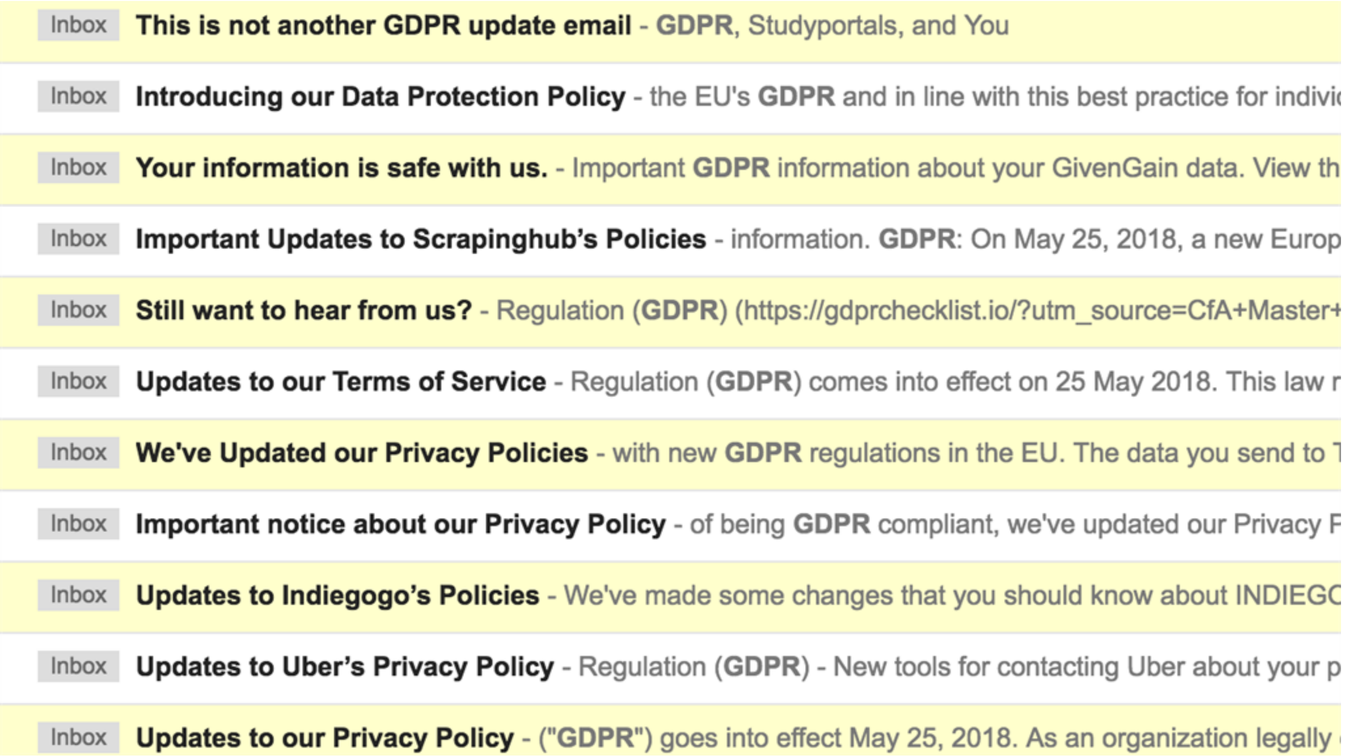
Analyzing Privacy Policies Using Contextual Integrity Annotations

Yan Shvartzshnaider, Noah Apthorpe, Nick Feamster, Helen Nissenbaum

Your Inbox on May 25 ...



