Poster: Evaluating the effectiveness of privacy tools using Information Flow Experiments

Amit Datta*, Anupam Datta*, Lay Kuan Loh*, Michael Carl Tschantz† and Zheng Zong*
*Carnegie Mellon University, †International Computer Science Institute
{amitdatta, danpuam}@cmu.edu, {lloh, zzong1}@andrew.cmu.edu, mct@icsi.berkeley.edu

I. Introduction

With the growing prevalence of personalized content and an advertising based Internet economy, users are being increasingly tracked across website visits with the purpose of creating accurate interest and demographic profiles. These profiles are then used to serve targeted content and advertisements. Online data aggregrators are always developing new tracking mechanisms to obtain more fine-grained information and such aggressive tracking had led to numerous privacy concerns. To help protect user privacy, privacy champions keep building privacy preserving tools to thwart trackers. This arms race between tracking tools and privacy tools is exemplified by the steady stream of studies detecting new forms of tracking and proposing defenses [1]–[3].

Privacy preserving tools like DoNotTrack, opt-out cookies, Ghostery, Privacy Badger, AdBlockPlus, Tor, etc., all aim to improve user privacy by blocking tracking. Some of these tools have been studied in the past to check for their effectiveness against online behavioral advertising [4] and HTTP requests and cookies set [5]. However, these studies lack statistical guarantees and have no means to discover effects of tracking other than those stipulated in advance. Moreover, many new tools and tracking mechanisms (like fingerprinting) have become prevalent since these studies were published.

In this poster, we measure the effectiveness of some of the most popular privacy preserving tools in the face of newer tracking mechanisms. To include statistical rigor and not be restricted to discovering only pre-decided effects, we use the methodology used in AdFisher [6]. These methods are an application of the general information flow experiment (IFE) methods [7] to study the Google Ad System. We first perform an empirical completeness evaluation of the methods used in AdFisher and then design and run experiments using AdFisher to evaluate the effectiveness of privacy tools.

II. Information Flow Experiments

Tschantz et al. [7] show that information flow is a causal relationship and use randomized controlled experiments to establish causation. They use the permutation test, which tests for a null hypothesis stating that outputs from the experimental and control units are drawn from the same distribution. If the resulting p-value is smaller than the significance criterion (α), then they reject the null hypothesis, thereby concluding the presence of information flow. In their experiments on the Google Ad System, Datta et al. [6] used α=0.05, which indicates that their test would lead to unsound conclusions at most 5% of the time.[1]

Additionally, it is also important for a test to be powerful. The power of a statistical test is the probability that it correctly rejects the null hypothesis, when the null hypothesis is indeed false. The prior methodology for IFEs lacks power analysis, thereby lacking any completeness evaluation.

III. Empirical Completeness Evaluation

The soundness and completeness of information flow experiments relies on the type-I (false-positive) and type-II (false-negative) error rates of the underlying statistical test. The maximum acceptable type-I error rate is denoted by α, which is a measure of the unsoundness of the test. The completeness of a test is quantified by power, given by $1-\beta$, where β is the type-II error rate.

To empirically estimate the type-I and type-II errors of information flow experiments, we use Monte-Carlo simulations. Monte-Carlo simulations have been successfully used to estimate power for non-parametric tests [8]. In such Monte-Carlo methods, one first simulates populations so as to introduce a pre-determined effect. Thereafter, samples are drawn from the populations and the test performed on the datasets to check if the null hypothesis is rejected or not. This is repeated for a large number of iterations. If the effect size is set to zero, the proportion of iterations where the test incorrectly rejects the null hypothesis gives an estimate of type-I error rate. Similarly, by setting the effect-size to a non-zero value, the proportion where the test correctly rejects the null hypothesis gives an estimate of power (completeness rate). We use Cohen's percent nonoverlap (pno) as a measure of effect size [9]. The pno of two distributions is the area under the two distributions that overlap as a ratio of the total area under the two distributions. For example, if the distributions are identical, then pno $= 100\%$.

Using Monte-Carlo simulations, we perform an empirical soundness and completeness analysis of AdFisher instantiated with the permutation test and the automatically selected classifier-based test statistic. We limit the number of unique features to 1000 and the total features observed by each unit to 50, in accordance with their experiments, and compute estimates over different distributions over the features. We cover a range of distributions starting from the basic ones to more complex and noisy versions. For soundness, we expect a type-I error estimate of at most 5%, since we select a significance criterion of 0.05. We estimate the type-I errors by averaging over 100 Monte-Carlo simulations, each at different

[1] They apply Holm-Bonferroni correction to address multiple comparisons.

TABLE I. SOUNDNESS AND COMPLETENESS ESTIMATIONS

	Soundness	Completeness	
Distribution	Type-I error	Effect size (pno)	Samples
Poisson	0.022	$> 20\%$	40
Poisson with 5% noise	0.013	$> 26\%$	40
Normal	0.02	$> 5\%$ *when means differ*	40
		$< 50\%$ *when means do not differ*	-

TABLE II. EXPERIMENTAL TREATMENTS

Treatment A	Treatment B	Treatment C	Treatment D
¬ Visit sites	Visit sites	Visit sites	¬ Visit sites
¬ Privacy tool	¬ Privacy tool	Privacy tool	Privacy tool

TABLE III. EVALUATION OF PRIVACY TOOLS

Privacy Tool	$p(A, B)$	$p(B, C)$	$p(C, D)$	Conclusion
AdBlockPlus	0.022	0.004	0.062	Completely effective
Ghostery	0.064	0.141	0.936	Inconclusive

instantiations of the distributions (for example at 6 randomly selected values of the parameter λ for the Poisson distribution). We find that the empirical type-I error matches the set significance criterion, thereby providing a sanity check for the Monte-Carlo simulations. Table I shows our results.

