# Digital Footprints: Your Unique Identity

Juanita Blue
University of Ulster
Derry, Northern Ireland, UK
*boyle-j2@ulster.ac.uk*

Joan Condell
University of Ulster
Derry, Northern Ireland, UK
*j.condell@ulster.ac.uk*

Tom Lunney
University of Ulster
Derry, Northern Ireland, UK
*tf.lunney@ulster.ac.uk*

**In the digital age where Human Computer Interaction is creating large and entirely unique digital footprints, online accounts and activities can prove to be a valuable source of information that may contribute to verification that an asserted identity is genuine. Online social contextual data – or 'Digital identities' -- pertaining to real people are built over time and bolstered by associated accounts, relationships and attributes. This data is difficult to fake and therefore may have the capacity to provide proof of a 'real' identity. This paper outlines the design and initial development of a solution that utilizes data sourced from an individual's digital footprint to assess the likelihood that it pertains to a 'real' identity. This is achieved through application of machine learning and Bayesian probabilistic modelling *techniques.***

*Identity; Authentication; Digital Footprint, Privacy; Security*

## 1. INTRODUCTION

As individuals spend increasing amounts of time interacting online, the line between their physical lives and their digital lives is becoming increasingly blurred. Digital footprints map and record the activities that substantiate each individual's online life. This automated trail logs the habits, interests, events, relationships and communications that are intertwined with an individual's physical life. It combines this information with metadata to produce an entirely unique blend of information that has the capacity to prove that you possess a true identity. It achieves this though defined patterns and repetition that demonstrate that there is only one you and there are many components that represent you digitally.

The digital age has witnessed a significant rise in identity theft, where perpetrators use the identities of others to essentially violate the law (Spalevic,& Ilic, 2017). This is an important issue, as fake identities present terrorists and criminals with the opportunity to commit various types of crime, while concealing their true identities (Kean, et al., 2004).

An individual's true identity is comprised of basic components including a personal identity, represented by standard identifiers and also a social identity. A social identity refers to a person's biographical history that gathers over their lifetime (Vignoles, 2017). Research conducted in the area by Wang et al. has indicated that the use of non-standard attributes that relate to an individual's social behavior may contribute to the authentication or refutation of identities within identity resolution techniques that indicate where identities are fake (Wang, et al., 2006) (Li, et al., 2010). In a cyber context, this social identity is demonstrated via an individual's digital footprint. Therefore, it may be surmised that the same behaviours possess the potential to aid authentication of true identities.

This paper outlines the initial design and development stages of smart digital identification, where data sourced from an individual's 'digital footprint' is analysed through machine learning and probabilistic modelling to ascertain the likelihood of a 'true' identity.

## 2. BACKGROUND

In seeking a solution to prove identity through a digital footprint, various associated areas were explored. These included identity resolution, digital footprints and Bayesian probability.

### 2.1 Identity Resolution

Identity resolution is a process of semantic reconciliation that determines whether a single identity is the same when being described differently (Fish, 2009). Conventional records consist of multiple attributes (Wang, et al., 2006) (Li, et al., 2010), identity resolution identifies where two records relate to one individual by comparing the content of individual corresponding fields (Köpcke & Rahm, 2010). However, the accuracy of these attributes cannot be relied upon (Li, et al., 2010) and thus, they do not present a reliable source of information against which identity authentication can be performed. Identity resolution is used largely to detect identity theft and fraud.

Machine learned techniques automatically extract patterns and identify annotated matches. Distance/similarity measures between two records are defined for various attributes, resulting in the output of an over-all 'distance score'. Li, et al. developed an algorithm for detection of fake identities by comparison of several personal identifiers; combining them to produce a similarity score (Li, et al., 2010).

