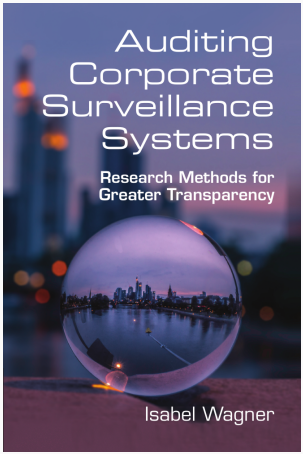
DATA ANALYSIS

Isabel Wagner

De Montfort University



DATA ANALYSIS & PRESENTATION

- Aims of data analysis
  - Are differences between experimental treatments and control statistically significant?
  - What causes the differences?
  - Trends (especially longitudinal studies)?
  - Pre-processing needed? Heuristics, machine learning
- Aims of data presentation
  - Visualize results for human audience

Quantitative measures

MEASURES FOR TRACKING

- Prevalence: number of trackers in a website/app
- Reach: percentage of all websites tracked by a given tracker
  - In practice: “all” websites = all websites in a sample
  - Reach on domain level vs reach on corporation level
- Prominence: reach adjusted for website rank
  - $\text{Prominence}(t) = \sum_{\text{present}(s,t)} \frac{1}{rank(s)}$
  - Robust to number of websites in study, good for comparisons
- Concentration of power: measure (lack of) competition in tracking ecosystem
  - Market share of tracker: $s_{t_i} = \frac{\text{prominence}(t_i)}{\sum_{j=1}^{N} \text{prominence}(t_j)}$
  - Herfindahl–Hirschman Index $HHI = \sum_{i=i}^{N} s_i^2$
- Penetration: percentage of users who see given tracker during a period of time (e.g., 1/2/5/10 days)

MEASURES FOR FINGERPRINTING

- Anonymity set size: number of browsers with the same fingerprint (indistinguishable by fingerprinter)
- Entropy: level of identifying information in a fingerprint
  - Fingerprinting attribute X, frequency of its values $P(x_i)$
  - Entropy $H(X) = -\sum_{i=0}^{n} P(x_i) \log_b P(x_i)$
  - Entropy measured in bits ($b = 2$): one additional bit doubles probability for fingerprinter to identify a browser
  - Entropy depends on number of samples N, not good for comparisons
  - Normalized entropy $H_N(X) = \frac{H(X)}{\log_b(N)}$

MEASURES FOR USER PRIVACY

- Identifiability/unicity/uniqueness
  - How many pieces of information does adversary need to uniquely identify a user?
  - Example of mobile apps: dataset D contains combinations of apps installed by users; U is the set of apps installed by a specific user
  - Uniqueness is size of subset $S \subset D$ that contains U
  - Other examples: spatiotemporal points, purchases, click traces, browsing histories
- Scale and sensitivity of information leaks, e.g.:
  - Number of attributes contained in each ad request[1]
  - Type and frequency of PII leaks[2]

[1] S. Nath, "MAdScope: Characterizing Mobile In-App Targeted Ads," in *Proceedings of the 13th Annual International Conference on Mobile Systems, Applications, and Services*, ser. MobiSys '15, Florence, Italy: ACM, 2015, pp. 59–73. DOI: 10.1145/2742647.2742653.

[2] J. Ren, M. Lindorfer, D. J. Dubois, *et al.*, "Bug Fixes, Improvements, ... and Privacy Leaks - A Longitudinal Study of PII Leaks Across Android App Versions," in *Proceedings 2018 Network and Distributed System Security Symposium*, San Diego, CA, USA: Internet Society, Feb. 2018. DOI: 10.14722/ndss.2018.23143.

MEASURES FOR FAIRNESS & DISCRIMINATION

- Discrimination
  - Direct: algorithm takes protected characteristic as input (disparate treatment)
  - Indirect: algorithm output correlates with protected characteristic (disparate impact)
- Fairness
  - Individual: similar individuals should be treated similarly
  - Group: Subsets of the population, according to protected characteristic, should be treated similar to entire population
- Representation ratio: based on disparate impact measure[3]
  - Frequency of users being selected (e.g., for ad targeting) when they have vs. do not have a sensitive attribute
  - $rr_S(T, R) = \frac{|T \cap R_s|/|R_s|}{|T \cap R_{\neg s}|/|R_{\neg s}|}$, where s sensitive attribute, T target audience, R relevant audience
  - Disparity $\text{disp}_s(T, R) = \max(rr_S(T, R), \frac{1}{rr_S(T,R)})$
  - Targeting is discriminatory if disparity exceeds a threshold, e.g., 1.25

[3] T. Speicher, M. Ali, G. Venkatadri, *et al.*, "Potential for Discrimination in Online Targeted Advertising," in *Conference on Fairness, Accountability and*

MEASURES FOR BIAS

- Overlap between sets (e.g. extent of personalization of search results)[4]
  - Jaccard index: similarity between two sets
  - Cosine similarity: similarity between two vectors
  - Overlap coefficient: Jaccard variant for sets of very different sizes
  - Edit distance, e.g., Damerau-Levenshtein distance, Kendall Tau distance: similarity between ranked lists
  - Rank-biased overlap: overlap adjusted by rank of website
- Political bias: very difficult, mostly based on proxy measures[5]
  - E.g., measure political bias of a news outlet as political bias of its audience, as estimated by Facebook and available via their advertiser interface

[4] A. Hannak, P. Sapiezynski, A. Molavi Kakhki, *et al.*, "Measuring Personalization of Web Search," in *Proceedings of the 22nd International Conference on World Wide Web*, ser. WWW '13, Rio de Janeiro, Brazil: ACM, 2013, pp. 527–538. DOI: *10.1145/2488388.2488435*.

[5] F. N. Ribeiro, L. Henrique, F. Benevenuto, *et al.*, "Media Bias Monitor: Quantifying Biases of Social Media News Outlets at Large-Scale," in *Twelfth International AAAI Conference on Web and Social Media*, Palo Alto, California, USA: AAAI Press, Jun. 2018.

Heuristics

WHY HEURISTICS?

- Need to process raw collected data to find interesting response variables
  - Tracking on corporation level: who owns a domain?
  - Which cookies are used for tracking?
  - Which third parties are trackers?
- Heuristics are practical approaches to extract response variables *at scale*
- Usually not optimal
  - May label some tracking third parties as benign (false negatives)
  - May label some benign third parties as trackers (false positives)

WHO OWNS A DOMAIN? WHO OWNS AN APP?

- whois lookup
  - But: use of whois privacy solutions increasingly common, 40% in 2017
- Combinations of (imperfect) methods
  - Loading the website and checking for redirects
  - whois registration email address
  - *Organization* field in TLS certificate
  - Manual inspection of website
- Hand-curated lists of parent-subsidiary relationships
- Crunchbase, Hoovers, opencorporates
- App store listing: name and contact of developer
- App binary: at least one host name contacted by the app belongs to owner

WHICH COOKIES CAN IDENTIFY USERS?

- Four-step heuristic
- Exclude session cookies and cookies with lifetime shorter than 90 days
  - But: maybe disregard lifetime criterion because cookie lifetime can be updated when cookie is accessed
- Parse cookie key and value based on common delimiters (-&:)
  - Multiple values can be stored in both key and value
- Discard parsed values shorter than 8 characters
  - Cannot hold enough information for unique identifiers
- Compare parsed values across entire experiment
  - Identifying cookie values remain the same throughout
  - But differ between measurement machines and virtual personas
  - Ratcliff-Obershelp algorithm checks for level of difference
  - Thresholds for similarity score: 33%, 55%, 66%

WHICH THIRD PARTIES ARE TRACKERS?

- Third parties: domain names different from first-party website
- Not all are trackers, e.g., content delivery networks
- Public blocklists
  - Trackers related to advertising: EasyList
  - Non-advertising trackers: EasyPrivacy
  - Other blocklists: Ghostery, Disconnect, Fanboy, Pi-hole
  - But: manually curated; web-focused, may miss mobile trackers
- Domain classification services
  - Is domain listed as *advertising* or *tracking*?
- Tracker behavior
  - Cookie values in URL parameters
  - Tracking keywords in URL parameters, e.g., *usermatch*, *rtb*, *cookiesync*
- Invisible pixels: no meaningful content, therefore likely trackers

When does cookie synchronization happen?

