Computers and Electrical Engineering 107 (2023) 108627

Contents lists available at ScienceDirect

Computers and Electrical Engineering

journal homepage: www.elsevier.com/locate/compeleceng

S.U.S. You're SUS!—Identifying influencer hackers on dark web social networks☆

Anum Atique Paracha [a], Junaid Arshad [b,*], Muhammad Mubashir Khan [a]

[a] Department of Computer Science & IT, NED University of Engineering & Technology Karachi, Pakistan
[b] School of Computing and Digital Technology, Birmingham City University, Birmingham, UK

ARTICLE INFO

Keywords:
Dark web
Threat intelligence
Social Network Analysis
Semantic analysis
Linear regression
Feature engineering

ABSTRACT

Dark web is an obscured part of the Internet, specifically used for sharing exploits, data breaches, and other means of cybercrime. Dark web forums provide opportunities to share such data and exploits and assign user reputation and credibility through participation in discussions and sharing data, exploits, and hacks. Such activities can help develop metrics to enable identification of influential mal-actors facilitating efficient and effective defense against emerging cyber threats, particularly zero-day exploits. This paper proposes an AI-inspired framework to identify influencers on dark web social networks (INSPECT) through intelligent analysis of user-profiles, interactions, and other activities. INSPECT framework leverages Feature Engineering, Social Network Analysis, Semantic Analysis, and K-means clustering and calculates an influencer score representing the users' significance within these forums. INSPECT has been evaluated using CrimeBB dataset comprising user profiles and activities within dark web forums to assess its effectiveness in identifying influential users on the dark web forums.

1. Introduction

Dark web is the non-indexed network hidden through layers of encryption followed by the development of Onion Routing for anonymous networking [1]. Onion Routing was designed for the bi-directional network accessibility with embedded privacy and anonymity known as Dark net. Used interchangeably with *Darknet*, dark web is the integral part of the Deepweb which comprises of almost 90% of the Internet [2]. Although technologies underpinning dark web were aimed at achieving privacy-aware use of public networks, such forums are increasingly used for conglomerate illicit and hidden services, forums, and marketplaces utilized for unethical and illegal activities and communication [3]. The research [4] highlighted the increasingly important role of dark web markets and communities by critically analyzing their impact on socio-economic and geographical interpretation. Further, the dark web became an attractive market for drugs distribution and gave rise to illicit use specifically in the era of COVID-19 [5]. The socialization, externalization, combination, and internalization (SECI) model defined in [6] interprets the illicit community development, knowledge enhancement and sharing practices on the dark web markets that contributes to illegal businesses, providing an anonymous platform for the illegal communities to prosper.

Typically, anonymous routers, various encryption schemes and dynamic URL changing mechanisms are used to achieve anonymity and stealth for this part of Internet from open access [5]. Table 1 presents a summary of important characteristics

☆ This paper is for CAEE special section VSI-webc. Reviews processed and recommended for publication to the Editor-in-Chief by Guest Editor Dr. M. Manikandan.

* Corresponding author.
E-mail address: junaid.arshad@bcu.ac.uk (J. Arshad).

https://doi.org/10.1016/j.compeleceng.2023.108627
Received 27 April 2022; Received in revised form 4 February 2023; Accepted 8 February 2023
Available online 20 February 2023
0045-7906/

Table 1
Dark web traits.

S.No.	Dark web traits	Technical aspect
1	Encrypted browsing	Dark web based browsers are designed to encrypt data with the multiple layers of encryption
2	Dynamic routing	VPN is required with specific ports to access and transmit data to the dark web
3	PGP encryption	Asymmetric encryption is provided to restrict the data readability
4	Anonymity	With random routing and VPN based access, the dark web users remains anonymous
5	Dynamic addressing	The URL addressing of resources and websites on the dark web is not static and keeps on changing

Table 2
Dark web reported incidents and their impact.

Source	Reported incidents and attacks	Impact
Y. Zhang et al. [7], 2019	Selling stolen credit card data on dark web marketplaces	Generate annual revenue of $300 million
Y. Zhang et al. [7], 2019	Providing rental DDoS services by dark web markets	Received annual profit of around $864 million
Sagar Samtani et al. [8], 2022	BlackPOS for target breach	Can reach to loss of more than $10 millions

of the dark web which facilitate its use for users requiring anonymity whilst also making it difficult to trace activities (illicit or benign) over the forums and platforms.

Further, criminal activities recorded over dark web account for more than 50% [2] which shows the usage and significance of this part of the public Internet. Table 2 present some recent examples of cases where dark web has been used for illicit purposes such as marketing exploits and sale of stolen data. The volume of cyber-threats and exploits has increased significantly — the daily cyber-attack rate reached to 4000 in 2016 [9]. For instance, in [3], the designed crawler has crawled a list of 0.5 million leaked Yahoo accounts credentials, posted in 2014 but identified by Yahoo in 2016. The significant increase in volume and variety of cyber-attacks occurring over recent years have highlighted that the conventional security controls are inadequate to achieve resilience within cutting-edge computing systems [10]. An emerging approach to enhance defense against cyber-threats is to improve Cyber Threat Intelligence (CTI) by developing innovative methods and techniques such as crawlers and spyware to harvest malicious activities on the dark web [10]. A number of approaches to CTI include significant human intervention and therefore error-prone and time consuming, which limits their effectiveness to develop effective defense mechanisms against emerging cyber-threats [11]. Although advancements within machine learning and AI have enhanced existing approaches [10], these are limited with respect to validating and verifying the extracted data and the source from where the intelligence is extracted. Intelligence data extracted from CTI get seriously affected by the scammed services and the fraudulent data, specifically designed and added, over these dark web forums, for diverting the crawlers and CTI services. Therein, the focus of our research is to address the significant challenge of not only extracting useful CTI from dark web forum, but to verify the extracted content and the targeted sources for CTI.

Critically evaluating the existing studies, [3,12] have developed the dark web crawling systems based on the semantic interpretation of the data posted on the forums to automate the CTI but the issue persists is to validate the profiles before extracting CTI information. [10] has proved that 90% of the posts are irrelevant out of 1 million. [13] has developed a system for the behavioral analysis and accounts recognition of the cybercriminals by investigating their social network analysis and semantic analysis of the users and [14] aggregated centralities measures of the users to identify the user types but the dataset utilized is outdated. The study [7] has provided the system to identify the key players or the key hackers by identifying the level of information from their conversation only with an accuracy achieved of almost 80%. But none of the state-of-the-art has exactly identify the influence score of the users that validate the authenticity of the users and their exact impact in the network.

In this paper, we present our efforts to address this challenge through development of a novel AI-inspired framework to identify influencers on the dark web social networks (INSPECT). The INSPECT framework investigates the hidden community of crime and provides the thorough and significant profiles that are influencers to that network and crucial source to depict the focused area of community and are also authenticated source of information transformations in the network. This way, identifying and omitting spam users and fake profiles and fraudulent data is reach in most authenticated manner.

The INSPECT framework first filters the inactive users or no mentioned ratings and then semantically infer the posted contents by the involved users which helps in identification of scrap or spam data for removal. This further categorize the users into dormant profiles and impactful users. Following this semantic analysis tightens the density of the social network efficiently. Applying social

network analysis on the updated network helps in reaching the social and geographical position of each entity of the network and its social influence from five different perspectives. Afterwards, linear regression algorithms are used to compute or predict the exact score of the profile based on the reflecting information and the community services. The resultant influential score of the users is more authenticated instead of user rating that is concluded on individual justification. Another resultant branch transformed from the INSPECT framework is the group identification of the users. Using clustering technique on the results achieved by the social network analysis, the network is segregated into transparent communities on the basis of influence and the infer knowledge.

1.1. Problem statement

The general secrecy surrounding dark web is primarily due to additional layer of encryption and anonymous routing applied within this part of the internet. Further, the continuous location changing property of the non-indexed contents of the dark web makes it more difficult to approach and target the active social networks and marketplaces [5]. Social site monitoring and profiling marketplaces are some of the techniques developed to access these untraceable networks and crawl the social communities, various activities and shared data [2].

Similarly, social networks and marketplaces are developed over the non-indexed networks for the communication and advertisement of illicit goods and hacking assets [11]. Specifically, the exploits including zero-days are the core part of such products. However, to utilize intelligence gathered from the dark web forums and marketplaces to enhance identification and protection against threats and malicious contents is a non-trivial challenge.

Although, research community has explored this challenge and made important advancements, however, the combination of malactors and benign users forms the spam forums and data to fool the scraping algorithms and CTI bots. Although, many security crawlers and CTI scraping tools are integrated with the dark web forums but if the retrieved data comprises of fake information it will be much damaging and results in waste of resources, time and often distract the focus of the investigation. In this way, the dedicated exploiters will be successful in attaining their objectives.

Examples of existing efforts for automated systems and crawlers such as [3,10,12] use different approaches to design crawling systems or algorithms that access dark web social networks and extract contents to identify threats and provide intelligence data but lack the ability of verifying and validating sources points and only target to specific hidden websites.

In view of the above, this research is focused at advancing the state of the art by conducting an in-depth analysis of dark web forums to identify influence of users within such networks. Our approach to analysis of dark web forums includes feature engineering, social network analysis, and machine learning with the aim to develop a model to compute influence of each user in the forum utilizing connections, interactions and posted information. In doing so, we hypothesize that identifying influencer hackers on dark web forums will enable us to filter and identify credible threat intelligence which can be used to strengthen defense against emerging cyber threats.

1.2. Major contributions

Increasing number of incidents and limitations of contemporary CTI techniques identified above highlight the significance of a threat intelligence system which can identify emerging threats by mining dark web forums. Our approach to achieve this goal is to focus on identifying users on dark web forums who can be regarded as *influencers* i.e. users who have the prestige and credibility as perceived by other users of the forum. By doing so, our hypothesis is that monitoring activities of such users on dark web forums will enable us to gather credible intelligence about emerging threats which can be used to enhance security systems.

Major contributions of this paper are:

1. Highlight the significance of dark web forums as an important cyber threat intelligence platform to enhance cyber defense mechanisms. Through the innovative use of cutting-edge technologies, this paper presents a method to identify meaningful intelligence from dark web forums which can help strengthen cyber defense mechanisms.
2. A novel framework (INSPECT) is proposed to analyses data from dark web forums to identify influential users and their associations through deep analysis of their activities such as posts, likes, and comments. The proposed framework drives its uniqueness by not only focusing on the user metrics but exploring hidden correlations among the user records, actions and reach. The proposed framework predicts an *influencer score* for users of the dark web forums by taking into account different features which represent their activities on the forums. Specifically, we use key features includes social network properties using centralities measures, Semantic Analysis using Latent Dirichlet Allocation for filtering noisy users' posts.
3. We present a detailed analysis of the INSPECT framework to assess its effectiveness to aid cyber defense measures through identification of influential mal-actors on dark web forums. The detailed analysis comprises of the user traits including user id, name, related threads and forums, posted/replied texts and user reputation, contributions and the text content posted on various threads.

Rest of the paper is organized as follows. Section 2 includes a discussion about the fundamental concepts with respect to the dark web, social network analysis and clustering. Section 3 presents a critical analysis of existing literature followed by Section 4 which includes dataset description and transformation of dataset into relational form which enables its use for this research. Section 5 presents details of the INSPECT framework including a detailed architecture as well as textual and algorithmic definitions of all major components of the system. Section 6 explains the experimental details and an analysis of the evaluation of the INSPECT framework. Section 7 includes a discussion of the limitations of the INSPECT framework and highlights open research challenges with respect to the use of dark web. Section 8 concludes the paper whilst also identifying opportunities for future indentation.

2. Background

Dark web is mostly intended to be used for non-ethical purposes as it is anonymous, the server's IP and addresses are encrypted, the user's identity is hidden and the routes taken by the server's include anonymous locators to be involved [1]. Therefore, the accessibility to the dark web is not straight forward as compared to the public Internet as it requires special software to access and browse. For instance, some specific routers, browsers or Web Engines are needed that include the required encryption and anonymity schemes. The Onion Router (ToR) [5], is one special type of browser that can used to reach to the contents available at dark web. The *Dark web* comprises of marketplaces and social forums used for the marketing products available on the marketplaces. Dark web forums and dark web marketplaces are in social contact and forums are the backbone of the markets that are utilized for different business requirements and trading of the objects including mostly illegal goods and tradings including Hacking exploits, malicious information and assets etc.

2.1. Dark web (Forums and marketplaces)

The most commonly utilized platform for the social communication and marketing of the illicit products and hacking assets, are the forums on the dark web. Hacker communities and the exploiters, from the beginners to experts, advertise and demonstrate the details of the products to be placed in marketplaces [15]. Hacking assets, various exploits, zero-days, stolen data; specifically financial data are most common product types that are not only advertised over marketplaces for selling but socialized and discussed over the forums [16].

Hacktivist activities including, exposing system vulnerabilities, advertising hacking tools and exploits, coordinating on various discussions are conducted over the dark web forums. These social forums communication generate digital traces, and therefore can be utilized for CTI effectively [8]. Not only the content of the dark web forums is important but, the networks developed on the forums and data of the contributors are of equal significance. For instance, the knowledge of the posting author, the impact of the query/response, the social influence of the users, all the immensely useful traits established by the dark web forums.

