Université de Montréal

IDENTIFYING KEY THREAT ACTORS BASED ON THEIR EXPERTISE

Par

RUELLAN Estelle

École de Criminologie

Mémoire présenté en vue de l'obtention du grade de Maitrise

en Criminologie, option Générale

Avril 2024

Université de Montréal

Unité académique : École de Criminologie

Ce mémoire intitulé

Identifying key threat actors based on their expertise.

Présenté par

Estelle RUELLAN

A été évalué(e) par un jury composé des personnes suivantes

Benoit Dupont
Président-rapporteur

Masarah Paquet-Clouston
Directrice de recherche

François Labrèche
Codirecteur

Andréanne Bergeron
Membre du jury

Résumé

L'avènement du Big Data a rendu la collecte et l'analyse des renseignements sur les cybermenaces difficiles en raison de leur volume, conduisant la recherche à se concentrer sur l'identification d'acteurs clés. Cependant, ces études délaissent l'expertise dans l'identification de ces acteurs. L'expertise est pertinente puisqu'elle est étroitement liée au succès criminel. Cette recherche s'appuie sur une évaluation proactive de l'expertise potentielle envers des types d'attaque afin identifier les acteurs clés correspondant dans les forums de cybercrime. En étudiant 4 441 acteurs sur des forums de cybercrime, cette étude utilise l'algorithme de détection de communautés Leiden et partitionnement K-means, en plus d'un cadre criminologique, afin d'identifier les acteurs clés experts selon leur type d'attaque de prédilection. Les analyses révèlent plusieurs résultats pertinents. Premièrement, les types d'attaque agissent comme des catalyseurs de communautés d'intérêt, transcendant les frontières des forums. Deuxièmement, les acteurs clés identifiés dans cette étude représentent moins de 2% de la population et, constituent une minorité prometteuse pour l'allocation des ressources dans l'industrie du renseignement. Troisièmement, en adoptant une opérationnalisation criminologique de l'expertise intégrant l'évaluation objective du niveau de compétence, de l'engagement et du taux d'activité, l'étude introduit un cadre plus holistique pour l'étude des cybercriminels. Enfin, l'intégration des fondements criminologiques à l'approche hybride propre à la littérature en informatique dans l'analyse des communautés de *threat actors* a permis une nouvelle compréhension de cette population et de leur expertise. Ce faisant, cette étude contribue à combler le manque d'intérêt criminologique pour l'identification et l'étude des acteurs clés selon leur expertise.

Mots-clés : hackers clés, acteurs clés, expertise cybercriminelle, renseignement cyber menace (CTI)

Abstract

The advent of Big Data has made the collection and analysis of cyberthreat intelligence challenging due to its volume, leading research to focus on identifying key threat actors; yet these studies have failed to consider the expertise of these actors. Expertise is relevant as it is closely bound to criminal success. Hence, this research relies on a technical expertise in attack patterns to identify key threat actors in hacking forums. Specifically, studying 4,441 actors from cybercrime forums, this study leverages Leiden community detection, K-means and a criminological framework to identify areas of expertise and detect their related expert key actors. The analyses reveal several key contributions. First, attack patterns act as catalysts of cybercrime communities of shared interest, transcending forum borders. Second, key actors identified in this study account for less than 2% of our population and represent a promising scarcity for resources allocation in cyber threat intelligence production. Third, by adopting a criminological operationalization of expertise, integrating an objective assessment of skill level, commitment, and activity rate, the study introduces a more comprehensive framework for understanding cybercriminals. The focus on expertise results in more complete profiles of experts that are actionable for cyberthreat prevention. Combining criminological theoretical foundations with previous literature's hybrid approach in the analysis of threat actor communities, this study contributes to a new comprehension of threat actor populations and their expertise. Consequently, this research contributes to bridging the criminological gap regarding the identification and study of key actors based on their expertise.

Keywords: key hacker identification, cybercrime expertise, cyber threat intelligence

Table of Content

Table of Tables

Table of Figures

Acronym List

CVE: Common Vulnerabilities and Exposures

CAPEC: Common Attack Pattern Enumeration and Classification

CoI: Community of Interest

Remerciements

Tout d'abord, j'aimerais remercier ma famille pour m'avoir permis de réaliser mes études au Canada et m'avoir soutenu pendant ces 5 années. Merci à toi ma petite maman pour avoir relu mes travaux depuis le début de ma scolarité.

J'aimerais remercier Masarah et François pour leur patience et leurs conseils tout au long de la direction de ce mémoire. Merci Masarah pour m'avoir fait confiance et m'avoir guidée dans plein d'autres projets au-delà de ce mémoire.

Merci à toi, Max, pour ton soutien autant technique qu'émotionnel mais aussi pour tes critiques toujours pertinentes dans ce projet et tous les autres. Merci pour toutes les heures que tu as passées à relire ce même travail pendant un an.

Merci aux Zouzous pour toutes les rigolades, les débats et les pratiques de présentations dans le salon qui ont su animer nos soirées Montréal malgré la neige et le manque de soleil. Vous avez rendu cette année magique les gars, j'ai gagné en skills de présentation grâce à vos talents de HEC boys.

Introduction

Cybercrime has become an omnipresent threat in today's digital age, causing significant financial loses and disrupting organizations worldwide (World Economic Forum, 2023). Traditionally, criminal investigations have focused on specific events and adopted a reactive approach (Eck and Rossmo, 2019; Samtani and Chen, 2016). However, the scale of the (potential) damages stressed the change towards a more proactive approach to cyberthreats (Kugler, 2009; Geers, 2010).

With the sheer volume of data on the internet, cyber threat intelligence production has become a real challenge. The amount of data and users on the internet makes the collection and production of cyber threat intelligence extremely time-consuming and resources intensive. As a result, identifying specific and relevant threat actors to focus on for the production of credible intelligence has become a great challenge for the cyber threat intelligence production industry (Marin & al., 2018; Huang & al., 2021). Research has named this problem: the key hacker identification problem (Marin & al., 2018).

The key hacker identification has drawn attention from both computer science and criminology research. This body of research has studied key actors under three perspectives: the social network of key actors (Samtani and Chen, 2016; Samtani & al., 2017), the content of key actors' discussions (Holt and Kilger, 2008; Benjamin and Chen, 2012; Zhang & al., 2015; Fang & al.,2016) and finally, the combination of both social network and discussion content of key actors (Abbasi & al., 2014; Grisham & al., 2017; Marin & al., 2018; Johnsen and Franke, 2020; Huang & al., 2021). But the skill level, and expertise, of key actors appears almost absent from this body of literature.

The internet population comprises varying levels of knowledge (Marin & al., 2018; Huang & al., 2021), yet it's the proficient and respected actors that are of interest. Skill level, and more generally, expertise is closely related to criminal success (Bartol and Bartol, 2014). Threat actors with expertise in their field are more likely to be successful, making them more dangerous. Given their higher success rate in their misdeed, it becomes logical to prioritize the threats posed by

these expert threat actors for the production of intelligence (Motoyama & al., 2011; Bartol and Bartol, 2014; Marin & al., 2018).

Bouchard and Nguyen (2011) propose a contemporary classification of criminals based on a two-facets conceptualization of expertise with skill level and commitment being its foundations. The framework contains four classes: professionals (skilled and committed), average career criminals (committed but unskilled), pro-amateurs (skilled but uncommitted) and amateurs (unskilled and uncommit. It allows to distinguish the nuances of expertise profiles among actors. With the concept of skill level, and thus expertise, being at its foundation, the framework becomes highly relevant for the identification of key actors because it allows distinguishing highly skilled actors within their community.

Taking inspiration from Bouchard and Nguyen's (2011) framework from criminology, this research aims to contribute to the key hacker identification problem by identifying areas of expertise towards attack patterns in cybercrime forums and their related key expert actors.

First, we identify areas of expertise in the form of communities of interest towards attack patterns. To do so, we use a community detection algorithm on a bimodal network linking actors to their attack patterns of choice. Communities of interest towards attack patterns allow to map the potential area of an actor's expertise. An actor's expertise in a certain attack pattern gives them a deeper understanding of that attack pattern making them a threat to their chosen field and thus key to cyber threat intelligence.

Then, we detect key actors based on their expertise in their chosen area. Identifying key threat actors based on their expertise among the mass of internet users could allow for a better allocation of resources while producing more efficient intelligence. Aligning with Bouchard and Nguyen and previous literature, actors' expertise in their chosen field is measured through three facets: skill level, commitment to the chosen field, and activity. Key actors emerge as the ones scoring high on all three variables: the experts in their field.

Drawing from both criminological theoretical foundations and established methods in computer science, this study provides a new perspective on the key hacker identification problem. The following sections will delve deeper into the existing literature on cybercrime prevention and key hacker identification. The methodology used for data collection and analysis will be explained, followed by a presentation and discussion of the research findings. Finally, this research will conclude by summarizing the key contribution to the field.

Literature Review

The aim of this section is to provide the information and scientific background needed to understand the present research. First, cybercrime forums are introduced. Next, hacking forums are introduced along their different features. The body of literature on proactive threat identification on hacking forums then sets the scene for the problem of identifying key actors for cyberthreat intelligence. Finally, existing research attempting to remedy this problem precedes the theoretical framework structuring this project.

The Internet: where cybercrime converges.

Goldsmith and Brewer (2015) have argued that the Internet is a source and facilitator of criminal interactions. Indeed, the Internet acts as a source of ideas and information, offering an individual empowerment environment by facilitating learning and thus enabling individuals to commit crime more autonomously (Holt, 2007). According to the authors, forums and social networks are part of the new means of social encounters that the Internet has enabled to emerge.

Online meeting places, such as forums and chat rooms, represent parameters for the convergence of online offenders that can be used for commercial, social, or even learning purposes (Holt, 2007; Leukfeldt, Kleemans and Stol, 2017). Individuals joining these forums and chat rooms can find accomplices and co-offenders, buy, or sell various crime-related products, seek expertise in cybercrime and even acquire new skills (Holt, 2007; Leukfeldt, Kleemans and Stol, 2017). Forums are thus platforms whose main objective is to enable communication between like-

minded individuals, regardless of their geophysical location, facilitating the emergence of hacker[1] communities (Nunes & al., 2016; Shakarian, Gunn, and Shakarian, 2016).

Hacker communities are free social networks that facilitate the exchange and distribution of information (Holt, 2007). Many of these forums are freely accessible, so you don't need to be a member to access them. In other words, anyone with access to the Internet can consult the information exchanged via these open forums. Other communities gather on the darknet, enabling their members to explicitly protect their identities and thus counter law enforcement surveillance (Macdonald & al, 2015).

In a nutshell, the darknet is intentionally hidden from users, search engines and browsers thanks to "The Onion Router" (Tor) network. Used primarily for underground communications, Tor was originally created by the U.S. Naval Research Laboratory in collaboration with the non-profit organization (NGO) the Free Haven Project (Moore & Rid, 2016). The Tor software is a free open network that provides routing services to its users so they can browse and exchange information online anonymously. By deploying transactions via different servers, known as "relay nodes", Tor's aim is to protect users from online surveillance that threatens personal freedom and privacy (Dingledine, Mathewson & Syverson, 2004). Each node in the network transmits the information it has received from the previous node to the next. So, even though the nodes know neither the origin nor the final destination of the information (each node knows only the information of nodes n-1 and n+1), the Tor network makes tracking almost impossible, while still being functional (Huang and Bashir, 2016). This is why Tor-hosted forums are the most popular environment for illicit communications, and therefore the most commonly used data source for studies of cybercriminals (Abbasi et al, 2014).

Hacking Forums and vulnerability trade

On the darknet and its counterpart, the clearnet, there are cybercrime forums specialized around a particular theme, for example, hacking. Among the best-known and most popular hacking forums

[1] To align with established terminology in the literature, the term «hacker» will be employed throughout this literature review. However, it is important to clarify that within the context of this study, « hacker » specifically refers to active users on hacking forums.

are exploit[.]in, breached[.]co, xss[.]is and nulled, to name but a few. Hacking forums are highly prized by hacker and other curious communities, as they serve as a platform for the exchange of malicious technical knowledge and know-how (Nunes & al., 2016; Shakarian, Gunn, and Shakarian., 2016; Biswas & al., 2022). In addition, some hacking forums even allow the exchange of hacking tools and exploits (Nunes & al., 2016; Biswas & al., 2022).

Some hacking forums features an market/auction space. These market spaces provide a space dedicated to bringing together buyers and sellers of products related to computer vulnerabilities, stolen credentials or hacking services to name but a few (Paquet-Clouston & al., 2018). The commercialization space or auction platform of those hacking forums are referred to as Vulnerability Black Markets (VBM) in literature (Radianti, Rich and Gonzalez, 2009). There are numerous instances of forums with VBM in different parts of the web, and some even exceed 15,000 users (Radianti, Rich and Gonzalez., 2009). Among the best known is exploit[.]in, a russian hacking forum featuring an auction platform. In October 2023, exploit[.]in recorded over 100,000 visits (Similar Web, 2023).

On these hacking forums, so-called “active” vendors can be identified by their explicit posts. They may be individuals who have found vulnerabilities in systems, or individuals who write scripts for exploits and malware (Radianti, Rich and Gonzalez, 2009). “Active” buyers can be identified by their posts or responses to sellers’ posts. Among the “active” buyers identified by Radianti, Rich and Gonzalez (2009) are spammers, hackers for hire, malware writers and exploiters, to name but a few. Finally, there are more discreet buyers who don’t publicly announce their request in a post. Instead, they simply reply to buyer posts with a private communication request to conduct their business out of sight (Radianti, Rich and Gonzalez., 2009).

Allodi (2017) compares cybercrime market forums with legitimate bug-bounty programs. Bug-bounty programs reward people for reporting bugs, especially those associated with vulnerabilities. For cybercrime market forums, the author focuses on a prominent Russian cybercrime market (referred to as RuMarket in the article), whose real name is not disclosed for security and anonymity reasons. Allodi (2017) highlights several interesting facts. Firstly, the author notes that the market is clearly expanding both in terms of members, exploits and also the

number of exploit transactions. Allodi (2017) observes a positive relationship between market activity and the deployment of an exploit in a cyberattack. In other words, an exploit that has received a great deal of attention from the Russian cybercrime market community will be more likely to be used than exploits that have received less attention. Also, vulnerabilities with critical dangerousness scores (score> 9.0) exchanged on cybercrime market forums are more likely to be exploited than those with lower scores.

Allodi also produces an estimate of the price of exploits identified thanks to their vulnerability within the cybercrime market studied. The author then compares the results with those of Finifter, Akhawe and Wagner (2013), who evaluated the rewards offered in Google's bug-bounty program: the Chrome Vulnerability Reward Porgram (VRP). The estimated prices of exploits sold on the Russian market are similar to, or even higher than, the rewards offered in legitimate bug-bounty programs. Comparing the average rewards of Google's (VRP) program with the estimated selling prices on the Russian market forum studied, Allodi reports that the estimated average selling price is twice as high on the Russian cybercrime market forum as on the VRP program (Finifter, Akhawe and Wagner, 2013; Allodi, 2017).

Considering this discrepancy, a financial incentive to direct one's activities towards the underground economy emerges. Indeed, according to Allodi (2017), exploit providers would have every incentive to turn to, and participate in, the underground cybercriminal economy rather than contribute to legitimate programs since underground market forums seem to pay more. Moreover, in the latter, a vulnerability can be sold several times over, by different suppliers to different customers, which is impossible in legitimate programs. Indeed, the publication of a vulnerability within the latter creates an association between the individual who discovered the vulnerability and the discovered vulnerability (Allodi, 2017). In this way, a vulnerability can only be published, and monetized, once within a legitimate bug-bounty program.

The existence of hacking forums featuring a market or exchange space for products related to computer vulnerabilities, or in other words, places of interest for the sale of Common Vulnerabilities and Exposures (CVEs) and exploits, represents a major problem in terms of cybersecurity, but also a source of invaluable information. Indeed, as previously stated, these

places enable the creation of a community of hackers regardless of their geophysical location. These forums serve as exchange platforms where malicious actors share knowledge and know-how, and discuss the latest CVEs discovered. They also take on the role of commercial platforms where exploits, tools and compromised data can be shelved.

Sometimes hackers also mention specific CVEs when discussing, or when offering products derived from their attacks for sale. Since CVE nomenclature is standardized (CVE-XXX-YYYYY), CVEs are recognizable regardless of the language in which they are written. The presence of CVEs on these forums is therefore an interesting way of tracking the evolution of vulnerabilities and attack vectors, as well as the attention paid to them within the hacker community.

As problematic as it is pertinent for cybersecurity, the gathering of cybercriminal activities into a focal point emphasizes its relevance in the realm of cyber intelligence production.

Proactive Identification of Cyberthreats on Forums

Capitalizing on this gathering of cybercriminal activities, research has focused on the proactive identification of cyberthreats to tackle cybercrime activities. The focus of the main papers on CVEs, exploits, and hacking forums has, for the most part, been on developing 1) innovative methods for proactively identifying potentially at-risk systems (Nunes, Shakarian and Simari, 2018), 2) methods for predicting cyberattacks (Marin, Almukaynizi and Shakarian, 2019) and 3) models for predicting CVE exploitation (Almukaynizi & al., 2017, a, b).

In 2019, Marin, Almukaynizi and Shakarian introduce an intelligent tool capable of predicting imminent or near-future cyberattacks. Based on 7,800 posts mentioning at least one CVE, as well as over 230 records of past cyberattacks, the authors built an intelligent model capable of learning correlation rules between present CVE mentions and real-world cyberattacks. Two factors are used in order to weight the activities deemed relevant to the model: the socio-personal indicators of the actors mentioning CVEs (the actor's activity and expertise) and the technical indicators of the CVEs mentioned (existence of a known exploit or patch). The tool then uses these correlation

rules to predict the date (plus or minus 3 days) and method of an attack likely to occur with probability p.

Another prediction model is presented by Almukaynizi and colleagues in 2017 (a). Their model intends to predict the exploitation of a vulnerability using supervised machine learning algorithms. The authors use data mentioning at least one CVE from a variety of sources: darknet/deepnet[2] forums and markets focusing on hacking, white-hat communities, and vulnerability researchers to identify CVEs. They then identify CVEs that have already been exploited in cyber attacks, thanks to the attack signatures made available by Symantec antivirus.

From this data, the following characteristics are extracted for each CVE listed: the description of the CVE from the National Vulnerability Database (NVD), as well as the description of the CVE made by actors on darknet/deepnet forums, its dangerousness score (CVSS or Common Vulnerability Scoring System) , the language of the post(s) mentioning the CVE, the presence of a Proof of Concept exploit (small piece of code that proves that a particular security weakness can be exploited), the presence of mentions on the darknet and finally the presence of mentions in a community of vulnerability researchers (Almukaynizi et al. , 2017, a). The model is then trained on CVEs that have been exploited, and their features, to determine which features are key in predicting exploitation. CVEs mentioned within dark/deepnet forums, white-hat communities as well as within vulnerability researcher communities are more likely to be exploited than those having only been mentioned on NVD. Also, vulnerabilities discussed on Russian forums are more likely to be exploited than those mentioned in Chinese, English or Swedish forums.

Almukaynizi and Colleagues (2017, b) take another perspective on CVE exploitation prediction, that of the social network and its metrics. The methodology used in this research focuses on predicting the likelihood of vulnerabilities being exploited. The researchers combine social network analysis with supervised machine learning techniques to achieve this objective. They use data from darknet/deepnet forums that explicitly mention CVEs. A social network is then constructed from actors' discussions on the forums, and various measures derived from the social network are calculated for each actor and fed into the algorithm. Measures include in and out

[2] part of the internet not indexed by search engines

degrees, and centralities. Following the example of Almukaynizi and colleagues in 2017 (a), the authors use Symantec antivirus data, to identify CVEs that have already been exploited and use them as training and test data.

In addition to social network metrics, other technical characteristics of CVEs and forums are integrated into the algorithm. These include CVSS (dangerousness Score), the source forum language, the CVE description and post content. By combining these features with social network metrics, the algorithm aims to improve prediction of vulnerability exploitation. The results of their study suggest that social network metrics are promising predictors and, combined with other features, provide a viable machine learning model in vulnerability exploitation prediction (Almukaynizi & al., 2017, b).

Thus, some research focuses on mentions of CVEs on the deep and darknet to predict CVE exploitation using intelligent algorithms (Almukaynizi & al., 2017, a, b; Marin, Almukaynizi and Shakarian, 2019). Such approaches provide insight into the use of open sources in cyber threat intelligence production. However, by relying on the presence of CVE mentions, this line of research ignores discussions in which no CVE is mentioned.

Nunes, Shakarian and Simari (2018) therefore set about the task of identifying systems at risk of attack without resorting to CVE mentions. Instead of relying on the use of CVE mentions, the authors capitalize on the content of posts as well as descriptions of items sold on over 300 forums and marketplaces on the deep and darknet. Systems at risk of attack are divided into three levels: platforms (e.g. hardware), vendors (e.g. Google) and products (e.g. Adobe Reader).

The authors present a hybrid system that incorporates hacker discussions to make informed decisions about systems at risk (Nunes, Shakarian and Simari, 2018). The hybrid system combines Defeasible Logic Programming (DeLP), a structured argumentation model, with machine learning. More precisely, DeLP is a reasoning framework combining logic programming (with logical rules and facts) and Defeasible Reasoning; the latter challenges established logical rules in the light of additional information. In this way, DeLP makes it possible to follow the reasoning, since it gives arguments when it overturns a rule.

Making use of the content of hacker discussions, as well as additional information such as forum or marketplace details, information on the authors of discussions and the hierarchy of systems (platforms, suppliers, and products), the system combining DeLP and machine learning builds rules and uses machine learning to identify potentially at-risk systems. The hybrid system presented by the authors demonstrates superior accuracy in identifying systems at-risk compared to an approach based solely on machine learning (Nunes, Shakarian and Simari, 2018).

In sum, the development of methods for predicting exploits and cyberattacks has been the subject of various research. Combining various data sources and leveraging intelligent algorithms, several tools and methods have been developed (Almukaynizi & al., 2017, a, b; Nunes, Shakarian and Simari, 2018; Marin, Almukaynizi and Shakarian, 2019). These findings suggest that the prediction of vulnerability exploitation is closely linked to a broader challenge, *that of identifying the hackers* that are relevant to intelligence: the key hacker identification problem.

The Key Hacker Identification Problem

With the advent of Big Data, the production of cyberthreat intelligence has become a real challenge, both in terms of collection and analysis. The sheer volume of data makes the process extremely time-consuming and energy intensive. In an attempt to remedy this problem, research has focused on the ***key hacker identification problem***, i.e. the identification of malicious actors considered key to the production of intelligence.

Identifying a small number of relevant actors and using them to find credible intelligence on cyber threats would enable efforts and resources allocated to intelligence gathering and production to be targeted more effectively. Hacker communities have members with varying levels of knowledge, but it's the more skilled and reputable members who seem to be of interest. Indeed, as the latter are usually more successful in their misdeeds, it becomes natural to prioritize the threats emanating from these highly skilled actors (Motoyama & al., 2011; Bartol and Bartol, 2014; Marin & al., 2018). As a result, key actors make up only a small proportion of the users who make up hacking platforms; the rest being unskilled or present out of mere curiosity (Marin

& al., 2018, b). In other words, a small number of individuals (the most qualified) would be responsible for the most serious threats within these forums.

It is precisely this distribution of key actors among other users that makes the problem of identifying key actors one of the most complex and important issues for the production of cyber threat intelligence.

One promising study was by Marin and colleagues (2018, a). Using several bi- and unimodal networks, similarity functions as well as community detection algorithms, the authors succeeded in developing a method for identifying communities of sellers of cybercrime-related products. The aim of their project was to explore a new method based on network analysis and machine learning to identify and validate communities of malware and exploit vendors present on darknet markets. The authors employ a seven-step methodology. Starting from a bimodal network (vendor-product) and categorizing products into 34 categories, they create a unimodal network (vendor-seller). Within the latter, suppliers with one or more products in common would be linked to each other. Next, the authors inferred communities, based on their product similarity. In total, Marin and colleagues observed 37 and 48 distinct communities in two sets of darknet markets.

Although highly relevant, the focus of this research is industry oriented. The authors develop an algorithm to identify similar communities within different markets so that intelligence agencies can keep an eye on them. Unfortunately, the details of these communities are not developed by the authors, who seem to concentrate on the identification of those communities to the detriment of their interpretation. Thus, the next section aims to paint a picture of the existing literature on the identification of key actors within hacker communities.

Identifying Key Actors in Hacker Communities

Research into the study of key actors and profiles within hacker communities has become increasingly popular with the rise in cybercrime. This craze has resulted in three research approaches. The first is based on social network analysis and focuses on actors with significant

centrality values (Samtani and Chen, 2016; Samtani & al., 2017). The second relies on content analysis of discussions, establishes metrics of activity and quality of interactions, and identifies actors excelling in these key metrics (Holt and Kilger, 2008; Benjamin and Chen, 2012; Zhang & al., 2015; Fang & al.,2016). Finally, the third combines the previous two approaches to study key actors (Abbasi & al., 2014; Grisham & al., 2017; Marin & al., 2018; Johnsen and Franke, 2020; Huang & al., 2021).

The Social Network Analysis

A small number of researchers have studied the social networks of actors present within hacking forums, i.e. how actors are connected to each other through their social interactions, to identify key actors (Décary-Hétu and Dupont, 2012; Samtani and Chen, 2016; Samtani & al., 2017).

Décary-Hétu and Dupont (2012) use Social Network Analysis (SNA) to assess its efficacy in identifying key actors and excluding potential individuals to optimize resource allocation for researchers. To do so, they focus on more than 4,700 one to one conversations between hackers involved in botnets from a Canadian police force investigation (Décary-Hétu & Dupont, 2012). The authors distinguish between arrested hackers, Persons of Interest (POIs), and others. Then they build a social network between arrested hackers and POIs (who were in contact with two or more arrested hackers) (Décary-Hétu & Dupont, 2012).

