Cyber Attack Taxonomy for Big Data Environment

Keerti Dixit
Institute of Computer Science
Vikram University, Ujjain

Umesh Kumar Singh, PhD
Institute of Computer Science
Vikram University, Ujjain

Bhupendra K. Pandya, PhD
Institute of Computer Science
Vikram University, Ujjain

ABSTRACT
Data is growing rapidly in the contemporary digital era from a variety of sources, including banking, enterprises, education, entertainment, etc. Because of its profound impact, it became a well-known method for several study fields, including semantic web, machine learning, computational intelligence, and data mining. Several corporate sectors rely on tweets, blogs, and social data to get proper analysis for information extraction. They are able to predict customer interests and preferences and use resources more effectively. Sometimes the same data causes issues that result in a problem known as big data. In this research paper we have addressed characteristics and various stages involved in big data. Further, we have developed cyber attack taxonomy for big data environment.

Keywords
Big data, attacks.

1. INTRODUCTION
The amount of data in the globe is increasing on a daily basis. Because of the widespread use of the internet, smartphones, and social media, data is rising. The word "big data" is increasingly widely used in our daily lives. Big data is a word coined in 2005 to define a broad range of bulky data sets that are about challenging to handle and analyze with typical data management tools due to their size and complexity. Big data is generated by a range of websites, social networks, multimedia archives, and IoT networks which connect a wide range of devices and sensors. Big data is gaining traction in a variety of industries, including telecommunications [1, 2], healthcare [3, 4], and many more, as a result of its enormous advantages. Big data has recently become a trendy topic with big implications, impacting industries all over the world. Big data analytics is viewed like cutting-edge and helpful approach for analyzing data that is both sophisticated and historical to uncover patterns that might aid in efficient decision-making by businesses and government agencies. In numerous industry sectors big data will play an essential role in future data management and operations [5, 6]. Several investigations have discovered various advantages of big data applications. Modern literature assessments on big data security, on the other hand, show that mischievous attackers targeting large data are increasing [7]. In big data area, the main concerns around privacy protection and security risk have yet to be thoroughly investigated [8, 9]. These problems drive fresh ideas and research to uncover unsolved problems that clear the path for more research and practise in the future [10, 11].

2. CHARACTERISTICS OF BIG DATA
Big data is a concept which is still being defined. It refers to any extensive volume of structured, unstructured and semi-structured data that can be mined for useful information.

The three Vs were primarily used to define big data [12]:

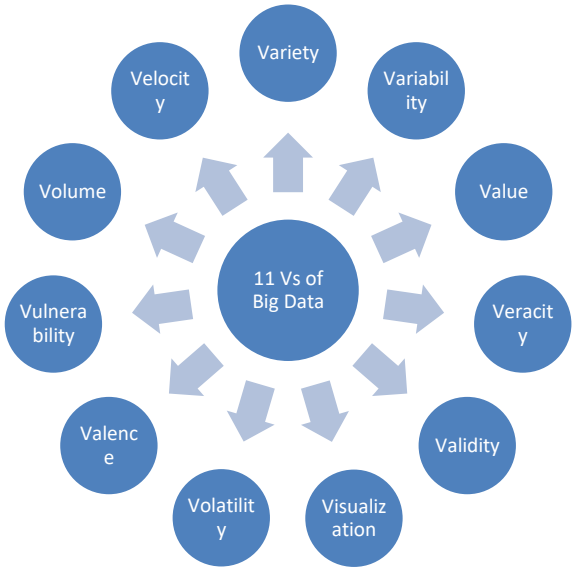

Fig. 1: Characteristics of Big Data

2.1 Volume:
Organizations get data from various sources. Previously, keeping such a massive volume of data would have been a challenge, but the advent of modern technologies like Hadoop has eased the strain. Big data has an impact on security and privacy in at least two ways, as noted below:

- Software tools and Traditional database systems are incapable in frequently monitoring and enforcing standard security procedures since data is stored in numerous locations in a distributed fashion.
- Any failure of a node or a cluster can have an impact on performance and data transactions in the cop-up time limitations, as well as expose security risks.

2.2 Velocity:
Data is generated at a rapid rate and must be processed quickly. The need to deal with massive amounts of data in real time drives RFID tags, sensors, cell phones, and smart metres; the velocity of big data has an influence on privacy and security since quicker cryptographic algorithms should be synchronized with real-time transaction processing. Furthermore, to maintain past data track by enforcing privacy policies which correspond to the large rate of data acquisition, security audits are needed.

