The impact of cloaking digital footprints on user privacy and personalization

THE IMPACT OF CLOAKING DIGITAL FOOTPRINTS ON USER PRIVACY AND PERSONALIZATION

Sofie Goethals
University of Antwerp
Antwerp, Belgium
sofie.goethals@uantwerpen.be

Sandra Matz
Columbia Business & Engineering Schools
New York, USA

Foster Provost
NYU Stern
New YOrk, USA

Yanou Ramon
University of Antwerp
Antwerp, Belgium

David Martens
University of Antwerp
Antwerp, Belgium

ABSTRACT

Our online lives generate a wealth of behavioral records -*"digital footprints"*- which are stored and leveraged by technology platforms. This data can be used to create value for users by personalizing services. At the same time, however, it also poses a threat to people's privacy by offering a highly intimate window into their private traits (e.g., their personality, political ideology, sexual orientation). Prior work has proposed a potential remedy: The cloaking of users' footprints. That is, platforms could allow users to hide portions of their digital footprints from predictive algorithms to avoid undesired inferences. While such an approach has been shown to offer privacy protection in the moment, there are two open questions. First, it remains unclear how well cloaking performs over time. As people constantly leave new digital footprints, the algorithm might regain the ability to predict previously cloaked traits. Second, cloaking digital footprints to avoid one undesirable inference may degrade the performance of models for other, desirable inferences (e.g., those driving desired personalized content). In the light of these research gaps, our contributions are twofold: 1) We propose a novel cloaking strategy that conceals "metafeatures" (automatically generated higher-level categories) and compares its effectiveness against existing cloaking approaches, and 2) we test the spill-over effects of cloaking one trait on the accuracy of inferences on other traits. A key finding is that the effectiveness of cloaking degrades over times, but the rate at which it degrades is significantly smaller when cloaking metafeatures rather than individual footprints. In addition, our findings reveal the expected trade-off between privacy and personalization: Cloaking an undesired trait also partially conceals other desirable traits.

***K*eywords** Privacy, Personalization, Explainable AI, Digital Footprints

1 Introduction and related work

A growing portion of our life happens online: We shop on Amazon, entertain ourselves on Netflix or Spotify and communicate with friend and family via Facebook or Whatsapp. Whether we like it or not, the digital traces we generate during these interactions provide the mediating platforms with an extensive and comprehensive picture of our personal habits and preferences [Matz et al., 2020, Kosinski et al., 2013]. In fact, research has shown that a person's digital footprints - including their Facebook Likes and status updates, smartphone records or credit card spending - can be used to predict highly intimate characteristics such as sexual or political orientation, personality traits, mental health, or religious views [Kosinski et al., 2013, Matz et al., 2017]. Given that most individuals consider these characteristics deeply private, the automated predictions of such traits without individuals knowledge or consent raises important concerns related to people's right to privacy and self-determination [Matz et al., 2020]. The act of drawing highly intimate inferences from seemingly innocuous data, for example, can be regarded as a intrusion of privacy, especially when individuals are neither aware of such inferences being made nor able to object to them. Moreover, the

psychological insights platforms (and other third parties) can glean from digital footprints allow them to influence their users' behaviors and decisions through mechanisms of personalization (an approach known as psychological targeting [Matz et al., 2017]).

On the one hand, such personalization approaches might be appreciated by consumers who receive more relevant products and services as a result of targeted advertising and product design [Tran, 2017]. On the other hand, the ability to predict people's intimate traits and influence their behavior raises serious concerns for individuals and society at large. In countries where homosexuality is illegal, for example, the ability to infer sexual orientation from Facebook Likes could become a death sentence [Cabañas et al., 2018]. Similarly, health assurance companies could attempt to identify people with unhealthy habits or specific health problems, resulting in higher premiums or even rejection of coverage altogether [Cabañas et al., 2018]. The perhaps most well-known case of such an abuse is that of Cambridge Analytica, the UK-based PR firm which claimed to have interfered in the 2016 US presidential election by targeting voters with psychologically-tailored advertising on Facebook [Appel and Matz, 2021].

Given the seriousness of these potential transgressions and privacy violations by more malicious actors, scientists, activists and policy makers have pushed for legislation that aims to prohibit the prediction of protected categories, such as race or religion. Facebook, for example, has faced years of criticism for offering advertisers 'interest' categories that have led to the exclusion of people of color from housing ads, fueled political polarization, and helped Big Pharma track users with specific illnesses [Waller, Angie and Lecher, Colin, 2022, Angwin and Parris Jr., 2016, Edelman, Gilad, 2019, Lecher, Colin, 2021]. In 2022, Facebook responded to the growing public pressure and changing regulatory landscape by removing the option to target users explicitly based on potentially sensitive traits such as health, race, sexual or political orientation, and religious beliefs [Waller, Angie and Lecher, Colin, 2022].[1]

However, as the non-profit news organization The Markup reported, Facebook's attempts at better protecting their users' privacy and preventing discrimination were only partially successful. For example, although *Hispanic Culture* was removed from the target categories available to advertisers, *Spanish Language* was not [Waller, Angie and Lecher, Colin, 2022]. Although Facebook has since removed additional interests and categories related to protected traits, we argue that playing whack-a-mole across many millions of pages and categories is destined to fail. This is partially the case because few users for which a protected trait is predicted, will actually like pages that explicitly reveal these traits. For example, less than 5% of users predicted as homosexual were connected with explicitly homosexual pages such as *No H8 Campaign*, *Being Gay* or *I Love Being Gay* [Kosinski et al., 2013].

Consequently, the mere act of eliminating certain prediction categories from the platforms prediction or targeting engines is insufficient. This is particularly true when ads or content are still targeted based on data-driven prediction models (e.g., inferring psychological traits) rather than individual interests. Even if *Gay Pride* is removed as an explicit targeting category and *No H8 Campaign* is removed as a data item for prediction, algorithms may still learn to use other digital footprints to target content or ads that would appeal to gay individuals.

In order to seriously limit the predictability and use of sensitive traits across all users, platforms would have to ban an unreasonable number of pages from their inference algorithms, among them many seemingly neutral pages that will be hard to justify and would likely evoke concerns related to freedom of speech and expression. Moreover, implementing such paternalistic measures may undermine an individual's agency over what they choose to reveal about themselves. For example, should we force users to restrict their online identities if they feel perfectly safe and comfortable about their lifestyle and sexual orientation, and would be delighted to receive advertisements about Drag Shows? In addition, generic one-size-fits-all approaches to restricting certain aspects of online behavior can have negative consequences for socially relevant causes that would benefit from personalization and civic engagement: climate activists and medical researchers, for example, have pointed out that the changes to Facebook's targeting platform have severely limited their ability to reach relevant audiences [Waller, Angie and Lecher, Colin, 2022].

In this paper, we examine a more individualized approach that offers users more control and transparency over their online identities, and can be tailored to and by the individual: cloaking certain digital traces that are relevant for the prediction of a particular individual. Chen et al. [2017] propose a "cloaking device" that reveals to users the digital footprints without which the prediction model would not have made the inference and allows them to restrict inference procedures from using them. Let us consider a digital footprint to be a specific aspect of online behavior that is stored about the individual on a technology platform, such as a particular song listened to on Spotify, or a specific page "liked" on Facebook. *Cloaking* a digital footprint means removing it from the set of data considered by an algorithm drawing inferences. In the common case of a machine learning model in an AI inference system, where the digital footprints are the features used by the model, cloaking the digital footprint represented by feature x for user u would mean setting the value of x to whatever would be the value if the system had not saved that digital footprint for that user (for example,

[1] Some of the interest categories that will be no longer available include 'Gay Pride', 'Islamic Calendar', and 'Lung Cancer Awareness' [Silberling, Amanda, 2021, Waller, Angie and Lecher, Colin, 2022]. A comprehensive list of removed Facebook pages can be found here: `https://www.propublica.org/article/facebook-lets-advertisers-exclude-users-by-race`.

setting the feature value to zero as an indication that the user did not like the page in question). This cloaking could be implemented by the platform, by providing users with the option to choose which inferences to avoid. Alternatively, such transparency could guide users to better decide which data they feel comfortable sharing in the first place; however, for many systems, the digital footprints are the result of simply using the system, so such self-cloaking would involve restricting one's own behavior (e.g., consider songs listened to on Spotify as digital footprints).

The reason why we (and previous authors) focus on cloaking the underlying features, and not the particular inferences, is because the latter cannot protect the user from closely-related inferences in the future [Chen et al., 2017]. This is in line with the current advertising options on Facebook: as mentioned before, it is no longer possible to target people based on certain private traits (e.g., sexual orientation). Hence, advertisers have to rely on associated interests (e.g. Facebook likes that have been empirically shown to be related to a certain sexual orientation) if they want to target a specific private group. The previously proposed cloaking strategy has been shown to be effective in avoiding inferences at the time the cloaking takes place, with relatively little burden on the users. However, prior work has not investigated how effective the cloaking would be over time, as users continue to leave new digital traces over time.

Digital footprints are usually high-dimensional, sparse, fine-grained behavioral data, for which models normally draw on a large combination of different features as evidence for a possible inferenceRamon et al. [2021b]. Therefore, as our results indicate, using a cloaking strategy solely based on singular fine-grained features will not be sufficient, as people will continue to live their lives and behave similarly in the future [De Cnudde et al., 2020, Chen et al., 2017]. For example, our analysis reveals that when we take a snapshot at a later moment, more than 80% of the people whom the models would target as republican will be subject to targeting again in the future despite cloaking based on the fine-grained features. We investigate methods to enhance the longer-term efficacy of cloaking, by grouping the fine-grained features into metafeatures (higher-level feature representations) and cloaking these metafeatures instead. Our results show that this approach increases the longer-term effectiveness of cloaking considerably and thus enhances desired the privacy protection over time.