First, centrality (in and out degree and flow betweenness centrality) and power social network metrics are used to determine the position of each actor in the network. The power metric measures the indirect influence of a node by considering its direct connections as well as the connections of its connections. Next, the authors employ a disruption algorithm, aimed at identifying key nodes of a network based on the impact of the removal of that very node, to identify the optimal node removal to maximize network activity disruption (Décary-Hétu & Dupont, 2012). They then compared the results of the algorithm with the arrests from the investigation and the position of each removed node in the network. The key nodes identified by the algorithm were those with high degree centrality and flow betweenness centrality and coincide, for the most part, with the arrested hackers from the investigation. However, the arrested hackers scored lower on the power metric (Décary-Hétu & Dupont, 2012).

In their 2016 study, Samtani and Chen used social network analysis of over 120,000 posts from the OpenSC hacking forum to discern the main actors involved in keylogging activities within the forum. The authors constructed a bipartite network linking actors to keylogging threads. Based on the degree as well as the betweenness centrality, the researchers classified actors according to their contributions to threads featuring keylogging. Their analysis revealed that key actors were often the oldest members of the forum and possessed a particularly high degree centrality (Samtani and Chen, 2016).

Following in the footsteps of Samtani and Chen (2016), Samtani and colleagues (2017) focused on identifying key hackers disseminating malicious assets (e.g., source code, attachments) using social network analysis. By analyzing over 400,000 posts from seven different forums, including Exploit[.]in and OpenSC, the authors first used machine learning and topic modeling algorithms (LDA) to identify new malicious assets. Next, they built a bipartite hacker-asset network to study the key hackers responsible for distributing these assets within the forums studied.

The findings of their study are similar to those of Samtani and Chen (2016): key hackers possess high degree centrality and high betweenness centrality within a network with disparate degree distribution. The majority of key hackers are senior members of their community and present a significant number of posts on the forum(s) (Samtani & al., 2017).

Thus, these studies highlight the effectiveness of social network analysis in revealing the main contributors to forums, notably with the use of mono and bipartite networks, as well as centrality measures.

Discussion Content Analysis

Another strand of literature has turned to the behaviors studied through reputation, the preferred topics and content of chats in these communities to identify key actors (Benjamin and Chen, 2012; Fang & al., 2016).

In the underground economy, where traditional trust mechanisms are absent due to the anonymity of these platforms, reputation becomes crucial for business (Lusthaus, 2012). Actors, unable to

rely on legal recourse or established brands, prioritize reputable individual to mitigate risk of being scammed (Lusthaus, 2012; Décary-Hétu and Dupont, 2013). This focus on reputation is particularly pronounced in online criminal marketplaces, where anonymity and transactions happening online make it even harder to assess someone's trustworthiness (Yip, Webber and Shadbolt, 2017).

To deal with this problem, hacker communities have developed reputation systems that function as a scorecard for trustworthiness (Dupont et al., 2016). These systems often tie reputation to a user's nickname, creating a public «record» of past transactions and interactions (Lusthaus, 2012; Yip et al., 2017). A high reputation score shows an actor is a reliable business partner and can potentially lead to more business and profitability (Motoyama et al., 2011; Décary-Hétu and Dupont, 2013; Yip & al., 2017). Furthermore, users' reputation in underground online communities is also closely linked to their knowledge and respective expertise (Zhang, Ackerman, and Adamic, 2007; Benjamin and Chen, 2012). Due to the uneven distribution of reputation (not everyone has a good reputation), it can serve as a valuable metric for identifying important actors within the underground economy (Dupont & al., 2016).In 2012, Benjamin and Chen examined two darknet forums with a reputation system to investigate which characteristics influenced actors' scores using linear regression. The key actors in this study were those with the highest reputation score. The authors used six variables extracted from the content of the posts in the communities under study: average post length, total number of posts, number of threads in which the actor is involved, number of replies, seniority, and attachments in their linear regression model to observe which ones influenced reputation (Benjamin and Chen, 2012).

Their results suggest that while being involved in multiple threads, sharing material (attachments) and total number of posts contribute significantly to reputation, while discussion quality (average post length, number of replies per thread) and an actor's seniority do not (Benjamin and Chen, 2012). In other words, key actors according to reputation, post more, are involved in more threads and share more ancillary material.

Finally, Fang and colleagues (2016) set out to identify key actors within 19 Chinese hacker communities hosted on Baidu Tieba based on their preferred topic. Using topic modeling (LDA),

the authors identified five main topics: trading, fraud identification and prevention, cooperation calls, informal conversation, and monetization. They then identified the ten most active users for each topic, i.e. those with the greatest membership of the topic. It is these users who are designated as key for each topic by Fang and colleagues (2016).

This second approach in identifying key actors focused on content, its quality as well as the activity and even seniority of actors within the forums to identify actors considered key. The next section adopts a hybrid approach, combining both previously presented approaches.

The Hybrid Approach

The final body of literature combines social network analysis with content analysis to study hacker communities and their key actors (Abbasi & al., 2014; Grisham & al., 2017; Marin & al., 2018; Johnsen and Franke, 2020; Huang & al., 2021).

Abbasi and colleagues (2014) identify actors deemed "experts" using a combination of topological analysis, interaction analysis and content analysis to characterize their specialty. The study uses interaction coherence analysis (ICA) to extract topological features and interactions between users. It also integrates content analysis of posts in the identification and analysis of experts. Four features are extracted from post content: topology, cybercriminal assets, specialized lexicons, and forum involvement. Cybercriminal assets include the number of attachments and source code, specialized lexicons relate to hacker-specific terminology, and forum involvement encompasses measures such as number of posts, thread involvement and user seniority.

The authors identify four groups: black market activists, founding members, technical enthusiasts, and average users. Due to their importance in the ICA interaction network, the first three groups were deemed key by Abbasi and colleagues (2014). Indeed, degree centrality was taken as an indicator of the importance of these key actors. The groups are fairly eponymous. Black market activists are focused on finding opportunities to exchange illicit products. Founding members include the forum's longest-serving users, with the most posts and interaction with others. Founding members have been identified as central figures, with many interactions between members and acting as a link between technical enthusiasts and black-market activists.

Technical enthusiasts are highly qualified and often include code in their messages. Average users, on the other hand, are participants who do not actively participate in the forum, so they have not been classified in any other group (Abbasi & al., 2014).

In 2017, Grisham and colleagues set out to identify key actors in darknet forums dealing with mobile malware. To do so, they employed deep learning-based content analysis to identify mobile malware attachments. Next, the authors leveraged social network analysis to identify the key actors spreading those mobile malwares (Grisham & al., 2017). The authors identify the authors of posts containing a mobile malware attachment to create a two-mode social network, linking the actors to the mobile malware-related threads in which they post messages. The two-mode network is then projected onto one mode to calculate social network metrics for each actor. Their results show that key actors are those with the highest degree and eigenvector centralities and are among the longest-serving actors in the forums on which they operate. The authors also reveal that these key actors often hold administrative positions in the forums on which they operate (Grisham & al., 2017).

In 2020, Johnsen and Franke also combined text analysis with latent dirichlet allocation (LDA) and social network analysis with centrality measures. In this article, text analysis is used to eliminate users with low technical skills. Indeed, their LDA approach enabled them to distinguish between content creators assumed to be highly technically proficient and content consumers, assumed to be low-skilled (who only write messages of appreciation). Of the 300,000 or so users of the Nulled forum they studied, 24% qualified as users with high technical skills (Johnsen and Franke, 2020).

A network based solely on the interactions of highly skilled users was constructed. And these users were analyzed according to their in- and out-degree, eigenvector centrality, closeness centrality and betweenness centrality. The authors then examine the preferred topics (obtained through LDA) of the most important actors for each measure. LDA content analysis makes it possible to inspect all the content posted by each actor and deduce their role (which could have been implied in their statements) to determine whether they are key (Johnsen and Franke, 2020). At the end of their analysis, a key actor profile emerged: key actors are those with the longest

tenure, who have been involved in forum life for the longest, and who occupy a high rank or position within the forum (Johnsen and Franke, 2020).

In their 2021 study, Huang and colleagues introduce a new framework, HackerRank, aimed at automating the analysis of key actors in underground forums. This framework uses a combination of content analysis and social network analysis to identify and classify key actors. The authors apply HackerRank to five forums: Nulled, HackThisSite, HiddenAnswers, BreachForum and Raid. First, user characteristics and preferred topics are extracted separately using content analysis. The authors define user content in terms of three aspects: activity (number of messages and replies), content quality (longer messages, elaborate replies, technical jargon) and knowledge dissemination capacity (keywords related to knowledge sharing and acquisition in messages and replies). Next, a social network graph based on user interactions is built. Finally, an improved algorithm for combining the results of the content analysis with the social network analysis is used to obtain a user ranking; the top-ranked users being considered key actors (Huang & al., 2021).

Their findings underline that leading hackers not only demonstrate a high level of influence on their forum network, but also publish very frequently. Their content is also distinguished by its high quality, as well as themes specific to each key actor (Huang & al., 2021). Finally, according to the authors, HackerRank outperforms traditional methods relying solely on content analysis or social network analysis in identifying key actors (Huang & al., 2021).

Finally, Marin and Colleagues 2018 (b) tackled the key hacker identification problem by adopting a global approach using a set of 25 features to predict hacker reputation scores within three darknet hacker forums. The highest-scoring actors are considered key in this study; their model score estimate is therefore compared with the actual reputation system score present in the forums under study (Marin & al., 2018, b).

The authors investigate the effectiveness of three approaches: content analysis, social network analysis and seniority-based analysis, both individually and combined. Features extracted from the content of actors' discussions reflect the expertise, activity within the forum as well as

behavioral tendencies (knowledge acquisition and provision behaviors) of each actor (Marin & al., 2018, b). Next, features derived from social network analysis are measures of centrality to assess the influence and structural position of each actor. And finally, the last features concern seniority and examine the temporal aspect of actor participation (Marin & al., 2018, b).

The authors then develop several algorithms to which the features are fed in order to evaluate the effectiveness of these features (individually and combined). The hybrid approach, which integrates content analysis, social network and seniority, proves to outperform approaches that rely solely on one category of features (Marin & al., 2018, b). Thus, like Huang and colleagues (2021), their model highlights the importance of a hybrid approach that integrates features derived from content, social network, and seniority analyses in the identification of key actors within darknet forums (Marin & al., 2018, b). The profile of actors identified as "key" by the algorithms is regrettably not developed in their study.

This section paints a picture of the existing literature on the identification of key actors within hacker communities. In the social network approach, different centralities are used to identify key actors. This first approach suggests that key actors possess a high degree and betweenness centrality within a network and are among the most senior and active members (Samtani and Chen, 2016; Samtani & al., 2017).

The content analysis approach shows that key hackers are very active members (Benjamin and Chen, 2012; Abbasi & al., 2014; Zhang & al., 2015) and with the most seniority (Abbasi & al., 2014; Zhang & al., 2015). They have specialized their discussions in a particular (Abbasi & al., 2014; Fang & al., 2016) and often sophisticated topic (Abbasi & al., 2014; Zhang & al., 2015). They are the main knowledge-sharing actors within their community and do not hesitate to share cybercriminal assets in their messages (Benjamin and Chen, 2012; Abbasi & al., 2014; Zhang & al., 2015).

Finally, the hybrid approach suggests that key actors are also those with the most seniority and post the most (Abbasi & al., 2014; Grisham & al., 2017; Johnsen and Franke, 2020; Huang & al.,

2021). They occupy a high rank or position within the forum (Grisham & al., 2017; Johnsen and Franke, 2020; Huang & al., 2021) and demonstrate great influence on their forum's network, which is illustrated by high centrality measures (Abbasi & al., 2014; Grisham & al., 2017; Huang & al., 2021). Their content is also distinguished by their high quality as well as themes specific to each key actor (Huang & al., 2021).

However, this body of literature focuses on the production of algorithms for the intelligence industry and neglects the criminological study of the actors identified by their algorithms. The next section aims to paint a picture of the criminological knowledge and profiles of these actors.

Criminological Profiles of Key Hackers

Among the first studies, Holt and Kilger (2008) propose a dichotomous vision: techcraft and makecrafts hacker. The makecraft hacker is a product creator, creating new scripts, tools, and products for malicious, benign, or beneficial purposes, depending on the user. The techcraft, on the other hand, is more of a product and knowledge consumer. The techcraft will apply the information available to the devices the techcraft is accustomed to interacting with, to accomplish various tasks (Holt and Kilger, 2008).

Next, Zhang and colleagues (2015) study the discussions of a hacking forum and deduce four user profiles. Based on the presence of one of two behavioral tendencies: knowledge acquisition and knowledge sharing, as well as other characteristics, users were classified into one of four profiles: guru hackers, casual hackers, learning hackers, and novice hackers. Guru hackers are the most knowledgeable, sharing their knowledge and advice with others. Casual hackers are more passive and observant. They are mainly interested in obtaining information that could be useful. The Learning profile is quite eponymous. These are the people who use the forum for educational purposes. Finally, Novices are the newcomers who generally join the forum on an ephemeral basis (Zhang & al., 2015). Gurus are the ones considered key or "expert" by the authors. Gurus mainly engage in discussions about sophisticated hacking techniques and share hacking software with others. They are also the oldest members of the community and interact with a large number of users (Zhang & al., 2015).

So, while research on key actors within hacker forums has studied them from various angles, none have thoroughly focused on the skills of these key actors. And yet, as previously stated, hacker communities have members with varying levels of skills. And it is the most skilled members who are of interest, since they are usually more successful in their misdeeds (Motoyama & al., 2011; Bartol and Bartol, 2014; Marin & al., 2018; Marin & al., 2018, b). The inclusion of skills -and subsequently expertise- in the study of key actors therefore becomes imperative to distinguish them within the broad community of users among which they evolve. This is where the classification developed by Bouchard and Nguyen in 2011 becomes relevant, as presented below.

Identifying key actors considering their skill level through expertise

Indeed, considering the skill level of actors when identifying key actors is imperative to distinguish them within the broad community. However, the skill level appears almost absent from the studies mentioned in the previous section.

Indeed, Zhang and colleagues (2015), Benjamin and Chen (2012), Abbasi and colleagues (2014), Marin and colleagues (2018, b) and Huang and colleagues (2021) use a concept akin to skill level: behavior towards knowledge. This concept is transcribed through two behavioral tendencies (knowledge acquisition and knowledge sharing) coded from the content of the actors' posts under study.

Although we understand that for sharing to take place, different levels of knowledge are required, the level of knowledge is not really central to their analyses, but much more instrumental. Indeed, the authors quantify behavioral trends in the presence of keywords related to knowledge sharing and acquisition in messages and replies (questions or answers in the post content, presence of tutorials or doubts about information, or presence of tips or queries). Admittedly, all the elements cited are implicit behaviors reflecting the request or sharing of information. However, none of these elements can be used to evaluate, quantify, or gauge a user's skill level, but simply to

identify a behavioral tendency towards knowledge. Thus, these studies are based on behavior towards knowledge, not the level of knowledge itself.

Other studies use metrics and measures that might be similar or close to the skill level, such as the specialization of discussions towards often sophisticated themes (Abbasi & al., 2014; Zhang & al., 2015; Fang & al., 2016; Huang & al., 2021). The use of technical lexicon and jargon is also recurrent in the key actor literature to reflect the quality of actor content (Abbasi & al., 2014; Marin & al., 2018, b; Huang & al., 2021), implying their skill level. Johnsen and Franke (2020), too, used post content to select presumed technically qualified actors by eliminating actors posting only acknowledgements and not actively contributing to discussions. Actors posting no acknowledgements were presumed to be technically qualified. This binary approach is interesting but falls short of an objective assessment of skill levels. Indeed, not all actors who don't offer thanks are in fact necessarily qualified; they may simply never thank their interlocutors for the information they provide, for example.

Bouchard and Nguyen's (2011) criminological framework differs from key actor profiles outlined above. Bouchard and Nguyen's (2011) perspective holds its advantage in the characteristics on which their profiles are based. The framework allows to differentiate individuals based on their skill level and commitment level in and towards their illegal activities. As stated before, the most successful, and thus dangerous, actors are the most skilled members of their community (Motoyama & al., 2011; Bartol and Bartol, 2014; Marin & al., 2018; Marin & al., 2018, b). With skill level being one of, but not the sole foundation of the framework, Bouchard and Nguyen's (2011) framework becomes highly relevant for the identification of key actors because it allows distinguishing highly skilled and committed actors within their community. In this framework, skill level is further conceptualized as one of two facets of expertise, as explained below.

An expertise-based Framework: Bouchard and Nguyen (2011)

Specifically, Bouchard and Nguyen (2011) propose a contemporary classification of criminals, deeming the criteria used in older classifications, such as Sutherland's The Professional Thief

(1974), as no longer relevant. Their classification is based on an analysis of the life and criminal trajectories of 13 individuals who grow or have grown cannabis, and draws on research into professional crime (Sutherland, 1937; Hobbs, 1997) and criminal success (Bouchard and Nguyen, 2010). The authors' classification is based on two key characteristics - skill level and commitment level - and contains four classes: professionals (skilled and committed), average career criminals (committed but unskilled), pro-amateurs (skilled but uncommitted) and amateurs (unskilled and uncommitted). It is presented in Table 1.

Professionals are individuals who are highly skilled and committed to their field of crime. Conversely, amateurs are unskilled and uncommitted. The authors give the example of an amateur: an individual who has just taken up cannabis cultivation without success. The other two classes are found in the middle of each criterion. Average career criminals are highly committed to their criminal activities despite low qualifications. Finally, pro-amateurs are those who are unable to make a living from crime despite being qualified. Although they have not committed themselves full-time to their criminal activity, pro-amateurs follow and impose a professional standard on themselves. Bouchard and Nguyen explain that pro-amateurs are neither amateurs nor professionals, but rather connoisseurs or aficionados; they are passionate about their activity, but in no way see it as a means of earning a living.

Table 1.

Bouchard and Nguyen (2011, p. 111) framework

	High Commitment	**Low Commitment**
High Skill Level	Professional	Pro-Amateur
Low Skill Level	Average Career Criminal	Amateur

This classification not only considers the notions of "professional" and "amateur" criminals, but also provides meaningful categories of offenders for all those who fall between these two extremes. Finally, the authors qualify their remarks by explaining that the classes are ideals and subject to variation.

Bouchard and Nguyen's (2011) skill level refers directly to an individual's expertise in their criminal field. To illustrate their point, the two authors mention professional athletes who have skills and know-how specific to their sport (Bouchard and Nguyen, 2011).

In the studies presented above, only Marin and colleagues (2018, b) use a concept of expertise to identify key actors. The authors divide expertise into several metrics: quality of engagement, cybercriminal assets and specialized lexicons. Quality of engagement measures the length of threads and replies, as well as knowledge acquisition and sharing behavior. Cybercriminal assets are measured by sharing attachments, and specialized lexicons measure the presence of keywords associated with darknet or technical jargon. Thus, the study considers the expertise of these actors, but unfortunately does not elaborate on the role of expertise in their algorithm, nor on the level and profile of expertise of the key actors identified.

Bouchard and Nguyen's (2011) framework offer a classification based on a two-facets conceptualization of expertise with skill level and commitment being its foundations. Once again, the criminological framework holds its advantage in the resulting four categories of the framework as they allow to distinguish the nuances of expertise profiles among actors. As the concept of expertise is key in Bouchard and Nguyen's (2011) classification, the aim of the next section is to review the literature on the concept of expertise so that the reader can become familiar with it. The section then turns to criminal expertise and the body of literature on the subject.

Expertise

In the *Cambridge handbook of expertise and expert performance*, Ericsson and colleagues (2006) define expertise as "the characteristics, skills and knowledge that distinguish experts from beginners and the less experienced" (p.3-4). Thus, the contemporary conceptualization of expertise refers to a wide range of cognitive knowledge and/or specialized skills in a particular domain (Ericsson, 2006a; van Gog, 2012; Bartol and Bartol, 2014; Nee, 2015).

Another approach to expertise research, the relative approach, uses the comparison between experts and novices to define it. This approach views expertise as a level of skill/knowledge that even a novice can achieve. Expertise is thus seen as a continuum of skill level from novice to expert (Hoffman & al., 1995; Hoffman, 1998; Chi, 2006; Nee and Ward, 2015). Reaching and maintaining the extreme end of the expertise continuum requires ongoing practice to keep one's knowledge and skills in one's chosen field up to date (Ericsson, 2006b). By defining experts in comparison with novices on a continuum, the relative approach is more permissive and requires less precision in defining expertise.

For decades, expertise has been studied from multiple angles. For example, within the legal professions, significant research has focused on chess players, pilots, and doctors, to name a few (Simon and Chase, 1988; Schmidt & al., 1990; Vicente and Wang, 1998). Some have turned their attention to the question of criminal expertise. The first and most important studies of criminal expertise focused on non-violent offences such as burglary (Wright and Logie, 1988; Wright & al., 1995). Since then, other researchers have joined the movement. Some have continued to focus on the expertise of burglars (Nee and Taylor, 2000; Nee and Meenaghan, 2006; Nee, 2015). While others moved beyond burglary to focus on other offenses such as violent crime (Topalli, 2005), arson (Butler and Gannon, 2015), carjacking (Topalli & al., 2015) and identity theft (Vieraitis & al., 2015).

Research by Nee (2015), Nee and Meenaghan (2006), Nee and Taylor (2000), and Wright and Logie (1988) focus on the cognitive processes and decision-making mechanisms used by burglars during target selection. These studies highlight the expertise of burglars in the recognition of visual stimuli and subsequent decision-making. Variables such as position, occupancy, access, cover, security measures or lack thereof, and apparent wealth collectively influence the target selection process (Wright and Logie, 1988; Nee and Taylor, 2000; Nee and Meenaghan, 2006; Nee, 2015). Environmental factors as well as the previously stated variables interact and influence each other in target selection. The bulk of the findings suggest that burglars use their past experiences to build up expertise (Nee and Taylor, 2000; Nee and Meenaghan, 2006). Burglars demonstrate expertise in assessing the criminogenic environment, as well as increased recognition of indicators that make a property a potential target (Nee and Meenaghan, 2006; Nee,

2015). Burglars select their targets with an ease and celerity characteristic of experts (Nee and Taylor 2000) and some even change their modus operandi once inside in order to mislead the police (Nee and Meenaghan, 2006).

In the same fashion as burglary, in carjacking, the expertise lies in the selection of the target, but also in the execution of the act. As Topalli and colleagues explained in 2015, carjackers operate under the constraint of rapid decision-making, relying on their perception to recognize opportunities that present themselves and disappear just as quickly. The expertise demonstrated by carjackers encompasses the rapid assessment of potential targets based on factors such as the value of a vehicle and the perceived resistance of the driver (Topalli & al., 2015). Then, during the actual act, carjackers demonstrate procedural expertise: the application of sophisticated scripts, developed with experience, enabling a rapid attack adapted to the situation (Topalli & al., 2015).

In 2015, Butler and Gannon suggested the existence of expertise among arsonists. Their expertise would manifest itself in two parts: knowledge of fire and avoiding detection. Knowledge of fire refers to the various techniques for starting and spreading fire, acquired and perfected through experience (Butler and Gannon, 2015). Discretion in starting a fire, selecting an out-of-sight location, or using an accomplice to obtain the necessary materials without being seen are all part of the expertise suggested by the authors to prevent the perpetrator from being apprehended by the police (Butler and Gannon, 2015).

Finally, Vieraitis and colleagues (2015) discuss the expertise present in identity thieves and its origin. Unlike street crime and delinquents, identity thieves acquire the skills specific to their crime through their experiences in the legal economy (positions as bankers, or insurance advisors...), as well as the illegal experiences (Vieraitis & al., 2015). Indeed, during their (il)legal experiences, they hone the practical, social, and cognitive skills characteristic of their expertise (Vieraitis & al., 2015). The expertise acquired may include an understanding of financial systems as well as the fabrication of false documents (practical skills), situational awareness and the ability to convince of one’s honesty and react in case of suspicion (social skills) or knowing

where to find the necessary information as well as recognizing problematic situations before they occur (cognitive skills) (Vieraitis & al., 2015).

Generally speaking, expertise is linked to, and sometimes even necessary for criminal success (Bartol and Bartol, 2014). Criminal individuals with expertise in their field carry out their misdeeds in a more sophisticated manner and are more likely to be successful. Thus, criminal expertise makes individuals more effective and more dangerous (Bartol and Bartol, 2014).

The criminal success of an individual can only be measured reactively, after the individual has executed or attempted to execute their activity. Assessing success requires analyzing past events with known outcomes to determine if they were successful. Consequently, this approach is inherently incompatible with proactive intelligence, which aims to prevent those events from happening. In the absence of criminal success indicators, expertise, which can be assessed proactively, becomes essential for identifying relevant actors in cyber threat intelligence.

So, like the expertise a burglar, arsonist or identity thief might possess, we hypothesize that some threat actors possess technical expertise related to one or more types of vulnerability or attack. In the context of this study, expertise would not refer to manual skills such as the botany required to grow cannabis, but rather to the knowledge required to understand and exploit a particular type of vulnerability.

Research Problem: Identifying Key Actors Based on Their Expertise

Previous literature has produced methods of identifying key actors in underground communities to produce cyber threat intelligence (Benjamin and Chen, 2012; Abbasi & al., 2014; Zhang & al., 2015; Fang & al., 2016; Samtani and Chen, 2016; Grisham & al., 2017; Samtani & al., 2017; Johnsen and Franke, 2020; Huang & al., 2021). While these methods prove useful to the intelligence industry, they neglect the criminological study of the key actors identified by their

algorithms. Other research has studied underground communities from various angles and established profiles of actors on these forums (Holt and Kilger, 2008; Zhang & al., 2015). However, none of these studies or those previously stated have thoroughly focused on the technical skill level, and subsequently, the technical expertise of these key actors (Holt and Kilger, 2008; Abbasi & al., 2014; Zhang & al., 2015; Fang & al., 2016; Marin & al., 2018, b; Huang & al., 2021).