### 2.2 Social Contextual Data & Identity Resolution

Modern sociological literature indicates that two components form individual's identity: a personal identity and a social identity. An individual's personal identity is acquired from birth and includes identifiers such as name and date of birth; officially assigned identifiers such as a national security number (NSN); current physical descriptions such as height and weight and also biometric data such as fingerprints. A social identity is a person's biographical history, gathered over their lifetime (Vignoles, 2017), describing the social context of their life experience. Incorporating both these aspects allows for a more comprehensive understanding of identity.

In deviating from the utilization of traditional identifiers, an individual's social contextual information possesses attributes that authenticate their undeniable identity. Recent studies have recognized the value of social context data such as relationships and social behaviours in identity resolution. Identity matching through social behaviour and social relationship features was developed by Li et al. in 2010 (Li, et al., 2010). Köpcke and Rahm also devised a categorical scheme that considered attribute-value-matchers that rely only on attributes that are descriptive and contextual matching to examine data gathered from social interaction links (Köpcke & Rahm, 2010).

### 2.3 Digital Footprint

In the current world, individuals now possess two identities. A "real world identity", that is verified by official paper documentation, as well as a "digital identity", that is defined by an individual's use of the internet, including search history, online services, forums, blogs, and social media (Park, 2017). This use extends to create links to the real life identity of an individual. A digital footprint represents an individual's online presence and provides evidence of their digital and real world identities. It logs the trail and artifacts left behind by individuals interacting in a digital setting (Fish, 2009). Digital footprints are persistent and link the past with the present, regardless of transitions and changes in an individual's life (Haimson, et al., 2016).

Online accounts provide many verified links to the attributes of real identities. Often these attributes are recounted across multiple accounts and sources. Almost every online account that is created requires an email address. Official online services require personal identifiers such as name, DOB, address and unique personal identifying numbers such as an NSN. Online shopping requires a postal address and payment details, searching the internet and the use of Google Maps often involves the use of an individual's current GPS location and potentially where they will be in the future and social media accounts represent confirmation of contacts, relationships (Xiang, et al., 2016) and professional and personal interests. Across several sources the same information relating to an individual is stored, reiterated and relied upon to conduct the simple tasks that form the operation of an individual's daily life.

A digital footprint is created unknowingly and with ease through automated logging such as the storage of cookies that has become an accepted aspect of being 'online'. However, it's direct descendent, a digital identity, in the social contextual form, is not so easily gained. The construction of a digital identity or reputation across multiple sources takes a significant amount of time to gather and its links to an individual's real world identity make it difficult to fake (Haimson, et al., 2016), as it is relied upon to conduct daily tasks. It requires multiple participants as it intertwines with external and official entities and it is bolstered by electronic records, email notifications, digital receipts, the lives of others and the metadata that forms components of the digital footprint and used to trace and record every online move.

As digital footprints map and record more and more aspects of an individual's real world life, they offer information and attributes that can be used to verify, validate and authenticate real identities, whilst also refuting those that are fake. These attributes include name, DOB, home address, phone number, email address, GPS locations, timestamps, financial information, professional affiliations, social relationships, personal health information, purchases, habits, interests and much, much more.

### 2.4 Probability

"Probability is the branch of mathematics that studies the possible outcomes of given events together with the outcomes' relative likelihoods and distributions" (Weisstein, 2017). Within this mathematical branch, Bayes Theorem/Rule provides a method of calculating the probability of an occurrence, when given the probability of another occurrence. Bayes Rule is used to relate conditionals of the format p(x|y) to the inverse, p(y|x) (Bernado & Smith, 2017).

$$p(x \mid y) = \frac{p(y \mid x)p(x)}{p(y)} = \frac{p(y \mid x)p(x)}{\sum_{x'} p(y \mid x')p(x')}$$

In simple terms this is the basis for Bayesian Inference where Bayesian networks can calculate the chance of something occurring based other related information. Bayesian networks are commonly used in diagnostic systems where information is incomplete (Fox, et al. 2017).