Importantly, the implications of cloaking digital footprints to reduce undesired inferences are not uniformly positive. As we discussed at the beginning of this paper, the same digital footprints may be used for different inferences and ultimate purposes. For example, a particular footprint might reveal not only sexual orientation but also the personality trait of Openness. While a user might be concerned about their data being used to predict their sexual orientation and subsequently discriminate against them, they might be appreciative of personalized services and ads that account for their level of openness to experience. That is, the same traces and mechanisms that may lead to discrimination, can also benefit users in the form of personalized content (e.g., individualized playlists, more relevant news etc.). This desired form of personalization not only leads to happier users but also higher engagement [Fernández-Loría et al., 2017] and ultimately higher platform revenue[Johnson et al., 2020].

In this paper we study the potential unintended consequences of cloaking. That is, when users decide to suppress certain digital footprints via cloaking mechanisms, the data available for other desired personalization tasks decreases, potentially reducing their accuracy and effectiveness via spill-over effects. To explore this privacy-personalization trade-off, we evaluate the impact of the two previously outlined cloaking strategies (using fine-grained features versus metafeatures) on the accuracy of unrelated prediction tasks that are not subject to cloaking. For example, we examine how cloaking for sexual orientation impacts the predictive performance of a model predicting personality using the same set of digital footprints. Insights into the nature of this trade-off are crucial to empower users to make informed decisions about their online activity and about where on the privacy-personalization trade-off they want to be.

In summary, our study offers three major contributions to the existing literature:

- We assess the longer-term effectiveness of cloaking digital footprints, measuring the percentage of targeted individuals whose privacy remains protected over time. The results show that the effectiveness of cloaking fine-grained features decreases steadily and markedly over time for most inference tasks.

- We introduce a new type of cloaking strategy based on metafeatures, and show that it enhances longer-term cloaking protection (as intended).

- We examine the privacy-personalization trade-off inherent in using cloaking to protect against unwanted inferences. Specifically, we show that cloaking for one task can affect the predictive performance of other personalization tasks. Moreover, the metafeature-based strategies affect other tasks more, highlighting the trade-offs faced by users: better longer-term privacy protection indeed can reduce desired personalization performance

2 Data

We use data from the MyPersonality project, which contains the liked Facebook pages of 220,489 volunteers in the United States, along with their scores on the Big 5 personality traits and some personal characteristics such as gender, age, sexual orientation and political preferences [Kosinski et al., 2013]. A Facebook like is a mechanism used by Facebook users to express their positive association with online content, and in this case we focus on the public pages they liked, which can relate to products, public persons, music, sport, books, restaurants, or public statements they agree with. Using this data, it is possible to create a user-like matrix X such that $x_{ij} = 1$ if user i liked page j. Behavioral datasets, such as Facebook likes, are usually very sparse as every user can only take a limited number of actions (in this case like Facebook pages), while the total number of possible actions is very large [Junqué de Fortuny et al., 2013]. As described in more detail below, we assess the impact of cloaking the likes that lead to the inferences of gender, political orientation and sexual orientation.[2] In this study, we use these as examples of the attributes individuals might wish to safeguard; the specific attributes deemed private of course will vary depending on the individual's preferences. The data is described in Table 1.

Table 1: Data description for the target variables that will be cloaked. We select only the instances that have a value for the correspondent trait.[3]The features are the Facebook pages that remain after pre-processing. *Active elements* shows the non-zero elements in the entire matrix; *Sparsity* is the percentage of active elements over the total number of elements in the matrix. *Average likes* is the average number of likes a person associated with this trait has. *Balance* is the percentage of instances with a positive value for the target variable.

Target variable	Instances	Features	Active elements	Sparsity (in %)	Average likes	Balance (in %)
Male	165,234	115,326	16,901,459	99.91	86.8	38.37
Female	165,234	115,326	16,901,459	99.91	112.0	61.63
Homosexual	22,477	115,326	2,197205	99.92	104.3	4.67
Lesbian	29,309	115,326	4,041,148	99.88	110.5	2.65
Democrat	36,534	115,326	4,190,576	99.90	134.0	17.27
Republican	36,534	115,326	4,190,576	99.90	124.2	10.24

Personality traits Trait models suggest that personality consists of a range of consistent and relatively stable characteristics (traits) that determine how an individual will think, feel and behave [Matz et al., 2016]. The Big 5 (BF) Model of Personality is the most widely accepted model and proposes five independent traits to capture individual personality differences [Costa and McCrae, 1992, Matz et al., 2016]. The five traits are: 1) *Extraversion*, the tendency to seek excitement and stimulation in the company of others, 2) *Openness*, the tendency to be intellectually curious, creative and unconventional, 3) *Neuroticism*, the tendency to experience negative emotions, and being anxious and nervous 4) *Agreeableness*, the tendency to be trusting, compassionate and cooperative 5) *Consciousness*, the tendency to be organized and efficient [Matz et al., 2016, Ramon et al., 2021a]. The Big 5 personality traits were established using the international Personality Item Pool (IPIP) questionnaire with 20 items [Goldberg et al., 2006, Kosinski et al., 2013]. The traits are recorded on a 5-point Likert scale, and we report the data description in Table 2. Research shows that digital footprints have a predictive power over personality traits which is in line with the typical strength of the relationship between personality and behavior, also known as the *personality coefficient* (a correlation between 0.30 and 0.40) [Meyer et al., 2001, Azucar et al., 2018].

[2]Only gender is still available as an explicit targeting option on Facebook, but machine learning models can still learn the other traits implicitly when optimizing a particular ad or content element.

[3]For the prediction task of homosexuality, only men whose data record has a value for sexual orientation will be considered, while for the prediction task of lesbian, only women with a value for sexual orientation will be taken into account.

Table 2: Data description of the Big 5 personality traits. We select only the instances that have a value for the corresponding trait.

Personality trait	Instances	Features	Active elements	Sparsity (in %)	Average score
Extraversion	137,529	115,326	14,513,946	99.91	3.55
Openness	137,529	115,326	14,513,946	99.91	3.88
Neuroticism	137,529	115,326	14,513,946	99.91	2.79
Agreeableness	137,529	115,326	14,513,946	99.91	3.54
Consciousness	137,529	115,326	14,513,946	99.91	3.42

3 Cloaking methods

Cloaking mechanism As described above, *cloaking* refers to the mechanism of changing user data so that—from the perspective of the inference procedure—it was as if the user did not exhibit this behavior. The cloaking mechanism introduced by Chen et al. [2017] relies on counterfactual explanations. These counterfactual explanations explain the decisions made by machine learning models in a human-understandable way, defined specifically as a minimal change to the feature values such that the system's classification decision is changed [Martens and Provost, 2014, Wachter et al., 2017, Verma et al., 2020, Fernández-Loría et al., 2022]. When using behavioral data, this corresponds to a minimal set of active features of the instance, where changing (just) these feature values to zero would lead the model to make a different decision [Ramon et al., 2020]. We apply counterfactual explanations instead of other explanation techniques as they give a direct way to alter the predicted outcome, in line with Chen et al. [2017]. Nevertheless, with suitable modifications other explanation techniques such as SHAP could also be adapted to support cloaking [Lundberg and Lee, 2017, Ramon et al., 2020]. We use the SEDC algorithm[4] to compute the counterfactual explanations, which uses a best-first heuristic search strategy to search for the smallest set of features to change [Martens and Provost, 2014, Ramon et al., 2020]. A change is defined as replacing the original feature value with the median value of that feature over the training data, which in the case of behavioral data will be 0 as this data is usually very sparse. For example, in Facebook data, there is no page that is liked by the majority of users, so the median value of every feature will be 0. Counterfactual explanations will then point to the Facebook likes a user has to cloak (or simply unlike). A user is considered to be successfully cloaked when his or her score falls below the predefined threshold [Chen et al., 2017]. The average size of a counterfactual explanation, the average number of likes that have to be cloaked to avoid positive inference for each prediction task, can be found in Table 3.

Table 3: Model statistics. *Positive rate* indicates the percentage of instances that are predicted as positive by the machine learning model. *AUC* is the model's on the task accuracy, as measured by the area under the ROC curve. *Explanation size* is the average number of likes that must be cloaked to avoid positive inferences.

Target variable	AUC (in %)	Positive rate (in %)	Explanation size (avg.)
Male	95.2	5.13	8
Female	95.2	4.75	6
Homosexual	89.4	5.72	4
Lesbian	77.8	7.88	2
Democrat	77.3	4.58	3
Republican	82.1	4.79	6

Metafeatures Dimensionality reduction methods reduce the size of the set of features used for modeling [Clark and Provost, 2019]. To group fine-grained features into higher-level metafeatures, we use Non-Negative Matrix Factorization (NMF) [Lee and Seung, 1999]. We chose this dimensionality reduction technique as the non-negativity constraint facilitates the interpretation of the extracted metafeatures, and has been shown to provide interpretable results for fine-grained data applications [Contreras-Piña and Ríos, 2016, Ramon et al., 2021b]. [5] We create 50 metafeatures for the Facebook likes and assign each page to the metafeature or topic for which it has the highest weight to ensure mutual exclusivity (each feature only belongs to one metafeature) [Ramon et al., 2021b]. An example of two metafeatures

[4]Python code available at `https://github.com/ADMAntwerp/edc`

[5]Note that there exist many other techniques to generate the metafeatures, but we do not compare them in this work, as our goal is to study whether using metafeatures can give better performance, rather than to figure out what sort of metafeatures performs best.

is shown in Section 4. The creation of these metafeatures is data-driven. Another option is to use the categories that Facebook assigned to the Facebook like pages itself. These categories are more broad such as 'Public Figure' or 'Musician/Band'. We will call these domain-based metafeatures, in line with Ramon et al. [2021b]. We include the results based on these metafeatures in Appendix A.

