August 20, 2014

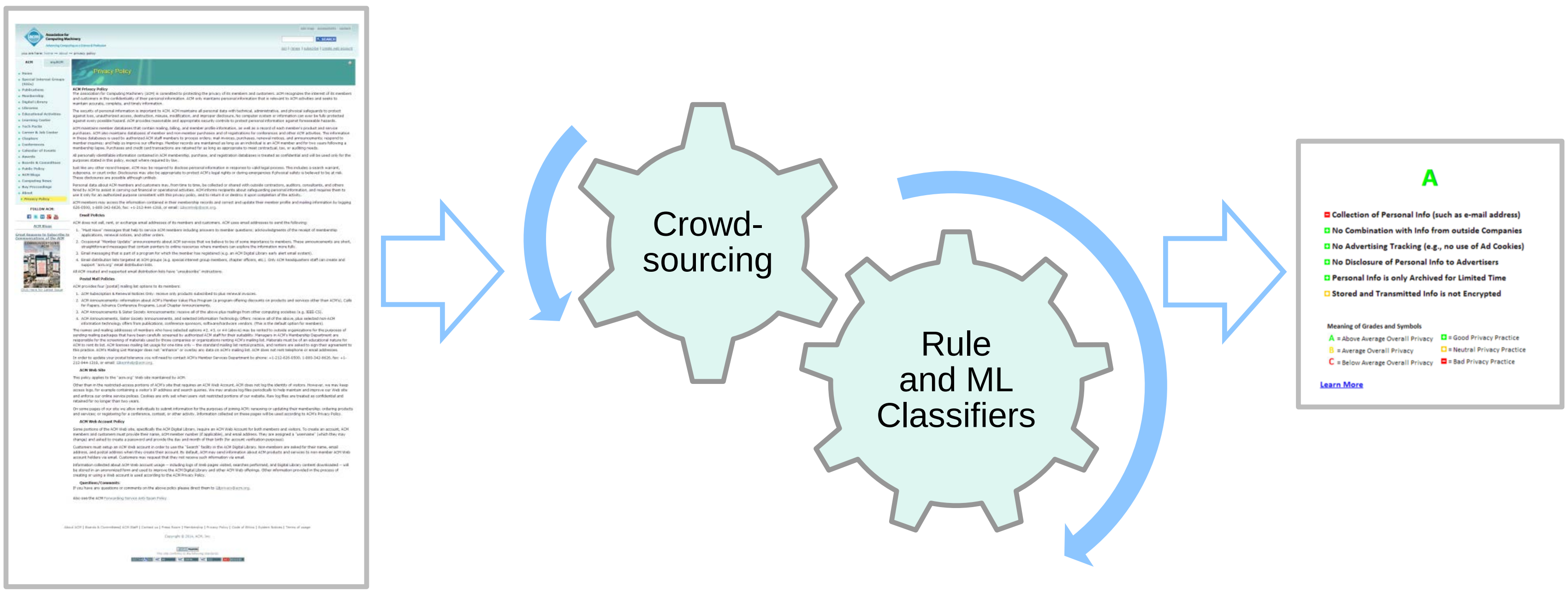
Privee: An Architecture for Automatically Analyzing Web Privacy Policies

Sebastian Zimmeck and Steven M. Bellovin

23rd USENIX Security Symposium, San Diego, CA

Columbia University
IN THE CITY OF NEW YORK

Problem: Not many Web Users read Privacy Policies
Solution: Privee—Automatic Privacy Policy Analysis



P rivee

Talk Overview

1. Problem
2. Privee
3. Performance
4. Reliability
5. Summary

Talk Overview

1. Problem

2. Privee

3. Performance

4. Reliability

5. Summary

- **Notice-and-choice Principle**
  → Notification of Privacy Practices
  → Consent

- Federal Trade Commission enforces Violations of Privacy Promises
  → "Unfair or deceptive acts or practices in or affecting commerce“ (15 U.S.C. § 45(a)(1))

