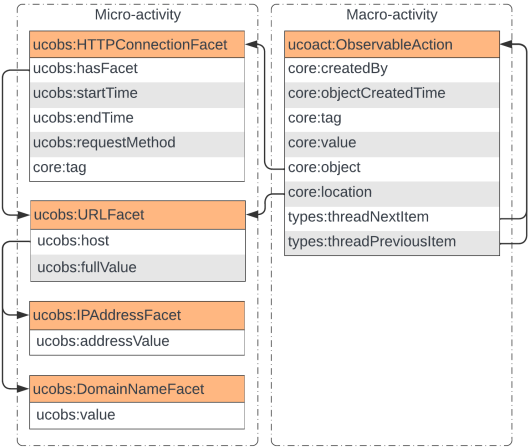
Dépôt de code en source ouverte

Extension Web Graphaméléon
https://github.com/Orange-OpenSource/graphameleon

Jeu de données Graphaméléon
https://github.com/Orange-OpenSource/graphameleon-ds

Annexes

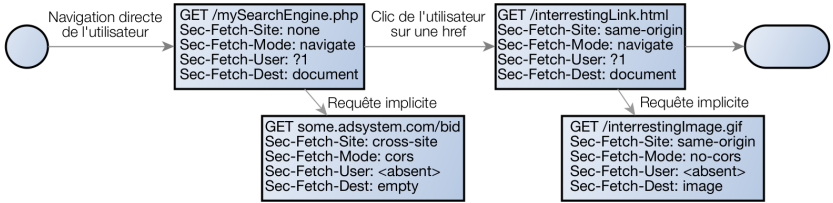
Modèle de données pour décrire les activités des utilisateurs et du réseau



- Le diagramme de classe définit les concepts et les propriétés utilisés pour la représentation sémantique des `micro`-activités (à gauche) et des `macro`-activités (à droite).
- Pour les micro-activités, les classes et les propriétés présentées décrivent avec précision une séquence de requêtes capturées au niveau du navigateur Web. La propriété `core:tag` contient les valeurs des ex. `core:tag "click"; core:value "clc-password-input";`).
- Les noms des concepts et des propriétés utilisés sont définis par le vocabulaire UCO. Espaces de noms :
  - `core` = https://ontology.unifiedcyberontology.org/uco/core#
  - `ucobs` = https://ontology.unifiedcyberontology.org/uco/observable#
  - `types` = https://ontology.unifiedcyberontology.org/uco/types#

Les “fetch metadata” pour déduire l’activité de l’utilisateur/équipement

Ce diagramme de séquence illustre, à travers un exemple fictif d’une session de navigation Web où l’utilisateur se connecte à un site de recherche et suit un hyperlien, le principe pour déduire l’activité de l’utilisateur/équipement à partir des quatre en-têtes HTTP “fetch metadata”.



Sémantique des “fetch metadata” :

- `Sec-Fetch-Site` = relation entre l’origine de l’initiateur de la requête et l’origine de la ressource demandée (p.ex. même site, inter sites)
- `Sec-Fetch-Mode` = mode de la requête (p.ex. l’utilisateur navigue entre les pages HTML vs requêtes secondaires pour charger des images et d’autres ressources)
- `Sec-Fetch-User` = envoyé uniquement pour les requêtes initiées par l’activation de l’utilisateur, et sa valeur sera toujours “`?1`” (p.ex. identifier si une requête de navigation à partir d’un document, d’un iframe, etc., a été initiée par l’utilisateur)
- `Sec-Fetch-Dest` = où et comment les “fetch metadata” seront utilisées pour une meilleure gestion des requêtes côté serveur (p.ex. iframe, composant vidéo). Les sous-documents de chaque page Web (requêtes implicites) sont identifiés en fonction de l’absence de valeur pour l’en-tête `Sec-Fetch-Dest`.

Données collectées par Graphaméléon

Types de données collectées par l'extension Web Graphaméléon en fonction du mode de capture (`micro`-activité vs `macro`-activité), et regroupées selon leur portée (requête vs interactions vs les deux):

Portée	Paramètre ou nom de l'en-tête HTTP	**Micro**	**Macro**
Requête	Method	✓	✓
	URL	✓	✓
	IP	✓	✓
	Domain	✓	✓
	Sec-Fetch-Dest	✓	✓
	Sec-Fetch-Site	✓	✓
	Sec-Fetch-User	✓	✓
	Sec-Fetch-Mode	✓	✓
Interaction	EventType	-	✓
	Element	-	✓
	Base URL	-	✓
Les deux	User-Agent	✓	✓
	Start time	✓	✓
	End time	✓	✓

Expérimentation 2 : correspondance au motif de référence

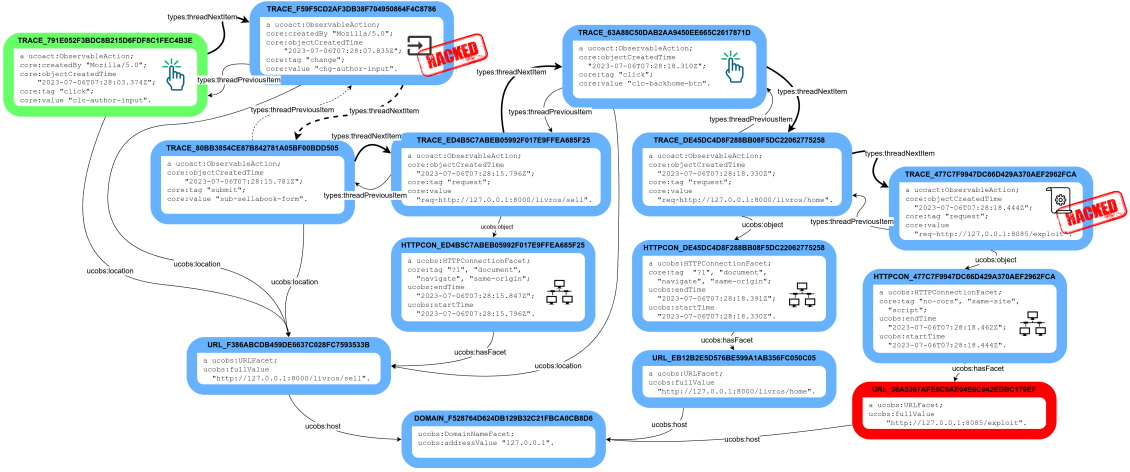
Comparaison des scores de correspondance au motif de référence (modèle d'activité du scénario de base) pour les modèles d'activité des scénarios "alternatif" et "attaque XSS":

		Alternatif	**Attaque XSS**
Token-Based	Alpha	0.886	0.968
	Alpha+	0.890	0.969
	Inductive	0.923	1.000
	Heuristic	0.923	1.000
Alignement	Alpha	-	-
	Alpha+	-	-
	Inductive	0.718	0.976
	Heuristic	0.718	0.976
Log Skeleton		0.684	0.999

Différentes techniques et algorithmes de vérification de conformité sont utilisés pour calculer les scores de correspondance:

- Les techniques "token-based" et "alignement" nécessitent une découverte préalable du modèle d'activité; les algorithmes "Alpha/Alpha+", "Inductive" et "Heuristic Miner" sont utilisés pour cela.
- La technique "Log Skeleton" fournit directement les scores de correspondance en utilisant les traces d'activité.

Expérimentation 2 : extrait du graphe `GPL_attack_scenario.ttl`



Source de données : https://github.com/Orange-OpenSource/graphameleon-ds/blob/main/exp-02/GPL_attack_scenario.ttl