4 Experimental set-up

We focus on cloaking the inferences gender (*male* and *female*), sexual orientation (*homosexual* and *lesbian*) and political orientation (*democrat* and *republican*). We train Logistic Regression models with ℓ_2-regularization with the Scikit-learn library (Python). The literature has shown that this is one of the best performing classification models for behavioral data [De Cnudde et al., 2020]. We use 66% of the data for training, and the remaining 33% for testing. We also exclude users with fewer than 10 likes and Facebook pages with fewer than 10 likes. For fine-tuning the hyperparameters of the model, we perform a grid search on the training set by using three-fold cross-validation, where we tune the regularization parameter C of ℓ_2-LR model. As is common in targeted advertising, we assume that a positive inference is drawn, which means that the user would be subjected to targeted advertising, when the model assigns the user a score which places him or her in a specified top quantile of the score distribution produced by the prediction model [Chen et al., 2017, Perlich et al., 2014]. For online targeting, a typical value for this quantile is between 90 and 100, and we base our threshold for positive inference on the top 95 quantile in the training set [Chen et al., 2017, Perlich et al., 2014]. The AUC and positive rate for each prediction task can be found in Table 3.

4.1 Longer-term cloaking protection

We study longer-term cloaking protection using the methodology described in Figure 1. We simulate a person's behavior over time by first dropping 50% of likes for each user at random.[6] After dropping the pages, we train a regularized logistic regression model for every prediction task on the reduced training set. We use this model to make predictions on the instances in the reduced test set and select the positively predicted instances. For these instances, we compare two cloaking strategies to remove positive inference:

1. Cloaking the individual likes, which we also call the fine-grained features (where we remove all the pages in the counterfactual explanation of that instance as described by Chen et al. [2017]). We call this strategy *FG*.
2. Cloaking the metafeatures, where we remove all the pages in the counterfactual explanation of that instance **and** the pages that belong to the same metafeatures as the pages in the counterfactual explanation. We call this strategy MF.

Using the second strategy also means that the number of liked pages will decrease substantially more than when using the first strategy, which we will show in Figure 6. We simulate an individual's behavior over time by gradually re-introducing more of the 50% of pages that initially were dropped. We measure the **longer-term cloaking protection** of both strategies by measuring the percentage of positively predicted instances for which cloaking this targeting task in the past successfully inhibited future inferences for that same task and individual.

4.2 Trade-off between privacy and personalization

As introduced above, although we may think myopically about a privacy-preserving action when taking it, such actions can have broader effects. In particular, cloaking data in order to inhibit one inference can have effect on other inferences—possibly ones that we would not want to inhibit. Therefore, consumers and platforms should be interested in what the effect of hiding larger parts of someone's traces is on the performance of other prediction tasks.

Define X as the initial complete data, and X_c as the cloaked data. To what extent does changing X to X_c affect the predictions of models for different (but desired) target variables?

We describe the set-up of our experiment in Figure 2. We examine the effect on a second set of prediction tasks when cloaking for the trait-prediction tasks we described above. The new tasks involve predicting an individual's ratings for the Big 5 personality traits, the accuracy of which we measure using Pearson correlation, which is the most commonly used measure of prediction accuracy for predicting these personality traits [Kosinski et al., 2013, Azucar et al., 2018]. We choose the Big 5 traits as the second set of tasks because they cover broad aspects of personality and are very well

[6]This means that the simulation uses the assumption that people's behavior over time is stable in the short run, as their liking behavior does not change—since the data does not include time stamps. Verifying this on time-stamped data would be an avenue for follow-up research.

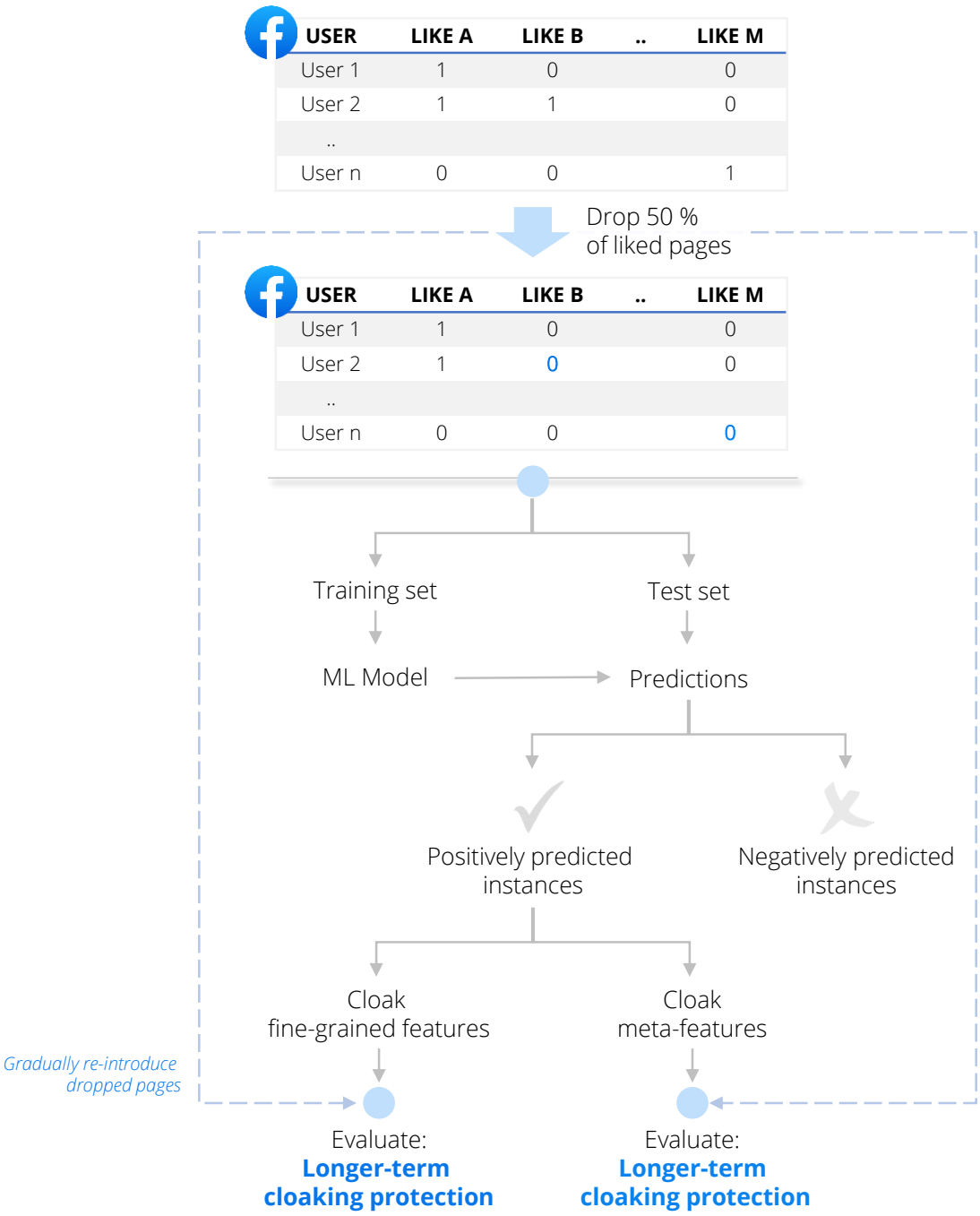


Figure 1: Experimental set-up to measure the longer-term cloaking protection.

understood. We compare the effect of not cloaking an individual's data, cloaking fine-grained footprints (FG), and cloaking metafeatures (MF).[7]

[7]The point here is not that inferences for the Big 5 traits are not privacy invasive; this of course will depend on the individual. Rather, the point simply is to examine the effects of cloaking some potentially sensitive inferences on other inferences that are broadly applicable.