According to Bartol and Bartol (2014), expertise and criminal success are closely bound. Criminal expertise render individuals more efficient, sophisticated, and successful in their misdeeds, making them more dangerous (Bartol and Bartol, 2014). As measuring the criminal success of threat actors can only be done reactively (post-attack), we rely on expertise to identify key actors. In the absence criminal success indicators, identifiying criminal individuals with expertise in their field becomes crucial for the production of proactive cyber threat intelligence as these individuals are more likely to succeed in their misdeed. By identifying individuals with technical expertise in cybercrime, we can produce proactive intelligence and thus be more effective in the prevention of cyber threats.

In order to contribute to the identification of key actors and address a gap in criminological interest in the study of these key actors within hacking forums, the present study draws inspiration from the profiles created by Bouchard and Nguyen (Bouchard & Nguyen, 2011). Bouchard and Nguyen's (2011) framework offer a classification based on a two-facets conceptualization of expertise with skill level and commitment being its foundations. The two authors' classification allows us to structure the analyses of this study according to the concept of expertise: a concept almost absent from previous literature, but central to the identification of key actors.

Bouchard and Nguyen's classification echoes the relative approach of expertise research. Firstly, the conceptualization of expertise as a continuum of skill levels is entirely consistent with the skill level criterion used in Bouchard and Nguyen's (2011) classification. Secondly, the process of maintaining expert status refers to the second evaluation criterion used by Bouchard and Nguyen (2011): commitment level. Continuous practice requires a certain level of commitment to

the area of expertise in question. Threat actors' discussions and extensive, assiduous activity around a single type of attack could thus be considered as continuous practice, enabling them to keep their knowledge and technical skills up to date in their preferred domain: a certain type of attack. On the contrary, threat actors spreading themselves too thinly over the types of attack they discuss could be seen as not being committed to a particular type of attack.

Drawing on Bouchard and Nguyen (2011) as well as previous literature, the present study identifies areas of technical expertise in attack patterns in cybercrime forums and their related key expert actors. Specifically, the first research objective is to:

> **(Obj.1) Identify areas of expertise in the form of communities of interest towards attack patterns**.

In this study, attack patterns describe techniques, tactics and methods used by malicious actors to exploit weaknesses in systems. Communities of interest towards attack patterns allow to map the potential areas of expertise and play an important role in the study of threat actors' technical expertise. An actor's technical expertise in a certain attack pattern gives him an in-depth understanding of that type of attack. In this way, an expert actor may be able to carry out sophisticated and damaging attacks that are difficult to detect and prevent. An actor's technical expertise would therefore make him a threat to his chosen field and industry, making him key for intelligence production.

The second objective is to:

> **(Obj. 2) Detect key actors based on their technical expertise level.**

According to Bouchard and Nguyen (2011), expertise is conceptualized through skill level and commitment. In this study, the four expertise profiles of Bouchard and Nguyen (2011) are used as expertise levels with 'Professionals' representing the highest level of expertise and 'Amateurs' being at the bottom of this scale.

Actors of hacking communities tend to not be active for a long time on forums (Hughes & Hutchings, 2023). Considering the ephemeral aspect of forums' population and previous literature suggesting that key actors are among the most active and senior members of their community (Benjamin and Chen, 2012; Abbasi and al., 2014; Zhang & al., 2015; Samtani and Chen, 2016; Samtani & al., 2017; Grisham and al., 2017; Johnsen and Franke, 2020; Huang and al., 2021), actors' activity should be considered in the identification of key actors. Following the literature review and echoing the relative approach on expertise, we consider a third variable to nuance Bouchard and Nguyen's (2011) expertise profiles: activity rate. Activity rate echoes the process of maintaining expertise from the relative approach on expertise. An actor's activity on forums over time can become an indicator of the time consumed for the practice of one's technical expertise. Activity rate then extends Bouchard and Nguyen's framework to nuance the expertise profiles based on seniority and diligence of posting activity. In this study, seniority refers to actors' experience on forums, the amount of time they have been active, whereas diligence refers to the consistent effort in contributing to their forum with specialized content. Those fitting in the 'Professional' class and striking the right balance between seniority and diligence are thus considered key actors in our framework.

By identifying actors with a particular technical expertise in a type of attack patterns, this study contributes to the understanding of cyberthreats by enabling the identification of key actors through their technical expertise. By adopting a two facets conceptualization of expertise level and adding activity rate to nuance the expertise profiles based on seniority and posting activity to identify key threat actors, the concept of expertise distinguishes this study from those previously presented. In so doing, it proposes a method for identifying key actors that would be complementary to those proposed in the existing literature, enabling a more holistic approach to **the key hacker identification problem.**

Contribution to the theoretical framework

Following the relative approach on expertise, the main contribution of this study to the criminological framework is the addition of a third dimension to our expertise framework : the actor's activity rate. The dimension extends Bouchard and Nguyen (2011) to nuance resulting

expertise profiles. Moreover, it allows to differentiate between expert based on their seniority and posting behavior, aligning with previous literature on the key hacker identification.

This study also provides an adaptation of the framework to cybercrime by leveraging CVE and CAPECS to measure the core concepts of Bouchard and Nguyen's (2011) framework: skill level and commitment. Since we can't interview each actor to measure their expertise directly, our adaptation measures the key concepts of expertise using proxy variables (CVE and CAPEC). As a result, the measured technical expertise remains theoretical, as it cannot be verified with absolute certainty.

Navigating Objectives: leveraging CVE and CAPECs for operationalization

To identify actors based on their level of technical expertise, we first need to identify their area of technical expertise, i.e. a specific type of vulnerabilities or attack patterns they show interest in. Taking inspiration from research on vulnerabilities and their exploitation (Almukaynizi & al., 2017, a, b; Marin, Almukaynizi and Shakarian, 2019), the present study leverages mention of vulnerabilities (CVEs). CVEs are identified by a unique formatted identifier, enabling them to be identified automatically in any language. This identifier makes it possible to process and analyze actors regardless of the language they use, without the need for translation and in-depth content review. The choice of CVEs fitted perfectly with the quantitative methodology of this study, as well as its aim: to identify key actors based on their level of technical expertise.

More specifically, the present study uses the attack pattern (CAPEC) corresponding to mentioned vulnerability (CVE) to obtain communities of interest towards attack patterns, mapping the different areas of technical expertise.
To better understand the distinction between CVEs and CAPECs, a technical aside is in order. CVEs, or Common Vulnerability Exposures, are computer security vulnerabilities which, once exploited, can allow entry into a system and subsequent access to private data or disruption of a

system. Several organizations such as MITRE[3] or NIST[4] keep a register of all CVEs known to date and update it daily with newly discovered vulnerabilities. In these dictionaries, information such as the description of the vulnerability, the existence of an exploit or the vulnerability's danger score is available. Each CVE has a unique standardized identifier (e.g. CVE-2019-12255).

CAPEC, or Common Attack Pattern Enumeration and Classification, is a publicly available and community-driven catalog of known attack patterns. These attack patterns describe techniques, tactics and methods used by malicious actors to exploit weaknesses in systems. As of June 2024, the CAPAC catalog counts 559 different CAPECS. Each CAPEC represents a distinct method that attackers might employ to compromise security. CAPEC is closely related to the Common Weakness Enumeration (CWE) framework. CWE identifies weaknesses in software, and CAPEC associates these weaknesses with specific attack patterns that exploit them (CAPEC, 2024).

CAPEC operates within a hierarchical structure where specific CAPEC instances are categorized as children under parent CAPECs (CAPEC, 2024). According to MITRE, the idea behind the hierarchy is that a child CAPEC is a refinement of the parent CAPEC's attack pattern. Much like a family, a parent CAPEC can have multiple children.

The choice between using CAPECs (Common Attack Pattern Enumeration and Classification) and CVEs (Common Vulnerabilities and Exposures) holds significant implications for understanding threat actors based on their interest and technical expertise. It is like deciding which map to use to understand how threat actors work. Opting for CVEs is like using a topographic map, you gain a broad view of the cyber threat landscape, observing vulnerabilities as features without explicit details on the connecting tactics. On the other hand, selecting CAPECs is akin to navigating with a detailed city map, where streets and roads represent the specific attack patterns, providing a clear view of how threat actors traverse the cyber terrain and connect various elements.

[3] https://attack.mitre.org/
[4] https://www.nist.gov/

A more criminology-oriented analogy would be that CAPECs are methods to break into houses or rob banks and CVEs are the specific points or vulnerabilities within a house or bank's security. Imagine CAPEC as a guidebook that explains a general method for breaking into houses, outlining the steps and methods involved in this type of crime. Now, think of CVEs as markings on the house's blueprint that indicate specific weaknesses within a house's security system. Each CVE represents a potential entry point or vulnerability, much like a loose doorknob, broken windows, the orientation of security camera or other areas that might be exploited during a break-in. So, CAPEC serves as the overall strategy for house intrusion, and CVEs highlight particular weak spots within a house that align with this strategy.

Just as a criminal might have a preferred method for breaking into houses, such as using a crowbar to force open a window or tricking the homeowner into opening the door, threat actors might also have their own preferred attack patterns or methodologies. Our study aims to understand threat actors based on the strategy they specialize in instead of the particular weak spot they exploit: a scammer is known for their modus operandi such as creating elaborate social engineering schemes rather than the weak spot they always use, the human's gullibility and desire to help.

While CVEs are essential for tracking and identifying specific vulnerabilities, CAPECs offer a more attacker-centric perspective, making them more suited for studying actors based on their potential technical expertise in a type of attack. This approach enables the analysis of a threat actor's attack patterns, thus providing insights into their modus operandi. The use of CAPECs then facilitates the identification of signature modus operandi associated with specific threat actors' attack profile, making it a more suited approach to the goal of this study.

On top of being attack-centric focused, several CVE can be linked to a single CAPEC, allowing us to reduce the density of the information while opting for an approach more suited to the goal of this study. Notably, CVEs, which represent specific vulnerabilities, are grouped within CAPECs, fostering a more readable and less dense representation of the threat landscape.

To identify areas of technical expertise in the form of communities of interest towards attack patterns **(Obj 1.)**, we focus on cybercrime forums posts and more specifically posts where actors mention a CVE. We link actors with the CAPECs corresponding to the CVEs they mentioned and build a bimodal network connecting CAPECs and actors. Leveraging the Leiden community detection algorithm, we plan on unveiling distinct communities underlying the bimodal network. Next, content analysis is applied to make sense of the identified communities of interest. Thus, this study differs from previous research in its use of CAPECs to study communities of interest and threat actor technical expertise. Mention of CVEs, and their corresponding CAPECs, in this study highlights communities of interest in certain attack patterns and allow for the mapping of areas of technical expertise.

Aligning with Bouchard and Nguyen actors' technical expertise level towards their attack patterns of choice is measured through two facets: skill level and commitment. These facets are conceptualized based on actors' area of technical expertise identified in the first objective, as explained in the methodology below. Then, taking inspiration from previous literature stating key actors are among the most senior and active actors (Benjamin and Chen, 2012; Abbasi and al., 2014; Zhang & al., 2015; Samtani and Chen, 2016; Samtani & al., 2017; Grisham and al., 2017; Johnsen and Franke, 2020; Huang and al., 2021), activity nuances the expertise profiles based on seniority and diligence of posting activity. Using a K-Mean clustering algorithm, those exhibiting the highest level of technical expertise ('Professionals' in Bouchard and Nguyen's (2011) framework) and striking the right balance in activity rate between seniority and posting activity are identified and considered key expert actors **(Obj. 2).**

In short, this methodology integrates network analysis, community detection, content analysis, seniority analysis along with the Bouchard and Nguyen criminological lens providing a robust framework to achieve our research subobjectives. In doing so, this study intends to offer valuable insights into the landscape of hacking communities and their key expert actors. The next section aims to detail the methodology used in this study.

Methodology

This section aims to present the methodology used in this study. First, the data collection process is explained, then the dataset is presented. Social and bimodal networks precede the final dataset. Then, the analysis employed to achieve the first and second objectives are detailed.

Data Collection

The data was collected using the Flare Systems search engine API. Flare systems[5] is an information technology (IT) security company that maintains a cyber threat intelligence platform by monitoring various online spaces. The Flare Systems search engine has been used in previous literature (Paquet-Clouston, 2021; Paquet-Clouston and al., 2022).

To build a network of actors and their CAPECs (attack patterns) of interest, we first needed to collect the posts of actors mentioning CVEs on darknet hacking forums. To do so, we queried the Flare Systems search engine application programming interface (API) for all posts coming from a darknet forum (i.e. exploit_in, xss_is, breached...) that mentioned “CVE-2023”, “CVE-2022”, “CVE-2021”, “CVE-2020” and “CVE-2019”. As actors of hacking communities tend to not be active for a long time on forums (Hughes & Hutchings, 2023), we decided to limit our search to CVEs published within the past five years to have a representative and substantive dataset while avoiding erroneous data by going further back.

The specific query selected all posts coming from a darknet hacking forum containing one of the CVE identifiers: “CVE-2023”, “CVE-2022”, “CVE-2021”, “CVE-2020” and “CVE-2019”. Some actors mentioned multiple CVE identifiers in their post, for example CVE-2022-22965 and CVE-2017-9798. Thus, while our search query only included CVEs from the past five years, the collected data also included older CVEs.

All post not in English were translated by the Flare Software. Once the translation for non-English post was done, we extracted the relevant information for each post, including the post ID,

[5] https://flare.io/

actor, source (forum), title, timestamp, and content. Then, mentions of CVEs were extracted from the content using regular expressions (*regex*).

The Dataset

In total, the data collected included 11,558 posts made by 4,441 different actors on 124 forums. The top 10 most popular forums in our dataset are presented in Figure 1.a) and 1.b). The posts date from January 8th, 2015, to July 31st, 2023, and mention 6,232 different CVEs.

Figure 1.a).

Distribution of Posts per Year

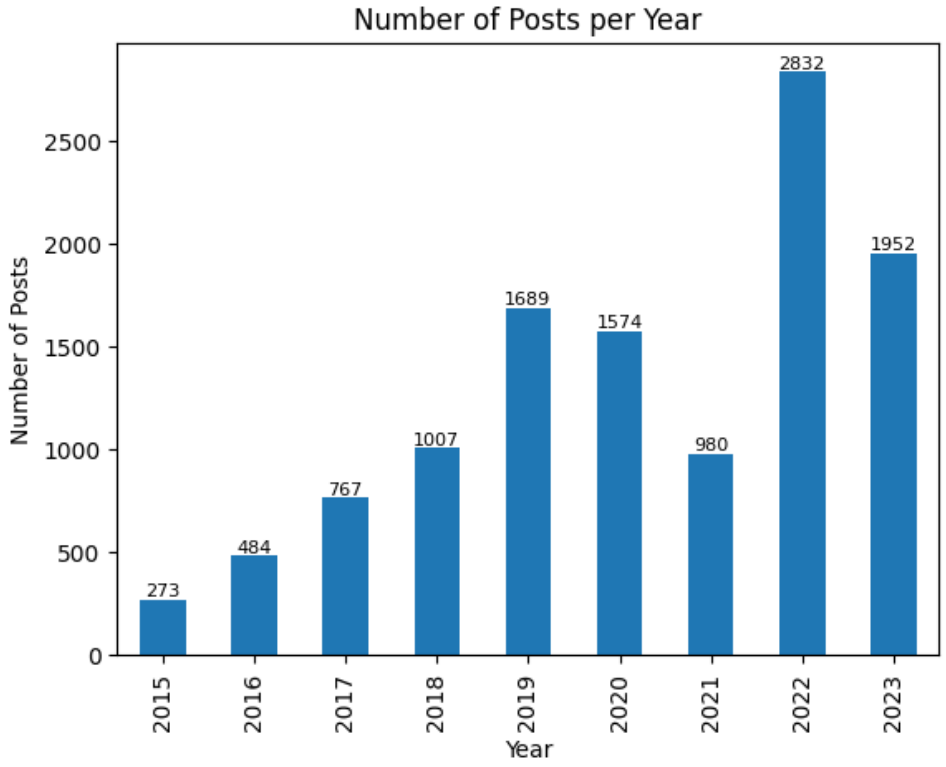
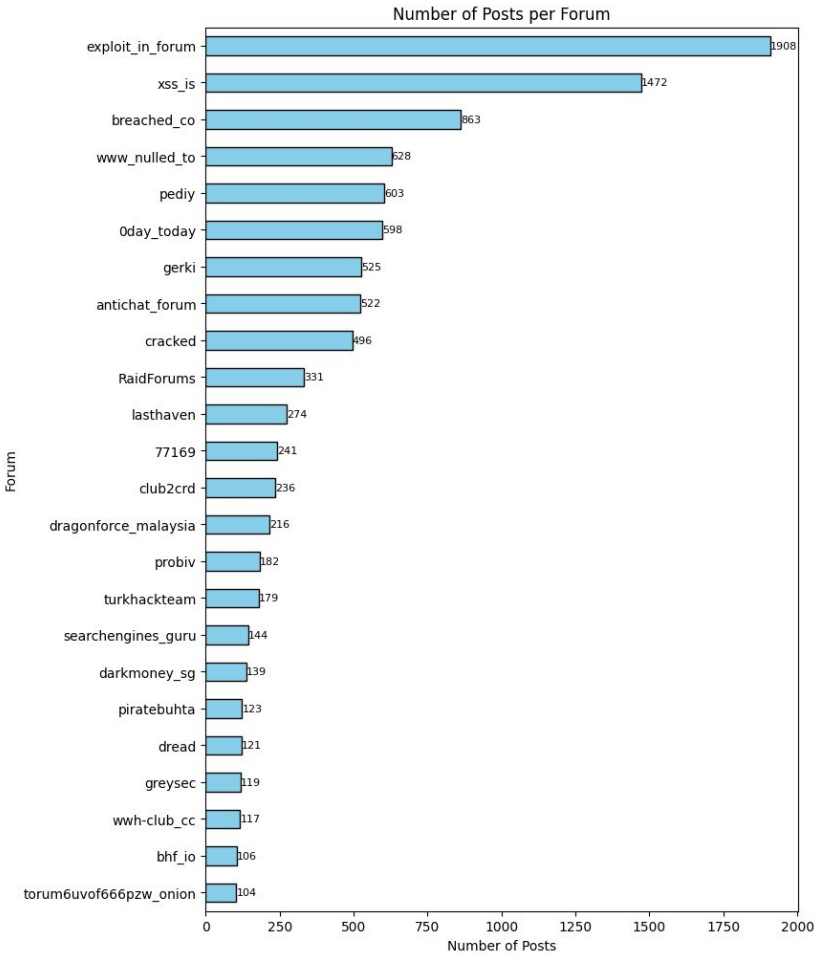


Figure 1.a) shows the yearly distribution of posts in our dataset. With 2022 being the year with the most posts followed by 2023 and 2019. The lower number of posts in 2023 could be explained by the time of the data collection, which happened at the beginning of summer 2023.

Figure 1.b) displays the top 25 forums with the most posts. The forum with the most posts is exploit_in with 1,908 posts (16.51% of total posts), followed by xss_is with a little less than 13% of all posts. We have a very uneven distribution of posts across forums, since 99 forums have less than 100 posts and 43 of them have less than 10 posts. The distribution for all 124 forums is available in Annex A

Figure 1.b).

Number of Posts per Forum



Further in the research the most popular forums were reviewed by two cybersecurity experts[6]. The forums were deemed as reliable sources of information on threat actors for current production of cyber threat intelligence.

To meet the first objective: identify areas of technical expertise in the form of communities of interest towards attack patterns, we created an actor-CAPEC bimodal social network with the data above, as explained below.

[6] This research was conducted as part of a MITACS research internship, which provided access to cybersecurity experts at SecureWorks. Experts were consulted throughout the research project to confirm the data sources and provide assistance during the analysis.

Social Network

A social network is a network of interconnected individuals that are linked by some sort or relationship (Wasserman and Faust, 1994; Borgatti, Everett and Johnson., 2013). Social networks can be represented as a graph, with vertices (nodes) representing actors and edges (links, ties) representing relationships between the actors. The nodes have their own attributes distinguishing them from one another. For instance, an attribute of a node can be its height, its hobbies, or its gender. Edges also have their own characteristics qualifying the type of link or relationship between two nodes (Borgatti, Everett and Johnson., 2013). For example, the relationship between Stephen Curry and Klay Thompson is that of teammates and Romeo and Juliet are lovers.

Social Network Analysis (SNA) allows the study of the most important actors in a network, the ones that are at the center of it. To identify those actors, SNA relies on centrality measures reflecting this central position in the network: degree, closeness and betweenness centrality. The degree of a node refers to the number of nodes linked to it (Wasserman and Faust, 1994; Borgatti, Everett and Johnson., 2013). The closeness centrality corresponds to the shortest path from a node to all other nodes. The idea behind closeness is that a central actor is the closest to the rest of the network (Wasserman and Faust, 1994; Borgatti, Everett and Johnson., 2013). On the other hand, betweenness centrality quantifies the number of shortest paths between two nodes that passes through a node. Betweenness centrality is used to identify nodes that play important roles in the flow of information in a network (Wasserman and Faust, 1994; Borgatti, Everett and Johnson., 2013).

The actor-CAPEC network developed in this study is a bimodal social network, as explained below.

Bimodal Social Network

Bimodal networks or two-mode networks are networks in which nodes can be categorized into two distinct groups or modes whereas a one-mode network only consist of one group of nodes. In bimodal social networks, modes are distinct set of entities or nodes (Wasserman and Faust, 1994). A mode represents the nature of the node: musical instrument vs musician, event vs participants, or in the case of this study, CAPEC vs actor. A bimodal social network allows the

study of ties between modes. Usually, one mode is the sender or initiator of the links, and the other is on the receiving end of it (Wasserman and Faust, 1994). In the context of this study, actors are the initiators of the link towards the CAPEC they interact with.

To build our bimodal network, we used the CVEs mentioned by actors. Each actor was assigned the CAPECs corresponding to the CVEs they mentioned: if actor A mentioned CVE-2022-45451, then actor A would be assigned the corresponding CAPEC of CVE-2022-45451, which is CAPEC 233. Once every actor had been assigned CAPECs, we built the bimodal network based on actor-CAPEC interactions. The bimodal network is unweighted, meaning that all edges between nodes have a weight of one regardless of the number of times an actor has mentioned the same CVE.

Considering the bipartite structure of bimodal networks, several centrality measures specific to bipartite networks have been developed. In this study, we will only use two of them: the in-degrees and out-degrees. The in-degree counts the number of links to a node from the opposite node mode. The in-degree will be used for CAPECs. The out-degree, on the other hand, counts the number of links from a node to the opposite node mode and will be used for actors as they are the initiator of the links.

The 500-in-degree filter

As mentioned above, based on the information collected, each actor was assigned the CAPECs corresponding to the CVEs they mentioned. Some actors mentioned CVEs that did not have CAPECs and were therefore removed from the network. These actors numbered 1,133, reducing the number of actors from 4,441 to 3,308.

Some CAPECs are present in a large number of attacks. However, in our analysis, we seek to identify precise attacks used by sub-groups of threat actors, and more general methods shared by a large number of actors are not beneficial to this aim. To go back to the criminology analogy, we don’t want to profile a criminal based on the fact that he picks locks or uses violence, which are both basic and very common methods that are known to many criminals. We would rather profile criminals that use more specific and complex attack strategies, such as elaborate social engineering schemes, embezzlement, or identity theft.

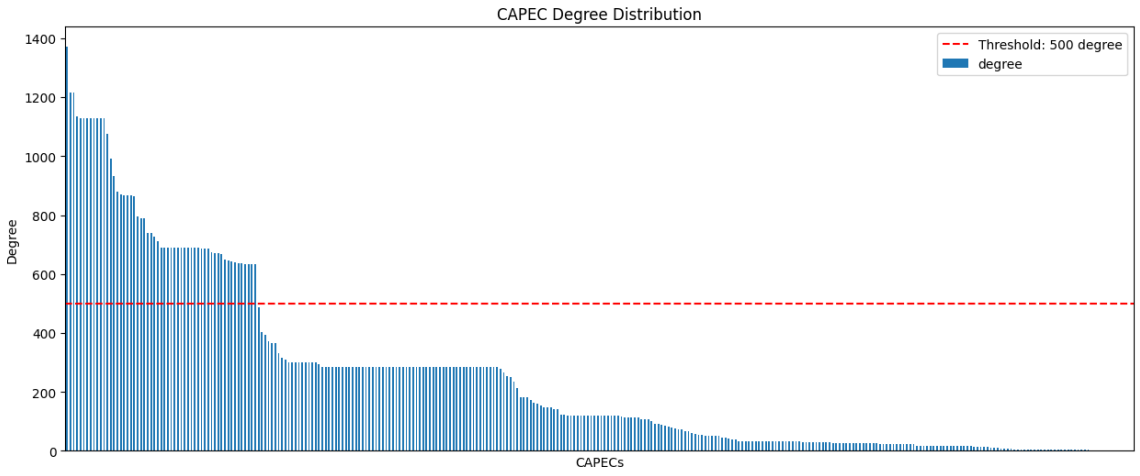
Hence, if a CAPEC is popular, shared by a large number of actors, it doesn't matter if it is specific or not, its relevance is limited and will pollute our attack profile analysis. For example, if every criminal in our dataset is able to perform embezzlement, then this modus operandi becomes irrelevant to classify criminals according to their speciality. By removing CAPEC shared by a large number of actors, we ensure that we categorize them by their more specific patterns that distinguish them from others.

We can also note global parents in CAPECs being shared by a large number of actors. As stated before, CAPEC operates within a hierarchical structure in which specific CAPECs are categorized as children under parent CAPECs (CAPEC, 2024). In other words, a child CAPEC is a refinement of the parent CAPEC's attack pattern. The hierarchical organization of CAPECs results in some CAPECs grouping together a very large number of more specialized attacks (or CAPECs), such as CAPEC 63: Cross-Site Scripting (XSS), CAPEC 88: OS Command Injection and CAPEC 66: SQL Injection. The latter are not useful for our analysis, since the more precise children of these attacks are the ones that will benefit us in identifying a specific attack profile.

