# Big Data Analytics for The Detection of Structure Based Attack over Social Media: A Review

Jitendra Patel [1], Ravi Singh Pippal [2],
[1,2]Department of Computer Science & Engineering
[1]RKDF University, Bhopal, India

**Abstract:** With the rapid increase in the technology, there is an astonishing growth in the usage of social media sites in the day to day human's life. Users uses social networking sites to communicate, share views and to interact with family, friends and to new people. Users can share, post and upload the user's information on social networking sites. In order to access the services of social networking site, user's needs to create a profile in it by providing all the necessary personal information. The unprecedented amount of user's information are stored in social media databases. These tremendous amount of crucial data attracts sniffier to perform attack over social network to break the privacy of data. This paper gives a bird eye over anonymization process that put a breach mark in the field of privacy preservation of social media over big data environment.

**Keywords:** Big Data, Social Media, Social Networking Sites, Structure Based Attack, Anonymization

## I Introduction

Today, data is rapidly generating at an unprecedented scale from wide range of sources. Adoption of new strategies is required for managing such huge data volume, as data has changed a lot over the last few years, to cope up with the increasing demand to deal with terabytes, petabytes, and now zeta bytes. This enormous generation of data has caused the arrival of a new era of data management, often referred to as "Big Data" [1].

Big Data environment is very huge, complex, unstructured, contains incomplete and noisy information, and is heterogeneous, which may changes the traditional statistical and data analysis approaches. However, it seems that big data makes it feasible to collect more data for extracting some helpful information, but the fact is that more data do not mean more helpful information [2]. As the data and demand for real time processing increases, it will create the need of massive storage space in the distributed environment to enhance high availability and scalability. Giant companies involved in the cloud computing such as Google, Amazon, and Facebook cannot handle the huge amount of data using traditional relational database for their business model [3]. The traditional relational schema is of less use for such applications and shifting to NoSQL database seems a much better approach. Social media become one of the prominent data sources for big data. Social networking websites like Facebook, Twitter, LinkedIn , etc. shown astonishing growth of data. In Q3 2017, there are more than 2.07 billion monthly active users on Facebook [4], and handles more than 30 Petabytes of user generated data. Five new profiles are getting created on every second [5]. 51% of Instagram users access it on daily basis.

Over 30 billion data is shared on Facebook every month [6]. A substantial amount of personal and social data is getting collected by the social networking sites. Generally, the data being stored in these sites are rich in content and relationships are quite important to various third party users. As the site contain sensitive data of the user, it is necessary to ensure that the released social media would not breach privacy of the social media users. So before publishing the data are made anonymized for the use of third party. Many anonymization technique like k-anonymity, I diversity, t-closeness are used for solving the privacy issue. These models may anonymize social networks tabular data but an attacker may re-identify user in anonymized social network data by creating social network graph. Preserving privacy in social network data is more challenging than anonymizing the regular tabular data in databases. That problem emerges because of the graph data diversity and complexity. Furthermore, an attacker may utilize different kind of global properties of a graph to perform privacy attacks. Likewise, if two graphs that have the same number of vertices and edges can be essentially diverse in their global properties. An attacker may launch different type of de-anonymization attacks on anonymized published social network to re-identify the users.

## II Big Data

Data's in Big Data environment are unable to handle and processed by the traditional systems because data volume is too big to be loaded into a single machine. Also most of the traditional data analytics tools developed for a centralized data analysis process cannot be applied directly to big data. Big Data contain more abnormal or delphic data. For instance, a user may have multiple accounts, or an account may be used by many number of users, which may degrades mining result accuracy. Therefore, many new issues for data analytics are coming up, such as privacy, fault tolerance, security issues, data quality, and storage issues.

It has been a difficult task for analyzing on such complex and voluminous data without choosing right hardware or software platforms and also if the user's requirements increases up to a certain limit scaling up of hardware platform should be impended [7]. The comparison of big data with conventional data [8]. Labels the adoption of big data technologies which aims to extracts the value out of huge volume of data with various formats by enabling high velocity capture, discovery and analysis [9]. Defined big data as the dataset whose size cannot be managed by the typical database software tools to record, store, handle and analyze.

