Volume 4, Issue 10, October 2014 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Personal Big Data Usage and Controls - Review

Sujata Jindal*
Computer Science Engineering Department
SGT Institute of Engineering & Technology,
Gurgaon, India

Dr. Ritu Sindhu
Associate Professor, CSE Department
SGT Institute of Engineering & Technology,
Gurgaon, India

Abstract: ***Personal Big Data plays a vital role in today's digital world. There are many opportunities and challenges that big data is providing. The need to protect personal big data has escalated alongside the rise of information age and mobile usage. The risk is quite high as typical user online today is not as technically knowledgeable in the mobile age. There are multiple consumers for big data from internet to mobile devices multiplying the risk. A seamless integration of our online big data is necessary to protect an individual's identity in digital world. It requires collection, control and thorough analysis of personal big data. Many social networks are mining the personal activities which are being leveraged for custom advertising, customized content etc. In this paper we give review of most common data miners, security concerns and methods they provide to control personal big data on various social networks like LinkedIn, Facebook, Google +, YouTube. It presents a comprehensive analysis of existing solutions and shows the possible research in this field.***

Keywords: ***Big data; personal big data; data mining; social web; social network; data privacy; data security; data control***

I. INTRODUCTION

A. Understanding Big Data and Personal Big Data

Big data describes a massive volume of structured and unstructured data that is so large that it's difficult to process using traditional database techniques. [1] "Personal Data" means any information that could identify a person directly or indirectly. The kind of personal data available have increased in recent times due to emergence of social media and growth in mobile devices. It is any information relating to a data subject, being an identified natural person or a natural person who can be identified, directly or indirectly, by means reasonably likely to be used by a data controller or by any other natural or legal person, in particular by reference to an identification number, location data, online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that person. [2]

B. Big Data Types

In its 2012 report on personal data: Rethinking Personal Data [3], WEF discriminated between three types of data-

Volunteered data

It is created and explicitly shared by individuals. It consists of photos, blog posts, tweets, emails etc. on social network profiles. Users give their implicit and explicit consent for various levels of use of data. Very few of them read the terms and policies. Consent is often a tick-box exercise that individuals race through when they sign up for a service. Users are providing data in return of services provided by these social networks.

Observed data

It is captured by recording the actions of individuals, e.g. location data when using cell phones, telephone usage behaviour, internet browsing preferences. Individuals may not aware of how much data is being captured and used about them.

Inferred data

It is data about individuals based on analysis of volunteered or observed information, e.g., credit scores. Inferred data has predictive capabilities that have become concentrated in the hands of a few large companies

While there is a great potential in using big data but the thing to worry is that more and more data is collected about people knowingly and unknowingly. Collected personal big data is analysed by various companies using different algorithms to draw results about user behaviour and to apply these correlations in future. They may derive wrong conclusions through their data analysis. So good news is that there is big data and bad news is that these collections may not be perfect every time from an individual's point of view. A number of open source platforms have grown up specifically to handle these vast amounts of data quickly and efficiently, including Hadoop, MongoDB, Cassandra, and NoSQL.

II. BIG DATA AND AN INDIVIDUAL

Today, data is more deeply woven into the fabric of our lives than ever before. We aspire to use data to solve problems, improve well-being, and make our lives easier. The collection, storage, and analysis of data is possible due to in increases

in processing power, the catering costs of computation and storage, and the growing number of sensor technologies embedded in devices of all kinds. In 2011, some estimated the amount of information created and replicated would surpass 1.8 zettabytes. [4] In 2013, estimates reached 4 zettabytes of data generated worldwide.

A. Personal Big Data Miners, Consumers

The sources and formats of data continue to grow in variety and complexity. A partial list of sources includes the public web; social media; mobile applications; federal, state and local records and databases; commercial databases that aggregate individual data from a spectrum of commercial transactions and public records; geospatial data; surveys; and traditional offline documents scanned by optical character recognition into electronic form. The advent of the more internet enabled devices and sensors expand the capacity to collect data from physical entities

Few have ever heard of Acxiom. But analysts say it has amassed the world's largest commercial database on consumers — and that it wants to know much, much more. Its servers process more than 50 trillion data "transactions" a year. Company executives have said its database contains information about 500 million active consumers worldwide, with about 1,500 data points per person. That includes a majority of adults in the United States. [5]

Carolinas HealthCare, which runs more than 900 care centres, including hospitals, nursing homes, doctors' offices, and surgical centres, has begun plugging consumer data on 2 million people into algorithms designed to identify high-risk patients so that doctors can intervene before they get sick. The company purchases the data from brokers who cull public records, store loyalty program transactions, and credit card purchases. Such data can be used in many different ways – e.g. frequent purchases at pizza outlets would mean that you may be at risk of weight gain, frequent purchases of sleeping pills could mean you are depressed or overworked and so on. [6]

Here are some examples of data miners, consumers (one way or other) from various common usage categories -

