Privee: An Architecture for Automatically Analyzing Web Privacy Policies

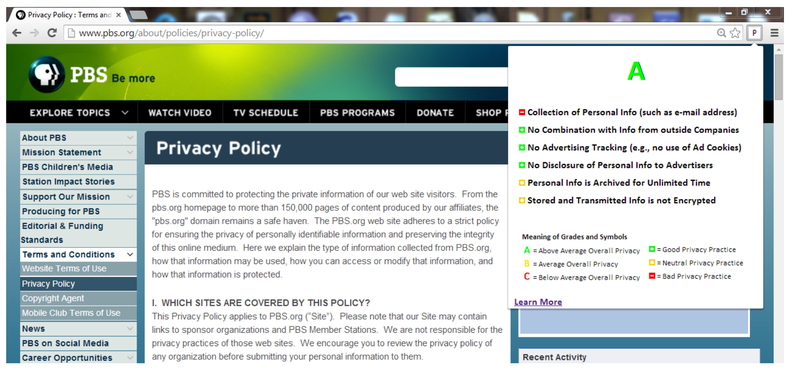
Sebastian Zimmeck and Steven M. Bellovin

1. Problem: Web Users do not Read Privacy Policies

Privacy policies on websites are based on the **notice-and-choice principle**. They notify Web users of their privacy choices. However, many users do not read privacy policies or have difficulties understanding them. The resulting **information asymmetry** leaves users uninformed about their privacy choices, can lead to **market failure**, and **calls the notice-and-choice principle into question** altogether.

2. Solution: Automatic Analysis of Privacy Policies

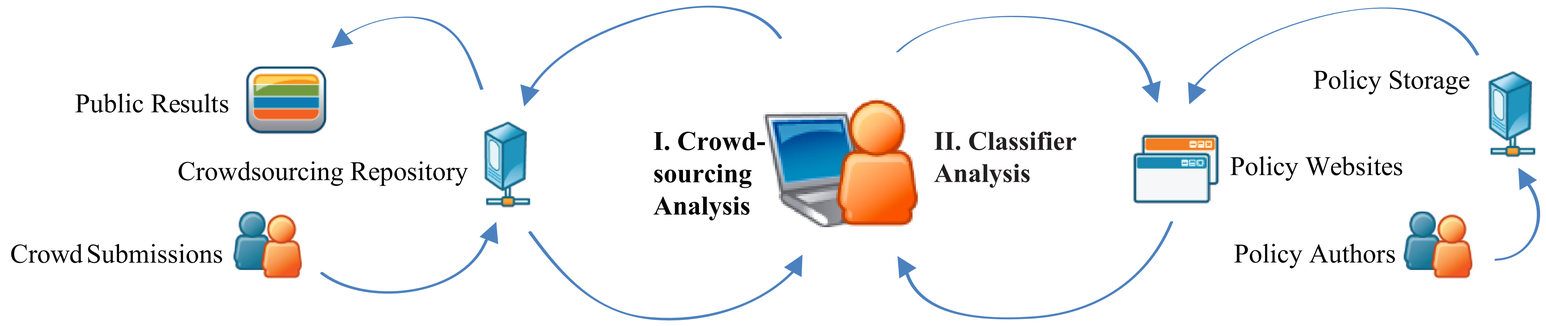
In order to increase privacy transparency we propose **Privee—a software architecture for analyzing essential policy terms** that combines crowdsourcing with rule- and machine learning-based classification techniques.



Our Privee browser extension (implemented for Google Chrome) performs a classifier analysis checking whether a privacy policy:

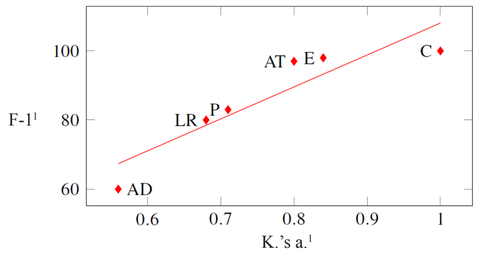
- allows **collection** of personal information from users;
- provides **encryption** for information storage or transmission;
- allows **ad tracking** by means of ad cookies or other trackers;
- provides a **limited retention** period for personal information;
- allows **profiling** of users by combining collected information with information from third parties;
- allows personal information **disclosure to advertisers**.

3. The Privee Concept



When a user requests a privacy policy analysis, the program checks whether analysis results are available at a crowdsourcing repository (to which crowd contributors can submit analysis results of policies). If results are available, they are returned and displayed to the user (**I. Crowdsourcing Analysis**). If no results are available, the policy text is fetched from the policy website, analyzed by automatic classifiers on the client machine, and the analysis results are displayed to the user (**II. Classifier Analysis**).

4. Performance and Inter-annotator Agreement



	Base.	F-1	K.'s a.
Overall	**68%**	**90%**	**0.77**
Collection	100%	100%	1
Encryption	52%	98%	0.84
Ad Tracking	64%	97%	0.8
L. Retention	74%	80%	0.68
Profiling	52.%	83%	0.71
Ad Disclosure	66%	60%	0.56

Our classifiers (Naive Bayes and/or rules) have an **overall F-1 score of 90%** (when compared to human annotators trained in privacy law). The baseline accuracy consists of always selecting the classification that occurred the most for the respective category in our training set.

To ensure the reliability of annotations we calculated the inter-annotator agreement by Krippendorff's alpha, which indicated for all categories fair or good agreement (except for Ad Disclosure). It is striking that **performance (F-1 score) correlates to agreement (Krippendorff's alpha).**

1. AD = Ad Disclosure, LR = Limited Retention, P = Profiling, AT = Ad Tracking, E = Encryption, C = Collection.

5. Semantic Diversity

	Extr. Text	Section
Mean Sem. D.	1.87	2.08
Significance (P)	**0.02**	**0.04**
Odds Ratio (Z)[2]	2.07	1.51
95% C. Int (Z)[2]	1.12-3.81	1.02-2.22

Our experimental results suggest that **classifier performance is inherently limited as it correlates to the same variable to which human interpretations correlate—the ambiguity of natural language**. We measured this ambiguity in form of **semantic diversity.** The less ambiguity in the extracted text for the classifier to analyze and in the section for the annotators to read, the fewer misclassifications and disagreements occurred, respectively.

2. Odds ratios and confidence intervals were normalized to Z scores.

6. Reference

Sebastian Zimmeck and Steven M. Bellovin, Privee: An Architecture for Automatically Analyzing Web Privacy Policies, 23rd USENIX Security Symposium, San Diego, CA, USA, August 2014.

COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK