Comparison of data fluctuations that lead to cyber security attacks: A comparison between surface, deep and dark net

Abstract:

The term "darknet" refers to the address space on the internet that is not being used, and users do not anticipate that this area will interact with their machines. Darknet is a source of cyber intelligence. In order to develop network security, it is necessary to conduct studies of the many dangers that comprise the network. In this research, we offer brand new machine learning classifiers that go by the name stacking ensemble learning. Their purpose is to evaluate and categorize darknet traffic. This novel approach employs predictions created by three different base learning techniques in order to deal with the issues relating to darknet attacks. The software was validated using a dataset that had more than 141,000 records and was derived from the CIC-Darknet 2020 database. The findings of the experiment indicated that the classifiers used in the investigation were able to easily differentiate between benign and malignant traffic. The classifiers have the ability to efficiently recognize known as well as unknown threats with a high degree of precision and accuracy that is greater than 99% in the training and 97% in the testing phases, with increments ranging from 4 to 64% based on the algorithms that are currently in use. As a consequence of this, the suggested system will become more reliable and accurate as more data is collected. Additionally, in comparison to other AI algorithms already available, the suggested system has the lowest standard deviation.

1. INTRODUCTION:

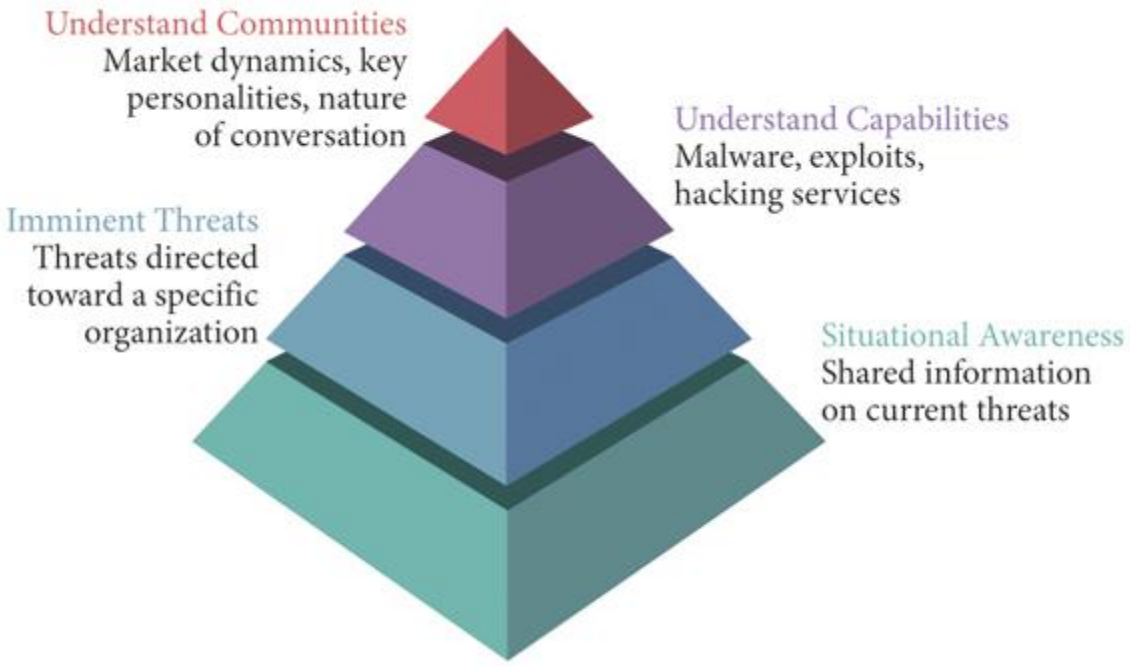
A portion of the Internet Protocol space that has been allotted and routed is referred to as a darknet or black web. These areas are home to servers and services that are currently not being used. It is possible that it will incorporate technologies that are undetectable and designed for the purpose of receiving messages. These kinds of systems do not appear to react to anything, and it is possible that they are part of an overlay network. Accessing this network may require using communication protocols and ports that are not considered standard.