- Identifying cookie values that are transmitted in HTTP requests, responses, referers
  - URL or referer contain identifier: requested domain learns identifier
  - Requested domain redirects to a third domain: third domain learns identifier
  - Identifier in location of HTTP redirect: location domain learns identifier
- But: only works if plain cookie value is used for synchronization
  - Fails if cookie values are encrypted or hashed
- Match URL parameters (not values) against documentation of cookie synchronization APIs, e.g., DoubleClick

When is a website fingerprinting the user?

- Analyze API calls made by embedded JavaScript
- Detect each fingerprinting attribute separately
- Individual attributes may be benign, but combination of many indicates fingerprinting
  - Canvas fingerprinting: *HTMLCanvasElement*, *CanvasRenderingContext2D*
  - Canvas size >16x16px, text in 2+ colors, 10+ different characters, configure text properties with *fillText* and *strokeText*
  - Call *toDataURL* and *getImageData*, but not *save*, *restore*, *addEventListener*
  - Canvas font fingerprinting: set font properties to 50+ values, call *measureText*
  - WebRTC fingerprinting: *RTCPeerConnection*
  - AudioContext fingerprinting: *AudioContext*, *OscillatorNode*

WHICH WEBSITE ELEMENTS ARE ADVERTISING?

- Analyzing DOM tree: difficult due to complex structure, deep nesting, dynamic changes from JavaScript
- Matching against filter lists (e.g., EasyList): simple, but manually curated lists may miss ads
- Matching platform-specific DOM elements (e.g., Facebook's *Sponsored* tag): good when it works, but subject to change
  - Facebook changed the text to "SpSonSsoSredS" and later "SpSpSononSsososSredredSSS" (interweaving invisible letters S and placing groups of one or two letters into separate span elements)[6]

[6] J. B. Merrill and A. Tobin, "Facebook Moves to Block Ad Transparency Tools —…," *ProPublica*, Jan. 2019.

WHAT ARE THE CATEGORIES/TOPICS OF ADS?

- Analyze ad creative (image, text): limited information, may not allow reliable classification
- Record ad URL and landing page:
  - Ad URL often points to advertising/analytics server which redirects to the landing page
  - Find landing page URL without clicking on ad to avoid inducing cost for the advertiser
  - Can extract landing page URL from ad URL parameters (*adurl=* or *redirecturl=*)
- Topic of landing page: look up in online tagging services (McAfee, Alexa, Google AdWords, Cyren)

WHICH TARGETING TYPES DO ADS USE?

- Need to distinguish between static ads, contextual ads, demographic or geographic ads, profile-based ads, retargeted ads
- Retargeted: ads for which virtual persona has previously visited ad landing page
- Behavioral/profile-based ads
  - Ads collected with clean browser can only be static, contextual, or geographic
  - Ads served to more than one persona: demographic or geographic if personas have dissimilar behaviors
- Contextual vs profile-based ads:
  - Contextual if similarity between ad and website is high
  - Profile-based if similarity between ad and user's interests is high

WHICH BIDDING TECHNOLOGIES ARE USED FOR ADS?

- Real-time bidding
  - Only some messages observable from client-side: message with ad and winning bid price, message to winning bidder
  - Compare HTTP requests with known RTB message formats, e.g., DoubleClick or IAB
- Header bidding
  - Most messages observable from client-side
  - Static analysis of website DOM to check for well-known header bidding libraries (e.g., *prebid.js*, *gpt.js*)
  - Detect DOM events within browser: add event listeners, e.g., for end of bidding phase (*auctionEnd*), determination of winning bidder (*bidWon*), or ad rendering (*slotRenderEnded*)
  - Detect HTTP requests to well-known Demand Partners, or with header bidding parameters (*bidder*, *hb_partner*, *hb_price*)

IS PII PRESENT IN NETWORK TRAFFIC?

- Plain-text PII: choose experiment design to allow control over values, choose unique values (e.g., names, email addresses)
- Encrypted or obfuscated PII:
  - Can be nested, for example: SHA-1 hash of Android ID, XOR'ed with a random key, encoded with base64, appended to a JSON string, encrypted with RSA, encoded again with base64
  - Attempt de-obfuscation: can be successful for (combinations of) encodings
  - Deterministic obfuscation: repeat experiment 3x: twice with identical PII, once with control PII. PII leak if parameters in HTTP requests are same for first two cases, different for third case
  - Differential analysis: reduce sources of randomness (e.g., by patching Android), repeat experiment to establish baseline network traffic, then vary PII and repeat experiment to detect non-determinism

What are the demographics of a website's users?

- Use Facebook's ad audience size estimation tool
- Target ads at the website's Facebook page, audience size estimate only includes users who have liked the website
- Reasonable proxy for actual website visitors
- To record demographics: target ad at website + demographic attribute (gender, age, etc.)
- Record audience size estimate for each demographic attribute

WHAT IS THE POLITICAL LEANING OF A WEBSITE?

- Content-based approach
  - Linguistic analysis, measure differential use of phrases
  - Difficult for small text samples (tweets)
- Rater-based approach
  - Human raters evaluate political leaning, e.g., Media Bias/Fact Check
  - May suffer from rater bias
- Audience-based approach
  - Homophily: political leaning of (news) website is similar to political leaning of its audience
  - For tweets: compute similarity between interest vectors for target users and prototypical left-/right-leaning users
  - Interest vector: tf-idf vector of topics that the user's followers have been tagged with
  - For websites: use Facebook ad audience size estimation, target ads at website + political leaning

Which third-party libraries are included in an app?

- Simplest approach: match library name against package names
- But: fails if library uses obfuscation which changes package names
- Signature-based approach
  - Features for signatures: call frequencies for Android APIs, reference/inheritance relationships between classes and methods, function call graph
  - Signatures for libraries need to be precomputed (can be expensive), change with library version
- Signature also useful to detect library release dates and versions
  - Latest release date (upper bound) for library is the app's release date
  - Bound can become tight if library is used in many apps

Statistics

DESCRIPTIVE STATISTICS

- Statistics that describe characteristics of response variables
- Central tendency: mean, median
- Dispersion: standard deviation, variance, percentiles, extreme values
- Five-number summary: smallest observation, 25% percentile (lower quartile), median, 75% percentile (upper quartile), largest observation
- Visualization:
  - Bar charts, box plots, scatter plots, histograms
  - Can break down response variable by values of input variable
  - Can show evolution over time
- Descriptive statistics say nothing about causes or statistical significance!

HYPOTHESIS TESTING

- Hypothesis: possible explanation for an effect
- Null hypothesis: experimental treatment has no effect
- Hypothesis tests allow to reject the null hypothesis, i.e., show that the experimental treatment has a statistically significant effect
- Three steps:
  - Compute test statistic
  - Compute p-value: probability that a test statistic at least as extreme as the observed one is sampled under the null hypothesis
  - Reject null hypothesis if p-value is below significance level α (often $\alpha = 0.05$ or $\alpha = 0.001$)
  - If p-value is above significance level: no further conclusions! (we **cannot** accept the null hypothesis!)