## 3. METHODOLOGY

Bayesi-Chain aims to provide an intelligent method of secure and tamper-proof identity authentication. This

is achieved by utilizing non-standard identity attributes sourced from an individual's digital footprint and encapsulating it in a blockchain inspired secure ledger. This solution is intended to be 'opt-in', meaning that individuals will submit their own data in order to potentially obtain a Bayesi-Chain Digital Identification document. Figure 1 depicts an overview of the design of the solution; a detailed description of each step follows.

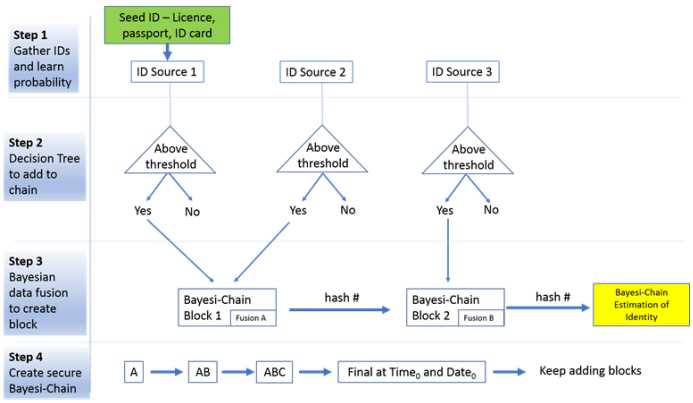


***Figure 1:*** *Bayesi-Chain Design Overview*

Combined methods of machine learning and probabilistic modelling will be utilized to estimate the reliability of identity sources pertaining to an individual's digital footprint; whereby weighted values associated with each of the attributes belonging to individual identity sources will produce an estimated score. Where the score is above an intelligently predetermined threshold, the combined standard and non-standard attributes will be fused into a 'block' with the current time and date stamp.

As additional identity sources are added to blocks, they will be combined with the previous 'block' by way of one-way hashing, producing a new hash value with each subsequent block. Further probabilistic modelling will be applied to estimate the overall 'Bayesi-Chain Estimation of Identity' score. An increased number of blocks in the chain results in a higher score. The higher the score, the higher the likelihood that the digital identification presented is 'real' and a valid authenticator of a true identity.

## 3.1 Step 1: Identity Sources & Identity Attributes

This section provides detail on Step 1 as depicted in Figure 1. It categorises various identity sources associated with an individual's digital footprint and details the associated non-standard attributes. It describes how identity sources and attributes may be weighted and identifies how correlations and commonalities between attributes from multiple identity sources possess the capacity to verify personal details and authenticate identity. Furthermore there is an overview that provides explanation of extracting an estimated reliability score from identity sources and their attributes.

### *3.1.1 Identity Sources*

An individual's digital footprint represents their online presence and is comprised of all their activities conducted on the internet. These activities span from cursory searches using engines such as Google or viewing a movie on Netflix, to more important tasks that facilitate the daily operation of life, such as communication by email or performing online banking tasks. It must be noted that any type of data associated with online accounts that is built up over time acts collectively as an efficient verifier of identity. The mere presence of data gathered over months and years is valuable, not necessarily the content of the data. For instance, an individual who wears a fitbit will gather data over extended periods including their heartrate, sleeping patterns, exercise regime and GPS locations. The existence of this extensive data that is also associated with other attributes such as their email address and mobile device including MAC address bolsters the probability that this data relates to a true identity.

Online activities provide varying degrees of valuable information. Based on the common requirement to create an online account or profile, certain sources may be considered more reliable than others. Online accounts such as online banking that previously required manual verification of paper identity documentation, or an online phone bill account that link to real-world information and payment card details may be considered more reliable than a social media account or subscription.

### *3.1.2 Identity Attributes*

Attributes provided by various identity sources can provide valuable links that verify information pertaining to an individual's true identity, in addition they provide validation of standard attributes such as residential address and phone number.

The unique set of attributes possessed by each source can be weighted based on the categorised reliability of the source. This will determine the weight that may be applied to each attribute.