2.3 Variety:
From organized and numerical data in relational databases to unstructured data, data is generated in a variety of formats. Data classification and access controls are necessary for various data sources as there is diversity of big data which has a greater influence on security and privacy.

Nonetheless, the big data notion has grown to include eight new aspects throughout time [13-21]:

2.4 Variability:

Data flows can be highly erratic, with frequent peaks, including rising speed and variety of data. Seasonal peaks or peak events can be difficult to manage on a daily basis, especially when dealing with unstructured data; variability characteristics of big data are accountable to IT security operations within different audit log collecting and in vigilance systems.

2.5 Value:

It's also important to remember that big data has to be valuable. In this regard, it is critical to ensure that the insights created are based on reliable data and lead to demonstrable changes; big data value is a critical aspect, and suitable more restrictions and permissions up to detailed evaluations are essential. It's crucial to establish proper security controls during the generation of such data insights.

2.6 Veracity:

It is critical to ensure high data quality and record data provenance in terms of inputs, systems, entities, and factors that affect data of interest; by increasing the big data veracity, decision-making risk can be reduced. This effects on privacy and security regulations, in terms of ensuring high data quality, by the adoption of suitable data control and timely access analysis mechanisms.

2.7 Validity:

Valid data is necessary for making accurate and timely judgments. As a result, we should avoid using corrupted data in data analysis. Because many organizations spend a significant amount of time cleaning data before doing data analysis, appropriate data governance policies are essential to ensure validity. This necessitates appropriate management of third-party vendors and associates who enforce data supply chain security.

2.8 Visualization:

The facts must be presented in a way that is both intuitive and visually appealing. Indeed, as we accumulate more data, the capacity to display it becomes increasingly vital, as it would otherwise be difficult to intuitively discern patterns or correlations between the data. Many advanced visualization technologies are now being combined with data analytics models to produce more relevant graphical interpretations of huge data, allowing for more effective decision making. Along with establishing access controls and privileges depend on responsibilities and roles of users, privacy and protection policies for the outputs from distinct visualization tools must be developed.

2.9 Volatility:

"Big Data Volatility" relates to the length of time that data will be valid and how long it should be stored. The privacy and security rules and processes for data destruction, preservation and timely re-assessment of security approaches are impacted by the volatility of big data.

2.10 Valence

It's more harder to find indirect relationships between data pieces, but they provide value to the business. The tenth dimension of big data, called Valence, is the outcome of these interconnections, which are comparable to atoms in a molecule. It is a unit that measures the density of data, and a unit of Valence is calculated as the ratio among number of data items that are actually connected and the connections which may be made in the data collection. For ongoing and forthcoming increase of big data systems, security and privacy management methods should retain a high degree of performance.

2.11 Vulnerability:

The Vulnerability aspect of big data is the last but most essential dimension, which refers to the privacy, security, and technical risks associated with diverse rich individualized data obtained by products and services adopting Internet apps, social networks, and IoT devices. Big data technology, procedures, and management are vulnerable due to security and privacy breaches, as well as a lack of standards. Recently reported breaches of big data privacy and security have sparked a lot of interest in the Vulnerability dimension. The earlier-mentioned dimensions, which necessitate ongoing monitoring, drive security and privacy policies and processes for incident management with relation to big data. Vulnerability checks and penetration testing will be conducted on a regular basis, taking into account the special characteristics of big data. The risks of vulnerable data leaking must be identified, and adequate procedures to ensure the confidentiality, integrity, and availability of big data systems and data must be implemented.

3. STAGES INVOLVED IN BIG DATA

3.1 Data Acquisition:

The acquisition of data is the initial stage towards Big Data. The rate of data creation is increasing exponentially as the medium grows. Smart gadgets, which have different types of sensors, generate data on a constant basis. The majority of this data is useless and may be eliminated; but, due to its unstructured nature, removing the data carefully is difficult. When this data is merged with other useful data and overlay, it becomes robust. Data is rapidly being collected and stored in the cloud as a result of the interconnection of devices through the Internet.

