Enhancing Transparency and Control when Drawing Data-Driven Inferences about Individuals

Daizhuo Chen
Columbia University

Samuel P. Fraiberger
New York University

Robert Moakler
NYU Stern School of Business

Foster Provost
NYU Center for Data Science

May 27, 2015

Working Paper CBA-15-01

Abstract

Recent studies show the remarkable power of information disclosed by users on social network sites to infer the users' personal characteristics via predictive modeling. In response, attention is turning increasingly to the transparency that sites provide to users as to what inferences are drawn and why, as well as to what sort of control users can be given over inferences that are drawn about them. We draw on the *evidence counterfactual* as a means for providing transparency into why particular inferences are drawn about them. We then introduce the idea of a "cloaking device" as a vehicle to provide (and to study) control. Specifically, the cloaking device provides a mechanism for users to inhibit the use of particular pieces of information in inference; combined with the transparency provided by the evidence counterfactual a user can control model-driven inferences, while minimizing the amount of disruption to her normal activity. Using these analytical tools we ask two main questions: (1) How much information must users cloak in order to significantly affect inferences about their personal traits? We find that usually a user must cloak only a small portion of her actions in order to inhibit inference. We also find that, encouragingly, false positive inferences are significantly easier to cloak than true positive inferences. (2) Can firms change their modeling behavior to make cloaking more difficult? The answer is a definitive yes. In our main results we replicate the methodology of Kosinski et al. (2013) for modeling personal traits; then we demonstrate a simple modeling change that still

gives accurate inferences of personal traits, but requires users to cloak substantially more information to affect the inferences drawn. The upshot is that organizations can provide transparency and control even into complicated, predictive model-driven inferences, but they also can make modeling choices to make control easier or harder for their users.

1 Introduction

Successful pricing strategies, marketing campaigns, and political campaigns depend on the ability to optimally target customers or voters. This generates incentives for firms and governments to acquire and exploit information related to people's personal characteristics, such as their gender, marital status, religion, sexual or political orientation. The boom in availability of online data has accentuated efforts to do so. However, personal characteristics often are hard to determine with certainty because of privacy restrictions. As a result, online marketers find themselves increasingly depending on statistical inferences based on available information. A predictive model can be used to give each user a score that is proportional to the probability of having a certain personal trait, such as being gullible, introverted, female, a drug user, gay, etc. [7]. Users then can be targeted based on their predicted propensities and the relationships of these inferences to a particular advertising campaign. Alternatively, such characteristics can be used implicitly in campaigns, via models trained on feedback from those who responded positively. In practice, usually a combination of model confidence and a budget for showing content or ads leads campaigns to target users in some top percentile of the score distribution given by predictive models [12].

Traditionally, online user targeting systems, particularly in digital advertising, have been trained using information on users' web browsing behavior [12]. However, a growing trend is to include information disclosed by users on social networks.[1] For instance, Facebook has recently deployed a system that allows third party applications to display ads on their platform using their user's profile information, such as the things they explicitly indicate that they "Like."[2]

While some online users may benefit from being targeted based on inferences of their personal characteristics, others may find such inferences unsettling. Not only may these inferences be incorrect due to a lack of data or inadequate models, some users may not wish to have certain characteristics inferred at all. To many, privacy invasions via statistical inferences are at least as troublesome as privacy invasions based on personal data [2]. In response to an increase in demand for privacy from online users, suppliers of browsers such

[1] https://www.facebook.com/about/ads/

[2] https://developers.facebook.com/blog/post/2014/10/07/audience-network

as Chrome and Firefox have developed features such as "Do Not Track," "Incognito," and "Private Windows" to control the collection of information about web browsing. However, these features provide neither clear transparency into what inferences are drawn and why, nor easy, fine-grained control over what information may be used for inference. Furthermore, as of now social networks such as Facebook do not have a strong analog to these features that would allow for transparency and control in how user information is used to decide on the presentation of content and advertisements.[3]

In this paper, as a means for providing transparency into the reasons why a particular inference is drawn about an individual, we draw on an idea introduced for explaining the reasons behind instance-level document classifications [9]. Specifically, what is a minimal set of evidence such that if it had not been present, the inference would not have been drawn? Let's call this an *evidence counterfactual*. The evidence counterfactual can be applied beyond document classification to the sorts of inference that interest us here. As a concrete example, consider that Manu has been determined by the system's inference procedure to be gay, based on the things that Manu has chosen to Like.[4] Keeping the inference procedure constant, what is a minimal set of Likes such that after their removal Manu would no longer be classified as being gay?

We then introduce the idea of a "cloaking device" as a vehicle to provide (and to study) control over inferences. Specifically, the cloaking device provides a mechanism for users to inhibit the use of particular pieces of information in inference; combined with the transparency provided by the evidence counterfactual a user could be given control over model-driven inferences. Importantly, the user can cloak particular information *from inference*, without having to stop sharing the information with his social network friends. Thus, hopefully, this combination will allow control with a minimal amount of disruption to the user's normal activity. However, this hope rests on the relationship between the evidence and the behavior of the predictive models.

In this paper we use these analytical tools to answer two main questions: (1) How much information must users cloak in order to significantly affect inferences about their personal traits? We find that generally a user does not need to cloak the majority of her information in order to inhibit inference. In fact, we find that for the most common (to our knowledge) online inference setting, users need to cloak only a small portion of the information recorded

[3]In 2014, Facebook developed a feature called "Why am I seeing this ad?" which gives users partial transparency on why they are being targeted. Users can also selectively cloak a particular categories of ads or advertisers; they can also modify their "ad preferences" to hide categories of information from being used for targeting. However it does not currently allow fine grained control over inferences of personal characteristics based on information displayed, which is the topic of the paper. We view this recent development by Facebook as strong support for the approach we propose here.

[4]We will capitalize "Like" when referring to the action or its result on Facebook.

about them. We also find that, encouragingly, false positive inferences are generally easier to cloak than true positive inferences. (2) Can firms change their modeling behavior to make cloaking more difficult? The answer is a definitive yes. In our main results we replicate the methodology of Kosinski et al. [7] for modeling personal traits; then we demonstrate a simple modeling change that still gives accurate inferences of personal traits, but requires users to cloak substantially more information to affect the inferences drawn. The upshot is that firms can provide transparency and control even into very complicated, predictive model-driven inferences, but they also can make modeling choices to make control easier or harder for their users.

The rest of the paper is organized as follows. Section 2 gives additional necessary background related to online user privacy, the evidence counterfactual, and control. Section 3 formalizes the concept of cloakability. Section 4 examines the effort needed to cloak various personal characteristics, using a dataset relating Facebook profiles to inferences about personal traits, showing the degree of cloakability observed across characteristics. The paper closes by discussing the results and their implications.

2 Privacy, Cloakability, and the Evidence Counterfactual

Online privacy is becoming an increasing concern for consumers, regulators and policy makers [16]. Treatments of privacy in the analytics literature often focus on the issue of confidentiality of personal characteristics (see [15, 11] for an overview). However, with the rapid increase in the amount of social media data available, statistical inference about personal characteristics is drawing attention [3, 2]. A series of papers have shown the predictive power of information disclosed on Facebook to infer users' personal characteristics [1, 7, 14]. The set of pages which users choose to "Like" on Facebook can predict their gender, religion, sexual or political orientation, and many more personal traits.

