Journal of Ambient Intelligence and Humanized Computing (2019) 10:3573–3602
https://doi.org/10.1007/s12652-018-1084-9

ORIGINAL RESEARCH

DeepSeq: learning browsing log data based personalized security vulnerabilities and counter intelligent measures

Chiranjib Sur[1]

Received: 4 July 2018 / Accepted: 28 September 2018 / Published online: 11 October 2018

Abstract
Personalization security is a concern with the rising ability to monitor and access public and personal data by organizations, mainly with gradual integration of human life with their devices. In this paper we have shown how simple browsing log data can jeopardize the identity and the personal integrity of a person along with analysis of preventive measures to protect them. As people get digitally enslaved, unknowingly browsing logs inherited certain unique behaviors of the people. It can be characterized and used for monitoring them and their aligned social, professional and organizational counterparts. It is quite a challenge for modern systems to keep attackers at bay and prevent them from gathering and analyzing activity data which can be used to identify specific, easy and valuable targets. Our analysis is based on modeling efficient systems for justification of the possible vulnerabilities and counter-measures through data driven approaches to learn and analyze such data and derive the extent these data can be exploited. Overall, we achieved an accuracy of 85% for identification of targeted characteristics using log data features using deep learning models, which achieved better than other learning models, thus effectively pointing out to the fact that there is severe non-linearity and combination possibilities in the data.

Keywords Browsing · Logs · Behavioural biometrics · Deep learning · Profiling · noise

1 Introduction

"We become what we repeatedly do."—Sean Covey. Apart from inheriting the fundamental identity in genetic structure, people reflect their identity in many other factors they do to sustain their existence. They are so unique that in modern days they create threat to personal integrity and researchers succeeded in developing procedures and intelligent systems to analyze, learn and recognize through them. Even if it is not so unique apparently, it is believed to have potential with better feature engineering. While permanent markers like finger, retina prints are invariant, variation is inevitable for face, gait etc. But still researchers try to identify people with definition and extraction of rich features which can uniquely identify them. These characteristics are revealed components of a person and many times, irreversible. While the permanent markers never change unless there occurs some physical transplant, variation markers do not change but varies. As researchers try to identify people through these markers, it becomes very important that they define and extract very rich set of features which can uniquely identify them. These characteristics are revealed components of a person and many times, irreversible. There are other behavioral instances for identifying a person and browsing history is one of them.

Browsing history is of utmost importance as it not only jeopardized the personal integrity of the person but also compromises information regarding habits, mood, requirement, personal strategies, ideas and several other kinds of personal credentials of the users. In this work we, for the first time, have justified the potential of browsing history as behavioral instance facilitating breach of identification of a person. In fact, it is not only human but even entities possess definite identities and behavioral instances, like signature of mobile devices or may be a fighter jet or an aircraft carrier or even submarine. Apart from the communication trends like mac numbers, bandwidth of operation, even their vibration, heat map is very unique. This document focuses on the analysis of the browsing history of several users belonging to different groups with respect to gender and age. The selection of the different age and gender group also focus the

✉ Chiranjib Sur
chiranjib@ufl.edu

[1] Computer and Information Science and Engineering Department, University of Florida, Gainesville, USA

study of the relative differences in quantity and quality in form and what can be inferred from those relations.

This study will also establish many unknown facts regarding the characteristics of the browsing history and has emerged as very special properties which can even help in tracking the usage of infrastructures by person without the need to make them provide their authentication credentials. This will also help in tracking people who are trying not to reveal their identity. In future, the analysis of these kinds of data can help in generating forensic evidence. Organization can bookmark such signatures and use it for different purposes. It can gather digital behavior, one kind of signature of people or it can be used for dishonest business opportunity through pushing products and services without properly and fairly competing in the market. The advantage of analysis of personalized crowd-sourcing behavior can eliminate other competitors from the market and create monopoly, which is not at all good for any economy. Other applications include website content presentation with prioritize display of services based on the trend of the user profiling data, search engine ranking learned from user profile preferences etc. However, the dark side jeopardized exploration of the internet space and fairness of the internet resources reaching to you. Modern day organizations push government to scrap net neutrality and impose totalitarian rules. However, policy making, fairness in algorithms and inclusion of noise on the user-side are some of the possible overcomes for the risk of vulnerability of a specific targeted group of users.

The rest of the document is arranged with literature review in Sect. 2, problem specification in Sect. 3, the details of data collection and preliminary characteristics of the distribution of the data along time line are provided in Sect. 4 and the learning capability of these data is provided in Sect. 5. The details of noise analysis is provided in Sect. 6. We have concluded in Sect. 7 with future prospects.

Contribution of this paper:

1. Browsing data as profiling data, first to do this analysis
2. Profiling of individuals
3. Profiling of characteristics (gender and age)
4. Security issue of these profiling through models
5. Concept of digital DNA and significance observation of characteristics through browsing data
6. DeepSeq or counter measures for identity and effectiveness evaluation

2 Related work

Research in security analysis has also changed its approach for looking into the problem. With more and more devices and people connected and innovation in characteristics of attacks, instead of investing tremendous effort on development of protection, researchers try to keep things open and inter-tangled or noised so that the profit made out of those can be minimized. A preventive measure is like adding noise or create capability of detecting the breach instead of trying stop the breach. Like any person can forge a credit card, but banks want to make the consumer aware and report any breach and thus involve a processing time for each transaction. Traditional model based analysis of security data is a common practice.

Personalization has been a common practice for recommendation and understanding the users through profiling of system and network activities. Like, Petrosyan (2018) sorted and categorized individuals and groups by their inclination on risky behavior or level of dangerousness, an essential security function, Atote et al. (2018) proposed an algorithm for privacy and security purposes for different profiles, with the integration of Information Dispersal Algorithm, through the use of vast data on profiles at any location at any time and would be achieved by the use of private cloud. These user profiling played an important role in recommending the best result to the user as per his requirement. Nowak et al. (2018) presented a system to detect abnormal network users behavior based on web pages features which were requested by a user like URL address, URL category, the day of week or time when the web page was visited. This can be an important security mechanism and can also be used to make personal user profiles. Koh et al. (2015) compared voluntary profiling to no profiling and showed that voluntary profiling can lead to counter-intuitive results. Also, it showed that consumers that do not participate in profiling and some that participate are worse off under voluntary profiling. Meng et al. (2018) developed a trust-based mechanism to detect insider nodes via behavioral profiling through four mobile and networking features to establish behavioral profiles. Euclidean distance was used for determination trust between two profiles. Nicol et al. (2018) discussed CPU and memory profiling of services and are commonly-used methods to identify potential performance and cost optimization. A scalable, language- and platform-independent framework was designed to enable on-demand CPU and memory profiling of microservices, and centralized storage, sharing, and analysis of the resulting data. Flesca et al. (2018) discussed relevant concepts and approaches for Big Data security and privacy, and identified research challenges to achieve comprehensive solutions to data security and privacy in the Big Data scenario. Lebiednik et al. (2018) introduced RF Power Profiling and studied the potential threat of spoofing attacks in Wireless Network-on-Chips due to malicious hardware trojans, and introduce Veritas, a drop-in solution that detects and corrects such spoofing attacks. Andersen and Karlsen (2018) discussed the process of creating and maintaining user profiles with a privacy preserving focus

as personalization can be used to improve the quality of a service for a user and can be used to better target its users by understanding her interest and her current context.

Big data requires new tools and techniques to capture, store and analyse it and is used to improve decision making for enhancing customer management. Anshari et al. (2018) introduced customer relationship management (CRM)'s strategies in supporting personalization and customization of sales, services and customer services. This had enabled business to become more aggressive in term of marketing strategy like push notification through smartphone to their potential target audiences. Park et al. (2018) attempted to model heterogeneous user traits and interests, including personality, boredom proneness, demographics, and shopping interests. Based on modeling results, it discussed various implications to personalization, privacy, and personal data rights. Logesh et al. (2018) developed a novel hybridization approach for aggregating recommendations from multiple recommendation systems to improve the effectiveness of recommendations and evaluated on the real-time large-scale datasets of Yelp and TripAdvisor. The outcome had improved performance over standalone and baseline hybrid approaches. Yang et al. (2018) aimed to predict passenger satisfaction over their rides and understand the key factors that lead to good/bad experiences and is based on in-depth analysis of large-scale travel data through profiling. Knowledge of websites of specific users or aggregates of users visit creates new opportunities of business. García-Dorado et al. (2018) proposed a counting method based DNSprints and can identify visits and their durations with false and true positives rates between 2 and 9% and over 90%, respectively, at throughputs between 800,000 and 1.4 million DNS packets per second in diverse scenarios. Ren et al. (2018) found strong correlations between users' demographics and their CPS behaviors, log-recorded cyber-physical behavior reflected well data captured in the corresponding questionnaire, different CPS behaviors contribute differently to the predictability of demographic attributes. Chen (2018) proposed a distributed representations of users' viewing and purchasing behaviors on an e-commerce website for recommender systems by leveraging on the cosine distance between the distributed representations of the behaviors on items under different contexts.

Karataş and Korkmaz (2018) used *k*-means clustering on Spark to determine whether the incoming network values are normal behavior using 400 thousand network data from KDD Cup 1999. Al-Gburi et al. (2018) proposed simpler definitions of security and privacy, boiling down to their most essential characteristics for personal immunity, more as security characteristics than privacy characteristics and as a right of complete immunity to be let alone. Lin et al. (2018) introduce network security-related data, including its definition and characteristics, and the applications of network data collection for network security in the context of big data and 5G and also presented a taxonomy of data collection technologies. Atli et al. (2018) proposed an intrusion detection system based on modeling distributions of network statistics and extreme learning machine to achieve high detection rates of intrusions. It aggregates the network traffic at the IP subnetwork level and the distribution of statistics are collected for the most frequent IPv4 addresses encountered as destination.

A user profiling and adaptation case with exhibition booth was proposed in Salem et al. (2010). Otebolaku and Andrade (2015) investigated context-aware recommendation techniques for implicit delivery of contextually relevant online media items through contextual user profile and a context recognition framework. Maleki-Dizaji et al. (2014) proposed adaptive agent-based modeling to address the incompatible effectiveness of retrieving the exact information, the users require. Liu et al. (2018) introduced a context-awareness-based intelligent method to predict users' intention to use a customized service and keep clients satisfied, depending on the contexts they provided. Park (2017) analyzed personal user preferences from resource usage history based on the Myers-Briggs type indicator, to recommend customized resources for classified user types. Azimi et al. (2017) studied IoT-enabled systems tackling elderly monitoring to categorize the existing approaches for elderly-centered monitoring. Kosmides et al. (2016) proposed location recommendations based on users' needs for user's possible future locations. Rafferty et al. (2016) presented a novel approach for video based content analysis, a mechanism to facilitate matching analyzed videos to dynamic activities/goals. Su et al. (2016) explored advertising formats and interactive modes related to the effectiveness of advertising by adopting the attention, interest, desire and action model.

Personalized security using internet and browsing logs Baglioni et al. (2003) identified navigational styles of web users by extracting information from the logs of a particular web server. However integration of such data and analyze them on a common platform will be challenging, however they did an interesting pre-processing by extracting semantic information like category of the site from the URL name itself. Song et al. (2013) used Gaussian Mixture Model (GMM) to fit biometric identification data of several users which can reflect behavior for authentication. However these kinds of probability based models are approximate generators and do not keep provision for variations in the data. The authentication was trained on the estimated distribution of the expected number of unique system level events like process creation registry key changes, file creation etc. Yang et al. (2015) leveraged on HTTP traffic data to cluster user behaviors into different groups based on application categories they use. These kinds of mobile internet user behavior characteristics can help in provisioning

information for resources of network operators. Nogueira et al. (2005) used hourly internet traffic of the users to classify individuals using techniques like discriminant analysis and neural networks. This solution will not be scalable as discriminant analysis performs unnecessary compression of data for the classifier. Freeman et al. (2016) created a probabilistic statistical model with the source IP, Geo-location, browser configuration, and time of day of the users to identify user behavior apart from password authentication to minimize least amount of damage in case of breach and before providing full privilege of the system. Duarte Torres et al. (2014) aimed at understanding search and browsing behavior of young people through topic progression modeling and clearly categorized applications and websites and used these to classify people based on their of age by statistically correlating the features that were generated with browsing and search behaviors along with ads clicking, click duration and query length etc. Mobasher (2007) used clustering, association rule discovery, sequential pattern mining, Markov models, and probabilistic mixture and hidden (latent) variable models to model web based personalization and also showed that a hybrid data mining framework version consisting of the best of each can perform much better leveraging the data from a variety of channels and provided much better solutions. Stolfo et al. (2000) used raw binary TCP dump packets data consisting of 41 high level features under the categories of basic features, content features, time based traffic features, and host based traffic features. It is a network intrusion detection problem where each connection is labeled either normal or malicious. Mcdaniel et al. (2006) used unsupervised clustering methods to analyze historic network communication for four different individual host profiles within enterprise network to capture historical communication patterns and generated malware profiles under different granularities to detect and block worms. Marella et al. (2014) discussed privacy and risk awareness from different browser plugins as people disclosed tons of personal data while browsing without getting aware of it and thus created their own potential threat. Gulyás et al. (2016) involved problem of identifying the attributes that can serve as a fingerprint of users given the size of the fingerprint, where fingerprint is typically a couple of their attributes people tend to use to browse the web.

Recommendation analysis using internet logs Castellano et al. (2009) consisted of several web personalization techniques starting from semantic content-based recommender system, collaborative filtering techniques and use of fuzzy schemes to involve overlapping criteria for decision makings. Artificial neural network Chang et al. (2009) had been used to group users into clusters and Kano's method to extract the implicit needs of those users in different clusters for website content recommendation based on the training data collected as part of the survey of expert users to avoid display of unnecessary things. Davidson et al. (2014) advocated client side personalization at the OS level through learning profile characteristics for each person using the Windows phone operating system considering the possible analogues for other mobile operating system. Leon et al. (2012) recorded behavior and attitudes of users while dealing with tools to limit online behavioral advertising like plugin that blocked requests to specific URLs. Malandrino et al. (2013) described the impact of increased awareness through a customized tool for users to understand privacy risks related to web browsing. Mobasher et al. (2002) discussed application of aggregate usage profiles in techniques like collaborative filtering for personalization as it emphasized on not only better pattern discovery but also derivation of better quality recommendations and helped in overcoming total reliance on user ratings, lack of scalability, with poor performance due to sparsity and high-dimensionality. Komiak and Benbasat (2006) investigated the cognitive and emotional trust balance perspective for analysis of personalized recommendation agents for determining the intention of the customers to adopt the recommended product and found that emotional trust played the important role and fully mediated the cognitive trust.

Phishing, malware and network intrusion Marforio et al. (2015) suggested personalized security indicators as a phishing detection solution for applications in mobile devices by experimenting with people on a real banking application environment usage. Jiang et al. (2014) proposed a scanning-free personalized malware warning system by learning from feedback generated from detection logs of usage. It created one "malware recommendation engine" which can rank the risk of each malware sample for each user to determine the potential of the different types of malwares with respect to a pool of susceptible individuals. Egelman and Peer (2015) discussed the need to study individual differences in decisions to help determine and understand privacy and security attributes. Individual differences that are described in the literature of psychology are much stronger predictors of personality traits. Wang and Stolfo (2004) used a fully automatic, unsupervised learning fashion based payload-based anomaly detector using the payload of network traffic. Automatic analysis of malware behavior using machine learning was described in Rieck et al. (2011) where both clustering and classification had been used as an incremental approach for behavior-based analysis with identification of new classes through clustering and allocation of new malwares to these classes. Wang and Goldberg (2016) worked on attacks that website fingerprinting possesses to Tor network.