DIFFERENT TYPES OF COMPARISONS NEED DIFFERENT TEST STATISTICS

- One-sample test: compare sample to known population
- Two-sample test: compare samples from two experimental conditions, e.g., experimental treatment vs. control treatment
- Paired test:
  - Compare samples from the same subjects, e.g., before/after treatment
  - Compare samples for matched treatment/control subjects (e.g., in observational studies)

STATISTICAL TESTS AND THEIR APPLICABILITY

- Common assumptions for test statistics are not satisfied in transparency research[7]:
  - Parametric models (system behavior follows known distribution)
  - Under null hypothesis (no effect), experimental units see independent and identically distributed responses
  - No cross-unit effects (treatment of one unit will not affect other units)
- Need to select *nonparametric* test statistics

Example:

Hypothesis (effect)	User data **is** used for marketing
Null hypothesis	User data **is not** used for marketing
Experimental unit	Browser instances
Experimental factor	User behavior
Constant factors	IP address, time of day, etc.
Response	Sequences of ads

[7] M. C. Tschantz, A. Datta, A. Datta, *et al.*, "A Methodology for Information Flow Experiments," in *2015 IEEE 28th Computer Security Foundations Symposium*, Verona, Italy: IEEE, Jul. 2015, pp. 554–568. DOI: 10.1109/CSF.2015.40

MULTIPLE TESTS OR MULTIPLE HYPOTHESES

- Hypotheses = possible explanations for effects
  - It can be reasonable to test multiple hypotheses on one set of recorded data
  - E.g., factors that can explain accuracy of Google's gender+age inference[8]
- Using multiple test statistics can also make sense
- Need to apply correction for multiple testing[9]
  - Holm-Bonferroni correction: adjust p-value so that probability of false rejection of null hypothesis is < 0.05
  - Benjamini-Yekutieli correction: adjust p-value so that fraction of false discoveries is < 0.05

[8] M. C. Tschantz, S. Egelman, J. Choi, *et al.*, "The Accuracy of the Demographic Inferences Shown on Google's Ad Settings," in *Proceedings of the 2018 Workshop on Privacy in the Electronic Society*, ser. WPES'18, Toronto, Canada: ACM, 2018, pp. 33–41. DOI: 10.1145/3267323.3268962.

[9] M. Lecuyer, R. Spahn, Y. Spiliopolous, *et al.*, "Sunlight: Fine-grained Targeting Detection at Scale with Statistical Confidence," in *Proceedings of the 22Nd ACM SIGSAC Conference on Computer and Communications Security*, ser. CCS '15, Denver, Colorado, USA: ACM, 2015, pp. 554–566. DOI: 10.1145/2810103.2813614.

Statistical tests: Pearson's chi-squared test

- Goodness of fit: compare distribution of observed data with theoretical distribution
  - Are ads on mobile devices targeted?[10]
  - Virtual personas with single interest
  - Record frequency of ads shown to each persona, distribution should be uniform if no behavioral targeting takes place
- Homogeneity: compare distribution of observed data for 2+ groups
  - For ads shown alongside search results, are there differences when searching for white-identifying names vs black-identifying names?[11]
  - Record frequencies with which ads contain target words (e.g., *arrest*)
  - Test difference between frequencies

[10] S. Nath, "MAdScope: Characterizing Mobile In-App Targeted Ads," in *Proceedings of the 13th Annual International Conference on Mobile Systems, Applications, and Services*, ser. MobiSys '15, Florence, Italy: ACM, 2015, pp. 59–73. DOI: 10.1145/2742647.2742653.

[11] L. Sweeney, "Discrimination in Online Ad Delivery," *Commun. ACM*, vol. 56, no. 5, pp. 44–54, May 2013. DOI: 10.1145/2447976.2447990.

STATISTICAL TESTS: KOLMOGOROV-SMIRNOV TEST

- Supports one-sample and two-sample tests
- Compare distribution of observed data with known distribution or with second set of observed data
- Does browser configuration affect the number of third-party requests?[12]
- Experimental treatment uses ad blocker, control treatment uses unmodified browser
- Null hypothesis: treatment has no effect, distributions of third-party requests are equal

[12] J. Mazel, R. Garnier, and K. Fukuda, "A comparison of web privacy protection techniques," *Computer Communications*, vol. 144, pp. 162–174, Aug. 2019. DOI: 10.1016/j.comcom.2019.04.005.

STATISTICAL TESTS: KRUSKAL-WALLIS, MANN-WHITNEY, WILCOXON SIGNED-RANK TESTS

- Do groups of observed samples have the same distribution?
- Mann-Whitney: 2 groups, independent samples
- Wilcoxon signed-rank: 2 groups, can compare dependent and paired samples
- Kruskal-Wallis: more than 2 groups, independent samples
- All tests can compare ordinal data (e.g., five-point Likert scales) because they work on ranks instead of numerical values
- ANOVA with Kruskal-Wallis and Mann-Whitney[13]
  - Do the characteristics of search results shown to voters influence the likelihood of voting for one of two political candidate?
  - Kruskal-Wallis to compare voting behavior of three groups pre-treatment
  - Mann-Whitney to compare voting behavior of each bias group with control group post-treatment

[13] R. Epstein and R. E. Robertson, "The search engine manipulation effect (SEME) and its possible impact on the outcomes of elections," *Proceedings of the National Academy of Sciences*, vol. 112, no. 33, E4512–E4521, Aug. 2015. DOI: *10.1073/pnas.1419828112*.

CONFIDENCE INTERVALS

- Confidence intervals estimate the value range for an unknown parameter (often the mean)
- Computed for specific confidence level, e.g., 0.95 or 0.99
- Confidence level indicates the probability that interval contains the true value
- Can be used to test specific hypotheses:
  - Null hypothesis: mean equals zero
  - If confidence interval does not contain zero, hypothesis can be rejected
  - Null hypothesis: mean for two groups is equal
  - If confidence intervals do not overlap, hypothesis can be rejected
  - Likelihood of rejecting true null hypothesis (significance level) = confidence level

REGRESSION

- Model relationship between response variable (dependent variable) and explanatory (independent) variables
- Can predict values of response variables
- Can explain to what extent explanatory variables “explain” response variables
  - More precisely: to what extent variation in the explanatory variables contributes to variation in the response variable
- Generic linear regression model: $y = X\beta + \epsilon$
  - y: vector of observations of response variable
  - X: matrix of values of explanatory variables
  - β: vector of regression coefficients estimated by the regression
  - ϵ: noise or error term, summarizes all influences that are not represented by explanatory variable

HYPOTHESIS TESTS IN REGRESSION

- Based on y and X, statistical estimation process estimates regression coefficients
- Performs hypothesis tests: is true value of each regression coefficient different from zero?
  - If not different from zero: corresponding explanatory variable has no explanatory power in predicting the response
  - Three significance levels for each hypothesis test: $p < 0.1$, $p < 0.05$, or $p < 0.01$
- What influences the rank of a hotel on a travel booking site?
  - X includes observations for hotel prices on different booking sites, star rating, user ratings, cancellation policy, ...
  - y: observed rank for each hotel

ESTIMATING CAUSAL EFFECTS

- Hypothesis tests: show significant *correlations* between input and response
- Not enough to show causation
- To show causes of effects: need hypothesis tests + suitable experiment design
- For active/experimental studies: permutation test with randomization and blocking
- For passive/observational studies: quasi-experiments based on propensity score matching or difference-in-differences method

PERMUTATION TEST

- Choice of test statistic: measure of distance between control and experimental group
  - E.g., number of ads related to a specific interest, cosine similarity, etc.
  - Can also be output of a machine learning classifier
- Permutation test[14]
  - Randomly permute the labels for recorded data, i.e. assign “control” and “experimental” randomly
  - Compute hypothetical value of test statistic for each permutation, compare with value for true labels
  - Rationale: if null hypothesis is true (no effect), then hypothetical test statistics shouldn’t be much different from actual value
  - p-value: proportion of permutations where hypothetical test statistic was $\geq$ to actual value

[14] 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, Apr. 2015. DOI: *10.1515/popets-2015-0007*.

BLOCKING FOR PERMUTATION TESTS

- Experimental units need to be assigned randomly to *control* and *experimental* groups
- Data collection for all experimental units should be in parallel to reduce noise
  - Can be expensive in terms of bandwidth, hardware
- Blocking:
  - Group experimental units into blocks
  - Parallel data collection within each block
  - Permutations performed within blocks
  - Execute multiple blocks in sequence to ensure sufficient number of experimental units

- Observational studies: cannot randomly assign treatment groups – treatment label is determined by observed data
- Find quasi-experiments in the data: sets of experimental units (subjects) with different treatment label, but similar control variables
- Control variables: alternative explanations for observed treatment
- Matching methods: find quasi-experiments systematically
  - Exact matches for control variables: results in low number of matches[15]
  - Match based on distance metric (e.g., Mahalanobis distance): matches on each control variable separately, can lead to low number of matches[16]

[15] E. A. Stuart, "Matching Methods for Causal Inference: A Review and a Look Forward," *Statistical Science*, vol. 25, no. 1, pp. 1–21, Feb. 2010. DOI: *10.1214/09-STS313*.