A recent study based on a survey of Facebook users found that they did not feel that they had the appropriate tools to mitigate their privacy concerns when it comes to social network data [4]. There is evidence that when given the appropriate tools, people will choose to give up some of the benefits they derive from their social network activity in order to meet their privacy concerns [6]. Besides being a conceptual tool to help with the analysis of control, the cloaking device can be a practical tool to achieve it.

Our notion of the evidence counterfactual is based on the work of [9] for explaining data-driven document classifications. Generalizing that work, consider any domain where the features taken as input can be seen as evidence for or against a particular non-default[5]

[5]The inference not being the default is important for explaining the reasons for model-based prediction.

inference. Consider also the increasingly common scenario [5] where there are a vast number of possible pieces of evidence, but any individual normally only exhibits a very small number of them—such as when drawing inferences from Likes on Facebook.[6] Thus, we can provide transparency by applying the methods presented by [9] to create one or more evidence counterfactual explanations for any non-default classification. Martens and Provost describe how to create evidence counterfactual explanations from any arbitrary predictive model. For our results below, we consider linear models, for which the procedure for computing the evidence counterfactual is straightforward, efficient, and optimal [9].

Given an individual, a specific model-based inference about the individual, and an evidence counterfactual explanation for why the inference was made, we can now describe the core design, use, and value of the cloaking device. The cloaking device allows the individual to hide (to "cloak") particular evidence, e.g., one or more Likes, from the inference procedure. Specifically, once a Like is cloaked, the inference procedure would remove it from its input, and therefore treat the user as if she had not Liked this item. The evidence counterfactual presents the user with a minimal set of Likes to cloak in order to change the inference made about her.

Consider the task of predicting whether or not a user is gay using Facebook Likes. While users might choose to disclose on the platform that they are gay, some may not wish to make this fact available to advertisers or others modeling online user behavior. A user who has not shared this status may not want it to be predicted by the system. In addition, a user who is in fact not gay may not want an incorrect inference to be drawn about him. Figure 1 illustrates two users, their probabilities of being gay as predicted by a model-based inference procedure, and the effect of removing evidence from their data. As evidence is removed by cloaking Likes, we see that removing fewer than ten Likes for one user results in a dramatic drop in the predicted probability of being gay, whereas for the same number of removals the probability is reduced hardly at all for the other user.

The cloaking device thus has two important dimensions of value. First, it provides us with a basis for studying the relationship between evidence and model-based inference, and thereby transparency and control, in settings such as these. Second it provides a practical

The default prediction is the prediction that is given when there is not enough evidence for predicting anything else, for example predicting that there is no fraud on a particular account. Thus, the *explanation* for a default prediction—that there is no evidence for any alternative—will be viewed as either trivial or unsatisfying. Usually the default inference is either the most common alternative or the least costly alternative, and very often these two concur. See [9] for further discussion and other nuances of explaining model-based inferences.

[6] As with predictive modeling projects generally, engineering the right representation often is key to top-level performance. So for example, one might code the lack of a particularly popular Like as positive evidence. We will only consider the presence of a Like in our results, but our qualitative results should generalize across such alternative representation engineering.

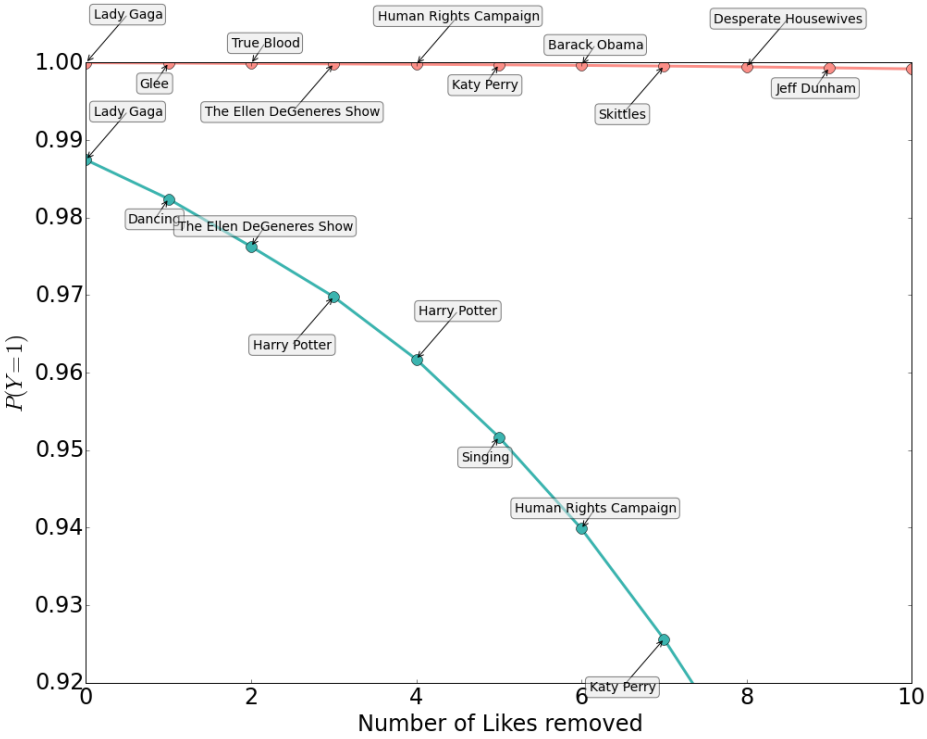


Figure 1: The predicted probability of being gay as a function of Like cloaking for two users. For each line, the leftmost point is the estimated probability of being gay for the user before cloaking. Moving left to right, for each user, Likes are removed one-by-one from consideration by the inference procedure in order of greatest effect on the estimated score. One user's probability drops dramatically with cloaking fewer than ten Likes; the other's is hardly affected at all.

device that could be implemented by social media sites (and others) to provide such transparency and control to its users. This paper focuses on the former, both for its own intrinsic interest and also as potential support for the latter.

3 A Model of Cloaking

In this section we describe the technical details of the cloaking device as used for the results in this paper. We do not know the value each user attaches to each piece of information he chooses to reveal on the platform. We assume uniformity, essentially quantifying the minimum amount of information to be removed not to be the target of a particular inference.[7]

As described above, cloaking is defined in the context of a predictive model used by an entity that engages in digital user modeling and inference, for example for targeting online content or ads. We assume the model to be fixed.[8] We consider a supervised classification or ranking task, which can be described by a linear model.[9] All of the features and targets in these models are assumed to be binary. In particular, our main model replicates the predictive modeling used by [7] and use their data on predicting personal traits from Facebook Likes. More specifically, the modeling procedure first reduces modeling dimensionality by computing the singular-value decomposition (SVD) of the matrix of users and their Likes, and choosing the top-100 SVD dimensions' vectors as the modeling dimensions (as has become standard practice with such massively dimensional data). Then logistic regression models are built on these dimensions to predict a variety of personal traits, as detailed below.