Personalized security and privacy Kasanoff (2002) described the demarcation of the line of personalization from privacy invasion. Taylor et al. (2009) discussed the effects of personalization and privacy concerns associated with information control and the associated compensation,

as the relation was very important in understanding the potential moderation that can persist between the privacy and behavioral intentions in online interactions. They showed that the offer of compensation had no effect on such relationship. Riecken (2000) described several aspects of personalization and tried to answer question with respect to personal view and its relationship with business. Sackmann et al. (2006) discussed on personalization in privacy-aware highly dynamic systems like ubiquitous computing with increasingly interconnected mobile communication where the challenge is to keep up with the constant growth of communicated data in different forms. Kobsa (2007) discussed privacy-enhanced personalization and how to tackle the tension between personalization and privacy and reconcile through coexistence as that will provide better consumer experience and will also provide a win–win situation for both the web vendors and the consumers. Brar and Kay (2004) described privacy-aware personalization in ubiquitous computing environments as previously the prospect of privacy in ubiquitous computing had been confined to ad-hoc, application specific or partially acknowledged. Mulvenna et al. (2000) had taken up the science behind personalization through personalize content mining analysis and how they were delivered to individual users and described how they will be ubiquitous in the future to come. Mcdonald et al. (2009) described different online privacy policies and their formats.

Security news and surveys Survey of supervised, unsupervised and hybrid learning techniques for internet traffic classification had been discussed in Nguyen and Armitage (2008). Eirinaki and Vazirgiannis (2003) covered survey of the different web mining techniques on the various kinds of attributes of the data for web personalization and its prospects. McAteer (2016) described how Google can stalk the entire life's history of a person. "Google's My Activity Page Is a Scary Reminder That Google Knows Everything About You" (Sathe 2016) and "This Creepy New Google Feature Lets You Stalk Your Entire Life's History" Purewal 2016 became very popular tweets of people, showing the concern of general population and how techie journalists aware people before it gets too late. There is always mixed reaction from the population, some are amused while others are concerned about personalization and security. Olivarez-Giles (2016) described how to use Google's my activity privacy tool which is important because Google collects data as people surf the web and use its apps and services and there are digital tracks, Google do not tell their users. Schaub et al. (2016) provided important information about the third party companies tracking users' web browsing behavior and this has raised privacy concerns. These tracking extensions do not have registered companies and there is less visibility of what they are collecting and what are these used for. This increased awareness of unauthorized tracking while the sense was that browser extension was claimed to protect the user. Our work is mostly based on browsing data and this approach has not been tried before. We establish no matter how the browsing history looks like, we can impose a complex transformation through representation learning of deep architectures and help categorize individuals. The browsing categories contain important information for these kinds of analysis.

3 Problem scope and future prospect description

What we are trying to do is to articulate diverse statistical and data driven models that can characterize certain groups of people with common interest, habituated through their log history or even real time browsing patterns and try to recognize them through the models. This kind of recognition, in the modern day, possesses capability to breach personal integrity and security measures are required to curb them. First part of our analysis has shown how it can used to characterize different users and there are certain patterns of their browsing that can easily be read. This will also reveal certain characteristics of the users and their requirements. These requirements make people enslaved and vulnerable because organizations exploit such opportunities. Is this bad ? One is to recommend and the other is to facilitate. The moment people gets facilitated, their capacity and capability shrink and there are chances that people get tamed to the control of hidden powers. Till now, researchers have investigated several kinds of personality measures and this is the first time, we have tried to do the same for log data. Not only that, we have also generated the remedy for this kind of situation and have elaborated in Sect. 6.

We mainly focused on analysis of browsing logs of users, diversified with respect to gender and/or age. The selection of different age/gender quantifies on generative and discriminating aspects on relative differences in qualitative and quantitative measures of data and what can be inferred from those relations. This will, by far, extent the study of vulnerability with psychological and behavioral facets of certain class of people who are less aware of such consequences and at the same time held immense wealth and responsible positions. The main motivation of this work is identification of certain people of certain classes, organizations, principles, age, gender, locations, profession, social, economic and cultural habit. This helps many organizations to hit their targets and learn patterns and dependencies under certain circumstances to create bottlenecks out of it. Apart from identification, there are other motivations like creation of digital database of individual/collective behavior and integrated with social media to detect any unforeseen issues or opportunities.

With "Internet of Things" coordinating integration and increased interdependence of services, till now there is only 1% things connected, but the projection is expected to reach 50 billion in 2020, involving and bringing together billions of people. With so many people and most importantly devices connected, it becomes necessity not only to identify them but also identify who are using it. This accompanies lots of business opportunities and there will be integration of various sectors like if you are browsing a certain food video, then it is more likely that it will provide you all the places, with reviews, prices and offers, where you can go and eat and also the sites where you can order to get it delivered with drones. It can even offer a holiday package (flights and hotels descriptions) to that place where the video was being made. With this kind of integration, it is very important to understand the sentiments (may also be psychology) of the people without speaking a word and there can be situations where the device or the person can be devoid of any social media interactions. There will be time when the device will itself be smart enough to browse the internet and inform the user what need to be done and from where. Likewise, it is also inevitable that the machine on the other site understands the device, its usage, its necessity and also the end user. Our study is very specific to some strata of people and can actually be extended to many more.

The main motivation of this work is identification of certain peoples, classes of peoples, people of certain organizations, people with certain principles, people of certain age, gender, locations, profession, social, economic and cultural habit. This helps many organizations to get to their targets, helps in taking surveys, behavior under certain circumstances and creates opportunities out of it. However, it can also be misused or misinterpreted. Engineering is all about defining the problem and the goal, and this era will also witness modeling of thoughts, psychology and mentality of human beings and thus create greater challenge for the security researchers as the analysis must accompany understanding both systems and human minds. Thus the data collection, to support the understanding, must be carried with utmost care and precaution. Most of our analysis is biased on the old age people who are more prone to attacks and while feature extraction we came across various new concepts related to browsing history and the habit of people. While the website type distribution is very helpful and generalized, other features related to browsing patterns and time spent can be also helpful. For example, a person spending more time on Britain news as well as share market news channel is more likely involved in trading and investments and if those people are monitored or at least predicted, then their collective trends can be traced much earlier based on other market parameters. While consequences of behavior are traceable, many times the real reason cannot be traced out and this kind of monitoring can be inductive.

Apart from identification, creation of digital database of people or individual behavior, integration of social media to detect any unforeseen issues or opportunities can be very specifically targeted as it resolves the ambiguity of person on different personalized security context. Like, if there are too many people making certain search query or visiting certain kinds of websites, then something actually has happened and necessary actions can be taken. At least, the feed can help in tracking the event in social media. Research in system and browser log history and the corresponding opportunities gained momentum with the increasing strength, capacity and capability of the browsers. People started contributing new applications integrated with the browser while the web-based applications replaced many traditional standalone software based architectures. In this open source era, with highly enriched APIs for the developers, it is easy to lay down innovations and requirements, whereas some tries to exploit the system with the same to revolve the wheel in their favor. The research in this narrow section exploded in the form of personalized security and personalized privacy agreement and was inevitable and large part of the community is getting involved in it. With too much of data collection and analysis, people are obsessed with their privacy while the researchers and organizations with privacy policies. There are efforts being made to redefine privacy and also keep provision for customization the levels. Like for many people, detection of what kind of news they are visiting can be their issue as it compromises their geographical integrity while for another person, his search keys can be a private issue. Research in this sector comprises in prevention, protection, designing, awareness, freedom to select or customizable, and more importantly to stop money launders from getting away from law. For the purpose of study, we have collected several daily log files of different users and the data description section will provide some procedural details and will also provide some analysis and comparison overview among them.

The participants used their same device for our data collection and happen to be from a certain geographical region. Thus device signature and information related to geographical status and related categorical information are redundant and avoided, but could have formed a very significant part of the analysis if data is collected without bias like without considering age group, gender and location. The data, being collected, is geographically motivated to analyze people of certain age and distinguish them from other disjoint age groups. The collected data is in form of log files and are sequences of time based distribution which can compress or expand and the distribution curve changes accordingly. The choice for the best time interval is a NP hard problem and need to be addressed as a separate problem and we find the time interval of quarter or half an hour is providing descent feature vector and the dimension is quite under control.

Quarter or half an hour of browsing is a reasonable browsing segment to characterize a person. No doubt that over expansion can sparse the vector which will make the learning difficult for many machine learning algorithms while over compression will suppress many information and significance. The best time interval will depend on the criteria on which it is being estimated. However, it suffers the same kind of fate as determining the number of bins for histogram generation. The purpose of this research is analysis of these features based on characterization and learning-prediction methodologies, while evaluation of the accuracy is the best judge of what kind of information these data can reveal. However, the main motivation of this research is not to design a classifier or win a prediction competition, but go through a thorough analysis of the data, its trends and what can be done with it. The quantitative analysis in the result section provides some interesting trends and must not be taken granted just on classification accuracies. This is the first time, we have tried to answer questions related to personal security and that even using the impact of browsing data on breaching personal identification. However, this is just the starting and the analysis provides enough evidence of what can be done with it. This analysis can also be extended to data related to application usage and other digital behavioral data from both personal systems and hand-held devices.

4 Data collection and analysis

The data collection was done with the approve of IRB, from the browsers of paid participants of different gender and age group. Participants were not informed about the research goals and was put up as an internet study limited to answering survey questions and installation of a browser plugin for understanding internet browsing. There was an effort to balance the age group and also the gender. But getting older adults, who are willing to go through the series of survey and are also a proficient computer user, was difficult. So does getting female older adults. They were made to installed a special plugin in their browsers. None of the users were informed what exactly we were doing to suppress the influence of the study and keep unbiased psychological behavior. Also the study dealt with the analysis of the natural vulnerability of the users and informing them will ruin the purpose of the study. The study for each participant consisted of 21 days, so the log files mostly have browsing data of 21 days. However, there were users who didn't remove the plugin even after the participation days were over and had more log data than usual or even less than 21 days when the participant agrees to discontinue with the study due to various reasons and circumstances. Till date, the distribution of data consists of 162 participants and out of that there are 102 younger and 60 older adult individuals, out of which 49% male, 51% female. We discarded those participants, whose data were too low, to be considered for learning. However for analysis, we have seen that at least 4 days of data were enough to identify a person with the next remaining days of data or to characterize a majority group of people based on GroupWtc feature. Though the analysis is framed on limited resources and collection, the same conclusion from large part of the samples provided us enough evidence that such possibilities cannot be overlooked.

When dealing with the log files, we have scraped all the data of the individuals into a matrix of time stamp and visited sites into feature vectors where the time has been decomposed in intervals (like in minutes, it makes the scale to 1440 features for a day) and then we have windowed them according to our size requirements. These are quite a large amount of features, but our modern personal systems and libraries of algorithms are scalable enough to handling this, if not more. This amount is just more than enough to represent the whole span of a day, but there is a limitation when it comes to the number of websites accessed at any time instance. It will be 0 for none, and 1 for both one or more, losing part of information that could have made some or a lot of differences. Hence we introduced a window distribution for each time interval. We tried to analyze with this binary representation of the time stamp as 0 representing no browsing and 1 means that any web request being made. There are a number of other features being defined and we will elaborate on each of them subsequently and also see how the real log file plots differ for each of the possible categories. The different profile data were expected to be balanced but in reality it is not, both with respect to gender and age, and hence we investigated on random sample sets from the represented. The data is shuffled random before dividing them for n-fold cross validation so that the influence and the dominance of both the best or worst is diminished as far as possible. This is the reason, regular shuffling before each testing and training division, is suitable to get better accuracy of the data.

The hierarchical clustering analysis in Fig. 1 for the different users can also be seen as the normalized cumulative summation of the time series data. For features, where distribution window is not used like the time, webtype, speed etc, smoothing based phenomenon like Gaussian smoothing etc. of the normalized cumulative summation curve can sometimes be helpful in providing better continuity of feature sets and also helps in better classification accuracy, but cannot be used for features like GroupWTc and GroupWTs as it won't make sense. Experiments has revealed that smoothing is found to be very efficient not only in achieving better accuracy for machine learning algorithms, but also in creating the window of seamless browsing which gets interrupted because of reading pauses and other session based browsing.

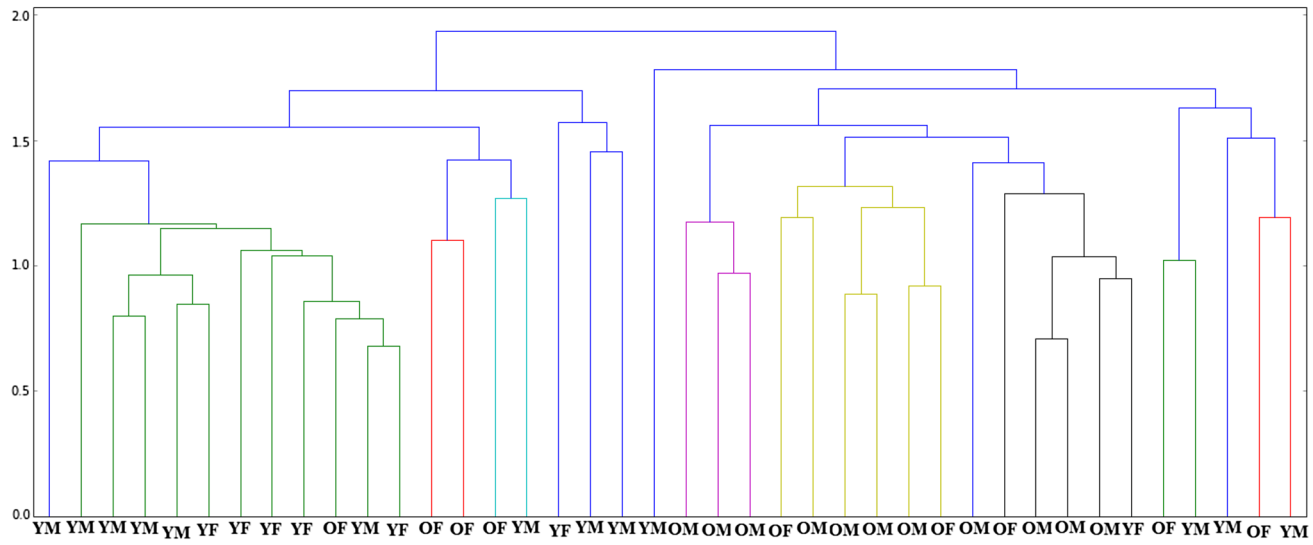

Fig. 1 Hierarchical clustering (Ward's method) of the different users based on distance metrics of the feature GroupWTc distribution matrix (*YM* young male, *YF* young female, *OF* older female, *OM* older male)

However, the hierarchical clustering shown in our analysis is devoid of such smoothing.

4.1 Description of different features spaces

The data log files, for each individual, were transformed to different conceptual feature space and this helped in defining browsing characteristics of the users. However, each of the defined features are completely disjoint and each of them is unique and has its significance. That means that span of one feature does not overlap with any other feature space. This is evident from the results of classification accuracy where we have analyzed quantitatively and also in many cases there were instances where combining the feature vectors actually enhanced the classification accuracy. This concept of dDNA is like DNA which can be defined as a sequence of ACGT, microarray, ChipSeq etc. All of them map to the same original feature subspace but yet they are different and have their own advantages, disadvantages, utility and analytic significance. While some of them are more useful in the diagnosis of diseases, others are just useful for mere comparison. In case of log files, the transformation to different subspace is also quite significant and the analysis will reflect how they, when combined, can actually help in performing better analysis and performance, mainly for the machine learning techniques. Before we justify why the log files are the digital DNA for a person, there is a requirement of a visual inspection of the plots of the different feature vectors.

Time stamp feature transforms the logs to sequence of activity throughout the time-line. With x time stamps and $\varphi()$ reflecting activity function, we define mapping function $f = \{T : (\varphi(TimeStamp) \rightarrow \mathbb{Z}_2 \rightarrow \mathbb{Z}_{\{0/1\}}) \in \mathbb{Z}^x \rightarrow \mathbb{Z}^t\}$.The preliminary and easiest way of analysis was using the time stamp throughout the day and such analysis provided very limited insight for most of the users as they tend to suppress many valuable data related to the users. However, it was, no doubt, a considerable feature vector for differentiation between many users. But time stamp cannot be regarded as a generalized feature and there is always chance that it can create confusion for the differentiator. This is because browsing history analysis must be time invariant and models must be designed accordingly. Time stamp features log the correlation of browsing pattern with activity of a time line and is not the time spend on a website.

