IST-Africa
IST-Africa 2017 Conference Proceedings
Paul Cunningham and Miriam Cunningham (Eds)
IIMC International Information Management Corporation, 2017
ISBN: 978-1-905824-57-1

Towards Quantifying and Defining Privacy Metrics for Online Users

Frans F BLAUW[1], Sebastiaan VON SOLMS[2]
University of Johannesburg, Cnr Kingsway and University Road,
Johannesburg, 2092, South Africa
[1]*Tel: +27 11 559 3241, Email: fblauw@uj.ac.za*
[2]*Tel: +27 11 559 2843, Email: basievs@uj.ac.za*

Abstract: With personal information being stored on information systems, there exists a need to determine the amount of information that is obtainable for an individual. In this paper, the authors work toward measuring the personal information of an individual as stored within different levels within multiple information systems. The proposal takes into account different identity elements and assigned a privacy score and weighting to it. This is done in order to quantify an individual's privacy. A total privacy score can then be assigned to an individual for them to ascertain what their level of "visibility" is on various information systems accessible on and off the Internet. A user can then make use of this score to further a potential risk assessment of their personal information.

Keywords: privacy, privacy metrics, identity, identity elements

1. Privacy – The Concern

Our information can be found all over the Internet in various forms. Whether intentionally published by ourselves on a social network or whether as part of a restricted Internet facing information system, the information is there. However, with our information so widely spread, it is difficult to get an overall image of what is available.

Persons with a better understanding of technology might often be able to estimate what is available online about them. However, in countries where technology is not widely spread and, in particular, the necessary education surround technology is not widely available, there might exist a particular ignorance with regard to an individual's online visibility and privacy.

To address this need, we will attempt to define a privacy metric that can quantify our level of privacy online. In short, we will give a number that defines how "visible" an individual is online.

To do so, we will first consult the legal aspect of privacy. This will show what the law has to say about privacy and an individual's right to privacy. Next, we will look at the general scope of the accessibility of information online by looking at the nonliteral layers of the Internet. The idea where the Internet is subdivided into different layers of information availability will be explored.

As this is not a brand new problem we are looking at, we will also see how others have attempted to quantify privacy before moving on the defining our own metrics.

This paper forms part of the initial stages of a much larger research subject.

2. Objectives

A need exists to quantify online privacy. In this paper, we will define a privacy metric that can place a value on the visibility of a person online. Visibility, similar to its dictionary

definition, will define an individual's "state of being […] seen" and the "degree to which [someone] has attracted general attention; prominence" [1].

This privacy metric should provide a simple, straight-forward means of calculating the visibility. A numerical value should be able to give an individual a concrete classification of their online visibility.

Let us start with the first building block we will require, the identity elements.

3. Legal Definition of Personal Information

As privacy of information is becoming more a legal issue, we will investigate what the law states about information privacy. For this purpose, we will investigate South Africa's Protection of Personal Information (POPI) act of 2013 [2], which bears a striking resemblance to the United Kingdom's Data Protection Act of 1998 [3].

The POPI Act applies to all companies that collect and process personal information. "Personal Information" has a very broad definition within the POPI Act. Any information (or as we will call it, an identity element) that can be used to identify any natural (or juristic) person is included within this definition. This includes basic elements such as race; gender; marital status; national, ethnic, social origin; age, sexual orientation, religion, language, etc.

Also included in this definition are histories such as education, medical, financial, employment and criminal history. Numbers and symbols that can be used to identity a person, such as email address, physical address, and telephone numbers are also obviously included.

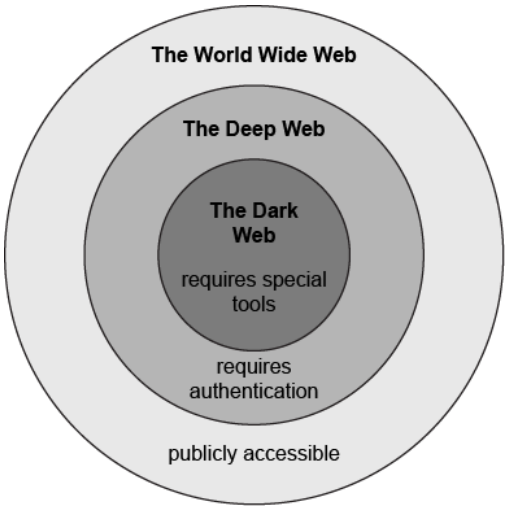
Finally, vaguer elements such as opinions and preferences can, according to the POPI Act, also be seen as a personal information and, as such, protected under law.

Using this definition, and its extremely broad scope of what can be considered personal information (or identity element), we will construct our classifications of identity elements for the purpose of our metric definition.

