IST-Africa
IST-Africa 2017 Conference Proceedings
Paul Cunningham and Miriam Cunningham (Eds)
IIMC International Information Management Corporation, 2017
ISBN: 978-1-905824-56-4

Big Data Privacy in Social Media Sites

Nobubele Angel SHOZI[1], Jabu MTSWENI[2]
[1]*Council for Scientific and Industrial Research (CSIR), Meraka Institute, Meiring Naude Road, Pretoria, 0001, South Africa*
Tel: +27128414426, Email: nshozi@csir.co.za
[2]*Council for Scientific and Industrial Research (CSIR), Defence Peace Safety and Security (DPSS), Meiring Naude Road, Pretoria, 0001, South Africa*
Tel: +27128414319, Email: mtswenij@gmail.com

Abstract: Social media sites have provided us with the ability to share information about our daily activities. However, users do not understand the implications of sharing their personal data on social media sites. Individuals on these social sites share information without realising that this could lead to their privacy being invaded. All of this has been made possible through the growth of data which is now a phenomenon known as 'big data'. At the same time, most developing countries such as those found in Africa do not have legislations in place to deal with the increasing challenges of online data and personal information privacy. This paper aims to discuss privacy issues in the social media context through highlighting how big data plays a role towards impacting on the privacy of individuals. This paper will also analyse privacy policies of social media sites and lastly an African perspective towards privacy is presented. This paper provides recommendations towards data privacy in general and in the African context.

Keywords: Big Data, Privacy, Social Media Sites, African Data Privacy, Privacy Policies

1. Background

Social media is the channel of electronic communication that is used to create and distribute content allowing for interaction between the content creator and the end user [1]. Examples of some social media sites are Facebook, YouTube, Twitter, Flickr, Instagram, Google+ and many others. In order to be part of a social media site, an individual must create an account/profile that will be used to identify the user. This process of creating an account requires certain mandatory personal information such as first name and last name, email address or mobile phone number.

The rise in the adoption and use of social media sites means that individuals are continuously creating data about themselves. The more information that is available about users online, the higher the chances of their privacy being invaded through people collating their information and being able to identify their cross-cutting profile, such as location, online behaviour, and interests, consequently threatening their safety and security.

This main objective of this paper is to highlight the role that Big Data plays in social media through showing the various privacy invasions that have occurred in social media due to the vast amount of information available. Secondly, this paper presents an analysis of privacy policies found on various social media sites to understand what information is collected and shared by these platforms and how this information could result in privacy invasion on individuals.

The rest of this paper is organized as follows: Section 2 defines the concept of Big Data through its characteristics. Section 3 discusses Big Data Privacy and the various techniques

that can be used to preserve privacy. Section 4 of this paper discusses Big Data Privacy in Social Media. Section 5 presents the analysis of the various social media policies. Section 6 highlights the privacy landscape in Africa and the paper is concluded with a summary and further research recommendations.

2. Big Data

Big data can be described with characteristics that are sometimes referred to as the 3V's (Volume, Velocity and Variety) [2]. These are briefly defined as follows:

- Volume refers to the size of the data, which is more than what ordinary storage medium could handle. This volume of data is normally then stored across various clusters of medium, including local and cloud storages.
- Velocity refers to the speed at which the data is generated and made available. Data from social media changes at every second, and ultimately contributes significantly to big data.
- Variety refers to the different formats in which the data can be found in.

Big data can therefore be described as data that are large in size, comes in at different speeds and in different formats. Other researchers have also added Veracity and Value as characteristics to further define big data and these are seen as important [3]. Veracity refers to the reliability and accuracy of the data, whilst Value refers to the benefits that could be drawn from the big data.

3. Big Data Privacy

To anonymize and protect data, various privacy protection methods can be employed [4]:

- **Suppression**: The direct identifiers are usually completed removed from datasets.
- **Generalization**: An example of generalisation is removing a date of birth and just replacing it with the year.
- **Aggregation**: This turns atypical records which are generally at most risk into typical records.
- **Data Swapping**: Data values for certain records can be swapped. For example, switching the values for age, race etc. This is done to protect sensitive and risk records.
- **Random noise**: Adding noise into the data can reduce the possibilities of re-identification. This is usually done to protect numerical data by adding a randomly selected number.
- **Synthetic data**: This is replacing original values with values simulated from probability distributions.