Similarly, dark web also facilitates its communities by offering hidden marketplaces in abundance. In correspondence to the dark web forums, dark web markets provide the transactional and financial services to the available products [8]. By incorporating crypto-currency as a form of secure and anonymous payments, these networks add another layer of anonymity and denies the identification of the user through transactional points.

These forums and market places are incorporated with various sophisticated user access control techniques to manage access. Therefore, only certain users with the token or credentials can gain the access to the systems. Marketplaces are intercommunicated with the hidden forums and these forums establish a baseline to these marketplaces.

2.2. Social network analysis

To identify the stature of the users within a social setting such as online forums and social media, Social Network Analysis (SNA) is one of the most effective solutions. SNA provides the ability to study and conduct deep analysis of the communal architecture of the specified community. The strong and tight coupling and interaction of the users. SNA is the complete and in-depth investigation and identification of the user's position. SNA calculates the communication route length of the nodes and the inter-dependent nodes occurred in-between that route. Centrality measures with variants defined below, identifies the flow of data in the network and the loop-in nodes required for roaming data in the network. SNA signifies the popularity and communication power of the node to rapidly flowing data in-between users and the whole group. To scale out users based on the social influence and strong community position, is very useful in identifying core area of interest of the registered group of users and in verifying and validating the pointed community and the effectiveness of the traced information. Due to the high-level similarities observed between social media forums and dark web forums, we believe SNA is a potentially effective mechanism to identify the hacktivist groups with maximum influence. Therefore, this research has utilized various techniques for the identification of the centrality, position and importance of each node in a particular network.

Measuring traits of SNA calculates the positional coordination of the node (forum user in this research) within the graph (forum network in this research) from various perspectives. Analyzing these features, results in coordination magnitude and the engagement proportion of the user in driving social network. Driving influence factor or the centrality of the contributor provides the significance of the user on the network as well as to other users. Thus, this analysis immensely helpful in proving the truthiness and significance of digital traces and network provision, highlights the core focus of the community and easy approach to other contributors.

In this research, we have used parameters such as *Degree of Centrality, Betweenness Centrality, Closeness Centrality, Eigenvector Centrality* and *PageRank* to conduct social network analysis in this research.

2.2.1. Degree centrality

Total number of direct links, either in-degree or out-degree, of the user with other users in the same network, is *Degree Centrality* of that user. It is used to analyze the direct connections of the user in the network. Degree Centrality is an effective way to identify the social impact by determining connected nodes. In this study, we use degree centrality as a major trait to measure the influence of the user. The mathematical formulation of the degree centrality is given in Eq. (1).

$$Cd(N_i) = \sum_{j=1}^{n} Xij/(n-1)(i \neq j) \tag{1}$$

Here, $Cd(N_i)$ is the calculated degree centrality score by the summation of direct links X_{ij} of the node N_i at position i in the complete network with other nodes n at position $jwherei \neq j$

2.2.2. Betweenness centrality

Interlink node for the flow of information, connection and communication (number of nodes needed to pass through a specific node for connecting to another node) is known as *Betweenness Centrality*. Calculating betweenness centrality of the user justifies the interlink dependency and communication paths developed by that user in that network. To identify the mediation role undertaking by the user, provides a baseline for detecting influence of a user in a network.

$$Cb(N_i) = \sum_{j<k} \frac{Gjk(Ni)}{Gjk} \tag{2}$$

Here, $Cb(N_i)$ is the score of betweenness centrality of node N at position i. Gjk is the path between two nodes at position j and at k where $Gjk(N_i)$ is the specified path via node N stand at position i for which we are calculating the betweenness centrality. $\sum_{j<k} \frac{Gjk(Ni)}{Gjk}$ is the summation of all the identified paths traverse by the node N_i that computes its betweenness centrality and n are all other nodes in the traversed network.

2.2.3. Closeness centrality

The sum of distances from one node to another node is known as *Closeness Centrality*. To calculate the closeness centrality of each individual user, it will be helpful in identifying the communication gap between users and how strongly users collaborated over the network.

$$Cc(N_i) = \frac{(n-1)}{\sum_{j=a}^{n} d(Ni, Nj)} (i \neq j) \tag{3}$$

Calculating the distance $d(N_i, N_j)$ between two nodes in the network with the minimum number of steps or crossing minimum nodes is known as closeness centrality. The Eq. (3) is the computation of summation of the distances between the node N_i to every other node N_j where $j \neq i$ and starting from a and a = 1. n is the total number of nodes in the network defined. Closeness centrality defines the reachability of the users and even their conveyed information that is the objective of this research.

2.2.4. Eigenvector centrality

To determine the importance of user not only by statistical analysis but also considering the importance of surrounding users is *Eigenvector Centrality*. Eigenvector Centrality is the recursive degree of Centrality of the user. This is the most important factor in determining the influential rate of the user in the specified network.

$$Ce(N_i) = 1/\Lambda \sum_{j \in M(N_i)}^{n} Nj \tag{4}$$

where N_j is the node at position j in the network. $Ce(N_i)$ is the eigenvector value calculated for the node N at position i and Λ is a universal vector. Here eigenvector centrality not only calculates the connected nodes but also the importance or centrality of those nodes that are connected providing the importance of node or user N_{Ni} with the significance measure of node N_j. Therein, it enables calculating importance of the users in the network through their correspondence and corresponding users.

2.2.5. PageRank

The likelihood representation defined by the probability distribution is known as *PageRank*. PageRank algorithm is the variant of the Eigenvector Centrality that is based on the in-degree users, the significance of the linked users and their link tendency. These are the major factor that are included to measure PageRank degree of each user in the defined network. PageRank is of equal importance as of Eigenvector to identify the influence of the in-degree user.

$$PR(N_i) = \sum_{N_j \in B(N_i)} PR(N_j)/L(N_j) \tag{5}$$

PR_i is the PageRank value for node N at position i computed by the summation of the calculated value of PageRank of node N at position j by dividing it with the linked tendency $L(N_j)$ of the node N_j. B_{N_i} is the set of all linked nodes. Using PageRank centrality, we are able to identify the significance of the user by the relative significance of the connected users.

2.3. K-Means clustering

Arranging data into homogeneous subgroups based on the similarity measures is the process of clustering. Clustering algorithm are mostly applicable where the data is not labeled with no availability of the ground truth. Iterative clustering mechanism that comprises of measuring squared distance from randomly assigned centroids, is known as K-Means clustering. To structure and segregate the non-labeled data, K-Means identifies the temporary centroids and determine the clusters based on the mean distances of the nodes from the specified centroid. Then, with each iteration of cluster formation, it calculates the minimum mean distances between the points of data and set the centroid position accordingly. K-Means clustering algorithm follows the approach of Expectation–Maximization. Based on the technique of Expectation–Maximization, the clusters follow the iterative mode for optimization. The social forums consists of various types of people having several types of expertise and information and here is no proper significance to distinguish community into different group. To recognize expertise level and signify groups of people, K-Means clustering algorithm is embedded in this framework. To identify the likely groups from social community, the INSPECT framework tends to perform well, based on similarity traits. To distinguish the users based on the expertise level, area of interest, collaboration rate and other features, K-Means clustering is considered the most suitable approach which extracts the major groups of people.

2.4. Feature engineering

Feature engineering is the metadata processing mechanism for improving and optimizing machine learning models to be applied. Preparing and structuring input data compatible with the processing model is the core objective of feature engineering. Feature engineering plays role to identify the more operative traits with their proper placement in the methodology to be more productive and efficient. Feature engineering mechanism transforms the algorithms effectively to maximize the efficiency of the algorithm and give effectual results. To structure the input data with immensely influencing features to optimize the performance and results of the applied model is the task achieved by the embedding feature engineering algorithm as the pre-processing step. *Imputation, Binning, Handling Outliers* and *Feature Split* are techniques for implementing feature engineering. Forums and the dedicated community over the dark web are critical to analyze and require optimized algorithms at pre-processing and processing layers of the designed framework.

2.5. Topic modeling or semantic analysis

For discovering data, correlate data relationship and data mining, Topic modeling is most efficient technique [17] and Latent Dirichlet Allocation (LDA) is proved to be highly effective method of Topic modeling or Semantic Analysis. Below given are the different text analysis strategies that can be used at various text examination scenarios. The processing overview of each of the technique is given as follows:

1. **Latent Semantic Indexing (LSI):** LSI uses the Singular value decomposition technique for dimension reduction and then identifies the semantics by mapping the spaces of the words to the extracted frequent text.
2. **Vector Space Model (VSM):** VSM is the simple keyword search algorithm to identify the words and their sequences, searched with maximum frequency terms.
3. **Probabilistic Latent Semantic Analysis (PLSA):** Derived from LSI, PLSI is a statistical approach for analyzing documents or the topics from the text. Instead of LSI, it reduces the dimensions of the text by incorporating latent class decomposition method
4. **Latent Dirichlet Allocation (LDA):** LDA is a generative model that probabilistically maps documents to the latent variables. It uses the sparse Dirichlet and probabilistic distribution to model the corpus [10].

For content analyzing to extract the discussed subject, LDA is the most suitable approach It is a generative model in which each text corpora or document belongs to multiple topics and multiple topics can be identified in one document Dark web forums comprise of highly dense networks and to divert the crawling algorithms, fake forums and large volume of spam data is logically attached to fool CTI systems [10]. To reduce the density of the network, we utilize LDA model and reduce the volume of network content by removing non-technical data.

2.6. Linear regression

An analysis estimation model that identifies the dependency and correlation of the independent variables and in result, predicts the resultant dependent variable. To statistically calculate the value, linear regression model is mathematically reliable and easy to interpret. To identify the mathematical value of reputation of the user in the network, we have compute the social impact of the user in the network, its popularity and contribution to connect other users and transference of information in the community. Centrality traits and SNA properties are directly proportional to the strong social profile and depicts high influence of the users in the network. To statistically map these Social Network traits as independent features will provide the score of influence of the user in the provided network by applying Linear Regression. To mathematically compute this value, Linear Regression works properly and compute the influence score by interpreting results of Social Network Analysis. To analyze and evaluate the impact of social traits of each entity on network, Linear Regression is the best fit. The deep and hidden networks, depends on the involved community and to properly derived the authority in the network, Linear Regression algorithm with multiple properties plays a vital role. To identify the credibility and the authority which further influence users in the provided network can be significantly calculated by the mechanism of Linear Regression [18].

3. Analysis of relevant existing work

Data from dark web forums has been used in existing research for multiple objectives. For instance, Akyazi et al.[19] presented a research study in which they have studied the criminal services provided by the hacking community by utilizing the secret forums of dark web. Using the CrimeBB dataset, the authors investigated the supply and demand for cyber crime services to the related community. As with social media platforms in the open Internet, users can have different usernames or identifiers across different platforms making investigations typically challenging. Cabrero-Holgueras & Pastrana [20] proposed a method to overcome this challenge by traversing and interpreting contents posted by the users. However, the research does not included the meta data analysis and the user traits to validate the posted data which can compromise the correlations identified.

With respect to use of dark web forums for cyber threat intelligence, Koloveas et al.[21], proposed a machine learning-based framework for threat intelligence which is focused on scraping data from multiple web sources including dark web. However, the approach lacks a data validation strategy to filter counterfeit data that can distort the outcomes. Although, a hybrid crawler is

designed for scrapping data but the login and accessibility complexities are not addressed. Further, the proposed framework focuses on analyzing contents using NLP techniques to develop intelligence data and does not take into account user profile and reputation which are vital to ascertain the trustworthiness of intelligence.

Kaur & Randhawa [1] and Sagar Samtani et al. [8], highlighted the role of dark web highlighting the use of dark web for the cyber crimes due to the properties of anonymity. In [1], the authors explored dark web from different aspects including features, the products such as gambling and hacking exploits available and incidents and criminal activities taking place over dark web. These hacking exploits primarily contain the critical and cutting-edge offensive cyber-attacks known as "Zero-days". Access to dark web forums is the initial step to uncover the stealth market of hacking exploits. But these social platforms and markets are hidden beneath certain layers of internet, therefore, to crawl these sites is itself a challenge. Whereas, in [8], developed a machine learning based proactive CTI mechanism to interlink the marketed exploits on the dark web forums to the system vulnerabilities. Also providing the importance of HackerForums, one of the dark web forum, in identifying the key hackers that is a strong justification of targeting HackerForum in our research.

Another study for providing proactive cyber threat intelligence and exploring hacking assets was conducted by N. Zhang et al. [11]. The researchers analyzed the importance of proactive cyber threat intelligence by focusing on the proactive approaches. Focusing on the dark web monitoring, by providing automated machine learning text-based CAPTCHA resolution, as a source of threat intelligence, this study demonstrates that proactive cyber threat intelligence can be more useful and dark web forums can play a key role in proactively identifying cyber threats.

To crawl hidden websites, forums and marketplaces, M. Pannu et al. [12] developed a scraping tool (Dark web Crawler) that is able to access the dark web through the use of The Onion Router (TOR) browser, and crawl the dark web websites and pages and after detecting the malicious data. To improve the effectiveness of CTI, researchers have also focused on automating the crawling mechanisms to extract data from dark web forums. In this respect, Isuf Deliu et al. [10], automate the process of gathering and filtering hacker forums and other social networks data, that is processed manually. In this research study, they have utilized various machine learning algorithms and showed that traditional machine learning algorithms (SVM) and Latent Dirichlet Allocation (LDA) for semantic processing of the gathered text data.