This feature is perhaps the simplest possible transform for the log file and can be regarded as the presence stamp of the user throughout the day and boils down to a binary series of ones and zeros. Ones reflect that some or any website was being visited at that moment while zeros signify that it is not. The following plot in Fig. 2 will make it clear how this feature vector stretch for the different classes and users. Normally it is expected that the older ones will have different time stamp than the younger ones and the male time stamp will differ from the females. There can be exceptions and anomaly like another systems and must not be generalized. The main importance of this features is to see whether timing information can be used for prediction and whether the distribution can be generalized and made to work as a time invariant way.

Web type Websites can be clustered into many categories and subcategories like social-media, news, sports, educational, programming etc. and based on the categories, a time line can be populated with the mapped values which will reflect the web types visited. These features map from infinite

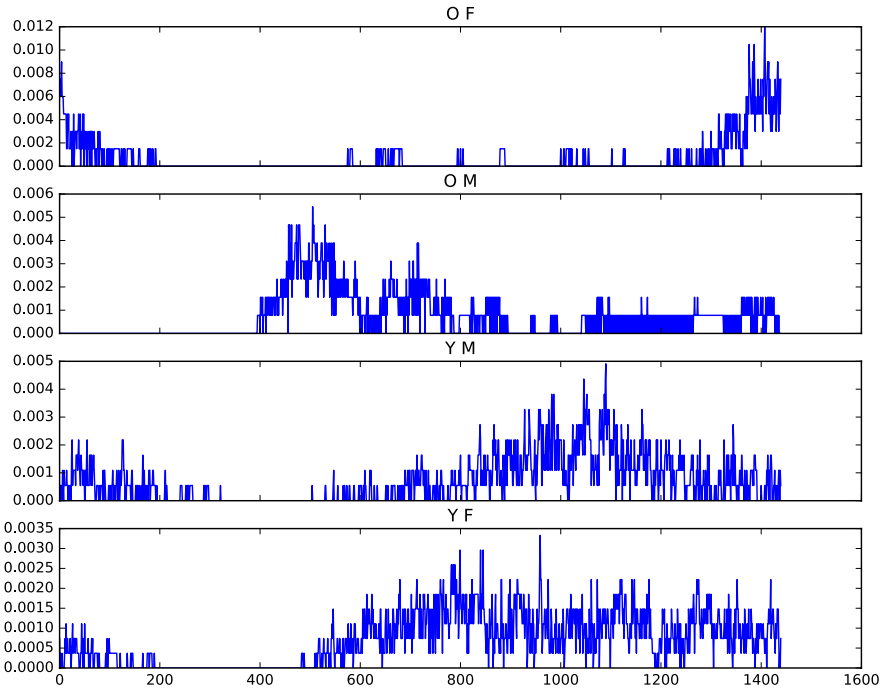

Fig. 2 Distribution of time feature, x axis (time), y axis (probability of time stamp at that time instant)

website space to a subspace which is bounded in some predetermined categories and change with region, country, ethnicity and other implicit factors. With m as $\{Website \rightarrow \mathbb{R}_m\}$, we define $f = \{W_T : (\varphi(Website) \rightarrow \mathbb{R}_m) \in \mathbb{R}^x \rightarrow \mathbb{R}^t\}$. More precisely, $f = \{W_T : ((\max(Website_j))_i \rightarrow \mathbb{R}_m) \in \mathbb{R}^t \rightarrow \mathbb{R}^t\}$ or even $f = \{W_T : ((Website_{last})_i \rightarrow \mathbb{R}_m) \in \mathbb{R}^t \rightarrow \mathbb{R}^t\}$, $\forall i \in [1, t]$.

Websites can be transformed into many categories and subcategories like that shown in Fig. 7 and based on these categories, a time line can be populated with the mapped values which will reflect the web types people visited and can actually characterize the behavior. Now a person can visit several websites at a certain instance of time unless it is very negligible, then which website must and should be considered. In fact, any of them can be considered but we considered the last one as many people tend to use search engines initially and perhaps the last one has the high probability of being the one being looked for. This series of features has been extended to better ones but they still suffer from the choice of the best time interval and dimensional explosion and is a NP hard problem. These set of features are a direct map from the infinite website space to a subspace which is bounded in a predefined space based on some predetermined expectation and can change with region, country, ethnicity and other implicit factors. The following figures will provide some greater overview of the features for different users. Later on in Sect. 4.4 we provided accuracy analysis for different machine learning techniques. While it comes to younger and the older ones, younger people will have disjoint domain of digital hanging out and will be different than the older ones. The same theory will dominate the males and the females.

It must be mentioned here that apparently, it will be difficult to define the best set of web types for any category of people and this can itself a problem to understand the psychology and necessity of people. However once such distribution window is sought out, differentiation and learning for machine learning algorithm can be much easy and smoother. We have done a distributional analysis on the percentage occurrence of different categories of people (Sect. 4.4) which clearly indicates how the posterior probability will be different for the different categories of people. However, Baÿesian analysis will not be fair judgment for determination of classification as we were able to analyze that there are more complex and non-linear features possible to feature out that can contain more information with respect to content and not mere probability of occurrence (Fig. 3).

Speed Time stamp feature may not infer the full information of the access details of the individual and a part of the information is lost. It conceals the aspect of visiting multiple sites in a time interval. In most intervals,

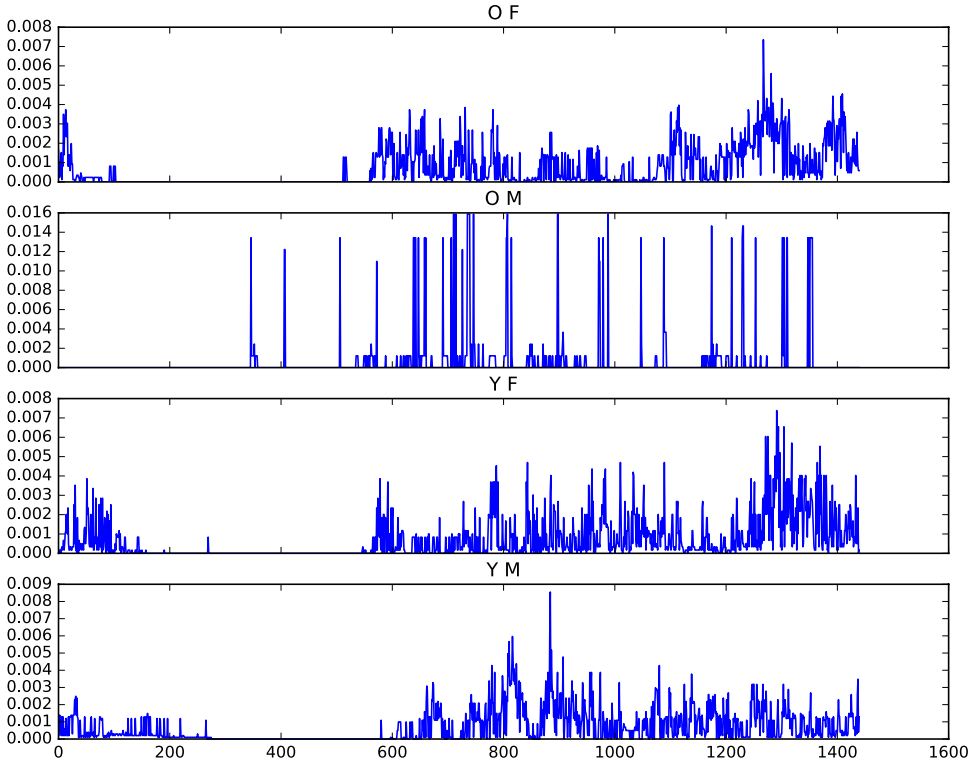

Fig. 3 Distribution of WebType feature, x axis (time), y axis (probability of webtype at that time instant)

considerable part of the distribution will definitely be different. $\forall i \in [1,t], f = \{S : (\sum_{j=1}^{j=x_i}(\varphi(TimeStamp) \rightarrow \mathbb{Z}_2)_j)_i \in \mathbb{Z}^t \rightarrow \mathbb{Z}^t\}$.

Time stamp feature may not infer the full information of the access details of the individual and a part of the information is lost. Even if the person visits one more than one site at a time instances, it is mapped same. However, a speed vector can actually have much significance when comparing the older and the younger and also the male and the female. At least if any time instance may not differ, but considerable part of the distribution will definitely be different. The next set of refined time stamp is the speed vector along the domain of time scale. The positional information signifies the timing factor in a day as the previous features. Figure 4 will reflect the visibility better across the different age and the gender groups.

Group web type count (GroupWTc) Previous features had one instance for each time interval and there was both qualitative and quantitative loss of information along that time scale. The group web type feature will make up on that and has all the categorical instances for a particular time interval. It will count the number of times all types of web site types requested in an interval. GroupWTc creates the perfect distribution scale for each time interval in the form of a two dimension convoluted in one. This is a true and complete feature description, however deciding efficient time interval can be challenging. So $\forall i \in [1,m]$, $\forall k \in [1,t], \quad f = \{G_C : (\sum_{j=1}^{j=x_i}((\varphi(WS) \rightarrow \mathbb{R}_m)_j)_i \in \mathbb{R}^m)_k \in \mathbb{R}^t \rightarrow \mathbb{R}^m\mathbb{R}^t\}$.

All the features, described previously, have one instance for each time interval and hence there was dearth of integration of different spread of information along the time scale. The group web type feature set will generate a full wave of both time and information scale and has all the categorical distribution instances for different time interval scale. This time interval is a follow up of time stamp which will contain a distribution of different categories instead of just one. Also it will count the number of times, that particular type of web site has been visited and will account for the variations in distribution in website types with the change in time. This can be called a complete feature description, however its utilization as features can only happen if the proper time interval has been considered. These set of features can be seen as a distribution of the web type for a segmented time fragments for a day and for any category, the distribution can change. It is being called as GroupWTc as the category representations are grouped for a time interval. How useful it is will be reflected by the graphical plots and also later by the accuracy it can provide through training. The formula that govern this set of features is provided above. Each of the plots in Fig. 5 is mainly marked for the timely distribution of the different categories of website visited in each window of time and it is the only feature which is getting the

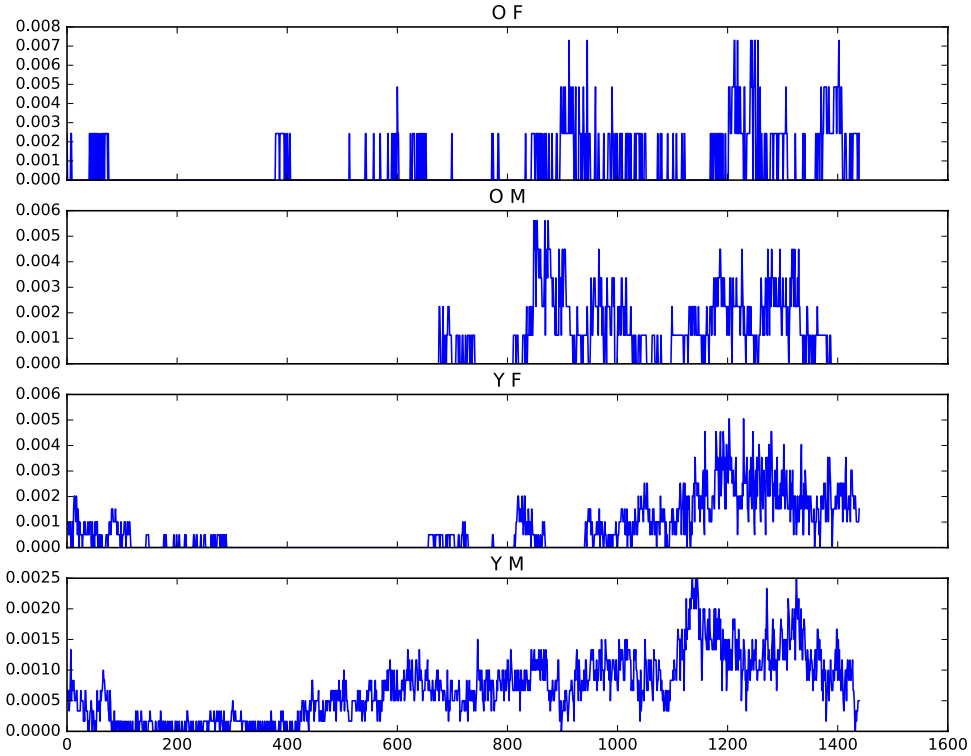

Fig. 4 Distribution of speed feature, *x* axis (time), *y* axis (probability of speed average at that time instant)

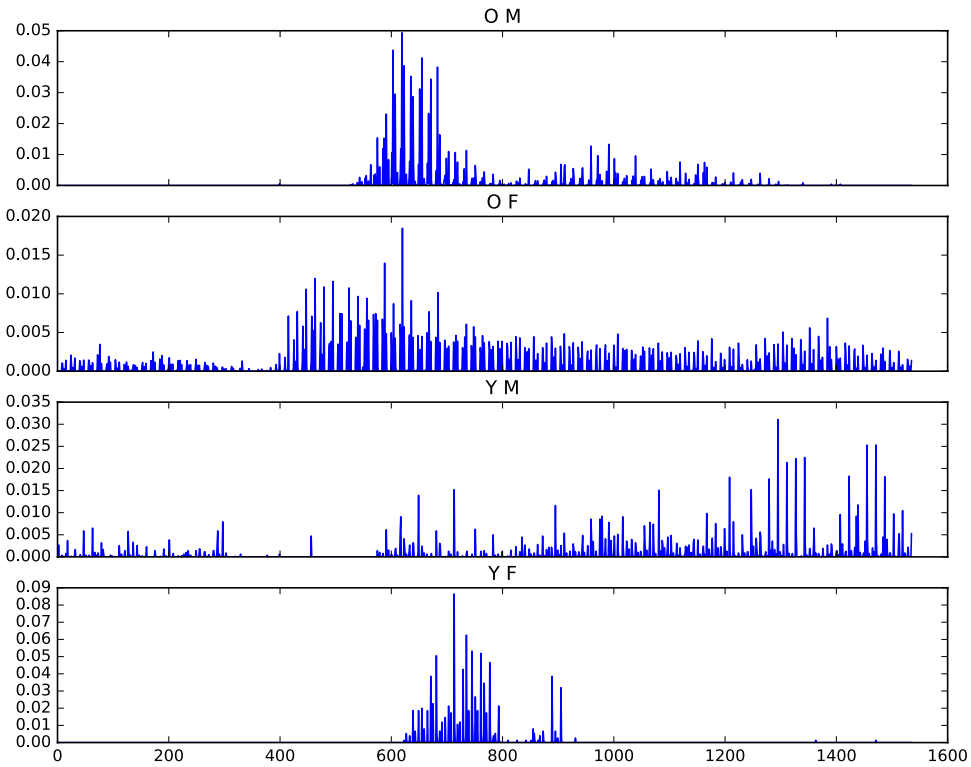

Fig. 5 Distribution of Group-WTc feature, *x* axis (time), *y* axis (probability of gwtc stamp at that time distribution window)

whole information structure intact without compromise but in a compressed time frame bound. However the perfect time interval is variable and the feature space is convoluted to a linear space with alternating spread of feature distribution in a windowed time scale. So feature space is defined as $\{timeWindow * distribution\}$ where "distribution" is defined as the number of requests being made for a certain web type category and "timeWindow" is defined

as the time interval and logically a 15 min time interval can be logical, people, at least, spend.

Group web type stamp (GroupWTs) This set of features is same as GroupWTc except that it will consider a stamp instead of count. The entry of this GroupWTs will be quite similar with GroupWTc, but the span will change and it can provide some new information and also suppress or normalize the over expressed feature stacks. The equation for GroupWTs goes like the following. So $\forall i \in [1, m], \forall k \in [1, t]$ $f = \{G_S : (\varphi(Website) \rightarrow \mathbb{Z}_2)_i \in \mathbb{R}^m)_k \in \mathbb{R}^t \rightarrow \mathbb{R}^m \mathbb{R}^t\}$.

When designing a classifier, the count may dominate the content and hence this GroupWTs feature is significant. The span and the entry of this GroupWTs will be quite overlapping with group web type count for particular cases, but the numerical value will change and it can provide some new information with respect to the distribution. The browsing pattern of people is more than just counts and content of the browsing is important than how many such content is requested. Like for a researcher, it is likely that a Journal site is visited and this marked his characteristics and no new information is created if several Journal sites are being visited. Figure 6 will definitely provide some insight of the difference it can provide for different gender and age groups and any defined categories.

Web site Another set of features was derived from the actual websites being listed. This is more generalized than the web type features as it reflects the original sites and will characterize more for the individuals. However, there are millions of sites and computing with million-dimension vector can be difficult. Dictionary d is $\{Website \rightarrow \mathbb{R}_d\}$, and $f = \{W_S : (\varphi(Website) \rightarrow \mathbb{R}_d) \in \mathbb{R}^x \rightarrow \mathbb{R}^t\}$. We have $f = \{W_S : ((\max(Website_j))_i \rightarrow \mathbb{R}_d) \in \mathbb{R}^t \rightarrow \mathbb{R}^t\}$ or even $f = \{W_S : ((Website_{last})_i \rightarrow \mathbb{R}_d) \in \mathbb{R}^t \rightarrow \mathbb{R}^t\}, \forall i \in [1, t]$.

This considers the same assumptions that are considered for the feature web type. However, this is unrealistic as there are millions of sites and computing with million-dimension vector space is unfeasible. For any two users, the feature vector will be restricted and feasible and some limited analysis has been done with it. The plots may not reflect what it should have been expected as the web sites are mapped to some numerical. This is the most diverse feature and the fact is that it will create the maximum difference possible as the physical web site will always differ for any two individuals and not only for younger and older, it will also be different for male and female of the same generation. While web sites provide a large scale of scalable feature spaces, it would be wise to understand the how the feature space need to be compressed or decompression which can help in defining the hyper-plane in the classifier and can help designing better model.

4.2 Digital DNA or dDNA

With the wide spread concern for personalized security, this is an effort to analyze quantitatively how the browsing log data can be utilized for categorization of the users

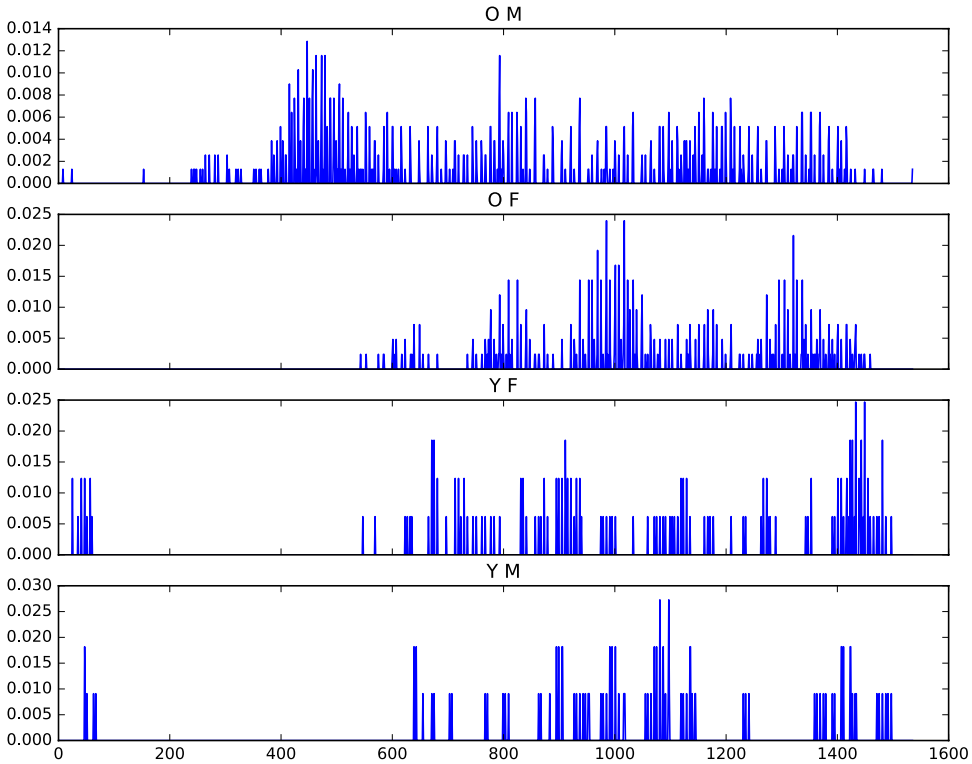

Fig. 6 Distribution instance of GroupWTs, x axis (time), y axis (probability of gwts stamp at that time distribution window)

and can detect their characteristics and even their identity. The analysis will not only help understand human behavior, but also help in designing preventive measures of further compromise of identity and credentials. Also it will help in understanding the requirements for the design of foolproof systems for the next generation utilities and applications. From the analysis in Sect. 5, we can claim that the log files and their corresponding extracted features can be regarded as the "digital DNA" or dDNA of the people. This is clear from the fact that dDNA is so unique that it is very difficult to combine different log file features from certain combined groups with same age or gender and then use for a classifier. To elaborate, say there are log files of 10 women and 10 men and we use them to train a classifier to detect gender, it will provide some accuracy of training, but if the number of males and females are increased (say 20 from each), the accuracy decreases. This clearly indicates that there is absolutely no trend in the features and adding more unique participant data is complicating the learning. Another reason is that the linear separatibity of classes gradually diminishes and only complicated and stochastic non-linearity can help in classification and has been demonstrated by deep learning architectures which can transform the features non-linearly in a complicated way and perform decent prediction. There are fundamental differences between biometrics and the log files like the dimension, span of the features and the inconsistency and the incongruity of the log files is quite different from the stable and invariable thumb and retina scans. Now if this is the case, the question is whether a multiclass classifier can be trained and predict to detect personal identification like any biometric authentication machine. The reality is that it depends. If the training data is rich and highly scalable algorithms like deep learning based neural network is used, it can escalate up to a certain extent. But there are some fundamental differences between biometric authentication instances like thumb prints, retina scans etc. and the log files. First is the dimension of the log files and the span of the features are much larger than biometric images which opens up the scope of exploration. While the second is that at the same time, there is inconsistency and provides the elements of variation into the system which apparently creates ambiguity on the veracity of the log files and is quite different from the stable and invariable thumb and retina scans. The term of "digital DNA" can have several meanings as they are associated with different parameters, but here the term is very confined to the characteristic usage of internet resources along time in a day and had been noticed that the distribution may be shifted but have varied very less.

However there are ways of merging ambiguities through relevance or context discoveries. Like DNA is the fundamental identity of an individual and each person can be identified uniquely through it. Even the difference between the gene of the parents and the progeny differ so much that it is possible to regard it as identifier, though there are large parts of the sequences that can be regarded as genetically common or similar. The browsing history time line can also be regarded as that. The building blocks of any gene are the proteins ACGT, while for the browsing data, the time stamp is linked with the category of the website visited. We have identified 16 such categories, while this value can differ with the difference in assumption and also on the definition of categorization. While defining or categorization, there are specific targets of the designers to get accomplished and accordingly they can be defined. However, our analysis is fully experimental without specificity and as a result the categories are fundamental and arbitrary. We have tried to establish the fact of uniqueness through several comparisons both visually and analytically. Several rigorous experiments are being performed, both through pairing up the users and also through grouping when some of the browsing data of different users are combined to classify the age groups and also their genders. The heat map in Fig. 11 shows that the individuals are unique enough to train the classifiers to differentiate one from the rest and the level of accuracy of being recognized by the classifier is very high. However, gradual mixing up of more and more users' log feature vectors degrades the accuracy which clearly indicates that they are confusing the classifier relative to the dual ones, clearly establishing the fact that random mixing the log files of individuals will never solve the problem of identifying the gender and the age group. Instead if dictionary of non-linear transformations, for different age and gender group, is created, then each individual can identify their own feature subspace where they belong both with respect to age and gender. Defining dictionary helps in dimension reduction because considering each feature as dictionary element will make it very large and unfortunately present systems and models can indulge but can't afford.

The heat map of the corresponding classification accuracy of each user with the others in Fig. 11, for randomly picked samples, makes it apparent the dissimilarity among samples of all types like for all the classes ranging from young male to young females and from older males to older females. Apart from visual analysis, we have done quantitative analysis on the overall in Table 1 based on different data mining and machine learning techniques to establish the fact that even machine identifies distinctly their difference in behaviors and can recognize them easily. Also digital database for specific targets can easily be made.

Log data of the participants, providing less recognition accuracy for the machine learning techniques, are found to produce better accuracy when the feature vector are processed with some kind of feature selection methods like

F1-score, RFE, etc. The enhancement in accuracy testifies the fact that the learning for the algorithms can be more accurate when a proper feature selection is introduced for the reduction in dimension and it must not be taken in granted that less accuracy means that the feature sequences are very close to each other. Instead it means that the model has not being designed properly. Dimension reduction techniques may contribute to better results, only if there are better scores for the new set of generated features. Defining usable non-linear dimension reduction techniques is difficult, however working proficiency of deep learning provides the fact that it is feasible. Dimension reduction techniques can be avoided as it introduces extra overhead for the systems and more emphasis must be given to the feature definition. Feature selection techniques can help classifiers like random forest and support vector machine as success of these algorithms depends on optimization of feature space and feature selection can be a great relief for the optimization techniques.

As different features are extracted from log data of the users and used for analysis, it may be possible that mapping of several users may converge to the same subspace of the new feature domain while still maintaining its separation in a different one. Such analysis will provide how these log data can be used for personalization breach and how to counter them. Also it can provide different measurements or evaluation functions if there is an effort to define better intelligent noise to dismantle the personal identification. If not at least it will definitely help in providing gender and age information to a great extent.

4.3 Proposition from log data

This work started with the hypothesis that the older people are more vulnerable and can attackers make out their targets through personalized analysis of log data from the browsers in real time without the need to transfer the data out of the browser or computer or capturing the data like network traffic etc. But when we looked into the data, it showed much more than that and can be easily extended such analysis for other community detection without explicitly detecting the relationship but just monitoring the behavior.

The problem is not only analysis of target specific cases like gender and/or age, but can be other service specific multiple classifications In Fig. 7, we have provided an overview of the different distributions, different classes of people posses and the things will get finer with the reclassification of the webtypes and the contextual data of the individual. The importance of features need to mined out of the data and there are specific trends based on region and happenings all around. Our primary assumption, that older people are more concerned with the financial aspects, is evident from the distribution than the other contemporaries and reflects the vulnerabilities and has been found to browse Financial Management sites more often than the other categories. If the browsing patterns for other categories are noticed, they also show some joint and disjoint patterns. Like an attacker can create Medical Havoc by making someone run to hospital by creating false health condition in the health monitoring device and create loss of money through unnecessary diagnosis.

Log data can be used to identify groups of people with specific contexts like locational, organizational etc The main analysis for designing identifier is to analyze a data driven scenario of crowd-sourced data and then use it for detection of individuals who reflect close similarity with the pattern and can be regarded as a possible target. These kinds of situations are detrimental and a possible threat to enterprise perimeters and can be used in weakening skill perimeters for organizations. These kinds of situations can easily automated with botnets, who can freely work in critical infrastructure attack and work as point of entry to any organization. Whereas a thingbot is something embedded in systems with access to internet connection that hackers integrate with part of a botnet of networked things.

It can be used for detection billions of people The features of browsing history is very unique and the way machine learning algorithms had responded, it is clear that they are unique computationally, though not apparently. This is because of certain patterns, people gets habituated to daily life and that becomes their digital signature. However like any other research, it is a matter of time that organization can predict with high probability that a person is someone.

Each individual has unique browsing behavior World wide web is vast but human periphery is limited and so is their requirement and liking. This is the reason people gets struck doing the same each and every day. At least the periphery or domain will vary but not deviate much.

SubCategories will fine tune feature diversity but can create sparsity When categorizing the features, we have concentrated mainly on obtaining the best set of web type features instead of considering the web sites. Consideration of web sites will create sparsity in data representations. This is the reason why creation of proper categorical division is important. Like, if two person can be differentiated with respect to social media, then re-categorizing to subcategories will be essential.

Browsing behavior is time invariant The behavior is invariant for any time of the day. That mean irrespective of when the person is browsing, the model can successfully classify. Browsing behavior, in active form or purposefully complete, will be time invariant unless deliberate random browsing. Thus when we had trained deep learning network, we were able to see such behavior and better accuracy than the other machine learning algorithms.

Fig. 7 Percentage of WebType features accessed by each class

The algorithm must be resistant to shift in time scaling. However, there can be differentiation in browsing between morning and afternoon.

Weekday vs weekend browsing data We performed a statistical analysis of the weekday with weekend browsing data and found there is significant difference which clearly established the fact that misclassification error may be due to this as we never tried to differentiate the weekdays with the weekend data for the purpose of training. In fact, whether such data will be significant or not will be decided if large amount of data is collected.

4.4 Categorical analysis log data

In this part we have mainly concentrated on the categorical statistical analysis of the data so that we can understand what kind of websites are being accessed by different groups of people starting from different age, gender, profession and other possible categories that were revealed by the participants when the data were being collected. Figure 7 shows the different percentage of website types web (2017), people of different age group and gender visit and we have used the same data to get the distribution for other categories as well. However since the data was collected on the specific region and aimed at certain hypothesis and a large part of the younger people happened to be from same region, the distribution acuteness is compromised but still reflects some information. The distribution creates the fact that a clear likelihood can be generated for many classes based on the present categorical divisions. Apart from gender and age, distinction between English speaker vs non-English speakers, different degree categories, different income range, different employment status, different race, different marital status, different living conditions are also present.

If we plot some box plot of the number of websites, webtypes and domains for each class based on weekends (Fig. 9) and weekdays (Fig. 8) based on a certain time window scale, we will clearly see some pattern of distribution. Figures 9 and 8 clearly denote that pattern deviates during weekends with more number of web type categories. However as usual, large amount of outliers in some of the cases can be problematic. However this amount of outliers and descent variation is inevitable as the statistics is taken on the whole population of data. However we have tried to see each class like male, female, old, young or like young male, old male etc. When it comes to the relation of statistical distribution with time, we found that the data distribution is not significant as the browsing features are variational and particular features or non-linear transformation of features will be important. This is the reason, why deep learning has performed very well in cases where other classifiers failed as reflected in Table 3 for age categories and Table 2 for gender categories.

Significant observation

– Normal trend is that people visited more kinds of websites during weekends than weekdays and the mean is constant during weekdays and high variance during weekends. This is due to causal browsing during weekend in comparison to specific browsing during weekdays.

