Privacy concerns of implicit secondary factors for web authentication

Joseph Bonneau
Princeton University
jbonneau@princeton.edu

Edward W. Felten
Princeton University
felten@cs.princeton.edu

Prateek Mittal
Princeton University
pmittal@princeton.edu

Arvind Narayanan
Princeton University
arvind@cs.princeton.edu

1. INTRODUCTION

Passwords remain the dominant mechanism for web authentication despite their well-known weaknesses. Economics makes them very difficult to replace [2], but they can be strengthened by analyzing a number of *implicit* secondary factors [5, 6]. Potential implicit factors include:

- Keystroke dynamics [7]
- Touchscreen dynamics [4]
- IP address and geolocation [1]
- Time of login
- Browser information such as user-agent and fingerprint

These implicit factors can help transform authentication from a binary decision problem (based on passwords alone) into a classification problem with a spectrum of possible decisions. For example, unusual values for implicit factors can be used as an indicator to detect merely suspicious logins for which additional explicit authentication actions (such as sending an SMS code) can be taken. Alternately, known values for implicit factors can be used as an indicator that it is safe to relax normal rate-limiting constraints and avoid frustrating users by locking them out due to typos (or worse, requiring password resets).

We highlight three distinct privacy issues in the next three sections. The first is well known from biometrics, whereas the second two appear specific to some web-based implicit factors. We observe that most of the published work on implicit factors has paid little or no attention to these issues.

We are limiting our focus to web-based authentication. Implicit factors are also commonly mentioned for use in mobile devices, typically touchscreen-based smartphones and tablets, but privacy concerns are fundamentally different as data can be stored on-device and authentication implemented at the OS level. However, some proposals involve mobile devices interacting with remote data-aggregating authentication services [3], in which cases these same issues may exist.

2. PERMANENCE AND SIMULATABILITY

While often not as completely permanent as biometrics, many implicit secondary factors change rarely or very slowly. Because most can be simulated once known, security may be significantly diminished once a signal is leaked. The degree of security loss depends on two factors:

1. **Permanence:** We have very little long-term data on how most implicit secondary factors change over time as most studies are conducted in a relatively short time frame. Some factors like “behavioral biometrics” (e.g. typing dynamics) may be effectively permanent. Other factors like IP address may change frequently for some users or be very static in other cases.
2. **Simulatability:** Some signals are relatively easy to simulate in software once known, such as keystroke dynamics or login times. Most of these signals appear difficult to simulate by human users, though this doesn’t appear to have been tested empirically. Other signals like IP address are more challenging to simulate.

If implicit signals are permanent and simulatable, this means we would like only “important” authentication services to have access to them. Another option is to hope that signals are sufficiently impermanent that only important service are able to keep tabs on a user’s most recent value. Interestingly, this means systems might be *more* secure if the signal changes frequently, even though the performance of the authenticator may decrease.

3. INHERENT SENSITIVITY

Some implicit signals are sensitive data in their own right. In particular, a user’s detailed pattern of login activity, including geolocation and time of logins, can reveal significant information. Other “history-based” authentication schemes, such as those which ask about recent email content or recent purchases, are similarly sensitive. Typically this is addressed by only proposing these systems in cases where the authentication service is already collecting this data in the regular course of business, but this assumption may limit high security to only these services.

Who are you?! Adventures in Authentication (WAY) 2014, July 9, 2014, Menlo Park, CA.

If this data proves useful for authentication, it might provide an incentive for more services to collect sensitive personal data, to keep it around for a longer period of time, or to make it available to more teams internally. This can potentially be addressed by privacy-preserving machine learning techniques but these techniques don't yet appear common in practice.

A particularly interesting example exists for federated authentication in the form of a user's exact pattern of logins at relying sites. An identity provider can potentially use this to build a profile of the user and detect requests to log in to unusual sites (or usual sites at unusual times). This suggests that federated authentication protocols without strong privacy, such as OpenID [9], might enable identity providers to offer better security than competing protocols such as Persona [8] which hide this data from the identity provider.