It is believed that there is something fishy about the traffic that is heading into the darkness. It is possible for it to be malicious, but it is also possible for it to be a misconfiguration. The monitoring systems can be set up in darknets in such a way as to attract potential attackers so that intelligence can be gathered from them. The intelligence of botnets (Al-Nawasrah et al. 2018; Alieyan et al. 2018) and malware (Al-Kasassbeh et al. 2020) is frequently lacking. Sadly, some individuals will resort to unethical practices in order to boost both their link count and their reputation, such as the dissemination of fake news through the use of texts, images, and videos (Sahoo and Gupta 2021).

Figure 1 Dark Web Vs Surface Web Vs Deep Web

Black hole monitors, dark space, network telescopes (Pang et al. 2004), and false traffic are all names that have been used to refer to a darknet. Darknet traffic, in addition to being far less than genuine traffic, is known to carry evidence of malicious activities. It is to everyone's advantage to perform an analysis of this data in order to determine the pattern of attacks in the actual network. Each assault has its own unique strategy for making use of the data that is already present in the network. When these patterns are identified, it is much easier to track them down to the attacks and clustering that correspond to them. It has been demonstrated to be useful at recognizing patterns among unclassified data, such as traffic on the dark web.

Fig 2. Surface web model



1.1. Surface web:

This is the Internet that everyone of us has come to know and adore. The part of the internet that we access on a regular basis for things like social networking and reading the news is what some experts refer to as the Common Web. (If you are someone who believes in conspiracies, you might find it interesting to learn about the eight levels of the web, all the way down to Level 8, which is referred to as "The final boss of the universe.") In this case, traditional web spiders make use of complex algorithms in order to collect data from hyperlinked pages, and you may peruse this data using search engines like Google or Yahoo.

In a strange twist of fate, parts of the Deep Web can be found on the Surface Web. Because search engines are unable to access the content of websites that require credentials to access, these websites are technically considered to be part of the Deep Web. What about a website like Facebook, which requires users to check in before access can be granted, but which nonetheless generates pages that can be indexed by search engines even when a user is not signed in? In 2007, Facebook began providing search engines with access to some of its listing data. In

addition, links on Facebook can be indexed; for instance, if there is a page anywhere that includes your Facebook URL, a search engine WILL be able to index the page in question.

- In some circles, the term "Clearnet" also refers to the "Visible Web" and "Indexed Web."
- A portion of the entire internet consisting of more than nineteen terabytes of information
- A few examples of the drawbacks are as follows: There are trolls and stalkers; there is exploitation, pornography, and violence; it is addictive, distracting, and a waste of time; there is identity theft and hacking; there is spam, advertising, and an invasion of privacy.

1.2. Deep Web

It is a common misconception that criminals, drug users, terrorists, and sexual deviants use the Deep Web as a playground. The intranet of a firm is comparable to the deep web in the sense that no one from the outside of the company may access the information that is contained within it.

- Approximately 95% is available to the general public. The contentious situation involving WikiLeaks brought to light information that had been hidden on the Deep Web for a number of years.
- Inaccessible to search engines that are considered "normal" You will need a specialized browser, such as TOR, in order to access the Deep Web. The Onion Router (TOR) is by far the most well-known portal, but there are others such as Freenet, GNUnut, and Hotspot Shield. The United States Naval Research Laboratory created the Tor network in the 1990s in order to make it easier for individuals within the company to communicate anonymously with one another.
- The Hidden Wiki is the primary gateway to the Deep Web; nonetheless, you should approach it with extreme caution.
- "Normal" search engines like Google are unable to index websites in their databases. DeepPeep and IncyWincy are two of the most well-known search engines on the Deep Web.
- It is possible to remain anonymous. In point of fact, many government agencies keep a close eye on specific websites. There is also the possibility of being attacked by computer hackers.

- The value of illegal activities is around one hundred million dollars. Check out this page to see an infographic that was developed by Norwich University. According to The SSL Store, the amount of money generated by cybercrime in 2018 will be at least $1.5 trillion, and this estimate is on the low end.
- Your emails and the information about your online banking are examples of Deep Web content, which refers to information that is not accessible to the general public and is not indexed by search engines.
- Even Facebook has its own address on the Deep Web, which users in countries where Facebook is blocked can use to access the website. Whenever you log in there, the anonymous Tor IP address that you have will be used.
- A repository of information that is not widely known, as well as a host for government databases, legal and medical documents, and websites that have not yet been indexed. For example, SciHub is a repository for more than 50 million scientific research publications that may be downloaded for free.

