Threat Intelligence : pourquoi le tri « intelligent » de l'information n'est pas une bonne solution

La valeur – et même tout simplement l'utilité – des informations qui seront produites par la Threat Intelligence dépendra en grande partie de la qualité et de l'exactitude des données sources qui seront collectées et stockées pour être analysées. Malheureusement, il est désormais courant de collecter bien plus de données qu'il n'est possible de traiter.

Plusieurs approches sont alors couramment proposées afin de réduire cette « infobésité » :

Les trois stratégies les plus courantes sont les suivantes :

1 – L'échantillonnage : il s'agit de n'analyser qu'un sous-ensemble de données statistiquement significatif
2 – Le filtrage : la mise en place de règles qui permette de supprimer dès la captation les informations jugées sans importance ou répétitives
3 – Le passage à l'échelle : trouver des outils et des technologies qui permettront de traiter toutes les données de manière efficace

Alors que tout le monde aimerait, dans l'absolu, tout capter et tout traiter (option 3), il s'agit rarement de la solution retenue, essentiellement à cause du défi technique à relever pour traiter une volumétrie très importante dans des temps limités.

L'échantillonnage (option 1), pour sa part, peut être un moyen efficace afin de connaître la nature statistique des données. Mais il n'est pas très utile lorsque l'analyste recherche une donnée ou un type de données en particulier.

Le filtrage, enfin (option 2), est une bonne stratégie à condition d'avoir une confiance absolue dans les méthodes de filtrage utilisées. Et notamment d'être certain qu'elles ne modifient pas le profil statistique des données collectées et permettent de conserver toutes les celles vraiment importantes. Mais il y aura toujours le risque d'avoir écarté hier une donnée qui semblait inutile alors qu'elle s'avère vitale aujourd'hui.

Ces trois stratégies peuvent être appliquées à de nombreux domaines et de nombreux types de données. Mais si l'on considère plus particulièrement le domaine de la sécurité, le fait de devoir faire face à un adversaire intelligent et imprévisible change considérablement la donne.
Par exemple, est-ce que l'on manque de données à cause du comportement de l'adversaire ou à cause d'une perte accidentelle ? L'adversaire a-t-il compris la stratégie de filtrage et n'est-il pas en train d'exploiter cette connaissance pour ne pas être détecté ?

Prenons l'exemple du filtrage dans le contexte de la cybersécurité.

Supposons que nous limitions la collecte des données réseau à 100 connexions pour chaque processus. À première vue, cela semble raisonnable. Le nombre moyen de connexions pour

chaque processus sur les postes de travail est beaucoup plus faible.

Cependant, en réalité, les modèles de données dans les environnements informatisés suivent une distribution agressive de la loi de puissance, et non une distribution linéaire ou même naturelle. Autrement dit, il sera difficile de dire si une information intéressante ne se situait pas dans la « longue traîne »… qui n'a pas été capturée !

Par exemple, alors que le traitement de texte n'ouvrira que très peu de connexions, parfois aucune, parfois une ou deux à l'occasion d'une mise à jour, le navigateur ouvre quant à lui des dizaines de connexions pour un seul site web visité… et il en aura facilement des milliers par jour.

Ce problème s'aggrave encore si l'on considère les serveurs – sur ces derniers un faible nombre de processus serveur spécifiques (par exemple la base de données oracle, le serveur de messagerie, etc.) recevront 99,999 9 % de toutes les communications entrantes. Tout filtrage basé sur le plafonnement supprimera alors la grande majorité des données.

En définitive, quelle que soit la stratégie choisie pour limiter la collecte de données, ces limites seront toujours présentes.

Le problème, bien sûr, c'est tout simplement que les données qui ne sont pas collectées ne peuvent fournir de visibilité, ne peuvent pas être utilisées pour de la chasse aux attaquants (le « *hunting* »), et ne contribuent en rien aux détections côté serveur.

Et même si un effort spécifique est réalisé pour ne cibler que les comportements potentiellement malveillants, d'autres problèmes demeurent, notamment à cause :

1 – Du principe d'incertitude : l'efficacité d'un filtre censé identifier les comportements malveillants n'est jamais garantie à 100 %

2 – De la loi de puissance : quel que soit le mécanisme de limitation mis en place, la nature de la loi de puissance conduira toujours à un filtrage arbitraire.

3 – De l'adaptabilité de l'adversaire humain : lorsque celui-ci comprend comment est limitée la collecte des données, il peut s'en servir pour contourner la protection.

4 – De la rupture de la chaîne forensique : à partir du moment où des règles de filtrage sont en place, il y aura toujours des données manquantes. Et en raison de la nature de la loi de puissance, il y aura probablement beaucoup de données manquantes ! Ce pourra empêcher de collecter des données qui s'avéreront par la suite essentielles lors d'une réponse aux incidents.

5 -De la perte de données non vérifiable : lorsque des données sont manquantes, il est impossible de déterminer si c'est par conception, par erreur ou due à un comportement malveillant

C'est pourquoi, alors que certains fournisseurs parlent de « filtrage intelligent », les problèmes ci-dessus montrent qu'en fait, il n'y a rien d'intelligent à filtrer !
Trop souvent, d'ailleurs, le filtre n'est pas conçu pour réduire le bruit, mais simplement pour pallier les limites technologiques des systèmes et économiser sur le coût de la solution.

Cependant, ce choix a un coût important pour l'intégrité des données, la qualité de la détection et

au final sur la valeur de sécurité fournie par le système.

C'est pourquoi seule la stratégie numéro 3 – tout collecter, tout traiter, tout conserver et donner toutes les données aux chasseurs – est viable. Bien entendu, cela demande plus d'effort et constitue un véritable défi technique : il est nécessaire de traiter les données suffisamment rapidement pour qu'elles demeurent pertinentes et de s'assurer que le système est capable de corréler rapidement des sous-ensembles de données différentielles, ce qui nécessite des algorithmes spécifiques.

Mais pour autant, toute autre option est limitante : lorsque l'on applique un filtrage arbitraire/intelligent/statistique, l'on va inévitablement introduire de la cécité dans le système.

Et les pirates l'exploiteront – soit délibérément en analysant les règles de filtrage pour les exploiter, soit par accident, parce qu'il est impossible d'avoir la certitude absolue qu'une information utile ne se dissimulait pas dans les données qui ont été ignorées.

Finalement, donc, l'approche vraiment intelligente pour filtrer les données… c'est de ne pas les filtrer du tout.