Inbox **This is not another GDPR update email** - **GDPR**, Studyportals, and You

Inbox **Introducing our Data Protection Policy** - the EU's **GDPR** and in line with this best practice for indivi

Inbox **Your information is safe with us.** - Important **GDPR** information about your GivenGain data. View th

Inbox **Important Updates to Scrapinghub's Policies** - information. **GDPR**: On May 25, 2018, a new Europ

Inbox **Still want to hear from us?** - Regulation (**GDPR**) (https://gdprchecklist.io/?utm_source=CfA+Master+

Inbox **Updates to our Terms of Service** - Regulation (**GDPR**) comes into effect on 25 May 2018. This law r

Inbox **We've Updated our Privacy Policies** - with new **GDPR** regulations in the EU. The data you send to T

Inbox **Important notice about our Privacy Policy** - of being **GDPR** compliant, we've updated our Privacy P

Inbox **Updates to Indiegogo's Policies** - We've made some changes that you should know about INDIEGO

Inbox **Updates to Uber's Privacy Policy** - Regulation (**GDPR**) - New tools for contacting Uber about your p

Inbox **Updates to our Privacy Policy** - ("**GDPR**") goes into effect May 25, 2018. As an organization legally

Problem

- Privacy policies are
  - Lengthy …
  - Hard to parse …
  - Written with legal lingo …
  - Hard to compare across versions …

Dima Yarovinsky, I AGREE, http://vizknowledge.aalto.fi/showcase/

Previous work

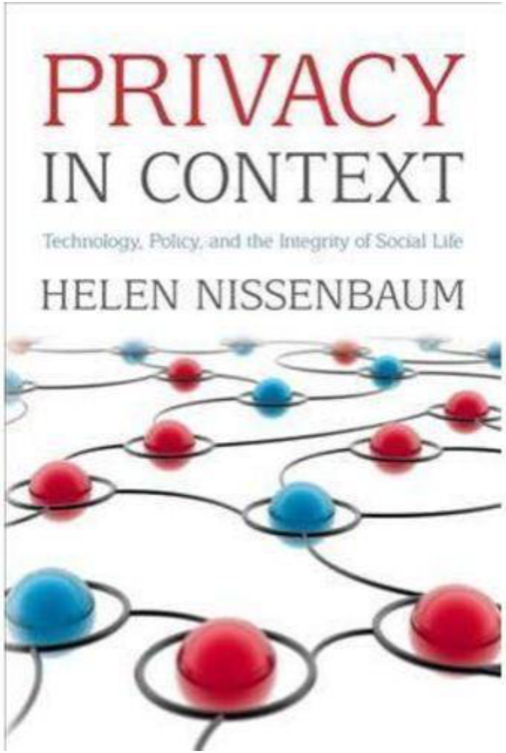
- Use NLP, ML to perform lexical and semantic analysis of privacy policy text
- Terms-of-service tracker
  - Tracking changes in policies
- Crowdsourcing and ranking privacy statements

USABLE PRIVACY.ORG
the usable privacy policy project

Terms of Service
Didn't Read

Methodology

- Use the CI framework to annotate policy statements that describe contextual information exchanges
    - **Sender.** Any entity (person, company, website, device, etc.) that transfers or shares the information.
    - **Recipient.** Any entity (person, company, website, device, etc.) that ultimately receives the information.
    - **Transmission principle.** Any clause describing the "terms and conditions under which [...] transfers ought (or ought not) to occur"
    - **Attribute.** Any description of information type, instance
    - **Subject.** Any subjects of the information exchanged in a flow. Subjects may be explicitly stated or implicitly described using pronouns and possessives.