- Privacy Policies as Contracts
  → Direct Relationship between the Contract Parties that allows for individualized Privacy Levels

- Only few Web Users read Privacy Policies
  → Information Asymmetry
  → Market Failure

Three Previous Approaches

Privacy Policy Languages

- Making Privacy Policies machine-readable for Computers to read

P3P

Privacy Finder

W3C® Platform for Privacy Preferences Initiative

Labels

- Expressing Privacy Policies in Label Format using Short Descriptions and Icons

KnowPrivacy

Privacy Icons

Crowdsourcing

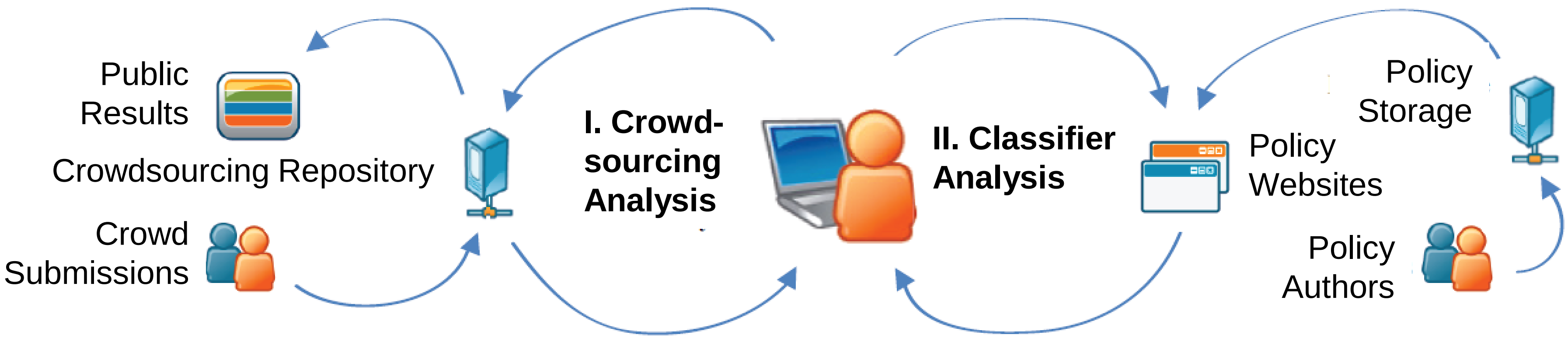
- Crowd Analysis of Privacy Policies and Submission of Results for Publication on the Web