To identify these generic and global CAPECs, we performed a visual analysis and determined that the most popular CAPECs were those mentioned by more than 500 actors, i.e. mentioned by more than 15% of total actors in the network. The choice of a 500-in-degree threshold was informed by the observation of an apparent 'elbow' in the data, where the in-degree distribution exhibits a notable change as shown in Figure 2.

Figure 2.

CAPEC in-Degree Distribution



Fifty-seven CAPECs exceeded the 500-in-degree threshold. Of those 57 CAPECs, 39 (68%) of them were general attack patterns and/or very popular ones (used by most actors), such as "Using Slash in Using Slashes" (CAPEC 79, 64, 78, 76) or "XML Injection" (CAPEC 250).

Out of the remaining 18, 16 CAPECs were related to one of the following big categories of attacks: "Buffer overflows", "SQL injections" or "XSS". Buffer overflow vulnerabilities are prevalent in many types of software and older software, making them an attractive target for the related attack. On top of having relatively easy attacks to execute, the most popular XSS and SQL injection vulnerabilities can be found in a wide range of web applications, making their related attack highly attractive for a broad range of attackers including those with limited technical skills. Even though these attack patterns are considered children of parent attack patterns, they seem to be popular in our dataset.

Only 2 out of the 57 CAPECs were outsiders: CAPEC 230 (Serialized payload data) and 231 (Oversized serialized payload data) were not considered particularly generic. However, they both had 670 different actors mentioning them, i.e 28.74% of all actors; making them very popular.

After the analysis of the most mentioned CAPECs, we decided to remove the 57 CAPECs that exceeded the 500-in-degree threshold. Filtering out the more general and popular attacks enables us to focus on the specializations of these attacks, rather than the global attacks (Insurance fraud vs. Fraud), and thus find threat actors with a more specific attack profile. For research reproducibility, the list of removed CAPECs is available in Table 2.

Table 2.

List of CAPECs removed from the analysis.

Category	CAPECs removed
General/Global	22, 108, 43, 85, 63, 13, 88, 7, 66, 136, 83, 250, 135, 3, 28, 72, 100, 126, 101, 35, 81, 104, 31, 473, 23, 153, 261
Popular	79, 64, 78, 76, 109, 67, 52, 53, 80, 267, 71, 120, 230, 231
Buffer overflows	45, 8, 9, 10, 14, 24, 46, 47, 42, 44, 123
XSS	209, 588, 73

This filter eliminated 987 actors from our dataset; their CAPECs having been eliminated from the analysis because they were too popular or general. Since these actors no longer had any CAPECs mentions, they were also removed from the dataset. This filtered dataset now contains 2,321 actors and 263 CAPECs.

Final Dataset

The bimodal actor-CAPECs network was built with this filtered dataset. Within this network, actors mentioning CAPEC X are linked to the latter. The network allows the visualization of relationships between actors and CAPECs, as well as those between actors through the exploitation of similar CAPECs. These relationships are the foundations of the identification of communities of interest towards a type of attack pattern. Finally, we computed the in and out-degree centralities for the bimodal network.

The network counts 263 CAPECs and 2,321 actors posting on a total of 116 different forums. Figure 3.a) presents the top 10 forums with the greatest number of posts mentioning a CVE (specialized post) and Figure 3.b) presents the top 10 forums based on the number of actors who posted a specialized post on them.

Figure 3.a)

Distribution of Posts mentioning at least a CVE per Forum

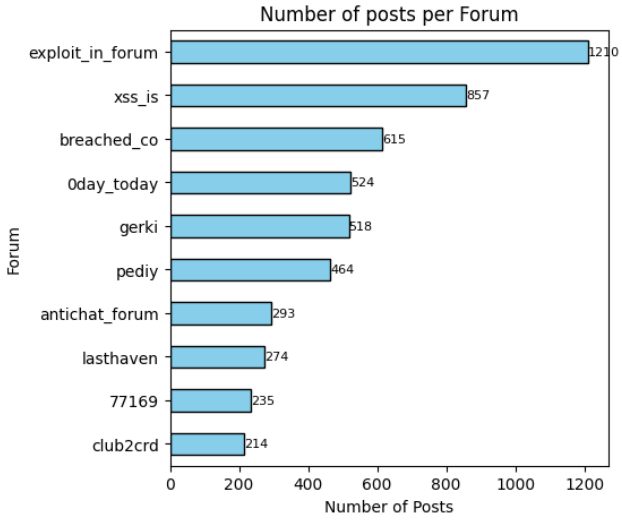
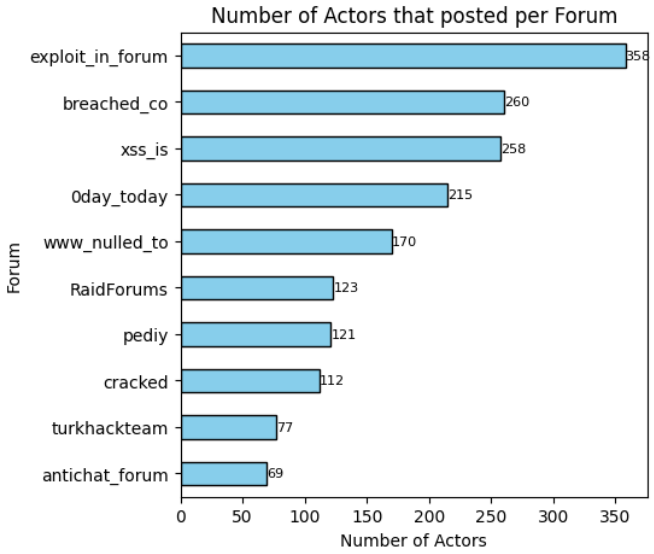


Figure 3.b)

Number of actors that posted per Forum



To provide an overview of the final dataset we selected four measures for actors displayed in Table 3: out-Degree, Number of specialized posts (mentioning a CVE), One timer and Number of specialized posts without one timers.

The out-degree of an actor is the number of CAPECs it interacts with in the network. Out-Degree is a metric indicating the level interaction of an actor within the network. Higher out-degree values suggest actors with broader connections to CAPECs, potentially playing a more central role in the network's structure and outlining potential communities of interest.

The number of posts refers to the number of specialized posts, i.e posts mentioning at least a CVE, with a unique ID per actor. This measures the posting activity of each actor. Actors with a higher number of specialized posts could be more active technical contributors, and their activity can be analyzed in terms of volume for further analysis of key actors.

The one timer variable is binary and discriminates between actors with a single specialized post and actor with multiple specialized posts in our final dataset. An actor that posted only once in

our final dataset is considered a one timer and is assigned the value one. Distinguishing between one timers and actors with multiple specialized posts is crucial for evaluating an actor's engagement and commitment to their CAPECs of interests.

Table 3.

Actors Overview

	count	mean	std	min	median	max
Out-Degree	2321	13.40	23.46	1	3	187
Nb specialized posts	2321	3.51	13.23	1	1	375
One timer	2321	0.56	X	0	X	1
Nb of specialized posts without one timers	1006	6.80	19.63	2	3	375

In our actor pool, actors are linked with, on average, 13 different CAPECs, while 50% of the actors are linked three or less CAPECs. At least 50% of our actors have posted only one specialized post in our dataset, while the actor with the most specialized posts has 375 of them. Overall, we have 56.70% of actors that posted only once and amongst those who posted multiple times, the average number of specialized posts is six.

As for the CAPECs, we picked the in-degree to provide an overall view of the CAPECs in our network. The in-degree of a CAPEC is the number of actors it has connections with in our network. A higher in-degree indicates that a CAPEC is connected to more actors, suggesting a higher level of interest in the network.

Table 4.

CAPECs Overview

	count	mean	std	min	median	max
In-Degree	263	118.22	119.40	1	54	478

CAPECs have connections with, on average, 118 actors, with the most popular one having connections with 478 actors, or 20.59% of total actors in our network. Once the network built, we focused on achieving the first objective. The next section details the analysis used.

Objective 1: Identify Areas of Technical Expertise in The Form of Communities of Interest Towards Attack Patterns

To identify areas of technical expertise in the form of communities of interest towards attack patterns, we applied a community detection algorithm to the actor-CAPEC network.

Community Detection Algorithm: Leiden

Community detection is the process of identifying communities, or clusters, in a network graph (Newman & Girvan, 2004; Bedi & Sharma, 2016; Anuar & al., 2021). Communities are defined as a group of nodes that interact more with each other than they do with other nodes in the rest of the network. To identify those communities, algorithms rely on the similarity between nodes, based on topological features and characteristics of the graph, and/or the strength of their connections (Newman & Girvan, 2004; Bedi & Sharma, 2016). These algorithms fall into two types of community detection methods, namely the agglomerative method and the divisive method. The former starts from an empty network and focuses on the addition of edges between similar nodes, based on a similarity measure. The latter is the other way around, it starts from a complete network and removes edges iteratively between dissimilar nodes, dividing the network into smaller communities with each iteration (Newman & Girvan, 2004; Anuar & al., 2021).

To judge the quality of the communities formed, a common measure introduced by Girvan and Newman in 2004 and known as 'Modularity' is used. Modularity is a measure of "the strength of the community structure" (Girvan & Newman, 2004). In other words, it measures how well a network can be divided into communities, such that the connections within each community are denser than the connections between communities. Modularity (denoted Q) ranges from minus one (–1) to one (1). A high modularity value indicates that there are strong intra-community connections and weak inter-community connections.

To identify community of interest, we use the Leiden community detection algorithm which falls into the agglomerative category. First introduced in 2019 by Traag, Waltman and van Eck, the Leiden algorithm is based on its predecessor, the Louvain algorithm (Blondel & al., 2008) and is more complex. On top of yielding higher-quality partitions, the Leiden algorithm performs well on small, medium, and large-scale networks.

Comprised of three phases, compared to two for Louvain, the Leiden algorithm is considered faster and returns higher quality partitions than the Louvain algorithm (Traag & al., 2019; Anuar & al., 2021). The three phases of the Leiden algorithm are the following: 1- modularity optimization process, 2- refinement of the partition and 3- aggregation of the network (Traag & al., 2019; Anuar & al., 2021). In phase one, nodes are moved from a community to another to find a partition that maximizes Modularity. The second phase is where the partition is refined, meaning that the algorithm merges small communities or divide larger ones to create a partition more in line with the underlying structure of the network, improving granularity and accuracy. In the last phase the algorithm combines smaller communities into larger ones until all communities have a certain minimum size. This step ensures that communities of the final partition are sufficiently large (avoiding communities of just one or two nodes) to be meaningful and represent well the underlying groups of nodes of the network (Traag & al., 2019; Anuar & al., 2021).

The selection of the Leiden algorithm was driven by its prevalence in criminological research, alongside the Louvain algorithm (Calderoni & al., 2017; Schaefer & al., 2017; Paquet-Clouston & Bouchard, 2023). Notably, the Leiden algorithm eliminates the need for explicit parameterization, such as the manual specification of the number of communities. Additionally, its primary aim is to discern dense clusters within the network. This approach aligns with established methodologies in criminology, ensuring robustness and consistency in our analyses.

Using the Leiden algorithm, different communities can be observed from the relationships between actors and CAPECs. Once we had performed the Leiden algorithm on the network, we extracted the CAPECs of each community to perform a qualitative analysis and see if the communities made sense in terms of their CAPECs of interest.

Content analysis: The interest behind each community

To make sense of the communities identified by the Leiden algorithm, this study uses content analysis. Content analysis is a widely used research method and consists of systematically analyzing the content of textual, visual, or audio information to identify, code and categorize recurring patterns or elements (Krippendorff, 2013; Hsieh & Shannon, 2005). This enables researchers to extract meaningful insights from the data, understand patterns, and draw conclusions about the studied subject (Elo & Kyngäs, 2008; Neuman, 2014). Its versatility makes it applicable across diverse fields, including social sciences (Weber, 1990), communication studies (Neuendorf, 2002), marketing (McQuarrie & Mick, 1999), and psychology (Elo & Kyngäs, 2008), where it serves to analyze and interpret the content of various materials. Thus, content analysis provides a structured and systematic approach to uncovering patterns and themes within large sets of data, contributing valuable insights to our first research objective.

Leiden is a tool allowing to discover communities underlying the network. To assess the coherence of these communities, we extracted their CAPECs of interest, the ones they were associated with inside the community, and examined whether they made sense and were meaningfully associated or simply assembled in an arbitrary manner. The former case would mean that our communities are interested in a particular strategy using certain related attack pattern whereas the latter would mean that our communities were just chatting about attack patterns that didn't make sense together.

The analysis involved two steps. First, a meticulous examination of the descriptive content associated with each CAPEC such as its description, its relationships with other CAPECs, its domain of attack and mechanism of attack was conducted. Through this method, distinct themes and attack pattern categories within each community were identified. Second, the themes and attack patterns categories for each community were reviewed and validated by two cybersecurity experts[7]. Once an agreement was reached, a nomenclature was assigned to these communities. Such nomenclature reflects the communities' respective interests in specific attack patterns, as illuminated by the content analysis.

[7] This research was conducted as part of a MITACS research internship, which provided access to cybersecurity experts at SecureWorks. Cybersecurity experts were consulted to analyze the themes and categories of types of attack identified. They made sure categories were coherent in the types of attack that were associated together.

Once communities of interest were established and interpreted, we focused on identifying key actors inside those communities.

Objective 2: Detect Key Actors Based on their Technical Expertise Level

Having established the distinct communities of interest (CoIs) based on shared CVE/CAPEC discussions in the previous section, key actors within each community could then be identified. Building upon the expertise framework proposed by Bouchard and Nguyen (2011), we aimed to identify individuals demonstrating high level of technical expertise, i.e. scoring high on both skill level and commitment towards their community's specific attack patterns. Then, drawing from previous literature stating that key actors are among the most senior and active actors (Benjamin and Chen, 2012; Abbasi and al., 2014; Zhang & al., 2015; Samtani and Chen, 2016; Samtani & al., 2017; Grisham and al., 2017; Johnsen and Franke, 2020; Huang and al., 2021), activity rate will come as an extension to our criminological framework to nuance expertise profiles based on their seniority and posting activity. Thus, key actors in our study are 'Professionals' exhibiting the highest level of technical expertise and striking the right balance in activity rate between seniority and diligence in their posting activity.

Skill Level

Skill level is one of two characteristics used to assess technical expertise level in our framework. The skill level assessed is a technical skill level. The technical skill level of actors is assessed through the CAPECs they are linked with. We leveraged the 'Skill Level Required' metric assigned to each CAPEC by MITRE, reflecting the 'Skills Level Required' to execute the attack. This metric has three discrete values: 'Low', 'Medium', and 'High', delineating degrees of skill proficiency necessary for successful execution. A CAPEC can have multiple skill level values, for example CAPEC 32 has a skill level of ['Low', 'High'] covering the different scenarios of the attack pattern and their respective skill level required.

Handling CAPECs missing a skill level value

However, among the 263 CAPECs in our network, 141 of them didn't have a skill level value assigned directly by MITRE, the organisation responsible for CAPEC. To address this limitation, we adopted a hierarchical approach wherein skill level values were propagated from direct child CAPECs to their parent entities missing a value. As stated before, CAPEC operates within a hierarchical structure in which specific CAPECs are categorized as children under parent CAPECs (CAPEC, 2024). In other words, a child CAPEC is a refinement of the parent CAPEC's attack pattern: the skill level required for the child technique is a reflection of the overall skill level needed for the parent category. Hence, we prioritized using skill levels from direct children CAPECs.

Parent CAPEC having multiple children were assigned their children's values to reflect the different scenarios that could arise from the parent CAPEC. Hence, a parent with two children has a skill level value of [child 1 skill level, child 2 skill level] and a parent with only one child but this child has 2 skill level values will also have two skill level values [child skill level 1, child skill level 2]. Using the child's skill level reduced the number of CAPECs without a Skill Level to 118.

Subsequently, for the subset of 118 CAPECs without any direct child entries with skill level values, we extrapolated skill level values from their direct parent CAPECs. As CAPEC follows a hierarchy, if a CAPEC has no direct children but has a parent, some level of complexity is likely inherited from that parent CAPEC. CAPECs within a broader category (parent) often share some foundational aspects with their children, making it a relatively fair assessment of the child CAPEC's skill level. Out of the childless 118 CAPECs, 32 of them had a direct CAPEC parent with a skill level value; allowing us to reduce the number of CAPECs without a value to 86.

Ultimately, the remaining CAPECs were assigned a skill level value through manual analysis and iterative refinement. Most of these CAPECs included different attack scenarios. Thus, they were first assigned a range of skill level as a value (ex: ['Low', 'Medium'] or ['Medium', 'High'] or ['Low', 'Medium', 'High']) to capture the changeability and potential skill level required to carry

out different attack scenarios. Then, the values assigned were reviewed and validated by two cybersecurity experts[8]. The list of CAPECs and their respective origin of skill level value is available in Annex B.

Skill Level Metric

Once we had at least a skill level value for each CAPEC in our study, we assigned each CAPEC its highest skill level scenario value. This decision ensures we don't underestimate actors in terms of skill level. We also avoid the risk to miss important actors because the CVE they mentioned were linked to CAPECs with a variety of skill level scenarios. We prefer to include intermediate actors rather than miss highly skilled ones.

Each actor was assigned its CAPECs' skill level value, measured with a list of skill level values, according to their CAPECs mentions in the network. In short, if an actor mentioned four CVEs, and those CVEs were linked to six CAPECs, this actor's list would consist of six skill level values ranging from 'Low' to 'High'. For instance, an actor that mentioned the following CVEs [CVE-A, CVE-A, CVE-B, CVE-C], which reference the following CAPECs [207, 207, 112, 111], will have a skill level list of ['High', 'High', 'Low', 'Medium'] where CAPEC 207 had two skill levels (['Low', 'High']) and was assigned the highest skill level value of ['High'] by the procedure explained above.

The overall distribution of skill level values amongst actors' list is available in Table 5 below. Additionally, Table 6 provides further statistics on the distribution, including the average proportion of 'High' values, the median, the 75th percentile, and the standard deviation.

[8] This research was conducted as part of a MITACS research internship, which provided access to cybersecurity experts at SecureWorks. Experts were consulted to make sure the skill level value assigned to each CAPEC was coherent with the skill level required to carry the corresponding attack in a real-life scenario.

Table 5.

Overall Distribution of Skill Level Values

Skill Level Value	Nb CAPECs	% of skill level values among all values in actors' list
Low	118 (44.87%)	57.71%
Medium	66 (25.09%)	24.14%
High	79 (30.04%)	18.14%

Table 6.

Skill Level Values Proportion Statistics

Skill Level Value	Average Proportion in actor's list	Median	75th percentile	std
High	29.07%	23.08%	50.00%	30.76%
Medium	36.12%	30.77%	50.00%	32.41%
Low	33.74%	33.33%	66.66%	31.72%

As shown in Table 5, 44.87% (118/263) of all CAPECs have been assigned a 'Low' skill level value and they occupy more than half (57.14%) of all mentioned CAPECs. Just over a quarter (25.09%) of all CAPECs have a 'Medium' value. The 'Medium' CAPECs account for 24.14% of all mentioned CAPECs. Finally, 30.04% of our CAPECs have a 'High' value. CAPECs with a 'High' skill level value amount for 18.14% of all mentioned CAPECs. As shown in table 6, the average proportion of 'High' values in actor's skill level list is 29.07% (std = 30.76). So, the average actor has approximately 30% of its list being 'High' values. Half of our actors have a proportion of 'High' occurrences less than or equal to 23.08% and 25% of them have a proportion of 'high' occurrences greater than 50%.

Each actor's list was transformed into a numerical list where 'Low'=1, 'Medium'=2 and 'High'=3. This way, an actor's list went from ['Low', 'Medium', High'] to [1, 2, 3].

The Actor's Skill Level Metric: the 70th percentile

To establish a single representative skill level value for each actor, the 70th percentile (7th decile) value from each actor's list was chosen as their skill level value. The choice of selecting a higher percentile value is first conceptual. Specifically, an actor with a certain proportion of 'High' values, regardless of other values in their list, is perceived as more technically proficient than an actor with only 'Medium' and 'Low' values. Consequently, the presence of 'High' values becomes pivotal in assessing the skill level of each actor.

Our decision to use a higher percentile was compared to weighted mean and median alternatives using the skill level value distribution. Considering the distribution of skill level values, particularly the imbalance between 'High' values and others, a weighted mean would not have accurately represented actors' skill levels. For instance, an actor with a list of [1, 1, 1, 3, 3, 3] would have been assigned a skill level of 2 ('Medium'), which does not reflect the significant presence of 'High' values in their list. Thus, the weighted mean was discarded.

Similarly, the choice of a higher percentile was contrasted with that of the median. Once again, the distribution of values and the average proportion of 'High' values guided our decision. With 50% of actors having less than 24% 'High' values in their lists, the median would also have been skewed downwards and thus not representative of an actor's skill level. For instance, with a list of [1, 1, 1, 1, 2, 2, 3, 3, 3, 3], the actor would have been assigned a skill level of 2, failing to accurately reflect the over 40% 'High' values in their list.

Given the limitations of weighted mean and median, we selected the 70th percentile. This choice aligns with the study's objective of identifying key actors with elevated skill levels. By assigning a 'High' value only to actors with over 30% 'High' values in their lists—corresponding to those above average—we ensure that only those demonstrating a significant level of 'High' occurrences are categorized as highly skilled actors.

The decision to refrain from selecting a lower percentile (meaning assigning 'High' to those with more than 30% of 'High' values) is also influenced by our theoretical framework, which

considers two variables: commitment and skill level. Key actors must score high on both variables to be considered key. Being too harsh on the assignment of the 'High' value could risk having too small of a highly skilled population for our analysis to be successful. Opting for the 70^{th} percentile allows for a 'High' assignment for actors above average while allowing for a large enough population for our analysis.

Given that the 70^{th} percentile represents the value separating the uppermost 30% of the data, it follows that each actor should ideally have a minimum of four skill level values to make sure of the applicability of this calculation. Indeed, with three values, the median and the third quartile, and thus the 70^{th} percentile, are the same thing. So, using just three values won't give us an accurate picture because the median and the 70^{th} percentile will be identical. Actors with fewer than four values would not provide a robust basis for calculating the third quartile, potentially leading to skewed or inaccurate assessments of skill level. Thus, actors with less than four skill level values in their list were filtered out.

Commitment

Commitment, the other characteristic assessing technical expertise level in our framework, was evaluated through an actor's focus within their communities of interests (CoI). It was operationalized using a majority threshold, as explained below.

The Majority Threshold

Commitment was operationalized by examining the percentage of an actor's posts primarily referencing (CAPEC) entries within their CoI. A post containing multiple CAPECs, the majority ($x \geq 50\%$) of which belong to the same CoI as the author, is considered as being 'in-interest'. Such a criterion is called a "majority threshold". For example, a post with five CAPECs, where three of them belong to the same CoI as the author is considered as 'in-interest'. On the contrary, a post containing CAPECs predominantly from CoIs other than that of the author is labeled as 'out-interest'. A post with a single CAPEC, and this CAPEC being in the same CoI as the author,

is considered ‘in-interest’. The degree of an individual’s commitment to their CoI is quantified as the percentage of in-interest specialized posts relative to their total number of specialized posts.

The selection of the majority threshold for CAPECs was influenced by two key considerations. First, employing a stricter engagement metric, wherein all referenced CAPECs must belong to the author’s CoI for a post to be deemed ‘in-interest’, posed the risk of undue selectivity, potentially resulting in the exclusion of actors meeting both skill level and commitment criteria. Second, it is plausible that generic or complementary CVEs may be mentioned in posts to supplement queries or scrutinize the compatibility of one CVE with another, for instance. However, the inclusion of such supplementary CVE and thus of the corresponding CAPECs, essential for many attacks, does not necessarily indicate deviation from an actor’s CoI. In some instances, it may signify that the actor wants deeper insights. A strict commitment metric, as described earlier, would penalize such behavior by excluding posts even if only one CAPEC outside the CoI is referenced among many. Moreover, a strict commitment metric would disproportionately penalize actors referencing a higher number of CAPECs in their posts, as they are more likely to mention a CAPEC outside their CoI. The choice of the majority appeared fair regardless of the number of CAPECs mentioned per post: it allows referencing of complementary CAPECs for deeper understanding without penalizing commitment.

An actor’s high focus on the attack pattern within their community is interpreted as a sign of their dedication to keeping their knowledge current and relevant. Thus, according to the relative approach on expertise, such dedication is considered a strong indicator of their commitment to maintaining expertise in this domain.

We also decided to eliminate the one timers from our key actor analysis. Since they only have one post, their commitment levels would either be 0 or 100%. This way, the commitment can’t be assessed on a single post because the step from 0 to 100 is too large. Filtering the one timers brought the dataset down to 1,006 actors.

Table 7.

Descriptive Statistics of the Number of specialized posts per actors without one timers

	count	mean	std	median	60th percentile	75th percentile	max
Nb specialized posts without one timers	1006	6.80	19.63	3	4	5	375

The literature review states that key actors are amongst those who post the most (Abbasi and al., 2014; Grisham and al., 2017; Johnsen and Franke, 2020; Huang and al., 2021), suggesting that those who contribute more frequently are likely to be more deeply engaged, or in our framework, committed to their communities. It is important to remember that the posts in our dataset don't represent all the post an actor posted, it only represents the number of 'specialized' posts, the ones that mentioned at least a CVE. While the literature review still applies, we have to keep this detail in mind when discussing the number of posts.

Considering the distribution of number of specialized posts in our final dataset, as shown in Table 7, with 50% of our actors with three specialized posts or less, it is evident that a significant portion of actors in our final dataset have relatively few specialized posts. In our case, filtering out actors with fewer than a threshold number of specialized posts seemed coherent in order to obtain a representative percentage of commitment.