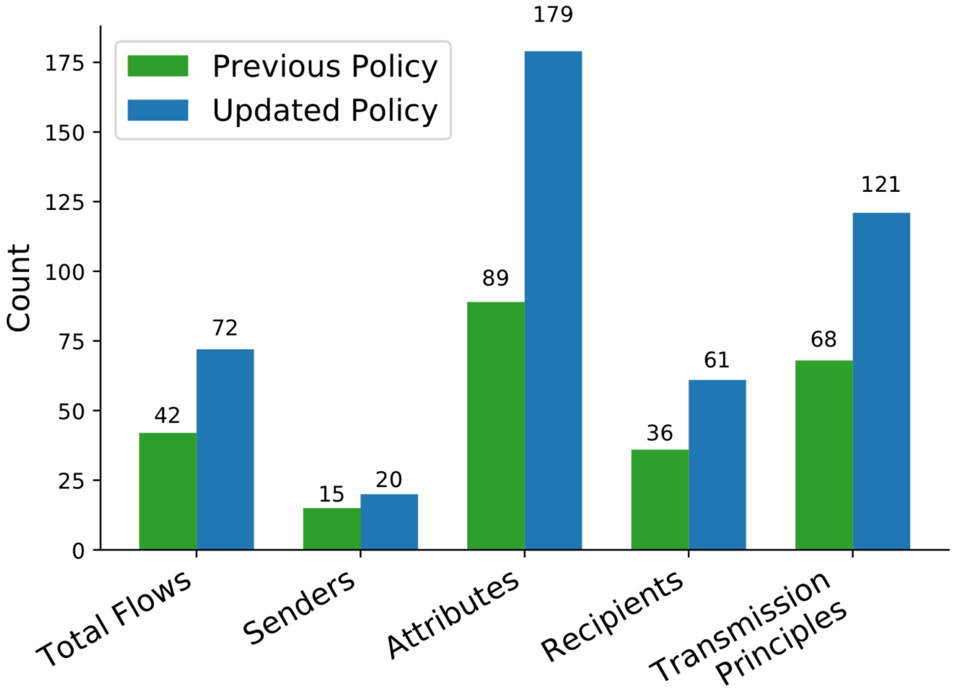


Analysis

- Compare CI parameters between privacy policies
- Identify incomplete information flows
  - Missing one or more parameters
- Identify information flows suffering from “CI parameter bloating”
  - Multiple CI parameters of the same type in the same flow
- Identify vague and ambiguous flows

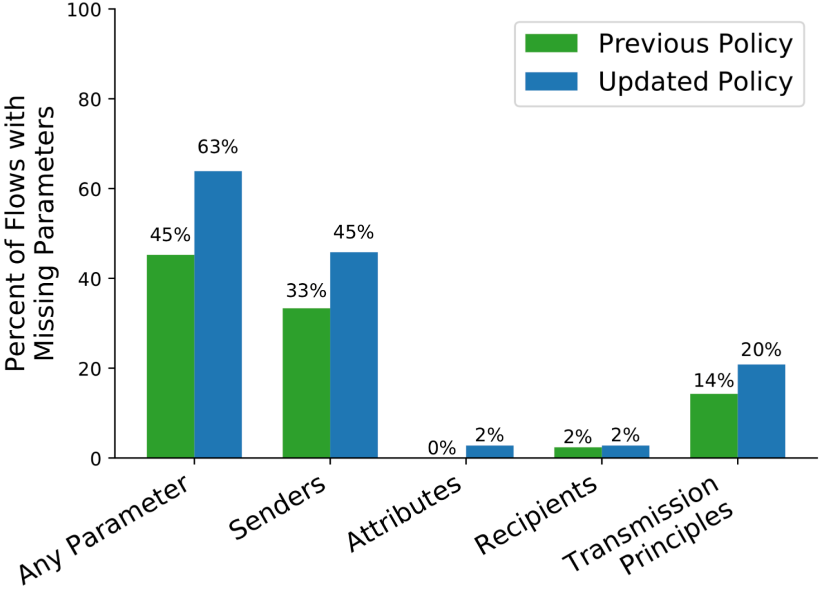
Facebook Case Study

- Use methodology to annotate and analyze the previous and updated versions of Facebook’s privacy policy
- Increase in the description of number of information flows
- **More information flows does not mean more clarity!**

