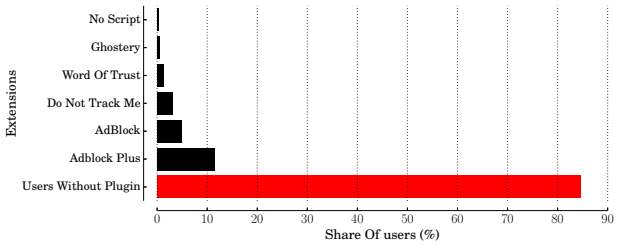

Figure 4: Shares of users adopting "popular" privacy-preserving extensions.

traffic, they have no visibility when encryption is enabled. Man-in-the-Middle solutions are also available, but appear to be very intrusive (and cumbersome).

We check the popularity of some extensions among Internet users by leveraging the ISP dataset. Fig. 4 show the results. 85% of households do not have any of these plugins installed on any device they use. Only AdBlock Plus seems to be quite popular, being installed in 11% of households. However, by checking the user-agents in the requests, we observe it is installed mostly on Personal Computers, leaving tablets and smartphones unprotected even in those households. This observation demonstrates, first, that the vast majority of users are not conscious of their information being exposed when surfing web, and, second, that those who actually care about this problem are not aware of means to supervise their surfing when using mobile terminals.

In summary, we lack a comprehensive solution capable of offering Internauts visibility and control of the information they send over the Internet, independent on device/OS they use. CROWDSURF's challenge is in offering a unified, not-optional, cross-device and cross-platform system.

6. DISCUSSION

CROWDSURF design presents some practical challenges that must be faced, and ingenuity must be used to find appropriate solutions. The research community as a whole is called to design efficient algorithms, and propose scalable implementations. CROWDSURF offers this possibility, allowing anyone to contribute.

i) Protecting CROWDSURF users' privacy: For CROWDSURF is vital to maintain a good degree of visibility on the traffic sample users decide to share. However, it is fundamental to guarantee that *all* users' sensitive information is filtered out from the traffic sample before this is shared. To overcome limitations of the current CROWDSURF's anonymization mechanisms, we are investigating a methodology based on differential privacy [8].

ii) Protection from malicious biases: CROWDSURF's suggestion system might be polluted by malicious services to increase their reputation or gain popularity among users. Clearly, mechanisms to to prevent this kind of attacks must be adopted. We leave this for our future work.

iii) Usability: A key factor to let CROWDSURF reach a wide population of users is to combine CROWDSURF features with a simple, easy to use, yet effective interface to quickly inspect contacted services, customize rules, generate suggestions, etc.

iv) Web industry reaction: We believe that CROWDSURF, if widely deployed, would not incite services to contrast or boycott it. Instead, we believe CROWDSURF would lead to a better, more transparent web, where providers would be induced to improve the reliability of their services. In turn, users would reward trustful providers with their loyalty. Moreover, CROWDSURF has the potential of creating room for new services and opportunities such as, e.g., a new "market of suggestions".

iv) Complexity: We are aware that CROWDSURF introduces a considerable amount of new "mechanisms" to guide users during their online activity. However, we believe that this represents a little technological cost in comparison to the large flexibility and advantages it would introduce.

7. CONCLUSION

This paper presented CROWDSURF, a crowdsourced holistic system which allows users and companies to supervise the information exchanged in the web.

We have shown that CROWDSURF is feasible. We presented real data to support how we can build a crowdsourced knowledge supported by automatic algorithms. As a proof of concept, we implemented CROWDSURF as a Firefox plugin, whose benefits can come at a marginal performance cost for the user. Second, we provided an example of algorithm for the automatic generation of suggestions for the users.

We are aware that our idea is ambitious, as, first, CROWDSURF shall pass through a long and difficult design and standardization process to get accepted as a compelling technology by the industry. It shall undergo a deep engineering effort to convince users about its reliability, usability and effectiveness. However, as the research community is becoming more and more conscious that data constitutes a vital asset in modern Web, we are confident that the unified solution offered by CROWDSURF represents a good starting point to protect (and possibly endorse) such asset.

8. REFERENCES

[1] D. Naylor, A. Finamore, I. Leontiadis, Y. Grunenberger, M. Mellia, K. Papagiannaki, and P. Steenkiste, "The Cost of the "S" in HTTPS," in *ACM CoNEXT*, 2014.

[2] G. Acar, C. Eubank, S. Englehardt, M. Juarez, A. Narayanan, and C. Diaz, "The Web Never Forgets: Persistent Tracking Mechanisms in the Wild," in *ACM SIGSAC*, 2014.

[3] B. Krishnamurthy, K. Naryshkin, and C. E. Wills, "Privacy leakage vs. Protection measures: the growing disconnect," in *W2SP*, 2011.

[4] T.-F. Yen, Y. Xie, F. Yu, R. P. Yu, and M. Abadi, "Host Fingerprinting and Tracking on the Web: Privacy and Security Implications," in *NDSS*, 2012.

[5] H. Metwalley, S. Traverso, M. Mellia, S. Miskovic, and M. Baldi, "The online tracking horde: a view from passive measurements," in *TMA*, 2015.

[6] L. Popa, A. Ghodsi, and I. Stoica, "HTTP As the Narrow Waist of the Future Internet," in *ACM HotNets*, 2010.

[7] C. Riederer, V. Erramilli, A. Chaintreau, B. Krishnamurthy, and P. Rodriguez, "For Sale : Your Data: By : You," in *ACM HotNets*, 2011.

[8] B. C. M. Fung, K. Wang, R. Chen, and P. S. Yu, "Privacy-preserving data publishing: A survey of recent developments," *ACM Comput. Surv.*, vol. 42, pp. 14:1–14:53, June 2010.

[9] H. Metwalley, S. Traverso, M. Mellia, S. Miskovic, and M. Baldi, "CrowdSurf: Empowering Informed Choices in the Web," *ArXiv e-prints*, Feb. 2015.

[10] Hackers steal vast eBay user database, including passwords, `http://www.bdlive.co.za/world/americas/2014/05/23/hackers-steal-vast-ebay-user-database-including-passwords`.

[11] H. Metwalley, S. Traverso, and M. Marco, "Unsupervised detection of web trackers," in *IEEE GLOBECOM*, 2015.

[12] D. J. Newman, "The double dixie cup problem," *American Mathematical Monthly*, 1960.

[13] S. Zimmeck and S. M. Bellovin, "Privee: an architecture for automatically analyzing web privacy policies," in *Proceedings of the 23rd USENIX conference on Security Symposium*, pp. 1–16, USENIX Association, 2014.

[14] Google, Microsoft, and Amazon are paying to get around Adblock Plus, `http://www.theverge.com/2015/2/2/7963577/google-ads-get-through-adblock`.