By choosing a threshold of four specialized posts minimum, we keep a substantial proportion of the population while getting commitment levels that are easier to work with, given the majority threshold. If we included those with only three posts, their commitment levels would either be 33% (1/3), 66% (2/3), or 100% (3/3). These levels are unclear for evaluating commitment because the jumps from one level to another are too big. However, setting the threshold at four gives us smoother and more meaningful commitment levels. With four posts, the commitment level can either be 25% (1/4), 50% (2/4), 75% (3/4), 100% (4/4) for those with very few posts.

Filtering out actors with fewer than four posts and fewer than four skill level values in their list reduced our final dataset to 359 actors.

Activity rate

Considering forums' population is ephemeral (Hughes & Hutchings, 2023) and that key actors are amongst the most active and senior actors (Benjamin and Chen, 2012; Abbasi and al., 2014; Zhang & al., 2015; Samtani and Chen, 2016; Samtani & al., 2017; Grisham and al., 2017; Johnsen and Franke, 2020; Huang and al., 2021), activity rate was added as the third variable of this analysis.

Activity rate combines both seniority and posting activity and comes as an extension of Bouchard and Nguyen's framework. Activity rate allows to nuance our expertise profiles based on seniority and diligence of posting activity. In our analysis, activity rate is operationalized as follows:

activity rate = nb specialized posts / total activity time in nb of days

The total time of an actor's activity corresponds to the amount of time between its oldest and latest post in our dataset. The number of specialized posts accounts for an actor's posting activity and the total activity time in number of days accounts for an actor's seniority. The activity rate is the number of specialized posts (posting activity) divided by the total time of an actor's activity (seniority). An actor with 10 posts, with the oldest dating back from March 20th, 2021, and the latest being from December 20th, 2021, will have a total activity time of 275 days. The activity rate for this actor is thus 0.036 (10/275).

Key actors in our study will then be 'Professionals' exhibiting the highest level of technical expertise, i.e. high scores on both skill level and commitment, and striking the right balance in activity rate between seniority and diligence in their posting activity.

Sample

Table 8 presents the descriptive statistics of our sample of actors. Our sample for key actor identification consists of 359 actors. The average actor has 36.68% of its posts committed to its CoI and shows a skill level of 2.19 ('Medium'). We can see from the median of the length of skill level value lists that half our final actors have 25 or less skill level values in their list and posted six times or less specialized posts. A quarter of our sample has more than 50% of their posts committed to their CoI and showcases a 70th percentile skill level value of 3.

Table 8.

Descriptive Statistics of Sample

	mean	**std**	**min**	**median**	**75th percentile**	**max**
Length Skill Level values list	99.42	255.76	4	25	85	3449
Skill Level 70th percentile value	2.19	0.64	1	2	3	3
Nb of specialized posts	14.55	31.37	4	6	10	375
% commitment	36.68	29.61	0	25	50	100
Activity time (days)	449.07	545.02	1	227.00	690.00	2669.00
Activity rate	0.72	1.90	0.002	0.04	0.20	14.00

Finding Key Actors

An actor exhibiting the highest level of technical expertise ('Professional'), i.e. scoring high on both skill level and commitment towards its CoI, and striking the right balance in activity rate between seniority and diligence in their posting activity will be considered key in our framework. To identify those key actors based on our three variables, we used a clustering algorithm. By grouping actors based on our three variables we can identify clusters of actors exhibiting similar characteristics. To do so, we employed the K-means clustering algorithm[9].

The K-means algorithm groups data points into a partition of a predefined number of clusters (k) to discover patterns and structure in the data. This clustering is based on iteratively minimizing the Euclidian distance between each data point and the centroid[10] of its assigned cluster (Jain, 2010). This way, by allocating data points to the nearest centroids in each iteration, the K-means algorithm aims to minimize the overall distance between the data points and their respective cluster centroid (Jain, 2010); thus, grouping similar data points together.

Typically, selecting the optimal number of cluster (k) implies running K-means for different values of k and the partition that appears the most meaningful is selected (Jain, 2010).

In our selecting process, we first employed the silhouette score, a popular metric for evaluating the clustering quality, and computed models with different k values. The silhouette score is a measure of cluster cohesion and separation, where higher scores indicate a higher clustering quality (Rousseeuw, 1987). The silhouette score is the average difference between 1- the distance between a point and its cluster's centroid and 2 - the distance between this point and the nearest foreign cluster's centroid. As the difference increases, the clusters become more distinct, with data points showing greater cohesion within their respective clusters. The silhouette score allows us to compare models with different k values and identify the ones that create high quality clustering. Figure 4 displays the silhouette score for each k model.

[9]We experimented with several clustering algorithms, including Hierarchical Clustering, K-means, and DBSCAN. Upon evaluating their respective performances, the K-means algorithm performances were equal or better compared to the others. Consequently, we selected K-means as the preferred clustering method for the purposes of our investigation.
[10] The centroid is the point with the average coordinate of all the points within that cluster. It is often refered to as the 'center of the cluster'.

Assessing statistical differences between clusters

To make sure our clusters present statically significant differences, their respective skill level, commitment, and activity rate were measured and compared. Due to the abnormal distributions, we used the non-parametric Kruskall-Wallis test and Dunn's post hoc tests. Tests were computed using the python scikit.stats[11] and scikit_posthocs[12] packages. The threshold of significance was set to 0.05.

Kruskall-Wallis H tests were used for comparing multiple clusters. When significant, ($p < 0.05$), we used the eta-squared to measure effect sizes. The Eta-squared formula below was computed:

$$\text{Eta-squared } (H) = (H - k + 1) / (n\_total - k)$$

Where H is the result of the Kruskall-Wallis test, k is the number of clusters and n_total is the total number of actors in our analysis. An eta-squared ($0.01 < \eta^2 < 0.06$) is considered a small effect, an eta-squared ($0.06 < \eta^2 < 0.14$) represents a moderate effect and $\eta 2 > 0.14$ represents a large effect. (Tomczak & Tomczak, 2014).

When Kruskall-Wallis tests were significant, to identify the pairwise significant differences, we used the Dunn's post hoc test with a Bonferroni correction. Bonferroni correction was applied to adjust the p-value and reduce the chances of false positives. Finally, effect sizes were computed using Hedge's g, using the formula below, for pairs of clusters with significant post hoc test.

$$g = (\bar{x}_1 - \bar{x}_2) / \sqrt{[((n_1-1).s_1^2 + (n_2-1).s_2^2) / n_1+n_2-2]}$$

n_1 and n_2 are the sample sizes and s_2^1 and s_2^2 are the sample variances and $n_1 + n_2 - 2$ the number of degrees of freedom. A g ($g \approx 0.20$) is considered a small effect, ($g \approx 0.5$) is considered a medium effect and ($g \geq 0.8$) a large effect (UCLA, accessed March 28, 2024).

[11] https://docs.scipy.org/doc/scipy/reference/stats.html
[12] https://scikit-posthocs.readthedocs.io/en/latest/generated/scikit_posthocs.posthoc_dunn.html

Having presented the methodology, the next section will focus on the results of this study.

Ethical Considerations

The study has been approved the ethics committee at the University of Montréal (project 2023-4678) under minimal risks. The study required asking for a waiver of consent in line with Article 5.5A of the Canadian Tri-Council Policy Statement on research ethics. To ensure participants' confidentiality and privacy, real pseudonyms of the actors are not displayed throughout the text.

Results

This section presents the key findings of this research. It starts with a presentation of the bimodal network actor-CAPEC. Subsequently, to answer the first research objective, the distinct communities of interest are presented. Finally, we identify key actors and analyze their distribution among communities of interest to answer the second research objective.

The Bimodal actor-CAPEC Network

The CAPEC-actor bimodal social network counts 2,584 nodes (2,321 actors and 263 CAPECs), and 31,093 edges. The network has a mean bilateral degree (in and out degrees combined) of 24 and a density of 0.009, meaning that less than 1% (0.9%) of possible connections between nodes are in the network. On average, actors are connected to 13 different CAPECs while CAPECs have links with 118 actors, as shown in Table 9.

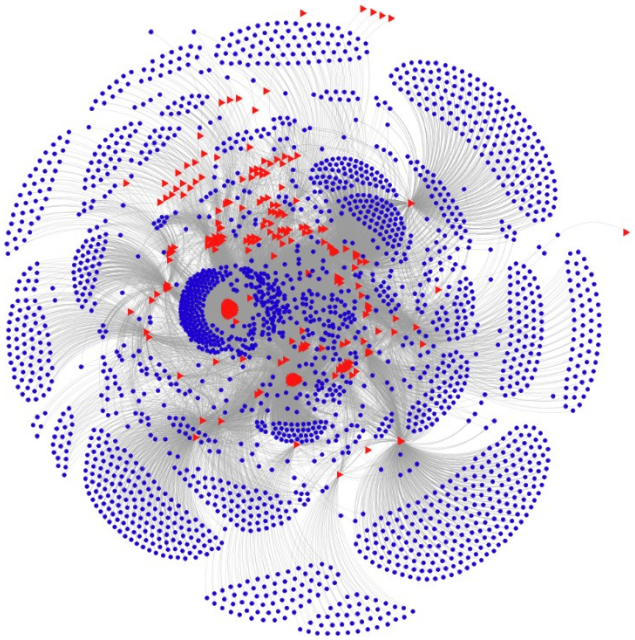
Table 9.

Actor-CAPEC network characteristics.

	count	**Mean degree**	**Std degree**
Actors	2321	13.40	23.46
CAPECs	263	118.22	119.40

Figure 4.

Bimodal actor-CAPEC Network.



Note. The representation of the graph uses the Fruchterman Reingold projection with the following settings in Gephi: zone=10000; Gravity=7.0; Speed=5.0.

Figure 4 depicts the network graph with a color per mode: the actors are in blue and the CAPECs are in red. Within the Fruchterman-Reingold projection, nodes positioned closer to the center generally exhibit higher connectivity, potentially serving as hubs or mediators facilitating information flow (Newman, 2018). Interestingly in our network visualization, CAPECs (represented in red) occupy central positions. This finding aligns with their inherent role as the connective tissue linking actors within communities of interest.

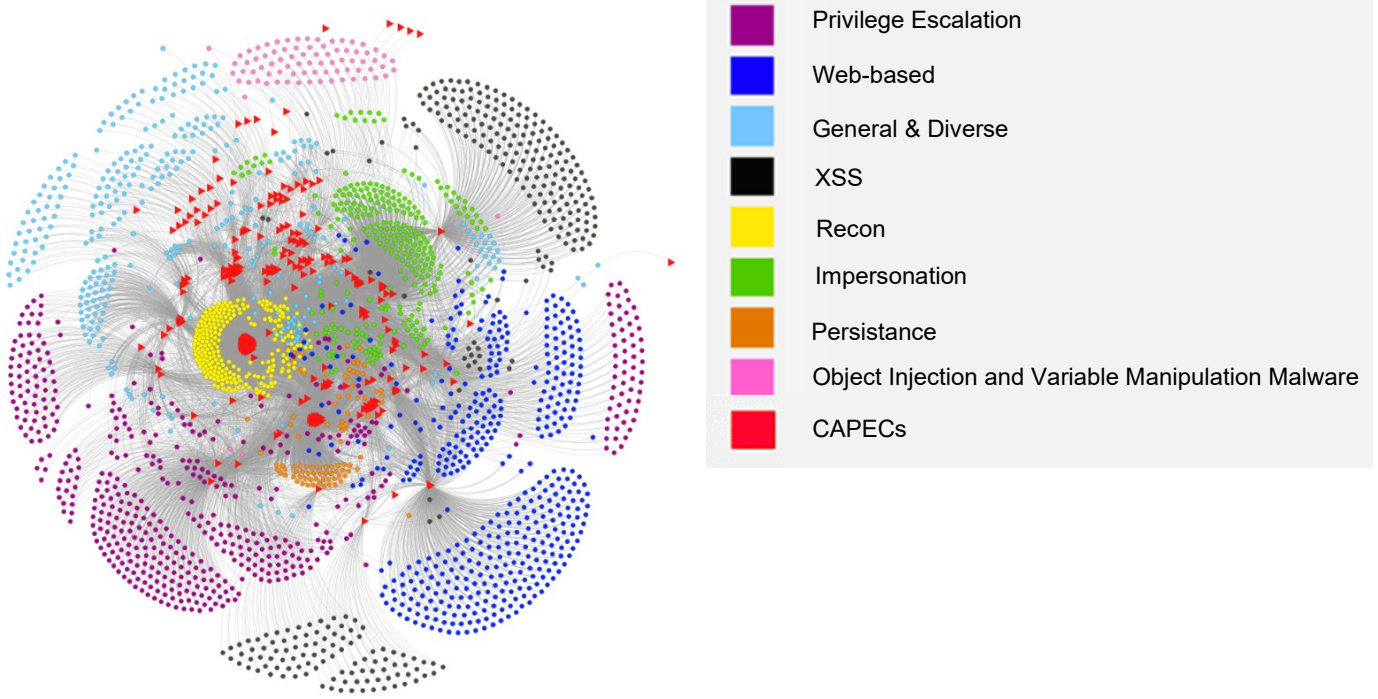
While the Fruchterman Reingold projection aims to minimize edge crossings, node positions can vary slightly between runs due to the algorithm being non-deterministic (Fruchterman & Reingold, 1991). Thus, while the observation of CAPECs position on the spatial representation of the network is interesting, the specific placements shouldn't be overinterpreted based on a single visualization.

Mapping areas of potential Technical Expertise: Communities of Interest and Their Preferred Attack Patterns

Through iterative application of the Leiden algorithm, we identified eight distinct communities within the network. The final partition achieved a modularity score (Q) of 0.473, the highest among the iterations, exceeding the well-established threshold of 0.3 and indicating a substantial level of cohesion within the identified communities (Newman & Girvan, 2004). Figure 5 depicts the network using the same projection settings but this time the actors are colored according to their respective communities (presented below). CAPECs remain red for clarity.

Figure 5.

Bimodal actor-CAPEC Network Colored according to Communities of Interests



As explained in the methodology, we performed content analysis to extract the interest in specific attack patterns for each community. Table 10 presents the communities and their attack patterns of interest.

Table 10.

Communities and their attack pattern of interest

Community	Attack Pattern interest
0	Privilege Escalation (PrivEsc)
1	Web-based
2	General and Diverse
3	XSS
4	Reconnaissance and Scanning (Recon)
5	Impersonation
6	Persistance
7	Object Injection and Variable Manipulation Malware (OIVMM)

To provide a better overview of the communities of interest, the communities' characteristics are presented in Table 11, including the number of nodes, number of actors, number of CAPECs, the percentage of one timers, the average out-degree per actor and the average number of posts per actor. The average out-degree represents the number of unique CAPECs an actor is linked with. Then, each community is interpreted, and their respective characteristics are presented below.

Table 11.

Communities of Interest (CoI) Overview

Community n°	Community of Intrest	nodes	CAPEC	actors	% one timers	mean out-degree per actor	std	mean nb of specialized posts	std
0	PrivEsc	544	19	525	65.14	4	7.11	2	4.76
1	Web-based	497	26	471	71.97	5	12.98	3	18.33
2	General	431	103	328	56.10	14	33.15	7	24.89
3	XSS	319	10	309	71.52	2	1.18	1	1.46
4	Recon	298	55	243	51.44	61	9.04	3	6.99
5	Impersonation	296	25	271	54.61	12	7.88	3	5.49
6	Persistence	116	22	94	41.49	26	25.76	5	7.96
7	OIVMM	83	3	80	85.00	1	0.31	1	1.62

Community 0 is related to privilege escalation attack pattern. Privilege escalation is when an attacker gains more access or control over a system than they should have. In simple terms, it's like someone sneaking into a restricted area and getting higher-level permissions. The PrivEsc community is the most populated with 525 actors, i.e. 22.62% of all actors, but has the third least number of CAPECs (19). In this community, 65.14% of actors are one timers. On average, PrivEsc actors have links with four CAPECs and posted twice in our final dataset.

Community 1 is focused on web-based attacks. Web-based attacks are attacks that target web interfaces. Web-based attacks are like digital break-ins, where criminals exploit weaknesses in websites or online systems, causing disruption or stealing valuable digital possessions. The Web-based community counts 471 members, and 26 CAPECs, making it the second largest community in our network. However, more than 70% of its population is a one timer (71.97%) making it second overall for one timer percentage. The average member of this community posted three times and has links with five CAPECs.

Community 3 specializes in XSS attacks. XSS or Cross-Site Scripting is when attackers inject harmful JavaScript code into a website or application that is then seen by other users. In the digital world, attackers inject harmful code into a website, and when others visit the site, they unwittingly execute the code, allowing the attacker to carry out their plans. With a percentage of 71.52% of one timers among its 309 actors, the XSS community holds third place in one timer percentage. The XSS community counts 10 CAPECs, which is the second lowest number of CAPECs. On average, members of this community have posted once and have links with two CAPECs.

Community 4 only contains CAPECs about reconnaissance and scanning. This is like criminals studying a neighborhood before a robbery. It involves gathering information about a target system to find weaknesses. The Recon community has the second higher number of CAPECs with 55. The percentage of one timers is also among the lowest with 51.44% of their 243 members having posted only once. The Recon members sit at the top for average number of

CAPECs: they have link with 61 CAPECs, with an average number of posts of three. The average member mentions around 20 CAPECs in a single post.

Community 5 is about impersonation attacks, combining CAPECs about authentication bypassing and spoofing. Put simply, just as someone might pretend to be a police officer to gain trust, in the digital world, attackers can pretend to be someone they're not to trick others into giving them access or information. The Impersonation community is populated by 271 actors and 25 CAPECs. In this community, just over half (54.61%) of actors are one timers. On average, Impersonation members have links with 12 CAPECs and posted three specialized posts.

Community 6 is focused on the persistence step techniques of an attack or a malware. This is when attackers solidify the attack's foothold on the system by writing it onto the disk to make sure their "presence" in a system lasts for a long time, even after the initial attack or even a restart. The persistence community has the lowest percentage of one timers with 41.49% of its 94 actors being one timers. The Persistence community counts 22 CAPECs. However, the members of this community have the second highest average number of CAPEC they are linked with, with 26 CAPECs, and posted, on average, five specialized posts.

Community 7 is related to object injection and variable manipulation malware (OIVMM) attacks. Think of a computer program as a complex board game. Each game piece (object) and rule (variable) are carefully defined. A player (attacker) can join the game and sneakily alter the rules or tweak the game pieces without others noticing. By changing the rules of the board game, the attacker gets an unfair advantage. This way, much like cheating in a game to achieve an unfair advantage, the OIVMM can lead to unauthorized access, information theft or control over the digital system. The OIVMM community is the smallest of our network. With only 80 actors and three CAPECs, this community holds the first place in terms of percentage of one timers with 85% of its members having posted only once. This percentage is reflected in the average number of post and CAPECs they are linked with, both being equal to one.

And lastly, Community 2 is a diverse community that couldn't settle for a specific attack pattern. This community contains a myriad of different CAPECs that are not specific to certain types of

target or attack patterns. With more than 100 CAPECs related to a myriad of different attack patterns unrelated to one another, this community seems to be home of the only diverse/versatile community. Having 103 CAPECs places this community at the top for number of CAPECs. However, it places third for the number of actors with 328. This community has 56.10% of one timers. Actors posted, on average, seven specialized posts and had connections with, on average 14 CAPECs.

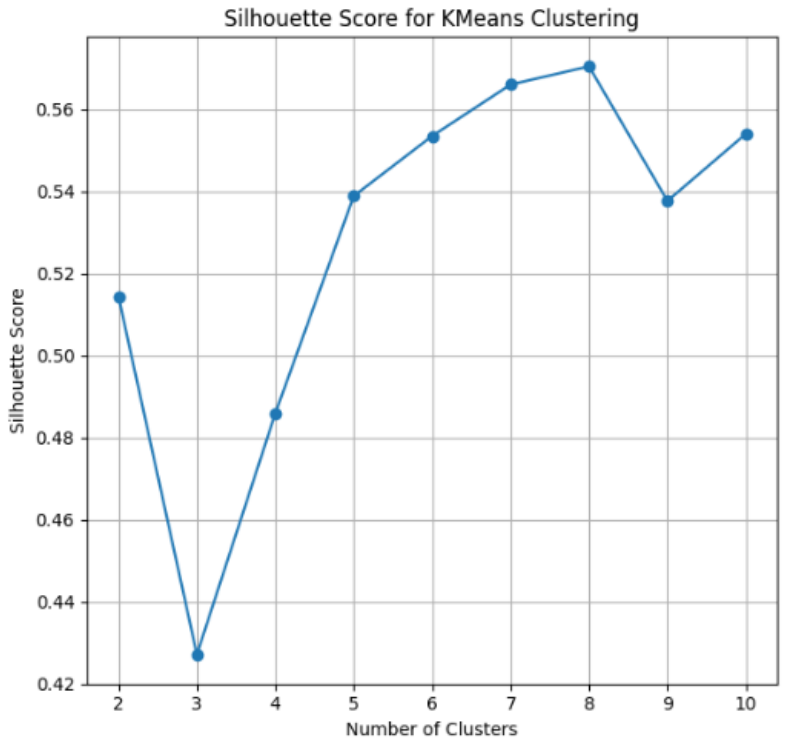
Then, having established the distinct communities of interest based on shared CVE/CAPEC discussions, we shifted our focus to identifying the key actors based on their technical expertise in their respective community. Building upon the expertise framework proposed by Bouchard and Nguyen (2011) and previous literature, we identified individuals demonstrating the highest level of technical expertise, i.e. high skill level, high commitment towards their CoI, and the right balance between seniority and diligence in their posting activity (activity rate).

Unveiling the Spectrum of Actors and Key Actors

To identify our key actors based on their technical expertise level towards their attack patterns of choice, we used the K-means clustering algorithm on our three variables. Figure 6 displays the silhouette score for models with different k (number of clusters). Our analysis identified one optimal model based on silhouette score: with k=8 clusters (Silhouette = 0.569). When selecting a model, it is important to consider not only its performance in terms of accuracy but also its interpretability (i.e., to what extent the clusters make sense). In our case, the model with k=8 clusters was the one with the highest silhouette score, and was the best in terms of interpretability. Thus, the model with k=8 clusters emerged as the optimal choice.

Figure 6.

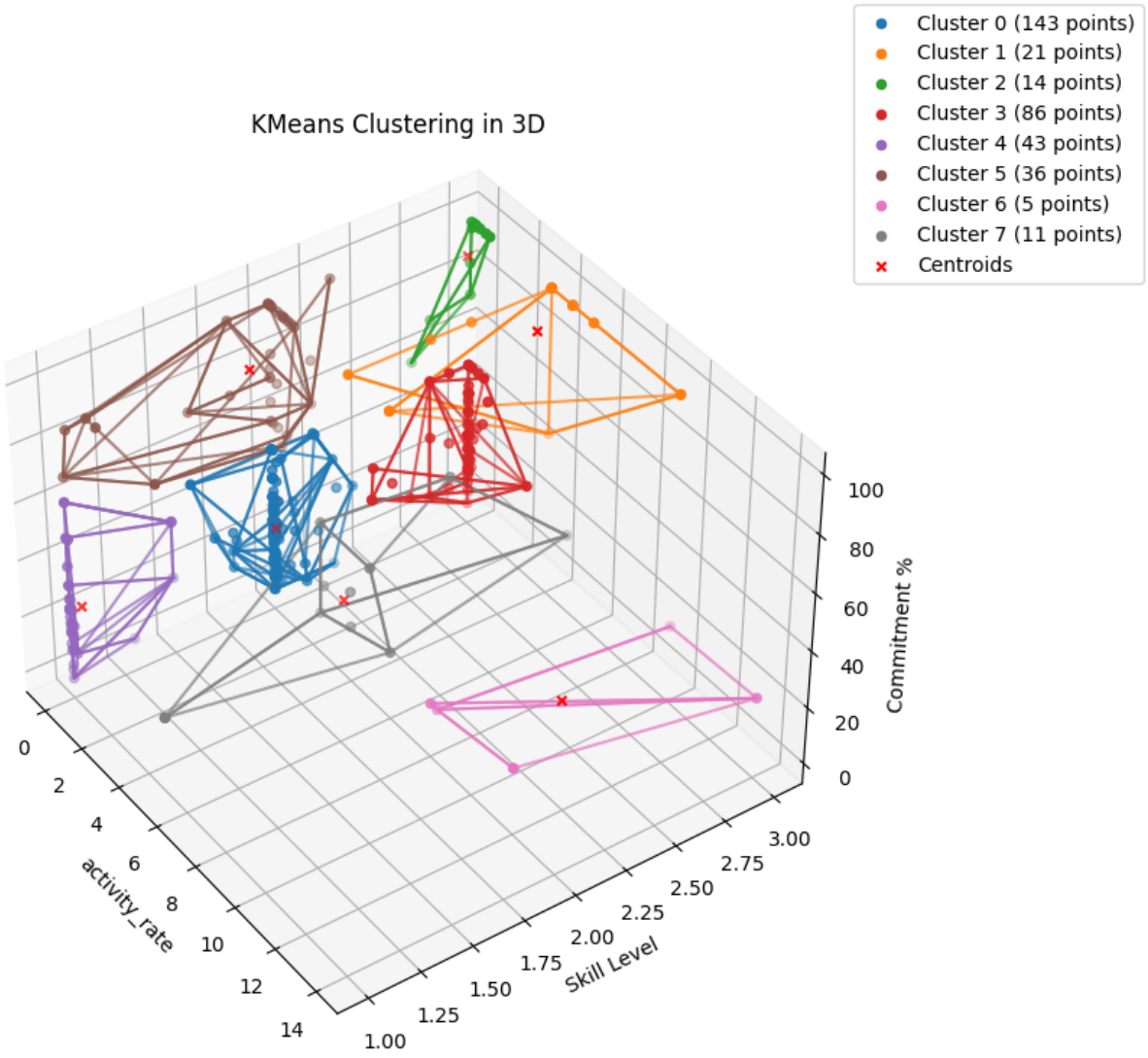
Silhouette Score for models with different k values



The partition of the chosen model is shown in Figure 7. It illustrates the resulting clusters from the chosen model, with each cluster having its assigned color. Each cluster is plotted on a 3-dimensional space, with activity rate on the x-axis, skill level on the y-axis and commitment percentage on the z-axis.

