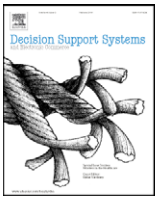
Decision Support Systems 48 (2010) 548–558

Contents lists available at ScienceDirect

Decision Support Systems

journal homepage: www.elsevier.com/locate/dss



Toward user patterns for online security: Observation time and online user identification

Yinghui (Catherine) Yang [a,*,1], Balaji Padmanabhan [b]

[a] Graduate School of Management, University of California, Davis, AOB IV, One Shields Ave., Davis, CA 95616, USA
[b] College of Business, University of South Florida, 4202 East Fowler Ave., Tampa, FL 33620, USA

ARTICLE INFO

Article history:
Received 24 December 2008
Received in revised form 21 October 2009
Accepted 6 November 2009
Available online 14 November 2009

Keywords:
Web usage mining
Behavioral signatures
Online security
User identification
Biometrics
Electronic commerce

ABSTRACT

Research in biometrics suggests that the time period a specific trait is monitored over (i.e. observing speech or handwriting "long enough") is useful for identification. Focusing on this aspect, this paper presents a data mining analysis of the effect of observation time period on user identification based on online user behavior. We show that online identification accuracies improve with pooling user data over sessions and present results that quantify the number of sessions needed to identify users at desired accuracy thresholds. We discuss potential applications of this for verification of online user identity, particularly as part of multi-factor authentication methods.

1. Motivation

Humans are believed to have many unique characteristics such as fingerprints and handwriting styles. We use the term "signatures" here to refer to distinguishing characteristics that are behavioral (e.g. writing styles), as opposed to characteristics that are physiological (e.g. fingerprints). The applications of methods for unique identification are significant, ranging from forensics and law enforcement to novel biometrics-based access to personal information that protects user privacy and mitigates fraud. The development and perfection of such unique distinguishing characteristics continues to be an important area of research.

Given the vast impact technology has in everyday life, there has naturally been interest in recent years on whether there might be unique signatures in technology mediated applications. Ref. [24] shows that users have distinct ways in which they use computer keyboards and that users have unique keystroke dynamics. Ref. [7] extends this work to the use of mouse movements in addition to keystroke dynamics and note that the combination can often be used to uniquely identify humans. Ref. [20] shows that authors have unique writing styles that enable identifying them from text. In a similar vein, Ref. [19] shows that users have unique writing patterns when they author content for online message boards. A recent article [10] published in *Nature* that studied user mobility patterns with cell phone GPS data showed that, perhaps not surprisingly, most users tend to have quite predictable daily mobility patterns.

In the same spirit, are there unique "clickprints" based on how users browse, or consume content online? This is an open question, the answer to which can have significant implications for applications such as online fraud detection and product recommendations.

If individuals can be identified based on online patterns, even to a reasonable extent, then there are important server-side and client-side applications. As an example of a server-side application, if a firm can identify a user who is not explicitly signed in there may be opportunities for targeted recommendations. Of course in this case the firm will have to consider online privacy needs of their customers and the firm's own privacy policy before any such action.

On the client-side there may be important practical applications of such technology that may mitigate online fraud and identity theft, issues that are known to be important to consumers [15]. For instance, users may opt-in to download client-side software from a trusted third party[2] that will track client-side activities to build user identification models. Such models may be used to provide behavioral authentication services on behalf of the user. For instance, when this user makes a large online brokerage transaction, the financial institution may, in real time, query the client-side software for a "user score". If the returned score suggests that the user is unlikely to be who they claim to be, the firm may then proceed to seek additional

* Corresponding author. Tel.: +1 530 754 5967.
E-mail addresses: yiyang@ucdavis.edu (Y.(C.) Yang), bpadmana@coba.usf.edu (B. Padmanabhan).
[1] Tel.: +1 813 974 6763.

[2] e.g. a firm such as Verisign that is known to provide certification and authentication services.

0167-9236/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.dss.2009.11.005

information. Such an application may offer users real benefits such as fraud and online identity theft mitigation, while being sensitive to privacy concerns due to its opt-in nature and limited data (a user score) that it reveals, with consent, to third parties. In this research we do take such a "user-centric" perspective — the data we analyze is user-centric browsing data and the results in this research are relevant to client-side applications such as the one noted here.

Related to the client-side security application, a key US federal agency, the Federal Financial Institutions Examination Council (FFIEC) recently issued guidance entitled *Authentication in an Internet Banking Environment*[3]. This document notes that: "existing authentication methodologies involve three basic "factors":

- something the user *knows* (e.g., password, PIN);
- something the user *has* (e.g., ATM card, smart card); and
- something the user *is* (e.g., biometric characteristic such as a fingerprint)."