To address challenges with respect to data gathering, filtering and validation Deliu et al. [10], proposed a hybrid approach, designed using a machine learning algorithm to investigate the hacker forums automatically overcoming the challenges encountered in manual Cyber Threat Intelligence. The research paper highlights challenges to achieve reliable Cyber Threat Intelligence as well as the significant cost and performance trade-offs to achieve this. To reduce the cost and improve the performance, the study introduced a hybrid approach to automatically engage with the hacking forums and identify the subjects of the discussion using SVM, while using Latent Dirichlet Allocation algorithm for clustering. The testing phase showed the results that the model is effective to be utilized for extracting the threats and cyber intelligence and can be used for further Intelligence processes and mitigation purposes.

Schafer et al. [3] proposed a highly automated analytical framework (Black Widow) that collects and analyzes large amount of unstructured data. This research identifies that the dark web is the core source of extracting the data of hacking exploits and attacks for the CTI and Black Widow provides the platform for information gathering and transform that information for the proper utilization by providing analyzing framework. It analyzes and infers relationships between forums and authors and also used for cyber security related threats detection. Validation of scraped data is at the high priority to not get scammed by the exploiters over the forums and social platforms. The solution to this critical issue is to identify the key players over the dark web forums and authenticate data against that identity. Our study is focused on this challenge and we present our efforts in this regard in the later sections of this paper.

Zhang et al. [7] proposed a automatic and intelligent system named *iDetective* for analyzing dark web forums with the aim to identify the users that are key players over that forum. The system involved the heterogeneous user representation network and semantics to identify the users. Player2Vec technique is embedded in the system to learn user representation in the system in appropriate way. For further evaluation and validation, a number of different forums are tested with this proposed system. Similarly, Rios et al. [22], highlighted a problem of noise and interruptions of irrelevant posts and users in an educational or knowledge sharing forums. While applying graph analyzing algorithms to predict or identify the influential identities, such noisy documents divert the research predictions. This paper uses Latent Dirichlet Allocation (LDA) and Fuzzy Concept Analysis (FCA) to filter out the noisy and disruptive posts and users and generate more accurate graphs. But still these validation techniques do not accommodate the needs of verification. Only 40%–50% evaluation results can be achieved via these techniques.

Table 3 summarizes a comparative analysis of notable recent literature relevant to the proposed work. Existing approaches propose various CTI methodologies and intelligence systems with embedded ML techniques but, no method involves the verification and validation strategies to filter out the fraudulent forums, identities and contents that deceive the CTI systems and intelligence bots. This study primarily focuses to identify the influential users that plays a vital role for verifying the social platforms as well as the digital traces conveyed over those forums. Furthermore, proposed solutions, including [3,10,12], focuses to provide defensive security methods and detection systems for minimizing threats and securing the targeted systems. But these existing CTI solutions lack the verification strategies which proves the authentication of the scraped contents and social sites it follows. To address this challenge, our approach proposes a solution that includes identifying influencing users to extract the details of the respective community and the digital traces and which will further be utilized for verifying CTI extracted data.

Table 3
Comparative analysis of existing work.

Research paper	Problem statement	Data acquisition	Proposed solution	Achieved results	Limitation
I. Deliu et al. [10], 2018	To automate the extraction of CTI from Hacker forums as manual processes are error prone	Nulled.IO hacker forum data is used in this research study	Fetching CTI from Hacker forums using Hybrid model (SVM and LDA)	With this model, the Researchers had identified almost 90% irrelevant content out of 1 million posts	With increase size of vocabulary of LDA, the processing time is also increased and Human intervention is also required at certain stages
Y. Zhang et al. [7], 2019	Identifying socially influenced users in dark web forums	Data utilized from Nulled community and HackForum	Developed and intelligent system named: "iDetective", for automating analysis of user networks over underground forums	The accuracy measure of the system to identify key-players is almost 80%	Manual processes involved to divide, filter and labeled user data that is extracted from the designed system
U. Akyazi et al. [19], 2021	Impact and usage of cybercrime forums for providing services towards cyber crimes	HackForum Data leveraged from CrimeBB dataset	Machine Learning based classifier is designed to analyze the supply and demand of the cyber crime services offered over the cybercrime forums	Classifier is designed with 9 different models and most accurate results of 76% are achieved with LinearSVC model	The study only trace the heading and first post of the thread to the cybercrime services provided in the respective thread
J. Cabrero-Holgueras et al. [20], 2021	Identifying multiple accounts operated under same entities to provide helping hand towards threat intelligence analysis	This research uses CrimeBB dataset	The methodology comprised of manual identification of features set mostly consists of credentials and unique ids. A four step feature processing framework is designed followed by stylometry analysis to map the similar user accounts	With the proposed framework, the authors were able to achieve the accuracy rate of 60% with the error of 15% in results	Manual selection of features and manual investigation of the accounts are the major limitations. Also the results proved to be accurate only up to 60%
Sebastian et al. [22], 2017	Proper identification of the key influencers that are not properly addressed by SNA	To evaluate the given network analysis approach, Plexilandia (Online Social Network) is used	Semantically analyze the influence of posted data for identifying key players instead of detecting social activities	For the applicability of semantic filtration, we have removed up to 50% of the bogus users	The accuracy rate of the filter is not more than 50%
Sagar Samtani et al. [8], 2022	Mining hacking exploits from dark web forums and interlink with vulnerabilities for CTI	Crawler is structured in the Research model to crawl data	A deep learning based operational system for scraping data for CTI by interlinking exploits with identified vulnerabilities	With DNN based approach, the precision of the system is 20%–41% more than non deep learning models	Linking non published vulnerabilities and exploits is still a challenge and Validation of the collected exploits to the linked vulnerability is still not addressed
S. Kaur et al. [1], 2020	Detailed survey of the dark web, playing role as the biggest crime market	VPN connection with ToR is established to crawl data from dark web	An insight to provide the accessibility of dark web through various browsers and details of criminal activities over dark web	Detailed study to access dark web via various browsers and servers. Enormous aspects of utilizing dark web specifically for illicit usage	Although a detailed study of the exploits and defense mechanism is provided but no information regarding exploiter is pointed out
M. Schafer et al. [3], 2019	Challenges in CTI and information gathering from dark web	Puppeteer is used as crawler through VPN to extract dark web data	A highly automated modular system for scraping data from dark web forums and structured that data	Collect data of 7 services with 100,000 users. within 2 days of monitoring time, years of data has been collected	No data validation and semantic analyzing approaches utilized

(*continued on next page*)

Table 3 (*continued*).

Research paper	Problem statement	Data acquisition	Proposed solution	Achieved results	Limitation
M. Pannu et al. [12], 2018	To discover and detect malicious websites from dark web	This study developed a crawling mechanism with attached database to crawl dark web	A dark web crawler that identify malicious websites from dark web and store their data in linked versioned database	A scraping tool with versioning capability is developed for crawling malicious websites from dark web	This tool still lack capability to access websites that are multi-lingual and embedded with verification locks like CAPTCHA

Table 4
In-depth analysis of CrimeBB dataset and some other open-source datasets.

Dataset	Entity data			Feature		
	Member	Forum	Post	Completeness	Availability	Normalized
CrimeBB dataset	✓	✓	✓	✓	✓	✓
Darknet Market Archives by Gwern Dataset	✗	✓	✓	✗	✗	✓
Azsecure dataset	✓	✗	✓	✗	✗	✗
Kaggle dataset	✓	✗	✓	✗	✓	✗

4. Data collection, rationale and pre-processing

The availability of appropriate data is a major challenge to conduct the research presented here. Further, the challenge is not only limited to technical difficulties in gathering such data but extends to ethical and legal issues which can be non-trivial to navigate depending upon the data being collected as well as organizational policies around such activities.

Therefore, we have utilized CrimeBB dataset [23] collected by the Cambridge Cyber Crime Center which has been collected in 2017. This data set is one of the prominent dark web datasets available and has been used in other research such as [19,20]. In order to primarily achieve a robust framework that is able to identify the features and behavior of the forums and users and detect the influencers within the forums, CrimeBB dataset is the most suitable option as this dataset is focused on the dark web forums and suitable to investigate cyber threat intelligence research.

CrimeBB dataset provides the complete hacking community data that provide details with respect to the users, user groups, forums and the activated threads on those forums. To deal with the issues of completeness and accessibility, CrimeBB crawler was designed to traverse and catch all the data of the forum and its sub-modules including users, sub-forums and threads. Completeness here marked for traversing and storing all the respect data and meta-data with crawled instance. Similarly, availability describes the issue of login and accessing complexities that are highly used in the dark web forums to build inaccessibility for crawlers and unknown users. CrimeBB dataset crawler focused on these features while designing the crawling strategy, the details of the CrimeBB crawling mechanism is provided in [23]. Other open-source datasets analyzed in this study are Darknet Market Archives by Gwern Dataset, and Kaggle dataset for darknet forums . The comparison of the above identified dataset to be used in INSPECT framework is provided in Table 4.

The above discussion, develops a strong reason to choose CrimeBB dataset among others in this study. Although, a customized crawler can be designed by ourselves but the focus point of this research is to follow up with the influential entities that are trademark in criminal activities and not the crawling mechanism. In this way, using CrimeBB dataset under ethical bindings is the best approach.

4.1. Features metrics

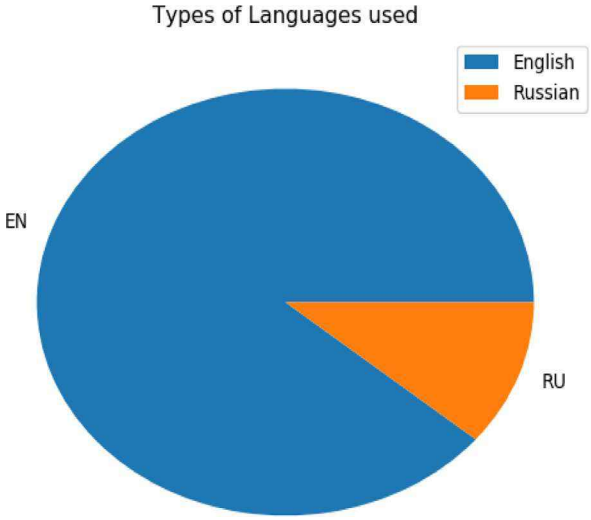
The original CrimeBB dataset is complex and comprises of 197 forums with 20597 average number of threads and 7 attributes or features for each forum. Therefore, we adopted feature engineering to facilitate identification of the most prevalent features that can be used to highlight influential users of the hidden forums. Forums are of various nature and domains and designed on heterogeneous architectural patterns. Although the focus of this research is on identifying influencer hackers within dark web forums, the INSPECT framework can be applied to other similar forums. The structured features set, with Feature Engineering, in this study tends to work with almost every hacking domain forum leading to identify key participants. The features selected to be used in the INSPECT framework are presented in the Table 5 whereas Table 6 presents all the features available within the dataset. The features are selected in a manner that represent properties of the forum, thread, user and their posts rather than the behavior or the configurations.

As the original dataset is in the relational format, ID of the entities is used to identify and link the segregated data. Feature engineering here is used to statistically analyze and normalize the numerical data available as the number of posts, reputation of the user, number of forums and members of the forums.

The targeted forums in the CrimeBB dataset comprise of various domains as specified in [23], but the focus of this research is to identify the influencer hackers within the dark web community. Therefore, we have selected *HackForum*, which contains

Table 5
Subset of features used in INSPECT framework.