1.3. Dark web

In the same way that readers of regular news outlets are drawn to ugly accounts of genuine crimes and accounts of disasters, the Dark Web possesses a certain allure that compels users to explore its depths. It is frequently confused with the Deep Web and is sometimes referred to as the Deep Web's malevolent counterpart. There is a plethora of lurid folklore that circulates about the horrible, bloodcurdling things that take on "down there." However, if the whole truth were known, it is not entirely negative. The "Dark Web" is a term that refers to the portion of the internet that allows users to remain anonymous, do business in private, and access information that is not readily available to the general public via traditional means. Some people who are fed up with the intrusive marketing strategies of firms like Google or Facebook turn to the Dark Web because it enables them to conceal their search behaviors, regardless of whether or not those behaviors involve illegal activity.

- According to estimates provided by security professionals, there are between 10,000 and 100,000 active sites at any given time.
- Also referred to as the Dark Web, the Deep Web, the Invisible Web, or the Hidden Internet

- Criminals take use of the anonymity provided by the Dark Web in order to sell firearms, drugs, and even human beings; but, governments and the United Nations also use it in order to safeguard political dissidents and chase criminals.
- If a data breach caused your personal information to become public, it will be offered for sale on this website.
- White hat hackers and law enforcement have made inroads tracking down criminals, including Ross Ulbricht, an American former drug trafficker and Darknet market operator who created and ran the Silk Road website, an underground marketplace primarily for drugs, from 2011 until his arrest in 2013. In 2014, the FBI was responsible for the shutdown of Silk Road.
- Enables the free and open discussion of ideas in an environment free from censorship
- Facilitates the collection of threat intelligence, such as information regarding the planning of terrorist attacks, money laundering, and other types of illegal activity
- Ensures the safety of journalists as well as the sources they cite
- Gives users the chance to access data from a variety of sources, giving them an advantage over their competitors when making business decisions and determining emerging technological and market trends. If you have the appropriate tools, you will be able to track the activities of your rivals and discover what your clients and employees are saying about you behind the scenes, despite the privacy settings they have in place.
- Access is permitted, but any illegal behavior may result in legal repercussions.

2. Comparison Of dark, Deep and Surface Web:

The terms "surface web," "deep web," and "dark web" are often used to describe different layers of the internet, each with distinct characteristics. Let's explore the differences between them:

2.1. Surface Web:

- **Accessibility:** The surface web refers to the portion of the internet that is indexed by traditional search engines (e.g., Google, Bing, Yahoo). It's easily accessible and includes websites that are meant for public consumption.
- **Content:** This is where most everyday internet users operate. Websites like social media platforms, news sites, blogs, and e-commerce sites are part of the surface web.
- **Examples:** Google, Facebook, Wikipedia, Amazon, and most other websites that you access through search engines.

2.2. Deep Web:

- **Accessibility:** The deep web comprises content that is not indexed by standard search engines. It includes password-protected sites, databases, and other content that is not meant for public consumption.
- **Content:** This part of the internet is often associated with private or proprietary databases, academic resources, email accounts, and other content that is not freely accessible to everyone.
- **Examples:** Online banking, private email accounts, subscription-based academic databases, and other restricted-access sites.

2.3. Dark Web:

- **Accessibility:** The dark web is a small, intentionally hidden part of the deep web. Accessing the dark web typically requires specific software, such as Tor (The

Onion Router), which anonymizes users and allows them to access ".onion" websites.

- **Content:** The dark web is known for its anonymity, and it hosts a range of legal and illegal activities. While it has legitimate uses, it is often associated with activities like illegal drug trade, hacking services, and other illicit transactions.
- **Examples:** Marketplaces like Silk Road (historically, as it was shut down), forums for hackers, whistleblowing platforms like WikiLeaks, and various other hidden services.