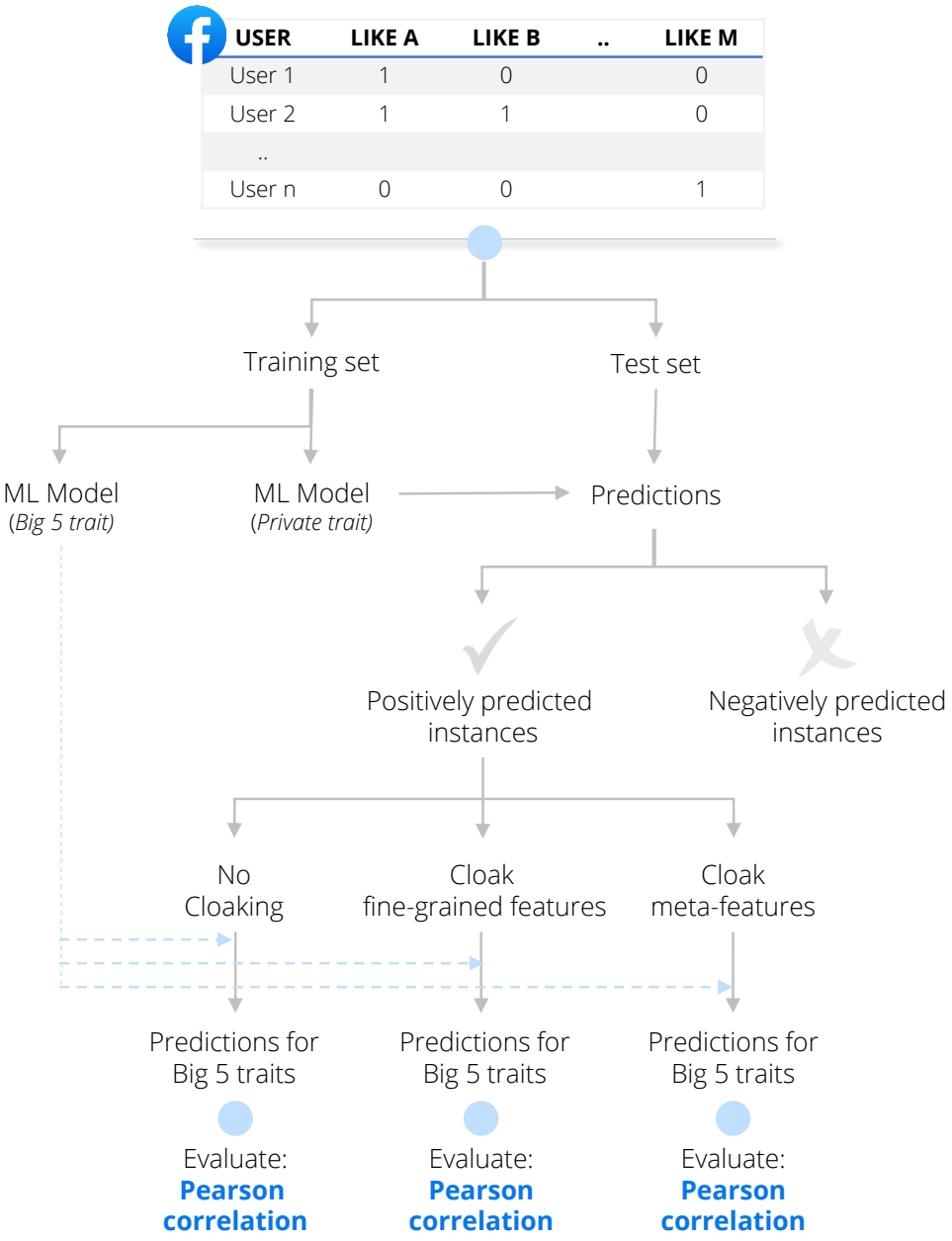


Figure 2: Experimental set-up to measure the impact of cloaking a private trait on other prediction tasks.

5 Results: Longer-term cloaking protection

Imagine a random person, John, who has been using a particular technology platform and thereby leaving digital footprints. The platform's political-orientation model gives him a high enough score as a *republican* in order for him to receive corresponding political ads. John no longer wants to receive advertisements related to his political orientation; maybe because he no longer identifies as such, or he wants to keep his political orientation private, or he simply finds the advertisements annoying. We want to see if, as he subsequently continues to like pages, John gets targeted as republican again after using the cloaking device described by Chen et al. [2017].

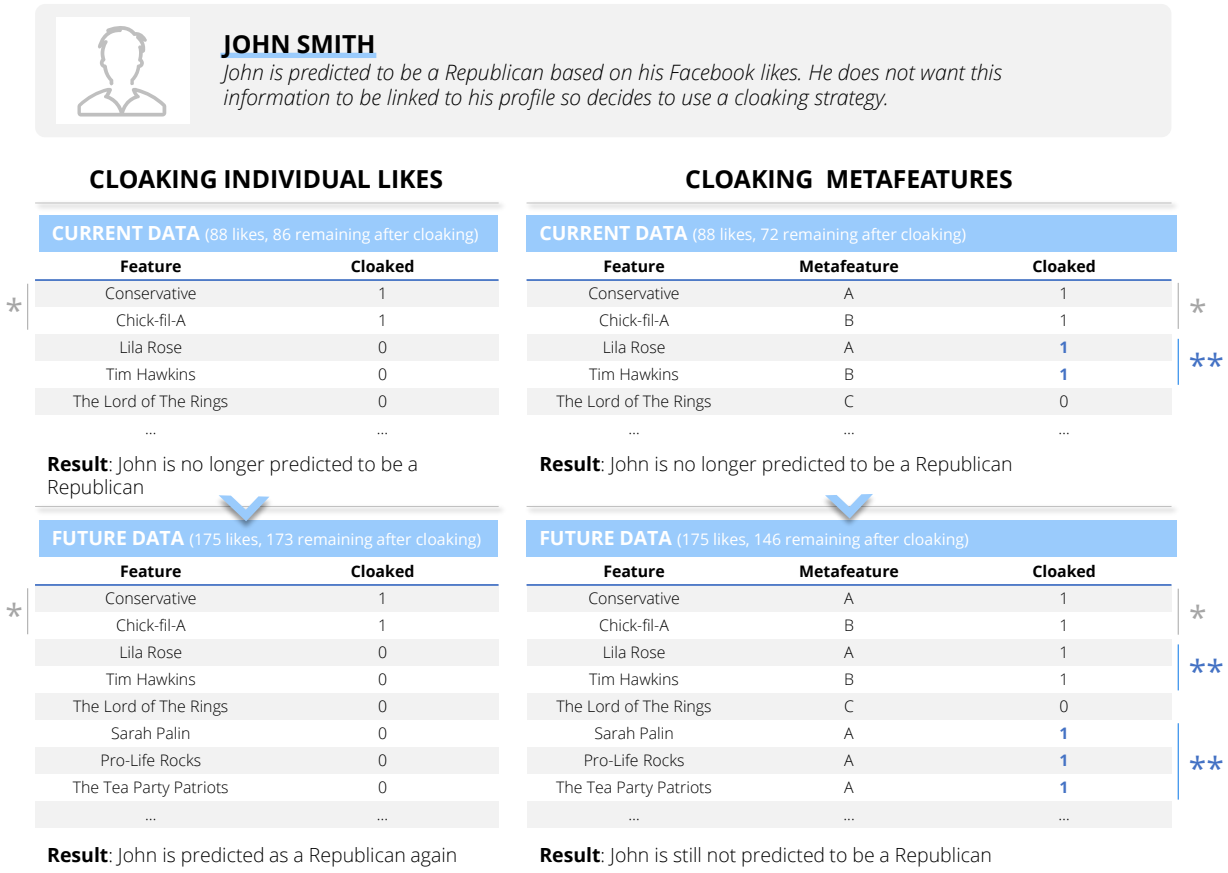


Figure 3: Example of John. The column *Cloaked* signals the pages that are cloaked for each strategy and point in time.
: Original features cloaked to ensure John is not predicted as republican.
*: Additional features cloaked because part of same metafeature as the original features.

We represent John in Figure 3. As described above, we simulate the point where the model uses half of his digital traces (88 likes) as the current footprint data, and the point with all his digital footprints as his future data (175 likes). Using the current data, John is predicted as a republican.[2] John wants to inhibit this inference and receives the following advice (based on counterfactual explanations to bring him under the threshold): *If you would hide the likes 'Conservative' and 'Chick-fil-A', you would no longer be predicted as republican.* After cloaking these pages, John has 86 likes remaining and is no longer predicted as republican. [3] We move on to the future point. John, who has remained active, has liked 87 new pages (essentially doubling his digital traces). Even though the two liked pages from his initial counterfactual explanation are still cloaked, John gets re-predicted as republican due to his new digital footprints. [8] This illustrates that cloaking is not necessarily robust in the longer term, as individuals continue to leave new digital footprints. Note that Chen et al. [2017] pointed out as a limitation of the cloaking design that if cloaking does not also cover closely associated features, one might end up being targeted again in the future [Chen et al., 2017].

We introduced cloaking based on metafeatures in an attempt to (partially) address this lack of robustness. Recall that cloaking metafeatures also cloaks other footprints that are (estimated to be) closely related to those suggested by the counterfactual explanation. So for our current example, the Facebook page 'Conservative' belongs to metafeature A, and the Facebook page 'Chick-fil-A' belongs to metafeature B. Typically, metafeatures such as these[8] are interpreted by

[2]Prediction score = 0.161, which is above the targeting threshold of 0.148.

[3]Prediction score = 0.140, which is below the threshold of 0.148.

[8]His prediction score on the full data is 0.260; after cloaking the two likes in his explanation, his prediction score is 0.225. This above the targeting threshold of 0.194. The threshold is different now because after everyone in the dataset has acquired new digital traces, the scores for the top 5th percentile will be different.

[8]Specifically, those created by embedding the original data in a lower dimensional space.

looking at the top weighted fine-grained features for each metafeature [Wang and Zhang, 2012, O'callaghan et al., 2015, Contreras-Piña and Ríos, 2016]. These are shown in Table 4.

Table 4: Interpretation of two metafeatures generated with NMF by showing the 10 features with the highest coefficients for each metafeature.

Metafeature A	**Metafeature B**
Being Conservative	*The Bible*
Sarah Palin	*Jesus Daily*
Conservative	*"I'm proud to be Christian" by Aaron Chavez*
Glenn Beck	*Casting Crowns*
Fox News	*Chris Tomlin*
Tea Party Patriots	*Third Day*
Mitt Romney	*TobyMac*
FreedomWorks	*Jeremy Camp*
Sean Hannity	*Switchfoot*
John McCain	*Skillet Music*

Metafeature A clearly is related to right-wing politics, and metafeature B to Christianity. Metafeature cloaking hides not only those likes (fine-grained features) in the counterfactual explanation, but also all the likes that belong to the same metafeature as each of these likes. When we also cloak all the likes in the associated metafeatures, 14 additional pages are cloaked. These include 'Tim Hawkins' and 'Lila Rose'.[9] Subsequently, when John likes pages in the future, the new pages associated with those same metafeatures will be hidden as well. For John, this leads to also hiding pages such as 'Sarah Palin', 'Tea Party Patriots' and 'Pro-Life Rocks'. In total, 29 new pages are cloaked in the future and the result is that John will not be predicted as republican even after leaving his future footprints.[10]

Moving beyond the specific example of John, we compare the longer-term cloaking protection of the two cloaking strategies in hiding gender, political orientation and sexual orientation. As shown in Figure 4, cloaking the fine-grained features offers less protection over time than cloaking the metafeatures. People get targeted again relatively quickly. For example, when cloaking *male*, after adding 10% new likes, only 57.6% of instances are still successfully cloaked. After adding all their new likes (and thus doubling their digital traces), only 21.5% are still successfully cloaked. On the other hand, when we cloak the metafeatures instead, we see that 86.6% are still successfully cloaked when the digital traces are doubled.