3.2 Data Extraction:

All of the data that has been generated and collected is unnecessary. It has a lot of info that is redundant or unimportant. A rudimentary CCTV camera, for example, polls sensors constantly to obtain information about the user's movements. The data supplied by the activity sensor, on the other hand, is redundant and useless when the user is inactive. The issues in data extraction are twofold: first, because of the nature of the data collected, determining which data to preserve and which to reject is becoming increasingly dependent on the generated data. For example, similar film from a security camera having same frames can be discarded, while similar data provided by a heart-rate sensor should not be discarded. Second, the lack of a unified platform poses its own set of problems. Because there is such a vast variety of data, putting it all together on a single platform to standardize data extraction is a huge difficulty.

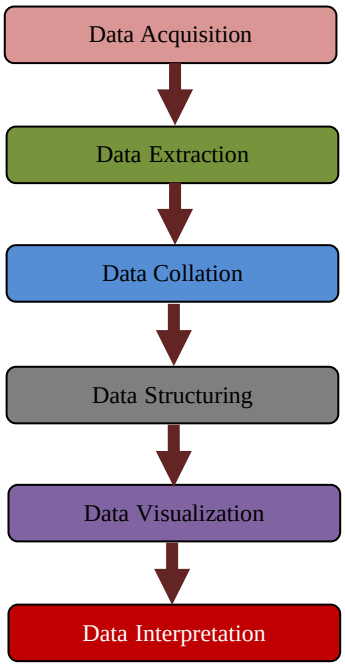

Fig. 2: Stages in Big Data

3.3 Data Collation:

For analysis or prediction, data from a single source is frequently insufficient. Multiple data sources are frequently merged to create a larger picture to study. A health monitor application, for example, frequently pulls data from the heart-rate sensor, pedometer, and other devices to summarise the user's health statistics. Similarly, weather prediction software gathers data from a variety of sources to indicate daily humidity, temperature, precipitation, and other factors. Convergence of data to generate a larger picture is typically regarded a critical aspect of processing in relation to Big Data.

3.4 Data Structuring:

After all of the data has been aggregated, it is critical to present and store the data in an organised fashion for future usage. It's crucial to structure the data so that queries can be run on it. Data structuring is the process of organising data in a specific structure. Various new systems, such as NoSQL, are rapidly being employed for Big Data Analysis since they can query even unstructured data. Because giving real-time results is a major concern with big data, the structuring of aggregated data should be done quickly.

3.5 Data Visualization:

After the data has been formatted, queries are run on it, and the output are shown in a visible manner. Data analysis is identifying areas of interest and producing outputs depend on structured data. For example, to calculate a relationship between average temperatures and water consumption rates, data comprising average temperatures is displayed alongside water consumption rates. This data analysis and presentation prepares it for consumer consumption. Because we cannot use raw data to acquire observationss or judge patterns, “humanizing” the data grow into even more critical.

3.6 Data Interpretation:

The final phase in Big Data processing is interpretation, which involves extracting useful information from the processed data. There are two forms of information that can be obtained: Retrospective analysis entails gaining knowledge of previously occurring events and behaviours. For example, statistics on a show's television viewership in various places can assist us in determining the show's popularity in those areas. Prospective analysis entails evaluating patterns and predicting future trends based on previously collected data. Prospective analysis is used to predict the weather using big data analysis. Problems that arise as a result of such interpretations include the prediction of false and misleading trends. This is especially problematic because crucial decisions are becoming increasingly reliant on data. In case a specific evidence is devise in contrast to the likelihood of being determined with a certain disease, for example, it may lead to erroneous information about the symptom being caused by the disease itself. As a result, data interpretation insights are crucial, and they are also the major purpose to deal with huge data.

4. CYBER ATTACKS ON BIG DATA ENVIRONMENT

Cyber-based attacks on big data environments are increasing at an exponential rate, affecting local and remote systems and causing considerable costs by stealing and destroying essential infrastructure. Attackers may take advantage of flaws in the database and utilize them to compromise the system's security. Because databases contain highly private information about organizations, it is critical to ensure that they are secure. The picture depicts the developed attack taxonomy for big data environments.

4.1 Injection Attacks:

These attacks typically target flaws in input validation or query input parameters that haven't been sanitized. In the big data world, there are following types of injection attacks [22]:

- Server side injection (SSI): By injecting scripts into HTML pages or remotely running inconsistent instructions, a server side injection attack allows for the exploitation of a web application. The attacker uses user input fields to influence the user into using the attacker's programme. In case the web server allows server-side injection without real validation does the attack succeed. SSI instructions are placed in input fields and transmitted to the web server. The web server parses the page before supplying it, and then it runs the directives. After the page is uploaded from browser, the attack result is displayed.
- NoSQL Injection: With harmful inputs, NoSQL Injection performs a critical role in online applications that change the performance of query structures. This changed query at run time could lead to security issues like unauthorized access to programme resources and changes to sensitive data.

