An Authentic Self: Big Data and Passive Digital Footprints

L. Y. Williams, D. M. Pennington

Faculty, MSIT, MSCM, School of Business and Information Technology
Purdue Global University

Lecturer, Department of Computer and Information Sciences
The University of Strathclyde
Glasgow, United Kingdom

email: lwilliams4@purdueglobal.edu, diane.pennington@strath.ac.uk

Abstract

The ability to allow users to create online communities of interest and to share a variety of personal information, collectively referred to as social media, is gradually being built into an expanding range of applications. Some of these applications, such as computer operating systems, were not originally intended to collect information from the user. Thus, users may not be aware that their digital information is being collected. Devices such as smart televisions, smart cars, and even smart grids, are now collecting massive quantities of user data without the user's knowledge.

Users of social media, and the internet in general, leave fragments of their activities and intentions behind them across an increasing range of technologies. These fragments collectively and passively create a hidden identity built up from metadata of which the user is mostly unaware.

Given that the user builds this hidden identity during the normal course of their day, without editing elements that the user may not wish to share with others, might the passive digital footprint more accurately reveal the individual's genuine or authentic self than the individual realises?

We propose that an aggregated, passively collected digital portrait of a user's unconscious but connected activities may reveal a more genuine view of that person's self than would be deduced from sources over which the user has conscious control. This more accurate and potentially revealing portrait of the individual requires a review of how privacy has been classically defined in both legal as well as ethical constructs.

Keywords

Passive digital footprint, data analytics, privacy, identity, Internet of Things, linkage attack, Big Data

1. The ubiquitous collection of personal data

Akamai (2016), a global leader in content delivery systems, estimates that as of the first quarter in 2016 there were well over a billion internet users. Internet users now have access to an unprecedented range of media content as well as tapping into connected services that allow them to monitor and control devices attached to the Internet of Things [IoT]. As IoT devices are used to deliver and manage these services, the devices themselves are collecting usage data. Many IoT systems are not immediately recognizable as a connected device and are often deployed as an embedded or mobile application connected to the internet using a variety of wireless technologies, such as radio-frequency identification [RFID] and Bluetooth.

Users of technology now feed a constant stream of personal data to dozens of applications and devices, either consciously by using social media and communal gaming or unconsciously through their interaction with their vehicles, smart televisions, household equipment, and cyberphysical systems [CPS] deployed in their community. Sadeghi et. al. (2015) note that "smart products do not only collect data during their production but also when they are deployed and used by customers", with the data being stored either by the vendor or on the device itself.

Although exacerbated by the relatively recent explosion of IoT deployment, the privacy concerns surrounding ubiquitous data collection are not new. In 2007, a pair of researchers at The University of Texas at Austin revealed the vulnerability of private data, even data collected in a supposedly anonymous fashion. Greengard (2008, pg. 17) states that this research "proved that it was possible to identify individuals among a half-million participants by using public reviews published in the Internet Movie Database (IMDb) to identify movie ratings within Netflix's data. In fact, eight ratings along with dates were enough to provide 99% accuracy, according to the researchers". Note that this research was done 11 years ago; the technology has advanced considerably since that time.

The aggregation technique used by the University of Texas researchers is known as a linkage attack. A linkage attack combines individually non-sensitive or anonymized data with data from other records until a highly accurate aggregated profile of the user is assembled.

Lawfully deployed data analytics present no less of a threat to an individual's privacy. Over the past couple of decades, several attempts have been made to protect an individual's personal data from data mining exposure. In the United States, this type of safeguard usually falls within the scope of various governmental security standards or regulations such as HIPAA (2017), which protects the privacy of Americans' personally identifying health information. Unfortunately, these types of standards do not exist in many business contexts.

Many internet users are unaware of how exposed their personal data may be. Most users now realize that they should protect personal data residing on their computers with anti-malware software and should also be leery of unsolicited email. At the same time, internet users have been slow to realize that every time they use an

internet-connected device of any type, either consciously or unconsciously, a trail of artefacts recording that activity is left behind. Digital artifacts can include obvious, consciously created information, such as email, instant messaging, and social media posts. More insidious is unconsciously created data which includes web search histories, cell phone or Global Positioning System [GPS] locations, even metadata gleaned from commonly used office applications.