For inference we simulate what is to our understanding the most common method of taking online actions based on such models. Specifically, we assume that a positive inference is drawn—e.g., a user would be subject to targeting—if the model assigns the user a score placing him in a specified top quantile (δ) of the score distribution produced by the predictive model.[10]

More formally, let $x_{ij} \in \mathbf{x}$ be an indicator equal to 1 if user i has Liked a piece of information j and 0 otherwise. For the main results we build the SVD-logistic regression model described above; then we convert it to a mathematically (and functionally) equivalent linear logistic regression (LRSVD) model in the original features $\mathbf{x}$, via the transformation

[7] An extension to this work that we do not consider here could consider minimum-cost cloaking, removing the subset of evidence (Likes) with minimal cost to the user.

[8] For the sake of simplicity, we assume either that new models are put into production infrequently, or that Likes are not cloaked from model learning. Beyond the scope of this paper, there are interesting possible dynamics between large numbers of users cloaking evidence from *learning* and the changes in the resultant models.

[9] For extensions to nonlinear models see [9].

[10] For example, for targeting online ads, a typical value for δ would range between $90\% - 100\%$. Perlich et al. [12] describe in detail online targeting with predictive models based on fine-grained user data.

described in the appendix A. This transformation facilitates direct manipulation of the original Likes. From now on unless stated otherwise we will consider this linear logistic model.

Let β_j be the coefficient in the model associated with feature $j \in \{1; ...; J\}$. Without loss of generality, assume that these are ranked by decreasing value of β. Each such coefficient corresponds to the marginal increase in a user's score if he chooses to Like feature j. Let s_i be the model output score given to user i, which ranks users by their probability of having a characteristic s. It is given by

$$s_i = \sum_{j=1}^{J} \beta_j x_{ij}. \tag{1}$$

For simplicity, let's call those users for whom the positive inference is made the "targeted" users. For a particular set of users, define the cutoff score s_δ to be the score of the highest-ranked user in the quantile directly below the targeted users. Thus the set of targeted, top-ranked users T_s for classification task s is

$$T_s = \{i | s_i > s_\delta\}. \tag{2}$$

To analyze the difficulty or ease of cloaking for each user in the targeted group, we iteratively remove Likes from his profile until he is successfully cloaked. For our linear models we do this by iteratively subtracting from his score the coefficient of the feature that is present in his data instance that has the largest coefficient in the model. Figure 1 shows two examples. A user is considered to be successfully cloaked when his score falls below s_δ.[11,12]

Figure 2 shows the discriminative power associated with each Like in our data for the task of predicting if individual male users are gay. The ten points with associated text labels are Likes that have the largest coefficients from the LRSVD model. The top ten Likes for the user shown in red in figure 1 are shown here as red points. Six out of this user's top-10 Likes overlap with the top ten for the entire task. This highlighted user is the user that the LRSVD model predicts as having the highest probability of being gay.

To quantify Like removal and the difficulty of cloaking, we let $\eta^s_{i,\delta}$ represent the effort to cloak user i from the top $\delta\%$ of the score distribution for a characteristic s. $\eta^s_{i,\delta}$ is defined

[11]If the targeted group is defined by a fixed threshold score (such as the estimated probability being above a fixed threshold), this is straightforward. If the targeted group is defined instead based on the actual quantile, then when a user is removed from the targeted group another user takes his place. In this paper we consider users in isolation and do not consider the effects of cloaking on sets of users.

[12]More generally, for non-linear models the evidence counterfactual would reveal a minimal set of Likes such that their removal would successfully cloak the individual [9].

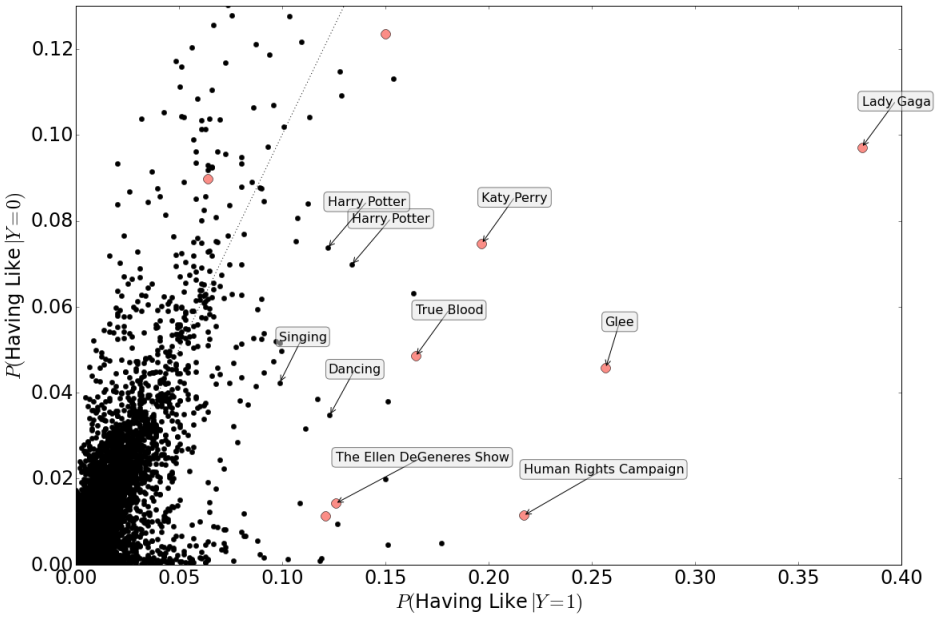


Figure 2: The discriminative power of Likes on Facebook when determining if a user is gay ($Y = 1$). Labels are given to the top ten Likes as sorted by their corresponding coefficients from the LRSVD model. Points colored in red are the top ten pages Liked by the user with the highest probability of being gay as predicted by the LRSVD model. This is the same user that appeared in red in figure 1.

precisely in algorithm 1; it is the minimum number of Likes that must be removed to move i below the threshold. All else being equal, the effort to cloak a user is smaller when (i) the coefficients of his removed features are larger, (ii) the threshold score is larger, and/or (iii) his predicted score is smaller.

Algorithm 1: Algorithm to determine the amount of effort needed to cloak a user for a particular predictive task.

$\eta^s_{i,\delta} \leftarrow 0$
$j \leftarrow 1$
Sort coefficients β in descending order as $1...J$
while $s_i > s_\delta$ **do**
  $s_i \leftarrow s_i - \beta_j$
  $\eta^s_{i,\delta} \leftarrow \eta^s_{i,\delta} + 1$
  $j \leftarrow j + 1$
end

The absolute effort to cloak a particular classification task s is given by averaging $\eta^s_{i,\delta}$ across users in T_s,

$$\eta^s_\delta = \frac{\sum_{i \in T_s} \eta^s_{i,\delta}}{|T_s|}. \tag{3}$$

Alternatively, we can examine the relative effort to cloak a task for user i, defined by normalizing the absolute effort by the total quantity of information revealed by the user,

$$\pi^s_{i,\delta} = \frac{\eta^s_{i,\delta}}{\sum_{j=1}^{J} x_{ij}}. \tag{4}$$

We can then define the relative effort to cloak a classification task s by averaging this measure across users in T_s,

$$\pi^s_\delta = \frac{\sum_{i \in T_s} \pi^s_{i,\delta}}{|T_s|}. \tag{5}$$

For the rest of this paper we use $\delta = 0.90$ to indicate that the top 10% of users are being targeted.[13]

4 Results

Let us now examine the effort required to cloak the inferences of a variety of personal characteristics, based on data on Facebook users. We first describe the data, and then

[13]For other values of δ the results hold qualitatively.

proceed to assess the effort required to cloak user characteristics.