Cross-site scripting: One more kind of injection attack is cross-site scripting, which puts mischievous script in legitimate websites. When a user visits a specific website, mischievous scripts can be present on the web server or exclusively installed by the user. It mostly relies on resources provided by the intended web server, such as third-party cookies. Typically, an attacker will use Cross-site Scripting to steal cookies, compromise personal data, run a mischievous script to deflect the session, steal the session, etc.

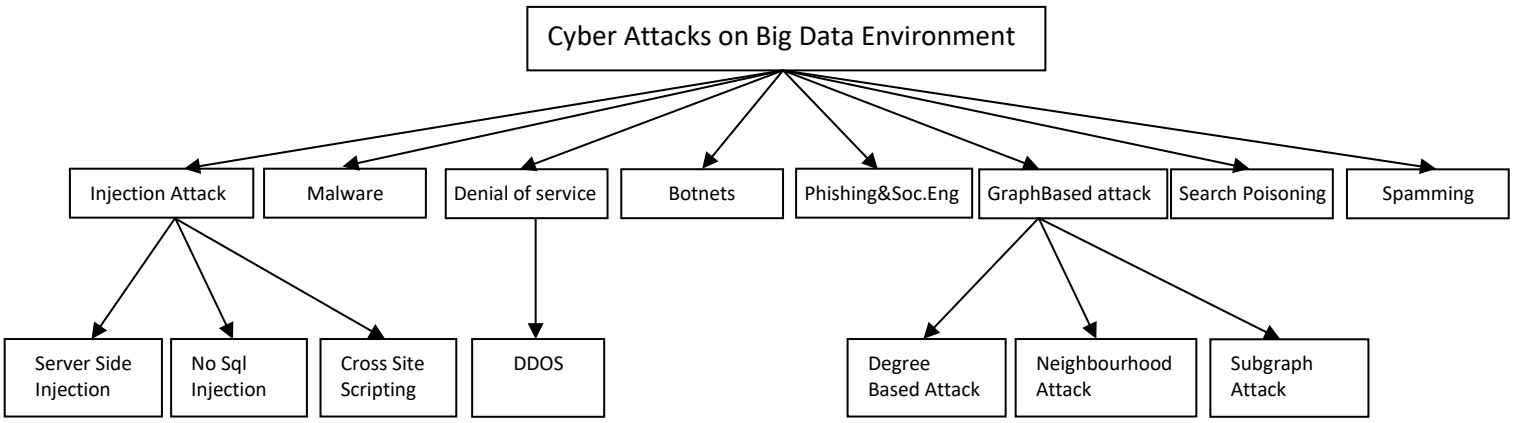

Fig. 3: Cyber Attacks on Big Data

4.2 Malware:

Malware is software that is designed to carry out and spread malicious activities, such as viruses, worms, and Trojan horses. Viruses need human interaction to spread, worms are self-replicating, and Trojans are not. Malware can cause data or operating system corruption, spyware installation, identity theft, or hard disc space theft, among other things [23].

4.3 Denial of service (DoS):

A DoS attack prevents users from accessing any network resource or a system. In most cases, it takes over the target by accessing its resources, like application buffers, CPU cycles, and so on. An attacker can also take advantage of the vulnerability by break through the target server and bringing the services decline. A large number of DoS attacks have been launched from various distributed host systems throughout the internet. And an attacker's ability to access the targeted resources from a system has gotten more challenging. Distributed denial of service (DDoS) attacks are a type of DoS attack. When a DDoS attack generates a large amount of traffic compared to the resources available to the targeted computer, the attack might cause the machine to degrade its efficiency or perhaps block any services. DDoS attacks are complicated and difficult to prevent [24].

4.4 Botnets:

Botnets are adversary-controlled networks of malware-infected systems [25]. Attackers employ bot software with a combined command and govern system to operate these zombies (bots) and combine them into a botnet network [26].

4.5 Phishing and Social Engineering:

One more type of cyber-attack is phishing, in which the attacker manipulates people to get their personal information. Attackers lure people in by sending lottery winning mail, legitimate bank mail, and messages from a false social media account [27].