Location tracking can reveal more about an individual's activities and relationships than simply where the person is located at any given point. As the Me and My Shadow website (2017) comments:

> Location data can also be used to map out your relationships with others. If you and another person, or other people, are in the same place at specific times of the day, it's possible to infer what relationships you have with these people - if, for example, they are co-workers, lovers, roommates, or family members.
>
> Or, to take another example, if you are a government employee and are in the same cafe as a specific journalist, you could be flagged as a leaker.

Location data is collected by a myriad of devices and sensors. Mobile phones log location information and can easily serve as a GPS device. Mobile location services use the GPS chip in the phone which then communicates with standard GPS satellites. This type of tracking leaves a range of artefacts that are accessible to anyone or anything, such as phone apps, that have access to the device. Mobile devices that are capable of connecting to Wi-Fi collect additional location artefacts and keep logs of all the locations where the device connected to Wi-Fi, giving an even more detailed picture of where the device has been.

2. Examples of unconsciously provided personal data

During an average internet session, a user may be asked several times by an array of entities to provide personal data for a variety of reasons. Perhaps a user is asked to provide a valid name and address to a website to make a purchase, or the user may have to provide an e-mail address to another website to subscribe to a newsletter. These are examples of providing information consciously, but a rapidly increasing percentage of personal data is now collected unconsciously.

A growing number of online entities that were originally designed as social media platforms are evolving into large scale data analytics sites. In this trend, the social media serves as a front end for users while the genuine purpose of the site is to gather personal information both consciously and unconsciously. Drachen et. al. (2012) note that free-to-play [F2P] games provided by Facebook and Google Play aid their ability to analyse player behaviours. This analysis in turn increases potential monetization by understanding player interaction with the games. Ahmad and Srivastava (2014) state that the immersive nature of online gaming environments

encourages players to engage with the game, revealing a range of behaviours, such as cooperation or deceit. Given the exposure of player characteristics, these gaming environments provide unprecedented opportunities for studying human behaviour in a granular fashion.

Chen et. al. (2015) provide an overview of an even more widespread and invasive vector of unconsciously provided personal information. The Internet of Things integrates classically networked industrial objects with a massive number of personal objects such as household video cameras and smart televisions. While clustering behavioural patterns can provide general insight into object usage, using the public internet as a transmission conduit makes identifying individual users of objects relatively trivial.

Another major source of unconsciously provided personal information resides in mobile devices such as smartphones and tablets. Most cell phones manufactured within the last four or five years allow the handset to switch between "Wi-Fi" (IEEE 802.11 standard) and cellular radio signals. While 1st and 2nd generation cellular signal is relatively impervious to useful interception, 3rd generation signal is somewhat more vulnerable. Given that 3rd generation signal [3G] enhances the ability to transmit data as well as voice, the potential for extracting unconsciously provided personal information from both signal as well as device storage becomes obvious. Goda et. al. (2015) outline the various types of information available on a mobile device and the methods for data extraction from both signal transmission and the handset. Among the types of information that can reside in a smartphone, Casey (2011) describes:

> Although compact, these handheld devices can contain personal information including call history, text messages, e-mails, digital photographs, videos, calendar items, memos, address books, passwords, and credit card numbers. These devices can be used to communicate, exchange photographs, connect to social networks, blog, take notes, record and consume video and audio, sketch, access the Internet, and much more. As the technology develops, higher data transmission rates are allowing individuals to transfer more data (e.g., digital video), and the computing power in these devices enables us to use them in much the same way as we used laptop systems over the past decade. Because these devices fit in a pocket or bag, they are often carried wherever a person goes and can be used to determine a person's whereabouts at a particular time.

Insel (2017) comments on other human characteristics that may be derived from mobile device use. The author notes that the manner in which a device is used, including typing and scrolling patterns, can be used to deduce a range of behavioural aspects. Keyboard interaction with a mobile device can reveal the subject's reaction time, attention span, and memory reliability. This data can then be used to deduce the user's behavioural tendencies, cognition, and even mood at the time of use.