4.1 Data

Our data were collected through a Facebook application called myPersonality.[14] It contains information on 164,883 individuals from the United States, including their responses to survey questions and a subset of their Facebook profiles. Users can be characterized by their sexual orientation, gender, political affiliation, religious view, IQ, alcohol and drug consumption behavior, personality dimensions, and lifestyle choices. (Users do not necessarily reveal all of these personal characteristics.) For these users we also know their Facebook Likes.

The personal characteristics are the target variables for the various modeling and inference problems. Some personal characteristics were extracted directly from users' Facebook profiles, whereas others were collected by survey. Binary variables are kept without change. Variables that fall on a Likert scale are separated into two groups, users that have the largest Likert value and users that have any other value. Continuous variables are represented as binary variables using the 90^{th} percentile as a cutoff. Multi-category variables are subsampled to only include the two most frequent categories, with the instances representing the other categories discarded for the corresponding inference task. Notice also that the feature data are very sparse; for each characteristic a user displays less than 0.5% of the set of Likes on average. Table 1 presents summary statistics of the data.

4.2 Replicating the prior prediction results

We first replicate the predictive modeling and inference procedure reported by [7]. Specifically, we build the predictive models on the SVD dimensions in Python using logistic regression as implemented in the scikit-learn package. For each model, we choose the regularization parameter by (5-fold) cross validation, as is the state-of-the-art practice [13]. Appendix B reports the predictive performance across the set of tasks. The results concur with those reported by [7]. As in the original paper, the predictive performance is quite strong across the classification tasks.

4.3 Main result: How hard is it to cloak?

Table 2 reports the efforts to cloak users that belong to the target group (*i.e.* those users in the top 10% of users as ranked by model score). First, we will focus on the "All" columns (in the next section we break down the results by true positives and false positives). The

[14] Thanks to the authors of [7] for sharing the data.

results show that although users on average display hundreds of Likes, on average they need to cloak fewer than 10 to successfully inhibit inference. This corresponds to cloaking only about $2-3\%$ of of a user's Likes, on average. Digging a little deeper, the prediction tasks are sorted in table 2 by π, showing that the averages give a fair picture: with only a couple exceptions the proportion of information needed to inhibit inference is around $2-4\%$. The actual numbers of Likes that must be removed vary more, as the top-decile users have different total numbers of Likes, but nevertheless we see no extreme outliers.

To put these results in context it would be useful to know how strongly the cloakability of a trait is related to the statistical dependency structure of the data-generating process. One might think that people who indeed hold a particular trait would exhibit it throughout their behavior, and in particular throughout the things that they Like. How do these cloakability results compare to what one would expect if Likes and the trait were not actually interrelated?

To draw this comparison, we conduct a randomization test to assess both qualitatively and quantitatively whether cloakability on these real individuals is indeed harder than it would be in the absence of this statistical interdependency. We first create a sampling distribution to be used to randomly assign Likes to individuals. We want only to remove the interdependency between the Likes and the dependency between the target and the Likes, so we retain the general popularity of Likes as follows (otherwise, due to the skew in popularity, individuals would have collections of oddly unpopular Likes). For each prediction task (personal trait), we assign to each Like a weight equal to the fraction of users for that task who have that particular Like; we then normalize the set of weights so that their sum is equal to one in order to create a sampling distribution. Then, for each user, we draw from this distribution a set of Likes without replacement. For each user we draw the same number of Likes as in the original dataset. Thus, in the resultant population the popularity distribution over the Likes is the same as in the original data, and the numbers of Likes that people have is the same, and the relationship between the number of Likes and the target trait is the same. However, there are no statistical dependencies among the Likes or between the Likes and the trait. This procedure is repeated 1,000 times and each time we apply the same procedure as above to the new population, computing the values of $\eta_{0.9}$ and $\pi_{0.9}$. This results in a distribution over $\eta_{0.9}$ and $\pi_{0.9}$ when the dependencies are removed.

Figure 3a shows the difference between $\eta_{0.9}$ in the no-dependency population and the true $\eta_{0.9}$. Quantitatively, for all tasks we find that the actual absolute effort to cloak is always higher ($p < 0.01$, sign test) than cloaking would be if Likes were randomly assigned. Qualitatively, we see that indeed cloaking seems very easy in the random case. In all but three cases, one needs to cloak fewer than two Likes on average to inhibit inference. In all cases, inference can be inhibited by cloaking fewer than four Likes on average. The figure

shows that generally the statistical dependency structure renders cloaking several times harder than it would otherwise be.

Figure 3b shows the difference in the relative effort to cloak, $\pi_{0.9}$, between the randomized setting and the true setting. Here the highest level result is the same: in every case the relative effort is no worse than in the true setting ($p < 0.01$, sign test). However, some of the differences quantitatively are not as striking as in the comparison of absolute effort. In fact, in one case ("is lesbian") the difference is essentially zero. This seeming paradox is explained by the fact that the numbers of Likes for the true top-decile individuals can be quite different from the numbers of likes for the top-decile individuals in the no-dependency setting. So, for example, the actual top-decile individuals for "is lesbian" have twice as many Likes on average as the top-decile individuals in the randomized setting.

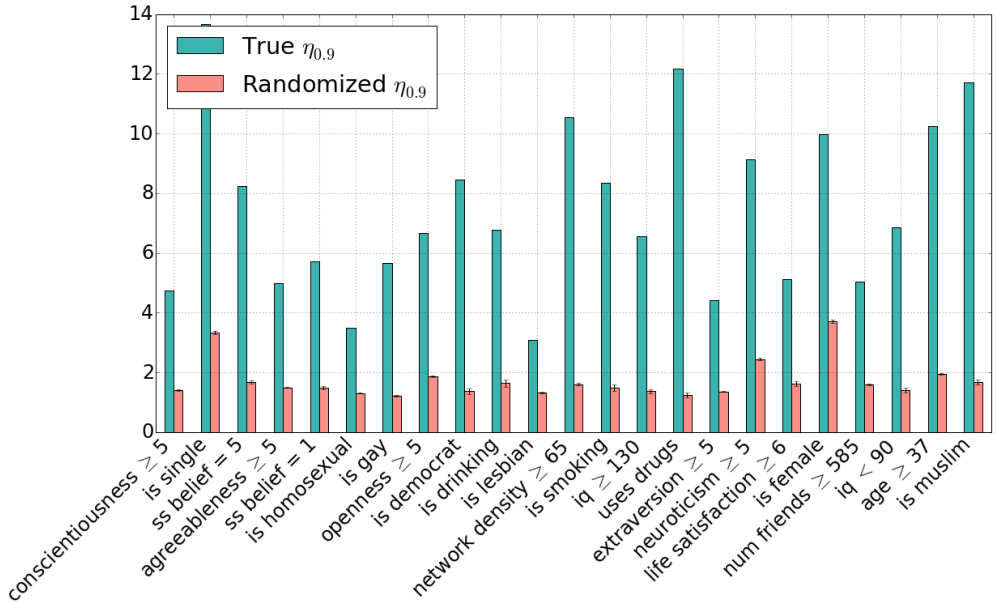
The upshot is that although in an absolute sense it is relatively easy to inhibit inference by cloaking Likes, the statistical dependence structure among the Likes and the predicted trait makes it more difficult than it would be without such structure. This has an important implication to which we will return in the discussion section.