Figure 7.

Partition of the model with k=8 clusters



Bouchard and Nguyen's framework is based on two variables, namely skill level and commitment. Thus, the clusters were interpreted into technical expertise levels according to their centroids' characteristics on both skill level and commitment. The population we aim to identify fits in the 'Professional' class, the highest level of technical expertise: actors scoring high on both ends.

Considering forums' population is quite ephemeral (Hughes & Hutchings, 2023) and that key actors are amongst the most active and senior actors (Benjamin and Chen, 2012; Abbasi and al.,

2014; Zhang & al., 2015; Samtani and Chen, 2016; Samtani & al., 2017; Grisham and al., 2017; Johnsen and Franke, 2020; Huang and al., 2021), and as mentioned above, activity rate was added as the third variable of this analysis. Thus, activity rate comes as an extension to Bouchard and Nguyen's framework to nuance the expertise profiles based on seniority and diligence of posting activity. **Those fitting in the 'Professional' class and striking the right balance between seniority and diligence are thus considered key actors in our framework**.

Table 12 provides an overview of the clusters. The table presents each cluster with its interpretation according to the Bouchard and Nguyen's framework, as well as each cluster's centroid, population, and population percentage.

Table 12.

Clusters Overview

Cluster	Bouchard & Nguyen (2011) framework	Centroid [Skill; Commitment; Activity]	Nb of actors	% of sample population
0	Amateurs	[2.00; 22.47; 0.11] [Mid; Low; Discrete]	143	39.83
1	Likely Professionals	[2.81; 97.62; 5.14] [High; High; Hyperactive]	21	5.85
2	Professionals	[2.96; 90.37; 0.28] [High; High; Active]	14	3.90
3	Pro-Amateurs	[2.96; 25.32; 0.12] [High; Low; Discrete]	86	23.96
4	Amateurs	[1.05; 24.32; 0.05] [Low; Low; Discrete]	43	11.98
5	Average Carreer Criminals	[1.86; 84.81; 0.50] [Low; High; Active]	36	10.02
6	Pro-Amateurs	[2.38; 18.46; 10.67] [Mid; Low; Hyperactive]	5	1.39
7	Amateurs	[1.95; 24.51; 4.14] [Mid; Low; Hyperactive]	11	3.06

Professionals

Professionals are actors scoring high on both skill level and commitment percentages. Professionals represent the highest level of technical expertise in our framework. Overall, the 'Professionals' consists of 2 clusters and accounts for less than 10% of our sample (9.75%). Clusters 1 and 2 are of the main interest in our study as they each fit the 'Professionals' class in our criminological framework. They each contain a relevant population for cyber threat intelligence production nuanced by their seniority and diligence in their posting activity.

Cluster 1 gathers the short-lived professionals. Indeed, the population of this cluster has one of the top activity rates and scores among the highest for both skill level and commitment percentage as can be seen with its centroid: [2.81; 97.62; 5.14]. These actors possess top-tier skills and are devoted to their attack pattern of interest, meeting both criteria for 'Professionals' in the framework of Bouchard and Nguyen (2011). On top of that, they are very active. With an activity rate of 5.14 on average, this population posts 5 times a day. The reason behind such a high activity rate is their limited period of activity. All actors in this population have only been active for a day in our sample. With an average of 5 specialized posts a day, despite being new, they have quickly established themselves as active contributors in a very short period of time. However, considering their short activity time, the label **'Likely Professionals'** seems more suited for this population as we don't have enough data to fully consider them 'Professionals' yet. Indeed, despite showing the highest level of technical expertise from our framework and a very active contribution, they fail to check the box for seniority. Thus, this cluster is the home of new elite and very active actors we should keep an eye on, **prospective key actors** in a sense. Monitoring this cluster can provide insights into the behavior of new and highly active members within the community. They account for 5.85% (21 out of 359) of our sample.

Cluster 2 is the home of a more senior elite in our sample. With a centroid characterized by a skill level of 2.96 and a high commitment level of 90.37%, these actors exhibit Top-tier skills and a strong dedication to their community of interest. Thus, Cluster 2's population also meets both criteria to be considered **'Professionals'** according to Bouchard and Nguyen's framework.

However, they score low on the activity rate scale (activity rate = 0.28). The reason behind this is that they are older or even senior members, with a longer period of activity, making their activity rate plummet. With an average period of activity of 159 days and average posting rate of one post every three to four days (average activity rate = 0.28), the senior elite strikes the right balance between seniority and activity rate to be considered key actors: they exhibit the highest level of technical expertise in our framework and are among the most senior and diligent members of their community. They represent 3.90% (14 out of 359) of the sample.

Pro-Amateurs

Pro-Amateurs are actors scoring high on the skill level but relatively low on the commitment scale. Pro-Amateurs represent the second level of technical expertise, just below Professionals. The Pro-Amateurs of our sample account for 25.35% of our sample and consist of Cluster 3 and Cluster 6.

Cluster 3 represents a more discrete population of highly skilled seniors. This population is characterized by a centroid with the highest skill level (2.96) but a commitment percentage barely above 25% and one of the lowest activity rates (0.12). Its centroid allows to paint the portrait of those inside the cluster: actors with top-tier skills with a tendency to explore various attack patterns rather than focusing on one. With such scores, this cluster's population falls under the **'Pro-Amateurs'** category of Bouchard and Nguyen's framework. Their low activity rate is also due to their long period of activity; on average they were active for 488 days with some having a track record of more than 2,500 days. However, despite their seniority, these actors don't contribute frequently. With an average activity rate of 0.12, these actors tend to share specialized content intermittently. They gather 23.96% of the sample. Given their significant presence/experience and their top skills, these actors are worth monitoring. Nevertheless, this cluster proves less useful as a profile since its population isn't committed to a single attack pattern. Therefore, while it merits some attention, its monitoring could be less beneficial than the monitoring of cluster 1 and 2.

Cluster 6 gathers a more short-lived hyperactive population in our sample. The population of this cluster has a mid to top-tier skill level (2.38), meaning they oscillate between medium and high skilled CAPECs, and shows the lowest commitment percentage (18.46%) of all clusters. With such characteristics, this population fits the **'Pro-Amateurs'** class. Moreover, they have the highest activity rate by far (10.67) meaning they post, on average, 10 times a day. However, all actors in this cluster have been active only for a day. Despite their new presence, they have actively contributed with specialized content. This cluster consist of only 5 actors (i.e. 1.39% of our sample).

Average Career Criminals

Average Career Criminals are actors scoring low on skill level but high in terms of commitment. Average Career Criminals represent the second lowest level of technical expertise in our framework. A single cluster fits this class: Cluster 5.

Cluster 5 musters the average career criminals of our sample. Its centroid indicates a low to mid-tier skill level (1.86), a high commitment (84.81%) as well as a relatively high activity rate (0.5). Despite their skill level neighbouring the mid-tier, these actors exhibit a high commitment to their attack pattern of interest, checking the boxes to be considered **'Average Career Criminal'** according to Bouchard and Nguyen's classification. Average Career Criminals constitute just over 10% of our sample (10.02%).

Amateurs

Amateurs is the last class of our framework. Amateurs are those who score low on both scales and represent the lowest level of technical expertise. The amateur population is the largest of our sample, with more than half (54.87%) of our sample, amateurs are scattered in three clusters: Cluster 0, Cluster 4 and Cluster 7.

Cluster 0 corresponds to a discrete amateur population in our sample. Its centroid presents a skill level of 2, a commitment of 22% and an activity rate just above 0.10. These actors have mid-tier

skills, can't seem to settle for a single attack pattern of interest, hence their low commitment rate. These characteristics make them very suited for the **'Amateurs'** class. Finally, they don't offer a very active specialized contribution. Discrete amateurs are the most numerous out of our 8 clusters, accounting for a little less than 40% of our sample (143/359 = 39.83%).

Cluster 4 represents the lowest skilled discrete amateurs. Characterized by a centroid with the lowest skill level (1.05) and activity rate (0.05) among all clusters and a commitment percentage of 24.32%, these actors possess the lowest skills and display a curiosity to explore various attack patterns, resulting in a low commitment level. This population is also the least active of our clusters. The characteristics makes them fit the **'Amateurs'** category of our criminological framework perfectly and gathers 11.98% of our sample.

Cluster 7 represents the hyperactive amateur population of our sample. This population's centroid has a medium skill level (1.95) as well as a low commitment (24.51%). Nevertheless, cluster 7 exhibits a high activity rate with an average of 4.14. Their characteristics makes them suitable to be hyperactive **'Amateurs'** in our criminological framework. Hyperactive amateurs constitute 3.06% of all actors.

Statistical differences between clusters

To determine whether the differences between clusters were significant, we computed statistical tests. First, the Kruskall-Wallis test were significant and showed a large effect for all three variables (Skill level- H=309.61, p=0.000; $\eta 2$=0.86; Commitment - H=173.82, p=0.000; $\eta 2$=0.47; Activity rate - H=117.75, p=0.000; $\eta 2$=0.31). Post hoc tests suggest that professional clusters (C1 (likely professional) and C2 (professionals)) with key actors showed significant differences on all three variables with almost all clusters with moderate to large effects size. The statistical test results, box plots and pairwise post hoc tests as well as effect sizes are available in Annex C.

Peering into the Actor Kaleidoscope: Professionals Distribution within Community of Interest

Having presented the distinct clusters based on their technical expertise level from Bouchard and Nguyen's (2011) framework within our sample, we now turn our focus to the distribution of professionals within Communities of Interest (CoI) to offer insights into the distribution of expertise levels among CoI. Figure 8 depicts the distribution of clusters in our communities of interest, where professionals are in black and likely professionals are in red.

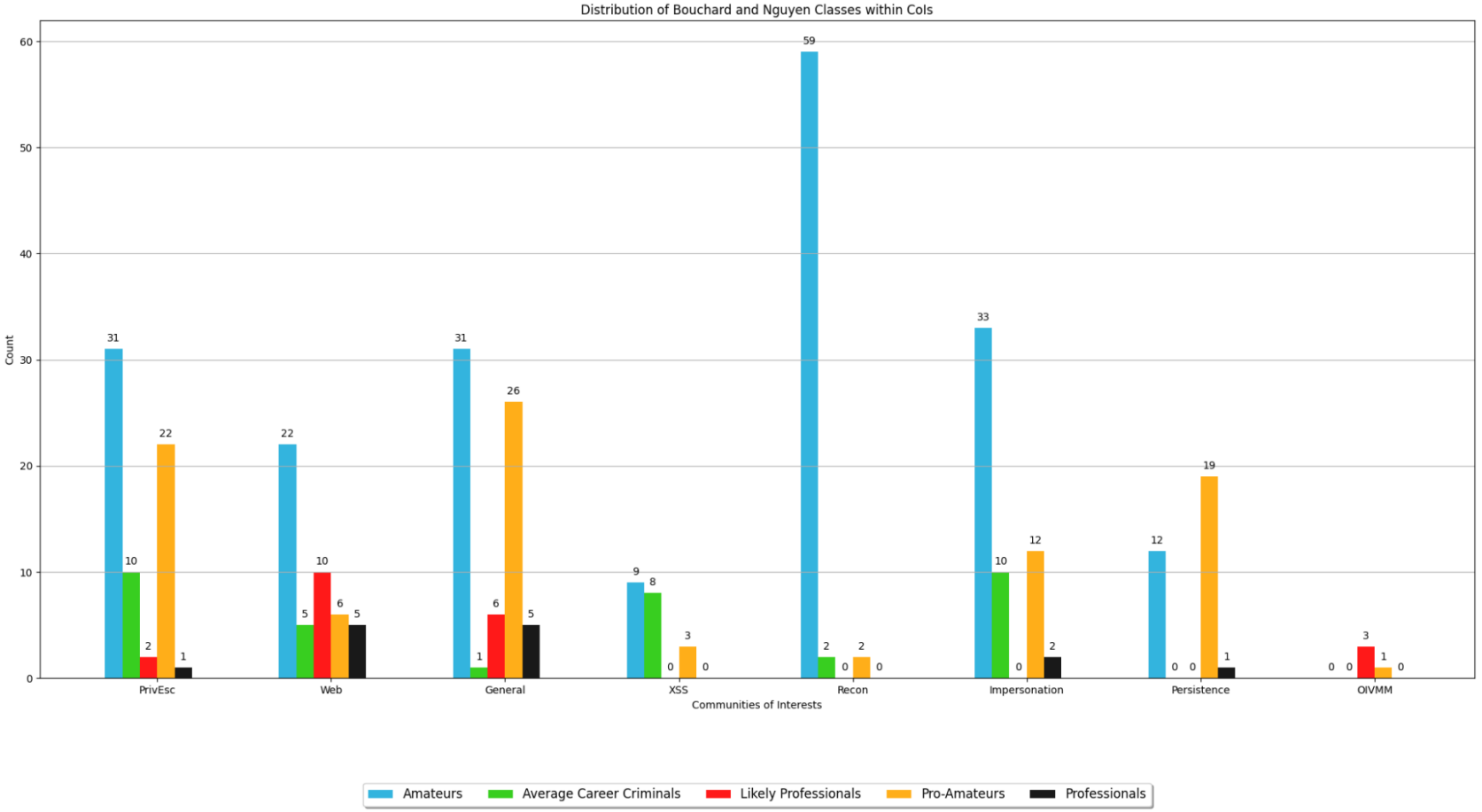
Likely Professionals (or prospective key actors) are present in four CoIs out of eight: PrivEsc, General, Web and OIVMM. Professionals (or key actors) are present in five out of eight CoIs: PrivEsc, Web, General, Impersonation and Persistence. Interestingly, the PrivEsc, Web and General CoIs contain both professionals and likely professional actors. This suggests that some communities attract both seasoned and new elite prospects.

Overall, Professionals are but a few in the population of our sample and are unevenly distributed among CoIs. Considering the population included in the key actor identification analysis was a small proportion of our final dataset (15.47%), the proportion of key actors becomes even smaller when put into perspective with their CoI. Key actors represent a very small proportion of the CoI they are part of, suggesting the presence of technical expertise in CoIs but in very little quantity.

Finally, only two communities don't have any Professionals or Likely Professionals: XSS and Recon. Amateurs and Pro-Amateurs represent the majority of our CoIs' population, followed by the Average Career Criminals. The Recon CoI stands out as it is mostly composed of Amateurs, while other CoIs, apart from the OIVMM CoI, have a more well-balanced population between these three categories

Figure 8.

Distribution of Clusters within Community of Interest (CoI)



Discussion

To address the gap in criminological interest towards the identification and study of key actors, **this study identified areas of technical expertise in attack patterns in cybercrime forums and their related key expert actors based on three facets: skill level, commitment, and activity.** Our research yielded four key findings: 1) The identified attack patterns act as catalysts of cybercrime communities. 2) A criminological take on operationalizing expertise level brings new insight for the study of cybercrime. 3) Expertise-based profiles of key actors allow for more targeted intelligence. And lastly, 4) key actors represent a promising scarcity for resources allocation in the production of cyber threat intelligence.

Attack patterns as Catalysts of Cybercrime Communities of Shared Interest

By analyzing CVE/CAPEC discussions, we successfully identified distinct communities based on shared interests towards attack patterns, unveiling the landscape of potential areas of technical expertise. The Leiden community detection algorithm allows to observe communities underlying the network but is based in the structural properties of the network. While Leiden is a powerful tool for identifying communities based on mathematical criteria (Traag & al., 2019; Anuar & al., 2021), it does not guarantee that these communities will be meaningful or relevant to the research question at hand.

To map the CAPEC-actor network, we first relied on CVE mentions made by actors, before relying on the association between CVEs and CAPECs provided my MITRE. By using CVEs as a proxy for CAPECs, we relied on an external entity to do the association between CAPEC and CVE and we therefore relied on their precision in doing so. Thus, the possibility of information loss during the CVE-CAPEC association process is a consideration not to be dismissed for the findings of this study.

With this consideration in mind, an interesting aspect of our analysis is that the structural communities identified by the Leiden algorithm transcend the groupings based on network connections. As stated before, our network is based on CVE mentions as proxy to CAPECs. Then, it is possible that our communities were simply a collection of actors discussing the same CVEs proxying CAPECs; grouping CAPECs that lack coherence when discussed together, leaving us puzzled as to why they are being discussed together.

However, our content analysis revealed that these communities not only mirror network connections (the mentions of similar CVEs amongst actors) but also reflect shared interests in specific attack patterns. This suggests that the structural communities are not solely a product of the network topology (random CVEs being mentioned together) but also reflect a shared cybercrime interest in specific attack patterns. The Leiden algorithm's ability to identify communities centered on specific attack patterns suggests that these communities reflect specialized discussions about specific attack patterns. This finding is particularly significant, as it suggests that despite the potential lack of precision in the association between CVE and CAPECs, the communities identified by the Leiden algorithm still align with real-world and coherent interests in specific attack patterns.

Internet, and forums have been known to be gathering places for cybercriminal activities (Goldsmith and Brewer, 2015; Nunes & al., 2016; Shakarian, Gunn, and Shakarian, 2016; Leukfeldt, Kleemans and Stol, 2017; Biswas & al., 2022). While this research uses several forums' data, our communities are based on the area of technical expertise, i.e. the type of attack pattern of interest, regardless of the forum actors were in. Obtaining communities reflecting shared cybercrime interest suggests that actors do in fact gather around shared interest towards a type of attack pattern; interest which transcends the forum barrier. The communities identified in this study go beyond forum boundaries, revealing another dimension/aspect to cyber threat intelligence. This finding is significant because it opens a new perspective on studying cybercrime activity and identifying key actors: their area of technical expertise, i.e. their interest towards a type of attack pattern.

A Criminological Take on Operationalizing Expertise Level

We operationalized the level of technical expertise following a criminological framework and previous literature. Adopting a hybrid approach using network analysis, content analysis as well as seniority and posting behavior, we developed three metrics to identify key actors based on their technical expertise level towards their attack patterns of choice.

This study differs from previous literature in the use of a skill level metric as one of the foundations for key actor identification. While research on key actors within hacker forums has studied them from various angles (Holt and Kilger, 2008; Fang & al., 2016; Benjamin and Chen, 2012; Abbasi & al., 2014; Zhang & al., 2015; Fang & al., 2016; Samtani and Chen, 2016; Grisham & al., 2017; Samtani & al., 2017; Marin & al., 2018; Johnsen and Franke, 2020; Huang & al., 2021), none had thoroughly focused on their skill level. This study proposes a framework considering an objective assessment of theoretical skill level, based on CVEs and their corresponding CAPECs. In doing so, it contributes to the creation of a more complete framework for key actor identification.

Next, the operationalization of commitment in our study comes with some considerations. The commitment metric we developed captures an actor's engagement towards its attack pattern of interest. While this metric is complementary to the activity rate, it shapes the profiles of the actors we consider key. Key actors are those posting consistently about a single type of attack requiring a relatively high skill level: elite specialists or in other terms, experts in a single domain.

Finally, the main contributions of this study are the adaptation of the framework to cybercrime as well as the addition of a third dimension (activity rate) to Bouchard and nguyen's (2011) framework. This study proposed an adaptation of the framework to cybercrime by leveraging CVE and CAPECS to measure the core concepts of expertise according to Bouchard and Nguyen's (2011) framework: skill level and commitment. In this adaptation, echoing previous research and the relative approach on expertise, activity rate was added as the third dimension of our expertise framework. Activity rate, considering both seniority and posting activity, allowed us to differentiate between diligent experts (Professionals) and short-lived experts (Likely Professionals) based on their diligence and time committed to their cybercriminal activities.

Aligning with previous literature (Benjamin and Chen, 2012; Abbasi and al., 2014; Zhang & al., 2015; Samtani and Chen, 2016; Samtani & al., 2017; Grisham and al., 2017; Johnsen and Franke, 2020; Huang and al., 2021), this study suggests that the activity rate should also be considered in a framework studying cybercriminals. In the case of cybercrime and even criminology, one's expertise in a specific area should then be nuanced by the diligence of one's cybercriminal activities overtime (activity rate).

Expertise-based Profiles of Key Actors for Targeted Intelligence

To contribute to the key hacker identification problem, this study identified key actors based on their level of technical expertise in their attack patterns of choice. In our study, key actors exhibit a high level of technical expertise, scoring high on both skill level and commitment, and they strike the right balance in activity rate between seniority and diligence in their posting activity.

Aligning with previous literature, key actors (Professionals, not to be confused with Likely Professionals) identified in this study are active and senior actors (Benjamin and Chen, 2012; Abbasi and al., 2014; Zhang & al., 2015; Samtani and Chen, 2016; Samtani & al., 2017; Grisham and al., 2017; Johnsen and Franke, 2020; Huang and al., 2021) who are specialized in their discussions on attack patterns (Abbasi & al., 2014; Fang & al., 2016) as shown by their high commitment percentage.

Key actors in this study differ from previous criminological profiles. Previous literature held behavioral tendencies towards knowledge as their primary focus (Holt and Kilger, 2008; Zang & al., 2015) while this study focused on a two facets assessment of technical expertise level.

Hold and Kilger's (2008) dichotomous vision proposed two profiles based on the behavior towards products and knowledge: the makecraft is a creator and the techcraft is more of a consumer. While Holt and Kilger focused on behavior towards knowledge based on their discussion content, our focus on two variables to identify actors exhibiting the highest technical expertise level contrasts. Identifying actors based on their technical expertise level allowed us to

differentiate them into four different expertise profiles, using Bouchard and Nguyen's (2011) framework. Moreover, the addition of the activity rate added more depth to our profiles based on actor's seniority and diligence in posting activity, ultimately identifying key expert actors.

Zang and colleagues' (2015) Guru hackers were considered key in their study. Professionals (key actors) in our study resembles the Gurus in their seniority and high activity (Zhang & al., 2015). However, they can't be compared on other aspects, as the authors used qualitative analysis of post content as well as their interactions with actors to define Gurus; both aspects being absent in this study. Nevertheless, Zang and colleagues' four profiles have some similarities with our criminological profiles in their hierarchy about activity and knowledge. Gurus are the most knowledgeable and active senior just like our Professionals have the highest skill level and are diligent in their posting activity. On the opposite side of the scale, Zang and colleagues' Novices are the least knowledgeable and have ephemeral presence just like our Amateurs have the lowest skill level and are very discrete in their posting activity.

Overall, key actors identified in this study have more dimensions (skill level, commitment, and activity) and depth compared to previous criminological profiles. They allow for the classification of all populations based on their level of technical expertise as well as the differentiation of actors based on their seniority and diligence in posting activity. Instead of a qualitative analysis of the behavioral tendencies towards knowledge, our framework focuses on actors' technical expertise towards their attack patterns of choice. The resulting key actors are experts (and likely experts) in a certain attack pattern, making them more actionable for intelligence production and allowing for a more targeted approach. Nevertheless, the experts in this study lack a qualitative analysis of their posts' content such as the one done in Holt and Kilger (2008) and Zang and colleagues (2015). Future research could incorporate qualitative analysis of actors' posts' content to the expertise profile provided in our framework to account for the unspecialized content of those actors. The study of how key actors became key and if they will maintain this status overtime represents an interesting avenue for future research as well.

Key Actors: A Promising Scarcity for Resources Allocation in the Production of Cyber Threat Intelligence.

Key actors, or Professionals, represent less than one tenth (9.75%) of the sample and the distribution of these experts is uneven across CoI. This finding suggests the presence of technical expertise in specific attack patterns; however, this technical expertise is very rare and represents a very small fraction of the population interested in this area. Some CoI don't even have experts and are instead the home of the Amateur population like the Recon CoI.

One important thing to note is the actors making it to the sample for the detection of key actors account for a small fraction of the overall population in our final dataset. Indeed, 84.53% (1,962/2,321=0.845) of our final dataset didn't meet the criteria to be included in the sample for final analysis. Considering that our key actors represent less than 10% of our sample and the sample itself accounts for 15% of our final dataset, then the real proportion of key actors represents 1.5% of our final dataset (9.75%*15.47%=1.50%). This finding aligns with previous literature on the key hacker identification problem stating that key actors make up only a small proportion of their platform, the rest being unskilled or just curious (Marin & al., 2018, b).

The detection process in this study allowed to reduce the population of interest for intelligence production to just a small fraction of the final dataset. This finding is significant and represents one of the main contributions of this study to the key hacker identification problem and is encouraging for the allocation of resources in cyber threat intelligence production. Identifying an expert population of key threat actors comprising 1.5% of the initial population could reshape resource allocation in cyber threat intelligence production, streamlining efforts for a more effective intelligence. Reducing the population of interest for intelligence production could allow for a better allocation of the resources poured into the process and thus produce a more effective intelligence.

Limits & Future Works

Despite the strengths of this study, several limitations warrant consideration. First, we rely on CVE mentions for our data collection causing the loss of potential key posts and thus key actors not mentioning any CVEs. To account for the unspecialized content of actors, future research could collect actors' non-specialized posts on top of those mentioning a CVE to study actors' whole contribution in their community.

Following this limit, we rely on CVEs as proxy for CAPECs. We follow MITRE's CVE-CAPEC mapping and rely on its precision in the association between CAPEC and CVE. While following MITRE's mapping brings credibility to the mapping used in this study, the potential loss of information in the association CVE-CAPEC is a limit of this study. Next, this research uses the skill level required metric provided by MITRE. We then rely on MITRE, once again, as the basis for our skill level metric computation. However, MITRE's precise computation process for the skill level required metric isn't publicly available, making us rely on what could be considered a 'black box'.

Another limit to this work comes from the data sources. First, all forums, regardless of their quality, were ingested following the same criteria, even though the quality of the discussions that take place there can vary greatly. Future research could be more selective in their sources and choose to only include high quality forums in their dataset. Secondly, this work focuses on the concept of communities. However, some forums in our dataset have been considered as markets in previous literature. As the structure of discussions and relationships in a market differs from that in a discussion forum, the processing of our data, regardless of its source, presents a limitation to this work. Future works could study the differences in emerging profiles depending on their source forum by studying markets and discussion forums separately.