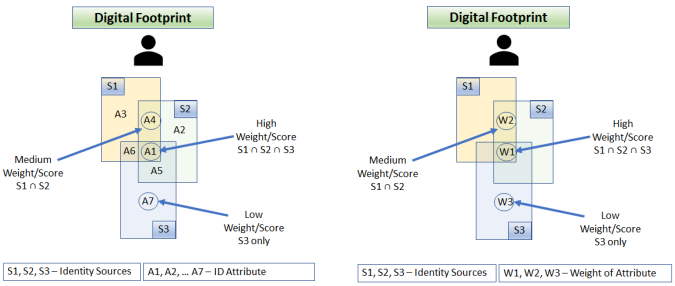


***Figure 2: Identity Sources with Common Attributes***

Figure 2 shows where attributes are found to be common across several identity sources, their score can subsequently be increased as their frequency increases the likelihood that they are associated with a true identity.

Identical attributes that are repeatedly associated with an individual's identity will be bolstered by their common use. These attributes will experience an increase in score, while attributes not verified by alternate data sources will gain no additional score. An example of an attribute that is common across several identity sources is an email address linked to multiple online accounts.

*3.1.3 Weighting of Identity Sources Attributes*

Initially standard weights sourced from 'Police Vetting' forms may be applied to input the reliability of various identity sources, based on the subset of attribute types they possess. Weighting of identity attributes will be intelligently calculated using machine learning techniques. This calculation will be based on the weight of the identity source, the type of attribute and the value's frequency of presence across multiple identity sources.

*3.1.4 Calculating Probability Values Sources*

To calculate the estimated reliability score for each identity source, Bayesian probabilistic modelling will be applied to the associated subset of attributes and their learned weights. This will output an estimated value for the reliability of the identity source. As the reliability of further identity sources is estimated, machine learning will again be invoked to identify a suitable threshold to determine if an identity source is considered reliable.

### 3.2 Step 2: Reliability Threshold & Decision Tree

This section provides detail on the process involved in Step 2, depicted in Figure 3. Inclusion of identity sources within the final Bayesi-Chain digital identification document will be dependent on the intelligently determined threshold.

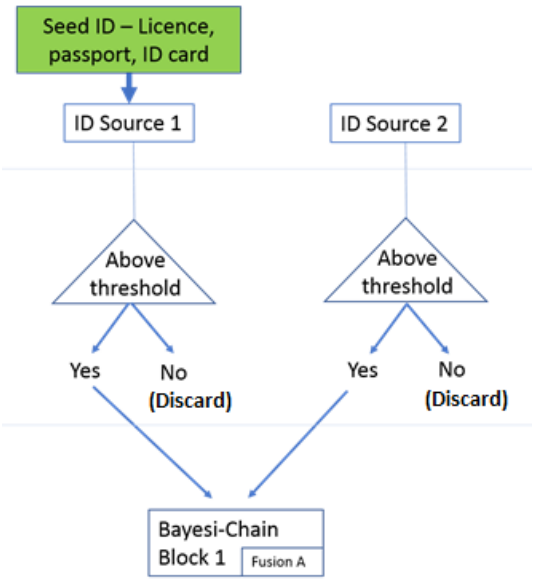


***Figure 3:*** *Reliability Threshold & Decision Tree*

Identity sources that produce a reliability estimation score below the threshold will be discarded. Those that produce a reliability estimation score above the threshold will move to Step 3 for data fusion and hashing as depicted in Figure 3.

## 4. DEVELOPMENT PROGRESS & RESULTS

This paper outlines the design and initial development stages of the aforementioned smart digital identification solution. Currently, an initial test environment has been created in the form of a relational database which includes both 'real' and 'fake' identities. This synthetic database is comprised of a core table that stores traditional personal identifiers including full name, date of birth, gender and home address. Each record's primary key links the records to a secondary table that documents relevant online accounts that have been nominated as potential identity sources. This table subsequently links to several other tables that store the attributes associated with each online source.