4.4 Cloaking true positives vs. false positives

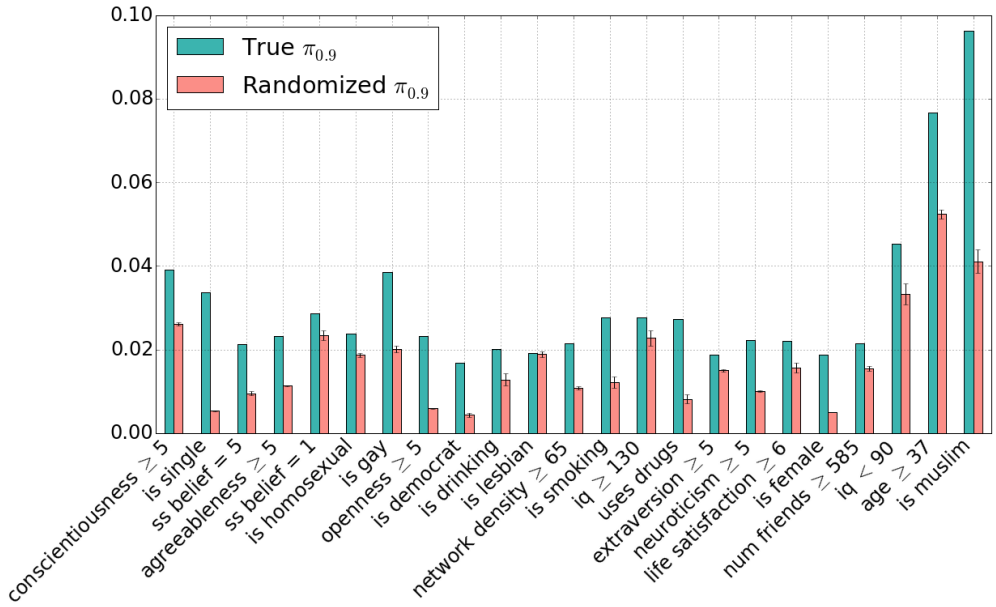
At the outset we introduced the idea that there are multiple settings where one might want to inhibit inference. Possibly the most important distinction is between inhibiting an inference that is in fact true (a true positive inference) and inhibiting an inference that is false (a false positive inference).

Based on the prior results, one might expect that a false positive inference would be easier to cloak because the statistical dependency to the (positive) trait is by definition missing. Thus, in a sense the false positive user "accidentally" received the inference, similarly to how the top-decile randomized users "accidentally" ended up in the class. In neither case was the presence of the trait reflected in the behavior of the user. However, there is an important distinction: in the randomized setting the statistical dependencies also were broken among the Likes, as opposed to simply between each Like and the target trait. For false positives, intuitively there still may be strong statistical interdependencies between the Likes—so if one has some Likes that trigger the inference by the predictive model, one may have many Likes that trigger the inference.

Thus, in addition to measuring the cloakability across all users in the targeted group, table 2 also reports the same results for true-positive (TP) and false-positive (FP) users separately. The results show that cloaking is indeed generally more difficult for true-positive users than for false-positive ($p < 0.05$, sign test). The differences in cloakability between



(a) Comparison for absolute effort to cloak ($\eta_{0.9}$).



(b) Comparison for relative effort to cloak ($\pi_{0.9}$).

Figure 3: Comparison between absolute ($\eta_{0.9}$) and relative ($\pi_{0.9}$) effort to cloak in the LRSVD model. Results from the normal cloaking procedure are compared to those of a randomization test for each task. Error bars depict the 95% confidence interval.

true-positive and false-positive users are shown in figure 4.

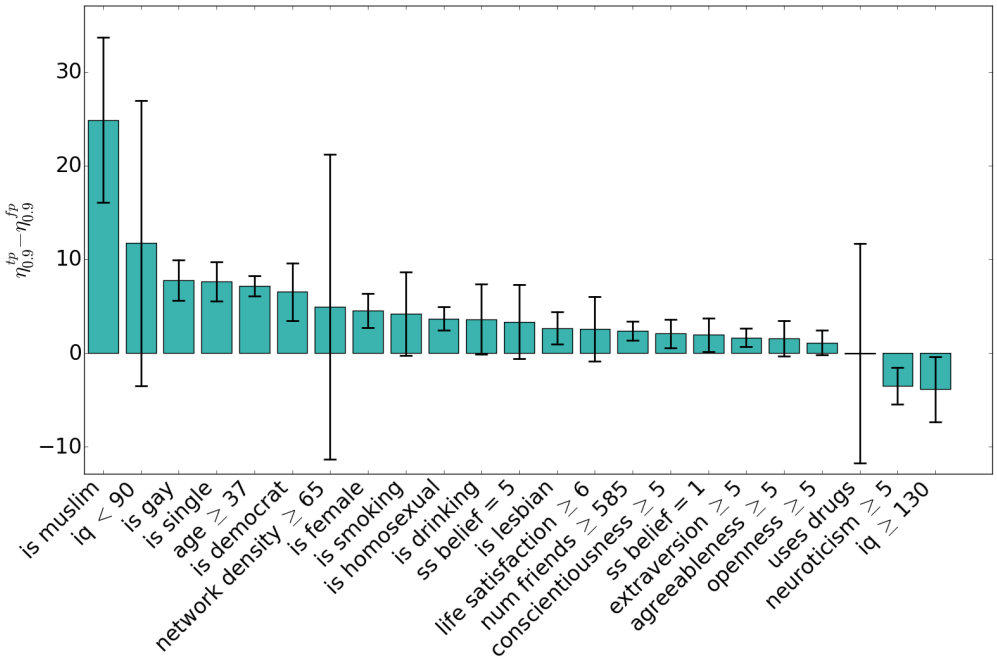
These results may provide some intuitive satisfaction. It is relatively easier to "fix" an incorrect classification, than to "hide" from a correct inference. The most striking example of this is in prediction for the "is muslim" trait. On average, to inhibit the positive inference for someone who actually is Muslim, 28 Likes have to be cloaked. This is almost twice as many as for any other trait. On the other hand, to inhibit the "is muslim" classification for a non-Muslim, only 3 traits need to be cloaked. This suggests a line of future inquiry: does this illustrate a case of a strong dependency between a personal trait and the individual's choice of actions? Or is there some alternative explanation having to do with the subtleties of predictive modeling? Other such examples can be see, although to a lesser extent, for "age ≥ 37", "IQ < 90", and "is gay".