Terms of Service Didn't Read

TOSBack

→ Low Industry Adoption Rate and User Interest

1. Problem
2. **Privee**
3. Performance
4. Reliability
5. Summary

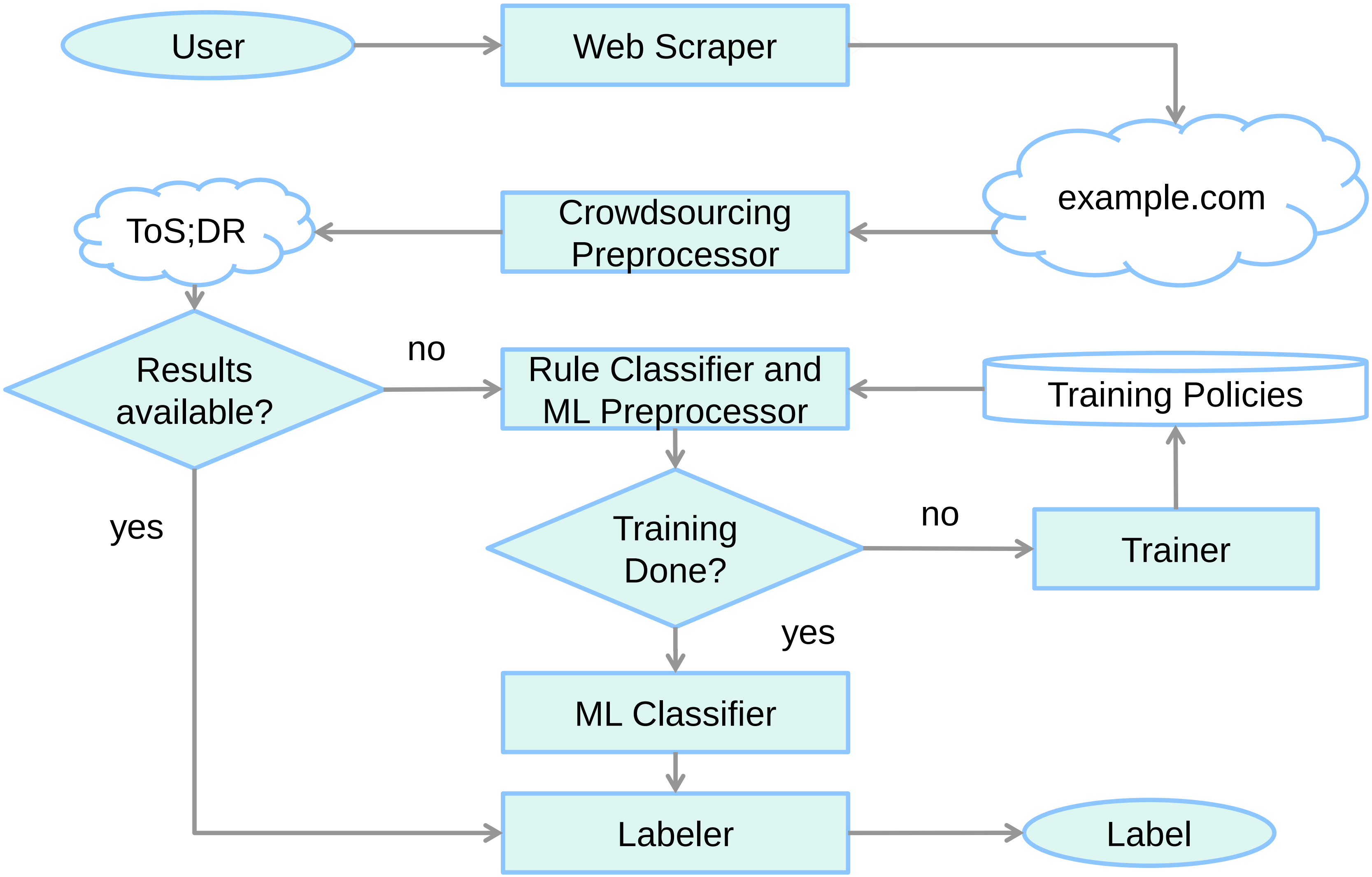


P rivee

I. Crowdsourcing Analysis: If Policy Analysis Results are available at a Crowdsourcing Repository, they are returned to the User

II. Classifier Analysis: Otherwise, the Policy Text is obtained from the Policy Website, automatically classified on the client machine, and results are returned to the User