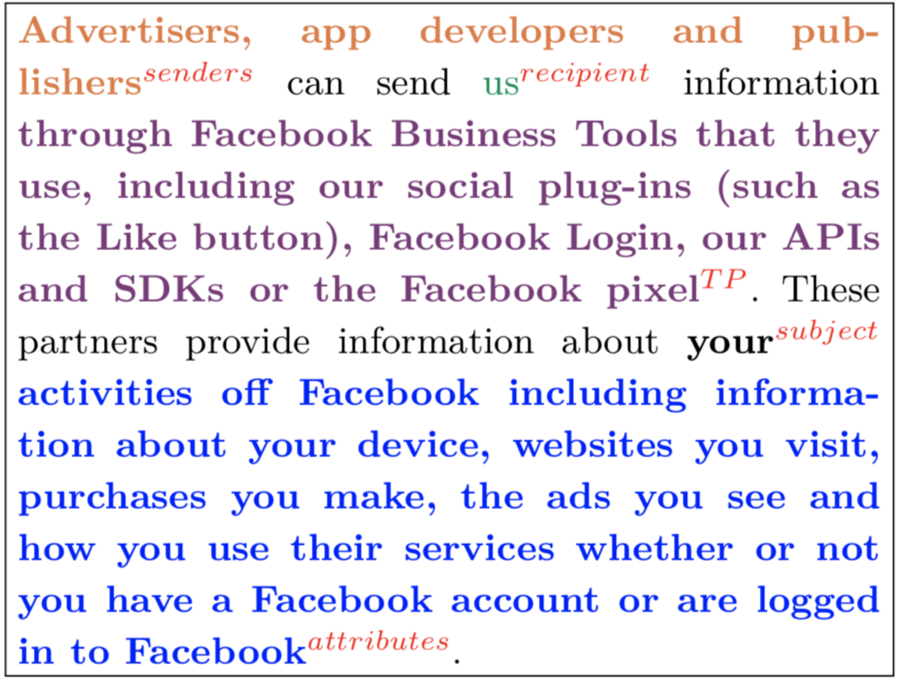


Analysis: Incomplete Information Flows

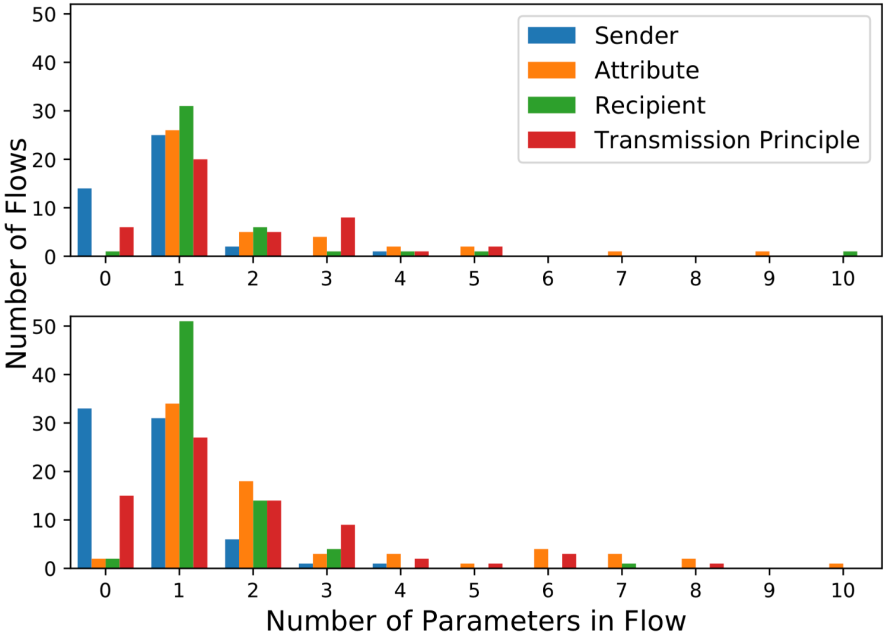
- Previous policy
  - **45%** (19/42) of flows are missing one or more parameters.
- Updated policy
  - **68%** (49/72) of flows are missing one or more parameters.
- **Failing to specify parameters introduces ambiguity, leaving consumers un-informed about company behavior.**



Analysis: CI Parameter Bloating



Advertisers, app developers and publishers^{senders} can send us^{recipient} information through Facebook Business Tools that they use, including our social plug-ins (such as the Like button), Facebook Login, our APIs and SDKs or the Facebook pixel^{TP}. These partners provide information about **your**^{subject} activities off Facebook including information about your device, websites you visit, purchases you make, the ads you see and how you use their services whether or not you have a Facebook account or are logged in to Facebook^{attributes}.