A comprehensive analysis of this question is beyond the scope of this paper, however we can offer an initial view. Besides the statistical dependency relationships discussed above, the observed differences in cloakability for the true-positive and false-positive users can also be attributed to the interaction between two factors: variance in predicted probability and the order in which each model ranks the users subject to prediction. For some tasks we find that the predicted probabilities for all users in the targeted group are tightly clustered; other tasks have a wide range or probabilities. Within the targeted group, each model finds itself discriminating between TP and FP users differently. Some models see a majority of TP users being ranked above FP users, while others find TP and FP to be mixed. If a majority of FP users find themselves ranked below their TP counterparts, ceteris paribus they will be easier to cloak simply because they are closer to the threshold. Additionally, if the variance in predicted probability is large, and many FP users fall at the lower end of the targeted range, again the FP users will find it easier to cloak themselves from inference.

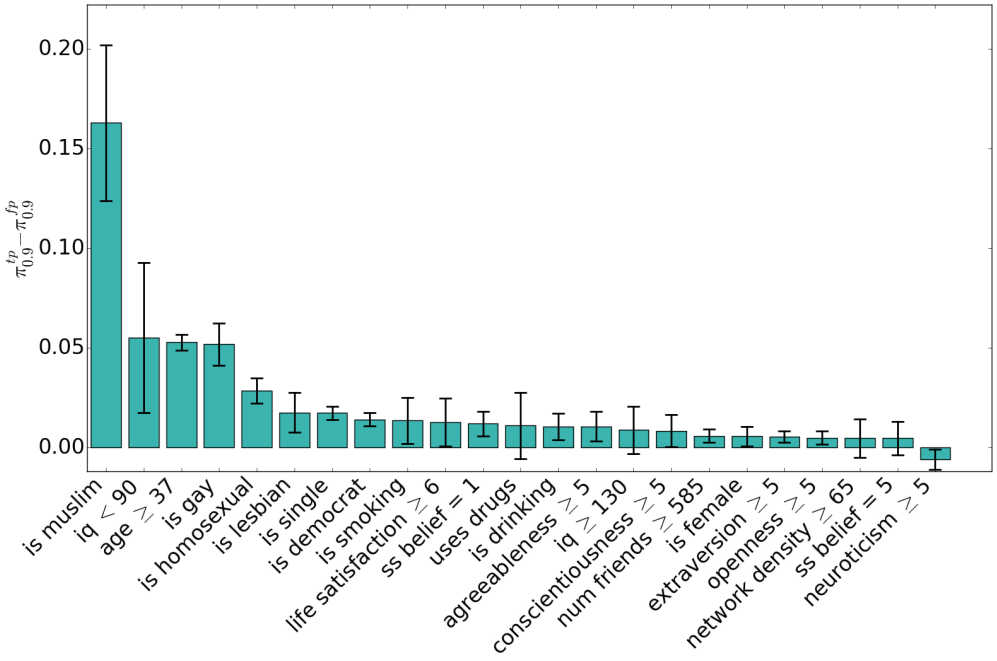
5 Discussion

In the previous section, we showed that inhibiting inference requires cloaking only a relatively small amount of personal information—only around seven (3%) out of one's hundreds of Likes on average, and that the statistical dependence structure among the Likes and the predicted trait makes cloaking more difficult than it would be without such dependency structure. However, we showed this for a particular predictive model and modeling procedure—even though it is a best-practices modeling procedure, we did not show that cloaking would be easy using any predictive model.

Could it be that organizations could make different modeling decisions that would allow them still to predict accurately and offer transparency and control with a cloaking device,



(a) Difference in cloaking, $\eta_{0.9}$.



(b) Difference in cloaking, $\pi_{0.9}$.

Figure 4: Difference in cloaking, $\eta_{0.9}$ and $\pi_{0.9}$, for true positive and false positive users using the LRSVD model. Error bars depict the 95% confidence interval.

but make it much harder for the users actually to cloak themselves? Those who run some organizations may be quite happy to provide transparency and easy control, either because they believe it is simply the right thing to do, or because they believe that it will increase user/customer satisfaction, or even because they believe it will be more profitable as the targeting actually will be better. Others may want to give the semblance of transparency and control, but actually dissuade users from manipulating their profiles to cloak. To explore whether a targeter can manipulate cloakability through modeling choices, let's briefly examine two alternative model choices.

The naive Bayes model (NB) is a linear model quite similar to logistic regression,[15] but with a certain particularity. Naive Bayes assumes that the pieces of evidence taken as input (the Likes) are conditionally independent of each other given the target (the trait). Mechanically, the algorithm for inducing the NB model from data treats each Like independently. When the Likes in fact are highly correlated, this creates a pathology in predictive behavior: the resulting inference model will tend to "double count" when users present correlated Likes. However, our unscrupulous targeter may decide to use this pathology to its advantage. The model will tend to give extra high scores when correlated evidence is presented, and because of the double counting, the user would have to cloak multiple Likes to achieve the same effect as in a model that does not exhibit this pathology (like the LRSVD model).[16]

For completeness, in addition to the LRSVD model and the NB model we also will examine a straightforward logistic regression model (LR) trained on the full (non-SVD) raw Like feature space. We would expect the results for LRSVD and LR to be similar, but that the NB model would require significantly more cloaking to inhibit inference.

Table 3 presents the values for our cloaking measure across different models.[17] As expected, the cloaking efforts required for the LR and LRSVD models are similar. In contrast, cloaking is indeed substantially more difficult for NB. Rather than needing to cloak only a half-dozen or so Likes, for the NB models users on average have to cloak 57 Likes. This is on average 15% of a user's Like set. At the extreme, an average person classified as "is Muslim" has to cloak 50% of her Likes! A person classified as "conscientiousness ≥ 5" has to cloak 44% of her Likes. Classified as "is female"? With the NB model you'll have to cloak over 377 (25%) of your Likes to escape that classification.

In summary, a targeter wishing to make cloaking more difficult could do so without

[15] Indeed equivalent under certain assumptions [10].

[16] Technically, since many Likes that supply evidence of a user being part of the positive class are highly correlated with one another, the NB modeling will essentially assign all of these Likes high coefficients whereas the LR modeling spreads the overall impact across the coefficients of the correlated Likes (in one way or another depending on the type and degree of regularization).

[17] The predictive (generalization) performance for the NB model is slightly lower than that for the logistic regression models. For details, see appendix B.

imposing any restrictions on their users by changing their predictive model choice. While it is clear that Like pages do not conform to the independence assumption inherent to naive Bayes, we find that across all tasks (with the exception of "is female") that the difference in predictive performance (measured by the area under the ROC curve (AUC) as in [7]) between LRSVD/LR and NB models is only 5% on average. Thus, by taking a small loss in predictive performance, it is possible to make cloaking significantly more difficult.