Social Networking and Media

Examples - Facebook, Google+, Twitter, LinkedIn, YouTube, Blogs, Review Sites (Glassdoor, Tripadvisor etc)

Internet Utility/Transactions

Examples – Online shopping (Amazon, Flipkart etc), Product comparison, review sites (Snapdeal, Junglee), Newsletters, Subscriptions, Memberships

Mobile and other internet connected devices

Examples: Chat (Whatsapp, FB chat), GPS (Google Map), In-app activity (App Likes, Purchases, Usage)

Gaming

Examples: Xbox, PS3/PS4, Wii

Advertising

Examples: Google AdWords, Website Advertising

III. WHY PERSONAL BIG DATA IS IMPORTANT

The reason why personal data is so important is well known to all of us, but there is not enough awareness about importance of personal big data that we all leave behind while browsing through websites, using gaming devices, using apps on smart phones, or posting status on social networking sites, or while shopping online. All of this is the digital foot prints that we leave behind which gives deep insight into people's behavior, nature, location, status and a lot more. When this data is aggregated and integrated, it gives an even more detailed picture about an individual or group of people.

There is ever-increasing need to tell the individuals about the data the data-miners are collecting, how it is collected and, how it is being used or shared. E.g. If a women is enrolled as daily runner in some mobile health check app, she must know what is happening with her information if she is running daily, or sick or has low stamina or such? Where is the information being stored, and how?

The behavioral marketing, advertising and search industries are following people all over the web, collecting information about them and their activities without their awareness and we as the personal big data contributors need to be in control of it and ensure that it is not being misused, or even if it's being used we need to be aware of it.

IV. PERSONAL BIG DATA PRIVACY

A. Consent and opt-in/opt-out mechanisms

- When system collects personal information, it requires individual's consent so that it can collect, use, or even store and shares it. However, there are exceptions to this rule, and/or some systems would find workarounds to take implicit consent from individual by different means. There are three main types of consent an individual may exercise:
  - Explicit Consent
  - Implicit Consent
  - Opt-out Consent
- The individuals often sign up to online services not knowing how their data will be stored, protected, shared and so on. While legally they have given organizations their consent to the stated rules for usage, only a very few individuals actually read these privacy policies or terms of services completely. Individuals therefore have little visibility into the practices of the organizations they are putting their trust in – until their data is breached or misused.
- An important concern of individuals is how to manage their online identity and the different aspects of their digital lives. The lack of contextual control and permissions represents another cause for concern among individuals. One

of the few ways for individuals to keep different parts of their digital lives separate is to use different names and e-mail addresses for different contexts, and to use pseudonyms, or to prevent their data being captured or linked to them in the first place, this has been a regular practice for quite long, but in recent times introduction of linked web and common sign-on features; the anonymity is compromised.

B. Tools at help?

Although there are not many tools at help, but some tools are under research & development that can help individuals understand how their personal big data is being used.

Xray [7] The researchers at Columbia Engineering have developed Xray , a new tool that reveals which data in a web account such as emails, searches, or viewed products are being used to target which outputs such as ads, recommended products, or prices. Its main objective is to make the online use of personal data more transparent. The current Xray system works with Gmail, Amazon and YouTube.

ABP (AdBlock Plus) [8] It can block annoying ads (allowing only acceptable ads), tracking and malware. It works on the basis of filter lists. These filter lists are extensive set of rules that tell the AdBlock Plus which elements of websites to block. Filter lists can also be used to block tracking and malware.

Collusion for Chrome [9] It graphs the spread of user data from sites to trackers, in real time, to expose and optionally, to break these hidden connections.

Privacyfix It checks your privacy exposure on Facebook, Google and LinkedIn. Block over 1200 trackers from following your movements online. It allows you to manage Facebook, Twitter, LinkedIn and Google from one dashboard.

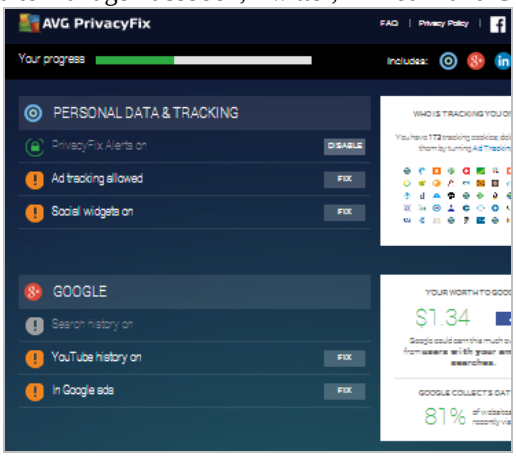

Figure 1: PrivacyFix helps understand how different websites use data.
Source: www.privacyfix.com

WolframAlpha helps individuals visualize vast quantities of social data about them to gain insight into their social network. Apart from this it uses its expert-level knowledge and algorithms to answer questions, generate reports and do analysis across thousands of domains.