[16] S. Jiang, R. E. Robertson, and C. Wilson, "Bias Misperceived:The Role of Partisanship and Misinformation in YouTube Comment Moderation," *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 13, pp. 278–289, Jul. 2019.

PROPENSITY SCORE MATCHING

- Propensity score = probability of experimental unit having treatment label, based on values of all control variables
- Estimation of propensity score, e.g., with logistic regression (all control variables included as independent variables)
- Treatment subjects are matched to control subjects with closest propensity score (within threshold)
- Hypothesis tests to evaluate quality of matches
- Data analysis can proceed with standard hypothesis tests or regression
- Propensity-score-stratified regression: simulates randomized blocked trial by grouping subjects with similar propensity scores[17]

[17] E. Foong, N. Vincent, B. Hecht, *et al.*, "Women (Still) Ask For Less: Gender Differences in Hourly Rate in an Online Labor Marketplace," *Proc. ACM Hum.-Comput. Interact.*, vol. 2, no. CSCW, 53:1–53:21, Nov. 2018. DOI: 10.1145/3274322.

DIFFERENCE-IN-DIFFERENCES METHOD

- Useful when data includes cases before/after treatment was applied
- For example:
  - Treatment variable indicates whether subject uses an ad blocker or not
  - Subjects make the decision to install an ad blocker during the observation period
  - Observed data includes responses for all subjects without treatment, and for some subjects with treatment
- Regression model to analyze change over time in response variable for control group and treatment group[18]
- Treatment effect: difference in changes between the two groups

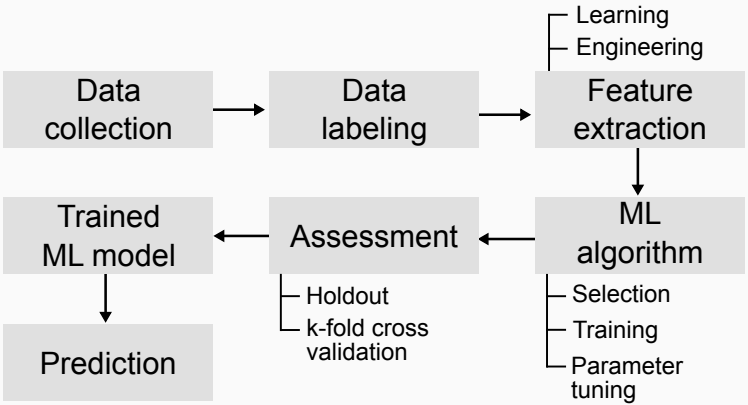
[18] A. Li, A. Wang, Z. Nazari, *et al.*, "Do podcasts and music compete with one another? Understanding users' audio streaming habits," in *Proceedings of The Web Conference 2020*, ser. WWW '20, Taipei, Taiwan: Association for Computing Machinery, Apr. 2020, pp. 1920–1931. DOI: *10.1145/3366423.3380260*, S. Zhao, A. Kalra, C. Borcea, *et al.*, "To be Tough or Soft: Measuring the Impact of Counter-Ad-blocking Strategies on User Engagement," in *Proceedings of The Web Conference 2020*, ser. WWW '20, Taipei, Taiwan: ACM, Apr. 2020, pp. 2690–2696. DOI: *10.1145/3366423.3380025*.

Machine learning

TYPES OF MACHINE LEARNING APPROACHES

- Supervised machine learning
  - Classification: predicts categories
  - Regression: predicts numeric values
- Unsupervised machine learning
  - Clustering: groups data

MACHINE LEARNING PROCESS



TRAINING OF MACHINE LEARNING MODELS

- Train on subset of collected/labeled data, evaluate performance of model on remainder
- Holdout method
  - Split data into fixed train/test sets, often 70%/30%
  - But: a *lucky* split can influence model performance
- k-fold cross validation
  - Split data into k random subsets
  - Perform k experiments, each uses $k - 1$ subsets for training and one subset for testing
  - Results are averaged to estimate true model performance
  - Often $k = 5$ or $k = 10$
- Test/train/validation splits (often 3:1:1 ratio)
  - Validation data used to tune the model (number of epochs, hyperparameters)
  - Test set only used once at very end to evaluate performance

- Performance metrics are computed based on the numbers of true positive samples (TP), true negatives (TN), false positives (FP), and false negatives (FN) observed during the testing phase
- Precision p: portion of positive results that are truly positive: $p = \frac{TP}{TP+FP}$
- Recall r (true positive rate): how many of the truly positive results were classified correctly: $r = \frac{TP}{TP+FN}$
- Accuracy a: portion of all data points that were classified correctly: $a = \frac{TP+TN}{TP+TN+FP+FN}$
- F1 score: harmonic mean between precision and recall: $F = 2 \times \frac{p \times r}{p+r}$

Research questions

- How prevalent is tracking? How many of a user's website visits are tracked?
- What is the reach of top trackers? How many websites are tracked by a tracker?
- To what extent do ad blockers and tracker blockers reduce exposure to tracking?

Corresponding technical questions

- Which of the HTTP requests sent by a user's browser **are used to** track the user?
- Which of the HTTP requests sent by a user's browser **allow** tracking the user?

HEURISTICS FOR DETECTING TRACKING

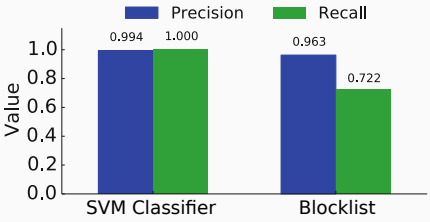
- Focus on cookies: HTTP requests that send a cookie value with an identifier allow tracking[19]
  - Discard session cookies and cookies with short expiration date as non-tracking
  - Parse keys/values stored in each cookie
  - Discard short values (< 8 characters) as non-tracking
  - Across experiment results: If values remain the same for each measurement instance but differ between instances, label as tracking (*difference* determined using threshold on similarity score)
- Focus on third parties: HTTP requests to third parties that are known trackers allow tracking[20]
  - Look up in public blocklist, e.g., EasyList or EasyPrivacy

[19] G. Acar, C. Eubank, S. Englehardt, *et al.*, "The Web Never Forgets: Persistent Tracking Mechanisms in the Wild," in *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*, ser. CCS '14, Scottsdale, Arizona, USA: ACM, 2014, pp. 674–689. DOI: 10.1145/2660267.2660347.

[20] S. Englehardt and A. Narayanan, "Online Tracking: A 1-million-site Measurement and Analysis," in *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, ser. CCS '16, Vienna, Austria: ACM, 2016, pp. 1388–1401. DOI: 10.1145/2976749.2978313.