Even with such data preserving techniques, an individual's privacy is still not safe. The very characteristics of big data can be identified as challenges. For example, the large volume and varying variety of this data can make it difficult for the current traditional methods (hardware and software) to safely and securely handle these large amounts of data [5]. Apart from the characteristics of big data, one of the major challenges and one which is usually identified as being the most important is privacy [6][7].

There has been a number of big data privacy invasions in the recent years that have shone the spotlight on what could occur to individuals due to their data being either compromised or being combined with other related data sources. The following are a few examples:

Netflix: Netflix published data sets that contain movie rating records, which list the movie rated, assigned rating, date or rating and the unique id of the person who rated it. The

researchers were able to demonstrate that by knowing an individual's movie viewing habits, they can potentially identify the individual's personal record [8].

American Online (AOL): AOL released a dataset that had replaced the AOL username and IP address with unique identification numbers to allow researchers to correlate different search queries to individual users [9]. From this dataset, one user was uniquely identified by researchers [9].

Target: Target is a large American retailing company that used the power of data fusion and correlation to identify assumed pregnant women to send them targeted coupons that are related to pregnancy such as baby-related products [10].

4. Big Data Privacy in Social Media

When it comes to big data privacy in social media, its structure can be likened to that of the 'Iceberg model' depicted in Figure 1.

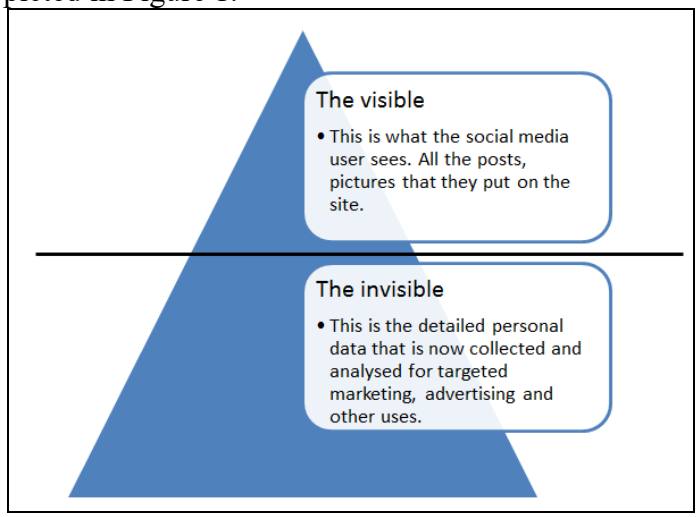

Figure 1: Social Media Iceberg Model (Adapted from [11])

All the user sees is the 'visible' part of the iceberg which is the information that they contribute, the fun activities that they are able to perform on the site. The 'invisible' part of the iceberg however represents all the data mining and data collection that is done on the user's data for targeted marketing and for other means by the social media site.

The primary goal of social media sites is to share information with other people. Users can share parts of their daily lives through status updates, photographs and videos. Smartphones are used by the users to take the pictures and videos, these can have location information and metadata embedded in them. This not only give people a front row seat into private information but it gives the ability to leverage this information and using it in malicious ways to invade the privacy of individuals.

Facebook has been highly criticized over privacy issues. In 2010, two personal information aggregation techniques called connections and instant personalization were criticised as they allowed anyone to be able to access personal information that may have not been intended for the public [11]. It also emerged that Facebook uses location based services to suggest friends to users [12].

5. Case Study: Social Media Privacy Policies

Most social media sites have a privacy policy that details the type of information that they collect and use. The usage of this information is however not detailed in the policies. Table 1 provides a systematic comparison of Facebook [13], Twitter [14], Instagram [15], LinkedIn [16] and YouTube [17] privacy and security policies. This is done to determine

what information each social media platform collects based on what they have detailed in their policies. These platforms have been chosen as they are some of the more popular social media platforms in the world [18].