Table 1. Comparative overview of different web

	Surface Web	Deep Web	Dark Web	Darknet
Description	Content that search engine can find	Content that search engine cannot find	Content that are hidden intentionally	-
Known as	Visible web, Indexed web	Invisible web, Hidden web, Deep web	-	Underbelly of internet
Constitutes	Web	Web	Web	Web
Contents	Legal	Legal + Illegal	Illegal	Network
Information found	4%	96%	-	-
Browsers	Google chrome, Mozilla firefox, opera	-	TOR browser	Freenet, Tor GNUnet, I2P, OneSwarm, RetroShare

It's important to note that while the dark web is often portrayed in a negative light due to its association with illegal activities, it also serves as a platform for individuals in oppressive regimes to communicate freely and for whistleblowers to share information without fear of retaliation. The key distinction lies in the level of anonymity and the intent of use on these different layers of the internet.

3. Comparison of Data Fluctuations That Lead To Cyber Security Attacks:

Data fluctuations in the context of cybersecurity refer to changes, variations, or anomalies in the normal patterns of data flow and behavior within a system or network. These fluctuations can sometimes indicate potential security threats or attacks. Here's a comparison of different types of data fluctuations that may lead to cybersecurity attacks:

1. **Traffic Spikes:**
    - **Normal Fluctuation:** Occasional spikes in network traffic may be normal, such as during peak usage times.
    - **Cybersecurity Concern:** Sudden and unexplained spikes could indicate a distributed denial of service (DDoS) attack, where an attacker overwhelms a system with traffic to disrupt its services.
2. **Unusual Login Patterns:**
    - **Normal Fluctuation:** Users may log in from different locations or at varying times.
    - **Cybersecurity Concern:** Rapid and unexpected changes in login patterns could suggest unauthorized access, potentially due to a compromised account or a brute force attack.
3. **Data Access Patterns:**
    - **Normal Fluctuation:** Different users and systems may access various data at different times.
    - **Cybersecurity Concern:** Abrupt changes in data access patterns may indicate a data breach or an insider threat.
4. **System Resource Usage:**

- **Normal Fluctuation:** System resource usage can vary based on user activity and workload.
- **Cybersecurity Concern:** Unusual resource spikes might signify a malware infection or a malicious process consuming excessive resources.

5. **Outbound Network Traffic:**
    - **Normal Fluctuation:** Outbound traffic can vary based on legitimate activities like software updates.
    - **Cybersecurity Concern:** Unexpected outbound traffic may indicate data exfiltration by malware or a compromised system communicating with a command and control server.
6. **Configuration Changes:**
    - **Normal Fluctuation:** System configurations may change due to updates or legitimate administrative actions.
    - **Cybersecurity Concern:** Sudden and unauthorized configuration changes can be a sign of an attack, such as privilege escalation or an attempt to create a backdoor.
7. **User Behavior Anomalies:**
    - **Normal Fluctuation:** Users may exhibit different behaviors over time.
    - **Cybersecurity Concern:** Drastic changes in user behavior, especially in privileged accounts, may indicate an account takeover or an insider threat.
8. **Data Integrity Changes:**
    - **Normal Fluctuation:** Changes in data integrity may occur due to legitimate updates or modifications.
    - **Cybersecurity Concern:** Unexpected and unauthorized changes in data integrity could be a sign of a cyberattack, such as data manipulation or ransomware.

9. **Anomalous System Events:**

    - **Normal Fluctuation:** Systems generate various logs and events during regular operations.
    - **Cybersecurity Concern:** Unusual or multiple system events may indicate an ongoing attack, especially if they involve known attack vectors or signatures.

Monitoring and analyzing these fluctuations through advanced threat detection systems, anomaly detection algorithms, and security information and event management (SIEM) tools can help organizations identify and respond to potential cybersecurity threats more effectively.

4. Challenges and Ethical Concerns:

In the linguistics domain, Ferguson addressed some significant challenges when studying the Dark Web content:

(1) Inconsistency of the language used in communications between community members and forum discussions; this inconsistency is intended in Dark Web communities as a type of anonymity procedure.
(2) Weak grammatical, spelling, and idiomatic context (also intended).
(3) Individuals deliberately do not use particular terms or use them only in specific cases and ways.
(4) The cultural dynamics of Dark Web communities: members come from worldwide; thus, they do not follow standard terminology or normative cultural context to contribute to the community.

Similarly, Queiroz and Keegan indicated that hackers use constantly changing and evolving technical terms that contain semantic differences, in addition to abbreviations and misspellings, which require frequent development of the analysis model to keep pace with these changes. Moreover, it urges to adopt different modeling approaches for each social network; in other words, the model developed for a network may not perform similarly on another network due to terms changes. In another work, Queiroz et al. justified the notion of "Concept Drift" caused by

the mentioned changes in hackers' terms. Furthermore, they introduced an approach to overcome this drift by updating and retraining the model with temporal features and weighting.