## III Social media as Big Data

Social media become one of the prominent and relevant data sources for big data. Social media data produced a wide varieties of Internet applications and websites, with most popular being Facebook, Twitter, LinkedIn, YouTube, Flicker and Instagram. Every day trillions and trillions of data are being generated in social media websites.

With growth of these social media sites allow users to be connected and creates an environment for interacting, sharing and collaborating communication. Information that are generated has been spread to many different areas such as everyday life (e-business, e-tourism, hobbies, friendship, etc.), education, health, and daily works. Exponential growth of social media has produced a challenging issue for conventional data analysis algorithm techniques ,such as machine learning, data mining, statistics, and so on, due to their high computational complexity for large datasets.

## IV Security Issue in Social Media

Social media becomes the best way for getting the information about a person because users are sharing lot of their information on social media. Social media users have uploaded their personal information in user-profile and share lots of personal information. As much information available on social media attracts the attackers to steal the personal information of users. Because of that attackers use a different kind of attacks to access the user-profile data from social media and that problem raises the security issues in social media [10].

### 4.1 Identity Theft Issue

In identity theft, an attacker steals the private information (username, email id, birth date, location) of a user. Users are having their private information in user-profile and using less profile privacy that attracts attacker to steal information from user-profile. In identity theft issue, attacker steals the private information of targeted users and creates a new profile on same social networking site or other social networking sites and use that profile for criminal or non-criminal activities. For creating new profile, attacker may use all the private information of targeted user as well as their profile picture. Creating a new profile of users without the awareness of them is called profile cloning [11]. Profile cloning is classified into two types:

#### 4.1.1 Existing Profile Cloning

Attacker creates victim's fake profile on the same social networking site. The fake profile contains the same private and public information as original to increase reliance. Using this fake profile attacker sends a friend request to the user friends and most of the friend requests are accepted by that user's friends. If friend request is accepted by that particular user's friends then the attacker is able to get their profile information also.

#### 4.1.2 Cross-site Profile Cloning

Attacker creates victim's fake profile on another social networking site where user profile does not exist. After creating a profile on other social networking site attacker send the friend requests to user friends those are available on that social networking site. Once a friend request is accepted by victim's friends then the attacker may access their profile information as well.

### 4.2 Spam Issues

In spam issue, social media users get the wall post, messages or news feeds that contain the hyperlink or advertisement link. Once a user click on any of that link then the users is redirected to the malicious or phishing websites and these websites have malware and viruses [10, 12]. This kind of spam is more effective than the traditional email spam because most of the users are aware of email spam. Users have marked that type of email as spam. But in case of social media, users are facing spam issue because here spam is coming in the form of a message, news feeds or wall post and these spam contains malicious or phishing websites link. If a user click any of those spam links then the user is redirected to the malicious or phishing websites and give user-profile access or user information to the attacker.

### 4.3 Malware Issues

In malware issue, the attacker creates a fake profile of celebrity person or hyperlink and injects malware with that profile, hyperlink. If a user try to connect with these fake profile or hyperlink then malware code spread to the user's profile without the awareness of user and send users private information to the attacker [13]. Such type of malware is also spread to the user's friends profile and send their information to the attacker. The malware issue may take place by creating a fake profile, cross-site scripting, shortened and hidden link by an attacker. Attackers create a fake profile of celebrities or popular people on social networking, these fake profiles contain malware. If a user visited one of these fake profiles then malicious code spread into the user profile.

### 4.4 Privacy Issue

Sharing information on social media may give popularity but it also raises the privacy issue for the users. On social networking site users uses their real name to represent their profile and that username is exposed publicly to the other users [10, 13]. Social media users profile may be indexed by the search engine (example: Google, Bing). If an attacker wants to target a user and attacker know the username then the attacker may easily search user by using their profile name.

After getting the targeted user profile, attacker sends spam or malware to the user profile for getting access to targeted users. Privacy issue arises because many social media users are not using the proper privacy mechanism. Users have to hide their private information and that is not visible to another user. If the personal information shared with close friends then this level privacy is secured the user profile information from the other users of social media [14].