4. Scope of Visibility

On the Internet, there are generally considered to be three layers. First is the Public World Wide Web (WWW), the Deep Web and then the Dark Web [4]. Figure 1 shows a graphical representation of these layers.

Figure 1: Logical layering of the Internet



The WWW is the first point of contact of anyone with an active connection to the Internet. Information is commonly indexable by search engines (such as Google and Bing)

and thus publically accessible. This information made available on the WWW is usually done so voluntarily for the purpose of the explicit dissemination of such information.

The next layer is the Deep Web. Often confused with the Dark Web, the Deep Web is merely the part of the Internet that is not indexable and not directly accessible by the public. Normally information in the Deep Web is restricted by a portal that requires some sort of authentication and as such only accessible by users with the required authorisation.

The final layer is the Dark Web. As the name suggests, this part of the Internet is intentionally hidden from anyone without a standard browser and requires special tools (such as the TOR Browser) to be accessed.

4.1 The World Wide Web (WWW)

As we said, information on the WWW is generally made available with the explicit purpose of sharing this information with the general public. In the category, there exist sites that share information such as encyclopaedias (e.g. Wikipedia); companies that share products, services and their contact information (e.g. Amazon); social media sites (e.g. Facebook, Twitter, YouTube, etc.) allow users to share any information they desire; and news distribution sites (e.g. CNN).

Many of the sites in this category can have their information restricted and can thus partly fall under the Deep Web, discussed next.

4.2 The Deep Web

There exists a lot more information in the Deep Web. This information can be shared with explicitly defined users or it exists for the sole purpose of specific users to be able to complete some task. In this category, there are company portals that can only be accessed and used by employees of a specific company; online banking sites that allow authorised users to complete transactions online; and restricted information systems that allow clients to access a service they are authorised to use. Information restricted on the Deep Web is generally never made public.

As mentioned before some sites, such as social media sites and online commerce, can exist in both the WWW and Deep Web depending on the specific requirement or privacy settings of the user. For example, on social media, a user can choose whether their post (Facebook [5]), tweet (Twitter [6]) or video (YouTube [7]) is made available to anyone on the Internet or only a restricted subset.

Commerce sites generally showcase their products on the WWW, but specific information pertaining to a specific user (previous transactions, payment options, etc.) only exists behind a restricted portal.

4.2 The Dark Web

On the Dark Web, the same types of services that exist on the WWW and the Deep Web can exist. However, the Dark Web is generally more notorious for having caches of illegally obtained personal information. This can be information such as stolen or leaked credit card information, or personal identification information. These caches can be bought or swapped.

Generally, one would not want to find your information on the Dark Web as it can then be used for illegal purposes. [8]

5. Related Works

Effort has been made to quantify a person's online privacy and visibility. However, the majority of explicit metrics rely on surveys or information that is explicitly made available by an individual on, for example, social networks. We will briefly look at two distinct examples.

Braunstein et al. at Google [9] proposed the idea of "measuring privacy without asking about it" by using privacy surveys. They found that something as simple as the wording in a privacy survey could impact the outcome of the survey as was not the best means to determine a user's privacy online. As such, a more objective means that measure privacy needs to exist without the need of a user's input or opinion.

For social networks Becket et al. [10] attempted to define a metric for quantifying an individual's privacy on a social network. They attempted to provide a framework for identity privacy risk on social networks and thus reduce information loss. It does this by inferring attributes of a user based on those of their friends. Attributes are personal information such as age, employer, relationship status, etc., to which weights are applied.

We will use this idea of attributes with specific weights for our proposed metric (in the next section).

6. Proposed Metrics

For our proposed metrics, we will attempt to quantify how much personal information of a person is on the Internet – be it in the public WWW, Deep Web, or Dark Web.

Our privacy metric starts at zero and increases as more information is available online. And so, the closer to zero the more "invisible" the person will be according to the metric.

Our metric consists of two parts, both regarding the identity element. The first is the visibility of the identity exposure. Here we look at the exposure, or visibility, of the identity elements. The second part looks at each identity element itself.

6.1 Identity Exposure

The exposure of identity elements relates to how easily each identity element can be accessed. The accessibility of the element will be used as a weighting for each individual identity element (discussed in the next subsection).

Looking back at Section 4 of this paper, we see that the Internet itself is inherently divided into separate layers. We use the idea of these layers to define our weightings. In Table 1 below, we see each visibility and its weighting.