6 Conclusion

In this paper we develop a method to give online users transparency into why certain inferences are made about them by statistical models, and control to inhibit those inferences by hiding ("cloaking") certain personal information. We use this method to examine whether such transparency and control would be a reasonable goal, by assessing how difficult it would be for users to actually inhibit such inferences. The method is applied to data from a large collection of real users on Facebook, where prior work has shown that predictive models can infer their personal characteristics with high accuracy from their Likes.

The results show that the amount of effort users must exert in order to successfully hide themselves is quite small. Although it is higher than if there were no statistical dependency among the Likes and the personal traits, the users still need only to cloak about a half-dozen of their hundreds of Likes on average to inhibit inference of a personal trait. Users for whom the inference made is actually wrong have an even easier time cloaking the inference.

However, organizations engaging in such modeling can alter their modeling choices to make cloaking increasingly difficult. The results show that, at the expense of a small amount of predictive performance, targeters can choose different types of predictive models that will leverage the interdependence of features to inflate cloaking difficulty. In extreme cases, even a simple modeling change can, for certain traits, raise the amount of Likes needing to be cloaked up to hundreds of Likes (from a half-dozen!). In these extreme cases, the increase in the number of cloaked Likes can result in having to cloak 20% of a user's profile (from 2%).

We propose three directions for future research. First, instead of treating all features as having a uniform weight, the relative importance for each can be factored into the decision criteria if known. This allows for cloaking to be measured using metrics beyond the minimal set we have already investigated. As real users may be unlikely to view all of their decisions as being equally important to them, the results for such an analysis may be quite different from what we have already seen. Second, as mentioned previously, we do not have a clear answer as to whether there is a strong dependency between a personal trait and an individual's choice

of actions. Modifications to our randomization test and drawing on behavioral knowledge for these user traits may offer insight into this question. Third, as digital data becomes increasingly centered on inherent network structures, expanding the set of features to utilize network based measures can have a dramatic effect on inference and cloakability. [8] shows how collective inference can improve the performance of a predictive model in the context of networked data. In our setting, utilizing network data could lead to not only removing features, but to suggesting the removal of friends in order to avoid being targeted.

Acknowledgements

Thank you very much to Michal Kosinski, David Stillwell and Thore Graepel for sharing their data. Thanks to Wally Wang for helpful discussions at the outset of this project. Foster Provost thanks Andre Meyer for a Faculty Fellowship. We also thank the Moore and Sloan Foundations for their generous support of the Moore-Sloan Data Science Environment at NYU.

References

[1] Bachrach, Y., Kosinski, M., Graepel, T., Kohli, P., and Stillwell, D. Personality and patterns of facebook usage. In *proceedings of the 3rd annual ACM web science conference* (2012), ACM, pp. 24–32.

[2] Barocas, S. *Panic Inducing: Data Mining, Fairness, and Privacy*. Phd dissertation, New York University, 2014.

[3] Harald Schoen, D. G.-A., Metaxas, P. T., Mustafaraj, E., Strohmaier, M., and Gloor, P. The power of prediction with social media. *Internet Research, Vol. 23 Iss: 5, pp.528 - 543, (2013)*.

[4] Johnson, M., Egelman, S., and Bellovin, S. M. Facebook and privacy: It's complicated. *Symposium On Usable Privacy and Security (SOUPS), July 2012*.

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

[6] Knijnenburg, B., Kobsa, S. M., and Jin, H. Counteracting the negative effect of form auto-completion on the privacy calculus. *Thirty Fourth International Conference on Information Systems, Milan 2013*.

[7] Kosinski, M., Stillwell, D., and Graepel, T. Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences 110*, 15 (2013), 5802–5805.

[8] Macskassy, S., and Provost, F. Classification in networked data: A toolkit and a univariate case study. *The Journal of Machine Learning Research 8* (2007), 935–983.

[9] Martens, D., and Provost, F. Explaining documents' predicted classifications. *MIS Quarterly 38(1), 73-99, 2014.*.

[10] Mitchell, T. M. *Machine Learning – Additional Chapters*, 1 ed. McGraw-Hill, Inc., New York, NY, USA, 1997.

[11] Pavlou, P. A. State of the information privacy literature: where are we now and where should we go. *MIS quarterly 35*, 4 (2011), 977–988.

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

[13] Provost, F., and Fawcett, T. *Data Science for Business: What you need to know about data mining and data-analytic thinking.* " O'Reilly Media, Inc.", 2013.

[14] Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., Shah, A., Kosinski, M., Stillwell, D., Seligman, M. E., et al. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one 8*, 9 (2013), e73791.

[15] Smith, H. J., Dinev, T., and Xu, H. Information privacy research: an interdisciplinary review. *MIS quarterly 35*, 4 (2011), 989–1016.

[16] White House. Consumer data privacy in a networked world: A framework for protecting privacy and promoting innovation in the global economy. *Washington, D.C.: White House.*.

Appendices

A Singular Value Decomposition (SVD)

The performance of a Logistic regression model can be improved by reducing the set of features if it is very large or if the data are sparse. A common technique is to use a singular

value decomposition (SVD).

Let M be a feature matrix which contains n records and m features. M can be decomposed into:

$$M = U\Sigma V^*. \tag{6}$$

In the above decomposition, U is an $n \times n$ unitary matrix, Σ is an $n \times m$ diagonal matrix composed of the singular values of M sorted in descending order, and V^* is the $m \times m$ conjugate transpose of the unitary matrix V. To reduce the space, we can choose to only include a subset of the first k features from the matrix Σ when training a new model.

A model trained on this reduced feature space will not yield coefficients for each of the original features. A simple transformation will allow for a mapping between a model trained on the SVD space to the original set of features before the reduction. Let β_{SVD} be the set of coefficients from the linear model trained on the SVD space and let β be the coefficients on the original set of features. We map from one to the other by:

$$\beta = \beta_{\text{SVD}}\Sigma^{-1}V^*. \tag{7}$$

B Classification Performance

Table 4 reports the AUC across classification tasks and across different predictive models.

Task	Number Users	Number Likes	% Positive	Average Likes
age $\geq$ 37	145,400	179,605	0.127	216
agreeableness $\geq$ 5	136,974	179,440	0.014	218
conscientiousness $\geq$ 5	136,974	179,440	0.018	218
extraversion $\geq$ 5	136,974	179,440	0.033	218
iq $\geq$ 130	4,540	136,289	0.130	186
iq $<$ 90	4,540	136,289	0.073	186
is democrat	7,301	127,103	0.596	262
is drinking	3,351	118,273	0.485	262
is female	164,285	179,605	0.616	209
is gay	22,383	169,219	0.046	192
is homosexual	51,703	179,182	0.035	257
is lesbian	29,320	175,993	0.027	307
is muslim	11,600	148,943	0.050	238
is single	124,863	179,605	0.535	226
is smoking	3,376	118,321	0.237	261
life satisfaction $\geq$ 6	5,958	141,110	0.125	252
network density $\geq$ 65	32,704	178,737	0.012	214
neuroticism $\geq$ 5	136,974	179,440	0.004	218
num friends $\geq$ 585	32,704	178,737	0.140	214
openness $\geq$ 5	136,974	179,440	0.043	218
ss belief = 1	13,900	169,487	0.178	229
ss belief = 5	13,900	169,487	0.079	229
uses drugs	2,490	105,001	0.172	264

Table 1: Summary statistics of the dataset. Number of Likes indicates how many unique Like pages are associated with a given task. Percent positive are how many positive instances there are for each task. Average Likes indicates the average number of Likes a user associated with the given task has.