The Privee Browser Extension



A Detailed Look at Preprocessing and Classification

Binary Classification

Rule Classifier

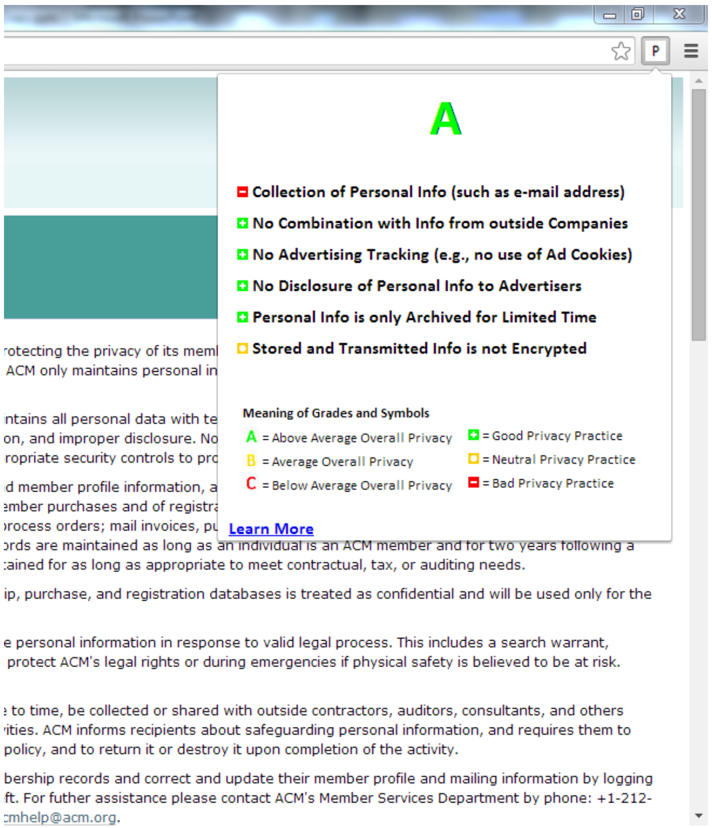
- Return a Class if Regular Expression matches Bigram that is nearly always associated with a certain Privacy Practice (e.g., “Ad Network" is nearly always associated with Ad Tracking Practice)

ML Preprocessor

- If the Rule Classifier did not return a Class, match Regular Expressions to collect Vocabulary on which the ML Classifier will run (e.g., all Bigrams that contain the Word “Ad“, “Marketing”, or “Behavioral”)

ML Classifier

- Find out whether Class (e.g., “Ad Tracking)” or complement class (e.g., “No Ad Tracking”) is more likely (if ML Preprocessor did not extract any Vocabulary, select Complement Class)



Six Binary Classification Categories

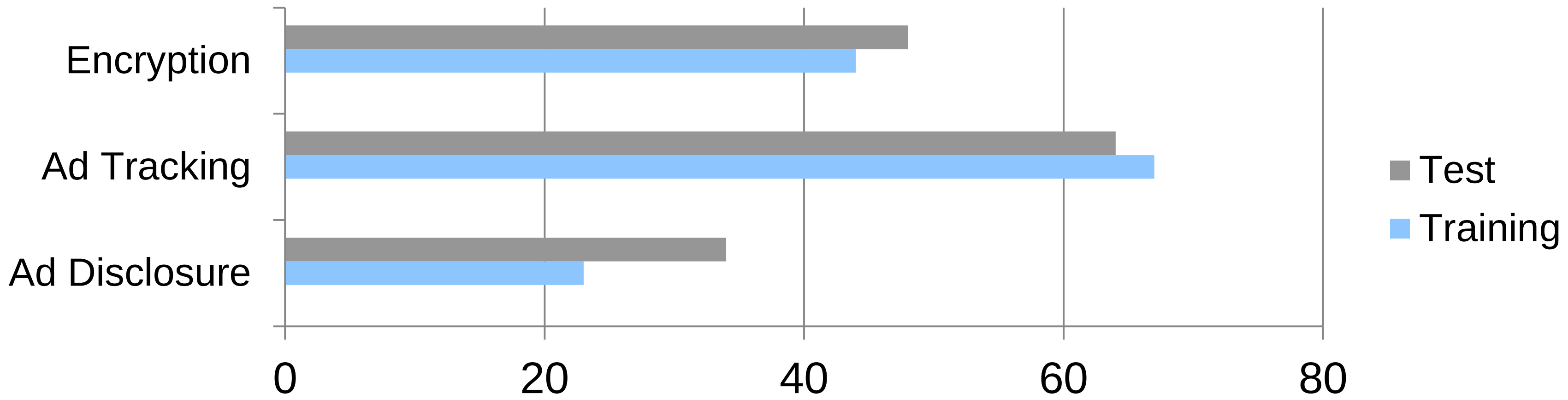
1. **Collection** of Personal Information from Users
2. **Profiling** of Users by combining own Information with 3rd Party Information
3. **Ad Tracking** by Means of Ad Cookies or other Trackers
4. **Ad Disclosure** analyzes Personal Information to Advertisers
5. **Limited Retention** Period for Personal Information
6. **Encryption** for Information Storage or Transmission