We see the same patterns when cloaking *female* and political orientation (*democrat* and *republican*). Sexual orientation, especially *lesbian*, is more effectively cloaked over time than other tasks when using fine-grained features; cloaking the metafeatures is still a more effective longer-term cloaking strategy, but the difference between the strategies is be smaller.[11] We hypothesize that this could be because there are fewer people whose true target label is *lesbian* in the targeted population (*True Positives*).

We analyze whether there is a difference in longer-term cloaking protection between correctly predicted people (True Positives) and people that were incorrectly predicted as the target variable (False Positives). Are people who were correctly inferred more likely to reveal themselves again over time? In Figure 5a, one can observe that the longer-term cloaking protection of True Positives is in fact lower. This aligns with intuition, given their higher likelihood of repeating behaviors that could result in the same prediction. This difference almost disappears when we assess the longer-term cloaking protection with metafeatures (Figure 5b).

We also present the results of two additional cloaking strategies in Appendix A. The first option involves using the categories assigned by Facebook to the like pages themselves, which we refer to as domain-based metafeatures. The advantage of using domain-based metafeatures is that they are readily available and by design comprehensible. However, as shown in Figure 8, the data-driven metafeatures created by NMF are more effective in avoiding inferences over time, and in addition our analysis reveals that on average they hide fewer pages than the domain-based metafeatures. We conjecture that this is because they more accurately capture general patterns of behavior. For example, when examining the metafeatures in Table 4, we see that they are strongly associated with right-wing politics and Christianity, which are

[9]This brings the prediction score further down to 0.120.

[10]The prediction score on the future data after cloaking the metafeatures is 0.126, which is well below the threshold of 0.194.

[11]We see that for the prediction task of *lesbian*, for a very small number of individuals, cloaking the metafeatures instead of the fine-grained can lower the number of successfully cloaked individuals, even without adding additional data.

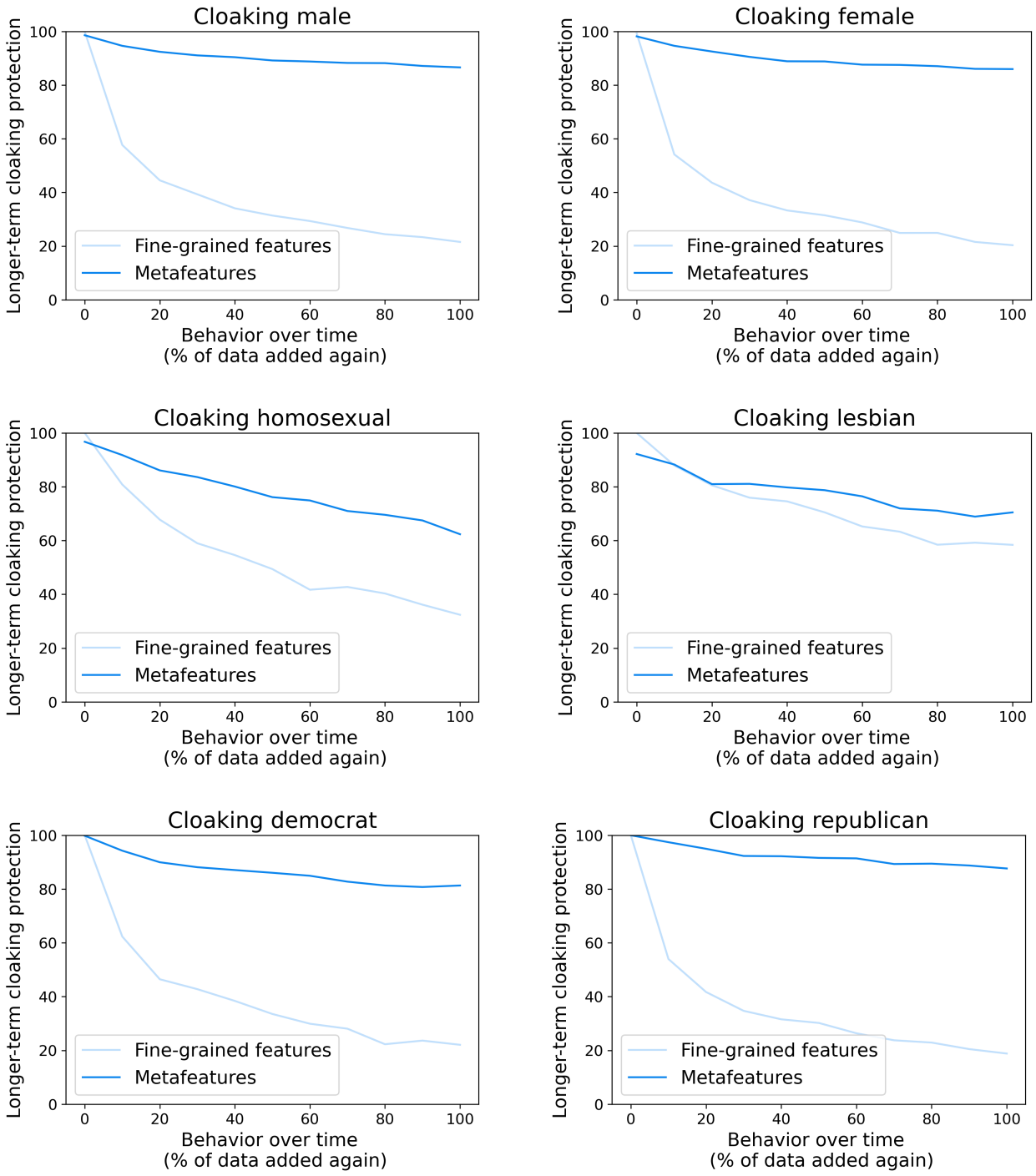
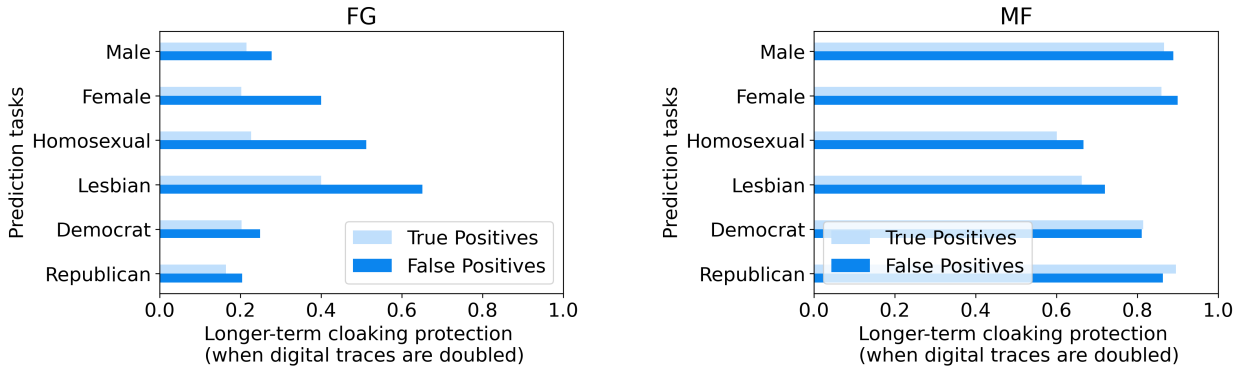


Figure 4: Longer-term cloaking protection. We measure the longer-term cloaking protection as the percentage of positively predicted instances for which cloaking this targeting task successfully inhibits future inference. The population taken into account constitutes the intersection of individuals predicted as positive when using 1/2 of the data, and when using the full data.We measure the evolution over time on the x-axis by gradually re-adding the dropped pages.

both highly predictive of being a Republican. On the other hand, domain-based metafeatures such as 'Public Figure' may be too general to capture these specific patterns.



(a) (b)

Figure 5: Is there a difference in longer-term cloaking protection between True Positives and False Positives? We measure this at the point where all the initially dropped data is re-added and the digital traces are thus doubled.

Another cloaking strategy we experiment with involves adding a *tolerance level* to the initial counterfactual explanations. This means that instead of using the threshold of the decision-making system to generate the counterfactual explanations for cloaking, we employ a **lower** threshold. It is to be expected that when we bring someone just below the threshold (which is what counterfactual explanations do), the chances of them crossing the threshold again are relatively high. Therefore, we explore bringing individuals not only below the 95% threshold but also below the 90% quantile (while still using the 95% quantile as threshold for predictions). This approach should provide an additional layer of protection from future targeting. As depicted in Figure 9, it does indeed offer extended protection initially, but on average, the protection of the cloaking strategy still diminishes rapidly as more likes are accumulated over time. This strategy has no impact on the pages that will be liked in the future, and this is clearly evident in the results. This highlights one of the major advantages of using metafeatures as a cloaking strategy.

