Counterfactual Metrics for Auditing Black-Box Recommender Systems for Ethical Concerns

Nil-Jana Akpinar* [1], Liu Leqi* [1], Dylan Hadfield-Menell[2], Zachary Lipton[1]
[1] Carnegie Mellon University, [2] EECS CSAIL, MIT, * Equal contribution

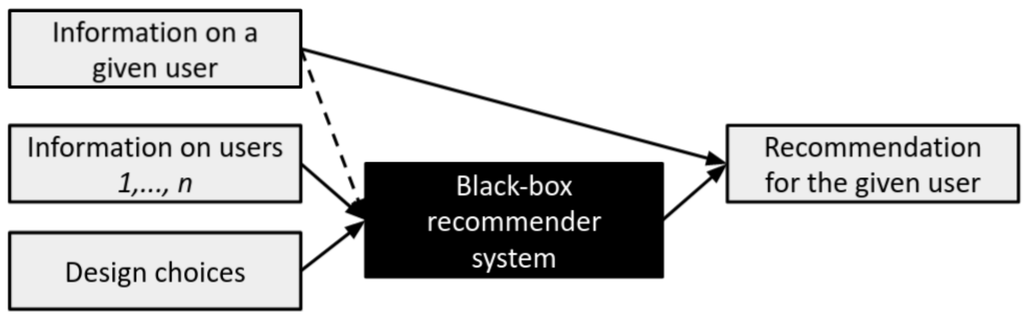

Figure 1:Dependency graph of the recommendation for a given user.

Contributions

Motivation

Most works audit recommender systems by exclusively looking at the output recommendations. Yet, outputs depend on a multitude of factors (Figure 1). How do we know if potential harm is *caused by the recommender system* rather than other factors?

Contributions

We propose counterfactual metrics for harm auditing which are based on an interventional perspective asking how recommendations would change if the information on one or several users was different. This allows us to disentangle the effects of the recommendation algorithm we want to audit from the impact of other factors.

Why are entirely observational metrics problematic?

Example of diversity

- Suggesting a wide variety of content is regarded as desirable because it honors users' multi-faceted interests, avoids algorithmic profiling, and avoids increasingly narrow recommendations which can facilitate a filter bubble problem.
- Diversity is generally measured in terms of observational quantities such as inverse similarity in recommendation slates.
- But what if a user's interests are truly narrow? Picture a researcher who is using Twitter only for academic purposes and is only recommended academic content. Is the algorithm really the culprit for narrow recommendations? Is there any harm inflicted?
- We need to control for user preferences and other external factors to measure potential harm inflicted by recommendation algorithms.

How do counterfactual metrics work?

Setting

- Training data $\mathcal{D} = \{\tau_1, \ldots, \tau_n\}$ where τ_i is the user information (e.g. ratings, demographics, reviews, etc.) for user $i \in [n]$.
- For any user with information τ which may or may not belong to the training data, the recommender system $\mathcal{A}$ outputs the individual's next step recommendation $\mathcal{A}(\mathcal{D}, \tau) \in \mathbb{C}$ where $\mathbb{C}$ is a finite set of all recommendations.
- We refer to $\mathcal{A}(\cdot, \cdot)$ as the recommender system and $\mathcal{A}(\mathcal{D}, \cdot)$ as the recommendation policy under training data $\mathcal{D}$.

Steps to obtain counterfactual metric

- Decide treatment space $\mathcal{W}$ that contains permissible new training data sets (e.g. add or remove user, add or delete ratings).
- Outcome of interest is next-step recommendation Y. We formally assume a random treatment W. $Y^w = \mathcal{A}(w, \tau)$ denotes the recommendation the user with information τ would have obtained if the recommender were trained using data set w. We assume consistency, i.e. $Y = Y^w$ if $W = w$.
- Define counterfactual metric on the potential outcome Y^w.

Reachability and Stability

Definition 1 (Individual-level Reachability) We say an item is reachable by a user if there is an allowable modification to their rating history that causes the item to be recommended [2]. Let τ be a vector of length $|\mathbb{C}|$ where $\tau_j[k]$ denotes j's rating of item k if available. To audit whether item k is reacheable by user j with $\tau_j \in \mathcal{D}$, we consider:

- Treatment space $\mathcal{W} = \{\tau' : \sum_{t \in [|\mathbb{C}|]} \mathbf{1}\{\tau_j[t] \neq \tau'[t]\} \leq B\}$ contains new user information that deviates from τ_j by at most B entries, where $B \in \mathbb{N}$ is a pre-specified budget.
- The outcome of interest is $Y^w = (\mathcal{D}', w)$ where $w \in \mathcal{W}$ and $\mathcal{D}' = \{\tau_1, \ldots, \tau_{j-1}, w, \tau_{j+1}, \ldots, \tau_n\}$.
- The reacheability metric is defined to be

$$\max_{w \in \mathcal{W}} \mathbb{P}_{\mathcal{A}} \left(Y^w = k\right),$$

which gives the maximal probability for user j to *reach* item k by modifying their own information τ_j. We note that $\mathbb{P}$ is used to indicate stochasticity in $\mathcal{A}$.

Definition 1 (Individual-level Stability) We propose this metric to capture how stable a user's recommendations are to other users' behaviors. In settings in which we are interested in user j's recommendation, we specify the following:

- Treatment space $\mathcal{W} = \{\mathcal{D}' : \mathcal{D}'$ differs from $\mathcal{D}$ for at most B users and τ_j remains unchanged$\}$, where $B \in \mathbb{N}$ is a pre-specified budget.
- The outcome of interest is $Y^w = (w, \tau_j)$ where $w \in \mathcal{W}$.
- The stability metric is defined to be

$$\max_{w \in \mathcal{W}} d\left(\mathcal{A}(\mathcal{D}, \tau_j), Y^w\right),$$

where $d : \mathbb{C} \times \mathbb{C} \to \mathbb{R}_+$ is a pre-speficied measure of distance between two recommendations.

Ethical Considerations

Previous work identified a taxonomy of ethical concerns for recommender systems: inappropriate content, privacy, autonomy & personal identity, opacity, fairness, and wider social effects [3].

Setting our metrics in context: Reachability

- Measure of users' agency over their recommendations which is relevant for autonomy and personal identity.
- Missing reachability could point towards over-categorization of users. Categories do often not align with recognizable social attributes which can lead to negative user experience [1].

Setting our metrics in context: Stability

- Connection to user autonomy and non-comparative fairness.
- If changes in user information drastically change the outputs for an unrelated user, we argue the user is not granted sufficient autonomy over their recommenadations. In addition, the arbitrariness of outputs is related to ideas around leave-one-out unfairness [4].

References

[1] Matthias Leese. "The new profiling: Algorithms, black boxes, and the failure of anti-discriminatory safeguards in the European Union". In: *Security Dialogue* 45.5 (2014).
[2] Sarah Dean, Sarah Rich, and Benjamin Recht. "Recommendations and User Agency: The Reachability of Collaboratively-Filtered Information". In: FAT*. 2020.
[3] Silvia Milano, Mariarosaria Taddeo, and Luciano Floridi. "Recommender systems and their ethical challenges". In: *AI Soc.* 35.4 (2020).
[4] Emily Black and Matt Fredrikson. "Leave-One-out Unfairness". In: FAccT. 2021.

tinyurl.com/counterfactual-audits nakpinar@andrew.cmu.edu leqil@andrew.cmu.edu