Table 1: Identity Exposure Weightings

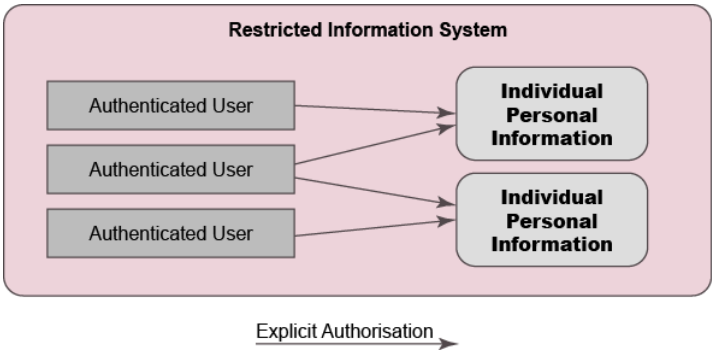
Visibility	Weighting
Private with Full Authentication	1.0
Private with Part Authentication	1.5
Private with Public Authentication	3.0
Public with Authentication	4.0
Public	4.5
Unknowingly Obtained	6.0

Private with Full Authentication

This weighting refers to information systems that are fully controlled by authentication methods. Access to an individual's information also needs to be explicitly granted – i.e. authorisation to each individual's information needs to be granted on a per-individual level.

Information stored here has no inherent way of becoming public (unless stolen, of course). These systems include company information systems and government information systems.

Figure 2: A restricted information that requires explicit authentication

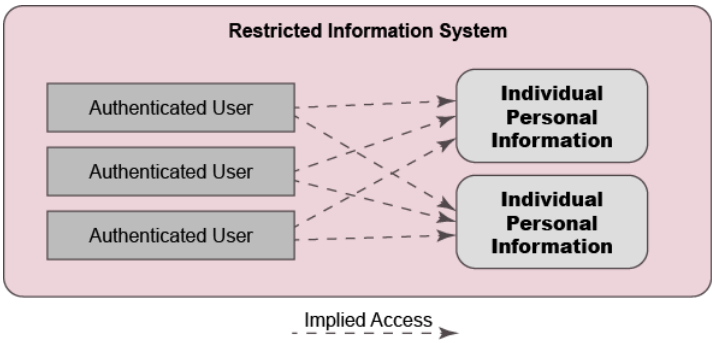


Private with Part Authentication

This weighting is similar to the above, but differs in that authentication is only required to the system itself, and not to each individual's records. That is, an entire group of authorised users can access each record on the system, with no known way to explicitly define granular access.

As this layer is slightly less controlled, it carried a higher weighting but is still considered secure.

Figure 3: A restricted information system that allows implies access



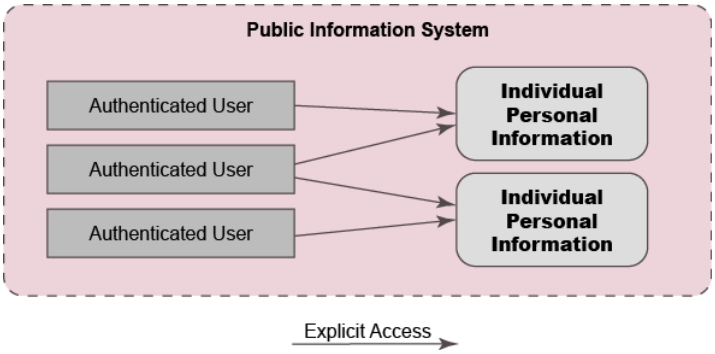
Private with Public Authentication

This layer might sound somewhat like an oxymoron, however, this refers to information that an individual themselves decided to make private and only authorise access to a select group of people, who require an easily obtained public account.

For example, on Facebook, a post can be made private to anyone except the user's friends, but such a friend can sign up for a Facebook account themselves without requiring a lot of authentication.

At this level, we are already looking at information that can very easily be made public, and will thus carry a higher weighting.

Figure 4: A public information system where information needs be explicitly granted access

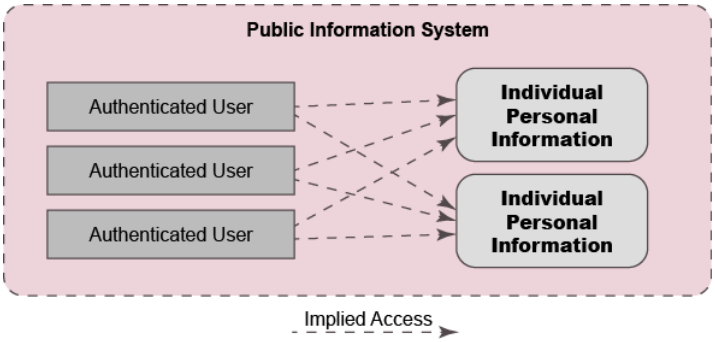


Public with Authentication

In this layer, anyone from the public can access the information, but a (normally free) user account is required. As such, it might be easier to track who accessed whose information.