The analysis of the different social media policies have been grouped into the following main themes:

- **Login Details:** this is the username, password and email address which was used to register the account.
- **Profile information:** this is the information that is provided on the user's profile such as first name, last name, phone number, address, picture etc.
- **User content:** This is the information that the user posts or shares on the social media site.
- **Information from other users:** this is information that users' user's friends post or share on a social media platform and this information might be related or linked to the registered user (e.g. tagged photos or comments by friends on user's posts).
- **Payment information:** this is the information such as credit information used to pay for the online services or applications.
- **Location information:** this is information about the user's location collected either through the Web browser or GPS-enabled devices, such as mobile phones.
- **Device and mobile information:** this is the information about the computer or mobile device used to access the social media site. This can be browser information or IP and MAC addresses that identify the connecting device or computer on the Internet.
- **3rd party services:** this information refers to 3rd party applications for measuring network traffic and analysing trends usually used by social media sites.
- **Cookies:** These are used to collect information about how the social media site is used by the connecting user.
- **Direct/Private messages:** these are messages that are shared between individuals privately using inboxes or messengers.
- **Address book information:** this is information the user uploads to the social media site to help them find their friends.

Table 1: Data Collected by Social Media Sites

Category	Social Media Site				
	Twitter	Facebook	Instagram	LinkedIn	YouTube
Login Details	X	X	X	X	X
Basic profile information	X	X	X	X	X
User Content	X	X	X	X	X
Information from other users	X	X	X	-	X
Payment information	X	X	-	-	X
Location information	X	X	X	X	X
Device and mobile information	X	X	X	X	X
3rd party services	X	X	X	X	X
Cookies	X	X	X	X	X
Device/Private messages	X	X	-	-	-
Address book information	X	X	X	X	-

Table 1 draws comparisons on how the data is used by the social media sites. The 'X' denotes that the social media site collects the mentioned category of information. The '-'

denotes that this social media site does not collect the mentioned category of information. The subsection below and Table 2 show how the collected information is used by the social media sites:

- **Communication purposes**: this information is used to communicate with the user through for example sending emails when policies change..
- **Improve service**: information can be used to conduct research and development for the improvement of the social media sites.
- **Sharing with affiliates**: information is shared with other companies that are owned and controlled by the social media sites. For example Facebook can share information with Instagram and WhatsApp.
- **Sharing information with 3rd parties**: information is shared with 3rd parties. This is usually the public profile version of a user's profile that is shared.
- **Advertising**: information is used to improve advertising and measurement systems so relevant adverts are shown and measured for effectiveness.

Table 2: Data Usage by Social Media Sites

Category	Social Media Site				
	Twitter	Facebook	Instagram	LinkedIn	YouTube
Communication purposes	X	X	X	X	X
Improve service	X	X	X	X	X
Sharing with affiliates	X	X	X	X	X
Sharing information with 3rd parties	X	X	X	X	X
Advertising	X	X	X	X	X

The sub-section below shows what the privacy policies state about deleting accounts on the social media sites, when the user decides to deregister from the specific social media platform. It is apparent that in most social media platforms once personal information has been posted; it can be used for various purposes, but also even after account deletion the information may remain persistent at various platforms raising even more personal information privacy challenges.

- **Facebook [13]:** *"You can delete your account any time. When you delete your account, we delete things you have posted, such as your photos and status updates. Keep in mind that information that others have shared about you is not part of your account and will not be deleted when you delete your account"*
- **Twitter [14]:** *"You can also permanently delete your Twitter account. Your account will be deactivated and then deleted. Keep in mind that search engines and other third parties may still retain copies of your public information, like your user profile information and public Tweets, even after you have deleted the information from the Twitter Services or deactivated your account".*
- **Instagram [15]:** *"Following termination or deactivation of your account, Instagram, its Affiliates, or its Service Providers may retain information (including your profile information) and User Content for a commercially reasonable time for backup, archival, and/or audit purposes"*
- **LinkedIn [16]:** *"Information you have shared with others or that others have copied may remain visible after you have closed your account or deleted the information from your own profile. Groups content associated with closed accounts will show an unknown user as the source. In addition, you may not be able to access, correct, or eliminate any information about you that other members copied or exported out of our*

services, because this information may not be in our control. Your public profile may be displayed in search engine results until the search engine refreshes its cache".