MACHINE LEARNING FOR DETECTING TRACKING (1)

- Which HTTP requests allow tracking?[21]
  - Features based on all cookies received during one website visit
  - 3 most important features after recursive feature elimination:
    - Minimum cookie lifetime
    - Number of third-party cookies
    - Sum over value length * cookie lifetime
  - Train/test data: 500 requests each, half to tracking third-party, half to non-tracking third-party
  - Support Vector Machine, binary classifier
  - Blocklist has low recall: misses many trackers



[21] T.-C. Li, H. Hang, M. Faloutsos, *et al.*, "TrackAdvisor: Taking Back Browsing Privacy from Third-Party Trackers," in *Passive and Active Measurement*, J. Mirkovic and Y. Liu, Eds., ser. Lecture Notes in Computer Science, Cham: Springer, 2015, pp. 277–289

MACHINE LEARNING FOR DETECTING TRACKING (2)

- Which third parties are trackers?[22]
  - Extract text from website visits to tracker domain and search engine result page for "about <domain>"
  - Features are words, bi-grams, trigrams
  - Feature vector = token frequencies
  - Train/test data: 2000 domains from Alexa (non-tracking), 2000 domains from EasyList (tracking)
  - SVM binary classifier, precision: 0.95, recall: 0.95
  - Low overlap with blocklists: partly due to their inclusion of mobile tracking

	ATS overlap 2,121 (100%)
McAfee	451 (21.0%)
OpenDNS	780 (36.0%)
VirusTotal	1,081 (50.0%)
EasyList	818 (38.0%)
hpHosts	1,652 (77.0%)

[22] A. Razaghpanah, R. Nithyanand, N. Vallina-Rodriguez, *et al.*, "Apps, Trackers, Privacy, and Regulators: A Global Study of the Mobile Tracking Ecosystem," in *Proceedings 2018 Network and Distributed System Security Symposium*, San Diego, CA: Internet Society, 2018. DOI: 10.14722/ndss.2018.23353

Detecting anti-ad blockers: based on websites[23]

- Does this website use an anti-ad blocker?
- Training data: benign sites from Alexa, sites with anti-ad blockers from anti-ad block filter lists
- Both visited with and without ad blocker
- Features: differences between paired visits, e.g., changes in URL, changes in number of HTML tags, tag attributes, lines, words, characters, cosine similarity of entire HTML
- Random forest classifier achieves precision of 94%

[23] M. H. Mughees, Z. Qian, and Z. Shafiq, "Detecting Anti Ad-blockers in the Wild," *Proceedings on Privacy Enhancing Technologies*, vol. 2017, no. 3, pp. 130–146, Jul. 2017. DOI: *10.1515/popets-2017-0032*.

DETECTING ANTI-AD BLOCKERS: BASED ON JAVASCRIPT[24]

- Is this JavaScript snippet an anti-ad blocker?
- Training data: 1 million benign JS snippets, 372 anti-ad block scripts from filter lists
- 1.7 million features based on abstract syntax tree of JS
  - Text elements
  - Literals
  - JavaScript keywords
  - Keywords from JavaScript Web API
- Feature selection:
  - Remove duplicates
  - Remove features that do not vary much
  - Rank according to chi-square correlation
  - Select top-n features (e.g., n=1000)
- AdaBoost classifier achieves recall of 99%

[24] U. Iqbal, Z. Shafiq, and Z. Qian, "The Ad Wars: Retrospective Measurement and Analysis of Anti-adblock Filter Lists," in *Proceedings of the 2017 Internet Measurement Conference*, ser. IMC '17, London, United Kingdom: ACM, 2017, pp. 171–183. DOI: *10.1145/3131365.3131387*.

DETECTING AD PRICES[25]

- What are the prices in encrypted real-time bidding auctions?
- Two sources of ground truth:
  - Prices from cleartext auctions
  - Prices from researcher-run ad campaign
- 200+ features: time of auction, HTTP headers, ad content, DSP, publisher, user's interests and location
- Reduce features with dimensionality reduction and feature importance estimation
- Regression model: performs badly because prices are highly variable
- Random forest classifier achieves precision of 83% – predicting 4 price buckets instead of numeric price

[25] P. Papadopoulos, N. Kourtellis, P. R. Rodriguez, *et al.*, "If You Are Not Paying for It, You Are the Product: How Much Do Advertisers Pay to Reach You?" In *Proceedings of the 2017 Internet Measurement Conference*, ser. IMC '17, London, United Kingdom: ACM, 2017, pp. 142–156. DOI: *10.1145/3131365.3131397*.

DETECTING PII IN NETWORK TRAFFIC[26]

- Does an outgoing HTTP request contain personally identifiable information?
- Ground truth:
  - Traffic from 950 apps
  - Label instances of unique, researcher-controlled PII in traffic
- Features:
  - Split HTTP requests by delimiters: *„;/()[]*
  - Construct bag-of-words
  - Remove features with low frequency
- General-purpose decision tree classifier achieves accuracy of 80%, app-specific classifiers are better

[26] J. Ren, A. Rao, M. Lindorfer, *et al.*, "ReCon: Revealing and Controlling PII Leaks in Mobile Network Traffic," in *Proceedings of the 14th Annual International Conference on Mobile Systems, Applications, and Services*, ser. MobiSys '16, Singapore: ACM, 2016, pp. 361–374. DOI: 10.1145/2906388.2906392.

DETECTING PARKED DOMAINS[27]

- Is a given domain a parked domain?
- Training data:
  - 3,000 verified parked domains
  - 3,000 verified nonparked domains from Alexa list
- 21 features, including:
  - Average/maximum link length
  - Average HTML length
  - External link ratio
  - Presence of redirection mechanisms
- Random forest classifier achieves true positive rate of 97%

[27] T. Vissers, W. Joosen, and N. Nikiforakis, "Parking Sensors: Analyzing and Detecting Parked Domains," in *Proceedings 2015 Network and Distributed System Security Symposium*, San Diego, CA, USA: Internet Society, Feb. 2015. DOI: 10.14722/ndss.2015.23053.

DETECTING SENSITIVE WEBSITES[28]

- Is a website sensitive, and if so, what is its category?
- Sensitive categories in data protection sense: health, ethnicity, religion, sexual orientation, and political beliefs
- Training data: labeled URLs from Curlie (crowdsourced website taxonomy)
- Features:
  - Tf-idf representation of content and metadata of each website
  - Limited to top 5,000 features
- Multinomial Naive Bayes classifier achieves
  - F1 scores of 90% for nonsensitive websites
  - F1 scores between 55% and 91% for the topics of sensitive websites
- Note that content of a domain's landing page is not a reliable indicator

[28] S. Matic, C. Iordanou, G. Smaragdakis, *et al.*, "Identifying Sensitive URLs at Web-Scale," in *Proceedings of the ACM Internet Measurement Conference*, ser. IMC '20, Pittsburgh, PA, USA: Association for Computing Machinery, Oct. 2020, pp. 619–633. DOI: *10.1145/3419394.3423653*.

DETECTING MALICIOUS WEBSITES THAT USE CLOAKING[29]

- Cloaking (or black-hat search engine optimization):
  - Website shows different content depending on visitor (search engine spider vs human)
  - Search results look benign
  - But human visitor gets malicious content
- Training data:
  - List of cloaking domains, e.g., domains with counterfeit luxury storefronts
  - Website crawls with different configurations: mimicking search engines and humans
- Features:
  - Similarity of visible text
  - HTML, screenshots, embedded links
  - Request trees
- Decision tree classifier achieves accuracy of 95%

[29] L. Invernizzi, K. Thomas, A. Kapravelos, *et al.*, "Cloak of Visibility: Detecting When Machines Browse a Different Web," in *2016 IEEE Symposium on Security and Privacy (SP)*, San Jose, CA, USA: IEEE, May 2016, pp. 743–758. DOI: *10.1109/SP.2016.50*.

DETECTING MALICIOUS INCLUSIONS (E.G., MALVERTISING)[30]

- Third-party content inclusions can be malicious on benign websites
- Training data:
  - Top 200,000 sites from Alexa
  - Label resource loads as benign/malicious with VirusTotal's URL scanning service
- Features:
  - Type of top-level domain, number of subdomains, Alexa rank
  - Numbers of nonalphabetic characters, unique characters in domain name
  - Randomness in domain name, whether domain name belongs to ad network, CDN, URL shortener
- Imbalanced dataset: train separate Hidden Markov model (HMM) for each class, estimate HMM parameters with Baum-Welch algorithm
- HMM can model interdependencies between resources in inclusion sequence
- HMM classifier with higher likelihood determines label, achieves recall of 93%

[30] S. Arshad, A. Kharraz, and W. Robertson, "Include Me Out: In-Browser Detection of Malicious Third-Party Content Inclusions," in *Financial Cryptography and Data Security*, J. Grossklags and B. Preneel, Eds., ser. Lecture Notes in Computer Science, Springer Berlin Heidelberg, 2017, pp. 441–459.