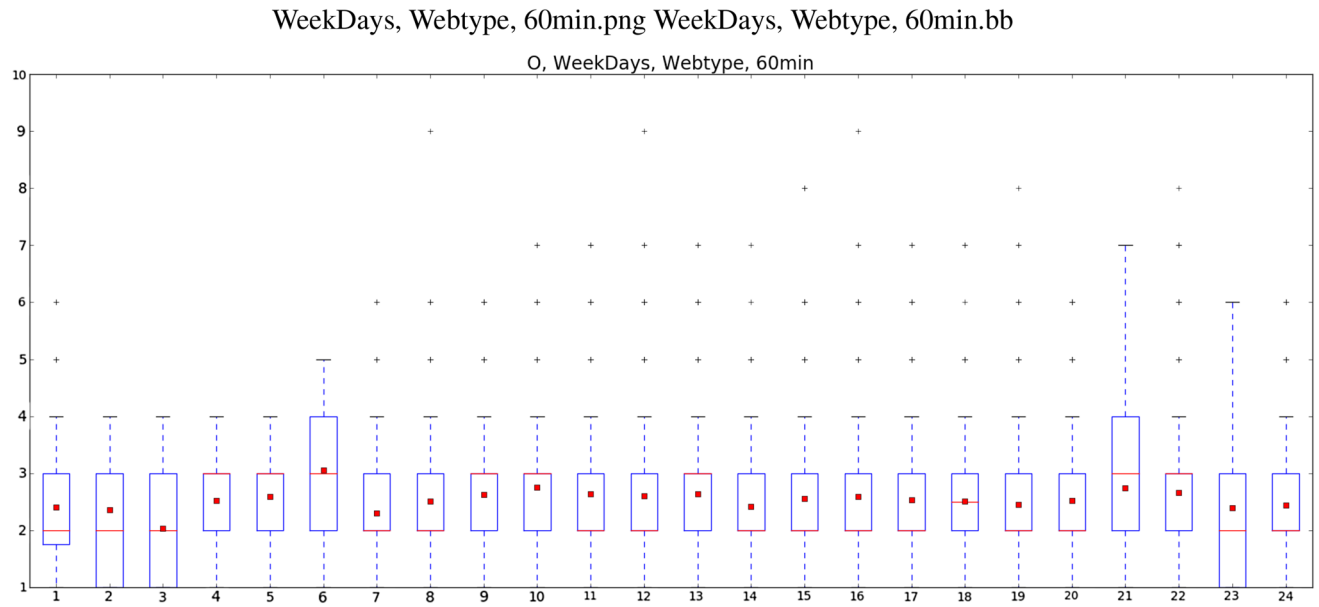

Fig. 8 Percentage of WebType features accessed by older people during weekdays

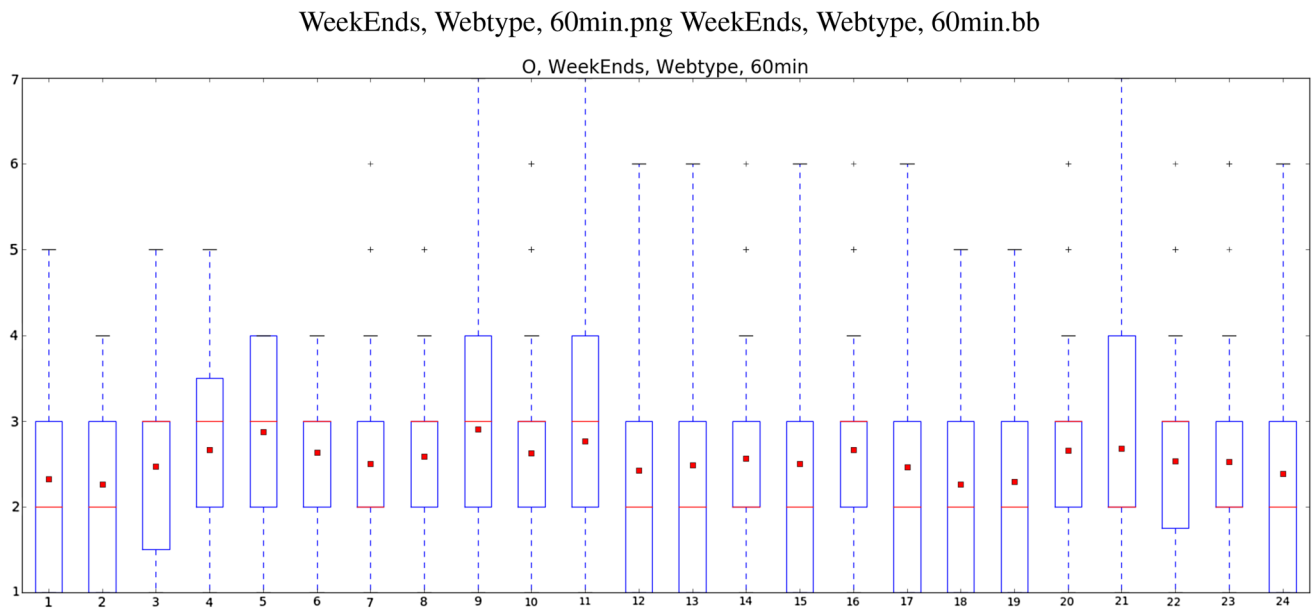

Fig. 9 Percentage of WebType features accessed by older people during weekends

- Old people visit more different types of websites in comparison to younger people.
- Old men browse much more than older women and even some younger people.
- African American people visit to financial websites is high. In gender, it is older male and then older female. people with Masters and Doctorate have high visit of financial sites. Retired people more visit on Financial sites. People with income more than 70K USD visit financial site more often. People with Children visit financial site more often, followed by people with spouse. Married people visit more financial sites than others.
- Professionals visit more TV and Video sites and clothing and recipes.
- Asian and people with Masters have high visit for programming sites.
- People with roommates visit more educational sites and people living alone visit more programming sites.
- Older males have high access to references.
- Younger males have high access to soccer.
- Older females have high visit to shopping category.

5 Analysis, methodology and results

Our main contribution lies in the approach of extraction of wide range of features and investigate the scope of such features in personalized security whereas some of these features can be used to define camouflaged opportunities for the users. Learning browsing pattern can help in identification and this learning can also help in hiding identity, characteristics and digital behavior for users. Later this concept has been used for the analysis and development of optimized noise.

Classifier characteristics for log data must be independent of time dependent data frames that is shift resistant. For that reason, feature reduction techniques like PCA, ICA and other feature selection heuristics should not be enforced but regularized as they kill many components which can be insignificant in training but play important and decisive role. The extracted feature space clearly reflects signature of high non-linearity. Non-linear dimensional techniques like kernelization, manifold transform etc, which operated and succeeded on low dimensional data, do not promise much as the non-linearity they provide is semi-scattered and uni-folded. Contrary, deep learning provides much enhanced feature mixing and scope of a robust adaptive non-linearity. However there are overfiting concerns for deep neural networks.

In reality normal machine learning techniques hardly scale up for multiple class problem and that is also what we found from experimentation that we can never train up one classifier to detect multiple people and there is a gradual decline of accuracy when the number of individuals go up. Tables 2, 3 where grouping people in (30 of each class) witnessed deterioration in accuracy for most of them, but deep learning performed well. A deep learning network with better network and more data perform much better. One reason is the non-linearity combination of features while another is that they use features as discriminative objects while in deep learning, they are generative features. Non-linearity indicates that there is a considerable overlapping of the feature

subspace for a linear classifier to act in the region. When we use deep learning techniques, the features were transformed to another feature space through the help of the alternative linear and non-linear functions of the hidden layers and the behavior of the classifiers scales up. Discriminative features work on classification locally whereas generative features helps globally. This is the reason there had been considerable paradigm shift in practice from using discriminative features to generative features through deep learning architectures. However, it is foreseen that the performance can be enhanced and tuned once the mapping of the websites to its corresponding web type category is refined through decompression, smart and optimal representations, so that it is useful and at the same time it is well within the control of the classifier.

The various analysis, implications and capacity of the collected data are being made here. However, before going into the details, it must be mentioned that the algorithms, with the same set of arguments, are applied on each features so that comparison is unbiased. With different grid search techniques, the different parameters of the algorithms can be tuned to get the better estimations. The main reason of doing this, is that the analysis is not to design the best classifier(s), but to analyze the capacity of the data and what can be done with it mainly from the prospect of personalized security. Figure 7 made some sanity checking of the distribution of data for different classes and we had a visualization

Table 1 Analysis for accuracy for different classifiers for different features extracted from log files

	Perceptron			SVM			Random Forest			Deep Learning			Logistic regre.			Naïve Bayesian		
	μ	σ^2	> 9.5	μ	σ^2	> 9.5	μ	σ^2	> 9.5	μ	σ^2	> 9.5	μ	σ^2	> 9.5	μ	σ^2	> 9.5
		*	> 8.5		*	> 8.5		*	> 8.5		*	> 8.5		*	> 8.5		*	> 8.5
Time	0.92	0.51	39.4	0.93	0.55	42.0	0.93	0.45	44.8	0.88	0.62	24.1	0.93	0.50	47.0	0.92	0.53	36.2
			84.8			84.5			87.6			39.6			86.7			83.2
WType	0.89	0.64	23.4	0.89	0.68	25.9	0.89	0.59	25.8	0.94	0.25	39.6	0.90	0.64	29.0	0.95	0.35	53.7
			70.3			69.6			72.3			85.1			74.0			92.4
Speed	0.89	0.59	20.0	0.88	0.77	22.4	0.89	0.57	22.5	0.89	0.54	29.0	0.89	0.74	26.1	0.87	0.82	17.6
			71.7			69.1			71.0			42.7			73.3			65.0
GroupWTc	0.92	0.43	37.4	0.92	0.48	36.0	0.92	0.37	39.0	0.93	0.45	29.5	0.93	0.47	41.1	0.88	0.89	21.5
			86.5			84.0			88.4			85.6			86.6			67.1
GroupWTs	0.96	0.21	77.3	0.97	0.23	71.4	0.95	0.30	55.0	0.94	0.67	41.3	0.97	0.22	74.0	0.95	0.33	62.7
			95.2			97.0			94.1			79.5			97.2			94.1
Time+	0.89	0.59	23.3	0.89	0.61	26.7	0.90	0.52	30.0	0.91	0.26	36.4	0.90	0.60	29.6	0.90	0.67	28.8
WType			72.4			72.4			77.0			63.8			75.8			73.5
Time+	0.89	0.58	22.7	0.89	0.66	23.0	0.92	0.49	34.4	0.92	0.39	28.9	0.90	0.63	28.6	0.89	0.62	22.2
Speed			74.4			72.4			84.4			81.3			76.9			73.5
Time+	0.92	0.54	37.9	0.92	0.54	37.9	0.93	0.46	41.2	0.91	0.52	25.6	0.93	0.52	45.1	0.94	0.45	48.6
GroupWTc			84.4			84.4			87.0			79.2			86.4			89.8
Time+	0.94	0.40	48.7	0.94	0.44	50.6	0.93	0.43	45.2	0.90	0.41	21.5	0.94	0.42	55.2	0.93	0.42	45.1
GroupWTs			89.2			89.0			88.3			81.0			91.0			88.6
WType+	0.90	0.53	25.6	0.90	0.62	30.2	0.92	0.49	40.3	0.88	0.75	23.2	0.91	0.60	33.6	0.90	0.62	28.2
Speed			79.1			75.1			84.7			61.6			78.7			75.4
WType+	0.92	0.40	33.5	0.91	0.51	31.8	0.93	0.41	43.1	0.91	0.35	38.9	0.92	0.48	38.8	0.93	0.46	39.0
GroupWTc			87.0			82.5			89.1			79.0			85.7			86.4
WType+	0.89	0.63	24.8	0.90	0.71	31.0	0.93	0.46	42.7	0.92	0.42	42.7	0.90	0.68	33.9	0.89	0.67	24.0
GroupWTs			72.3			72.4			86.6			85.1			75.1			71.1
Speed+	0.92	0.43	41.6	0.93	0.44	42.9	0.94	0.34	53.4	0.91	0.45	24.7	0.94	0.42	48.0	0.94	0.41	47.2
GroupWTc			83.0			87.5			92.8			82.8			89.3			90.5
Speed+	0.90	0.57	21.9	0.89	0.59	21.2	0.93	0.39	36.5	0.91	0.81	29.7	0.90	0.58	26.7	0.90	0.63	29.0
GroupWTs			77.0			74.8			88.8			87.4			79.3			78.0
GroupWTc+	0.94	0.38	47.0	0.93	0.48	45.4	0.94	0.38	51.8	0.94	0.35	59.8	0.94	0.42	53.8	0.95	0.37	53.1
GroupWTs			89.7			86.8			91.7			89.9			90.8			92.4

> 9.5 (or 0.85) means the percentage of pairs of individuals with 5-fold cross validation accuracy above 0.95 (or 0.85)

Table 2 Accuracy of gender (30 males and 30 females grouped as male vs female class) for different classifiers for different features extracted from log files

Algorithm ⇒	Perceptron			SVM			Random forest			GBDT			Deep learning			Logistic Regre.			Naïve Bayesian		
Features ⇓	10	20	50	10	20	50	10	20	50	10	20	50	10	20	50	10	20	50	10	20	50
Time	0.85	0.67	0.68	0.84	0.65	0.69	0.88	0.67	0.72	0.75	0.71	0.64	0.85	0.85	0.85	0.87	0.67	0.70	0.79	0.65	0.66
WType	0.78	0.60	0.63	0.72	0.60	0.61	0.82	0.65	0.71	0.82	0.68	0.60	0.84	0.85	0.85	0.82	0.58	0.63	0.80	0.63	0.64
Speed	0.69	0.68	0.70	0.66	0.70	0.65	0.70	0.68	0.74	0.65	0.60	0.62	0.84	0.84	0.84	0.69	0.61	0.66	0.73	0.66	0.64
GroupWTc	0.67	0.65	0.68	0.65	0.64	0.65	0.65	0.74	0.73	0.72	0.66	0.61	0.82	0.82	0.86	0.82	0.69	0.66	0.77	0.68	0.69
GroupWTs	0.82	0.70	0.67	0.73	0.69	0.66	0.77	0.74	0.71	0.75	0.68	0.60	0.82	0.82	0.85	0.81	0.75	0.69	0.80	0.77	0.70
GrWTc+GrWTs	0.83	0.71	0.69	0.83	0.76	0.71	0.85	0.76	0.73	0.86	0.72	0.75	0.85	0.83	0.85	0.78	0.72	0.69	0.80	0.79	0.69

Table 3 Accuracy of age (30 older people and 30 young people grouped as old vs young) for different classifiers for different features extracted from log files

Algorithm ⇒	Perceptron			SVM			Random forest			GBDT			Deep learning			Logistic Regre.			Naïve Bayesian		
Features ⇓	10	20	50	10	20	50	10	20	50	10	20	50	10	20	50	10	20	50	10	20	50
Time	0.80	0.75	0.74	0.82	0.71	0.66	0.79	0.77	0.74	0.73	0.69	0.68	0.85	0.84	0.84	0.92	0.77	0.76	0.83	0.79	0.77
WType	0.74	0.73	0.70	0.81	0.66	0.62	0.85	0.71	0.70	0.80	0.74	0.72	0.82	0.85	0.85	0.78	0.68	0.67	0.83	0.76	0.75
Speed	0.72	0.70	0.69	0.72	0.69	0.66	0.80	0.73	0.71	0.75	0.71	0.69	0.82	0.84	0.83	0.70	0.69	0.67	0.77	0.73	0.72
GroupWTc	0.80	0.80	0.78	0.79	0.75	0.77	0.79	0.83	0.79	0.82	0.70	0.65	0.85	0.86	0.85	0.79	0.77	0.66	0.86	0.83	0.82
GroupWTs	0.81	0.78	0.75	0.80	0.75	0.75	0.79	0.78	0.74	0.80	0.79	0.69	0.85	0.84	0.84	0.81	0.81	0.77	0.86	0.82	0.79
GrWTc+GrWTs	0.83	0.80	0.81	0.82	0.80	0.79	0.83	0.83	0.82	0.81	0.75	0.71	0.84	0.85	0.85	0.84	0.81	0.76	0.89	0.84	0.83

of behavior and distribution of the different parametric features throughout a timeline in Figs. 2, 3, 4, 5 and 6. We also had a better understanding of the data beyond the numerical information through the heat map in Fig. 11, where randomly picked 100 people are being used to get a one vs one accuracy and we have estimated the mean, variance and the percentage of the dual above a certain threshold in Table 1. However, for the heat map in Fig. 11, the matrix for 60 people clearly indicates, no matter what are the feature sets, they are still very distinct and unique for each individual and the high range of accuracy clearly denotes that a digital database for each individual can easily be made, like any biometric databases like gait etc.