4.6 Graph based Attack:

In social network data, graph-based attacks are common [28]. Even when user details is deleted from the graph in social media, attackers might leverage background knowledge to deanonymize individuals. The major types of graph based attacks are degree based attacks, neighbourhood attacks, and sub-graph attacks [29].

- Degree based Attack: A minor adversarial paradigm is the degree-based attack. In this scenario, in the actual graph G=(V, E) and falsely anonymized social network graph G=(V, E) (V, E), an attacker has deg(V) degree knowledge of a targeted vertex V. The attacker exploited this degree information as context to re-analyzed the vertex v in the anonymised social network graph G.
- Neighbourhood Attack: Sensitive information related to vertex and edges with no identities are contained in published social media. It is feasible to redefine the intended victim in an anonymised graph if the adversary or attacker has earlier knowledge of the intended victims, their neighbours' information, and the relationships between the neighbours in a neighbourhood attack.
- Sub-graph Attack: The attacker updates the real graph through introducing a complete subgraph before publishing in a subgraph attack. The attacker can find that subgraph when the unknown social network graph is released. An attacker can use subgraph to re-identify another connected user in an anonymised social network graph.

4.7 Search Poisoning:

The fake use of Search Engine Optimization strategies to artificially improve the ratting of a webpage is known as search poisoning [30]. Frequently utilised search phrases are typically used to fraudulently route readers to short-term websites. The first case of poisoning was recorded in 2007 [31], and it was quickly followed by a slew of others.

4.8 Spamming:

Spamming is the practise of delivering unsolicited mass messages to a large number of people [32, 33]. Anti-spam strategies like as munging, access filtering, and content screening are all crucial.

5. CONCLUSION

Big data approaches' ultimate goal is to be able to find meaningful and usable information in a timely manner. To develop effective analytics, there is a demand for the ability to evaluate a large amount of diverse data as well as the ability to identify, store, access, and retrieve large amounts of data. Data is becoming more vulnerable to cyber attacks since it grows exponentially every day. In this research paper an overview of big data technology is being given. We have talked about different types of attacks that can happen with big data. We

have also proposed a cyber attack taxonomy for big data environment.

6. REFERENCES

[1] L. Xu, et al., "WCDMA Data based LTE Site Selection Scheme in LTE Deployment[C]", in Proc. 2015 International Conference on Signal and Information Processing(ICSINC), Beijing, China, Oct. 2015.

[2] X. Cheng, et al., "A Novel Big Data Based Telecom Operation Architecture[C]", in Proc. 2015 International Conference on Signal and Information Processing(ICSINC), Beijing, China, Oct. 2015.

[3] M. Herland, T. M. Khoshgoftaar, et al., "Survey of Clinical Data Mining Applications on Big Data in Health Informatics[C]", in Proc. 2013 12th International Conference on Machine Learning and Applications (ICMLA), Miami, FL, 2013, pp. 465-472.

[4] M. Viceconti, P. Hunter, et al., " Big Data, Big Knowledge: Big Data for Personalized Healthcare[J]", IEEE Journal of Biomedical and Health Informatics, vol. 19, no. 4, pp. 1209-1215, July 2015.

[5] T. Huang, L. Lan, X. Fang, et al., "Promises and challenges of big data computing in health sciences", Big Data Res., 2 (2015), 2–11.

[6] J. Frizzo-Barker, P. A. Chow-White, M. Mozafari, et al., "An empirical study of the rise of big data in business scholarship", International Journal of Information Management, 36 (2016), 403–413.

[7] B. Nelson, T. Olovsson, "Security and privacy for big data: A systematic literature review", In: 2016 IEEE International Conference on Big Data (Big Data), (2016), 3693–3702

[8] F. Deng-Guo, Z. Min, L. Hao, "Big Data Security and Privacy Protection", Chinese Journal of Computers, 37 (2014), 246–258.

[9] M. Li-chuan, P. Qing-qi, L. Hao, et al., "Survey of Security Issues in Big Data", Radio Communications Technology, 41 (2015), 1–7.

[10] X. Jin, B. Wah, X. Cheng, et al., "Significance and challenges of big data research, Big Data Research", 2 (2015), 59–64.

[11] N. B. Kshetri, "The emerging role of Big Data in key development issues: Opportunities, challenges, and concerns", Big Data & Society, 1 (2014), 1–20.

[12] O. Ylojoki, and J. Porras, "Perspectives to Definition of Big Data: A Mapping Study and Discussion", Journal of Innovation Management, vol. 4, no. 1, pp. 69-91, 2016. http://hdl.handle.net/10216/83250.