DETECTING MALICIOUS APPS[31]

- Is a given app malicious or benign?
- Requested permissions indicate the *potential* to perform malicious operations
- Features: binary indicator for presence/absence of each permission
- Ranking of feature importance: mutual information or Pearson correlation coefficient
- Identifying most risky sets of permissions: principal component analysis
- Random forest classifier achieves true positive rate of 94%

[31] W. Wang, X. Wang, D. Feng, *et al.*, "Exploring Permission-Induced Risk in Android Applications for Malicious Application Detection," *IEEE Transactions on Information Forensics and Security*, vol. 9, no. 11, pp. 1869–1882, Nov. 2014. DOI: *10.1109/TIFS.2014.2353996*.

Natural language processing

- Preprocessing steps
  - Split into sentences
  - Convert to lower-case
  - Stemming (e.g., Porter-Stemmer): normalize word endings so that *walk*, *walks*, *walked*, and *walking* are recognized as the same word
- Representation as numerical feature vector
  - Bag-of-words: number of times each word occurs in a text
  - Construct vocabulary: all words that occur in a set of documents
  - Feature vector for one document: vector of word frequencies based on vocabulary
  - Word embeddings based on neural networks: see software packages Word2vec, GloVe, BERT, fastText, or Gensim

DETECTING TOPICS OF WEBSITES[32]

- Heuristics: online tagging services
- But: inaccurate if website has multiple categories, category may not be in tagging service, low coverage for less popular domains
- Topic modeling approach
  - List of topics: 1.932 Google AdWord categories
  - Corpus of document for each topic: top 10 Wikipedia articles for each topic
  - Preprocessing of corpus: extract 1,000 most relevant words (tf-idf)
  - Preprocessing of websites: extract visible text, metadata, remove stop words, apply stemming algorithm
- Matching score for each topic T on website W (K_i^T is i-th most relevant word for topic T): $\text{score}_T = \sum_{i=0}^{1000} \sum_{K_i^T \in W} \frac{1}{i}$
- Highest-scoring topic is website's inferred topic
- Accuracy evaluated with user study: 61%

[32] B. Weinshel, M. Wei, M. Mondal, *et al.*, "Oh, the Places You've Been! User Reactions to Longitudinal Transparency About Third-Party Web Tracking and Inferencing," in *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, ser. CCS '19, London, United Kingdom: ACM, 2019, pp. 149–166, DOI: 10.1145/3319535.3363200

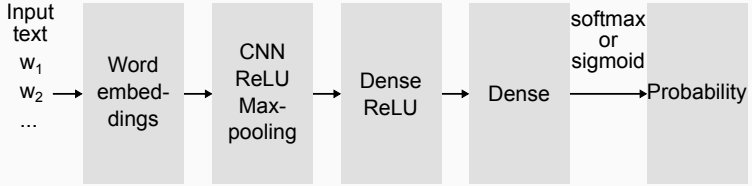
DETECTING TOPICS OF ADS: VIA LANDING PAGES[33]

- Idea: Landing page has more information than the ad itself, should be easier to label
- Training data: 1,000 sites from each Alexa category
- Alexa category is ground truth label – no need for manual labeling!
- Features: title, keywords in HTML header
- Bag-of-words with stemmed words and 2-grams
- Multi-class logistic regression achieves 76% accuracy

[33] S. Nath, "MAdScope: Characterizing Mobile In-App Targeted Ads," in *Proceedings of the 13th Annual International Conference on Mobile Systems, Applications, and Services*, ser. MobiSys '15, Florence, Italy: ACM, 2015, pp. 59–73. DOI: 10.1145/2742647.2742653.

DETECTING TOPICS OF ADS: VIA CATEGORIZATION[34]

- Does an ad belong to specific category or not?
- For example, is it a political ad or not?
- Continuous bag of words based on text in Facebook ads
- Convolutional Neural Network achieves 94% accuracy



[34]M. Silva, L. Santos de Oliveira, A. Andreou, *et al.*, "Facebook Ads Monitor: An Independent Auditing System for Political Ads on Facebook," in *Proceedings of The Web Conference 2020*, ser. WWW '20, Taipei, Taiwan: ACM, Apr. 2020, pp. 224–234. DOI: *10.1145/3366423.3380109*.

WHICH ATTRIBUTES ARE SENSITIVE?[35]

- Facebook assigns large number of *ad preferences* to each user from pool of 120,000+ preferences
- Some may be sensitive in data protection sense, e.g., ethnicity, health, sexual orientation, etc.
- To identify candidate sensitive ad preferences:
  - Compute semantic similarity between ad preferences and words from sensitive categories
  - List of words: use list of controversial issues on Wikipedia
  - If similarity score is above threshold (e.g., 0.6): human determines final label (sensitive or not)
- From 120,000 candidate ad preferences: 4,400 are potentially sensitive, 2,000 verified as sensitive by human expert

[35] J. G. Cabañas, Á. Cuevas, and R. Cuevas, "Unveiling and Quantifying Facebook Exploitation of Sensitive Personal Data for Advertising Purposes," in *27th USENIX Security Symposium (USENIX Security 18)*, Baltimore, MD, USA: USENIX Association, 2018, pp. 479–495.

WHICH APP REVIEWS ARE ABOUT PRIVACY/SECURITY?[36]

- Training data:
  - 4,000 Android app reviews
  - Label as security/privacy relevant is review contains keyword that indicates privacy/security issue (e.g., names of Android permissions, protected resources)
  - Preprocessing: remove stop words, apply stemming algorithm
- Features:
  - Bag-of-words based on character n-grams (to limit influence of typos)
- Imbalanced training data: use SMOTE to oversample smaller class
- SVM classifier achieves accuracy of 93%

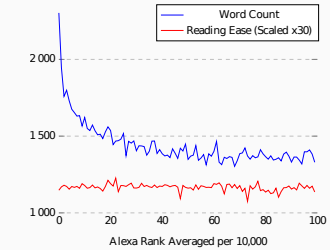
[36] D. C. Nguyen, E. Derr, M. Backes, *et al.*, "Short Text, Large Effect: Measuring the Impact of User Reviews on Android App Security & Privacy," in *IEEE Symposium on Security & Privacy*, San Francisco, CA, USA: IEEE, May 2019, p. 15.

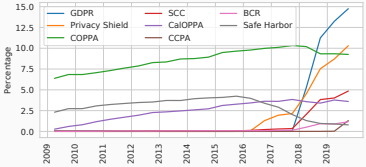
Research questions

- How usable are privacy policies?
- What do policies say about rights for users vs. for the website?
- What effect did the GDPR have on privacy policies on the web?
- Does a web service adhere to its own privacy policy?

HEURISTICS FOR ANALYZING PRIVACY POLICIES

- Usability: length and readability[37]
- Effect of new regulations: frequency of key words[38]
- Policy compliance: are third parties embedded in a website mentioned in the policy?