- **YouTube [17]:** *"After you delete information from our services, we may not immediately delete residual copies from our active servers and may not remove information from our backup systems."*

As shown in Table 1, social media sites have privacy policies which specify what information they collect from an individual's profile and how information is used. These policies serve more of an advantage to the social media sites than the individual. Users have to be made aware how sharing information on these sites can impede on their privacy. This however can be difficult as users want to use these sites and easily consent to the privacy policies without reading these.

For example: in order to create a Facebook account, it is necessary to provide the following information: first name, last name, mobile number or email address, date of birth and gender. Amongst all the surveyed social media sites, Facebook requires the most personal information about its users.

From the analysed social media sites, LinkedIn is the only site that did not collect information from 3 out of the 11 typical sources that social media sites use to collect data. In terms of how the data is used, all of the surveyed social media sites used the data for communication purposes, to improve their service, to share with affiliates, to share with 3rd parties and used it for advertising purposes. This shows how social media data is shared and the avenues in which users' personal information privacy can be violated in the big data era.

Even though all the sites state that a user can delete your information, not all the information is deleted. This is because if a user has shared this information or if people have commented/copied this information or if this information was made public than it is now in the public domain and the platform has no obligation to delete this as it does not belong to the user and they do not own it either.

In the social media space, there are some countries that are coming with legislations to regulate what users post on social media, including what data and personal information social media organisations are permitted to collect from users. However, progress in this regard is still limited to developed countries. In the following section, we provide the data/information privacy legislation landscape in Africa to show the gaps that exist in addressing the challenges associated with big data privacy in social media platforms,

6. Data Privacy Legislation in Africa

The rise of big data and the interconnectedness of systems to people have all given rise to the concerns towards individuals' privacy. This has not exempted developing countries as they also form part of this growing global community. The purpose of this section is to provide a brief summary of the data/information privacy legislation landscape in Africa.

Africa is considered the second largest continent in the world and also has the second largest population after Asia [19]. The growth in Internet usage has resulted in a number of African governments realising the importance of data privacy and protecting the privacy of their citizens. However, out of the 54 countries in Africa, less than 50% currently have any data protection or privacy laws that have been passed or bills that are yet to be passed in their respective parliaments. Four African countries have been selected for the purpose of studying their data/information privacy law/bill. The selected countries are – Nigeria, Egypt, South Africa, Cape Verde and Madagascar. Nigeria, Egypt and South Africa were chosen as these are the top 3 countries with the highest GDP in Africa [20]. The GDP determines the health of the economy and therefore could have a direct relation to how the country invests towards their ICTs. Cape Verde was chosen as it was the first African

country to implement a data privacy law in 2001 [21] and Madagascar was chosen as it is the latest African country to adopt a data law in 2015 [22].

6.1 Nigeria

The right to privacy is detailed in the Nigerian Constitution under section 37. The section states that "The privacy of citizens, their homes, correspondence, telephone conversations and telegraphic communications is hereby guaranteed and protected" [23]. However, apart from this declaration in the constitution there is no other law that pertains to protecting the privacy of individuals' information or data in Nigeria except provisions in some federal laws.

The Consumer code of practice regulations of 2007 which is issued by the Nigerian Communications Commission (NCC) does refer to specific regulations such as 'Information will be protected against improper or accidental disclosure' and 'Information may not be transferred to any party excepted by applicable laws or regulations' amongst others [24]. However, this does not specific deal with the handling of personal information posted or handled by social media platforms.

6.2 Egypt

Egypt does not have a formal data/information privacy protection law. However, the topic of privacy is covered in a number of other Egyptian legislative documents. Act 57 of the Egyptian constitution addresses issues of private life. It states that communication methods such as telegraph, postal, electronic correspondence, telephone calls and other forms of communication are inviolable and their confidentiality is guaranteed and may only be monitored or inspected with a court order [25]. The Egyptian parliament is working on a Cybercrime bill that might stipulate that social media website such as Facebook be regulated to deal with violations that infringe on Egyptian individuals privacy [26].