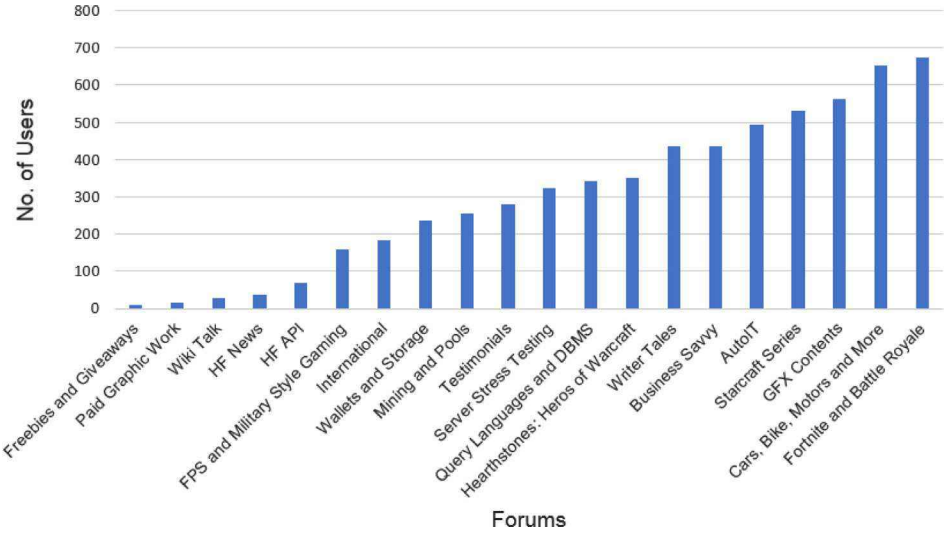
Feature name	Feature type	Related to	Description
Name	Varchar	Site	Name of the site
NumMembers	Integer	Site	Registered members on the site
NumForums	Integer	Site	Included forums on the site
IdForum	Integer	Forum	Primary identifier of the forum
Title	Varchar	Forum	Name of the forum
NumThreads	Integer	Forum	Number of threads generated
Site	Integer	Forum	Foreign identifier of the site
IdThread	Integer	Thread	Primary identifier of the thread
AuthorName	Varchar	Thread	Registered name of the author
Author	Integer	Thread	Foreign identifier of the author
Forum	Integer	Thread	Foreign identifier of the forum
Heading	Varchar	Thread	Registered name of the thread
NumPosts	Integer	Thread	Total number of posted data
IdPost	Integer	Post	Primary identifier of the post
Author	Integer	Post	Foreign identifier of the author
Thread	Integer	Post	Foreign identifier of the thread
Content	Varchar	Post	Text of the post
AuthorNumPosts	Integer	Post	Total number of posts of respective author
AuthorReputation	Integer	Post	Reputation of the author
AuthorName	Varchar	Post	Registered name of the author
IdMember	Integer	Member	Primary identifier of the member
Username	Varchar	Member	Registered name of member
TotalPosts	Integer	Member	Total number of posted data by member
Reputation	Integer	Member	Reputation of the member



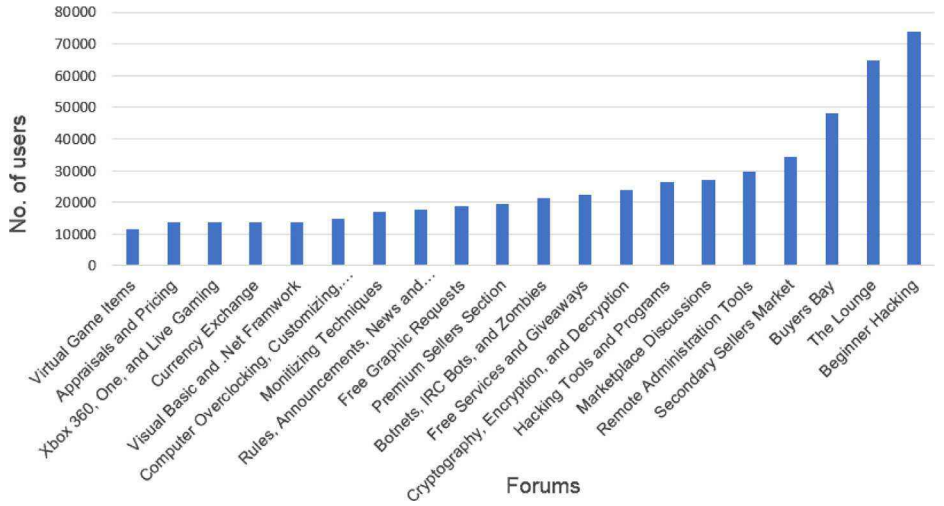
Clip 1. Segregation of forum data based on language used.

conversations from the dark web community related to exploiters and hackers of different levels of expertise. Another very important feature of the HackForum is that, it is the oldest forum to be crawled by CrimeBB dataset. These two features provides us a strong baseline to select HackForum for manipulating this study. CrimeBB dataset provides critical traits of the users including rating, active timestamp, generated threads and related posted contents. The data included in the forums is either in English or Russian - Clip 1 presents the proportion of distinct languages within forums.

Table 6 provides the detailed features of users, forums and correlated relation. This includes profiling details which reflect users' reputation, prestige, activity, time duration and interest of others via cited posts. Simultaneously, number of opened threads, posted contents and heading traces describes the usage and robustness of the forums. Users statistics are illustrated in Clips 2 and 3, which depict the least interested and most interested users in the specified forums of the HackForum board. The registration ratio of users not only reflects the users' interest but also the popularity of the forum. Simultaneously, Clips 4 and 5 demonstrates *liveliness* of the forums and how threads are currently registered and working over specific forums.



Clip 2. Minimum registered users for each forum.



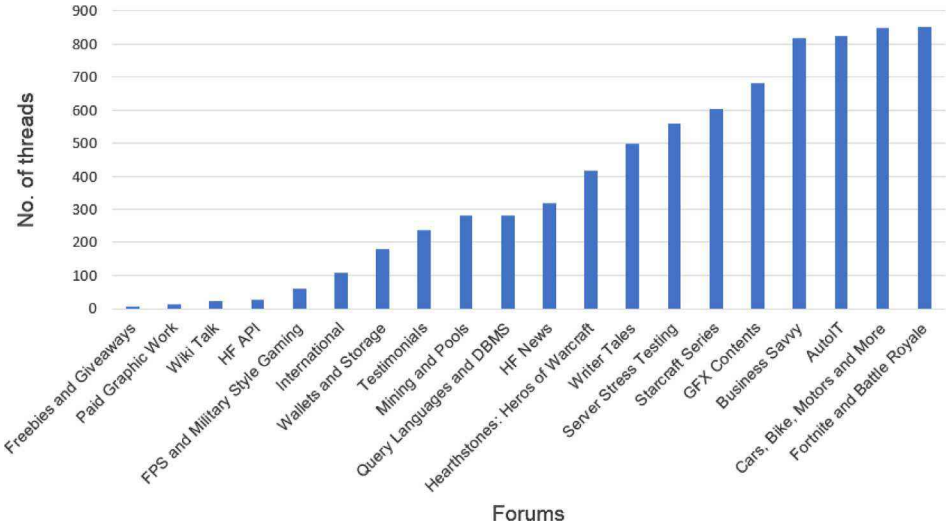
Clip 3. Maximum registered users for each forum.

Through an in-depth analysis of the statistics gathered and detailed survey of the forums, the most suitable forum is *Beginner-Hacking* which primarily comprises of hacking community and the posted contents have detailed discussions, queries and solutions of hacking exploits and assets. This is most active forum and consists of the maximum number of registered users.

4.2. BeginnerHacking forum

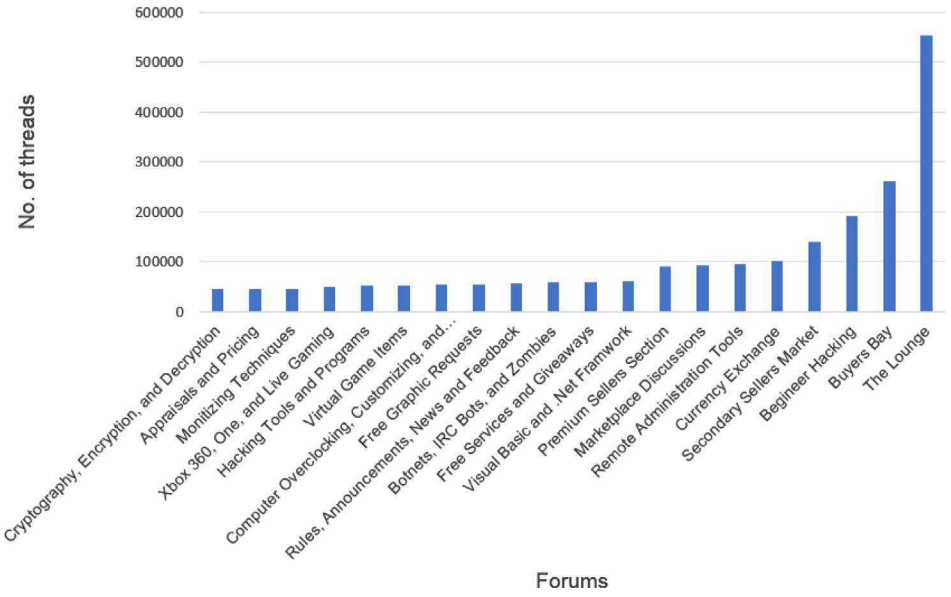
As this research is specific to identifying influencer hackers, we have selected *BeginnerHacking* forum. This forum contains an extensive amount of data and has the maximum number of registered users. Further, it contains the most number of conversations as compared to the other forums within the HackerForum board of CrimeBB data. Another important aspect is that the *BeginnerHacking* forum contains enormous information with respect to the hacking assets and beginner's queries. To evaluate the context of information concerning the hacking domain. We have experimented with the text content of *BeginnerHacking* forum, and the results are summarized in Clip 6.a and 6.b. Clip 6.a depicts the details embedded in the contents and subjective interest of the specified users.

The above-experimented results, given in Clip 6.a and 6.b, proved that the *BeginnerHacking* forum extensively used by hacker community. As this study mainly focuses on the hacking assets and the exploit details, we have analyzed the data and keywords relevant to cyber-attacks and the hacking domain. The outcome of this investigation is depicted in Clip 6.b.



Clip 4. Minimum Registered threads for each forum.

Clip 4: Minimum Registered threads for each forum



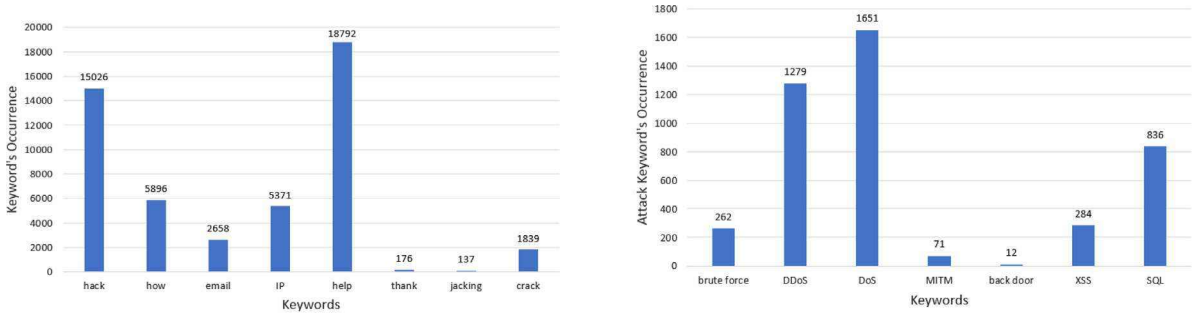
Clip 5. Maximum Registered threads for each forum.

From the graphical representation in Clip 6.b, it is clear that this forum has large amount of data not only for the basic queries and knowledge of cyber threats and its background, but the content also contains large volume of attacks data and related conversations. Further, through manual analysis of the Beginner Hacking forum data, it can be observed that the forum users represent different levels of expertise i.e. beginners to experts which fulfills the requirements of the current research.

5. An AI-inspired framework to identify influencer hackers on the dark web social networks (INSPECT)

Cyber threat intelligence is a crucial element in an organization's defense profile.

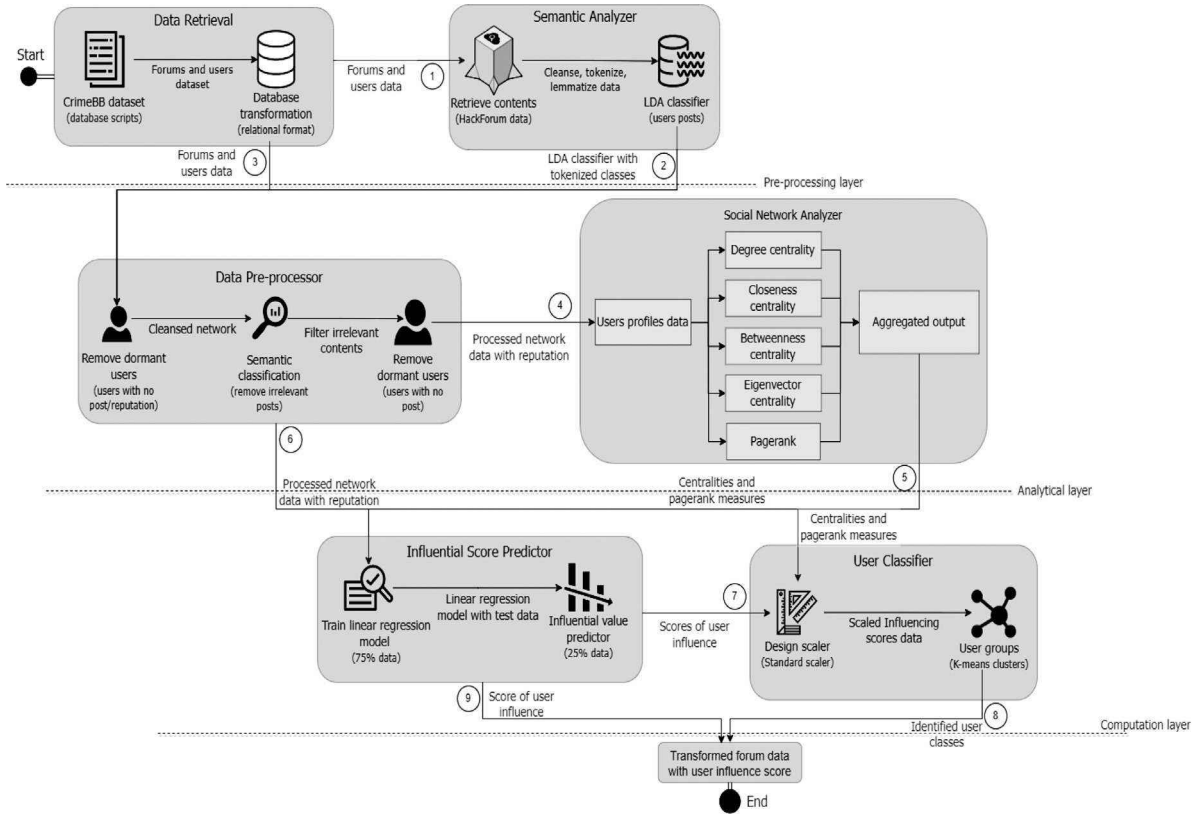
One of the most critical issues identified in the existing CTI algorithms is the verification and validation of crawled data to filter the spam information. Bots and scammer algorithms set up by mal-actors are focused on deceiving the CTI scrapers by incorporating fraudulent resources such as spam data and fake profiles [24]. In this context, the primary focus of the proposed approach is to identify the influence of the users by tracking their footsteps, posted content, and their reachability in the forum. To identify the



(a) Keywords utilized in contents

(b) Attack keywords utilized in contents

Clip 6. Attacks and keywords utilized in contents.