[13] R. Kune, P. Konugurthi, et al, "The Anatomy of Big Data Computing, Journal Software –Practice & Experience", vol. 46, no. 1, pp. 79-105, 2016. doi: 10.1002/spe.2374.

[14] I. Lee, Big Data: Dimensions, "Evolution, Impacts, and Challenges", Business Horizons, vol. 60, no. 3, pp. 293-303, 2017. doi: 10.1016/j.bushor.2017.01.004.

[15] C. Yang., Q. Huang, et al., "Big Data and cloud computing: innovation opportunities and challenges", International Journal of Digital Earth, vol. 10, no. 1, pp. 13-53, 2017. doi: 10.1080/17538947.2016.1239771.

[16] U. Sivarajah, M. Kamal, et al., "Critical analysis of Big Data challenges and analytical methods", Journal of Business Research, vol. 70, pp. 263-286, 2017. doi: 10.1016/j.jbusres.2016.08.001.

[17] S. Owais, and N. Hussein, "Extract Five Categories CPIVW from the 9V's Characteristics of the Big Data", International Journal of Advanced Computer Science and Applications, vol. 7, no. 3, pp. 254-258, 2016. http://thesai.org/Downloads/Volume7No3/Paper_37-Extract_Five_Categories_CPIVW.pdf.

[18] G. B. Tarekegn, Y. Y. Munaye, " Big Data: Security Issues, Challenges and Future Scope",International Journal of Computer Engineering & Technology, 7 (2016), 12–24.

[19] M. Benjamin, et al, "Eigenspace Analysis for Threat Detection in Social Networks". In: 14th International Conference on Information Fusion, (2011), 1–7, IEEE.

[20] B. A. Kumar, S. Maninder, " Data mining-based integrated network traffic visualization framework for threat detection", Neural Computing and Applications, 26 (2015), 117–130.

[21] F. Almeida, "Big Data: Concept, Potentialities and Vulnerabilities", Emerging Science Journal, 2(2018), 1–10.

[22] Atefeh Tajpour, et al, "SQL Injection Detection and Prevention Techniques", International Journal of Advancements in Computing Technology Volume 3, Number 7, August 201110.4156/ijact.vol3.issue7.11

[23] Adebayo, Olawale Surajudeen, et al, "Malware Detection, Supportive Software Agents and Its Classification Schemes", International Journal of Network Security & Its Applications (IJNSA), Vol. 4, No. 6, November 2012, pp.33-49, ISSN: 0974-9330.

[24] Q. Gu and P. Liu, "Denial of Service Attacks", Technical Report, http://s2.ist.psu.edu/paper/DDoS-Chap-Gu-June-07.pdf

[25] B. Stone-Grass, M. Cova, et al, "Your Botnet is My Botnet: Analysis of a Botnet Takeover", CCS'09,November 2009, Illinois, USA

[26] M. Bailey, E. Cooke, et al, "A Survey of Botnet Technology and Defenses", CATCH'09, Cybersecurity Applications and Technology, March 2009, Washingtin DC, USA

[27] S. Gupta, A. Singhal, et al, "A literature survey on social engineering attacks: Phishing attack," in Computing, Communication and Automation (ICCCA), 2016 International Conference on. IEEE, 2016, pp. 537–540.

[28] J. H. Abawajy, M. I. H. Ninggal, et al, "Privacy preserving social network data publication," IEEE communications surveys & tutorials, vol. 18, no. 3, pp. 1974–1997.

[29] N. K. Singh and D. S. Tomar, "Privacy preservation of social media services," Exploring Enterprise Service Bus in the Service-Oriented Architecture Paradigm, p. 236, 2017.

[30] L. Lu, R. Perdisci, and W. Lee, "SURF: Detecting and Measuring Search Poisoning", CCS'11, October 2011, Illinois, USA.

[31] L. Vaas, "Malware poisoning results for innocent

searches", 27th November, 2007, http://www.eweek.com/c/a/Security/Malware-Poisoning-Results-for-Innocent-Searches

[32] M.T. Banday and J.A. Qadri, "SPAM – Technological and Legal Aspects", Kashmir University Law Review, Vol. 8, No. 8, 2006.

[33] B. Whitworth, "Spam and the social technical gap," IEEE Computer,vol. 37, no. 10, pp. 38-45, Oct. 2004.