Queiroz and Keegan [4] added two more challenges in the CTI field. One is the lack of ground-truth datasets that researchers need to evaluate their modeling approaches and validate their results. The second are the ethical considerations when dealing with the data. Unlike common social media platforms (such as Facebook and Twitter), there is no explicit agreement in hacking forums and chat rooms explaining to the user that their data may be used by third parties (such as researchers). Additionally, the sheer volume of data makes it difficult to obtain explicit consent for the use of participants' data in research. These considerations call for researchers to make careful decisions about how to use the acquired data.

In the technical particularity of the Dark Web, Akhgar et al. addressed the following challenges:

(1) The nature of the web in general: the web consists of different types of media besides textual data, most commonly image, video, and audio.
(2) The published multimedia is in different languages and colloquialisms or accents, using different terminologies.
(3) The complexity of accessing criminals' social networks and closed groups: investigators often need to wait several weeks before obtaining approvals to join these networks. Moreover, they need to make their profiles look authentic, and their stories sound realistic and believable by administrators of the websites under study.

Due to the technical nature of the Dark Web, developing crawlers that collect and analyze the required data can be complicated. Furthermore, researchers must consider efficient precautionary measures since their employed techniques and tools themselves face the risk of being disclosed and vulnerable to cyberattacks.

In particular, Pastrana et al. discussed the ethical issues when collecting and analyzing data from underground forums. Ethical considerations require research studies involving human participants to be reviewed by a Research Ethics Board (REB). The importance of such reviews

is to consider the potential harm, how to reduce or avoid consequences, and protect the researchers from possible responsibilities. Moreover, Pastrana et al. differentiate ethical issues of collecting the data from analyzing the data. Their justification for this separation is due to the nature of each process. Collecting the data is to understand forum behavior as a computer system, whereas analyzing the data involves understanding human beings related to the collected data. In the former, researchers should consider some technical risks such as breaking terms of services of the platform or overcoming crawling prevention measures like CAPTCHAs. They suggest that if the benefits surpass the potential harms, it is ethically reasonable to break such measures. On the other hand, using TOR for research purposes cannot avoid making the researcher's device itself a relay on the network.

Researchers can consider several measures to mitigate potential harm :

- Avoiding identification of individuals (such as publishing their usernames)
- Introducing the results objectively
- Avoiding the disclosure of sensitive personal data (like credit card numbers of victims)
- Protecting the researcher: for example, by avoiding making comments that offend the community and taking precautions not to download malicious content, which can cause security or legal issues, such as malware, child pornography, or terrorist materials
- Hiding the name of the platform from which the researcher collected and analyzed the data
- The cultural dynamics of communities on the Dark Web: Members of these communities originate from all over the world; as a result, they do not adhere to conventional language or normative cultural context when contributing to the community.

In a similar vein, Queiroz and Keegan highlighted that hackers employ continually changing and evolving technical terminology that contain semantic variances, in addition to abbreviations and misspellings, which necessitate periodic improvement of the analytic model to keep pace with these changes. In addition to this, it strongly recommends using distinct modeling methodologies for each social network. To put it another way, the model that was established for one network might not perform similarly on another network due to variations in the terminology. In a different piece of research, Queiroz et al. justified the concept of "Concept Drift," which was

created by the changes in hackers' words that were previously described. In addition, they came up with a strategy to combat this drift by continually updating and retraining the model with different temporal characteristics and weighting.

Queiroz and Keegan introduced two further problems to the CTI field in their research. One problem is that academics do not have access to enough ground-truth datasets, which they require in order to analyze and confirm the outcomes of their modeling efforts. The second factor is to take into account any ethical concerns when working with the data. There is no explicit agreement in hacking forums and chat rooms stating to users that their data may be used by third parties (such as researchers). This is in contrast to other social media platforms, such as Facebook and Twitter, where users are required to agree to terms before using the network. In addition, the sheer amount of data makes it difficult to acquire participants' express consent for the use of their data in study. In light of these factors, it is necessary for researchers to deliberate carefully over the best way to utilize the data that has been gathered.