6 Results: Trade-off between privacy and personalization

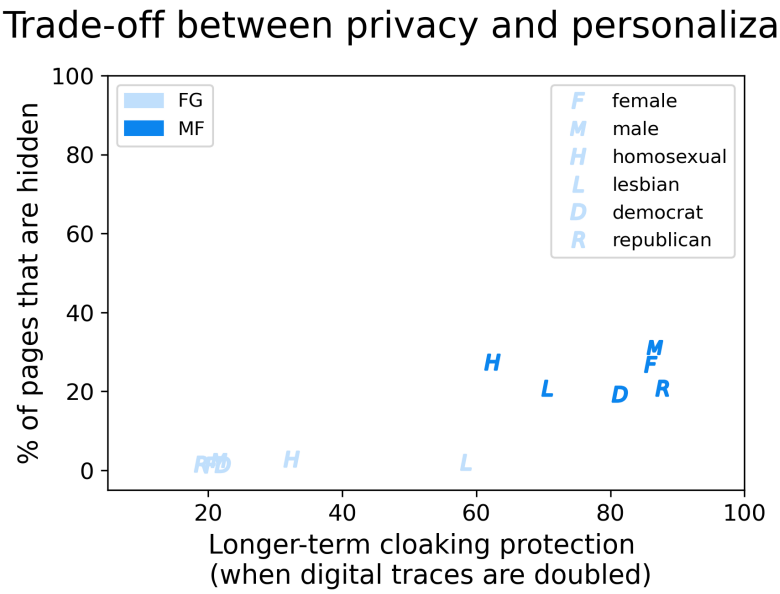
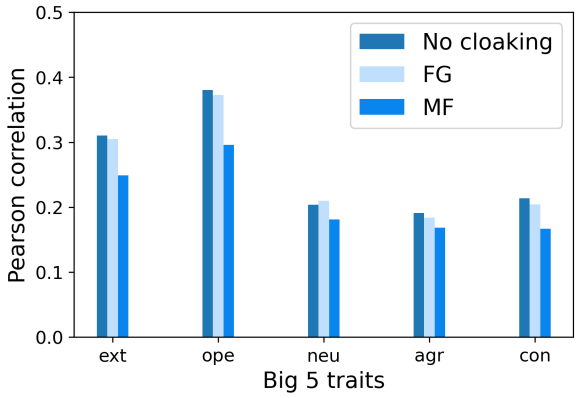
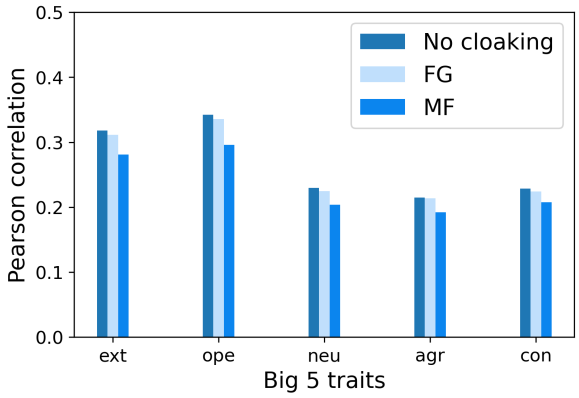


Figure 6: What percentage of people's digital footprints are hidden with each cloaking strategy? We measure loss in personalization by the average % of someone's likes that have to be removed, and privacy protection as the level of longer-term cloaking protection at the point when the individual's digital footprints are doubled.

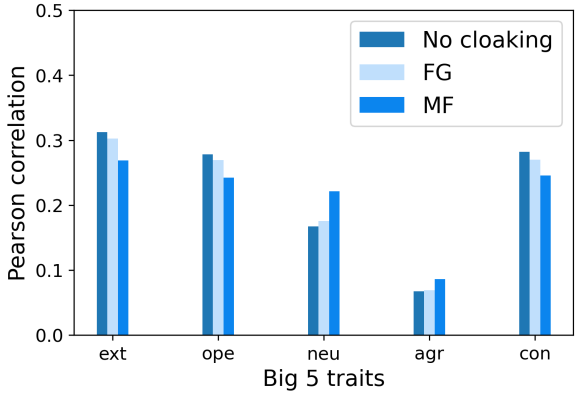
Cloaking metafeatures hides larger portions of an individual's digital footprints, resulting in increased privacy protection but potentially losing the benefits of personalization. We assume users do not want to lose all personalization; otherwise, an individual could simply cloak all his liked pages and no inferences would be made (this could still be a viable option for some users, although this will be a bad outcome from the perspective of the advertising platform). Figure 6 illustrates that cloaking metafeatures results in a substantial increase in privacy, but also in a substantial increase in the number of pages that are being hidden than when using fine-grained features. In the example of John, after cloaking the fine-grained features, he has 173 likes left for personalization, while after cloaking the metafeatures, he only has 149 likes left. Therefore it is important to assess the impact of cloaking an individual's sensitive traits on the ability to predict other things about the individual.



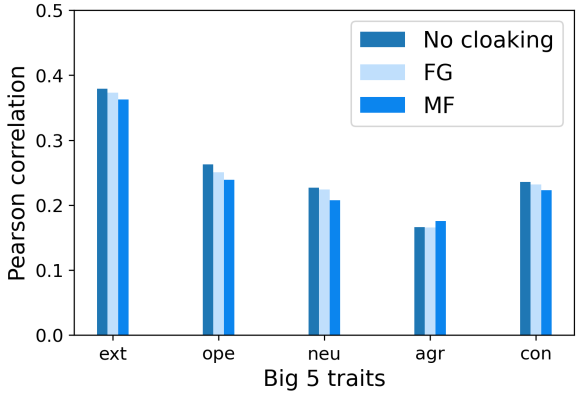
(a) Male



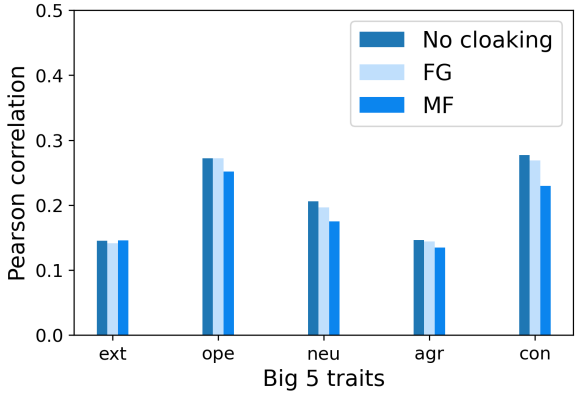
(b) Female



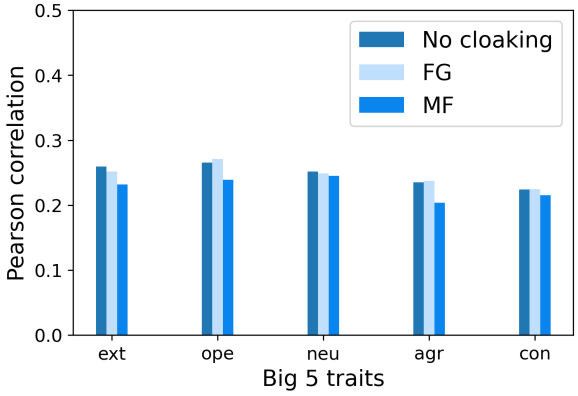
(c) Homosexual



(d) Lesbian



(e) Democrat



(f) Republican

Figure 7: Effect of cloaking on the predictive performance of the Big 5 traits.

We follow the set-up described in Section 4.2 to measure the impact of cloaking sensitive traits on the predictive performance of other prediction tasks (in this case the Big 5 traits). Figure 7 shows that the impact of cloaking metafeatures on the predictive performance of the Big 5 traits is larger on average than the impact of cloaking fine-grained features. For both strategies, the impact is largest when cloaking gender, followed by political orientation. The impact of both cloaking strategies on the predictive performance seems fairly *small* in most cases, but we cannot truly judge the losses in value (corresponding to the small losses in predictive power) in a study such as this.

7 Discussion and Conclusion

The digital traces we leave every day enable those who collect them to make intimate inferences about who we are. While such inferences might lead to desired personalization outcomes, they also pose a considerable threat to individuals' privacy and self-determination. In this paper, we examined the effectiveness and impact of two related privacy-enhancing cloaking strategies which conceal a portion of users' digital footprints to limit the ability of platforms to make predictions about underlying psychological traits. Although previous work has shown that such cloaking mechanisms can be effective in the short-term [Chen et al., 2017], our findings suggest that their effectiveness rapidly declines over time. That is, as people continue to generate traces after the cloaking has been implemented, the system quickly relearns to draw those same inferences from the new data. We introduce a new cloaking strategy - one that is based on cloaking metafeatures rather than individual footprints - and show how this strategy offers better longer-term privacy protection.

In addition, our findings also highlight potential trade-offs between privacy protection and personalized services. That is, while individuals might be interested in cloaking certain aspects of their identity (e.g., their sexual orientation), they might appreciate the benefits they receive from sharing other parts (e.g., their openness). We show that cloaking a particular trait likely has spillover effects on other traits that were not intentionally targeted. Although the trade-off between personalization and privacy is not a new idea, there are few (if any) empirical analyses of the actual trade-offs introduced by different privacy-enhancing techniques, including cloaking.[12]

The extent to which trading off personalization for enhanced privacy protection is desirable will depend on the specific context and preferences of the user. While some users might be willing to forsake targeted advertising for higher levels of privacy, others might favor convenience and service over the ability to conceal potentially unwanted aspects of their identity. The same is true for companies who might trade-off the ability to get highly granular consumer insights on all levels for a higher likelihood that consumers will stay on the platform and refrain from opting out of tracking and personalization altogether. We argue that different forms of cloaking can provide solutions that operate between the two extremes. On the one hand, they allow companies to keep collecting large amounts of data and monetize it within the boundaries set by users. On the other hand, users gain control over the level of personalization they feel comfortable with, while having the ability to inhibit unwanted inferences.