6.3 South Africa

The right to privacy in South Africa is protected by the common law and by Section 14 of the South African Constitution which states "Everyone has the right to privacy, which includes the right not to have the privacy of their communications infringed" [27]. A new comprehensive data privacy law was enacted in South Africa in 2014, which is known as The Protection of Personal Information (POPI) Act. This act has been passed but not yet in effect.

The POPI Act is built on eight (8) principles for governing the processing of personal information [28]: (1) **accountability** – information must be obtained in lawful manner, (2) **information processing limitation**– the act limits the processing of information beyond specified purpose, (3) **purpose specification** - information can only be used for specified purpose, (4) **further processing limitation** – further processing must be in accordance with initial reason that information was collected (5) **information quality** – the quality of information must be ensured, (6) **openness** - person processing information must have a degree of openness, (7) **security safeguard** – proper security safeguards must be ensured and (8) **data subject participation** – the data subject must be able to participate and access any information about them held by the data parties.

6.4 Cape Verde

The Data Protection Act of the Republic of Cape Verde was enacted in 2001 as law that protects individuals with regard to the processing of personal data [21]. The act was updated in 2003. The act covers areas pertaining to data quality and lawfulness when processing data, processing sensitive data, the combining of personal data. It also details the

rights of the data subject such as their right to accessing their information, their right to being able to object. Security and confidentiality during data processing is addressed to ensure that the data controller/processers ensure that there are adequate technical measures to protect personal data.

6.5 Madagascar

The Data protection law (Law no. 2014-038) is the main law in Madagascar that relates to the protection of personal data [22]. The law was adopted in 2015, however it is yet to be published in the Official Gazette of the Republic of Madagascar in order for it to be in effect. There is currently only a French version published and available to the public. This law does not contain any sections related to social media personal information privacy.

7. Conclusion and Recommendations

This paper has highlighted the various big data privacy invasions examples that have occurred and also the ones associated with social media. Through the users being aware of what constitutes privacy policies of social media sites and understanding that the information they share can be possible cause privacy invasion – they can better protect their information and decrease the possibilities of these invasions occurring.

With the initiation of data privacy laws in some Africa countries, there can be a number of contraventions in terms of how social media sites currently use individual's data. For example, in South Africa the POPI act states that information needs to be kept in South Africa. In this case, social media sites would need to find alternative ways to storing user's data in South Africa and not have the data stored in servers outside the country. The implications of this information being sent to 3rd parties who could potentially be outside the country are yet to be identified and investigated.

Recommendations and future work for the continuation of this research, it is recommended that the effects that the privacy laws in Africa will have on social media sites need an in-depth investigation. A landscape of the African data privacy laws needs to be presented to better understand how Africa has progressed or is progressing towards protecting the privacy of its citizens through privacy laws. Research in the area of data privacy in an African context is lacking and this is an area which can have growth. It is recommended that users are made more aware how their data is being used and what is contained in the privacy policies of the social media sites they use, this requires efforts from both the user and the social media sites. These sites need to make their policies easy to read and understand and the user needs to take the initiative to want to be informed about how these policies impact them.

References

[1] Ellison, N. B. (2007). Social network sites: Definition, history, and scholarship. Journal of Computer-Mediated Communication, 13(1), 210-230.

[2] Laney, D. (2001). 3D data management: Controlling data volume, velocity and variety. META Group Research Note, 6, 70.

[3] Yin, S., & Kaynak, O. (2015). Big data for modern industry: challenges and trends [point of view]. In Proceedings of the IEEE, 103(2), 143-146.

[4] El Emam, K. (2013). Guide to the de-identification of personal health information. CRC Press.

[5] Machanavajjhala, A., & Reiter, J. P. (2012). Big privacy: protecting confidentiality in big data. XRDS: Crossroads, The ACM Magazine for Students, 19(1), 20-23.

[6] Shozi, N. A., & Mtsweni, J. (2016). Big Data Privacy and Security: A Systematic Analysis of Current and Future Challenges. In 11th International Conference on Cyber Warfare and Security: ICCWS2016 (p. 296). Academic Conferences and publishing limited.

[7] Michael, K., & Miller, K. W. (2013). Big data: New opportunities and new challenges [guest editors' introduction]. Computer, 46(6), 22-24.
[8] Narayanan, A., & Shmatikov, V. (2008). Robust de-anonymization of large sparse datasets. In 2008 IEEE Symposium on Security and Privacy (sp 2008) (pp. 111-125). IEEE.
[9] Barbaro, M. & Teller, J. (2006). A Face Is Exposed for AOL Searcher No. 4417749. Retrieved 1 March 2017 from http://www.nytimes.com/2006/08/09/technology/09aol.html
[10] Duhigg, C. (2017). How Companies Learn Your Secrets. Nytimes.com. Retrieved 3 March 2017, from http://www.nytimes.com/2012/02/19/magazine/shopping-habits.html?pagewanted=1&_r=1&hp
[11] Debatin, B., Lovejoy, J. P., Horn, A. K., & Hughes, B. N. (2009). Facebook and online privacy: Attitudes, behaviors, and unintended consequences. In Journal of Computer-Mediated Communication, 15(1), 83-108.
[12] Jardin, X. (2016). Privacy invasion? Facebook is using your phone's location data to suggest friends. Retrieved from http://boingboing.net/2016/06/27/privacy-invasion-facebook-is.html
[13] Facebook. (2016) Data Policy. Retrieved 09 December 2016, from https://www.facebook.com/full_data_use_policy
[14] Twitter. (2016). Twitter Privacy Policy. Retrieved 09 December 2016, from https://twitter.com/privacy?lang=en
[15] Instagram. (2016). Privacy Policy. Retrieved 09 December 2016, from https://www.instagram.com/about/legal/privacy/
[16] LinkedIn. (2016). Your Privacy Matters. Retrieved 09 December 2016, from https://www.linkedin.com/legal/privacy-policy
[17] YouTube. (n.d). Privacy Policy. Retrieved 09 December 2016, from https://www.google.co.uk/intl/en-GB/policies/privacy/
[18] Smart Insights. (2017). Global Social Media Statistics Summary 2017. Retrieved 10 March 2017, from http://www.smartinsights.com/social-media-marketing/social-media-strategy/new-global-social-media-research/
[19] United Nations (UN). (2015). World Population Prospects – Key findings and advance tables. Retrieved February 25, 2017 from https://esa.un.org/unpd/wpp/Publications/Files/Key_Findings_WPP_2015.pdf
[20] The World Bank. (2015). Gross Domestic Product 2015. Retrieved 25 February 2017 from http://databank.worldbank.org/data/download/GDP.pdf
[21] Data Protection Act of the Republic of Cape Verde. (2001). Retrieved 25 February 2017 from http://www.cnpd.cv/leis/DATA%20PROTECTION%20Law%20133.pdf
[22] Data Protection Laws of the World. 92017). Madagascar. Retrieved 25 February 2017 from https://www.dlapiperdataprotection.com/index.html?t=law&c=MG
[23] Constitution of the Federal Republic of Nigeria [Nigeria]. (1999). Retrieved 24 February 2017 from http://www.refworld.org/docid/44e344fa4.html
[24] Nigerian Communications Commission (NCC). (2007). Consumer Code of Practice Regulations 2007. Retrieved 24 February 2017, from http://www.ncc.gov.ng/docman-main/legal-regulatory/regulations/102-consumer-code-of-practice-regulations-1/file
[25] The Constitution of the Arab Republic of Egypt 2014. Retrieved 3 March, from http://www.sis.gov.eg/Newvr/Dustor-en001.pdf
[26] Al-Monitor. (2017). Will Egypt's cybercrime law overstep boundaries?. Retrieved 10 March 2017, from http://www.al-monitor.com/pulse/originals/2016/06/egypt-enacts-cyber-crime-law-preserve-national-security.html
[27] The Constitution of the Republic of South Africa, Act 200 of 1993. (1993). Bill of Rights Section 14. Retrieved 25 February 2017, from http://www.gov.za/documents/constitution/chapter-2-bill-rights#14
[28] South African Government Gazette. (2013). Protection of Personal Information Act. Retrieved 25 February 2017, from http://www.justice.gov.za/legislation/acts/2013-004.pdf