Several communities have a significant proportion of actors who have posted only once, « one-timers ». However, it is important to note that this study does not consider posts without CVE mentions; it is possible that an actor has continued to post in the topic of interest without mentioning CVEs. It is also possible that an actor posts only once but continued to read discussions on the topic, just as he might have posted once and then lost interest. With that being

said, the choice of including actors with a single so-called technical post (i.e. post mentioning a CVE) in our communities remains a limitation.

The imputation of skill level required values to valueless CAPECs also constitutes a limit of this study. Despite the use of the hierarchical structure of the CAPEC framework for imputation, both approaches (using child or parent skill level) involve some level of estimation. The true skill level for a CAPEC might not perfectly match either the child or parent's skill level. Moreover, each CAPEC was then assigned its highest scenario's value. This way, overestimating CAPECs skill level required, thus impacting actor's skill level, is a possibility and a limit of this study. Exploring alternative sources or methodologies for assessing the skill level required beyond MITRE's metrics also represents an interesting avenue for future research. This could involve collaborating with industry experts to develop more transparent and publicly available computation processes for skill level metrics.

Next, the operationalization of skill level and commitment metrics both have their limits. The assumption behind the operationalization of skill level, namely that an actor mentioning a CVE has the theoretical skills to exploit the related CAPECs, represents a limit. An actor could be asking a question about the CVE and thus not have the skill level required to exploit the vulnerability and its CAPEC, at least for now. An actor asking a question about a CVE may indeed want to acquire knowledge, but this does not necessarily indicate a low skill level. By asking a question about a particular CVE, actors show that they have the skills to take an interest in that CVE, and by obtaining an answer, they could increase their skills. The question could therefore be a precursor to upgrading one's skills.

This work proposed an objective assessment of a skill level in the sense that it avoids any human biases behind the qualitative analysis of actors' content. The skill level assessed is theoretical because it is impossible de check if actors actually have their assigned skill level in real life. Both metrics (skill level and commitment) have a threshold specifically suited for our sample and its particular distribution, making those thresholds hardly suitable or not re-usable with a different sample. Future research could focus on the elaboration of a skill level and commitment metrics that would be suitable for all analysis.

The adaptation of the criminological framework has its own considerations. Since we couldn't interview each actor to measure their expertise directly, our approach assessed technical expertise

through proxy variables. As a result, the expertise measured is a theoretical and hypothetical, technical expertise. Additionally, it is even more challenging to determine whether this expertise is applied for criminal or legitimate purposes, making the expertise measured in this study neither a part of criminal or legitimate expertise. All three dimensions of expertise are dependant on the recorded activities in our data. Quantifying the activities of said individual is limited to the traces they leave on forums, making the measured activity heavily dependent on forum presence. Thus, the expertise measured in this work does not equate to a real holistic expertise, but is a measure of a theoretical technical expertise based on the recorded and available activity of a said individual. The expertise measured in this work is also an expertise at a given time, and therefore static expertise. It would be interesting to study the evolution of expertise over time for the studied actors to understand how one's expertise changes over time.

The measurement of expertise (even if it is a theoretical expertise) used in this work is limited. Indeed, without having the entire content of an actor, this measurement is inherently limited. Therefore, the term « contribution » which would denote the technical input of each actor might be more suitable to the context since we can't verify if actors actually have the level of expertise they were assigned based on their technical posts. The term « contribution » would reflect less on theoretical (and potentially absent) skills and align more closely with what is directly measured. Nonetheless, despite being a limitation, the decision to retain the term « expertise » was made to align as closely as possible with the theoretical framework of this work.

The possibility of identifying cybersecurity analysts, forum administrators, or law enforcement investigators as one of the (Likely) Professionals is quite present. Even though they aren't very active, identifying Likely Professionals helps to pinpoint a population worth monitoring despite their low activity levels because they demonstrate high level characteristics. The aim of this research is not to differentiate the roles of various participants, but to identify which participants might be interesting from a cyber threat intelligence perspective. This population includes new individuals on the forum as well as those who may have posted only once and then remained inactive, such as potential law enforcement investigators, for example. Distinguishing between undercover investigators and actual malicious actors is too complex to achieve with the quantitative method proposed in this work. Consequently, this research would identify a cybersecurity professional in the same terms as a malicious actor if both demonstrated the same

characteristics (i.e. high skill and significant commitment). A subsequent, more qualitative analysis would be necessary to distinguish between the two. However, this research does not claim to perform this subsequent analysis. Future research could dive deeper into the full content of those identified as Professionals and Likely Professionals to understand their role and study more closely this population identified as key.

This study proposes an identification of key actors based on their technical expertise; however, it does not have any ground truth regarding if those identified actors are really key in their communities or not. The study doesn't provide any objective metric vouching for identified key actors. Future research could study if there are overlap in the actors identified as key by various methods and study how key actors according to our framework are positioned in the broader cybercrime ecosystem compared to key actors identified in other studies.

Our framework differs from previous literature in the variables and metrics used to identify key actors. First, our data only had actors' specialized posts without following up on the replies or the evolution of any discussion thread. Due to the lack of data on interactions between actors, our analysis did not consider centrality measures which is a limit of this study. Second, the full focus on CVE mentions and the absence of consideration of post's content also differentiate this study from existing literature on key hacker identification. The data used in this study only consists of specialized posts, mentioning at least a CVE, meaning we didn't have the full extend of an actor's real activity on a forum. The posts included in this analysis are then decontextualized, meaning they were taken out of their respective discussion thread, leading to a loss of context information in the process. Therefore, our metrics and key actor identification are based solely on specialized and technical posting activity. As our analysis does not consider post content outside of CVE mentions, it is highly possible that there are key actors out there posting valuable content without CVE mentions. In those cases, our framework would fail to recognize those actors as key due to their lack of CVE mentions.

The limited proportion of the population used for key actor analysis also warrants consideration. To ensure consistency in our metrics, we only included actors who posted a minimum of four times and who referenced at least four CAPECs. This criterion was met by just 15% of our initial population, resulting in the exclusion of more discreet actors and those who referenced few or no CVEs in their content. Consequently, our analysis focuses on a small subset of active actors who

reference CVEs in their posts, representing only a small fraction of the overall active population within the source forums.

The choice to include activity rate as the third dimension to our framework can also discriminate against more discrete actors. It is possible that some relevant actors just don't post a lot but are still very relevant to cyber threat intelligence. Again, in those cases, our framework would fail to recognize those actors as key due to the lack of posts, they would fall under the « likely professional » umbrella.

However, using posts that mention CVEs guarantees that the post is talking about at least a direct vulnerability. Although we do miss posts that do not reference a CVE, using posts mentioning CVEs gives us a more solid collection of posts which should include less noise or trivial discussions. This focus on CVEs not only allows for a fast processing of the metrics, but also allows our metrics to be less subjective to interpretation of posts' content, making the whole process more scalable and less subject to human biases. Future research could collect all posts for each actor, regardless of CVE mentions and perform a qualitative analysis on posts' content to complement the profile identified within our framework.

Finally, this study aligns with previous literature in the use of a hybrid approach. Leveraging a bimodal network, CVE mentions, seniority and posting behaviors our framework offers a nuanced understanding of technical expertise based on skill level, commitment dynamics and activity within each area of technical expertise. Our framework is then complementary to those previously stated in the literature and should be used in addition/parallel to some interactions and social network analysis as well as posts' content analysis. The combination of existing methods with our framework in the identification of key actors could allow the identification of actors with a more complete threat profile. Combining existing computer-science-based methods and content analysis methods with our framework to identify key actors having a more complete threat profile appear as the most relevant avenue for future research.

Conclusion

Previous literature, encompassing methods for identifying key actors in underground communities (Benjamin and Chen, 2012; Abbasi & al., 2014; Zhang & al., 2015; Fang & al., 2016; Samtani and Chen, 2016; Grisham & al., 2017; Samtani & al., 2017; Johnsen and Franke, 2020; Huang & al., 2021) have predominantly focused on algorithmic approaches for the production of intelligence; overlooking the criminological study of the actors identified as key by their algorithms. Despite various other studies adopting different angles to profile actors on underground forums (Holt and Kilger, 2008; Abbasi & al., 2014; Zhang & al., 2015; Fang & al., 2016; Marin & al., 2018, b; Huang & al., 2021), none have extensively examined the expertise levels of these key actors.

Expertise and criminal success are closely bound, those exhibiting expertise are more successful in their activities, making them more dangerous (Bartol and Bartol, 2014). It then becomes valuable to prioritize threats coming from these experts actors for cyber threat intelligence since they are more likely to be met with success (Motoyama & al., 2011; Bartol and Bartol, 2014; Marin & al., 2018). To contribute to the identification of key actors and address a gap in criminological interest in the study of key actors within hacking forums, the present study has drawn inspiration from the profiles created by Bouchard and Nguyen (2011) to identify key actors based on their technical expertise.

To do so, we first identified areas of technical expertise in the form of communities of interest towards attack patterns. Leveraging CVE mentions from actors' posts and their corresponding CAPECs, we built a bimodal network before using the Leiden algorithm to identify communities of interest. Then, using the k-means algorithm, we detected key actors based on their technical expertise level in their area through two facets: skill level and commitment towards the area of technical expertise. And activity rate allowed to nuance the expertise profiles based on their seniority and posting activity. Resulting clusters were then interpreted at the light of Bouchard and Nguyen's (2011) framework.

This study yielded four key findings. First, attack patterns act as catalysts of cybercrime communities of shared interest. The communities identified in this study transcend forum

boundaries and opens a new perspective in the study of cybercrime and key actors: their area of technical expertise in a type of attack patterns.

Second, we adopted a criminological take on operationalizing expertise level, bringing new insights for the study of cybercrime and criminology as a whole. The use of two facets: an objective skill level assessment and commitment in the area of technical expertise, as well as the addition of activity rate in our framework to nuance our expertise profiles contribute to the creation of a more complete framework for the study of cybercriminals.

Third, the expertise-based profiles created in this study allow for a more targeted intelligence. Our framework allows for the classification of all populations based on their level of technical expertise and differentiate between actors based on their seniority and diligence. The focus on technical expertise results in more complete profiles of experts in a specific type of attack patterns, making them more actionable for intelligence production and cyberthreat prevention.

And four, key actors represent a promising scarcity for resources allocation in production of cyber threat intelligence. The detection of relevant experts for credible intelligence representing just a fraction of the overall population of internet users could reshape resource allocation and allow for a more effective intelligence.

Finally, this study has employed a hybrid framework to develop threat actor profiles based on a criminological take of expertise. The integration of criminological theoretical foundations with previous literature's hybrid approach in the analysis of threat actor communities has enabled a new understanding of threat actor populations and their technical expertise. In doing so, this study contributes to addressing the gap of criminological interest in the identification and study of key actors based on expertise.

Aknowledgements

This research was conducted as part of a MITACS research internship in collaboration with Secureworks. Secureworks is an American cybersecurity company that delivers threat detection, investigation, and response services to organizations of all sizes. Secureworks provided help with the data collection as well as with the analysis.

References

Abbasi, A., Li, W., Benjamin, V., Hu, S., & Chen, H. (2014). Descriptive Analytics : Examining Expert Hackers in Web Forums. *2014 IEEE Joint Intelligence and Security Informatics Conference*, 56-63. https://doi.org/10.1109/JISIC.2014.18

Allodi, L. (2017). Economic Factors of Vulnerability Trade and Exploitation. *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, 1483-1499. https://doi.org/10.1145/3133956.3133960

Almukaynizi, M., Nunes, E., Dharaiya, K., Senguttuvan, M., Shakarian, J., & Shakarian, P. (2017). Proactive identification of exploits in the wild through vulnerability mentions online. *2017 International Conference on Cyber Conflict (CyCon U.S.)*, 82-88. https://doi.org/10.1109/CYCONUS.2017.8167501

Analyse du trafic, classement et audience exploit.in [mars 2024]. (s. d.). Similarweb. Accessed april 17th, 2024, available at https://www.similarweb.com/fr/website/exploit.in/

Anuar, S. H. H., Abas, Z. A., Yunos, N. M., Zaki, N. H. M., Hashim, N. A., Mokhtar, M. F., Asmai, S. A., Abidin, Z. Z., & Nizam, A. F. (2021). Comparison between Louvain and Leiden Algorithm for Network Structure : A Review. *Journal of Physics: Conference Series*, *2129*(1), 012028. https://doi.org/10.1088/1742-6596/2129/1/012028

Bartol, C. R., & Bartol, A. M. (2014). Criminal behavior: A psychological approach (p. 672). Upper Saddle River, NJ: Pearson.

Bedi, P., & Sharma, C. (2016). Community detection in social networks. *WIREs Data Mining and Knowledge Discovery*, *6*(3), 115-135. https://doi.org/10.1002/widm.1178

Benjamin, V., & Chen, H. (2012). Securing cyberspace : Identifying key actors in hacker communities. *2012 IEEE International Conference on Intelligence and Security Informatics*, 24-29. https://doi.org/10.1109/ISI.2012.6283296

Biswas, B., Mukhopadhyay, A., Bhattacharjee, S., Kumar, A., & Delen, D. (2022). A text-mining based cyber-risk assessment and mitigation framework for critical analysis of online hacker forums. *Decision Support Systems*, *152*, 113651. https://doi.org/10.1016/j.dss.2021.113651

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, *2008*(10), P10008. https://doi.org/10.1088/1742-5468/2008/10/P10008

Borgatti, S. P., Everett, M. G., & Johnson, J. C. (2018). *Analyzing Social Networks*. SAGE.

Bouchard, M., & Nguyen, H. (2011). Professionals or Amateurs? Revisiting the Notion of Professional Crime in the Context of Cannabis Cultivation. In *World Wide Weed* (p. 109-126). Routledge.

Butler, H., & Gannon, T. A. (2015). The scripts and expertise of firesetters : A preliminary conceptualization. *Aggression and Violent Behavior*, *20*, 72-81. https://doi.org/10.1016/j.avb.2014.12.011

Calderoni, F., Brunetto, D., & Piccardi, C. (2017). Communities in criminal networks : A case study. *Social Networks*, *48*, 116-125. https://doi.org/10.1016/j.socnet.2016.08.003

CAPEC - New to CAPEC? (s. d.). Accessed april 17th, 2024, available at https://capec.mitre.org/about/new_to_capec.html

Chi, M. T. (2006). Two approaches to the study of experts' characteristics. The Cambridge handbook of expertise and expert performance, 21-30.

CVE - CVE. Accessed may 3rd, 2023, available at https://cve.mitre.org/

Décary-Hétu, D., & Dupont, B. (2012). The social network of hackers. *Global Crime*, *13*(3), 160-175. https://doi.org/10.1080/17440572.2012.702523

Décary-Hétu, D., & Dupont, B. (2013). Reputation in a dark network of online criminals. *Global Crime*, *14*. https://doi.org/10.1080/17440572.2013.801015

Dingledine, R., Mathewson, N., & Syverson, P. (2004). Tor : The Second-Generation Onion Router: *Defense Technical Information Center*. https://doi.org/10.21236/ADA465464

Dupont, B., Côté, A.-M., Savine, C., & Décary-Hétu, D. (2016). The ecology of trust among hackers. *Global Crime*, *17*, 1-23. https://doi.org/10.1080/17440572.2016.1157480

Eck, J. E., & Rossmo, D. K. (2019). The new detective. *Criminology & Public Policy*, *18*(3), 601-622. https://doi.org/10.1111/1745-9133.12450

Elo, S., & Kyngäs, H. (2008). The qualitative content analysis process. *Journal of Advanced Nursing*, *62*(1), 107-115. https://doi.org/10.1111/j.1365-2648.2007.04569.x

Ericsson, K. A. (2006, a). An introduction to Cambridge Handbook of expertise and expert performance: Its development, organization, and content. Cambridge Handbook of Expertise and Expert Performance, 1-19.

Ericsson, K. A. (2006, b). The Influence of Experience and Deliberate Practice on the Development of Superior Expert Performance. In *The Cambridge handbook of expertise and expert performance* (p. 683-703). Cambridge University Press. https://doi.org/10.1017/CBO9780511816796.038

Ericsson, K. A., Hoffman, R. R., & Kozbelt, A. (2018). *The Cambridge Handbook of Expertise and Expert Performance*. Cambridge University Press.

Expertise. (2023, september 20). https://dictionary.cambridge.org/fr/dictionnaire/anglais/expertise

Fang, Z., Zhao, X., Wei, Q., Chen, G., Zhang, Y., Xing, C., Li, W., & Chen, H. (2016). Exploring key hackers and cybersecurity threats in Chinese hacker communities. *2016 IEEE Conference on Intelligence and Security Informatics (ISI)*, 13-18. https://doi.org/10.1109/ISI.2016.7745436

Faust, K. (1997). Centrality in affiliation networks. *Social Networks*, *19*(2), 157-191. https://doi.org/10.1016/S0378-8733(96)00300-0

Finifter, M., Akhawe, D., & Wagner, D. (2013). *An Empirical Study of Vulnerability Rewards Programs*. 273-288. https://www.usenix.org/conference/usenixsecurity13/technical-sessions/presentation/finifter

Fruchterman, T. M. J., & Reingold, E. M. (1991). Graph drawing by force-directed placement. *Software: Practice and Experience*, *21*(11), 1129-1164. https://doi.org/10.1002/spe.4380211102

Geers, K. (2010). The challenge of cyber attack deterrence. *Computer Law & Security Review*, *26*(3), 298-303. https://doi.org/10.1016/j.clsr.2010.03.003

Goldsmith, A., & Brewer, R. (2015). Digital drift and the criminal interaction order. *Theoretical Criminology*, *19*(1), 112-130. https://doi.org/10.1177/1362480614538645

Grisham, J., Samtani, S., Patton, M., & Chen, H. (2017). Identifying mobile malware and key threat actors in online hacker forums for proactive cyber threat intelligence. *2017 IEEE International Conference on Intelligence and Security Informatics (ISI)*, 13-18. https://doi.org/10.1109/ISI.2017.8004867

Hobbs, D. (1997). Professional Crime : Change, Continuity and the Enduring Myth of the Underworld. *Sociology*, *31*(1), 57-72. https://doi.org/10.1177/0038038597031001005

Hoffman, R. R. (1998). How Can Expertise be Defined? Implications of Research from Cognitive Psychology. In R. Williams, W. Faulkner, & J. Fleck (Éds.), *Exploring Expertise : Issues and Perspectives* (p. 81-100). Palgrave Macmillan UK. https://doi.org/10.1007/978-1-349-13693-3_4

Hoffman, R. R., Shadbolt, N. R., Burton, A. M., & Klein, G. (1995). Eliciting Knowledge from Experts : A Methodological Analysis. *Organizational Behavior and Human Decision Processes*, *62*(2), 129-158. https://doi.org/10.1006/obhd.1995.1039

Holt, T. J. (2007). subcultural evolution? Examining the influence of on- and off-line experiences on deviant subcultures. *Deviant Behavior*, 28(2), 171-198. https://doi.org/10.1080/01639620601131065

Holt, T. J., & Kilger, M. (2008). Techcrafters and Makecrafters : A Comparison of Two Populations of Hackers. *2008 WOMBAT Workshop on Information Security Threats Data Collection and Sharing*, 67-78. https://doi.org/10.1109/WISTDCS.2008.9

Hsieh, H.-F., & Shannon, S. E. (2005). Three Approaches to Qualitative Content Analysis. *Qualitative Health Research*, *15*(9), 1277-1288. https://doi.org/10.1177/1049732305276687

Huang, H.-Y., & Bashir, M. (2016). The onion router : Understanding a privacy enhancing technology community. *Proceedings of the Association for Information Science and Technology*, 53(1), 1-10. https://doi.org/10.1002/pra2.2016.14505301034

Huang, C., Guo, Y., Guo, W., & Li, Y. (2021). HackerRank : Identifying key hackers in underground forums. *International Journal of Distributed Sensor Networks*, *17*(5), 15501477211015145. https://doi.org/10.1177/15501477211015145

Hughes, J., & Hutchings, A. (2023). Digital Drift and the Evolution of a Large Cybercrime Forum. *2023 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)*, 183-193. https://doi.org/10.1109/EuroSPW59978.2023.00026

Jain, A. K. (2010). Data clustering : 50 years beyond K-means. *Pattern Recognition Letters*, *31*(8), 651-666. https://doi.org/10.1016/j.patrec.2009.09.011

Johnsen, J. W., & Franke, K. (2020). Identifying Proficient Cybercriminals Through Text and Network Analysis. *2020 IEEE International Conference on Intelligence and Security Informatics (ISI)*, 1-7. https://doi.org/10.1109/ISI49825.2020.9280523

Krippendorff, K. (2018). *Content Analysis : An Introduction to Its Methodology*. SAGE Publications.

Kugler, R. L. (2009). Deterrence of cyber attacks. *Cyberpower and national security*, *320*, 309-340.

Leukfeldt, E., Kleemans, E. R., & Stol, W. (2017). Origin, growth and criminal capabilities of cybercriminal networks. An international empirical analysis. Crime, Law and Social Change, 67, 39-53. https://doi.org/10.1007/s10611-016-9663-1

Lusthaus, J. (2012). Trust in the world of cybercrime. *Global Crime*, *13*(2), 71-94. https://doi.org/10.1080/17440572.2012.674183

Macdonald, M., Frank, R., Mei, J., & Monk, B. (2015). Identifying Digital Threats in a Hacker Web Forum. *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015*, 926-933. https://doi.org/10.1145/2808797.2808878

Marin, E., Almukaynizi, M., Nunes, E., & Shakarian, P. (2018, a). Community Finding of Malware and Exploit Vendors on Darkweb Marketplaces. *2018 1st International Conference on Data Intelligence and Security (ICDIS)*, 81-84. https://doi.org/10.1109/ICDIS.2018.00019

Marin, E., Almukaynizi, M., & Shakarian, P. (2019). Reasoning About Future Cyber-Attacks Through Socio-Technical Hacking Information. *2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI)*, 157-164. https://doi.org/10.1109/ICTAI.2019.00030

Marin, E., Diab, A., & Shakarian, P. (2016). Product offerings in malicious hacker markets : 14th IEEE International Conference on Intelligence and Security Informatics, ISI 2015. *IEEE International Conference on Intelligence and Security Informatics*, 187-189. https://doi.org/10.1109/ISI.2016.7745465

Marin, E., Shakarian, J., & Shakarian, P. (2018, b). Mining Key-Hackers on Darkweb Forums. *2018 1st International Conference on Data Intelligence and Security (ICDIS)*, 73-80. https://doi.org/10.1109/ICDIS.2018.00018

Moore, D., & Rid, T. (2016). Cryptopolitik and the Darknet. *Survival*, 58(1), 7-38. https://doi.org/10.1080/00396338.2016.1142085

Motoyama, M., McCoy, D., Levchenko, K., Savage, S., & Voelker, G. M. (2011). An analysis of underground forums. *Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference*, 71-80. https://doi.org/10.1145/2068816.2068824

National Institute of Standards and Technology. (2023, mai 1). NIST. https://www.nist.gov/

Nee, C. (2015). Understanding expertise in burglars : From pre-conscious scanning to action and beyond. *Aggression and Violent Behavior*, *20*, 53-61. https://doi.org/10.1016/j.avb.2014.12.006

Nee, C., & Meenaghan, A. (2006). Expert Decision Making in Burglars. *The British Journal of Criminology*, *46*(5), 935-949. https://doi.org/10.1093/bjc/azl013

Nee, C., & Taylor, M. (2000). Examining burglars' target selection : Interview, experiment or ethnomethodology? *Psychology, Crime & Law*, *6*(1), 45-59. https://doi.org/10.1080/10683160008410831

Nee, C., & Ward, T. (2015). Review of expertise and its general implications for correctional psychology and criminology. *Aggression and Violent Behavior*, *20*, 1-9. https://doi.org/10.1016/j.avb.2014.12.002

Neuendorf, K. A. (2017). *The Content Analysis Guidebook*. SAGE.

Newman, M. E. J. (2018). *Networks* (Second Edition). Oxford University Press.