Overall Grade Assignment: A, B, or C

Talk Overview

1. Problem
2. Privee
3. **Performance**
4. Reliability
5. Summary

Percentage of Policies in Test and Training Set that allow for a certain Practice



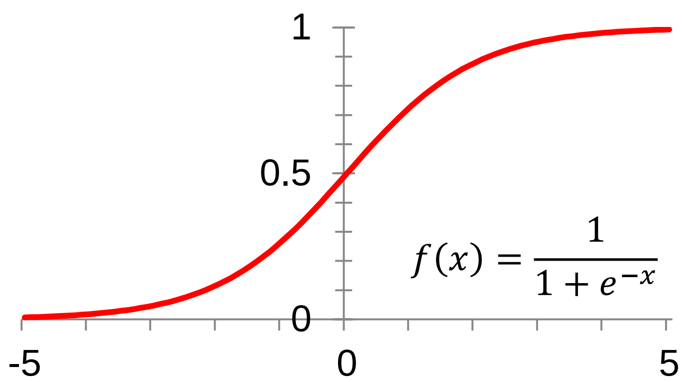
Classification Performance Results

	Baseline	Accuracy	Precision	Recall	F-1 Score
Overall	68%	84%	94%	89%	**90%**
Encryption	52%	98%	96%	100%	98%
Ad Tracking	64%	96%	94%	100%	97%
Ad Disclosure	66%	76%	69%	53%	60%

Binary Logistic Regression

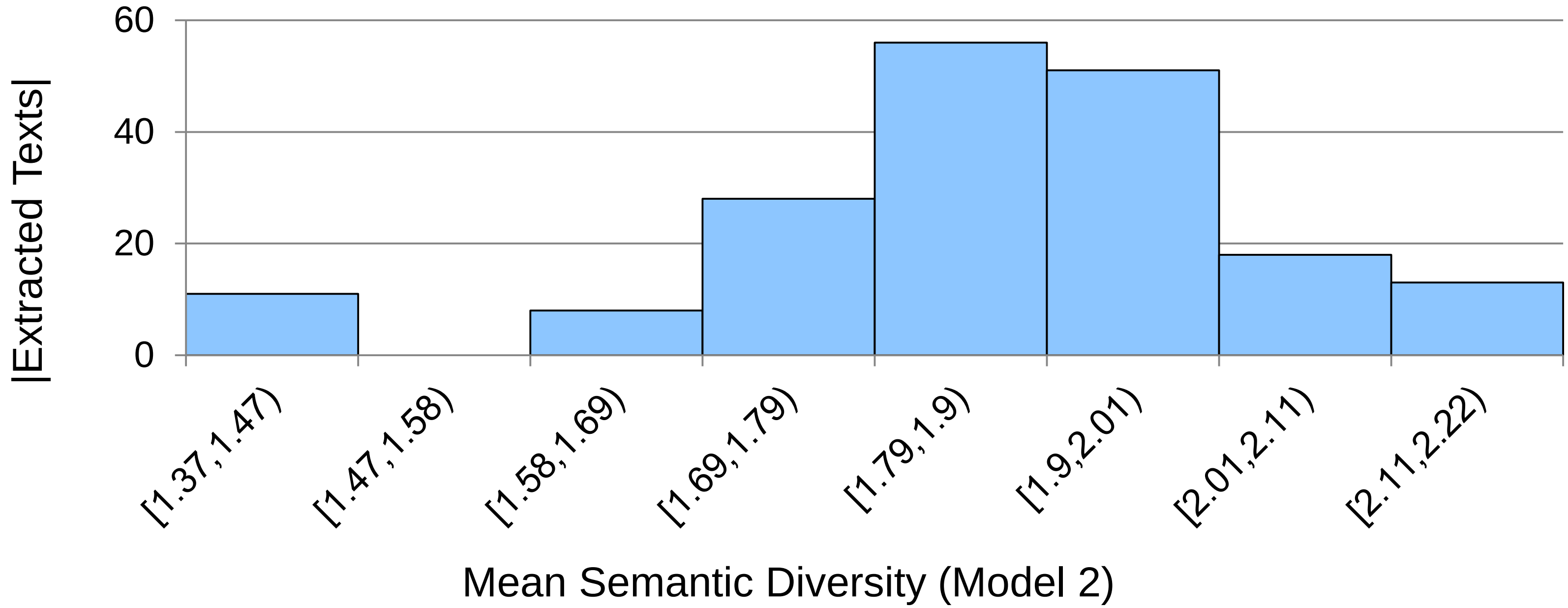
Model 1 (per Policy)
- Dependent Variable: Misclassification
- Independent Variables:
  (1) Policy Length (in Words)
  (2) Semantic Diversity (Mean)
  (3) Annotator Disagreement