Clip 7. An in-depth representation of the INSPECT framework with process flow and its components.

influential entities, we have thoroughly studied the social networks of the dark web by developing a novel framework that comprises Social Network Analysis (SNA) to compute the centrality position of the users in the network, Semantic Analysis to extract the thoroughness of information in their posted text, influencing score predictor, and the user classifier to group users according to their influence factor. The graphical interpretation of the framework is presented in Clip 7.

The CrimeBB dataset comprises of detailed and complex features set. To select and transform usable features according to the INSPECT framework, feature engineering is used. This dataset consists of almost 0.7 M users from top 20 different social forums as illustrated in Clips 2 and 3 across approximately 0.5M threads as highlighted in Clips 5 and 5. The details of the features present within this data are shown in Table 6.

This INSPECT framework consists of three layers with dedicated processing components. The initial layer pre-processes the data by transforming it into the relational format followed by the semantic analyzer. The endpoint connector then forward the results of the pre-processing layer to the analytical layer which cleanse data and passes response to Social Network Analyzer which investigates the social interpretation of the user in that network. The final and crucial layer is the computation layer which predict the influential

Table 6
CrimeBB dataset feature description.

Feature name	Feature type	Related to	Description
IdSite	Integer	Site	Primary identifier of site
URL	Varchar	Site	Locator of the site
Name	Varchar	Site	Name of the site
NumMembers	Integer	Site	Registered members on the site
NumForums	Integer	Site	Included forums on the site
LastParse	Date	Site	Date of last update
IdForum	Integer	Forum	Primary identifier of the forum
Title	Varchar	Forum	Name of the forum
NumThreads	Integer	Forum	Number of threads generated
URL	Varchar	Forum	Locator of the forum
LastParsed	Date	Forum	Date of last update
Site	Integer	Forum	Foreign identifier of the site
IdThread	Integer	Thread	Primary identifier of the thread
AuthorName	Varchar	Thread	Registered name of the author
Author	Integer	Thread	Foreign identifier of the author
Forum	Integer	Thread	Foreign identifier of the forum
Heading	Varchar	Thread	Registered name of the thread
NumPosts	Integer	Thread	Total number of posted data
URL	Varchar	Thread	Locator of the Thread
IdPost	Integer	Post	Primary identifier of the post
Author	Integer	Post	Foreign identifier of the author
Thread	Integer	Post	Foreign identifier of the thread
Timestamp	Date	Post	Posted time of the post on thread
Content	Varchar	Post	Text of the post
AuthorNumPosts	Integer	Post	Total number of posts of respective author
AuthorReputation	Integer	Post	Reputation of the author
LastParse	Date	Post	Date of last update
Site	Integer	Post	Foreign identifier of the site
CitedPosts	Integer	Post	Number of read posts
AuthorName	Varchar	Post	Registered name of the author
Likes	Integer	Post	Total number of likes on post
IdMember	Integer	Member	Primary identifier of the member
Site	Integer	Member	Foreign identifier of the site
Username	Varchar	Member	Registered name of member
RegistrationDate	Date	Member	Date of registration over forum
TotalPosts	Integer	Member	Total number of posted data by member
Reputation	Integer	Member	Reputation of the member
Prestige	Integer	Member	Prestige of the member
LastParse	Date	Member	Date of last update
URL	Varchar	Member	Locator of the member profile

score of the users and also classify the user community into sub-groups based of the influence rate and the acquired knowledge of the users. For designing this novel framework, the *BeginnerHacking* forum is selected as the study primarily focuses on identifying influencers that help to investigate the hacking assets. Applying feature engineering concerning the study domain listed out the critical traits of the users and the forum associated to supplementary leads to graphical analysis and contracts the density of the injected network by excluding dormant users with no active status and minimum interaction.

5.1. Pre-processing layer

The initial layer of pre-processing transforms the dataset into the prescribed format of INSPECT framework then compiles the Semantic analyzer to further process data semantically. It also contains the end-point to communicate with external systems. The first task within the current investigation is to retrieve data and conduct initial transformation to enable processing and computation by further layers of the proposed approach. The pre-processing layer achieves this through two modules which are dedicated to these objectives.

5.1.1. Forum data retrieval

The first module of this research framework is to retrieve data and relationally transform it to work for further processing in framework. The algorithmic formulation for data transformation is given in Algorithm 1.

As the framework works around the dark web forums and social community interactions, the transformation builds the relation between the entities including site, forums, users, posts/contents and threads along with the related meta-data. The output generated by the module of Data Retrieval provides relational structured data that is utilized by post-processing modules of the INSPECT Framework to properly compute and justify the results. As the social forums work around various entities as mentioned, relational

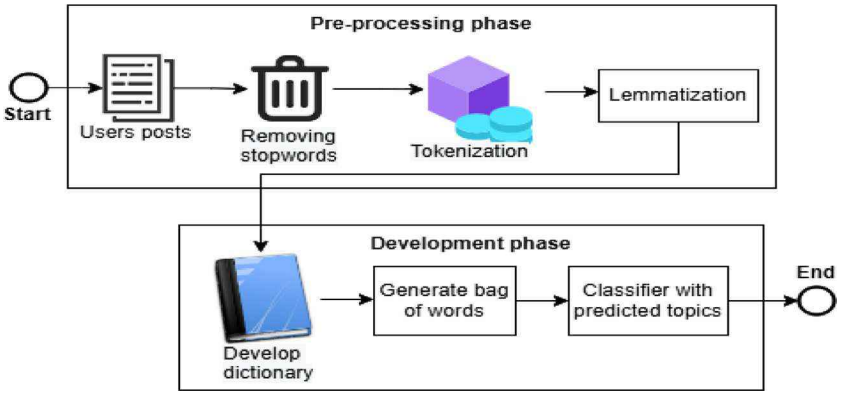
Algorithm 1 Forum Data Retrieval

```
var connection
var cursor
connection ← setup_database_connection
cursor ← connection.cursor()
userProfiles ← cursor.get_users_data()

profileCursor = 0
while userProfiles ≠ null do
    forum_network_retrieval ← add_user_node(userProfiles[profileCursor]) profileCursor+ = 1
end while

userPosts < −user_posted_contents()
profileCursor = 0
while userProfiles ≠ null do
    forum_network_retrieval() < −add_node_edge(userProfiles[profileCursor])
end while

user_social_network ← forum_network_retrieval()
```



Clip 8. LDA component model of INSPECT framework.

structure is the most suitable format to segregate and process the dataset in this framework instead of selecting any other structure (i.e. non-relational or file formats) for data setup.

5.1.2. Semantic analyzer

Analyzing text and extracting detailed information to obtain context is one of the most critical tasks to achieve. However, the information extracted as a result of the semantic analysis develops the baseline to; detect the context structure, the type of community, expertise level of the users, and the posted contents/replies of the users in the forums. Therein, semantic network analysis is used to conduct in-depth investigation of the social network community and filter irrelevant information to remove spam data and trivial (non-engaging) users. Some of the most appropriate methods designed for semantic analysis and topic modeling are: LSI, LDA, PLSI and VSM as discussed in Section 2.5, used in this research. The pictorial description is given in Clip 8.

Among these stated modeling strategies, LDA is the most effective and relevant technique. This framework maps the LDA-based semantic classifier over the posted contents and the shared information to extract the topic details that eventually help to remove the useless information and the noises added to the forum. The formulation of LDA classifier is given in Algorithm 2.

The structure of the Latent Dirichlet Allocation (LDA) classifier is developed by absorbing the conversations and queries shared in the discussions that appropriately relate to hacking terminologies. Then, this algorithm transforms the text to generate a bag of words by applying Stemming followed by Lemmatization to reduce text dimension morphologically. Subsequently, semantic distribution of the text contents generates the keywords classes based upon frequencies. These frequency classifications of keywords of the Beginner Hacking forum are in Table 7. The developed structure applied as the Semantic Analyzer results in retention of informational data and critical information in the social network.

5.2. Analytical layer

The second layer of the INSPECT Framework is dedicated to analytically process the data to analyze social impact of a user before classifying the user groups to predict the influential users. The analytical processing layer comprises of two modules with dedicated tasks assigned to each. Firstly, the data is analyzed for cleansing by applying features engineering, removing the dormant users and noisy contents with useless or no information with respect to threat intelligence. Another factor to be included is to conduct social

Algorithm 2 Semantic Analyzer

```
var connection
var cursor
connection ← setup_database_connection()
cursor ← connection.cursor()
userPosts ← retrieve_users_posts()

postCursor = 0
while userPosts ≠ null do
    userPosts ← content_cleanser(userPosts[postCursor]) postCursor+ = 1
end while

postCursor = 0
while userPosts ≠ null do
    userPosts ← lowercase_content(userPosts[postCursor]) postCursor+ = 1
end while

postCursor = 0
while userPosts ≠ null do
    userPosts ← tokenize_content(userPosts[postCursor]) postCursor+ = 1
end while

postCursor = 0
while userPosts ≠ null do
    userPosts ← lemmatization(userPosts[postCursor],
noun, adjective) postCursor+ = 1
end while

postCursor = 0
dictionary ← content_dictionary(userPosts)
while userPosts ≠ null do
    doc_term_matrix ← bag_of_words(content_dictionary(
userPosts[postCursor])) postCursor+ = 1
end while

LDA_model ← LDAModel(corpus = doc_term_matrix, id2word = dictionary, num_topics = 10, random_state = 100, chunksize = 1000, passes = 50)
```

Table 7
Contextual data extracted for defining topics of classes through Semantic Analysis (LDA component of INSPECT framework).

S.No	Classes keywords
0	0.044*“hack” + 0.037*“account” + 0.035*“site” + 0.029*“way” + 0.029*“password” + 0.027*“email” + 0.021*“address” + 0.020*“school” + 0.019*“server” + 0.018*“free”
1	0.043*“thing” + 0.034*“use” + 0.031*“guy” + 0.026*“much” + 0.025*“problem” + 0.025*“big” + 0.024*“new” + 0.024*“bad” + 0.020*“version” + 0.020*"page”
2	0.092*“thank” + 0.057*“help” + 0.045*“tutorial” + 0.042*“good” + 0.039*“work” + 0.032*“forum” + 0.032*“program” + 0.028*“search” + 0.025*“section” + 0.019*“helpful”
3	’0.079*“file” + 0.031*“people” + 0.029*“software” + 0.023*“possible” + 0.021*“old” + 0.018*“lol” + 0.016*“fake” + 0.014*“error” + 0.013*“name” + 0.012*“alot”’
4	’0.105*“nice” + 0.071*“great” + 0.047*“thank” + 0.044*“rat” + 0.031*“man” + 0.026*“post” + 0.025*“download” + 0.023*“link” + 0.022*“useful” + 0.019*"info”’
5	’0.042*“computer” + 0.040*“virus” + 0.028*“hacker” + 0.025*“tool” + 0.022*“time” + 0.020*“network” + 0.016*“simple” + 0.014*“change” + 0.013*“account” + 0.013*“day”’
6	’0.171*“net” + 0.151*“showthread” + 0.126*“php” + 0.121*“pid” + 0.087*“#” + 0.033*“https” + 0.024*“link” + 0.009*“tid” + 0.007*“search” + 0.007*“money”’
7	’0.114*“image” + 0.109*“smilie” + 0.089*“good” + 0.035*“img” + 0.029*“method” + 0.028*“list” + 0.026*“thank” + 0.024*“fud” + 0.020*“friend” + 0.019*“job”’
8	’0.080*“link” + 0.075*“com” + 0.060*“http” + 0.045*“www” + 0.014*“code” + 0.013*“user” + 0.013*“gmail” + 0.011*“file” + 0.010*“victim” + 0.010*“way”’
9	’0.058*“lot” + 0.057*“good” + 0.057*“port” + 0.049*“thread” + 0.031*“short” + 0.026*“open” + 0.025*“beginner” + 0.023*“luck” + 0.023*“stuff” + 0.021*”one”’

analysis of the network structure and calculate the social impact and position of the users in the network. Sample results gathered from the Social Network Analysis are given in Table 17. The in-depth details of each of the modules in Analytical layer are described below.

5.2.1. Data pre-processor

The social network forums are majorly populated with irrelevant and inactive users. Significantly, the hidden forums over the dark web network are merely used for in-demand illicit purposes. And to terminate the CTI bots and crawling algorithms, various spam information, and the users are embedded to fool the intelligence systems. Similarly, many fraudulent user-profiles attach

their responses to signify the influencing reflection. The incorporated LDA classifier in this novel framework plays a crucial role in separating the noisy and fake data to gain real context and topics from the text. At this pre-processing stage, the framework eliminates the inactive users with no logged communication and the illicit users identified by applying the LDA analyzer. The semantic analyzer will remove the illogical contents that will further transform the users inactive. The framework will detach those users. The ground truth utilized for experimentation and evaluation is the User Reputation. The user data with zero or null reputations is also of no use and removed from the processing data. The algorithmic interpretation is given in Algorithm 3

Algorithm 3 Data Pre-processor

```
connection ← setup_database_connection()
cursor ← connection.cursor()
userNetwork ← retrieve_network()

LDA_excluded_topics ← topic_indexes(2, 3, 7, 9)
while node in userNetwork do
    userPosts ← get_user_posts(node)
    while userPosts ≠ null do
        if post in LDA_excluded_topics then
            userNetwork ← remove_edge(node)
        end if
    end while
end while

while node in userNetwork do
    if user_node_edge(node) = 0 then
        userNetwork ← remove_user_node()
    end if
end while
```

Framework pre-processor will follow the data setup module and take the developed relational data having users, forum features, and contents in the forums from the designated database. Subsequently, the pre-processor aligned the LDA analyzer to process text contents semantically. This designed pre-processing algorithm depicts the ability to process semantics of the social network, identify topics, and reduce the density of the network by removing irrelevant data along with dormant and spam users to set the forum with the appropriate users and relevant information. This pre-processor makes the identification of the influencers more significant.

5.2.2. Social network analyzer

An in-depth investigation of the social network that provides the concrete details of each of the users and the analysis of the social interactions and calculates the communication links between various users that developed the network Centrality measures are the most impactful and usable features in analyzing the significance and social impact of the user in the forum. Centralities used in this framework measures the minimum communication paths required (Betweenness Centrality), 1st degree communicated nodes or users (Degree Centrality), the importance of users concerning others (Eigenvector Centrality), probabilistic distribution of information conveyed (Pagerank), and least number of users require for communication (Closeness Centrality). To signify and calculate the adherence of every user in the social arrangement and the conformity of every individual user of the social group, Social Network Analysis (SNA) tends to be the most effective method. After interpreting text and removing dormant users, the resultant network mainly consists of interconnected users and informational data. Calculating the exact position and social impact of every user in the network, Social Network Analysis (SNA) is embedded in the proposed framework. The processed and reassessed forum network with the features and contents work as the input to this module of Social Network Analysis which further measures the central position and social impact of each user in the network. Algorithms utilized in the mechanism of the SNA include Degree Centrality, Betweenness Centrality, Closeness Centrality, Eigenvector Centrality, and Page Rank. These algorithms identify the social impact and the correspondence of each individual with others in the specified network and calculate the respective centrality position of every forum user, which interprets the influential identities of the forum with maximum social reflection. In response, the social network analysis mechanism provides the segregated centralities scores of each user to define its position and influence from 5 different angles. The technical interpretation is given in Algorithm 4.

5.3. Computation layer

Computation layer is a critical element of the INSPECT framework and is envisaged to predict the influential score of each of the users and organize users into groups based on their influence, provided along with relevant meta-data. Also, this last layer of processing is responsible for categorizing users into specific groups concerning the calculated influence. The detailed technical information of the included modules is given below.

5.3.1. Influential score predictor

To identify, the influential user, a regression algorithm is incorporated to calculate the score of influence by applying it over the centrality scores individually for every user provided by the Social Network Analysis. To develop the statistical relation between the calculated centralities, which results in identifying the influence of the user, is to map the results of the Social Network Analysis over the Regression model. Centralities are the results of the user profile and its social impact in the network to helps in computing the rate of influence by propagating direct relation. Each of the centrality measures provided by SNA is directly proportional to the influence impact of the user in the respective social network. Hence, to manipulate the influential score of the users, a

Algorithm 4 Social Network Analysis

```
var degreeCentrality=0
var closenessCentrality=0
var betweennessCentrality=0
var eigenvectorCentrality=0
var pageRank=0
var closenessDistance=0

connection ← setup_database_connection()
cursor ← connection.cursor()
userNetwork ← retrieve_network()

while node in userNetwork do
    degreeCentrality ← degreeCentrality +
      node/(len(userNetwork) − 1)
    closenessDistance ← closenessDistance +
      distance(node_i, node_j)
    betweennessCentrality ← betweennessCentrality +
      path_node(node_i, node_j)
    eigenvectorCentrality ← eigenvectorCentrality +
      neighbourNode_centrality(Nj)
    pageRank ← pageRank + pageRank(j)/linkedNodes(j)
end while

closenessCentrality ← (len(userNetwork) − 1)/closenessDistance
eigenvectorCentrality ← 1/λ ∗ eigenvectorCentrality
```

linear algorithm works best in this particular scenario. The complete results of the social network analysis provide the input to the regression model to identifies the influence rate in response. Another possible techniques can be the Logistic Regression (results in binomial response) which depicts the best for classification problems [25]. To compute the influential score of each individual in the respective forum, Linear Regression applies to the gathered results from Social Network Analysis. This designed framework uses a Multi Linear Regression model to statistically compute the exact value of leadership score of the users. Influence or social consequences of the user in the respective group is directly proportional to each of the centrality measures, so applying Linear Regression collectively over the SNA results will provide an influential score of each entity of the social forum as their score of influence. The resultant score, reflects the overall collaboration, domination factor of the user on other users, authenticity of the activities, and shared information of the collaborator over the forum. The algorithmic steps are defined in Algorithm 5

Algorithm 5 Influential Score Predictor

```
x ← array(degreeCentrality, closenessCentrality, betweennessCentrality, eigenvectorCentrality, pageRank)

y ← array(userReputation)
MLR ← MultipleLinearRegression()
MLR ← dataSplit(train_data = 75%, test_data = 25%)
MLR_Train ← MLR(trainLR(x, y, 0.75))
MLR_Test ← MLR(predictLR(x, y, 0.25))
output LR_Test
```

5.3.2. User classifier

Another determined target of this framework is to recognize certain groups of users. The baseline for this clustering is the popularity and the shared level of knowledge. The framework designed in this study embedded 2 resultant approaches. At first state, user rating is suppressed by the influential score as this is the achieved value from the footstep traces of the users and centrality measure. Another focus of the study is to segregate and classify users into respective categories based on the influence rate and user interest and experience. Identifying the clusters of users will further approach to achieve the types of users, the maturity level of the network, critical use of the platform, and the severity of information posted over the forum. For classifying the users of the social forum supervised clustering algorithm, Support Vector Machine (SVM) is a supervised classification mechanism that plots the hyper-plane and maps the data points that are segregated by that hyper-plane, can be used. But, the only ground truth feature available, is the user reputation which cannot be incorporated in the classification track, because it is not a mathematically identified value rather only the user interpretations for each other. The developed framework prioritizes the unsupervised clustering algorithm and uses K-Means clustering for the classification of the users. The algorithm is given in Algorithm 6. The K-Means clustering algorithm is centroid-based and measures the least distances to generate clusters, so this is the most fitted technique to be incorporated in this use case. K-Means clustering algorithm incorporated with the proposed framework design that clusters users and defines 3 different sub-categories of the users within the respective forum. The given input to this clustering mechanism is the calculated centralities and the reputation score of the user from the dataset, which generates the users' clusters as the resultant data. These clusters then signify the user's characteristics and help to visualize the forum activeness, utilization, and the core theme of discussion on the forum.

Algorithm 6 User Classifier

$DataScaler \leftarrow StandardScaler()$
$data \leftarrow array(userReputation, userInfluence)$
$transformedData \leftarrow DataScaler(data)$

$KmeansModel \leftarrow KMeans(clusters = 3, initialization\_value = 10, maximum\_iterations = 300)$

$clusters \leftarrow KmeansModel(transformedData)$

Table 8
Classes identified as irrelevant and to be removed.

S.No.	Classes to be removed
1	0.092*“thank” + 0.057*“help” + 0.045*“tutorial” + 0.042*“good” + 0.039*“work” + 0.032*“forum” + 0.032*“program” + 0.028*“search” + 0.025*“section” + 0.019*“helpful”
2	’0.079*“file” + 0.031*“people” + 0.029*“software” + 0.023*“possible” + 0.021*“old” + 0.018*“lol” + 0.016*“fake” + 0.014*“error” + 0.013*“name” + 0.012*“alot””
3	’0.114*“image” + 0.109*“smilie” + 0.089*“good” + 0.035*“img” + 0.029*“method” + 0.028*“list” + 0.026*“thank” + 0.024*“fud” + 0.020*“friend” + 0.019*“job””
4	’0.058*“lot” + 0.057*“good” + 0.057*“port” + 0.049*“thread” + 0.031*“short” + 0.026*“open” + 0.025*“beginner” + 0.023*“luck” + 0.023*“stuff” + 0.021*”one””

Table 9
Configuration setup for experimentation.

Resource name	Details
Server	Name: Linux, Version: Debian, HD: 3568 MB, RAM: 8 GB
Database	Name: Postgresql, Version: 12.1
Language	Name: Python, Version: 3.8.6
APIs	Name: Networkx, nltk, numpy, spacy, genism, pyLDAvis, matplotlib, seaborn, psycopg2, sklearn, pandas

6. Experimental details and evaluation

The core objective of the INSPECT framework is to assess the influence of the users within dark web social network forums i.e. to identify influencer users within these networks. Another significant outcome achieved from this study is the identification and classification of users’ groups within these social environment. To evaluate the designed framework, thorough experimentation is conducted, which is described in the following sections.

6.1. Experimentation setup

To conduct the in-depth experiment of the INSPECT framework, an experimental testbed is formed which contains data from the CrimeBB dataset. Specifically, the data from Beginner Hacking forum is utilized such that 25% of the data from this forum is used for training phase and the rest of the 75% of the data is used for the test phase of the framework. Manual analysis is mandatory for classifying the generated classes for relevant extraction and prediction. Three classes are marked as irrelevant to skip out the noisy and irrelevant data. The classes marked to be mapped on the irrelevant data that need to be removed are shown in Table 8.

These classified groups of the classes are then utilized for experimenting and validating the LDA classifier and the noise-free social network. For the effective execution and testing of the developed framework, the environment has been tuned to the Linux system like database and python installations and set up the required APIs. The designed specifications are given in Table 9.

6.2. Experimentation and evaluation

The experiments to evaluate INSPECT framework used data about “Beginner Hacking” forum, forum included in the CrimeBB dataset, due to factors such as variety of attacks, diversity of users, and volume of the data. An in-depth investigation of the targeted forum is carried out as the pre-experimentation step, which is the thorough analysis of the design, community, activeness, and usage of the forum. In order to conduct a comparative analysis of the *influence score* of a user calculated by the INSPECT framework, the CrimeBB dataset has a *reputation* attribute for each user which can be utilized to assess the effectiveness of our approach. The analytical details of the Beginner Hacking forum taken from the CrimeBB dataset script are described in Table 10.

The statistics provided in Table 10 are computed from the postgreSQL database generated from the CrimeBB dataset. The pre-processing stage of the experimentation omitted the inactive or dormant users and the noisy information by applying the LDA Classifier. To interpret the respective network semantically, network contents are incorporated in the classifier designed with the LDA algorithm. The classifier corroborated the communicated data and lined-up with the frequency-based classes to detect various topics or subjects included in the network. The subject lines can be easily depicted in the identified classes by the semantic analysis classifier, shown in Table 6. Applying the LDA ontology-based filter, rigorously removed the non-related data or noises that is

Table 10
Analytical aspects of the Beginner Hacking forum.

CrimeBB dataset entity	Statistics
Total No. of users in the forum	74659
Total No. of posts in the forum	120454
Inactive users of the forum	4720
Users with zero (0) or no reputation in the forum	56886
Density of the forum	0.0049

Table 11
Network reduction and statistics results with INSPECT framework (LDA component).

Total number of posts: Initially in graph	120454
After applying LDA Filtering:	
Total Number of Posts(Edges in Graph)	57207
Average Posts per User after applying LDA Filtering	1.33

Table 12
Sample SNA (centralities measures) generated by INSPECT framework.