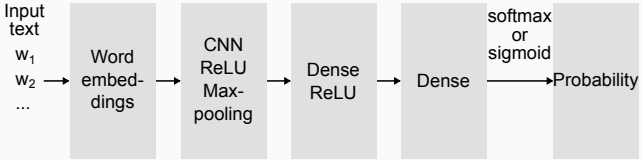


[37] T. Libert, "An Automated Approach to Auditing Disclosure of Third-Party Data Collection in Website Privacy Policies," in *Proceedings of the 2018 World Wide Web Conference*, ser. WWW '18, Lyon, France: International World Wide Web Conferences Steering Committee, 2018, pp. 207–216. DOI: 10.1145/3178876.3186087

[38] R. Amos, G. Acar, E. Lucherini, *et al.*, "Privacy Policies over Time: Curation and Analysis of a Million-Document Dataset," in *Proceedings of The Web Conference 2021*, ser. WWW '21, Ljubljana, Slovenia: ACM, Apr. 2021, p. 22. DOI: 10.1145/3442381.3450048. arXiv: 2008.09159

DETECTING PRIVACY POLICIES[39]

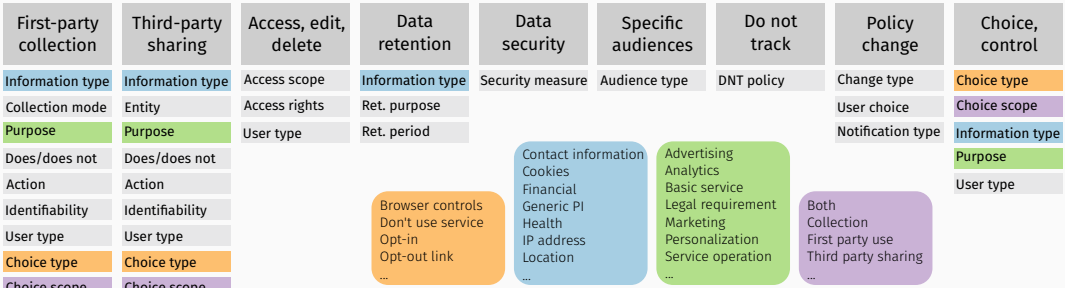
- Is given text a privacy policy?
- Training data: text from 1,000 privacy policies, text from landing pages of Alexa top 500 sites
- Features: tokenize text, use word embeddings
- Convolutional neural network architecture:



- CNN achieves 99% accuracy

[39] T. Linden, R. Khandelwal, H. Harkous, *et al.*, "The Privacy Policy Landscape After the GDPR," *Proceedings on Privacy Enhancing Technologies*, vol. 2020, no. 1, pp. 47–64, Jan. 2020. DOI: 10.2478/popets-2020-0004.

- Classify content of policy segments based on classifier hierarchy[40]
- Policy segmentation based on semantic relatedness graphs[41]



[40]H. Harkous, K. Fawaz, R. Lebret, *et al.*, "Polisis: Automated Analysis and Presentation of Privacy Policies Using Deep Learning," in *27th USENIX Security Symposium (USENIX Security 18)*, Baltimore, MD, USA: USENIX, 2018, pp. 531–548.

[41]G. Glavaš, F. Nanni, and S. P. Ponzetto, "Unsupervised text segmentation using semantic relatedness graphs," in **SEM 2016: The Fifth Joint Conference on Lexical and Computational Semantics : Proceedings of the Conference ; August 11-12 2016, Berlin, Germany*, C. Gardent, Ed., Stroudsburg, Pa.: Association for Computational Linguistics, 2016, pp. 125–130. [Online]. Available: `https://madoc.bib.uni-mannheim.de/41341`.

NATURAL LANGUAGE PROCESSING FOR ANALYZING PRIVACY POLICIES (CON'T)

- Training data: corpus of 115 privacy policies, labeled by law students (OPP-115)[42]
- Classifier: CNN (Convolutional Neural Network)[43] or BERT (Bidirectional Encoder Representations from Transformers)[44]
- Real-world dataset: collect policy texts from websites and Wayback Machine, 1996-2021
- CNN classifier to identify whether collected text is a privacy policy[45]

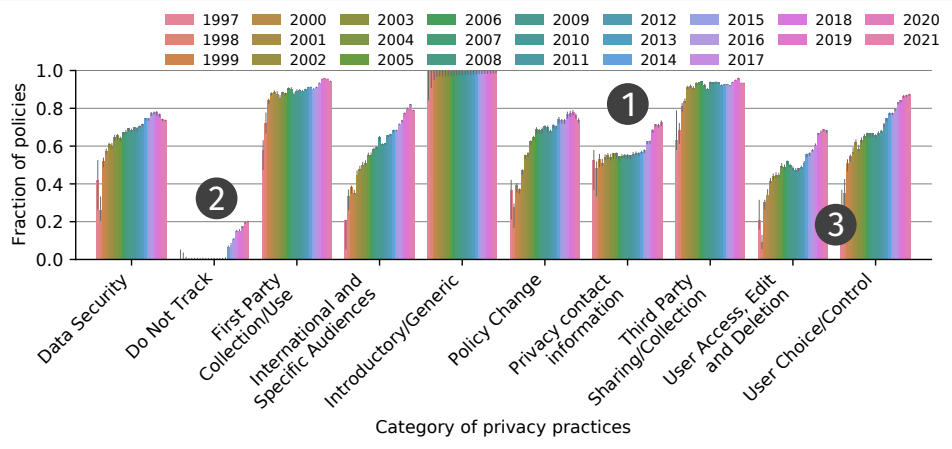
[42] S. Wilson, F. Schaub, A. A. Dara, *et al.*, "The Creation and Analysis of a Website Privacy Policy Corpus," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Berlin, Germany: Association for Computational Linguistics, 2016, pp. 1330–1340. DOI: 10.18653/v1/P16-1126.

[43] H. Harkous, K. Fawaz, R. Lebret, *et al.*, "Polisis: Automated Analysis and Presentation of Privacy Policies Using Deep Learning," in *27th USENIX Security Symposium (USENIX Security 18)*, Baltimore, MD, USA: USENIX, 2018, pp. 531–548.

[44] I. Wagner, "Privacy Policies Across the Ages: Content and Readability of Privacy Policies 1996–2021," *arXiv:2201.08739 [cs]*, Jan. 2022. arXiv: 2201.08739 [cs]. [Online]. Available: http://arxiv.org/abs/2201.08739.

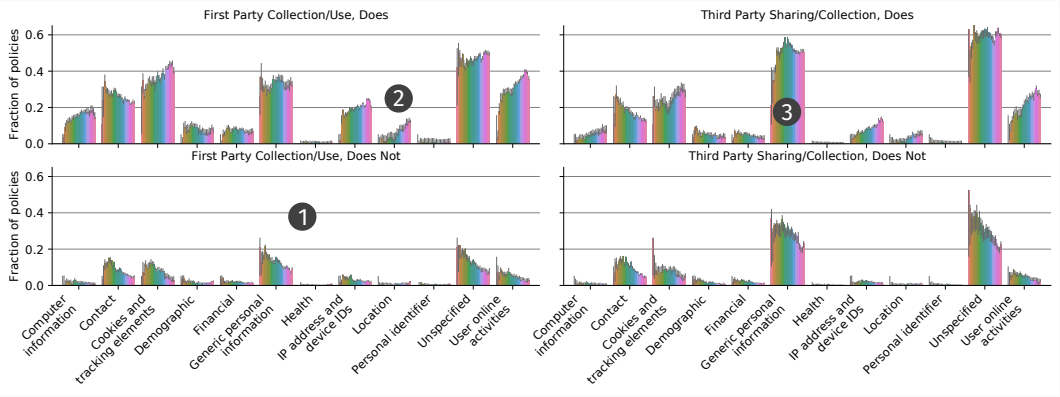
[45] T. Linden, R. Khandelwal, H. Harkous, *et al.*, "The Privacy Policy Landscape After the GDPR," *Proceedings on Privacy Enhancing Technologies*, vol. 2020, no. 1, pp. 47–64, Jan. 2020. DOI: 10.2478/popets-2020-0004.

RESULTS: CATEGORIES OF PRIVACY PRACTICES[46]



[46] I. Wagner, "Privacy Policies Across the Ages: Content and Readability of Privacy Policies 1996–2021," *arXiv:2201.08739 [cs]*, Jan. 2022. arXiv: *2201.08739 [cs]*. [Online]. Available: *http://arxiv.org/abs/2201.08739*.

RESULTS: FIRST-PARTY VS. THIRD-PARTY COLLECTION



DETECTING OPT-OUT CHOICES IN PRIVACY POLICIES[47]

- Training data: labeled privacy policies
- Active learning to reduce labeling effort:
  - Train classifier with small initial set of labeled policies
  - Classify policies not in training set, observe classifier confidence
  - Label additional policies and retrain classifier
  - Select additional policies based on entropy: labeling high-entropy policies (classifier is very uncertain) is most helpful
- Features: words and bigrams
- Logistic regression classifier achieves F1 score of 87%

[47] V. B. Kumar, R. Iyengar, N. Nisal, *et al.*, "Finding a Choice in a Haystack: Automatic Extraction of Opt-Out Statements from Privacy Policy Text," in *Proceedings of The Web Conference 2020*, ser. WWW '20, Taipei, Taiwan: Association for Computing Machinery, Apr. 2020, pp. 1943–1954. DOI: 10.1145/3366423.3380262.