Identity elements on this layer include information stored on social networks that require an account before information can be accessed. For example, Facebook requires an account before an individual's "public" posts can be viewed.

Figure 5: A publically available information system with open access to information, but requires authentication

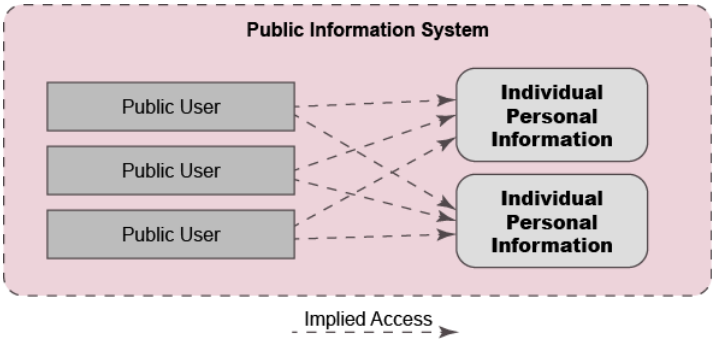


Public

This layer is fairly self-explanatory. Information on this layer has explicitly been made available to the public for use.

Regardless of intention, information is publically available and will thus carry a much higher exposure weighting.

Figure 6: A publically available information system with open access to information, no authentication required



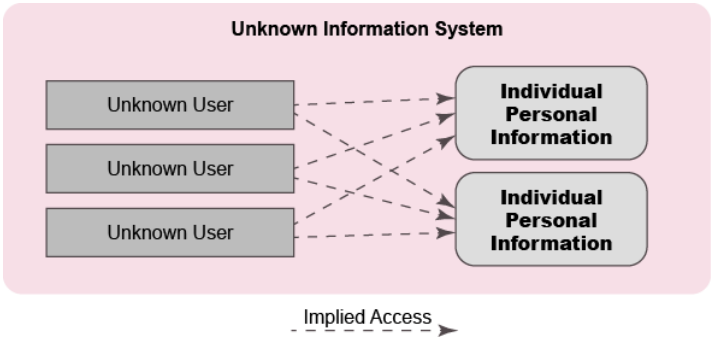
Unknowingly Obtained

As this layer's name suggests, the information here has been obtained (and possibly distributed) without the knowledge of the person in question. Here we look at information

that has been stolen (i.e. obtained through any unauthorised means) with no way to track the whereabouts of the information.

Regardless of intention or whether the data has been distributed to other parties, it can be seen that this data is the most vulnerable and carries the greatest exposure weighting.

Figure 7: Information that has been unknowingly obtained, and access is completely uncontrolled



6.2 Identity Elements

Next, we look at the specific identity elements and assign a point value to them. These points can then be multiplied using the weightings we described in the previous subsection. In Table 2 below, we can see a set of identity elements that we have identified that will generally be online.

Table 2: Identity Element Points

Identity Element	Points
Full Name	1
Gender	1
Race	1
Birthdate	1
Medical History (per line)	5
Criminal History (per indictment)	5
Employment Information	3
Phone Number	3
Passport Number	10
Personal Identification Number	10
Home Address	8
Financial History	8
Credit Card (partial)	13
Credit Card (full)	15
Family Information	5
Opinions	3

Elements that are easily obtained or deduced (such as name or gender and other demographics), we assign a lower point value.

Histories of a person – financial, employment, and criminal – will have a slightly higher value.

Information that can actively be used – contact details, passport details, and addresses – will carry an even higher point value.

Finally, we place the highest values on elements that are to do with finance. This includes elements such as credit card details.

For each identity element that is available on each unique, distinct site, the point value is weighted based on the exposure of the information on that site.

These point values assigned can be adjusted based on usage scenario and risk factor.

7. Case Example

Let us consider the two cases of Samantha and Ayabonga.

Samantha considers herself tech-savvy. She uses ABSA's online banking system, buys fairly regularly on Amazon.com, she also has a medical aid and uses the online portal of Discovery Medical Aid and is fairly active on social media. Her Facebook account is restricted to her friends online, however her Twitter is completely public. Considering these criteria, using our metrics, we can see Samantha's visibility in Table 3 and she has a total visibility of 2911 points.