The heat map plot in Fig. 11 also shows the different users marked by their age group and gender and indicates that their browsing histories are independent of each other and each can be differentiated from the rest with some accuracy. There are some cases where the accuracy is very low. The classification accuracy, in Fig. 11, is derived from the fivefold cross validation for each two individuals and used Naïve Bayesian classifier . However, a quantitative analysis of the different algorithms for 100 individuals (with 50 males and 50 females, which happened to be also 50 younger and 50 older adults) is also provided in Table 1 with mean, variance and percentage of pairs of individuals with fivefold cross validation accuracy above 0.95 (0.85). This will provide a clear indication of how the data behaved. Also, we have selected 100 individuals randomly through pre-shuffled data and the result of analysis is unbiased towards any particular 100 individuals. We have shown the grouping significance of 30 males and 30 females in Table 2 and also 30 older and 30 younger adults in Table 3 and have produced an overall comparison scenario in Table 1.

For user classification based on the person, we have sampled users from each of the classes. Classification with respect to age and gender had training set from a mixed number of users so that the true distribution and characteristics of the classes are generated. However, it must be declared that the analysis, that has been done, is based on the data we have collected from a specific geographical region and with certain criteria and objectives in mind. A generalization will involve more number of parameters like language, preference, geographical influence, business scope, life style of people living there, life objective of people involved and nonetheless the policies of the government and legal system involved for the specific region. Since our objective is specific and geospatially target based, our analysis is justified and may or may not be applicable with the change in the parameters specified above. In the next section, we have experimented with different kinds of noise which are generated and optimized through various algorithms and a thorough analysis of the collected data.

Table 4 Deep network (MLP) parameters

Layer no.	Size	Activation	Dropout
0	Input shape	None	0.2
1	1000	ReLu	0.2
2	800	ReLu	0.2
3	600	ReLu	0.2
4	400	ReLu	0.2
5	200	ReLu	0.2
6	Class shape	Softmax	0.2

Multi-layered perceptron (MLP) based deep learning models achieved an accuracy of around 85%, which is much higher than the other machine learning counter parts. Some cross-validation accuracy was sometimes low because of the fact that browsing is a soft authentication data and may contain outliers, which is acceptable in browsing behavior of users and is obvious sometimes. This might have lowered the average of the cross-validation. A typical details of deep learning model parameters is provided in Table 4.

The weights of the layers are initialization with normal distribution from $\{-0.1, 0.1\}$ and the regularizer used is $L2$ norm. Relu is used as activation function and RMSprop (or Adam) has been widely used as optimizer. We used 64 batch size and ran the training session for 20 complete epochs. The default value for the learning rate was kept at 0.001. We performed a wide range of parameter search to understand the optimal number of layers and optimal size of each layer. Random Forest was used with 1000 number of trees and gini information, max features had sqrt(number of features). We also applied Gradient Boosting Decision Tree (GBDT) for our analysis and found that the performance is not as good as Random Forest though both of them used similar kind of ensemble approach. The main reason is the over division of feature space by GBDT for finer classification, which is good for simpler and low dimension problems, but create over-fitting and more testing errors. Popular parameters of SVM were L2 as penalty, squared hinge loss, $C = 1.0$ and default stopping criteria for optimization. Perceptron used elasticnet penalty and $\alpha = 0.000001$, $\eta = 0.8$. Naive Bayesian and logistic regression (penalty L2) parameters were kept unchanged. The hyper-parameters were experimented for optimal. Most of the experiments were done with 10-fold cross validation. The outliers were not removed to understand their presence in the model and is also difficult to estimate. If outliers were removed, the performance should definitely be higher. The aim of this work is to find the best possible model and at the same time provide a thorough analysis and understanding of the interaction of different components (including unusual behaviors) on the model.

6 Noise for personalized security

This section discusses on how to deceive artificial intelligence through careful and intelligent mixing up noise with your browsing history, so that a browser can preserved integrity and identity of a person, in terms of age, gender, location, profession and all other criteria that can be defined or detected from the browsing history. There are apparently no criteria of how one should or can go about it as noise can vary and there is no indication of which part of the personal integrity is really important. Like for example, with older people vulnerable more to phishing attacks (Sur 2018), it is important to take care of the age group, while among the old people, women are more prone and so the gender group as well. But the question whether it is possible to get the transformation easily. Also whether it is possible to get the transform in both the direction instead of an unbalanced one. When it comes to browsing history, it follows a certain pattern or windows of parametric distributions. Each of such when collectively defined can create a conception of a dictionary and a collection of individuals can define an individual. Also it is apparent from the study, we have made from our collected browsing history collection, that younger ones visit more sites in a time span than an elder one. As adding noise will increase the browsing history, it is more likely and easier to transform a older adult data to younger adult data than vice versa. Several feature vectors have been generated from the browsing history data, each of them unique and different from the others. They also imply differently for different users. For the analysis and generation of noise, these parameters will be of utmost importance. The aim would be to get the best out of all. But that is quite challenging to achieve. Like for the same noise, each of these features vectors will land up differently based on the defined topology and metrics. As a matter of fact, different transformations are to be introduced and since the concept and the content of the dictionary comes from the data itself, it will depend on it solely.

When defining noise with respect to different defined criteria, it becomes a multi-objective optimization problem and it becomes very important to sought out some criteria that can rank them according to efficiency (maximization of accuracy) and profit (minimization of energy or resources consumed) achieved from individual and as a whole. No doubt, both are conflicting (non-dominant) and create a Pareto optimality and need heuristic solutions. It is a NP hard problem and needs efficient heuristics. Unlike many other problems where dimensional reduction can be applied, the scope and effect of such kind of transformation is very less feasible or may be trivial. Also the feature vectors actually represent a time line for each day and its transformation to a compressed version for some of the features might not be good idea. Now we will subsequently define and derive the noise functions through some case studies and that will make the procedure much easier to perceive and also how the global dictionary can be defined.

However, as we have defined the browsing history of the users as their "digital DNA", it is likely that define-one-fit-all is very least likely be successful and a dictionary structure with least error can be the solution. Each member of this Deep Learning based dictionary structure will transform a certain browsing data gene to another form which can certainly confuse and conceal the true identity of the individual. Now when it comes to security, there are certain priorities that need to be followed and the noise trend also should not be a trend itself. This is because if the same noise level contaminates the browsing history of the same person, then it can again be held along as unique mark. Instead a number of member transformations should be developed and should be selected randomly, to get the browsing pattern to several different levels. This is quite challenging to develop a system that can accommodate several uncorrelated transformations, taking care of the categories of the websites and at the same time take care of gender, age groups, profession, locations and requirements.

However, the noise quality and quantity need to be determined and is itself a topic of intense research and there are deployment issues. The quality and quantity is very specific and are dependent on various factors and parameters and must not be generalized. But a series of experimentation can create dictionary of such noise with variety and veracity. We can try to establish the fact that if proper noise is added to the user log data and is brought to a common platform where the log data of different users can not be differentiated through machine intelligence, then it may help in hiding the age and the gender group. This section should be dealing with the details of the procedure of the experiments but instead of focusing deep into the descriptions and formalities of the machine learning algorithm, we will go through the procedural details in brief and will concentrate on describing the noise generation and optimization techniques. The efficiency of the noise will be evaluated in this section through two parameters accuracy and power units.

How to deceive AI, through careful and intelligent mixing up junks in browser to preserve integrity and identity of a person, is a complex phenomenon. Complexity arises because of non-linear feature processing and thus requires more than just random generation. Here we have come up with a engine that can understand the browsing behavior without acquiring specifications and operates paralleling to complement the real identity with other website access. Browsing follows a certain parametric distributions. Each

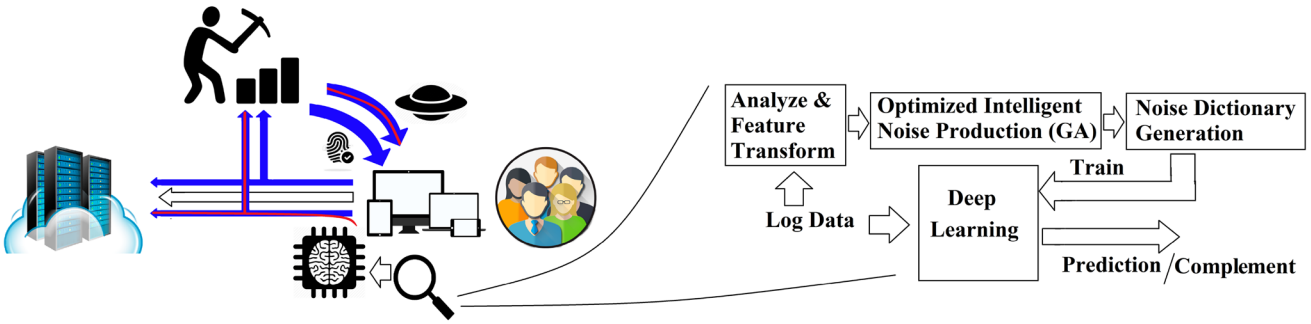

Fig. 10 Architecture of the intelligent prevention system in browser

of such when collectively defined can create a conception of a dictionary and a collection of individuals can define an individual. Algorithm 1 describes such a system (Fig. 10) capable of learning a noise dictionary model, optimized with genetic algorithm. We call it deep security network (DeepSeq) and is more than deep neural network.

6.1 Deep security network (DeepSeq)

Deep network based noise or noise from deep security network (DeepSeq) is a coordination of optimization and regression based Deep Learning or other embedded representations can also be used like Restricted Boltzmann Machine. Mathematically, $f(\mathbf{x}) = \{\mathbf{WtN} : \text{DNN}(\mathbf{x}) \in \mathbb{R}^n \rightarrow \mathbf{WtN} \in \mathbb{R}^m\}$ where **WtN** is the generated Webtype noise and satisfying conditions like {min Energy(**WtN**)} and MLA(**x**,DNN(**x**)) ~ MLA(DNN(**x**)) is not equal to $y_{\mathbf{x}}$ for data ($\mathbf{x}$,$y_{\mathbf{x}}$). DNN() is the Regression Deep Neural Network and MLA() is the category predictor and can be any machine learning algorithm. ($\mathbf{x}$,$y_{\mathbf{x}}$) is the training pair where $\text{MLA}(\mathbf{x}) \rightarrow y_{\mathbf{x}}$. Noticeably, we are optimizing noise with DNN() and to increase the precision and less variation, the likelihood is maximized as MLA(). Deep noise is not regression matched which is denoted as DNN(**x**). Instead, the most popular and optimized regression outputs are selected heuristically and a dictionary is created that representation the noise that have maximum influence in hiding the identity of the individual. Maximum influence is generated by the representation in the dictionary and not that generated by the DeepSeq. Hence, the job of DeepSeq representation is to maximize the likelihood and to select the best possible representation from the dictionary. In DeepSeq, we used a softmax layer MLA(DNN(**x**)) and converted the problem into a classification problem and the task of the DeepSeq is selection of the best from the dictionary. This is done through joint optimization of the mean square error and classification cross entropy.

The principle of DeepSeq is not deception of the algorithm through generation of look alike samples, but to hide own individual characteristics. Like if there are certain people with x_n trait and there are individuals with x_1, x_2 traits etc, x_1, x_2 must try to be x_n to hide their identity, instead of remaining as x_1, x_2. It leverages on the fact that noise mixing requires a very time efficient and adaptive mode of decision making and hence uses constant time based predictions from Deep Network. This also means that instead of defining "one-for-all" or "different-for-different", we are interested in "all-the-best" approach, where we shrink the search space and also search it in constant time. In real time, trained deep network operates as an adaptive control system capable of predicting the noise spontaneously with the browsing. Figure 10 shows the overall architecture of deep learning based optimized noise system and correspondingly what is happening in reality.

Require: LogData, Label, Wtype;

# Generate Prediction Model
for each $LogData_i$ **do**
  Train to Model {MLA($LogData_i$) → $Label_i$} satisfying f
  where f = arg min ∑ (MLA($LogData_i$)-$Label_i$)
  # Any Machine Learning Algorithm
end for

# Statistical Analysis of LogData
for each $LogData_i$ **do**
  for each $Wtype_i$ **do**
    Generate Distribution
    $(Distribution)_i \leftarrow (\mu, \sigma, \min, \max \text{ etc})_i \; \forall i \in \mathbb{Z} \in [1, m]$
  end for
end for

# Generate Initial Noise with Above Statistical Seed
for each $(\mathbf{WtN})_i$ with $i \in (1, 2, \ldots, 2c)$ **do**
  Generate $(\{WtN_1, \ldots, WtN_m\})_i$ for each $(\mathbf{WtN})_i$
end for

# Run Genetic Algorithm for Optimization
for each iteration $k \in [1, \ldots, K]$ **do**

  # Fitness Calculation
  for each $(\mathbf{WtN})_j$ **do**
    for each $LogData_i$ **do**
      FitnessTemp $\leftarrow MLA(LogData_i + (\mathbf{WtN})_j), Label_i)$
    end for
    Fitness$(\mathbf{WtN})_j \leftarrow \sum$ {FitnessTemp$\in \mathbb{Z}_{\forall \rightarrow False}$}
  end for

  # Discard the Least Efficient Noise
  $(\mathbf{WtN})_{\forall j} \in \mathbb{R}^{2c}\mathbb{R}^m \rightarrow (\mathbf{WtN})_{\forall j} \in \mathbb{R}^c\mathbb{R}^m$

  # Selection Randomly/Best and Crossover
  for {i1,i2 $\in \forall j \in (\mathbf{WtN})_{\forall j}$ } **do**
    tempFit → Fitness(Crossover($(\mathbf{WtN})_{i1}$,$(\mathbf{WtN})_{i2}$))
    if Fitness($(\mathbf{WtN})_{i1}$,$(\mathbf{WtN})_{i2}$) isBetter tempFit **then**
      $(\mathbf{WtN})_{\forall j} \cup$ Crossover($(\mathbf{WtN})_{i1}$,$(\mathbf{WtN})_{i2}$)
    end if
  end for

  # Selection Randomly/Best and Mutation
  for {i1 $\in \forall j \in (\mathbf{WtN})_{\forall j}$ } **do**
    tempFitness → Fitness(Mutation($(\mathbf{WtN})_{i1}$))
    if Fitness($(\mathbf{WtN})_{i1}$) isBetter tempFitness **then**
      $(\mathbf{WtN})_{\forall j} \cup$ Mutation($(\mathbf{WtN})_{i1}$)
    end if
  end for
end for

# Select the Best Noise for Each Sample
for each $LogData_i$ **do**
  for each $(\mathbf{WtN})_j$ **do**
    if MLA($LogData_i$ + $(\mathbf{WtN})_j, Label_i$) → False **then**
      if {Energy$(\mathbf{WtN})_j$ > Threshold } **then**
        TrainingSet $\leftarrow (LogData_i, \mathbf{WtN})_j)$
      end if
    end if
  end for
end for

# Training The Deep Noise Network
for each ($LogData_i$,$(\mathbf{WtN})_i$) with $i \in$ TrainingSet **do**
  Train to Model {DeepSeq($LogData_i$) → $\mathbf{WtN}_i$}, satisfying f
  where f = arg min ∑ (DeepSeq($LogData_i$) - $\mathbf{WtN}_i$)
end for

# User Study in Reality
for each $LogData$ **do**
  Transmit {$LogData$ + DeepSeq($LogData$)}
end for
return DeepSeq()

Algorithm 1: DeepSeq Based Dictionary Generation

From Table 1, we can see that irrespective of the gender and age, each individual is very different from the other and there are chances that they get bookmarked. Also behavior generalization could have been difficult. This is apparent from the different kinds of feature vectors being generated from the same browsing data spanned over a period of more than 20 days of time. However, the data and the criteria of noise changes with time, location and status of the region. There is a requirement where the system must keep track of the changes in behavior and update itself and the criteria of noise, when it comes to defending the personal integrity of an individual. From the analysis in Table 5, it can be seen that all the analysis is mainly focused around the machine learning algorithm evaluation functions. The main reason is that browsing history is a trend and not distinct features. Unlike biometric recognition, where the feature vector gets tallied with each and every specimen of the data base, gene protein matching engages with insertion, deletion and replacement without fully dependent on strict matches. So does the dDNA, where the topology may or may not be pertinent in time scale but the trend in transformed space definitely has important information. Perhaps this is the reason why the terminology dDNA is pertinent for the browsing history of any person. But yet again people will start debating on scalability as the number of people is large and is it feasible to contain large browsing history and trend of people? Browsing history in time scale is compressible and the compressing factor or time window is debatable. However, distributed machine learning and deep learning based infrastructure can promise dDNA based recognition.