## V Background and Literature Review

Sweeny *et al.* [15] proposed $k-$anonymity model, a dataset is said to be $k-$anonymous ($k$_1) if the published data have $k-$anonymity preservation if the information for each person contained in the published data cannot be categorized from at least $(k-1)$ persons whose information also appears in the published data. Machanavajjhala *et al.* [16] showed that $k-$anonymous dataset may leak user's privacy. If background knowledge is available to an attacker then $k-$anonymity may not protect the user's privacy in the $k-$anonymous dataset. In this paper author present $\ell-$diversity model, it is a group based anonymization that is used to provide privacy for data sets by reducing the granularity of a data representation. $l-$diversity requires that each equivalence class has at least $\ell-$well-represented values for each sensitive attribute. Li *et al.* [17] proposed $t-$closeness model, i.e., An equivalence class is said to have $t-$closeness if the distance between the distribution of a sensitive attribute in this class and the distribution of the attribute in the whole table is no more than a threshold $t$. The $t-$closeness model extends the $l-$diversity model by treating the values of an attribute distinctly by taking into account the distribution of data values for that attribute.

Liu *et al.* [18] demonstrate the degree based attack in social network data. In this paper author discussed that even after social networks data is anonymized does not guarantee user's privacy by simply removing the users identity. A user may be identified by node degree in the social network graph. The author proposed k-degree anonymous solution, a graph called $k-$anonymous if for every node v, there exist at least $k-1$ other nodes in social network graph with the same degree as $v$. In this paper, author is not aware of any effective metrics to quantify the information loss incurred by the changes of its nodes and edges.

Zhou *et al.* [19] have considered $d-$neighborhood node attack in social network data. Here $d$ represents the distance of neighbors from the targeted node and $d \geq 1$. If an attacker has the information of neighbors and relation between them of targeted user's node then the attacker may re-identify the targeted user's node from the social network data. In this paper, the author shows that the neighborhood attack may be real in practice and proposed an algorithm that can handle the $\ell-$neighborhood attack problem. The proposed algorithm can use minimum DFS code to get an isomorphic check. But as d increase the number of possible DFS tree for every component increase exponentially.

## VI Conclusion

Social media operator fetch and store data from the social media user for the purpose to share among a huge varieties of third party consumers. As the fetched information often contain sensitive data, network operator release the complete graph in an anonymized and sanitized versions. But it does not provide full guarantee of the user privacy. The attacker may use structural based attacks on social network data to re-identify the users and get the user's information. Mostly researches focus on social network data attacks and find out the different type of attack patterns (like; degree based attack, neighborhood attack, subgraph attack). Using these attacks, an attacker may re-identify the user in the social network data and acquire the user's information.

This work is based on controlling graph based attack in the social network graph. The graph based attack is based on the targeted neighbor's information and the relationship between them. Most researches happened for the social network anonymization by adding dummy edges and dummy vertices information in the social networks. Adding dummy data in the social networks creates the information loss and it caused for the inappropriate result in a research study. The proposed anonymization process will increase the number of isomorphic neighborhood networks by adding dummy edges in the social network graph. Therefore, a user may not be re-identified in social network graph based on its unique neighborhood network.

## REFERENCES

[1] J. Panneerselvam, L. Liu, and R. Hill, "Chapter 1 - an introduction to big data," in *Application of Big Data for National Security*, B. Akhgar, G. B. Saathoff, H. R. Arabnia, R. Hill, A. Staniforth, and P. S. Bayerl, Eds. Butterworth-Heinemann, 2015, pp. 3 – 13. [Online]. Available: http://www.sciencedirect.com/science/article/pii/B978012801967200001X

[2] C.-W. Tsai, C.-F. Lai, H.-C. Chao, and A. V. Vasilakos, *Big Data Analytics.* Cham: Springer International Publishing, 2016, pp. 13–52. [Online]. Available: https://doi.org/10.1007/978-3-319-44550-2_2

[3] V.-D. Ta, C.-M. Liu, and G. W. Nkabinde, "Big data stream computing in healthcare real-time analytics," in *2016 IEEE International Conference on Cloud Computing and Big Data Analysis (ICCCBDA)*, July 2016, pp. 37–42.