Table 3: Samantha's Online Visibility

		Qty	Pts	Wgt	Weighted Total
ABSA Bank	Full Name	1	1	1.0	1
	Financial History	312	8	1.0	2496
Amazon	Full Name	1	1	1.0	1
	Gender	1	1	1.0	1
	Birthdate	1	1	1.0	1
	Phone Number	2	3	1.0	6
	Home Address	1	8	1.0	8
	Credit Card (full)	2	15	1.0	30
	Transactions	18	3	1.0	54
Discovery	Full Name	1	1	1.0	1
	Gender	1	1	1.0	1
	Race	1	1	1.0	1
	Birthdate	1	1	1.0	1
	Phone Number	2	3	1.0	6
	Home Address	1	8	1.0	8
	Family Information	3	5	1.0	15
	Medical History (per line)	56	5	1.0	280
Facebook	Full Name	1	1	4.5	4.5
	Gender	1	1	4.0	4
	Race	1	1	4.0	4
	Birthdate	1	1	3.0	3
	Phone Number	1	3	3.0	9
	Opinions	56	3	3.0	504
Twitter	Full Name	1	1	4.5	4.5
	Opinions	1020	3	4.5	13770
TOTAL					**2911**

Next, we have Ayabonga. Ayabonga is a father of five and grandfather of 15 with whom he wants to keep contact. For this, he has a Facebook account to see what they do. He also uses the online portal of MTN to keep track of his airtime; however, he doesn't use it very often, his grandson set it up for him. We can see Ayabonga's visibility in Table 4 and he has a total visibility of 356.5.

Table 4: Ayabonga's Online Visibility

		Qty	Pts	Wgt	Weighted Total
Facebook	Full Name	1	1	4.5	4.5
	Gender	1	1	4.0	4
	Race	1	1	4.0	4
	Opinions	14	3	3.0	126
MTN	Full Name	1	1	1.0	1
	Phone Number	1	3	1.0	3
	South African ID	1	10	1.0	10
	Transactions	68	3	1.0	204
TOTAL					**356.5**

Even though it would have been fairly obvious that Samantha's would have a great online visibility, now we have a measured value of her visibility in comparison to Ayabonga's.

8. Future Research

The paper is truly only a stepping stone onto future research. With the knowledge gained from defining the metrics, we can continue investigating and setting up more finely grained metrics. We also intend to investigate what users deem to be important metrics by means of surveys, but also how information can be used against users.

What we learned from all these endeavours will then form of a larger investigation into a system that could potentially manage these metrics for users.

9. Conclusion

In this paper, we defined a metric for measuring an individual's online visibility in terms of their personal information. Looking at the legal definitions of personal information, we proceeded to define a subset of personal information for which we assigned scores.

We then took into account the visibility of that platform their personal information is shared on, and assigned a weighting to it. Tallying all the points, we can see a measured value of an individual's online visibility or exposure.

Future work on this will firstly refine the points and weighting system. Looking at the impact of different platforms will factor into the weighting of the platform. The actionability of personal information will also affect its point value.

The practicality of locating and classifying all the personal information held by different services will be addressed in future work. We felt it important to first look at what to do with this data once retrieved.

Providing a measurable metric for online privacy and exposure will no doubt give rise to further developments in privacy, but also the awareness of personal information available on a variety of information systems.

References

[1] Definition of *visibility*. Oxford English Living Dictionaries. Available from: https://en.oxforddictionaries.com/definition/visibility

[2] Republic of South Africa. (2013). Act No. 4 of 2013: Protection of Personal Information Act, 2013. Available from: http://www.justice.gov.za/legislation/acts/2013-004.pdf

[3] United Kingdom of Great Britain and Northern Ireland. (1998). Data Protection Act 1998. Available from: http://www.legislation.gov.uk/ukpga/1998/29/contents

[4] BrightPlanet. Clearing Up Confusion – Deep Web vs. Dark Web. Available from: https://brightplanet.com/2014/03/clearing-confusion-deep-web-vs-dark-web/

[5] Facebook. Basic Privacy Settings & Tools. Available from: https://www.facebook.com/help/325807937506242

[6] Twitter. Protecting and unprotecting your Tweets. Available from: https://support.twitter.com/articles/20169886

[7] YouTube. Change the privacy settings for your video. Available from: https://support.google.com/youtube/answer/157177

[8] Kassner, M. (2016). From the dark web to the 'open' web: What happens to stolen data. Available from: http://www.techrepublic.com/article/from-the-dark-web-to-the-open-web-what-happens-to-stolen-data/

[9] Braunstein A, Granka, L, Staddon, J. (2011). Indirect Content Privacy Surveys: Measuring Privacy Without Asking About It. Proceedings of the Seventh Symposium on Usable Privacy and Security

[10] Becker, J.L., Chen, H. (2009). Measuring privacy risk in online social networks. Web 2.0 Security & Privacy