Analysis of mobile apps

- Principle of taint tracking
  - Mark information from specific sources as *tainted* (e.g., API calls that access protected information)
  - Track use of this information through the app
  - Analyze whether information flows to problematic sink, e.g., the network
- Static analysis: inspect app metadata and bytecode without executing it[48]
  - Focus on possibility that tainted information flows to specific sink
- Dynamic analysis: execute app to analyze run-time behavior[49]
  - Actual occurrences of problematic information flows

[48] R. Binns, U. Lyngs, M. Van Kleek, *et al.*, "Third Party Tracking in the Mobile Ecosystem," in *Proceedings of the 10th ACM Conference on Web Science*, ser. WebSci '18, Amsterdam, Netherlands: ACM, 2018, pp. 23–31. DOI: *10.1145/3201064.3201089*.

[49] W. Enck, P. Gilbert, S. Han, *et al.*, "TaintDroid: An Information-Flow Tracking System for Realtime Privacy Monitoring on Smartphones," *ACM Trans. Comput. Syst.*, vol. 32, no. 2, 5:1–5:29, Jun. 2014. DOI: *10.1145/2619091*.

STATIC ANALYSIS: REACHABILITY ANALYSIS

- Can specific information flow to specific sinks?
- Sources: use of “dangerous” permissions, presence of privacy-sensitive API calls, tracking API calls
- Create control flow graph: for each API call, trace backward to find its source, forward to find its sinks
- Implementation in FlowDroid[50]
- Example reachability analyses:
  - Identify access to PII[51]
  - Identify API calls to embedded libraries[52]

[50] S. Arzt, S. Rasthofer, C. Fritz, *et al.*, “FlowDroid: Precise Context, Flow, Field, Object-sensitive and Lifecycle-aware Taint Analysis for Android Apps,” in *Proceedings of the 35th ACM SIGPLAN Conference on Programming Language Design and Implementation*, ser. PLDI ’14, Edinburgh, United Kingdom: ACM, 2014, pp. 259–269. DOI: *10.1145/2594291.2594299*.

[51] J. Gamba, M. Rashed, A. Razaghpanah, *et al.*, “An Analysis of Pre-installed Android Software,” in *2020 IEEE Symposium on Security and Privacy (SP)*, San Francisco, CA, USA: IEEE, May 2020, pp. 197–213. DOI: *10.1109/SP.2020.00013*.

[52] T. Book and D. S. Wallach, “A Case of Collusion: A Study of the Interface Between Ad Libraries and Their Apps,” in *Proceedings of the Third ACM Workshop on Security and Privacy in Smartphones & Mobile Devices*, ser. SPSM ’13, Berlin, Germany: ACM, 2013, pp. 79–86. DOI: *10.1145/2516760.2516762*.

- Principle: place hooks into API calls or system calls to trigger researcher-controlled analysis code when functions-of-interest are called
- Record parameter values of API calls: which information is given to ad or tracking libraries?[53]
- Taint tracking: identify taint sources + sinks, then track taint through execution
- How to track taint?
  - TaintDroid[54]: high-level concepts, i.e., variables within applications, messages passed between applications, method calls to native libraries, access to files
  - TaintMan[55]: low-level concepts, i.e., instructions on system level

[53] X. Liu, J. Liu, S. Zhu, *et al.*, "Privacy Risk Analysis and Mitigation of Analytics Libraries in the Android Ecosystem," *IEEE Transactions on Mobile Computing*, pp. 1–1, 2019. DOI: 10.1109/TMC.2019.2903186.

[54] W. Enck, P. Gilbert, S. Han, *et al.*, "TaintDroid: An Information-Flow Tracking System for Realtime Privacy Monitoring on Smartphones," *ACM Trans. Comput. Syst.*, vol. 32, no. 2, 5:1–5:29, Jun. 2014. DOI: 10.1145/2619091.

[55] W. You, B. Liang, W. Shi, *et al.*, "TaintMan: An ART-Compatible Dynamic Taint Analysis Framework on Unmodified and Non-Rooted Android Devices," *IEEE Transactions on Dependable and Secure Computing*, pp. 1–1, 2018. DOI: 10.1109/TDSC.2017.2740169.

COMBINED STATIC AND DYNAMIC ANALYSIS

- Static analysis for pre-filtering
  - Identify apps that use specific API calls
  - Observe results of these API calls in network traffic[56]
- Traffic analysis to trigger static
  - Observe traffic to detect PII
  - Use static analysis to find cause of PII transmission[57]
- Static analysis to guide dynamic analysis
  - Does call chain include third-party libraries?[58]
  - Is data transmission intended by user?[59]

[56] E. Pan, J. Ren, M. Lindorfer, *et al.*, "Panoptispy: Characterizing Audio and Video Exfiltration from Android Applications," *Proceedings on Privacy Enhancing Technologies*, vol. 2018, no. 4, pp. 33–50, Oct. 2018. DOI: *10.1515/popets-2018-0030*.

[57] J. Reardon, Á. Feal, P. Wijesekera, *et al.*, "50 Ways to Leak Your Data: An Exploration of Apps' Circumvention of the Android Permissions System," in *28th USENIX Security Symposium (USENIX Security 19)*, Santa Clara, CA, USA: USENIX, Aug. 2019, pp. 603–620.

[58] Y. He, X. Yang, B. Hu, *et al.*, "Dynamic privacy leakage analysis of Android third-party libraries," *Journal of Information Security and Applications*, vol. 46, pp. 259–270, Jun. 2019. DOI: *10.1016/j.jisa.2019.03.014*.

[59] Z. Yang, M. Yang, Y. Zhang, *et al.*, "AppIntent: Analyzing Sensitive Data Transmission in Android for Privacy Leakage Detection," in *Proceedings of the 2013 ACM SIGSAC Conference on Computer & Communications Security*, ser. CCS '13, Berlin, Germany: ACM, 2013, pp. 1043–1054. DOI: *10.1145/2508859.2516676*.

- Dynamic analysis is possible for any software that researchers can execute in controlled environment
- Leakage from Chrome browser extensions:[60]
  - Construct data and control flow graphs, link them to dynamic JS runtime objects
  - Propagate taint among dynamic objects, based on information from static analysis
- Detecting anti-ad blocker JavaScript:[61]
  - Record execution traces of JS code
  - Differences in execution traces indicate whether JS differentiates between sessions with/without ad blocker

[60] Q. Chen and A. Kapravelos, "Mystique: Uncovering Information Leakage from Browser Extensions," in *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*, ser. CCS '18, Toronto, Canada: ACM, 2018, pp. 1687–1700. DOI: 10.1145/3243734.3243823.

[61] S. Zhu, X. Hu, Z. Qian, *et al.*, "Measuring and Disrupting Anti-Adblockers Using Differential Execution Analysis," in *The Network and Distributed System Security Symposium (NDSS)*, San Diego, CA, USA, Feb. 2018. DOI: 10.14722/ndss.2018.23331.

Summary

- Many different tasks in data analysis
- Quantify interesting properties
- Use heuristics to extract interesting response variables from raw data
- Or: use machine learning and natural language processing for more precision and scalability
- Apply statistical methods to describe the data, test hypotheses, analyze causes
- Use static and dynamic analysis for code that can be executed by researchers

ABOUT THIS SLIDE DECK

- These slides are designed to accompany a lecture based on the textbook “Auditing Corporate Surveillance Systems: Research Methods for Greater Transparency” by Isabel Wagner, published in 2022 by Cambridge University Press.
-
-
- Please leave this slide intact, but indicate modifications below.
  - Version 2022-04
  - Improved version for release on book website (Isabel Wagner)
- Updated versions of the original slide deck are available online: corporatesurveillance.org