Model 2 (per extracted Text)
- Dependent Variable: Misclassification
- Independent Variables:
  (1) Text Length (in Words)
  (2) Semantic Diversity (Mean)
  (3) Annotator Disagreement



Semantic Diversity is statistically significant (P = 0.02) for whether a Misclassification occurs or not

Semantic Diversity



- Semantic Diversity is an Ambiguity Measure based on Latent Semantic Analysis
  It can range from 0 (highly unambiguous) to 2.5 (highly ambiguous)

- **Model 2: An increase of the Mean Semantic Diversity in an Extracted Text by 0.17 (One Standard Deviation) increases Likelihood of Misclassification by 2.07 Times**

Talk Overview

1. Problem
2. Privee
3. Performance
4. **Reliability**
5. Summary

- Measuring Classifier Performance requires a Gold Standard (i.e., Ground Truth)
- A Test Set of 50 Policies was annotated by three qualified Annotators and the Annotation on which at least two Annotators agreed was selected
- The higher the Inter-annotator Agreement, the more reliable the Gold Standard

Inter-annotator Agreement Results

	Disagreement	% Agreement	Krippendorff's α
Overall	8.12	84%	**0.77**
Encryption	6	88%	0.84
Ad Tracking	7	86%	0.8
Ad Disclosure	16	68%	0.56

Binary Logistic Regression

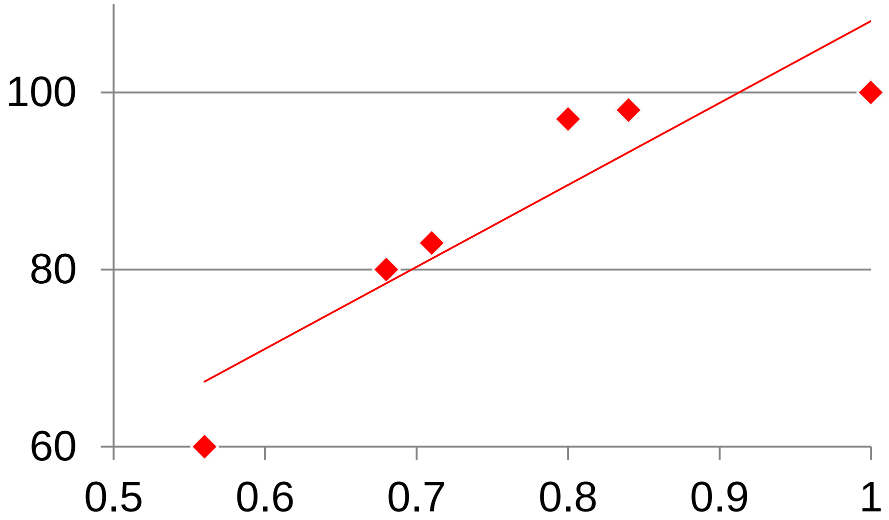
Model 3 (per Policy)
- Dependent Variable: Disagreement
- Independent Variables:
  (1) Policy Length (in Words), (2) Semantic Diversity (Mean), (3) Flesh-Kincaid Score