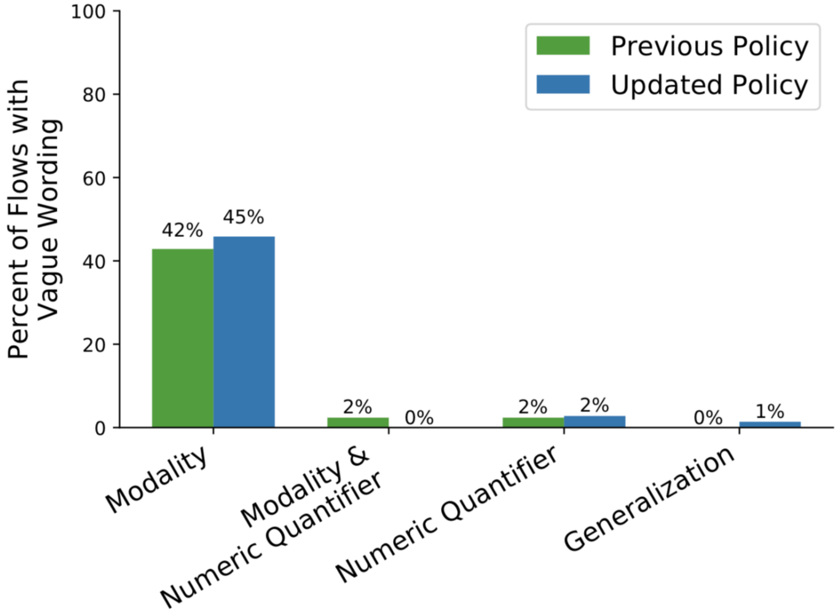


Analysis: Vague and Ambiguous Flows

- We identify information flows that use vague terminology as defined by Bhatia, et al.

- In both policies, “modality” vagueness dominates, occurring in close to **45%** of all flows.

- No reduction in vague terminology from previous to updated version.



J. Bhatia, T. D. Breaux, J. R. Reidenberg, and T. B. Nor- ton. A theory of vagueness and privacy risk perception. In Requirements Engineering Conference (RE), 2016 IEEE 24th International, pages 26–35. IEEE, 2016.

Crowdsourcing Annotations

- Constructed CI annotation as an Amazon Mechanical Turk task
- 99 out of 143 crowdworkers passed a set of 3 screener questions
- Crowdworkers annotated 48 policy excerpts
  - 16 excerpts from the pre-GDPR Google policy
  - 26 excerpt pairs from pre-GDPR and post-GDPR privacy policies of 16 well known companies (Amazon, Fitbit, The New York Times, Microsoft, etc.)
- Final “majority vote” annotation assigns each word in an excerpt to the CI parameter annotated by at least 50% of crowdworkers presented with that excerpt

Annotation Accuracy

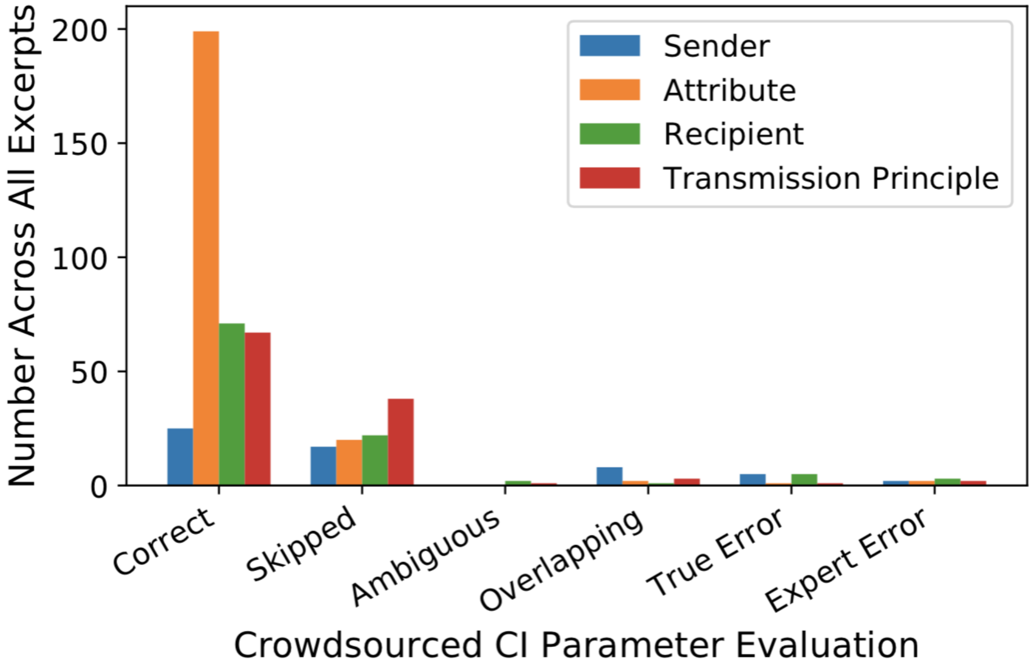
- Majority vote annotations **correctly** labeled
  - **43%** of senders
  - **89%** of attributes
  - **68%** of recipients
  - **60%** of transmission principles