The guidance notes that fraud and identity theft are often the result of exploiting single factor authentication systems and suggests that multi-factor authentication methods are stronger fraud deterrents. Indeed deterrence as a mechanism to improve IT security has been stressed in the IS literature. It is known that just the use of security software by firms can deter computer abuse from a network intrusion perspective [28]. While the accuracy of such systems matters, it is often the deterrence that comes from accuracy that actually contributes to better security [4]. The hypothetical client-side application discussed previously, if designed appropriately, may provide one such additional factor in a multi-factor approach to fraud deterrence. Designing such a system will require developing accurate user identification models. This in turn requires a deeper understanding of the factors that can result in better or worse identification accuracies. Our research in this paper focuses on one such factor as we note below in Section 2.

It should also be noted that there are efforts on the part of Internet service providers to improve security for all users. In this context recent research [35] has proposed certification mechanisms as a manner in which incentive alignment can be achieved. Indeed such efforts are complementary to better client-side approaches for security. Further, user and computer security within organizational settings has naturally attracted specific attention in research. Ref. [12] takes such a broader view of security in organizations and note that organizations should study employee "security behavior" in detail and discusses organizational mechanisms for this purpose. Related is the work of Ref. [11], where a genetic algorithm is used to determine an organization's optimal security profile to balance cost as well as risk.

2. Focus of this research

There are online user behavior theories, most notably the research in Web usage mining [1,26,32] which suggest that user behavior is not random and there is often a purpose that translates into revealed online behavior. However they do not provide specific answers on how unique the revealed behavior is. On the other hand there has been substantial work in biometrics over the last few decades that has specifically studied user identification from various characteristics such as fingerprints, handwriting or speech. Much of the biometrics work focusing on user identification has been experimental, and has collectively highlighted two intuitive aspects:

1. The quality of user data impacts identification accuracy. In the handwriting and fingerprinting literature, quality refers to image quality as measured by the resolution or number of pixels. There is evidence in this literature [16] that higher quality improves identification accuracies. For user identification from online behavior, quality of data can be measured by the features created from behavior. A large number of features can be generated from every click or page viewed online. The associated intuitive hypothesis here will be that better features result in better user identification models.
2. The quantity of user data impacts identification accuracy. In the speech recognition and handwriting literatures, quantity refers to how long this behavior is observed or monitored. The results show that if speech and handwriting can be observed "long enough", fairly accurate models can be built [14]. For user identification from online behavior, quantity is a measure of how much user data is observed. Intuitively we expect this result to hold as well. For instance from just one session of browsing data a user may not look sufficiently different from others, but over time there might be enough data to highlight differences.

In this paper we focus on studying the quantity aspect in online user identification. By doing so we hope this research can provide light on how much user data is needed before accurate user identification models can be obtained.[4] Insights into this issue are significant for client-side applications such as the one noted above. To our knowledge this is the first research paper that addresses this question. Further the methodology developed here can be used to study this aspect for any given quality of features used for identification.

For convenience we will use the term "aggregation" to describe the process of observing and collecting data over longer time periods. While time is a useful notion to intuitively measure "how much" data is needed, we will focus our analysis in this research to aggregation over multiple Web sessions as opposed to time periods. We note that the analysis methodology presented here can be directly applied to time if desired. However a user session is a commonly used unit of analysis when describing online behavior, and "long enough" in this context refers to observing a user over an adequate number of sessions.

In this paper we answer the following questions:

(Q1a). *Does aggregation result in improved user identification based on online behavior?*

While prior research in biometrics has shown the value of aggregation for other problems, to our knowledge this work is the first to present such analysis for online user identification.

Rather than treating this as a single hypothesis to test, we break this down into a series of tests as described below. Intuitively we expect that as the number of users in the population grows, user identification will be more difficult. Hence the significance of aggregation can be expected to depend on the number of users considered in a dataset. Further we restrict our consideration to a range of different aggregation levels and test if accuracies at a specific level of aggregation are lower than the accuracies at the immediate higher level of aggregation. Hence we answer Q1a in a table where the rows correspond to varying number of users (specifically $M=2$, 3, 5, 10, 25, 50 and 100) and the columns represent pairs of adjacent aggregations (2 over 1, 3 over 2,…, 10 over 9). Each cell represents a hypothesis that the accuracies corresponding to the higher level of aggregation are greater than the accuracies corresponding to the lower aggregation.

(Q1b). *What are the accuracy gains from aggregation for online user identification?*

Testing specific hypotheses related to Q1a will answer whether aggregation is useful. The empirical analyses needed for this will also

[3] http://www.ffiec.gov/pdf/authentication_guidance.pdf.

[4] Our approach in this paper differs from the basic PAC learning [25] formalism that has been used in machine learning to relate the amount of training data to prediction errors in a probabilistic framework. Unlike in traditional PAC learning, our approach generates a different set of features across varying time periods of aggregation.