3. Integrating data analytics with digital forensics

Digital forensics, data analytics, and linkage attacks share many of the same techniques for aggregating data to form a profile of a user, varying mostly by intent of use. Digital forensics is an extension of traditional evidence gathering and is associated with the investigation of data found on a variety of devices and networks. Data analytics is not necessarily concerned with identification of an individual or that person's activities but is rather a means of deriving patterns and relationships from large data sets i.e. Big Data. Linkage attacks can be defined as the "de-anonymization" of data provided by or about an individual, derived by analysing data relationships contained in two or more unrelated databases. Merener (2012, pg. 378) comments that anonymization typically removes "the variables that uniquely associate records to the corresponding individuals, such as name, email address, social security number". The author then notes that this step is insufficient to prevent algorithmic linkage of other variables in the data that can still serve as a type of digital fingerprint.

Shen et. al. (2014) discuss a more targeted type of linkage attack, a "User Identity Linkage Attack", which can link an individual's information across a range of online social networks. When data analytics and/or linkage attack methods are combined with digital forensics, the raw data derived by use of forensic techniques can then be analysed with a high degree of accuracy regarding the identity and activities of the individual under analysis. In addition, because a major goal of digital forensics is to provide nonrepudiation of evidence, the resulting profile gained by combining these tools is unlikely to have been modified by the individual, which lends a greater degree of veracity to the profile.

Digital forensics adhere to the legal rules of evidence as practiced in the location where the forensics operation takes place. Traditionally, digital forensics has been used to collect data that a suspect may have stored on physical media, such as a hard drive. Where digital forensics begin to raise concerns regarding an individual's passive digital footprint is when it is combined with sensor data provided by IoT devices. Caviglione et. al. (2017) note that common types of IoT implementations create an intersection between the digital world and the physical world through the use of sensors. The authors state:

> For example, IoT nodes can provide evidence of when a person was present in a room by investigating in-door presence sensor values. Obviously, such investigations are linked to further privacy issues, especially as sensors might be influenced not only by a single user but by an undefined set of influencers: several individuals could trigger a presence sensor in a room each day, not just the potential criminal.

This blurring of personal data as it exists in the digital world with its physical counterpart allows the investigator to assemble an extremely accurate portrait of a suspect. If this portrait were placed in the hands of a corporate or governmental entity, the privacy implications are disturbing.

4. Does unconsciously provided personal information reveal the "authentic self"?

Users of social media consciously craft an online identity and may compile a set of differing identities based on what the user perceives as different audiences. Matic's (2011, pg. 20) research indicates that "users constantly assess their online environment and based on their assessment they construct their own Internet playground where they choose roles they find suitable". So, what would unconsciously provided personal information reveal about an individual? Linking and analysing relationships between an individual's differing self-constructed identities would provide insights into that individual's consciously provided self-perception. If unconsciously provided personal information is added to that analysis, as is commonly done in digital forensics, the elements of intent, motivation, and exposure of deception come into play.

5. Discussion

Williams and Neal (2012) point out the similarities in technique between data mining and linkage attacks, stating that both types of information gathering and analysis pose a similar risk of personal data exposure. When digital forensics techniques are added, the combination of consciously and unconsciously generated data makes identifying an individual and that individual's activities, behaviours, and motivation relatively straightforward.

Analysis of a job candidate's social media information has become an accepted practice for human resource departments in the United States. Data analytics is also increasingly used by educational institutions to guide students toward courses and subjects where they may have a greater chance of academic success, based on their individual characteristics as revealed by the data. As Johnson (2014) notes, there are inherent ethical dilemmas in this type of "technosocial" system, which can include reducing an individual's autonomy over their own aggregated profile. Pistilli et. al. (2014) go further to state that acting on insights gained from data analytics presents the possibility of social engineering, which may or may not be advantageous or helpful to the individual whose data is analysed.

The use of continually refined data analytics, combined with the possible inclusion of digital forensics to add unconsciously provided data to the digital profile, provides a path to discovering more about an individual than would be uncovered through consciously provided information. Indeed, the combined data would likely provide deeper insight into an individual's characteristics, motivations, and actions than personal introspection might reveal. This ability to track a person's actions across a broad range of devices is unparalleled in history. The unguarded nature of much of this data provides an unprecedented opportunity to observe an individual's behaviour in a deeper manner than has ever been possible before.

