Cloudsweeper and Data-Centric Security

Peter Snyder and Chris Kanich
University of Illinois at Chicago
Chicago, Illinois, USA
{psnyde2,ckanich}@uic.edu

ABSTRACT

Most security online is binary, where being authorized to access a system allows complete access to the requested resource. This binary system amplifies the harm of giving access to an unauthorized individual and motivates system designers to strengthen access control mechanisms to the point where they become so strong as to be nearly insurmountable for illegitimate and legitimate users alike.

As a result, Internet users are required to jump through several hoops to access their data: ever longer passwords, multiple authentication factors, or time consuming CAPTCHAs. Users must always provide strong proof of their identity, regardless of whether they want to check their email for something as innocuous as a movie time or as serious as a medical test result. Not surprisingly, users often disable or refuse to use these tedious security options [2, 5, 7].

Users may be better served by a *data-centric* approach to security, where systems are sensitive to the differing security needs of data, even within a single account or collection. A data-centric approach can apply strong security only when the data being protected warrants it, while allowing users a less encumbered experience the majority of the time. Machine learning techniques can automate the detection of sensitive information, freeing users from the tedious task of sorting their data into low and high security categories. With less friction involved in securing their data, users may be more likely to use strong security where available, resulting in a more secure Internet for everyone.

We present *Cloudsweeper*, a tool that applies a data-centric approach to security to the specific case of plain text password sharing in Gmail accounts. Cloudsweeper detects and applies an additional layer of encryption to plain text passwords in a user's email account, while allowing the user to access the rest of their email archive as normal. Public use of Cloudsweeper shows that such a data-centric approach to securing data can be an effective way of providing users more security while still being acceptably convenient.

Categories and Subject Descriptors

H.3.2 [**Information Systems Applications**]: Information Storage; D.4.6 [**Operating Systems**]: Security and Protection; C.2.0 [**Computer-Communication Networks**]: General

Keywords

Cloud storage, email

1. SECURITY AT INTERNET SCALE

That the Internet is a dangerous place is unlikely to surprise anyone reading this article. Attacks, automated or manual, targeted or general, short-lived or persistent, are a constant occurrence online. Providing both usable and secure methods of accessing data in these online systems is an ongoing challenge.

Automated systems have a mixed ability to defend the average Internet user against attacks. Some attackers use highly automated systems to target as many people as possible. Attackers using this strategy expect to average a low return per attack, but hope to succeed against enough targets to make the effort profitable. Examples of such attacks include spam-based marketing and infecting computers for use in monetized botnets.

Because these attacks are high volume and automated, effective defenses can be created by generalizing from numerous examples. Automated systems like spam filters and intrusion detection systems can do a good job of mitigating these attacks transparently for users [4].

However, automated systems are much less useful when attackers follow a different strategy, one of spending a large amount of effort to extract maximum value from a small number of targets. One only needs to consider the regularity of successful "spear-phishing" campaigns or targeted point-of-sale attacks to see that current security approaches cannot effectively defend against these attacks. When only a few or zero examples of an attack exist before the damage is done, automated systems are much less likely to keep users safe.

Current best practice against these targeted attacks is for users to rely less on automated systems and to instead take direct responsibility for their data's security. And while solutions like multi-factor authentication and end-to-end

encryption promise security, the relatively low uptake of these technologies among Internet users suggests that these techniques require a level of direct management, tediousness, or technical knowledge that are not currently feasible for mainstream adoption.

2. DATA-CENTRIC APPROACH

An effective approach should combine the insights and convenience of automated defense systems with the security guarantees provided by heavyweight technologies like two-factor-authentication and encryption. One option is to focus defense on the data being requested just as much as the access request itself. Strong security techniques could then be applied only where most useful, so as to only bother users "when it really matters." While in the worst case, security would be decreased by this approach, effective security could increase by higher usage of the available tools.