4. LEGITIMATE SECONDARY USES

The sensors used to collect many potential implicit factors (touchscreen dynamics, mouse and keystroke dynamics, some aspects of fingerprinting) have legitimate secondary uses which is why they are available in the web platform in the first place. This implies that every website a user visits has the ability to sample them. This issue is largely distinct from biometrics, where few pieces of software need to know a user's fingerprint or iris pattern for any reason except authentication.

One solution is for the browser to only give a low-fidelity version of the raw data to some non-trusted sites, or to only allow the high-fidelity version of the signal to be used in some privacy-preserving authentication protocol. This seems very unlikely due to the difficulty of changing the web platform and agreeing on a standard. If a change of this magnitude were practical, then we could probably roll out a better technology than passwords.

Another solution is that sites that are used more often gain a more accurate picture of the user's signal because they interact with them far more often. This is plausible, but means that strong authentication with implicit signals is only available for a user's most visited sites. Some sites are important but rarely visited (banks) while others are often visited but perhaps less trusted (news websites).

This also raises the question of how efficiently a malicious website may try to more quickly extract the signal, for example by a typing game which quickly builds a profile of a user's typing that would otherwise take a long time to build up naturally. An implication is that research on these factors should consider how the ROC curves change based on an attacker who can interact with the user for a limited amount of time (less than the genuine prover).

5. IMPLICATIONS

All three of these issues point to an advantage for large web services. They already have fine-grained user data, mitigating the concern about collecting and using it for authentication purposes, and they interact with users frequently enough that they can learn a high-fidelity, up-to-date model of the user's behavior which can potentially recover from leaks or attackers gaining partial information. Thus, multidimensional authentication based on implicit factors may be a driver of centralization in authentication which suggests that the strongest authentication will be possible at large web services and not special-purpose authentication services.

Research on secondary factors for web authentication should keep these issues in mind. In particular, it would be helpful to study how implicit factors change over time, how quickly a signal can be extracted by a malicious adversary with access to the same platform as the legitimate authenticator (e.g. a malicious web site using the same browser), and to what extent potentially sensitive data can be stored in a sanitized form for use in authentication.

6. REFERENCES

[1] N. Akhtar and F. ul Haq. Real time online banking fraud detection using location information. In *Computational Intelligence and Information Technology*, pages 770–772. Springer, 2011.

[2] J. Bonneau, C. Herley, P. C. van Oorschot, and F. Stajano. The Quest to Replace Passwords: A Framework for Comparative Evaluation of Web Authentication Schemes. In *2012 IEEE Symposium on Security and Privacy*, May 2012.

[3] R. Chow, M. Jakobsson, R. Masuoka, J. Molina, Y. Niu, E. Shi, and Z. Song. Authentication in the clouds: a framework and its application to mobile users. In *Proceedings of the 2010 ACM Workshop on Cloud Computing Security*. ACM, 2010.

[4] A. De Luca, A. Hang, F. Brudy, C. Lindner, and H. Hussmann. Touch me once and i know it's you!: implicit authentication based on touch screen patterns. In *Proceedings of the 2012 ACM Annual Conference on Human Factors in Computing Systems (CHI)*, pages 987–996. ACM, 2012.

[5] R. Greenstadt and J. Beal. Cognitive security for personal devices. In *Proceedings of the 1st ACM Workshop on AISec*, pages 27–30. ACM, 2008.

[6] M. Jakobsson, E. Shi, P. Golle, and R. Chow. Implicit authentication for mobile devices. In *Proceedings of the 4th Usenix Workshop on Hot Topics in Security (HotSec)*.

[7] F. Monrose, M. K. Reiter, and S. Wetzel. Password hardening based on keystroke dynamics. In *Proceedings of the 6th ACM Conference on Computer and Communications Security*, CCS '99, pages 73–82, New York, NY, USA, 1999. ACM.

[8] Mozilla Foundation. Mozilla Persona, 2014. `http://www.mozilla.org/en-US/persona/`.

[9] D. Recordon and D. Reed. OpenID 2.0: a platform for user-centric identity management. In *DIM '06: Proceedings of the 2^{nd} ACM Workshop on Digital Identity Management*, pages 11–16, New York, NY, USA, 2006. ACM.