Degree centrality	Closeness centrality	Eigenvector centrality	Pagerank	Betweenness centrality
0.0000934142	0	2.848867967	0.0000279579	0
0.0007473	0.0004670	4.669×10^{-10}	0.000131916	4.3635226×10^{-8}
0.000116377 0	7.88×10^{-18}	7.42×10^{-06}	0	0
0.000698263	0.022047793	0.00825625	0.0000863	0.001645345
0.0000582	0	1.65×10^{-11}	0.0000454	0
0.000494603	0.01322453	0.0000676	0.0000106	0.0000648
0.00040732	0.000289232	2.31×10^{-10}	0.000100284	2.16×10^{-08}
0.00020366	0	7.88×10^{-18}	7.42×10^{-06}	0

embedded to perceive the fake influence or are helpful in fooling CTI algorithms and automated systems. The statistical analysis after implementing the pre-processing steps of the framework is illustrated in Table 11, given that Semantic analyzer has filtered almost 50% content marked to be irrelevant. Many studies like [3,12] have focused on the development of the semantic analysis based crawling mechanism to enrich CTI but validating the input sources and verifying the authenticity of the user profile is not addressed. INSPECT framework focuses on the identification of the influence of the user along with the semantic interpretation of their conversation, in order to verify the profile because utilizing its posted content for security enhancement purposes.

The resultant statistics given in Table 11 and the removed content, sample irrelevant posted given in Table 16, proved the efficiency and the effective classification strategy of the designed semantic classifier. Manual review of the network shows that only informative data and active users are retained in the network after semantic classification and filtration processes. To evaluate the effectiveness of the user's influence, the only ground truth available from the data is the *reputation* of the user. Although, the provided reputation is user based and rely completely on the individual intercommunication experiences, yet only comparable property is the user rating. The results of the LDA classifier over the respective forum significantly comply to the given ratings of the individual, which proves the effectiveness of the framework at this level.

Social Network Analysis (SNA) is the most crucial part of the designed framework as it analyses each user from five different perspectives and determine the exact location and social centrality of the user in the given network. To experiment with social network analysis techniques, the measures include *Degree Centrality* to calculate the direct connections, *Betweenness Centrality* which identifies the data flow connectors, *Closeness Centrality* to count down stepping nodes for expedite communication and information transformation, *Eigenvector Centrality* to measure influence by interpreting the surrounding nodes' importance and the *Page Rank* to determine the likelihood of the node. Sample results computed by SNA with the above given centralities are given in Table 12. Social network analysis computes the exact position of each of the user in the network.

Further on INSPECT has used the computed centralities of each of the user and calculate the influence of their profile in the forum with applicability of the linear regression. The ten highest influencing profiles in the network are given in Table 13 on the basis of their score of influence along with their reputation provided in the dataset. With the given results in Tables 13 and 14, it is proved that the reputation is just the aggregated score based on individual judgement and contradicts with the computed influence of the user, which is calculated on the semantic analysis and the social impact of the user.

For further evaluation, we have extracted the top 10 users with the highest reputation and their interlinked influence in the forum are given in Table 14 which shows that the user may have high reputation even with the low influence in the forum.

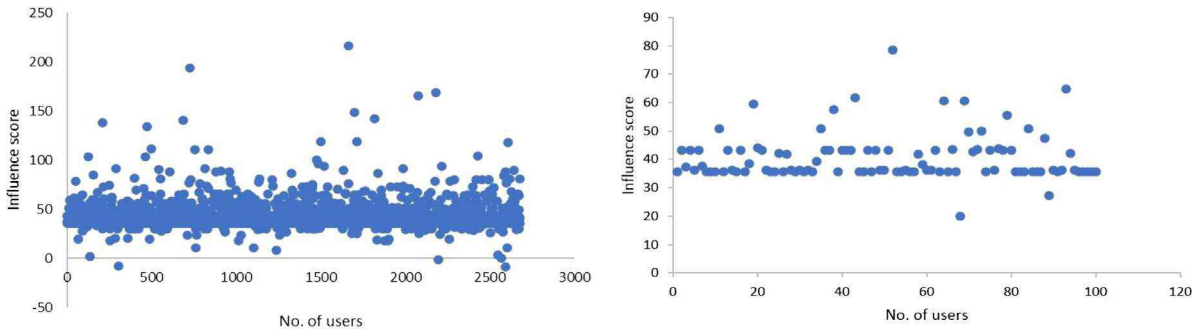
The analysis conducted using SNA signify an individual user's role and influence in the social environment. The computational results of the SNA describe the exact position of each user from five different angles and impact on other interconnected users. These results generate a baseline for computing the influential score of the user by incorporating the results of SNA with the Linear Regression algorithm. In parallel, clustering algorithm also subsequently follows the SNA computational results to determine the groups in the network.

Table 13
Top 10 influence scores and interlinked reputation of the users.

S.No	Score of influence	Reputation
1	216.2113832	−17
2	193.5286797	1
3	168.6974296	−3
4	165.5310948	4
5	148.5049773	1
6	142.091899	1
7	140.6749962	−3
8	137.658879	−10
9	133.7881543	650
10	118.4291162	−3

Table 14
10 highest reputation of the users and their computed influence scores.

S.No	Reputation	Score of influence
1	2818	36.23289493
2	2462	35.69027491
3	2328	44.80462414
4	2165	43.21198775
5	2071	56.26178173
6	2025	41.76766961
7	1871	35.69027491
8	1853	36.23289493
9	1840	41.76766961
10	1833	43.21198775



(a) Computed influence scores of each of the users (b) Computed influence scores of 100 random users

Clip 9. Influence scores for different user groups.

For calculating the influence score of the user in the given network, the algorithm of Linear Regression is mapped to the resultant scores of SNA techniques. As, Centrality measure is directly proportional to the influence factor, applying Linear Regression jointly on the subsequent results will generate the regression line and plot the influence scores of the users to determine the social impact and authenticity of the involved users. The proposed formula for computing score of influence is as follows

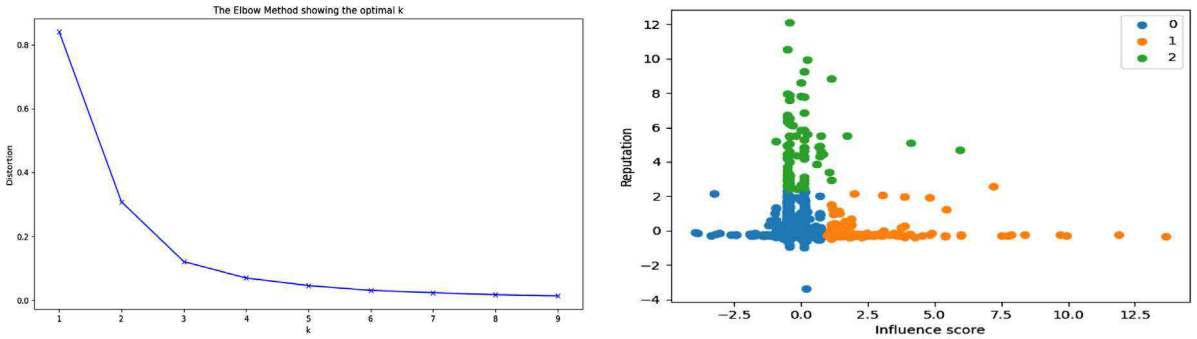
$$y = a + Bx \tag{6}$$

The influence rate (y) is dependent on the dependent centrality properties (x) and the Page Rank algorithm provided by the social network analysis. The direct and linear relation depicts that the user with high centralities must have high influence in the network. The computed score of influence depicts the profile and the impact of the user more appropriately than the user rating data. User rating is the average of the individual justification of others and does not attempt any formulated path. Nonetheless, the resultant influence score from this framework thoroughly investigated the footprints of the users, applied semantic analysis on the posted contents and identify the position of the user from various centrality angles in the network. The graphical interpretation of the influential score per user of the complete network is reflected in Clip 9.a, which depicted most of the users as common and some specific users with very high influence.

Clip 9.b, depicts the general determination of the users by providing the zoom-in reflection of initial 100 users of above Clip 9.a, which describes the common perspective of the forum.

Table 15
Comparative analysis of proposed work with relevant recent case studies.

S.No	Research	Dataset used	Semantic analysis	Social network analysis	Forensics users types	Computing user influence	Core intention
1	C. Chen et al. [13], 2021	DNM Corpus	To identify the potential topics discussed by user	Degree centrality, eigenvector centrality and local clustering coefficient are used	no	no	comparing and identifying replicated profiles
2	C. Peersman et al. [14], 2021	DNM Corpus	identify the theme of communication	Degree centrality, eigenvector centrality and local clustering coefficient are used	Enterpreneur, influencer and gatekeeper	no	analyzing the types of users
3	**This Study**	CrimeBB dataset	To filter the irrelevant content and identify the knowledge level of user	Degree centrality, closeness centrality, eigenvector centrality, betweenness centrality and pagerank algorithm are used	segregated users into three types	yes	computed the influence score of the user and identify the users groups



(a) Identified optimal number of clusters with Elbow method

(b) Clusters generated using K-Means clustering

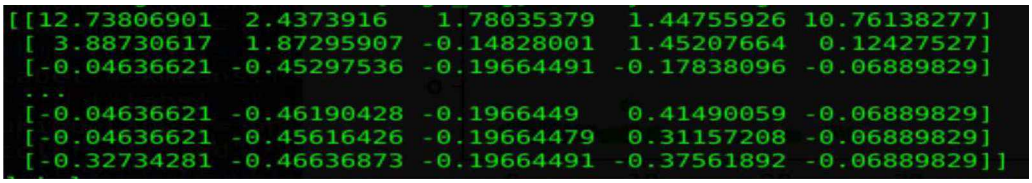
Clip 10. Influence scores for different user groups.

Comparing the proposed INSPECT framework with state-of-the-art research work is given in Table 15 which represents the current state-of-the-art analysis and justifies the intention of the INSPECT framework. The other two studies [13,14] have proposed their solutions using the DMN Corpus dataset collected between 2013–2015 while INSPECT framework is integrated with CrimeBB dataset that is updated and has crawled dark web data from decades up till 2017. Similarly, the social network analysis developed in existing literature only comprises of 2 centralities measures whereas this study interpret the user's position with five centralities paradigms. The most important point to focus is that we are unable to find the critically computed influence score of each of the users, which is the motive of this study. Tracing the existing literature and to the best of our knowledge, there is no existing study that calculates the influence score of the users based on the individual profile on the dark web forum. This makes the INSPECT framework novel.

This framework provides detailed analysis of the social environment and rigorously identified the traits, community and usage of the forum. To classify the community of the forum, K-means clustering is mapped to the subsequent results of calculated influence scores of the users which divides the registered users of the forum into three different groups. To identify the optimal number of clusters to be generated, the evaluation of K-Means clustering is obtained with the help of *elbow method*. Unlike supervised modeling strategies, K-Means does not have the availability of ground truth for analysis. Therefore, elbow method is used for identifying the adequate groups of users. The result obtained by implementing Elbow method is described in Clip 10.a.

With the results of the elbow method, given in graph 10.a, we have configured number of clusters to 3 and the clustering results can be visualize in Clip 10.b. The clusters are defined based on the generated influence of the users and the interlinked reputation of the users, provided by the CrimeBB dataset. To better interpret the clustering mechanism, the data is scaled with the **StandardScaler** and the sample of the resultant scaled data is given in Clip 11. The detailed synopsis of the identified users groups, in the form of clusters is given below:

1. **Cluster 0 - In Blue:** This cluster highlights the majority of the general users in the forum with low reputation and low influence in the community



Clip 11. Scaled influence score and reputation data to apply K-Means clustering algorithm.

Table 16
Samples of irrelevant posts, identified by INSPECT framework.

S.No.	Irrelevant posted contents
1	Great 10 year olds wanting to hack another 10 year olds email
2	Whats a good way to get stuff out of your victims :)
3	***CIT-ING***[https://hackforums.net/showthread.php?pid=22733851#pid22733851]***CITING***

Isnt that what hacking is about? lol getting infomation for fun, or to capitalize on it. AKA spying. |
| 4 | I would like this method. Thanks !!! |
| 5 | ***CIT-ING***[https://hackforums.net/showthread.php?pid=3838971#pid3838971]***CITING***

i bet i could find at least 75 post of this exact same question almost word for word.... "search" can be your best friend |
6	May I ask which RAT you are currently using? I may be able to assist you.
7	Look up some MSDOS Commands on Google, I'm sure you'll find some to your liking.
8	If there's any questions you need answered which aren't stated in the tutorials, throw me a PM.
9	Depends what you can bring to the table. I don't think we need any more blogs surrounding regurgitated exploits that are either old or already well-known.
10	Well I don't really consider it "hacking" but if you're looking for a quick buck you could try E-Whoring.

2. **Cluster 1 - In Orange:** This cluster justifies the INSPECT results and proves the mathematical interpretation of the user influence in the forum, either provided with low reputation score
3. **Cluster 2 - In Green:** These cluster points highlights those users with high individual reputation but intentionally have very low influence in the forum.

All the clusters are almost equivalent in size which reflects that almost 1/3 of the users are fraudulent and intended to be the influential identities in the community based on one-to-one interactions rather have very little to no influence at all. Around 30%–33% of the users are counted to be influential with semantic analysis and mathematical interpretation of the users position on the hacking ground with the very few outliers in the **cluster 1 - in orange**, that are the very high influencing profiles. The cluster points, given in **cluster 2 - in green**, justifies the investigation point of identifying the users influence based on their communication interpretation and knowledge level in the community rather than on the reputation that is only the individual scores. Furthermore, these clusters signify the majority types and behavior of the users in the forum and are helpful in identifying the intention of the forum.