Model 4 (per Policy Section)
- Dependent Variable: Disagreement
- Independent Variables:
  (1) Section Length (in Words), (2) Semantic Diversity (Mean), (3) Flesh-Kincaid Score

- **In Model 4 Semantic Diversity is statistically significant (P = 0.04) for whether a Disagreement occurs or not**
- **An Increase of the Mean Semantic Diversity in a Policy Section by 0.03 (One Standard Deviation) increases the Likelihood of Disagreement by 1.51 Times**

Correlation of Performance (F-1 Score) and Agreement (Krippendorf's α)

- **Performance and Agreement correlate to the same variable—Semantic Diversity**
- Further, as shown below, the Values of Krippendorff's α also correlate to the F-1 Scores

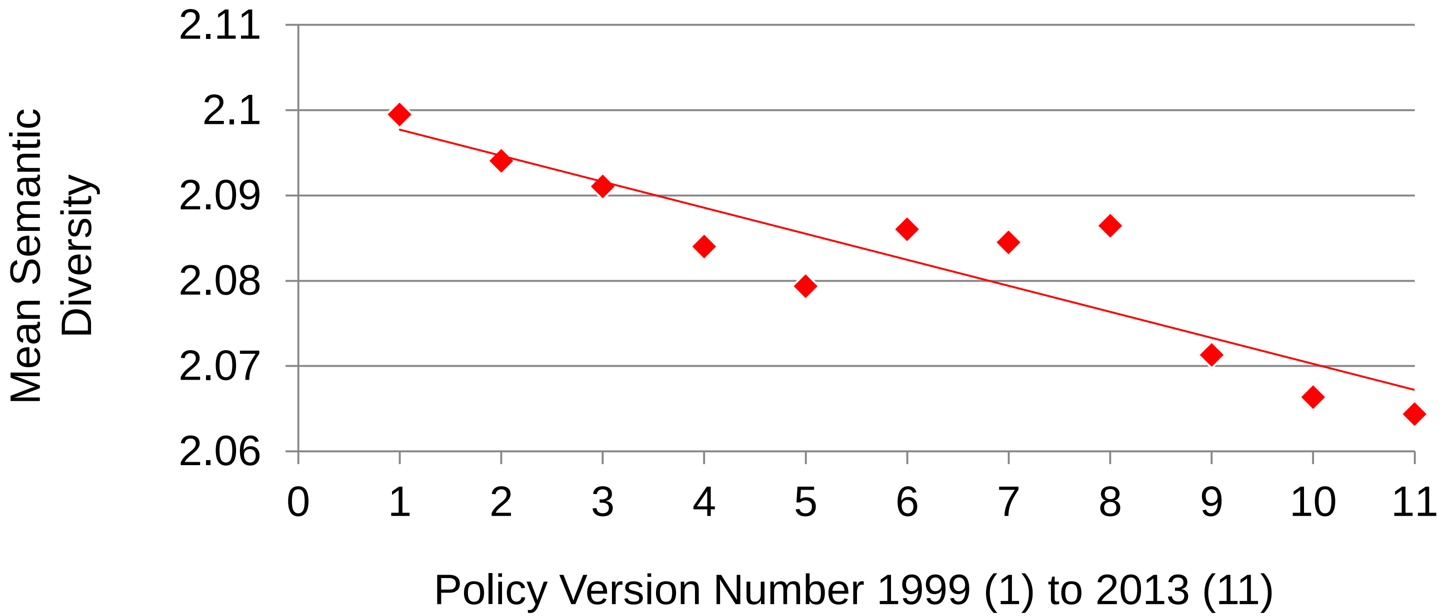


	F-1 Score	Krippendorff's α
Collection	100%	1
Encryption	98%	0.84
Ad Tracking	97%	0.8
L. Retention	80%	0.68
Profiling	83%	0.71
Ad Disclosure	60%	0.56

Will Policy Ambiguity impede the Notice-and-Choice Principle?