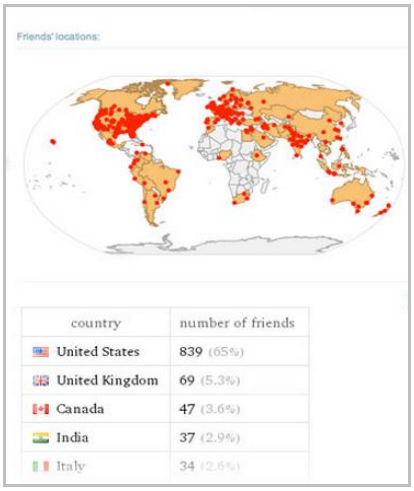

Figure 2: WolframAlpha's visualization of facebook data.
Source: http://www.wolframalpha.com/facebook/

Privacy Icons Privacy policies provided by various organizations are complex and sometimes hard to understand. Privacy icons provided by various companies make these privacy policies visual with icons that are easy to understand. Now people can understand how their personal data will be transacted, with just a glance. Policy icons are meant for any sites that store user data e.g. e-commerce sites, advertisers and social networks etc. and for users who voluntarily share their personal data.

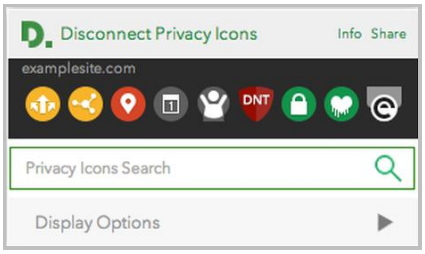

Figure 3: Disconnect Privacy Icons
Source: http://www.pcworld.com/article/2366840/new-software-targets-hardtounderstand-privacy-policies.html

V. CONCLUSIONS AND FUTURE WORK

Vital personal information is captured in various sizes and forms by data mining tools, systems and applications, which present plenty of opportunities to use that valuable asset for good use. Personal big data can be quite vast and dynamic, but is vulnerable at all the times by different means and is being misused. In this paper, we got an introduction to personal big data types, risks, mining tools, and common controls with illustrative examples to demonstrate how similar tools can be used to visualize and protect personal data.

While this paper shares key big data types, sources, miners, the list of sources and types continues to grow. The advancement in internet enabled devices has really fuelled the data collection over recent years and with expansion of devices that generate and broadcast personal data, the risk of theft, misuse and compromise are growing day by day as new and growing user base is not as technical and has little to no knowledge of such risks.

The need is to create a complete tool that not only gives user a visual view to personal data miners, person data linkages across web, data points being compromised, and it should give user complete control over same – in a simple, clear manner so that every user gets better understanding of risks and threats involved, specially users that are not as technical as some of us are.

ACKNOWLEDGMENT

The work presented here could not have been completed without the contributions of many supportive and knowledgeable people.

Foremost I want to express my deep gratitude to my guide Dr. Ritu Sindhu, Associate Professor (Department of CSE, SGT Institute of Engineering & Technology) for the continuous support and mentorship during my project. Her mentorship was paramount in providing a well rounded experience consistent with my long-term career goals. She encouraged me to not only grow as an experimentalist and a personal data scientist but also as an independent thinker.

I would like to thank faculty members and fellows at Department of CSE, SGT Institute of Engineering & Technology, for their inputs, valuable discussions and accessibility.

REFERENCES

[1] "What is Big Data? Webopedia." 2011. 29 Sep. 2014 http://www.webopedia.com/TERM/B/big_data.html
[2] European Commission, Proposal for a Regulation on the protection of individuals with regard to the processing of personal data and on the free movement of such data (General Data protection regulation), COM (2012) 11/4 draft.
[3] "Rethinking Personal Data: Strengthening Trust - Self Monitoring ..." 2012. 30 Sep. 2014 http://www3.weforum.org/docs/WEF_IT_RethinkingPersonalData_Report_2012.pdf
[4] John Gantz and David Reinsel, Extracting Value from Chaos, IDC, 2011, http://www.emc.com/collateral/analyst-reports/idc-extracting-value-from-chaos-ar.pdf
[5] "Acxiom, the Quiet Giant of Consumer Database Marketing ..." 2012. 5 Oct. 2014 http://www.nytimes.com/2012/06/17/technology/acxiom-the-quiet-giant-of-consumer-database-marketing.html?pagewanted=all
[6] "Hospitals Are Mining Patients' Credit Card Data to Predict ..." 2014. 5 Oct. 2014 http://www.businessweek.com/articles/2014-07-03/hospitals-are-mining-patients-credit-card-data-to-predict-who-will-get-sick
[7] "XRay · Transparency for the Web." 2014. 30 Sep. 2014 http://xray.cs.columbia.edu/
[8] Adblock plus - surf the web without annoying ads! http://adblockplus.org
[9] Chrome web store - collusion for chrome https://chrome.google.com/webstore/detail/collusion-forchrome/