Task	$\eta_{0.9}$ All	$\eta_{0.9}$ TP	$\eta_{0.9}$ FP	$\pi_{0.9}$ All	$\pi_{0.9}$ TP	$\pi_{0.9}$ FP
is democrat	8.462	8.533	2.000	0.017	0.017	0.003
is female	9.971	10.015	5.475	0.019	0.019	0.013
extraversion ≥ 5	4.428	5.944	4.300	0.019	0.024	0.018
is lesbian	3.075	5.437	2.829	0.019	0.035	0.017
is drinking	6.771	7.463	3.875	0.020	0.022	0.012
num friends ≥ 585	5.043	6.556	4.197	0.021	0.025	0.019
ss belief $= 5$	8.251	11.098	7.760	0.021	0.025	0.021
network density ≥ 65	10.545	15.308	10.388	0.021	0.026	0.021
neuroticism ≥ 5	9.140	5.667	9.173	0.022	0.016	0.022
life satisfaction ≥ 6	5.128	7.214	4.642	0.022	0.032	0.020
openness ≥ 5	6.674	7.677	6.571	0.023	0.028	0.023
agreeableness ≥ 5	4.985	6.508	4.957	0.023	0.033	0.023
is homosexual	3.493	6.572	2.888	0.024	0.047	0.019
uses drugs	12.161	12.143	12.176	0.027	0.033	0.022
is smoking	8.357	9.800	5.621	0.028	0.032	0.019
iq ≥ 130	6.566	3.429	7.283	0.028	0.035	0.026
ss belief $= 1$	5.738	6.880	4.946	0.029	0.036	0.024
is single	13.665	15.514	7.888	0.034	0.038	0.021
is gay	5.653	10.944	3.161	0.038	0.074	0.022
conscientiousness ≥ 5	4.746	6.746	4.670	0.039	0.047	0.039
iq < 90	6.867	16.318	4.582	0.045	0.090	0.035
age ≥ 37	10.259	13.011	5.847	0.077	0.097	0.044
is muslim	11.706	27.804	2.930	0.096	0.202	0.039
Mean	**7.465**	**9.851**	**5.572**	**0.031**	**0.045**	**0.023**
Median	**6.771**	**7.677**	**4.946**	**0.023**	**0.033**	**0.021**

Table 2: The effort to cloak different users' characteristics using the logistic regression with the 100-SVD-component logistic regression (LRSVD) model. Absolute efforts are presented in the left panel, and relative efforts are in the right panel. For each panel, we show in the first column the full set of users, in the second column only the true positive users, and in the third column only the false positive users (the negative users falsely targeted).

Task	$\eta_{0.9}$ LRSVD	$\eta_{0.9}$ LR	$\eta_{0.9}$ NB	$\pi_{0.9}$ LRSVD	$\pi_{0.9}$ LR	$\pi_{0.9}$ NB
is democrat	8.462	9.396	61.736	0.017	0.02	0.106
is female	9.971	11.619	377.436	0.019	0.020	0.259
extraversion ≥ 5	4.428	3.617	58.048	0.019	0.025	0.102
is lesbian	3.075	2.507	7.361	0.019	0.039	0.136
is drinking	6.771	5.398	17.145	0.02	0.021	0.082
num friends ≥ 585	5.043	4.748	52.599	0.021	0.025	0.106
ss belief $= 5$	8.251	4.692	18.432	0.021	0.041	0.062
network density ≥ 65	10.545	2.569	75.717	0.021	0.039	0.077
neuroticism ≥ 5	9.140	2.292	254.467	0.022	0.036	0.180
life satisfaction ≥ 6	5.128	4.061	10.297	0.022	0.072	0.083
openness ≥ 5	6.674	3.700	28.650	0.023	0.025	0.111
agreeableness ≥ 5	4.985	2.871	7.227	0.023	0.043	0.126
is homosexual	3.493	3.396	8.212	0.024	0.039	0.108
uses drugs	12.161	8.161	31.548	0.027	0.034	0.090
is smoking	8.357	7.012	26.190	0.028	0.032	0.135
iq ≥ 130	6.566	2.920	14.381	0.028	0.033	0.094
ss belief $= 1$	5.738	4.550	24.207	0.029	0.036	0.104
is single	13.665	10.233	105.794	0.034	0.028	0.125
is gay	5.653	9.073	20.597	0.038	0.150	0.153
conscientiousness ≥ 5	4.746	3.357	16.091	0.039	0.048	0.441
iq < 90	6.867	3.681	21.619	0.045	0.073	0.072
age ≥ 37	10.259	7.263	37.746	0.077	0.074	0.179
is muslim	11.706	8.934	31.090	0.096	0.101	0.465
Mean	**7.465**	**5.48**	**56.808**	**0.031**	**0.046**	**0.148**
Median	**6.771**	**4.55**	**26.19**	**0.023**	**0.036**	**0.108**

Table 3: The effort to cloak different users' characteristics using a logistic regression with 100 SVD components (LRSVD), a logistic regression (LR), and naive Bayes (NB) model. Absolute efforts are presented in the left panel, and relative efforts are in the right panel.

	LRSVD	LR	NB
age $\geq$ 37	0.868	0.904	0.816
agreeableness $\geq$ 5	0.604	0.590	0.587
conscientiousness $\geq$ 5	0.677	0.670	0.626
extraversion $\geq$ 5	0.680	0.671	0.590
iq $\geq$ 130	0.620	0.636	0.619
iq $<$ 90	0.631	0.625	0.571
is democrat	0.889	0.888	0.822
is drinking	0.782	0.790	0.683
is female	0.922	0.967	0.667
is gay	0.890	0.904	0.784
is homosexual	0.788	0.839	0.694
is lesbian	0.729	0.797	0.605
is muslim	0.949	0.949	0.894
is single	0.637	0.665	0.644
is smoking	0.785	0.792	0.673
life satisfaction $\geq$ 6	0.594	0.579	0.570
network density $\geq$ 65	0.609	0.575	0.518
neuroticism $\geq$ 5	0.673	0.603	0.523
num friends $\geq$ 585	0.717	0.734	0.625
openness $\geq$ 5	0.665	0.660	0.635
ss belief = 1	0.689	0.700	0.651
ss belief = 5	0.641	0.616	0.546
uses drugs	0.781	0.772	0.683

Table 4: Area under the ROC curve (AUC) for each classification task using a logistic regression with 100 SVD components (LRSVD), a logistic regression (LR), and a naive Bayes model (NB).