Newman, M. E. J., & Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical Review E*, *69*(2), 026113. https://doi.org/10.1103/PhysRevE.69.026113

Nunes, E., Diab, A., Gunn, A., Marin, E., Mishra, V., Paliath, V., Robertson, J., Shakarian, J., Thart, A., & Shakarian, P. (2016). Darknet and deepnet mining for proactive cybersecurity threat intelligence. *2016 IEEE Conference on Intelligence and Security Informatics (ISI)*, 7-12. https://doi.org/10.1109/ISI.2016.7745435

Nunes, E., Shakarian, P., & Simari, G. I. (2018). At-risk system identification via analysis of discussions on the darkweb. *2018 APWG Symposium on Electronic Crime Research (eCrime)*, 1-12. https://doi.org/10.1109/ECRIME.2018.8376211

NVD - Vulnerabilities. Accessed may 3rd, 2023, available at https://nvd.nist.gov/vuln

Paquet-Clouston, M., Décary-Hétu, D., & Morselli, C. (2018). Assessing market competition and vendors' size and scope on AlphaBay. *International Journal of Drug Policy*, *54*, 87-98. https://doi.org/10.1016/j.drugpo.2018.01.003

Paquet-Clouston, M., & Bouchard, M. (2023). A Robust Measure to Uncover Community Brokerage in Illicit Networks. *Journal of Quantitative Criminology*, *39*(3), 705-733. https://doi.org/10.1007/s10940-022-09549-6

Paquet-Clouston, M., Paquette, S.-O., Garcia, S., & Erquiaga, M. J. (2022). Entanglement : Cybercrime connections of a public forum population. *Journal of Cybersecurity*, *8*(1), tyac010. https://doi.org/10.1093/cybsec/tyac010

Paquet-Clouston, M.-C. (2021, août 30). *The Role of Informal Workers in Online Economic Crime*. Simon Fraser University. https://summit.sfu.ca/item/34606

Radianti, J., Rich, E., & Gonzalez, J. J. (2009). Vulnerability Black Markets : Empirical Evidence and Scenario Simulation. *2009 42nd Hawaii International Conference on System Sciences*, 1-10. https://doi.org/10.1109/HICSS.2009.504

Review of expertise and its general implications for correctional psychology and criminology. (2015). *Aggression and Violent Behavior*, *20*, 1-9. https://doi.org/10.1016/j.avb.2014.12.002

Rousseeuw, P. J. (1987a). Silhouettes : A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, *20*, 53-65. https://doi.org/10.1016/0377-0427(87)90125-7

Ruohonen, J., Hyrynsalmi, S., & Leppänen, V. (2016). Trading exploits online : A preliminary case study. *2016 IEEE Tenth International Conference on Research Challenges in Information Science (RCIS)*, 1-12. https://doi.org/10.1109/RCIS.2016.7549301

Samtani, S., & Chen, H. (2016). Using social network analysis to identify key hackers for keylogging tools in hacker forums. *2016 IEEE Conference on Intelligence and Security Informatics (ISI)*, 319-321. https://doi.org/10.1109/ISI.2016.7745500

Samtani, S., Chinn, R., Chen, H., & Nunamaker, J. F. (2017). Exploring Emerging Hacker Assets and Key Hackers for Proactive Cyber Threat Intelligence. *Journal of Management Information Systems*, *34*(4), 1023-1053. https://doi.org/10.1080/07421222.2017.1394049

Schaefer, D. R., Bouchard, M., Young, J. T. N., & Kreager, D. A. (2017). Friends in locked places : An investigation of prison inmate network structure. *Social Networks*, *51*, 88-103. https://doi.org/10.1016/j.socnet.2016.12.006

Schmidt, H. G., Norman, G. R., & Boshuizen, H. P. (1990). A cognitive perspective on medical expertise : Theory and implication [published erratum appears in Acad Med 1992 Apr;67(4):287]. *Academic Medicine*, *65*(10), 611.

Shakarian, J., Gunn, A. T., & Shakarian, P. (2016). Exploring Malicious Hacker Forums. In S. Jajodia, V. S. Subrahmanian, V. Swarup, & C. Wang (Éds.), *Cyber Deception : Building the Scientific*

Foundation (p. 259-282). Springer International Publishing. https://doi.org/10.1007/978-3-319-32699-3_11

Simon, H., & Chase, W. (1988). Skill in Chess. In D. Levy (Éd.), *Computer Chess Compendium* (p. 175-188). Springer. https://doi.org/10.1007/978-1-4757-1968-0_18

Statistical functions (scipy.stats)—SciPy v1.13.0 Manual. (s. d.). Accessed april 17th, 2024, available at https://docs.scipy.org/doc/scipy/reference/stats.html

Sutherland, E. H. (1937). The professional thief. *Journal of Criminal Law and Criminology (1931-1951)*, 161-163.

Topalli, V. (2005). Criminal Expertise and Offender Decision-Making : An Experimental Analysis of How Offenders and Non-Offenders Differentially Perceive Social Stimuli. *The British Journal of Criminology*, *45*(3), 269-295. https://doi.org/10.1093/bjc/azh086

Topalli, V., Jacques, S., & Wright, R. (2015). "It takes skills to take a car" : Perceptual and procedural expertise in carjacking. *Aggression and Violent Behavior*, *20*, 19-25. https://doi.org/10.1016/j.avb.2014.12.001

Tomczak, M., & Tomczak, E. (2014). The need to report effect size estimates revisited. An overview of some recommended measures of effect size.

Traag, V. A., Waltman, L., & van Eck, N. J. (2019). From Louvain to Leiden : Guaranteeing well-connected communities. *Scientific Reports*, *9*(1), Article 1. https://doi.org/10.1038/s41598-019-41695-z

UCLA. *FAQ How is effect size used in power analysis?*. Accessed march 28th, 2024, available at https://stats.oarc.ucla.edu/other/mult-pkg/faq/general/%20effect-size-power/faqhow-is-effect-size-used-in-power-analysis/

van Gog, T. (2012). Expertise. In N. M. Seel (Éd.), *Encyclopedia of the Sciences of Learning* (p. 1238-1240). Springer US. https://doi.org/10.1007/978-1-4419-1428-6_95

Vicente, K. J., & Wang, J. H. (1998). An ecological theory of expertise effects in memory recall. *Psychological Review*, *105*(1), 33-57. https://doi.org/10.1037/0033-295X.105.1.33

Vieraitis, L. M., Copes, H., Powell, Z. A., & Pike, A. (2015). A little information goes a long way : Expertise and identity theft. *Aggression and Violent Behavior*, *20*, 10-18. https://doi.org/10.1016/j.avb.2014.12.008

Wasserman, S., & Faust, K. (1994). *Social Network Analysis : Methods and Applications*. Cambridge University Press.

World Economic Forum. (s. d.). *The Global Risks Report 2023 18th Edition*. Accessed on April 17th, 2024, available at https://www3.weforum.org/docs/WEF_Global_Risks_Report_2023.pdf

Wright, R., & Logie, R. H. (1988). How Young House Burglars Choose Targets. *The Howard Journal of Criminal Justice*, *27*(2), 92-104. https://doi.org/10.1111/j.1468-2311.1988.tb00608.x

Wright, R., Logie, R. H., & Decker, S. H. (1995). Criminal Expertise and Offender Decision Making : An Experimental Study of the Target Selection Process in Residential Burglary. *Journal of Research in Crime and Delinquency*, *32*(1), 39-53. https://doi.org/10.1177/0022427895032001002

Yip, Webber, C., & Shadbolt. (2013). Trust among cybercriminals? Carding forums, uncertainty and implications for policing. *Policing and Society*, *23*. https://doi.org/10.1080/10439463.2013.780227

Zhang, J., Ackerman, M. S., & Adamic, L. (2007). Expertise networks in online communities : Structure and algorithms. *Proceedings of the 16th international conference on World Wide Web*, 221-230. https://doi.org/10.1145/1242572.1242603

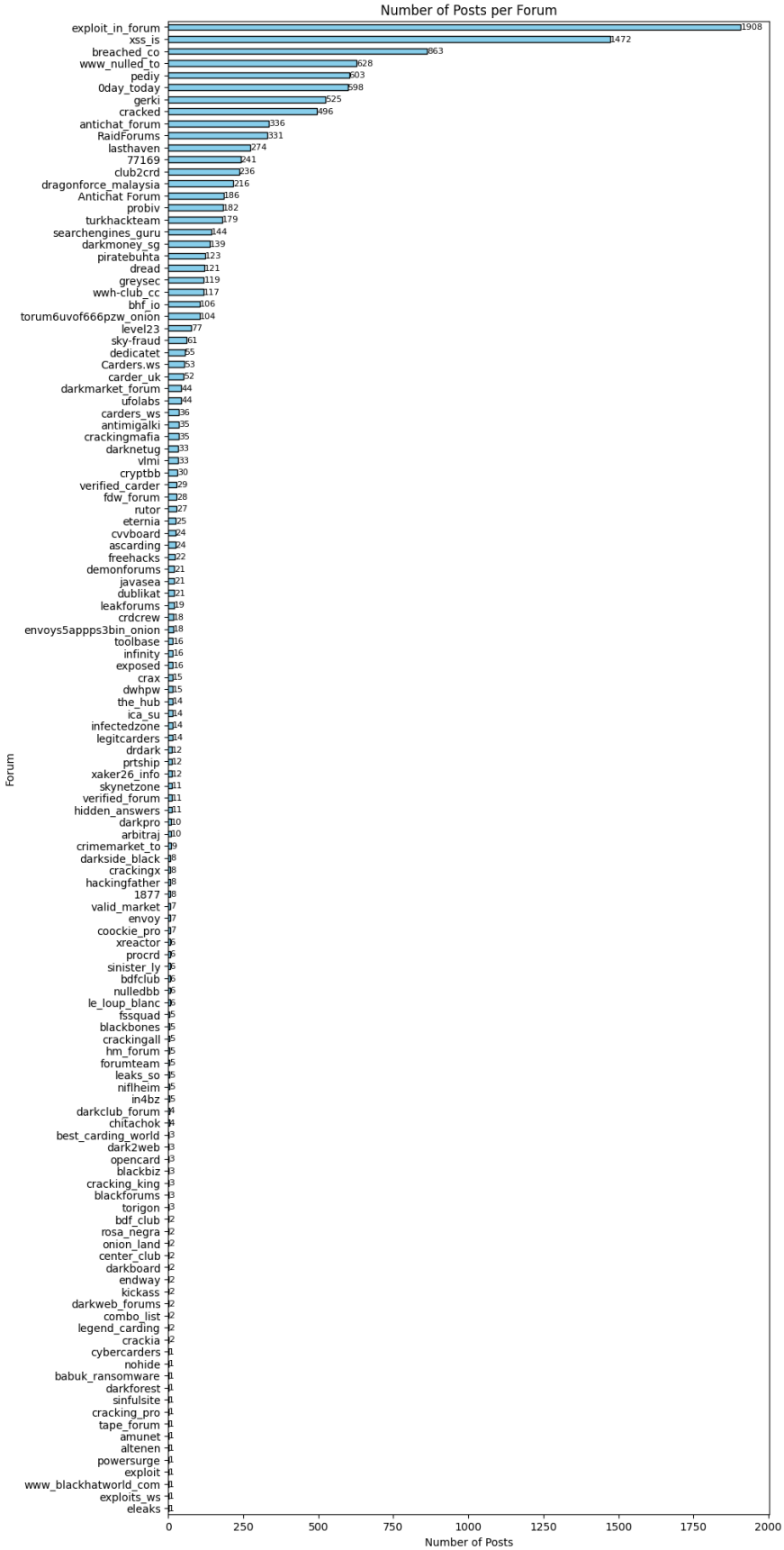
Zhang, X., Tsang, A., Yue, W. T., & Chau, M. (2015). The classification of hackers by knowledge exchange behaviors. *Information Systems Frontiers*, *17*(6), 1239-1251. https://doi.org/10.1007/s10796-015-9567-0

Annex A

The distribution of posts for all 124 forums is available in Figure 9.

Figure 9.

Distribution of posts per forum


Number of Posts per Forum
Forum
Number of Posts
exploit_in_forum 1908
xss_is 1472
breached_co 863
www_nulled_to 628
pediy 603
0day_today 598
gerki 525
cracked 496
antichat_forum 336
RaidForums 331
lasthaven 274
77169 241
club2crd 236
dragonforce_malaysia 216
Antichat Forum 186
probiv 182
turkhackteam 179
searchengines_guru 144
darkmoney_sg 139
piratebuhta 123
dread 121
greysec 119
wwh-club_cc 117
bhf_io 106
torum6uvof666pzw_onion 104
level23 77
sky-fraud 61
dedicatet 55
Carders.ws 53
carder_uk 52
darkmarket_forum 44
ufolabs 44
carders_ws 36
antimigalki 35
crackingmafia 35
darknetug 33
vlmi 33
cryptbb 30
verified_carder 29
fdw_forum 28
rutor 27
eternia 25
cvvboard 24
ascarding 24
freehacks 22
demonforums 21
javasea 21
dublikat 21
leakforums 19
crdcrew 18
envoys5appps3bin_onion 18
toolbase 16
infinity 16
exposed 16
crax 15
dwhpw 15
the_hub 14
ica_su 14
infectedzone 14
legitcarders 14
drdark 12
prtship 12
xaker26_info 12
skynetzone 11
verified_forum 11
hidden_answers 11
darkpro 10
arbitraj 10
crimemarket_to 9
darkside_black 8
crackingx 8
hackingfather 8
1877 8
valid_market 7
envoy 7
coockie_pro 7
xreactor 6
procrd 6
sinister_ly 6
bdfclub 6
nulledbb 6
le_loup_blanc 6
fssquad 5
blackbones 5
crackingall 5
hm_forum 5
forumteam 5
leaks_so 5
niflheim 5
in4bz 5
darkclub_forum 4
chitachok 4
best_carding_world 3
dark2web 3
opencard 3
blackbiz 3
cracking_king 3
blackforums 3
torigon 3
bdf_club 2
rosa_negra 2
onion_land 2
center_club 2
darkboard 2
endway 2
kickass 2
darkweb_forums 2
combo_list 2
legend_carding 2
crackia 2
cybercarders 1
nohide 1
babuk_ransomware 1
darkforest 1
sinfulsite 1
cracking_pro 1
tape_forum 1
amunet 1
altenen 1
powersurge 1
exploit 1
www_blackhatworld_com 1
exploits_ws 1
eleaks 1
0
250
500
750
1000
1250
1500
1750
2000

Annex B

The list of CAPECs and their respective origin of skill level value is available in Table 14.

Table 13.

CAPEC Skill Value Origin

id	origin
114	Parent CAPEC
115	Cybersecurity Experts
116	Child CAPEC
117	Parent CAPEC
125	Cybersecurity Experts
128	Cybersecurity Experts
129	Parent CAPEC
130	Cybersecurity Experts
131	Cybersecurity Experts
133	Parent CAPEC
137	Cybersecurity Experts
145	Cybersecurity Experts
148	Cybersecurity Experts
150	Parent CAPEC
151	Cybersecurity Experts
157	Parent CAPEC
160	Cybersecurity Experts
166	Parent CAPEC
168	Parent CAPEC
175	Child CAPEC
177	Cybersecurity Experts
181	Cybersecurity Experts
183	Child CAPEC
185	Cybersecurity Experts
187	Child CAPEC
190	Cybersecurity Experts
191	Parent CAPEC

194	Parent CAPEC
204	Cybersecurity Experts
206	Parent CAPEC
218	Cybersecurity Experts
221	Cybersecurity Experts
226	Cybersecurity Experts
228	Cybersecurity Experts
229	Cybersecurity Experts
233	Cybersecurity Experts
234	Cybersecurity Experts
242	Parent CAPEC
248	Cybersecurity Experts
251	Cybersecurity Experts
252	Parent CAPEC
253	Child CAPEC
263	Child CAPEC
276	Cybersecurity Experts
277	Cybersecurity Experts
278	Child CAPEC
279	Parent CAPEC
287	Cybersecurity Experts
290	Parent CAPEC
291	Child CAPEC
292	Cybersecurity Experts
293	Cybersecurity Experts
294	Parent CAPEC
295	Cybersecurity Experts
297	Child CAPEC
298	Cybersecurity Experts
300	Parent CAPEC
301	Cybersecurity Experts
302	Cybersecurity Experts
303	Child CAPEC

304	Parent CAPEC
305	Cybersecurity Experts
306	Cybersecurity Experts
307	Cybersecurity Experts
308	Cybersecurity Experts
309	Cybersecurity Experts
312	Cybersecurity Experts
313	Cybersecurity Experts
317	Cybersecurity Experts
318	Child CAPEC
319	Cybersecurity Experts
320	Child CAPEC
321	Cybersecurity Experts
322	Cybersecurity Experts
323	Cybersecurity Experts
324	Cybersecurity Experts
325	Cybersecurity Experts
326	Cybersecurity Experts
327	Cybersecurity Experts
328	Cybersecurity Experts
329	Parent CAPEC
330	Child CAPEC
383	Parent CAPEC
384	Child CAPEC
385	Cybersecurity Experts
386	Parent CAPEC
387	Child CAPEC
388	Cybersecurity Experts
389	Cybersecurity Experts
402	Parent CAPEC
441	Cybersecurity Experts
460	Parent CAPEC
463	Cybersecurity Experts

464	Cybersecurity Experts
469	Cybersecurity Experts
472	Cybersecurity Experts
478	Parent CAPEC
479	Cybersecurity Experts
480	Cybersecurity Experts
482	Cybersecurity Experts
486	Parent CAPEC
487	Cybersecurity Experts
488	Child CAPEC
489	Parent CAPEC
490	Cybersecurity Experts
491	Cybersecurity Experts
492	Child CAPEC
493	Child CAPEC
494	Cybersecurity Experts
495	Cybersecurity Experts
496	Parent CAPEC
497	Cybersecurity Experts
502	Cybersecurity Experts
503	Cybersecurity Experts
506	Child CAPEC
540	Cybersecurity Experts
549	Cybersecurity Experts
550	Cybersecurity Experts
551	Cybersecurity Experts
552	Parent CAPEC
555	Cybersecurity Experts
556	Cybersecurity Experts
558	Cybersecurity Experts
562	Cybersecurity Experts
563	Parent CAPEC
564	Cybersecurity Experts

573	Cybersecurity Experts
574	Parent CAPEC
575	Cybersecurity Experts
576	Cybersecurity Experts
577	Cybersecurity Experts
578	Cybersecurity Experts
586	Cybersecurity Experts
589	Child CAPEC
590	Child CAPEC
615	Cybersecurity Experts
633	Cybersecurity Experts
639	Child CAPEC
642	Child CAPEC
650	Parent CAPEC
651	Parent CAPEC

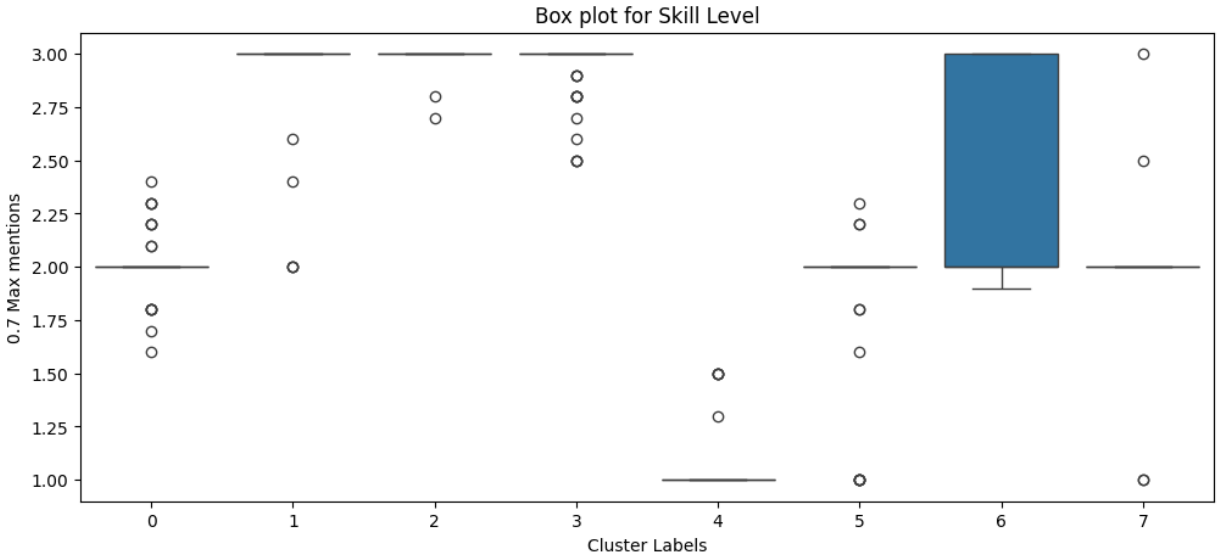
Annex C

Figures 10, 11 and 12 illustrate the distribution of skill level, commitment percentage and activity rate, respectively, per cluster.

Post hoc tests suggest that Cluster 1, 2 and 3 show a higher skill level compared to other clusters (C1mean=2.81, C2mean=2.96, C3mean=2.96). These three clusters showed significant differences with clusters 0, 4, 5 and 7 (p=0.000) and large effect sizes (|1.964| =< |*g*| =< |15.181|). On the other hand, cluster 4 exhibits the lowest skill level overall (C4mean=1.05), being statistically lower compared to every other cluster with a large effect size (p=0.000, |2.999| =< |*g*| =< |15.181|).

Figure 10.

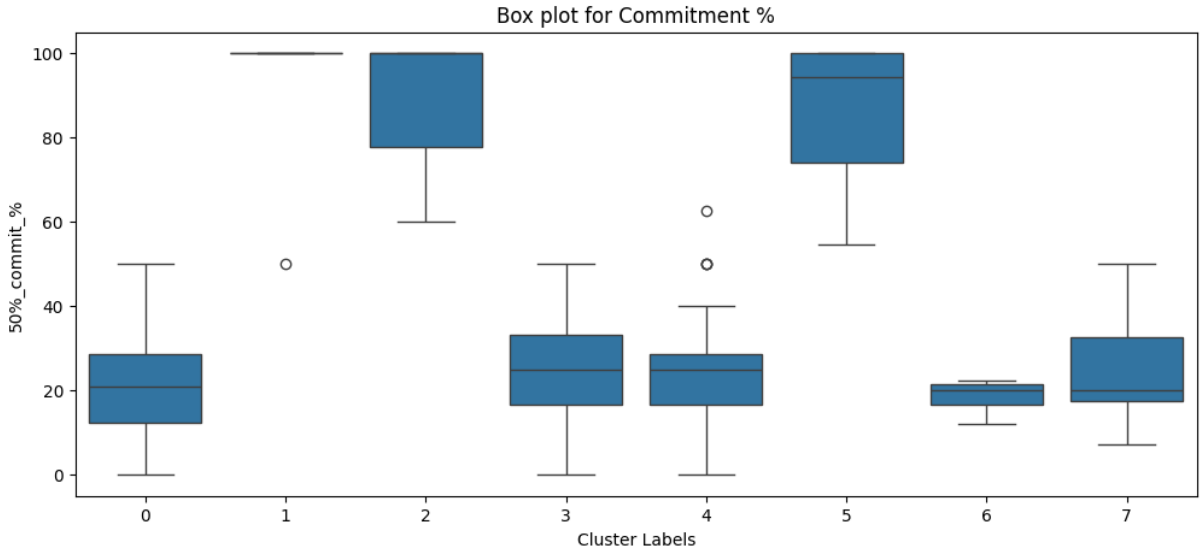
Skill Level Distribution per Cluster



Commitment Post hoc tests suggest that Cluster 1, 2 and 5 have the highest commitment percentage (C1mean=97.62%, C2mean=90.37%, C5mean=84.81%). They display significant differences with all other clusters ($0.000 < p < 0.01$, $|3.812| =< |g| =< |8.088|$).

Figure 11.

Commitment percentage distribution per Cluster



Post hoc test about activity rate show that cluster 1, 6 and 7 show the highest activity rate (C1mean= 5.14, C6mean= 10.67, C7mean = 4.14). These three clusters are significantly different from clusters 0,3,4,5 ($0.000 < p < 0.05$) with large size effects ($|3.593| =< |g| =< |23.191|$). Cluster 2 has a slightly higher activity rate (C2mean=0.28) compared to cluster 3 and 4 ($p<0.05$) with smaller size effects ($0.465 =< g =< 1.464$).

Figure 12.

Activity rate distribution per Cluster

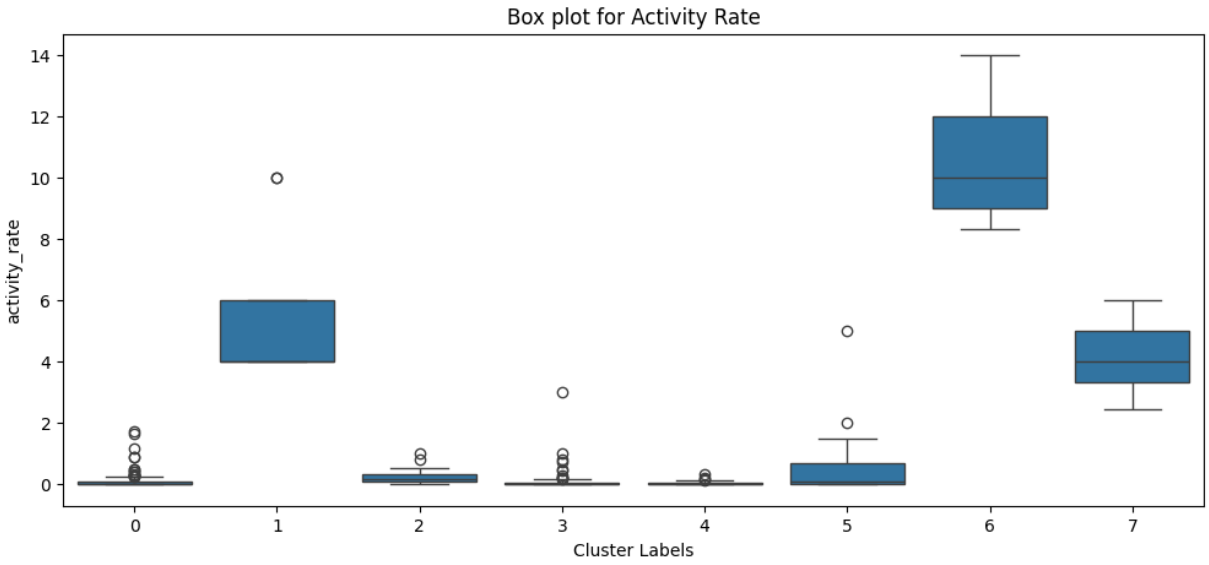


Table 14.

Post hoc tests and effect sizes for pairwise cluster comparison for Activity rate and Skill Level

activity/skill	0	1	2	3	4	5	6	7
0	X	-5.328***	-11.507***	-10.131***	9.286***			
1	-7.499***	X			7.252***	2.621***		1.964**
2		3.527	X		13.831***	3.553***		2.740**
3		5.972***	0.465*	X	15.181***	5.072***		4.777***
4		5.042***	1.464*		X	-2.999***	-6.091**	-3.219***
5		3.593***				X		
6	-23.191**		-9.407	-17.626**	-15.696**	-9.401*	X	
7	-10.338***		-4.868	-7.878***	-7.855***	-3.743***		X

Table 15.

Post hoc tests and effect sizes for pairwise cluster comparison for Commitment percentages

Commitment	1	2	3	4	5	6	7
0	-5.959***	-5.248***			-4.548***		
1			5.769***	5.825***		8.088***	6.403***
2			4.989***	4.905***		6.033**	4.974**
3					-4.210***		
4					-4.041***		
5						4.229**	3.812***
6							