In regards to the technical peculiarities of the Dark Web, Akhgar et al. tackled the following difficulties:

- The characteristics of the internet in general: in addition to textual data, the internet also contains other forms of media, the most frequent of which are audio, video, and still images.
- The published multimedia consists of a variety of linguistic styles, including colloquialisms, accents, and terminology.
- The difficulty of accessing criminals' social networks and closed groups: In most cases, investigators have to wait several weeks before they are granted authority to join these networks.

In addition, the administrators of the websites that are the focus of this research need to be convinced that their users' profiles are genuine and that the stories they tell are credible.

It can be difficult to design crawlers that are capable of collecting and analyzing the necessary data due to the technical nature of the Dark Web. In addition, researchers need to give careful

consideration to the most effective preventative measures because the strategies and technologies they use themselves run the danger of being made public and are susceptible to cyberattacks.

In particular, Pastrana et al. talked about the ethical problems that can arise while gathering and evaluating data from underground forums. Research projects that include people need to go under the scrutiny of a Research Ethics Board (often abbreviated as REB) before they can be published. The relevance of such evaluations lies in the fact that they take into consideration the potential for harm, offer suggestions on how effects can be minimized or avoided altogether, and shield the researchers from any potential liabilities.

In addition, Pastrana et al. distinguish between ethical concerns about the collection of the data and those regarding the analysis of the data. They say that this split is necessary due to the different characteristics of each phase. The purpose of collecting the data is to gain a knowledge of the behavior of the forum as a computer system, but the purpose of analyzing the data is to gain an understanding of human beings in relation to the acquired data. In the first scenario, researchers need to be aware of some technical dangers, such as violating the terms of service of the platform or bypassing crawling prevention mechanisms such as CAPTCHAs. They suggest that if the benefits of breaking such regulations are greater than the potential costs, then it is morally acceptable to do so. When using TOR for research purposes, on the other hand, it is impossible to prevent the researcher's device from becoming a relay on the network.

Researchers have a few options to choose from when it comes to mitigating potential risks:

- Preventing the identification of specific people, including the publication of their online handles.
- Presenting the findings in an objective manner; (3) Preventing the leaking of sensitive personal data (such as the credit card details of victims).
- Protecting the researcher by, for example, avoiding making remarks that are offensive to the community and taking steps not to download dangerous stuff that can cause security or legal difficulties, such as malware, child pornography, or terrorist materials.

- Hiding the name of the site from where the researcher gathered and analyzed the data. (6) Using caution while evaluating leaked data, as it may contain private messages, email addresses, IP addresses, and exclusive posts.

Conclusions:

Understanding and analyzing data fluctuations are crucial components of a proactive cybersecurity strategy. By comprehensively assessing data volume, velocity, variety, veracity, and security policy compliance fluctuations, organizations can fortify their defenses and stay one step ahead of cyber threats. Continuous monitoring, advanced analytics, and a commitment to cybersecurity best practices are integral in safeguarding the digital assets of businesses and individuals alike. Understanding the distinctions between the Surface, Deep, and Dark Web is crucial for navigating the internet responsibly. While the Surface Web is the most visible and user-friendly, the Deep and Dark Web layers offer increased privacy but come with their own set of risks and challenges. Users must exercise caution and awareness while exploring the diverse layers of the internet to ensure a safe and secure online experience.

References

Al-Kasassbeh M, Mohammed S, Alauthman M, Almomani A. Feature selection using a machine learning to classify a malware. Handbook of Computer Networks and Cyber Security: Principles and Paradigms. 2020:889-904.

Sahoo SR, Gupta BB. Multiple features based approach for automatic fake news detection on social networks using deep learning. Applied Soft Computing. 2021 Mar 1;100:106983.

Al-Nawasrah A, Al-Momani A, Meziane F, Alauthman M. Fast flux botnet detection framework using adaptive dynamic evolving spiking neural network algorithm. In2018 9th international conference on information and communication systems (ICICS) 2018 Apr 3 (pp. 7-11). IEEE.

Pang PT, Lu B. Regulation of late-phase LTP and long-term memory in normal and aging hippocampus: role of secreted proteins tPA and BDNF. Ageing research reviews. 2004 Nov 1;3(4):407-30.