6.3. Discussion

The pre-processor of the framework comprises of the semantic analyzer and the noise filter to remove the dormant and spam users and reduce the density of the network for further processing and accurate detection of the dominant user with prestigious profile and in-depth domain knowledge with respect to the community. The marked classes to remove the related data from the complete network worked accurately and remove all the text contents consists of the data with no impacted information. Sample text that is removed by the semantic classifier are shown in Table 16.

The dormant filter of the framework also removes the users from the network that are inactive and the users with no reputation set for them. These users are of no use and the density of the network reduced at large scale which efficiently follow-up with further computations.

Examining from the social context of the user with the help of SNA calculates the communication links and connection position in the network. This rigorous social analysis provides us the basis to calculate the influential score of the incorporated user. Applying linear regression over the SNA provided values, compute the impact factor of the user in the tested forum. With respect to the centralities measure, the influential score is more thorough and appropriate comparing to the user reputation given in the dataset. Some trivial values of the user social analysis and significance and the calculated authoritative measure are shown in Table 17, to provide a clear picture of the formulated value and the manual user analysis.

The mathematical indentation of each of the users provided by this research framework reflect the thorough and accurate position. This analytical measure depicts the profile knowledge and domination which will be useful for detaching the inappropriate

Table 17
Sample results generated by INSPECT of SNA, influence score and their inter-linked reputation.

Reputation	Degree centrality	Closeness centrality	Eigenvector centrality	Pagerank	Betweenness centrality	Influence score
67	0.000698263	0.022047793	0.00825625	0.0000863	0.001645345	16.59863404
70	0.0000582	0	1.65×10^{-11}	0.0000454	0	8.5869962
17	0.000494603	0.01322453	0.0000676	0.0000106	0.0000648	8.5869962
4	0.00040732	0.000289232	2.31×10^{-10}	0.000100284	2.16×10^{-08}	8.5869962
14	0.000320037	0.012193563	0.0000311	0.0000198	0.00107399	32.58067785

Table 18
Accuracy and effectiveness comparison of machine models for INSPECT framework use.

S.No.	Model	MSE	RMSE	Score
1	LinReg	44979.422900	212.083528	0.010000
2	KNN	47762.823474	47762.823474	0.050436
3	SVM	47587.877024	47587.877024	0.112391

and trivial users. In comparison to the reputation, this influence score is predicted efficiently and relevant. In this case study, regression model works more efficiently with respect to other classification models. The linear relation between centralities and the influence of user the is appropriately computed by incorporating linear regression algorithm. the algorithm of linear regression works more precisely and the results are less error-prone comparatively to the other scoring techniques as shown in Table 18.

7. Limitations of INSPECT framework and open challenges

This paper has presented our efforts with respect to using intelligent insights from dark web forums to strengthen defense against emerging cyber threats. There remain challenges in this research domain which we highlight below.

7.1. Limitations of INSPECT framework

Through the evaluation explained earlier in this paper, we have identified the following aspects of INSPECT framework which we aim to address in our future research.

- **Absence of ground truth** INSPECT framework is aimed at calculating the influence score of the users but no feature is provided by the dataset that can be marked as the ground truth to validate the effectiveness of the computed influence. In order to facilitate similar research, availability of such data is important.
- **Validation of hacker profile** The research presented in this paper is focused at identifying influential users of the dark web forums however further work is needed to validate user profiles to ascertain their influence on the dark web community.
- **Validation of groups developed through clustering** This study has identified the groups of the users in the forum based on their influence and the reputation provided by the dataset. This work can be strengthened by validating the groups identified.
- **Real-life applicability** The results of this study can be integrated and applicable to various real-life and security critical areas as defined in [4,5].

7.2. Open challenges in dark web

Some of the potential challenges that need to be considered to facilitate research utilizing dark web are highlighted below.

- **Standardization of approaches to conduct effective evaluation** The hidden structure of the dark web, randomize routing and its related components (forums, marketplaces), makes it difficult to define a standard mechanism or methodology to extract useful data to conduct evaluation across different approaches.
- **Challenges in access to meaningful data** The accessibility issues and the anonymous routing of the dark web makes it a challenge to gather the useful data from the dark web. Along with that, fake or spam data is included in the dark web which makes the extraction of the righteous data, a major challenge.
- **Ethical and legal concerns** Dark web has incredible volume of data which provides valuable intelligence in activities of malactors and can be used to learn about attacker behavior and strategies. However, accessing such data requires compliance with strict ethical and legal frameworks which are fundamental to cyber security research but sometimes can limit the depth of research that can be achieved.
- **Effective use of intelligence from dark web within protection systems** Leveraging outcomes of intelligent analytics of schemes such as the one presented in this paper, it is a complex task to integrate this intelligence within existing cyber defense platforms.

8. Conclusion and future work

Dark web is an important source of threat intelligence and can play a profound role in achieving proactive defense against emerging threats. The focus of research presented in this paper is to identify influential users (influencer hackers) within dark web forums which could facilitate intelligent mining of activities within such forums, leading to improved defense against cyber threats. Specifically, we have presented a novel framework (INSPECT) which utilizes clustering, social network analysis, different centralities, and Page Rank to calculate the *influencer score* of each user within a forum. The framework utilizes LDA and feature engineering techniques to remove noise from the data (users impersonating influencers). As demonstrated by the evaluation, the INSPECT framework is able to identify and rank users of dark web forums as per their influence whilst removing noisy and misleading data. We plan to continue this research by establishing linkage between influencer hackers and emerging exploits and malware shared on the dark web marketplaces.

Declaration of competing interest

No author associated with this paper has disclosed any potential or pertinent conflicts which may be perceived to have impending conflict with this work. For full disclosure statements refer to https://doi.org/10.1016/j.compeleceng.2023.108627.

Data availability

The authors do not have permission to share data.

Acknowledgments

The authors would like to pay gratitude to the Cambridge Cyber Crime Center, UK for the provision and accessibility to the CrimeBB dataset, which has been used to conduct this research.

References

[1] Kaur S, Randhawa S. Dark web: A web of crimes. vol. 112, (1). Springer Nature; 2020, https://www.researchgate.net/publication/338878596_Dark_Web_A_Web_of_Crimes.
[2] Nazah S, Huda S, Abawajy J, Hassan MM. Evolution of dark web threat analysis and detection: A systematic approach. 2020, https://ieeexplore.ieee.org/document/8756845.
[3] Schäfer M, Fuchs M, Strohmeier M, Engel M, Liechti M, Lenders V. Black-widow: Monitoring the dark web for cyber security information. In: International conference on cyber conflict. 2019, https://ieeexplore.ieee.org/document/8756845.
[4] Sutanrikulu A, Czajkowska S, Grossklags J. Analysis of darknet market activity as a country-specific, socio-economic and technological phenomenon. In: 2020 APWG symposium on electronic crime research (ECrime). 2021, https://ieeexplore.ieee.org/abstract/document/9493259.
[5] Bancroft A. Potential influences of the darknet on illicit drug diffusion. Springer; 2022, https://link.springer.com/article/10.1007/s40429-022-00439-2.
[6] Maras M-H, Arsovska J, Wandt AS, Knieps M, Logie K. The seci model and darknet markets: Knowledge creation in criminal organizations and communities of practice. Eur J Criminol 2022;1–26, https://journals.sagepub.com/doi/10.1177/14773708221115167.
[7] Zhang Y, Fan Y, Ye Y, Zhao L, Shi C. Key player identification in underground forums over attributed heterogeneous information network embedding. Graph Nerual Netw II 2019. http://cs.emory.edu/%7elzhao41/materials/papers/lp0110-zhangA.pdf.
[8] Samtani S, Chai Y, Chen H. Linking exploits from the dark web to known vulnerabilities for proactive cyber threat intelligence: An attention-based deep structured semantic model. MIS Q 2022;46:911–46, https://www.researchgate.net/publication/351426087_Linking_Exploits_from_the_Dark_Web_to_Known_Vulnerabilities_for_Proactive_Cyber_Threat_Intelligence_An_Attention-based_Deep_Structured_Semantic_Model.
[9] Sapienza A, Ernala SK, Bessi A, Lerman K, Ferrara E. Discover: Mining online chatter for emerging cyber threats. In: The third international workshop on cybersafety, online harassment, and misinformation. 2018, https://ieeexplore.ieee.org/abstract/document/8260822.
[10] Deliu I, Leichter C, Franke K. Collecting cyber threat intelligence from hacker forums via a two-stage, hybrid process using support vector machines and latent dirichlet allocation. IEEE Xplore 2018. https://ieeexplore.ieee.org/document/8622469.
[11] Zhang N, Ebrahimi M, Li W, Chen H. Counteracting dark web text-based captcha with generative adversarial learning for proactive cyber threat intelligence. ACM Trans Manag Inf Syst 2022;13:1–21. http://dx.doi.org/10.1145/3505226.
[12] Pannu M, Kay I, Harris D. Using dark web crawler to uncover suspicious and malicious websites. Adv Intell Syst Comput 2018;782:108–15, https://link.springer.com/chapter/10.1007/978-3-319-94782-2_11.
[13] Chen C, Peersman C, Edwards M, Ursani Z, Rashid A. Amoc: A multifaceted machine learning-based toolkit for analysing cybercriminal communities on the darknet. In: IEEE international conference on big data (Big Data). 2021, https://ieeexplore.ieee.org/document/9671906.
[14] Peersman C, Pencheva D, Rashid A. Tokyo, denver, helsinki, lisbon or the professor? a framework for understanding cybercriminal roles in darknet markets. In: APWG symposium on electronic crime research (ECrime). 2021, https://ieeexplore.ieee.org/document/9738782.
[15] Faisal MS, Daud A, Akram AU, Abbasi RA, Aljohani NR, Mehmood I. Expert ranking techniques for online rated forums. Comput Hum Behav 2018. https://www.researchgate.net/publication/330246782_Finding_Research_Areas_of_Academicians_using_Clique_Percolation.
[16] Du P-Y, Ebrahimi M, Chen NZH, Brown RA, Samtani S. Identifying high-impact opioid products and key sellers in dark net marketplaces: An interpret able text analytic approach. IEEE Explore 2019. https://ieeexplore.ieee.org/document/8823196.
[17] Jim M, Luo X, Zhu H, Zhuo HH. Combining deep learning and topic modeling for review understanding in context-aware recommendation. 2018, p. 1605–14, https://aclanthology.org/N18-1145.pdf.
[18] Machova K, Stefanik J. Authority estimation within social networks using regression analysis. 2017, https://link.springer.com/article/10.1007/s40595-016-0082-0.
[19] Akyazi U, van Eeten M, Ganan CH. Measuring cybercrime as a service (caas) offerings in a cybercrime forum. 2021, https://weis2021.econinfosec.org/wp-content/uploads/sites/9/2021/06/weis21-akyazi.pdf.
[20] Cabrero-Holgueras J, Pastrana S. A methodology for large-scale identification of related accounts in underground forums. 111, 2021, https://dl.acm.org/doi/abs/10.1016/j.cose.2021.102489.

[21] Koloveas P, Chantzios T, Alevizopoulou S, Skiadopoulos S. Intime: A machine learning-based framework for gathering and leveraging web data to cyber-threat intelligence. 10, 2021, p. 1845–54, https://www.mdpi.com/2079-9292/10/7/818.
[22] Rios SA, Aguilera F, Nunez-Gonzalez JD, Grana M. Semantically enhanced network analysis for influencer identification in online social networks. Neuro Comput 2017;326–327(6):50–9, https://www.sciencedirect.com/science/article/abs/pii/S0925231217315242.
[23] Pastrana S, Thomas DR, Hutchings A, Clayton R. Crimebb: Enabling cybercrime research on underground forums at scale. Creative Commons CC 2018;1845–54.
[24] Vu AV, Hughes J, Pete I, Collier B, Chua YT, Shumailov I, et al. Turning up the dial: the evolution of a cybercrime market through set-up, stable, and covid-19 eras. 2020, p. 551, 566, https://www.cl.cam.ac.uk/is410/Papers/turning_imc20_draft.pdf.
[25] Cai TT, Guo Z. Semisupervised inference for explained variance in high dimensional linear regression and its applications. J R Stat Soc 2020;82:391–419, https://rss.onlinelibrary.wiley.com/doi/10.1111/rssb.12357.

Anum Paracha is a Ph.D. student at the School of Computing and Digital Technology, Birmingham City University, UK. Her research interests are to investigate use of advanced machine learning techniques to mitigate emerging cybersecurity research challenges.

Junaid Arshad is an Associate Professor at the School of Computing and Digital Technology, Birmingham City University, UK. He received his Ph.D. in Computer Security from the University of Leeds, UK in 2011. His research interests include investigating security challenges for diverse computing paradigms such as distributed computing, cloud computing, IoT, and distributed ledger technologies.

Muhammad Mubashir Khan is a Professor in the Department of CS & IT NED University of Engineering and Technology, Karachi Pakistan. He is serving as Co-Principal Investigator of the National Center for Cybersecurity at NED University. His research interests include Network and Information Security, Cybersecurity, Blockchain and Quantum Key Distribution where he has published several research articles in prestigious journals.