As for completeness, we estimate how small of an effect size the tests can detect with a power of at least 95% for a reasonable number of samples. For the Poisson distributions, with just 40 samples, effect sizes as low as 20% are detected. For normal distributions with mean differences, 40 samples are enough to detect effect sizes as small as 5%. However, with only differences in the variance, even 340 samples achieved power levels of just 5%. This points out that the default instantiation of AdFisher is incapable of detecting differences in variance, likely because it trains a linear classifier. Thus, AdFisher is able to detect effect sizes as small as 26% for these distributions, if the differences are linear in nature.

IV. EXPERIMENTS

Design. To measure the effectiveness of privacy tools, we perform experiments comparing four treatment groups. The first treatment group (A) does not visit any website or install any privacy tool. The second treatment group (B) visits certain websites in order to induce a behavioral profile, but does not install any privacy tool. The third treatment group (C) visits the same websites as group B, as well as installs a privacy tool. The fourth treatment group (D) does not visit any websites, but just installs the privacy tool. Table II shows the different treatments. Once the treatments are applied, all the browsers collect ads from a third-party news website.

By observing a significant difference between A and B (i.e. $p(A, B) < 0.05$), we can conclude that website visits have a significant impact on the content, thereby indicating that visiting the websites leads to a behavioral profile being created and used to serve targeted content. By comparing B and C, we can evaluate if having a privacy tool installed has any significant effect on the measurements. We consider a tool to be *soundly effective* if $p(B, C) < 0.05$. We consider a privacy tool to be *completely effective*, if, in addition, $p(C, D) \geq 0.05$. We clarify that finding no difference does not guarantee that C and D does not have any difference, rather AdFisher's linear classifier is not able to find any consistent difference. We have empirically shown that Adfisher can detect pretty small differences in Section III. The complete effectiveness of the tool is with respect to AdFisher's detection methods.

If, however, we continue to find a significant difference between C and D, this would demonstrate that the privacy tool is not completely effective, and visiting websites still has some impact on recommended outputs in spite of using privacy tools. The elements of difference would point us to what outputs are being targeted.

Results. We evaluated AdBlockPlus and Ghostery in our experiments. We visited sites about automobiles and collected text ads served by Google on bbc.com/news. Table III shows the p-values from comparing the pair of treatments (A, B), (B, C), and (C, D) for each privacy tool. For AdBlockPlus, both $p(A, B)$ and $p(B, C)$ are < 0.05, thereby showing that it is soundly effective. In addition, $p(C, D) \geq 0.05$, thereby making it completely effective. However, $p(C, D)$ for AdBlockPlus is very close to achieving significance. Thus we suspect that increasing the sample size may allow us to show that AdBlockPlus is soundly effective, but not completely. Our experiments evaluating Ghostery were inconclusive, since none of the p-values were significant.

Thus, our experimental design enables us to evaluate the sound and complete effectiveness of any privacy tool. We perform an empirical completeness evaluation of the default instantiation of AdFisher to gauge the sensitivity of our methods. We acknowledge that these evaluations are limited to some standard distributions, and that they may not hold on the outputs we measure. Nevertheless, these results pave the way to evaluate more tools and at a larger scale, while still providing soundness and qualified completeness.

REFERENCES

[1] G. Acar, C. Eubank, S. Englehardt, M. Juarez, A. Narayanan, and C. Diaz, "The web never forgets: Persistent tracking mechanisms in the wild," in *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*, 2014, pp. 674–689.

[2] F. Roesner, T. Kohno, and D. Wetherall, "Detecting and defending against third-party tracking on the web," in *Proceedings of the 9th USENIX conference on Networked Systems Design and Implementation*, 2012.

[3] N. Nikiforakis, W. Joosen, and B. Livshits, "Privaricator: Deceiving fingerprinters with little white lies," in *Proceedings of the 24th International Conference on World Wide Web*, 2015, pp. 820–830.

[4] R. Balebako, P. Leon, R. Shay, B. Ur, Y. Wang, and L. Cranor, "Measuring the effectiveness of privacy tools for limiting behavioral advertising," in *Web 2.0 Security and Privacy Workshop*, 2012.

[5] J. R. Mayer and J. C. Mitchell, "Third-party web tracking: Policy and technology," in *Security and Privacy, IEEE Symposium on*, 2012.

[6] A. Datta, M. C. Tschantz, and A. Datta, "Automated experiments on ad privacy settings," *Proceedings on Privacy Enhancing Technologies*, vol. 2015, no. 1, pp. 92–112, 2015.

[7] M. C. Tschantz, A. Datta, A. Datta, and J. M. Wing, "A methodology for information flow experiments," in *Computer Security Foundations Symposium, IEEE 28th*, 2015.

[8] P. J. Mumby, "Statistical power of non-parametric tests: A quick guide for designing sampling strategies," *Marine pollution bulletin*, 2002.

[9] J. Cohen, *Statistical power analysis for the behavioral sciences*. Academic Press, 1969.