No, for the Majority of the Privacy Policies in our Test Set we observed a statistically relevant Decrease of Semantic Diversity over Time (P = 0.049)

Semantic Diversity of Symantec's Privacy Policy



Talk Overview

1. Problem

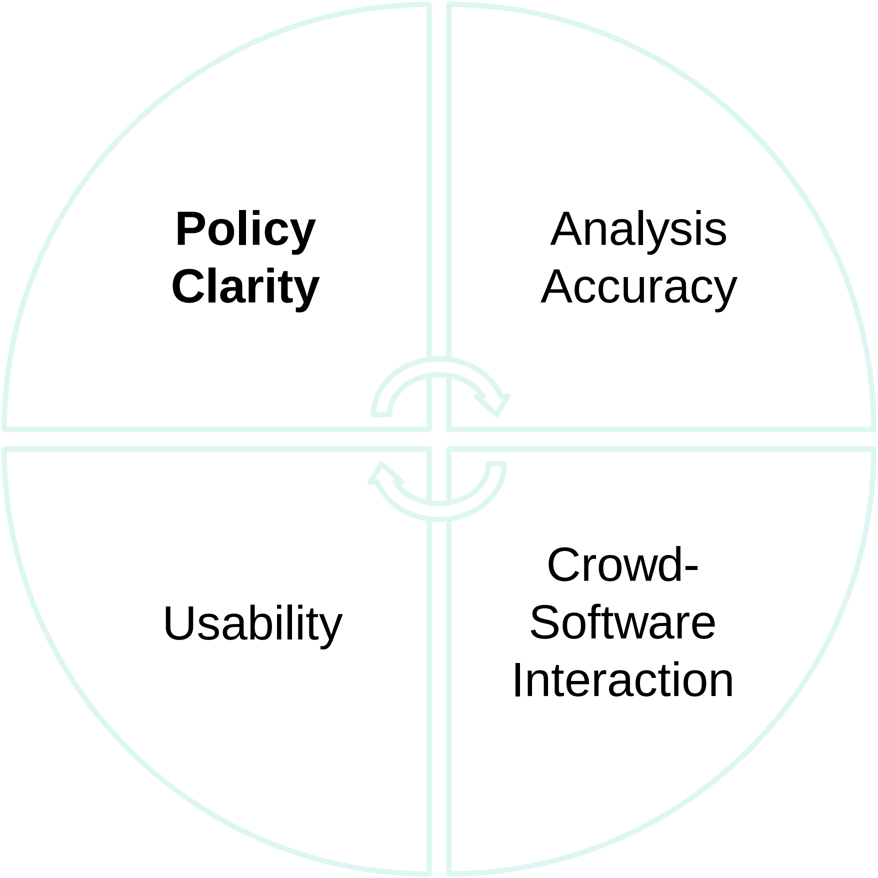
2. Privee

3. Performance

4. Reliability

5. Summary

Future Work


Policy Clarity
Analysis Accuracy
Usability
Crowd-Software Interaction

Highlights

- We introduced Privee—a novel Concept for analyzing Web Privacy Policies based on Crowdsourcing and Automatic Classification Techniques
- Our results suggest that the Automatic Classification of Privacy Policies as well as their Human Interpretation is limited by the Ambiguity of Natural Language
- As Policy Ambiguity seems to decrease over Time we remain optimistic that the Notice-and-choice Principle is workable and can be supplemented by Privee
- See http://www.sebastianzimmeck.de/publications.html for our Privee Extension