In other words, Internet users might be more likely to use defenses that are tailored to the security needs of their data instead of defenses that treated all data in a like-mannered fashion. Instead of being faced with a choice between lax security and strong security, a data-centric security approach could intelligently trade security for convenience depending on the sensitivity of the resource being accessed.

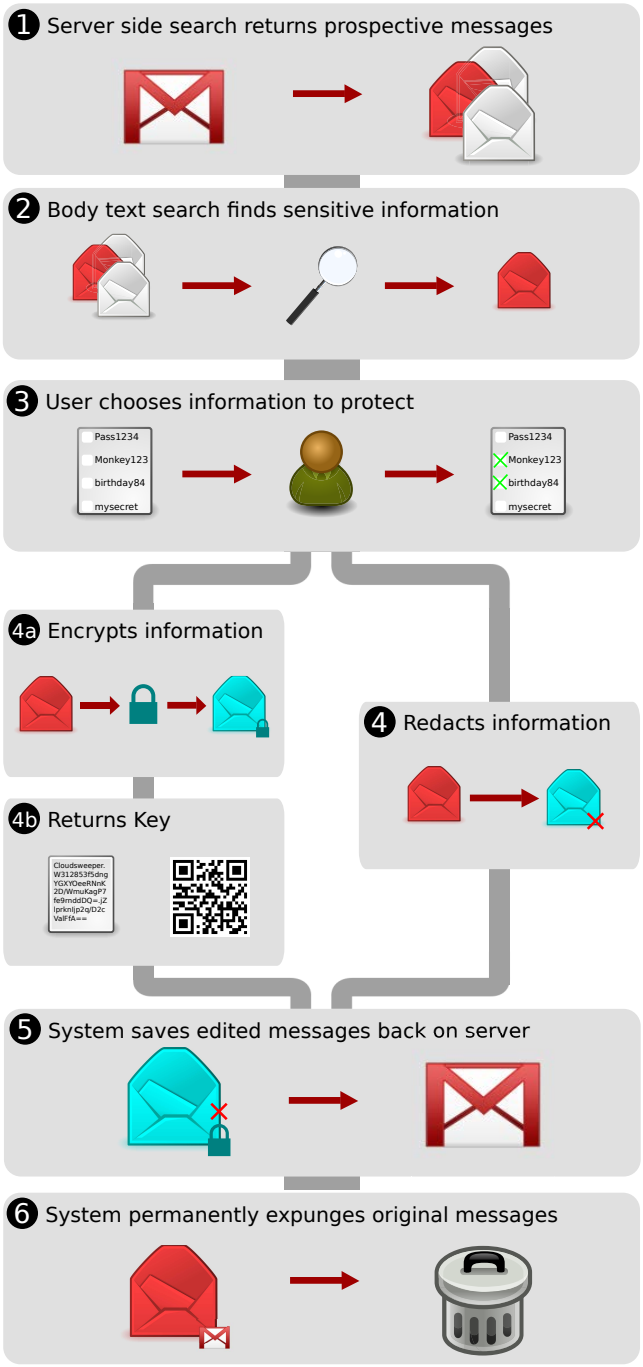
3. CLOUDSWEEPER

Cloudsweeper is a system we built to apply this data-centric security approach to the specific case of plain text password sharing in Gmail email accounts. The system is publicly available at *https://cloudsweeper.cs.uic.edu/* as a browser-based application. Over two thousand individuals have used Cloudsweeper, and the tool has encrypted or redacted forty thousand passwords in over one half million email messages. The tool is instrumented to take anonymous measurements of password sharing in email accounts, so that we can improve the data centric security mechanisms of the tool.

The risks of sending or storing plain text passwords in email are obvious; anyone who gains access to the email account also gains access to the resources the plain text passwords secure. It is also a surprisingly common practice, and one that grows in importance as the size of cloud based email accounts increases. [6] Large email archives containing years or decades of email are attractive targets for attackers looking to gain access to lots of sensitive information about an individual.