The defined noise will be another continuous web requests from the client side in parallel to the browsing. This will be the noise to deceive the browsing pattern of a person. But the noise is well defined instead of random. Like if a person in US visits news channel, his noise can be some news channel from India to deceive geographical identity while some sport news channel from Japan can hide his age identity, while some Stock Market news will deceive his profession. However all these are related to News category. So some Education site and Entertainment site are accessed to confuse the real pattern of the person. This is spontaneous and the selection of web access is random and highly uncorrelated to browsing pattern. But the noise description, we have defined, is a mapping from website to web type in the form of numerical $f = \{W : \varphi(Website) \rightarrow WebType, \phi(WebType) \rightarrow \mathbb{R}^n\}$ as analysis becomes easier. While we have defined different random noise in traditional ways, the criteria were ineffective because of the distribution of defined website types features and lack of intelligence involved.

Table 5 Accuracy of Machine Learning Techniques (+Feature Selected) for Different Noise

Algorithm ⇒	DL	+FS	RF	+FS	SVM	+FS	Percep	+FS	LR	+FS	NB	+FS
Without Noise	0.91	0.98	0.93	0.95	0.90	0.96	0.90	0.96	0.93	0.96	0.95	0.98
Gaussian Noise 1	0.90	0.96	0.74	0.26	0.89	0.93	0.89	0.98	0.86	0.93	0.40	0.89
Gaussian Noise 2	0.91	0.96	0.85	0.69	0.88	0.91	0.88	0.91	0.89	0.93	0.88	0.90
Correlated Noise	0.91	0.97	0.93	0.94	0.89	0.96	0.90	0.96	0.90	0.96	0.95	0.98
Intelligent Noise 1	0.65	0.76	0.78	0.47	0.67	0.83	0.77	0.85	0.70	0.85	0.25	0.57
Intelligent Noise 2	0.79	0.85	0.79	0.44	0.75	0.83	0.75	0.85	0.74	0.85	0.74	0.59
Intelligent Noise 3	0.79	0.83	0.79	0.44	0.75	0.83	0.75	0.85	0.74	0.85	0.74	0.59
Genetic algorithm-DeepSeq	0.61	0.62	0.55	0.62	0.62	0.68	0.61	0.65	0.63	0.65	0.63	0.51

6.2 Noise descriptions

Gaussian noise is the most traditional noise that is widely used in signal analysis and has been seen to have described majority of the noise distribution of the nature, effectively described by two of its parameters, mean and variance. Also, nature tends to follow Gaussian distribution for likelihood of events. But the higher probability of occurrence of some range over the other can inhibit the main purpose of noise selection and creation of confusion. Table 5 shows that such random noise does not help much. Also the noise being defined here is much optimized with respect to the browsing content.

Gaussian noise 2 (Throughout Active Timeline) operates with browsing only and is feedback with the browsing pattern of the person and the content of the noise is random. While random content can hardly spoil the robustness of analysts and the models, relevance and that even calculated is most effective.

Gaussian noise 1 (throughout timeline) operates both with and without browsing activity but is expensive for the user and the infrastructure and not a feasible solution in reality. We considered this to see the behavior of the robustness of the time invariant model. A model is not properly designed if it gets manipulated with this kind of noise.

Correlated noise relates with adding similar web types, equivalent to linear transform of feature spaces. Optimization of noise started with this but there was not proper intelligence involved. It is still random and intelligence was required on when, how much and whom. The importance of Correlated Noise is to see how far the range of features is important for the models. It is seen that such addition have little effect and proper content of the noise would be inevitable.

Intelligent noise with data driven approach where statistics of data is feed for generation of noise to help decision of optimized noise but the range and scope of generation is predefined. This type of generation will provide a better seed for the optimizer to get to better search spaces. However the magnitude is random and the perfection in magnitude and better combination generation through the use of the existing ones was produced by genetic algorithm.

Intelligent noise 1 (randomly generated) Noise distribution is generated randomly within ranges and the ranges are derived from the data through some procedure like subsampling and considering statistical analysis.

Intelligent noise 2 (randomly seeded with highest value + variance) This kind of noise distribution considers the maximum of the different web type counts while variance ranges are assigned as effectiveness is reflected with experiments.

Intelligent noise 3 (randomly seeded with average value + variance) This kind of Noise distribution considers average instead. Effective variance ranges are also determined with experiments. Intelligent Noise helps in generating and selecting one universal noise, which was further enhanced through heuristic search and fitness evaluation like genetic algorithm.

Optimized intelligent noise (Since random generation cannot provide a thorough sweep of the search space and also all combinations) genetic algorithm evolves and learns from previous experience, produces better combinations, segregates better partial solution segments and makes the life of optimization shorter. However, there are couple of other algorithms that can be used like simulated annealing, differential evolution etc. genetic algorithm still needs the statistical seed as it provides a concrete bound on the search space. While these procedures created one universal noise, procedure to create dictionary of Noise from existing and retrieve them in constant time was inevitable for better performance and to counter the fact that there is always apparent variance in the behavior of user. Whether the feature combination methods of deep learning are independent of such variance is yet another research issue in modeling. Evaluation criteria for the dictionary quality (like decrease in error gap between the browser and the dictionary elements) is based on least square error and can be defined as a regression data driven based prediction where the parameters are optimized through learning and not generated through closed form. Deep learning based noise prediction engine, which learns from a set of pairs

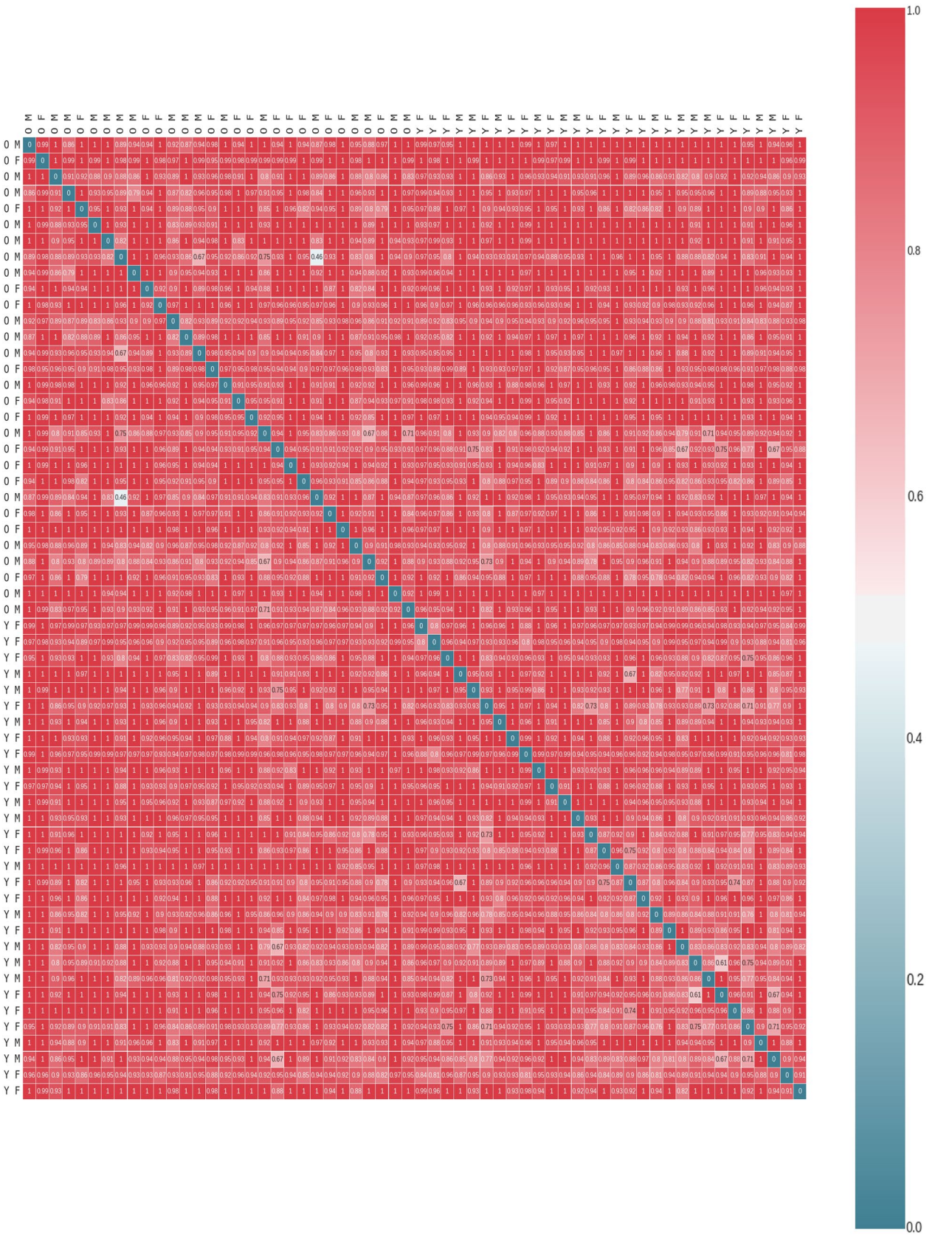

Fig. 11 Projection of ability for learning individual dDNA for different samples using deep learning

of (data,noise) combinations and arranges itself with the pattern, can generate them in constant time complexity and also can provide the best distortion in personalized browsing data.

6.3 Camouflage concept in noise

The concept of noise differs on several grounds and the main question is how noise optimization is different? In traditional signal processing, the noise is something which considers the detection parameters land up in foreign judgment. In disguise the internal structure of the parameters gets camouflaged to something else. But the problem gets worse if the parameters get them all land up in the same class of detection. This is what we have considered as noise for browsing data where the aim is to get all the participating classes to get to the same classification and that will make the classifier incapable of judgment and also in that case, data collection and characterization will tend to fail for any analyst to come to conclusion. This is equivalent to the phenomenon of camouflage where an animal changes its color according to the background. Here the purpose is to design such kind of camouflage where some will hide behind the others through processing. But the addition of noise is not individual specific but browsing content based context specific. The main reason of this kind of context specific camouflage is the processing power. Addition of noise will consume some browsing and processing of content and modern day computer will always reduce that as much as possible. For example, if C_1 class accessed x_1, x_2, x_3, x_4, x_6 web types and say C_2 accessed x_1, x_2, x_4, x_6 web types, then the noise will be to add x_3 to C_2 instead of x_5 to both the classes. The quantitative analysis will be decided by the participating optimization algorithm and while simulation with genetic algorithm has shown such behavior where the optimized noise will tend to get certain classes to a particular class so that the involved energy is minimum.

6.4 Optimization with genetic algorithm

Noise for browsing data is different for any other engineering application. Normally, the noise gets modeled to get a better prediction of deviations in the system and to overcome proper behavior. In this case, the noise is to ensure deformation and for that reason, there is absolute requirement for constraints to get the better out of it. Also there are cases, where constraints can not only help in better evaluation and produce bounded solutions, but also saves around resources that are directly or indirectly related to events. We have seen such situations in this case, but the optimization problem has no numerical quadratic form so that traditional optimization methods can be used. Also the maximization method for accuracy has no functional relation with the data. Hence, heuristic optimization method has been used for optimization and the accuracy in learning is used for judgment of optimization. The cross over of genetic algorithm occurs through certain pivot of the chromosome strings and mutation occurs at targeted point. In our case, we have used two or three pivots for crossover and only considered those chromosomes which have high ranges of obfuscation for the browsing history and for determination of accuracy in Table 5 we have only considered one noise model for all, while in deep learning, we have tried to come up with models where data driven noise model can be generated or prescribed from the existing dictionary of noise models. The main reason doing so is the genetic algorithm is a time and resource consuming optimization procedure. It can be paralleled easily but takes sometime to come up with near optimal solutions, which is not feasible for real time applications. However, deep learning based models can predict the noise in constant time and is quite feasible for real time applications. For training the deep learning model, we have used genetic algorithm to come up with optimized noise model based on the collected data and considering that this would be the behavior of the user, we have derived N such noise models and feed in to the deep learning architecture as $(Log, Noise)$ pairs and made it to learn them. Table 5 produced a comparison of the models and how much success we have achieved based on our procedure.

6.5 Dictionary of noise

Genetic algorithm based optimization has shown that one-for-all model hardly suitable for the best performance and with the gradual change in behavior, the noise model must change or rather adapt to the situation. For this reason we have come up with a deep learning prediction model that can learn the user behavior based on the output of the genetic algorithm and later on helps in prediction of the optimized noise. However the challenge is choosing the size of the dictionary and whether inclusion of lots of noise model is actually helping the users to hide their identity. But so far as our collected data is concerned we have noticed that inclusion of lots of noise data for the users actually has enhanced the prediction of the deep learning architectures. Figure 10 has provided the diagrammatic overview and Algorithm 1 has provided the overview of procedure to create dictionary of noise from existing user browser study with accuracy results in Table 5.

6.6 Evaluation for dictionary quality

Evaluation criteria for the dictionary quality is important and depends on the browsing pattern of the users. As a general they also depend on the different sets of categorical analysis initiated. Such initiation will help in understanding the

important features that are relevant and play a decisive role. For the part of optimization of noise for the different browsing activity data, we have directly used the trained machine learning model for evaluation and accuracy is used. Minimization of the overall accuracy is the ultimate goal through we have shown that the bound is constrained as the task is not to reverse sets of levels but to obfuscate some (minority) into the others (majority). But the amount of processing power consumed becomes important as there are some physical and separated activities involved. So the aim of the model is to take up the best of the combination of minimization of accuracy and minimization of power involved. Minimization of accuracy and minimization of power are non-dominant entities as minimization of one will increase the other and vice versa. In fact the fitness must involve both to create a multi-objective optimization. However the power model introduced can just be a reference and not absolute and may be a kind of activity can easily handled by the appropriate port without involving the browsers and processor much. The power model consists of the number of requests being made for each of the sample used for classification and the noise task can be regarded as decrease in error gap between the actual browser activity and the prescribed browser activity though the help of dictionary noise elements.

6.7 Deep learning prediction from dictionary

Deep Learning based prediction of noise from dictionary based on browsing status can be visualized in two folds. One is prediction from a dictionary where the scope of error is very less but is very inflexible. Another is the Deepseq kind of predictor where the work of deep learning architecture is to prescribe noise based on the stuffs it has learnt. Deepseq requires lot of data to get itself tuned to the level of optimization and after that it can perform well. It is highly efficient and flexible model and most of our result is based on this model as it can get itself adapted that the user can undergo. While dictionary based prediction consolidates on the fact that browsing of people is stable. Figure 10 has provided the overall architecture. The deep learning architecture is trained with the several (*Log*, *Noise*) pairs and also we have kept the noise model same for each users so that an uniformity can be brought and we can evaluate the system on a greater scale.

Feature Selection, denoted as +FS in Table 5 is used to see its effect with Noise. This also denotes that if the classifier considers selected webtype categories, then there is requirement to design noise focusing on that sector with some variations. Like for US male personalization, webtype categories like Japanese E-commerce is irrelevant, but us.fashionbunker.com (US e-commerce for female) can hide his gender while www.lavishalice.com (UK e-commerce for female) can hide his geographical identity. It has diverse behavior with high success rate for RF and RF+FS as it considers hierarchical feature consideration, but relatively low effect for others, but there is effect which establishes the fact that it can be used and more optimized version of dictionary is more effective than a single model. We have dealt with 16 categories but the scope of categories is vast and there will be difference in distribution for same group in US and any country in Asia. When it comes to noise, the aim is not to reverse the classification engine but to converge them to one and in computation, the noise are generated in that way. So if we have 20 samples of each class and we have an accuracy of 75% correctness, that means that 30 are classified to one and 10 to other. In that sense, the noise literally did a pretty good job.

6.8 Processing trade-off with noise

Noise for browsing is a parallel process for web site access produced to counter-conceal the credentials of the user through camouflage. When designing the optimization, we have simultaneously kept provision for efficiency of the noise so that it does not consume excessive power and processing cycles of the system. Power consumption and noise can seen as non-dominant as accuracy creates a maximization problem while power creates a minimization problem. So if the accuracy goes high, power requirement goes high and the problems gets down to a Pareto front where the aim is to select the amount of power the user is able to compromise to get itself camouflaged. For the analysis we have concentrated on accuracy as we seek intelligent optimization and the power rating calculation is hypothetical and linear and there can be techniques to handle it exclusively. Power and processing consumption is a non-linear phenomenon and there are techniques to handle them. However to make the calculation easier we have tried to sum up the power effect to linearly scale up to provide an overview of what it is like. Non-Dominant optimization or multi-objective optimization is important as the system goes complex and parameters are inversely related to each other. The most probable and effective way of dealing with this kind of optimization problem is through understanding the system and performing all scales of possible opportunities. The difference, this kind of analysis can create, is low power dissipation and effective solutions for the systems and also provide better security and conceal open credentials of the users.

7 Conclusions

Personal integrity can be hunted through browsing log of users. The more granular and intricately features are engineered, the more efficient will be the classification. Highly non-linear feature representations scale up for system. The purpose of this study is to feature out beneficial remedies

and prevention for people who are prone to different kind lucrative analogies and organization monitoring this can actually economically bankrupt others once they take control of the situation. We have introduced several types of features and have shown how individually they behave and also their collective strength and this clearly indicate how strong they can be once all of them are considered together and the trend shows that they are actually helping a stronger classifier. A perfect remedy of this kind of personalization attack is to mess up the log activities with some optimized intelligent noise which will be just enough to mix up the unique features for the different groups of classes and thus can prevent different autonomous models keep track and use as their prey. This is the first time, we have tried to answer questions related to personal security and the impact of browsing data on breaching personal identification. However, this is just the beginning and the analysis provides enough evidence of what can be done with it. This analysis can also be extended to data related to application usage and other digital behavioral data from both personal systems and handheld devices.

The more granular and intricately features are engineered, the more efficient will be the classification, achieving high accuracy of prediction. The feature representation is in thousands and with gradual increase in the scalability of the systems and also the algorithms, it won't be an issue. In fact, the feature vector must not be compressed as it will suppress important information and will not benefit in the long run, unless some of the features are identified as universal and are enough to represent the whole feature set. This is least likely to be the case for the time span of a day. These features may also characterize individuals or a class of individuals who are common in habitat, profession, behavior and liking. In some sense it is like having same majors, same hobbies and people having same interest. This will typically help in different business and commercial purposes, but at the same time can affect the personal integrity of users to external threats and unscrupulous activities. There are huge prospects of future works on this topic including finding correlation of browsing data and application usage data for identification of individuals, effects of changes in season or other factors including seasonal effects like sports season, Christmas and geographical influences etc. (Fig. 11)

References

(2017) Similarweb. https://www.similarweb.com/category

Al-Gburi A, Al-Hasnawi A, Lilien L (2018) Differentiating security from privacy in internet of things: a survey of selected threats and controls. In: Daimi K (ed) Computer and network security essentials. Springer, Cham, pp 153–172

Andersen A, Karlsen R (2018) Privacy preserving personalization in complex ecosystems. In: Linnhoff-Popien C, Schneider R, Zaddach M (eds) Digital marketplaces unleashed. Springer, Berlin, Heidelberg, pp 247–261

Anshari M, Almunawar MN, Lim SA, Al-mudimigh A (2018) Customer relationship management and big data enabled: personalization and customization of services. App Comput Inform. https://doi.org/10.1016/j.aci.2018.05.004

Atli BG, Miche Y, Kalliola A, Oliver I, Holtmanns S, Lendasse A (2018) Anomaly-based intrusion detection using extreme learning machine and aggregation of network traffic statistics in probability space. Cogn Comput 10:848–863. https://doi.org/10.1007/s12559-018-9564-y

Atote B, Zahoor S, Bedekar M, Panicker S (2018) Proposed use of information dispersal algorithm in user profiling. In: Mishra D, Nayak M, Joshi A (eds) Information and communication technology for sustainable development, vol 9. Springer, Singapore, pp 77–86

Azimi I, Rahmani AM, Liljeberg P, Tenhunen H (2017) Internet of things for remote elderly monitoring: a study from user-centered perspective. J Ambient Intell Human Comput 8(2):273–289

Baglioni M, Ferrara U, Romei A, Ruggieri S, Turini F (2003) Preprocessing and mining web log data for web personalization. In: Cappelli A, Turini F (eds) AI*IA 2003: advances in artificial intelligence. Lecture Notes in Computer Science, vol 2829. Springer, Berlin, Heidelberg, pp 237–249

Brar A, Kay J (2004) Privacy and security in ubiquitous personalized applications. School of Information Technologies. University of Sydney, Sydney

Castellano G, Fanelli AM, Torsello MA, Jain LC (2009) Innovations in web personalization. In: Castellano G, Jain LC, Fanelli AM (eds) Web personalization in intelligent environments, vol 229. Springer, Berlin, Heidelberg, pp 1–26

Chang CC, Chen PL, Chiu FR, Chen YK (2009) Application of neural networks and kanos method to content recommendation in web personalization. Expert Syst Appl 36(3):5310–5316

Chen HH (2018) Behavior2vec: generating distributed representations of users behaviors on products for recommender systems. ACM Trans Knowl Discov Data 12(4):43

Davidson D, Fredrikson M, Livshits B (2014) Morepriv: mobile OS support for application personalization and privacy. In: Proceedings of the 30th annual computer security applications conference. ACM, New York, pp 236–245

Duarte Torres S, Weber I, Hiemstra D (2014) Analysis of search and browsing behavior of young users on the web. ACM Trans Web 8(2):7

Egelman S, Peer E (2015) The myth of the average user: improving privacy and security systems through individualization. In: Proceedings of the 2015 new security paradigms workshop. ACM, New York, pp 16–28

Eirinaki M, Vazirgiannis M (2003) Web mining for web personalization. ACM Trans Internet Technol 3(1):1–27

Flesca S, Greco S, Masciari E, Saccà D (2018) A comprehensive guide through the italian database research over the last 25 years. Springer, New York

Freeman D, Jain S, Dürmuth M, Biggio B, Giacinto G (2016) Who are you? A statistical approach to measuring user authenticity. In: NDSS, pp 1–15

García-Dorado JL, Ramos J, Rodríguez M, Aracil J (2018) Dns weighted footprints for web browsing analytics. J Netw Comput Appl 111:35–48

Gulyás GG, Acs G, Castelluccia C (2016) Near-optimal fingerprinting with constraints. Proc Priv Enhanc Technol 2016(4):470–487

Jiang JY, Li CL, Yang CP, Su CT (2014) Poster: scanning-free personalized malware warning system by learning implicit feedback from detection logs. In: Proceedings of the 2014 ACM SIGSAC conference on computer and communications security. ACM, New York, pp 1436–1438

Karataş F, Korkmaz SA (2018) Big data: controlling fraud by using machine learning libraries on spark. Int J Appl Math Electron Comput 6(1):1–5

Kasanoff B (2002) Making it Personal: how to profit from personalization without invading privacy. Perseus Publishing, New York

Kobsa A (2007) Privacy-enhanced web personalization, the adaptive web: methods and strategies of web personalization. Springer, Berlin, Heidelberg, pp 628–670

Koh B, Raghunathan S, Nault BR (2015) Is voluntary profiling welfare enhancing? Management Information Systems Quarterly. p 52

Komiak SY, Benbasat I (2006) The effects of personalization and familiarity on trust and adoption of recommendation agents. MIS Q 30(4):941–960

Kosmides P, Demestichas K, Adamopoulou E, Remoundou C, Loumiotis I, Theologou M, Anagnostou M (2016) Providing recommendations on location-based social networks. J Ambient Intell Human Comput 7(4):567–578

Lebiednik B, Abadal S, Kwon H, Krishna T (2018) Spoofing prevention via rf power profiling in wireless network-on-chip. In: Proceedings of the 3rd international workshop on advanced interconnect solutions and technologies for emerging computing systems. ACM, New York, p 2

Leon P, Ur B, Shay R, Wang Y, Balebako R, Cranor L (2012) Why Johnny can't opt out: a usability evaluation of tools to limit online behavioral advertising. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM, New York, pp 589–598

Lin H, Yan Z, Chen Y, Zhang L (2018) A survey on network security-related data collection technologies. IEEE Access 6:18345–18365

Liu C, Park EM, Jiang F (2018) Examining effects of context-awareness on ambient intelligence of logistics service quality: user awareness compatibility as a moderator. J Ambient Intell Human Comput. https://doi.org/10.1007/s12652-018-1004-z

Logesh R, Subramaniyaswamy V, Vijayakumar V, Li X (2018) Efficient user profiling based intelligent travel recommender system for individual and group of users. Mob Netw Appl. https://doi.org/10.1007/s11036-018-1059-2

Malandrino D, Scarano V, Spinelli R (2013) How increased awareness can impact attitudes and behaviors toward online privacy protection. In: 2013 international conference on social computing (SocialCom). IEEE, pp 57–62

Maleki-Dizaji S, Siddiqi J, Soltan-Zadeh Y, Rahman F (2014) Adaptive information retrieval system via modelling user behaviour. J Ambient Intell Human Comput 5(1):105–110

Marella A, Pan C, Hu Z, Schaub F, Ur B, Cranor LF (2014) Assessing privacy awareness from browser plugins. In: Poster at the symposium on usable privacy and security (SOUPS)

Marforio C, Masti RJ, Soriente C, Kostiainen K, Capkun S (2015) Personalized security indicators to detect application phishing attacks in mobile platforms. arXiv:150206824 (**preprint**)

McAteer O (2016) This creepy new google feature lets you stalk your entire life's history. Elite Daily

Mcdaniel PD, Sen S, Spatscheck O, van der Merwe JE, Aiello W, Kalmanek CR (2006) Enterprise security: a community of interest based approach. NDSS 6:1–3

Mcdonald AM, Reeder RW, Kelley PG, Cranor LF (2009) A comparative study of online privacy policies and formats. In: Goldberg I, Atallah MJ (eds) Privacy enhancing technologies, vol 5672. Springer, Berlin, Heidelberg, pp 37–55

Meng W, Li W, Wang Y, Au MH (2018) Detecting insider attacks in medical cyber-physical networks based on behavioral profiling. Fut Gener Comput Syst. https://doi.org/10.1016/j.future.2018.06.007

Mobasher B (2007) Data mining for web personalization. In: Brusilovsky P, Kobsa A, Nejdl W (eds) The adaptive web, vol 4321. Springer, Berlin, Heidelberg, pp 90–135

Mobasher B, Dai H, Luo T, Nakagawa M (2002) Discovery and evaluation of aggregate usage profiles for web personalization. Data Min Knowl Discov 6(1):61–82

Mulvenna MD, Anand SS, Büchner AG (2000) Personalization on the net using web mining: introduction. Commun ACM 43(8):122–125

Nguyen TT, Armitage G (2008) A survey of techniques for internet traffic classification using machine learning. IEEE Commun Surv Tutor 10(4):56–76

Nicol J, Li C, Chen P, Feng T, Ramachandra H (2018) Odp: an infrastructure for on-demand service profiling. In: Proceedings of the 2018 ACM/SPEC international conference on performance engineering. ACM, New York, pp 139–144

Nogueira A, de Oliveira MR, Salvador P, Valadas R, Pacheco A (2005) Classification of internet users using discriminant analysis and neural networks. In: Next generation internet networks. IEEE, pp 341–348

Nowak J, Korytkowski M, Nowicki R, Scherer R, Siwocha A (2018) Random forests for profiling computer network users. In: Rutkowski L, Scherer R, Korytkowski M, Pedrycz W, Tadeusiewicz R, Zurada J (eds) Artificial intelligence and soft computing, vol 10842. Springer, Cham, pp 734–739

Olivarez-Giles N (2016) How to use google's new my activity privacy tool: search giant offers users a glimpse of the data it collects from web searches and other services. Wall Str J 1

Otebolaku AM, Andrade MT (2015) Context-aware media recommendations for smart devices. J Ambient Intell Human Comput 6(1):13–36

Park JH (2017) Resource recommender system based on psychological user type indicator. J Ambient Intell Human Comput. https://doi.org/10.1007/s12652-017-0583-4

Park S, Matic A, Garg K, Oliver N (2018) When simpler data does not imply less information: a study of user profiling scenarios with constrained view of mobile http (s) traffic. ACM Trans Web 12(2):9

Petrosyan D (2018) The dilemmas of surveillance profiling: the case of the united states. Fakulta sociálních věd. Univerzita Karlova, Prague

Purewal S (2016) Everything you need to know about google's my activity page: yes, google does know everything about you. CNET 10

Rafferty J, Nugent C, Liu J, Chen L (2016) An approach to provide dynamic, illustrative, video-based guidance within a goal-driven smart home. J Ambient Intell Human Comput. https://doi.org/10.1007/s12652-016-0421-0

Ren Y, Tomko M, Salim FD, Chan J, Sanderson M (2018) Understanding the predictability of user demographics from cyber-physical-social behaviours in indoor retail spaces. EPJ Data Sci 7(1):1

Rieck K, Trinius P, Willems C, Holz T (2011) Automatic analysis of malware behavior using machine learning. J Comput Secur 19(4):639–668

Riecken D (2000) Personalized views of personalization. Commun ACM 43(8):26–26

Sackmann S, Strüker J, Accorsi R (2006) Personalization in privacy-aware highly dynamic systems. Commun ACM 49(9):32–38

Salem B, Lino JA, Rauterberg M (2010) Smartex: a case study on user profiling and adaptation in exhibition booths. J Ambient Intell Human Comput 1(3):185–198

Sathe G (2016) Google's my activity page is a scary reminder that google knows everything about you. Gadgets360

Schaub F, Marella A, Kalvani P, Ur B, Pan C, Forney E, Cranor LF (2016) Watching them watching me: browser extensions impact on user privacy awareness and concern. In: NDSS workshop on usable security

Song Y, Salem MB, Hershkop S, Stolfo SJ (2013) System level user behavior biometrics using fisher features and Gaussian mixture models. In: Security and privacy workshops (SPW), 2013 IEEE. IEEE, pp 52–59

Stolfo SJ, Fan W, Prodromidis A, Chan PK, Lee W (2000) Cost-sensitive modeling for fraud and intrusion detection: results from the jam project. In: Proceedings of the 2000 DARPA information survivability conference and exposition. Citeseer

Sur C (2018) Ensemble one-vs-all learning technique with emphatic rehearsal training for phishing email classification using psychology. J Exp Theor Artif Intell. https://doi.org/10.1080/0952813X.2018.1467496

Su KW, Huang PH, Chen PH, Li YT (2016) The impact of formats and interactive modes on the effectiveness of mobile advertisements. J Ambient Intell Human Comput 7(6):817–827

Taylor DG, Davis DF, Jillapalli R (2009) Privacy concern and online personalization: the moderating effects of information control and compensation. Electron Commer Res 9(3):203–223

Wang K, Stolfo SJ (2004) Anomalous payload-based network intrusion detection. In: Jonsson E, Valdes A, Almgren M (eds) Recent advances in intrusion detection, vol 3224. Springer, Berlin, Heidelberg, pp 203–222

Wang T, Goldberg I (2016) On realistically attacking tor with website fingerprinting. Proc Priv Enhanc Technol 2016(4):21–36

Yang C, Zhang C, Chen X, Ye J, Han J (2018) Did you enjoy the ride: understanding passenger experience via heterogeneous network embedding. ICDE IEEE

Yang J, Qiao Y, Zhang X, He H, Liu F, Cheng G (2015) Characterizing user behavior in mobile internet. IEEE Trans Emerg Top Comput 3(1):95–106