Preliminary execution of identity resolution techniques that incorporate the additional social contextual data have proven successful in discerning between identities that are likely 'fake' and likely 'real'. These tests were conducted using standard prepared statements.

The next phase of the development will require researchers to apply machine learning techniques and Bayesian probabilistic modelling to data sources and attributes in order to intelligently determine reliability scores for identity sources, weight attributes and determine appropriate thresholds.

## 5. CONCLUSION

This paper has identified the importance of effective identity verification and authentication in preventing criminal and terrorist activity facilitated by fraudulent identities. It highlights methods by which fraudulent identities are gained and purported for nefarious purposes by those who intend to commit further illegal acts. It also highlights the necessity for law abiding citizens to protect and prove their own real identities.

Identity resolution methods that seek to authenticate or refute similar or duplicate identities have been reviewed. This emphasizes the substantial evidence that integration of an individual's social contextual data can greatly improve the success of identity resolution when paired with machine learning and Bayesian probabilistic modelling techniques. The inferred conclusion is that the same social contextual data could be used to authenticate a 'real' identity.

This paper outlines the design of an algorithm that aims to authenticate or refute identities based on information gained from digital footprints. It aims to demonstrate the feasibility of the concept, documenting the flow of each integrated step. Initial tests that have been executed have been successful in contributing to this end. The future work documented will continue to test the efficacy of the Bayesi-Chain Smart Digital Identification Solution.

**REFERENCES**

Bernado, J.M., Smith, A.F.M., “Bayesian Theory”, John Wiley & Sons, Chichester, (2000).

Fish, T., “My digital footprint”. AMF Ventures Limited, Futuretext, London (2009).

Fox, D., Hightower, J., Kauz, H., Liao, L. & Patterson, D.J., "Bayesian techniques for location estimation", Proceedings of the 2003 Workshop on Location-Aware Computing, pp. 16. Seattle, Washington, USA, (2003).

Haimson, O.L., Brubaker, J.R., Dombrowski, L., Hayes, G. “Digital footprints and changing networks during online identity transitions ”, Dept of Infomatics, University of California, Irvine, CA, Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, Pages 2895-2907, San Jose, CA (2016).

Kean, T.H., Kojm, C.A. Zelikow, P., Thompson, J.R., Gorton, S., Roemer, T.J., Gorelick, J.S. , Lehman, J.F., F.F. Fielding, F.F., Kerrey, B.,“The 9/11 Commission Report” (2004). URL: http://govinfo.library.unt.edu/911/report /index.htm

Köpcke, H., and Rahm, E., “Frameworks for entity matching: a comparison”. Data Knowledge Eng. 69, page 197–210 (2010).

Li, J., Wang, G.A., Chen, H., "Identity matching using personal and social identity features", Information Systems Frontier 13, page 101-113 (2010).

Park, M., “AR is on the verge of transforming the human-computer relationship”. VB, October (2017). URL: https://venturebeat.com/2017/10/30/ar-is-on-the-verge-of-transforming-the-human-computer-relationship/

Spalevic, Z. and Ilic, M., “The use of the dark web for the purpose of illegal activity spreading”, ЕКОНОМИКА, Vol. 63, January-March 2017

Vignoles, V.L., “Identity: Personal AND Social”. University of Sussex, Oxford Handbook of Personality and Social Psychology, Second edition, Oxford University Press, London (2017).

Wang, G.A., Chen, H.C., Xu, J.J., Atabakhsh, H., “Automatically detecting criminal identity deception: an adaptive detection algorithm”. IEEE Transport Systems Management, Part A-Systems Humans 36, page 988–999 (2006).

Weisstein, E.W., “Probability”. Mathworld, Wolfram Alpha (2017). URL: http://mathworld.wolfram.com/ Probability .html

Xiang, R., Neville, J., Rogati, M.,“Model relationship strength in social networks”, Purdue University, West Lafayette, IN, Proceedings of the 19th international conference on World wide web, Pages 981-990, Raleigh, NC (2010).