- False **negatives**
  - **30%** of senders
  - **9%** of attributes
  - **21%** of recipients
  - **34%** of transmission principles

- False **positives**
  - **26%** of senders,
  - **11%** of recipients,
  - **2%** of attributes
  - **6%** of transmission principles

Evaluating Crowdworker Errors

- Expert Errors
  - 11 cases where “ground truth” expert annotation was incorrect
- True Errors
  - 13 incorrectly labeled parameters
- Skipped Parameters
  - 117 unlabeled parameters
- Ambiguous Parameters
  - 3 cases where correct annotation was ambiguous
- Overlapping Parameters
  - 16 cases where a word contributed to multiple parameters



Discussion

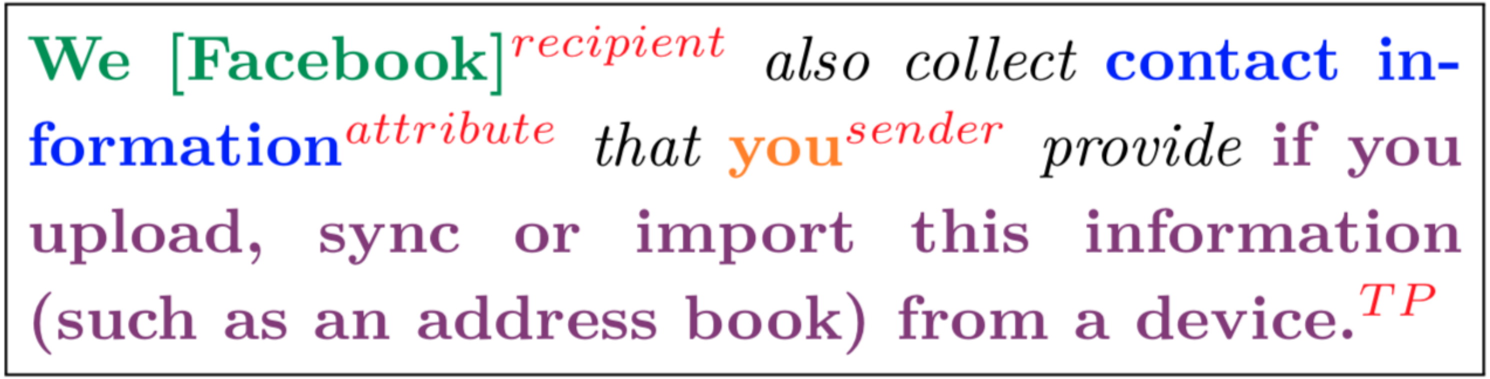
- Privacy policies are not written to intentionally fit the CI framework
  - Our crowdsourcing annotations showed promising results on a diverse privacy statements from privacy policies of 17 companies.
- Our annotation methodology deals only with statements describing information transfers
  - Annotating other statements will require additional methodologies to complement our approach

Conclusion

- The notion of an appropriate information flow in the CI framework lends itself well to user data privacy policies
- CI annotation is a stepping stone in a larger effort to improve readability and increase transparency in disclosure of information handling practices
- **Future goal:** produce a large corpus of privacy policies annotations to discover trends in within and across industries

Methodology: Example

- Annotate privacy statement and analyse the prescribed information flows using the theory of contextual integrity



We [Facebook]^{recipient} *also collect* **contact information**^{attribute} *that* **you**^{sender} *provide* **if you upload, sync or import this information (such as an address book) from a device.**^{TP}