[4] D. Noyes, "The top 20 valuable facebook statistics – updated july 2018," *Zephoria Digital Marketing*, 2018. [Online]. Available: https://zephoria.com/top-15-valuable-facebook-statistics

[5] M. Lister, "40 essential social media marketing statistics for 2018," *Wordstream*, 2018. [Online]. Available: https://www.wordstream.com/blog/ws/2017/01/05/social-media-marketing-statistics

[6] L. Vokorokos, M. Uchnár, and A. Baláž, "Mongodb scheme analysis," in *2017 IEEE 21st International Conference on Intelligent Engineering Systems (INES)*, Oct 2017, pp. 000 067–000 070. [Online]. Available: https://doi.org/10.1109/INES.2017.8118530

[7] S. M. Sanchez, "Simulation experiments: Better data, not just big data," in *Proceedings of the Winter Simulation Conference 2014*, Dec 2014, pp. 805–816. [Online]. Available: https://doi.org/10.1109/WSC.2014.7019942

[8] J. Gantz and D. Reinsel, "The digital universe in 2020: Big data, bigger digital shadows, and biggest growth in the far east (2012)," *URL: http://www. emc. com/collateral/analyst-reports/idc-the-digital-universein-2020. pdf*, 2012.

[9] J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, and A. Hung Byers, "Big data: The next frontier for innovation, competition, and productivity," *McKinsey & Company, Digital McKinsey*, 05 2011.

[10] P. Joshi and C. . J. Kuo, "Security and privacy in online social networks: A survey," in *2011 IEEE International Conference on Multimedia and Expo*, July 2011, pp. 1–6. [Online]. Available: https://doi.org/10.1109/ICME.2011.6012166

[11] L. Bilge, T. Strufe, D. Balzarotti, and E. Kirda, "All your contacts are belong to us: Automated identity theft attacks on social networks," in *Proceedings of the 18th International Conference on World Wide Web*, ser. WWW '09. New York, NY, USA: ACM, 2009, pp. 551–560. [Online]. Available: http://doi.acm.org/10.1145/1526709.1526784

[12] M. Huber, M. Mulazzani, E. Weippl, G. Kitzler, and S. Goluch, "Friend-in-the-middle attacks: Exploiting social networking sites for spam," *IEEE Internet Computing*, vol. 15, no. 3, pp. 28–34, May 2011. [Online]. Available: https://doi.org/10.1109/MIC.2011.24

[13] H. Gao, J. Hu, T. Huang, J. Wang, and Y. Chen, "Security issues in online social networks," *IEEE Internet Computing*, vol. 15, no. 4, pp. 56–63, July 2011. [Online]. Available: https://doi.org/10.1109/MIC.2011.50

[14] B. C. Fung, K. Wang, A. W.-C. Fu, and P. S. Yu, *Introduction to Privacy-Preserving Data Publishing: Concepts and Techniques*, 1st ed. Chapman & Hall/CRC, 2010.

[15] L. Sweeney, "K-anonymity: A model for protecting privacy," *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.*, vol. 10, no. 5, pp. 557–570, Oct. 2002. [Online]. Available: http://dx.doi.org/10.1142/S0218488502001648

[16] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam, "$\ell-$diversity: Privacy beyond $k-$anonymity," *ACM Trans. Knowl. Discov. Data*, vol. 1, no. 1, Mar. 2007. [Online]. Available: http://doi.acm.org/10.1145/1217299.1217302

[17] N. Li, T. Li, and S. Venkatasubramanian, "$t-$closeness: Privacy beyond $k-$anonymity and $\ell-$diversity," in *2007 IEEE 23rd International Conference on Data Engineering*, April 2007, pp. 106–115. [Online]. Available: https://doi.org/10.1109/ICDE.2007.367856

[18] L. Liu, J. Wang, J. Liu, and J. Zhang, *Privacy Preservation in Social Networks with Sensitive Edge Weights*, pp. 954–965. [Online]. Available: https://epubs.siam.org/doi/abs/10.1137/1.9781611972795.82

[19] B. Zhou and J. Pei, "Preserving privacy in social networks against neighborhood attacks," in *2008 IEEE 24th International Conference on Data Engineering*, April 2008, pp. 506–515. [Online]. Available: https://doi.org/10.1109/ICDE.2008.4497459