Current approaches to managing these risks fall in two general categories. One approach is to not trust the email provider with any secret material through the use of end-to-end encryption tools like PGP. This "maximal storage security" solution prevents attackers and eavesdroppers with access to the account from viewing any secrets [3], but comes with the significant downsides of losing the ability to outsource search or spam filtering to cloud systems, and making it difficult to access email, sensitive or otherwise, across multiple devices.

Another approach to protecting secrets shared in email is to make accessing the email account more difficult. Two factor authentication is a common example of this strategy. The security these systems provide come with their own downsides, such as user inconvenience and difficulties with automated or legacy login systems. These "maximal perimeter security" strategies also provide no harm mitigation once an attacker has gained access to the online account.



Cloudsweeper workflow. Only step 3 requires user interaction.

Cloudsweeper combines the benefits of both of these approaches with a data-centric security approach to mitigate the problem of plain text password storage in email. First, Cloudsweeper searches through the messages in a user's email account for pieces of text that may be passwords. Cloudsweeper then gives the account holder the option to redact or encrypt each password, while leaving the rest of

the message unchanged. Encrypted passwords can be later recovered through the use of a returned key, provided either as a QR-code or as a text string. These returned keys are the only way to decrypt the secured passwords in the account, providing the account holder the security that a compromise of her email will not also give an attacker access to further sensitive accounts.

Cloudsweeper provides this security benefit without requiring the user to manually examine the gigabytes of data in their email accounts, searching for the sensitive documents in a mostly non-sensitive haystack. Similarly, users are not faced with the inconvenience of needing to go through additional authentication or decrypting steps to find non-sensitive information in their email. We believe that this application of security is more in line with recent research in risk analytics, and has a good chance of being implemented by users who may not normally be sensitive to security concerns [1]. Furthermore, users do not lose the ability to access the now-secured sensitive data in their account unless they choose to permanently redact the information.

4. CONCLUSION

Cloudsweeper is an example implementation of the data-centric security approach described in this article. The same approach could be employed anywhere the security needs of data differ within a storage-based service, such as version control repositories or collaborative online documents. Systems that automatically deploy security technologies in proportion to the sensitivity of the data being requested at a level more granular than a binary yes/no can provide a high level of effective security to Internet users without requiring burdensome and unacceptable steps from the user.

5. ACKNOWLEDGMENTS

We would like to thank Brian Krebs for his feedback and publicity of the Cloudsweeper system. This work was supported in part by the National Science Foundation under grant DGE-1069311.

References

[1] Jim Blythe, Jean Camp, and Vaibhav Garg. 2011. Targeted risk communication for computer security. In *Proceedings of the 16th international conference on Intelligent user interfaces*. ACM, 295–298.

[2] Philip G Inglesant and M Angela Sasse. 2010. The true cost of unusable password policies: password use in the wild. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 383–392.

[3] I. Ion, N. Sachdeva, P. Kumaraguru, and S. Čapkun. 2011. Home is safer than the cloud!: privacy concerns for consumer cloud storage. In *Proceedings of the Seventh Symposium on Usable Privacy and Security*.

[4] Andreas Pitsillidis, Kirill Levchenko, Christian Kreibich, Chris Kanich, Geoffrey M. Voelker, Vern Paxson, Nicholas Weaver, and Stefan Savage. 2010. Botnet Judo: Fighting Spam with Itself. In *Proceedings of the Network and Diestributed System Security Symposium (NDSS)*.

[5] S. Schechter, A.J.B. Brush, and S. Egelman. 2009. It's no secret. Measuring the security and reliability of authentication via "secret" questions. In *Proceedings of the 2009 IEEE Symposium on Security and Privacy*.

[6] Steve Whittaker, Victoria Bellotti, and Jacek Gwizdka. 2006. Email in personal information management. *Commun. ACM* 49, 1 (2006).

[7] Alma Whitten and J Doug Tygar. 1999. Why Johnny Can't Encrypt: A Usability Evaluation of PGP 5.0.. In *Usenix Security*, Vol. 1999.