7.1 Practical Implications

One of the main practical implications of this research is the extent to which cloaking can assist individuals in making smarter decisions when it comes to protecting their privacy. While data protection regulations such as the General Data Protection Regulation (GDPR) in Europe or the California Consumer Privacy Act (CCPA) have pushed for increased consumer control, there is a growing body of research suggesting that - without any support - individuals struggle to act as responsible stewards of their personal data. Research on the *privacy paradox*, for example, reveals a stark discrepancy between a user's expressed concerns regarding online privacy and their actual behavior when sharing personal information. Despite expressing concerns about their privacy, individuals are often willing to share personal information online in exchange for personalized recommendations [Barth and De Jong, 2017]. For instance, even though 93% of USA citizens consider it important to maintain control over who can access their data, only a small fraction of people actually read the privacy policies of the services collecting their data [Madden and Rainie, 2015, Solove, 2012, Matz et al., 2020]. One (obvious) reason why consumers do not succeed in achieving desired levels of privacy is their lack of knowledge about how their data is actually being collected and used [Acquisti et al., 2020]. Related is the *acceptability gap*, which shows that users are more accepting of personalized services than of the collection of personal data required for these services [Kozyreva et al., 2021]. They overlook the relationship between them, and as a result, fail to engage in an adequate comparison of the value received from personalization to the value of keeping data private Kozyreva et al. [2021]. As a consequence, most people tend to overvalue the short-term benefits of actions, such as using an app, over the long-term privacy risks, which are delayed and intangible [Acquisti, 2004].

[12] Prior work has shown a trade-off between privacy protection and advertising effectiveness [Goldfarb and Tucker, 2011], but to our knowledge, not previously at the level of a specific prediction task.

Complicating matters further, research has suggested that people's apparent inaction regarding their privacy is also the result of them feeling that they have no control over the situation, and as a consequence simply give up (a phenomenon researchers have called *digital resignation*) [Draper and Turow, 2019, Acquisti et al., 2020]. Finally, it may simply be that the perceived cost of protecting privacy is simply too high: either not using a service or possibly navigating an ultra-complicated web of documents and settings. In all of these cases, providing transparency into how data is used and control over its use seems vital for consumer welfare and, in particular, for users to make informed privacy decisions [Matz et al., 2020]. Both privacy and transparency are essential prerequisites for establishing a trustworthy AI system [Liu et al., 2022].

In this paper, we introduce a tool that could help guide individuals in making choices on their privacy settings online. Since the implications of sharing personal data are often difficult to anticipate, let alone trade-off for immediate convenience rewards, we need easy ways for people to move the dial between oversharing and undersharing. Cloaking offers such a lever and might encourage platforms to offer more mechanisms for users to control data-driven inferences and personalization (including targeted advertising) either through editing of the data items—the digital footprints—that are stored about them or through an explicit cloaking mechanism that hides footprints from the AI inference systems specifically.

Notably, our cloaking methodology depends on the cooperation of the platforms that collect this kind of data (like Facebook, Google, Spotify). While platforms might try to resist the introduction of technology that limits their ability to commercialize consumer insights, we argue that introducing certain levels of consumer control in the form of cloaking could eventually benefit them in the long-run. As stricter data protection regulations are introduced around the world - often empowering consumers to revoke access to their personal data - platforms might be forced to provide sufficient transparency and control in order to retain users and prevent them from option out of data collection entirely. Moreover, a gradual shift to higher levels of platform-driven user control might prevent legislators from introducing more paternalistic regulatory actions.

Acknowledgments

This research was funded by Flemish Research Foundation (grant number 11N7723N). We would also like to thank Research Foundation-Flanders for their support for the research stay in New York (grant number V403523N). Foster Provost thanks Ira Rennert and the Stern/Fubon Center for support. This study was approved by the Institutional Review Board of the University of Antwerp (SHW2231, 7 February 2023).

References

Waller, Angie and Lecher, Colin. Facebook promised to remove "sensitive" ads. here's what it left behind. `https://themarkup.org/citizen-browser/2022/05/12/facebook-promised-to-remove-sensitive-ads-heres-what-it-left-behind`, 2022. Accessed: 2023-03-13.

A. Acquisti. Privacy in electronic commerce and the economics of immediate gratification. In *Proceedings of the 5th ACM conference on Electronic commerce*, pages 21–29, 2004.

A. Acquisti, L. Brandimarte, and G. Loewenstein. Secrets and likes: The drive for privacy and the difficulty of achieving it in the digital age. *Journal of Consumer Psychology*, 30(4):736–758, 2020.

J. Angwin and T. Parris Jr. Facebook lets advertisers exclude users by rac. `https://www.propublica.org/article/facebook-lets-advertisers-exclude-users-by-race`, 2016. Accessed: 2023-03-13.

R. E. Appel and S. C. Matz. Psychological targeting in the age of big data. In *Measuring and Modeling Persons and Situations*, pages 193–222. Elsevier, 2021.

D. Azucar, D. Marengo, and M. Settanni. Predicting the big 5 personality traits from digital footprints on social media: A meta-analysis. *Personality and individual differences*, 124:150–159, 2018.

S. Barth and M. D. De Jong. The privacy paradox–investigating discrepancies between expressed privacy concerns and actual online behavior–a systematic literature review. *Telematics and informatics*, 34(7):1038–1058, 2017.

J. G. Cabañas, Á. Cuevas, and R. Cuevas. Facebook use of sensitive data for advertising in europe. *arXiv preprint arXiv:1802.05030*, 2018.

D. Chen, S. P. Fraiberger, R. Moakler, and F. Provost. Enhancing transparency and control when drawing data-driven inferences about individuals. *Big data*, 5(3):197–212, 2017.

J. Clark and F. Provost. Unsupervised dimensionality reduction versus supervised regularization for classification from sparse data. *Data Mining and Knowledge Discovery*, 33:871–916, 2019.

C. Contreras-Piña and S. A. Ríos. An empirical comparison of latent sematic models for applications in industry. *Neurocomputing*, 179:176–185, 2016.

P. T. Costa and R. R. McCrae. Normal personality assessment in clinical practice: The neo personality inventory. *Psychological assessment*, 4(1):5, 1992.

S. De Cnudde, D. Martens, T. Evgeniou, and F. Provost. A benchmarking study of classification techniques for behavioral data. *International journal of data science and analytics*, 9(2):131–173, 2020.

N. A. Draper and J. Turow. The corporate cultivation of digital resignation. *New media & society*, 21(8):1824–1839, 2019.

Edelman, Gilad. How facebook's political ad system is designed to polarize. `https://www.wired.com/story/facebook-political-ad-system-designed-polarize/`, 2019. Accessed: 2023-03-13.

C. Fernández-Loría, F. Provost, and X. Han. Explaining data-driven decisions made by ai systems: The counterfactual approach. *MIS Quarterly*, 46(3):1635–1660, 2022.

C. Fernández-Loría, F. Provost, J. Anderton, B. Carterette, and P. Chandar. A comparison of methods for treatment assignment with an application to playlist generation. *Information Systems Research*, 34(2):786–803, 2017.

L. R. Goldberg, J. A. Johnson, H. W. Eber, R. Hogan, M. C. Ashton, C. R. Cloninger, and H. G. Gough. The international personality item pool and the future of public-domain personality measures. *Journal of Research in personality*, 40 (1):84–96, 2006.

A. Goldfarb and C. E. Tucker. Privacy regulation and online advertising. *Management science*, 57(1):57–71, 2011.

G. A. Johnson, S. K. Shriver, and S. Du. Consumer privacy choice in online advertising: Who opts out and at what cost to industry? *Marketing Science*, 39(1):33–51, 2020.

E. Junqué de Fortuny, D. Martens, and F. Provost. Predictive modeling with big data: is bigger really better? *Big data*, 1(4):215–226, 2013.

M. Kosinski, D. Stillwell, and T. Graepel. Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the national academy of sciences*, 110(15):5802–5805, 2013.

A. Kozyreva, P. Lorenz-Spreen, R. Hertwig, S. Lewandowsky, and S. M. Herzog. Public attitudes towards algorithmic personalization and use of personal data online: Evidence from germany, great britain, and the united states. *Humanities and Social Sciences Communications*, 8(1):1–11, 2021.

Lecher, Colin. How big pharma finds sick users on facebook. `https://themarkup.org/citizen-browser/2021/05/06/how-big-pharma-finds-sick-users-on-facebook`, 2021. Accessed: 2023-03-13.

D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755): 788–791, 1999.

H. Liu, Y. Wang, W. Fan, X. Liu, Y. Li, S. Jain, Y. Liu, A. Jain, and J. Tang. Trustworthy ai: A computational perspective. *ACM Transactions on Intelligent Systems and Technology*, 14(1):1–59, 2022.

S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

M. Madden and L. Rainie. Americans' attitudes about privacy, security and surveillance. 2015.

D. Martens and F. Provost. Explaining data-driven document classifications. *MIS quarterly*, 38(1):73–100, 2014.

S. Matz, Y. W. F. Chan, and M. Kosinski. Models of personality. *Emotions and Personality in Personalized Services: Models, Evaluation and Applications*, pages 35–54, 2016.

S. C. Matz, M. Kosinski, G. Nave, and D. J. Stillwell. Psychological targeting as an effective approach to digital mass persuasion. *Proceedings of the national academy of sciences*, 114(48):12714–12719, 2017.

S. C. Matz, R. E. Appel, and M. Kosinski. Privacy in the age of psychological targeting. *Current opinion in psychology*, 31:116–121, 2020.

G. J. Meyer, S. E. Finn, L. D. Eyde, G. G. Kay, K. L. Moreland, R. R. Dies, E. J. Eisman, T. W. Kubiszyn, and G. M. Reed. Psychological testing and psychological assessment: A review of evidence and issues. *American psychologist*, 56(2):128, 2001.

D. O'callaghan, D. Greene, J. Carthy, and P. Cunningham. An analysis of the coherence of descriptors in topic modeling. *Expert Systems with Applications*, 42(13):5645–5657, 2015.

C. Perlich, B. Dalessandro, T. Raeder, O. Stitelman, and F. Provost. Machine learning for targeted display advertising: Transfer learning in action. *Machine learning*, 95(1):103–127, 2014.

Y. Ramon, D. Martens, F. Provost, and T. Evgeniou. A comparison of instance-level counterfactual explanation algorithms for behavioral and textual data: SEDC, LIME-C and SHAP-C. *Advances in Data Analysis and Classification*, 14(4):801–819, 2020.

Y. Ramon, R. Farrokhnia, S. C. Matz, and D. Martens. Explainable ai for psychological profiling from behavioral data: an application to big five personality predictions from financial transaction records. *Information*, 12(12):518, 2021a.

Y. Ramon, D. Martens, T. Evgeniou, and S. Praet. Can metafeatures help improve explanations of prediction models when using behavioral and textual data? *Machine Learning*, pages 1–40, 2021b.

Silberling, Amanda. Facebook will no longer allow advertisers to target political beliefs, religion, sexual orientation. `https://techcrunch.com/2021/11/09/facebook-will-no-longer-allow-advertisers-to-target-political-beliefs-religion-sexual-orientation/`, 2021. Accessed: 2023-03-13.

D. J. Solove. Introduction: Privacy self-management and the consent dilemma. *Harv. L. Rev.*, 126:1880, 2012.

T. P. Tran. Personalized ads on facebook: An effective marketing tool for online marketers. *Journal of Retailing and Consumer Services*, 39:230–242, 2017.

S. Verma, V. Boonsanong, M. Hoang, K. E. Hines, J. P. Dickerson, and C. Shah. Counterfactual explanations and algorithmic recourses for machine learning: a review. *arXiv preprint arXiv:2010.10596*, 2020.

S. Wachter, B. Mittelstadt, and C. Russell. Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Harv. JL & Tech.*, 31:841, 2017.

Y.-X. Wang and Y.-J. Zhang. Nonnegative matrix factorization: A comprehensive review. *IEEE Transactions on knowledge and data engineering*, 25(6):1336–1353, 2012.

A Results of other cloaking strategies

A.1 Using domain-based metafeatures

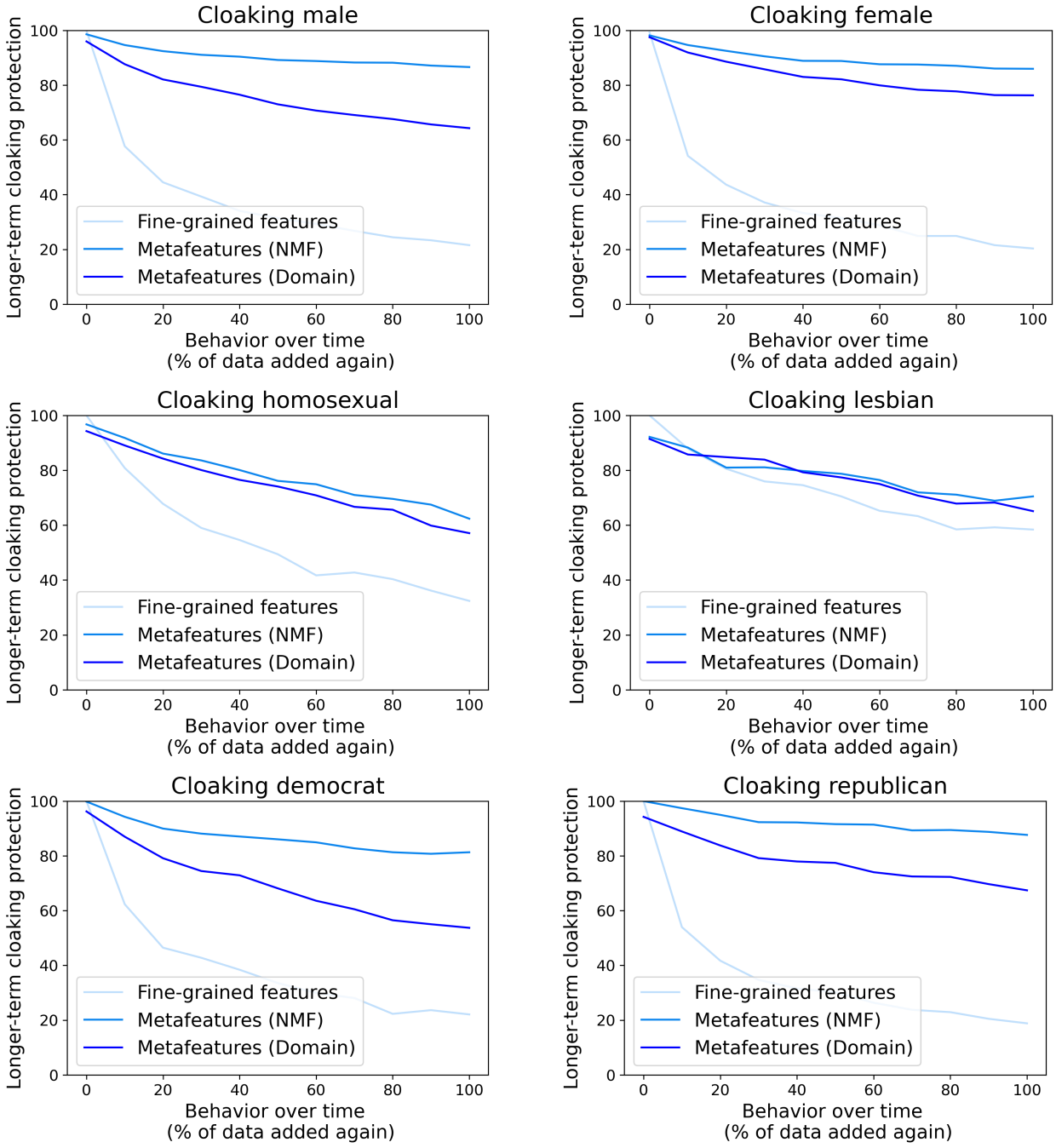


Figure 8: Longer-term cloaking protection over time when using domain-based metafeatures.

A.2 Using explanations with a tolerance

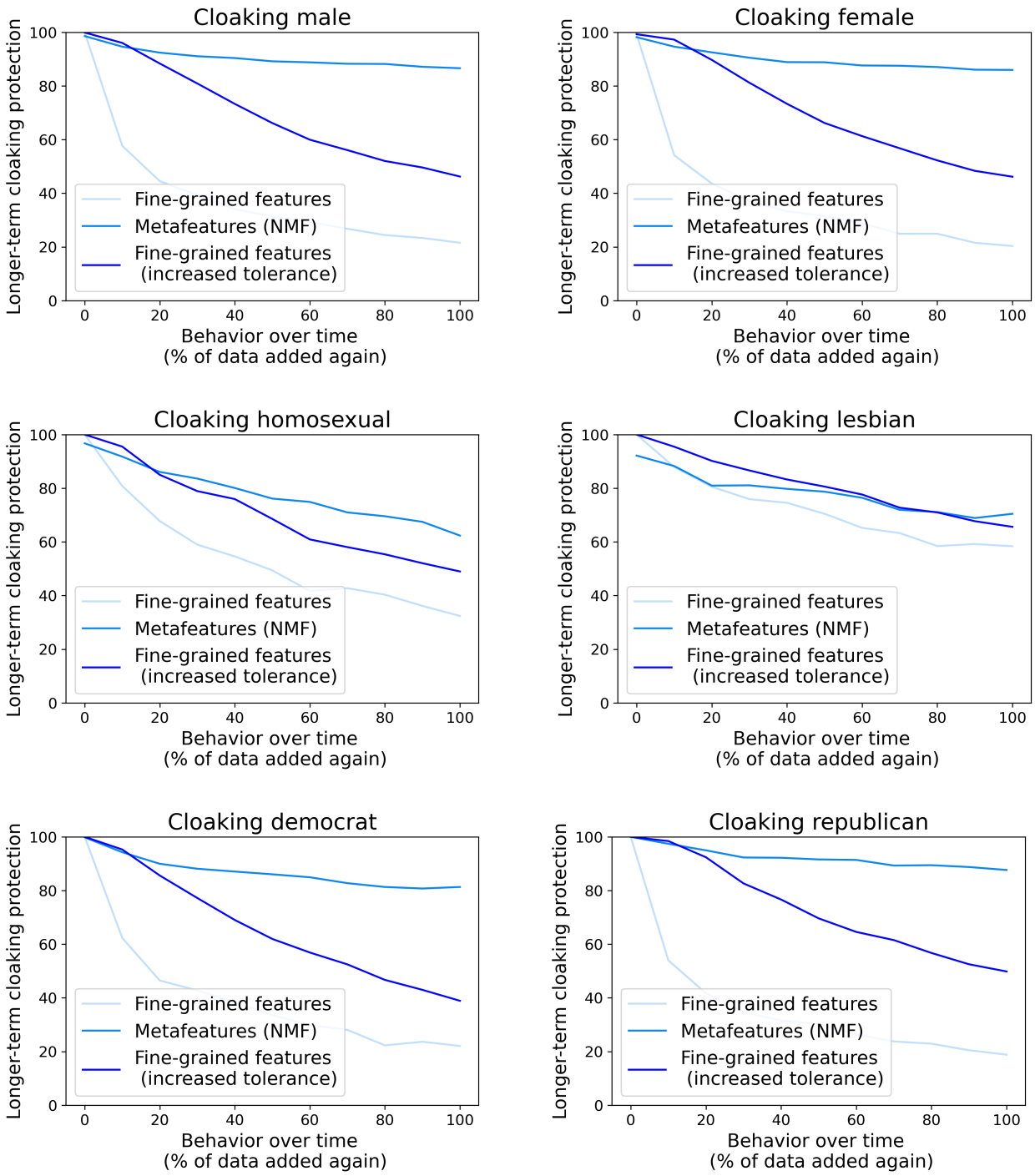


Figure 9: Longer-term cloaking protection over time when using explanations with an additional tolerance level.