Fairfield and Shtein (2014, pg. 39) state "[t]he nature of big data technology fundamentally challenges traditional methods of framing ethical principles". The traditional ethical and legal definitions of privacy are no longer sufficient for dealing with the overwhelming quantity and detail of personal data being generated by individuals both consciously and unconsciously. The changing nature of Big Data and the resulting analytically derived information will require legislators and businesses to develop more flexible and sophisticated ethical constructs to deal with the sheer quantity of personal data available, as well as the minefield of revelation embedded within that data.

5. References

Ahmad, M.A., & Srivastava, J. Behavioral data mining and network analysis in massive online games. *Proceedings of the 7th ACM International Conference on Web Search and Data Mining*, 673-674, 2014. doi:10.1145/2556195.2556196

Akamai's State of the Internet Report Q1 2016. (2016). Retrieved from https://www.akamai.com/uk/en/multimedia/documents/state-of-the-internet/akamai-state-of-the-internet-report-q1-2016.pdf

Casey, I. *Digital evidence and computer crime: Forensic science, computers and the Internet*. 2011. Waltham, MA: Academic Press.

Caviglione, L., Wendzel, S. and Mazurczyk, W. (2017). The Future of Digital Forensics: Challenges and the Road Ahead. IEEE Security & Privacy, 15(6), pp.12-17.

Chen, F., Deng, P., Wan, J., Zhang, D., Vasilakos, A.V., & Rong, X. Data mining for the Internet of Things: Literature review and challenges. *International Journal of Distributed Sensor Networks - Special Issue on Big Data and Knowledge Extraction for Cyber-Physical Systems*, 2015. doi:10.1155/2015/431047

Drachen, A., Sifa, R., Bauckhage, C., & Thurau, C. Guns, swords and data: Clustering of player behavior in computer games in the wild. *2012 IEEE Conference on Computational Intelligence and Games (CIG)*. doi:10.1109/cig.2012.6374152

Fairfield, J. & Shtein, H. (2014) Big Data, Big Problems: Emerging Issues in the Ethics of Data Science and Journalism, *Journal of Mass Media Ethics*, 29:1, 38-51.

Greengard, S. Privacy matters. *Communications of the ACM 2008, 51*(9), 17-18. doi:10.1145/1378727.1378734

Goda, B.S., Bair, J.W., & Costarella, C.E.. Cell Phone Forensics. *Proceedings of the 16th Annual Conference on Information Technology Education - SIGITE '15, 2015*. doi:10.1145/2808006.2808022

Health Information Privacy. Retrieved from http://www.hhs.gov/hipaa

Insel, T. R., 2017. Digital Phenotyping: Technology for a New Science of Behavior. *Journal of the American Medical Association*, 318:13, 1215 - 1216.

Johnson, J.A., 2014. The Ethics of Big Data in Higher Education. *International Review of Information Ethics*, 7, pp.3–10.

Matic, I., 2011. The Social Construction of Mediated Experience and Self Identity in Social Networking. *The International Journal of Interdisciplinary Social Sciences*, 5, 13 - 21.

Me and My Shadow. 2017. Location tracking. [ONLINE] Available at: https://myshadow.org/location-tracking. [Accessed 9 May 2018].

Merener, M.M. Theoretical Results on De-Anonymization via Linkage Attacks. *Transactions on Data Privacy*, 377-402, 2012. Retrieved November 2, 2017, from http://www.tdp.cat/issues11/tdp.a074a11.pdf

Pistilli, M.D., Willis, J.E., Campbell, J.P. (2014) Analytics Through an Institutional Lens: Definition, Theory, Design, and Impact. In: Larusson J., White B. (eds) Learning Analytics. Springer, New York, NY

Sadeghi, A.R., Wachsmann, C., & Waidner, M. *Security and privacy challenges in industrial internet of things*. Paper presented at the Proceedings of the 52nd Annual Design Automation Conference, San Francisco, California 2015.

Shen, Y., Wang, F. and Jin, H. (2014). Defending against User Identity Linkage Attack across Multiple Online Social Networks. In: *WWW'14 Companion*. Seoul, Korea: ACM, pp.375, 376.

Williams, L. and Neal, D. (2012). The digital aggregated self: A literature review. In: The International Conference on Cyber-enabled Distributed Computing and Knowledge Discovery 2012. Sanya, China: IEEE, pp.170 - 177.