Blend me in: Privacy-Preserving Input Generalization for Personalized Online Services

Alegria Baquero
Institute for Software Research
University of California, Irvine
abaquero@ics.uci.edu

Allan M. Schiffman
CommerceNet
Palo Alto, California
ams@commerce.net

Jeff Shrager
CommerceNet & Stanford
Palo Alto, California
jshrager@stanford.edu

***Abstract*—Users routinely disclose personal information to obtain the benefits of Personalized Online Services. As a result, personal data is distributed across uncounted and unaccountable remote databases. Data mismanagement, as well as privacy and security flaws undermine individuals' control and privacy of their personal data. Yet revealing detailed private data does not necessarily yield useful service personalization; often this functionality is only modestly dependent upon the accuracy of user-supplied input. We demonstrate knowledge-based *input generalization* wherein systematically perturbed user data is supplied to a personalized service to gain *forward privacy* for the user, while retaining the utility of the service's results.**

***Keywords*-Privacy, De-identification, Anonymity, Personalization, HIPAA, Data-mining**

I. Introduction

> If someone asks you a question they've got no business asking, you're under no obligation to tell them the truth.
> — *Leonard Schiffman*

Every day millions of users access Personalized Online Services (POSs) to make travel arrangements, check investments, compare products, arrange medical treatment, and so on. These services require the disclosure of personal information in order to obtain personalized results.

When users access a POS, their data is usually stored by the service provider, and these organizations may make secondary use of this data, or share it with other parties. Users concerned with data privacy have a natural instinct to conceal or misrepresent personal information when using websites that demand it [1]. However, it is a common misconception that removing *direct identifiers*—personally identifiable information, such as names and social security numbers—is enough to safeguard individuals' anonymity and privacy. Users may trust a POS to ethically (and competently) handle their personal information by de-identifying records before making secondary use of personal data. Nonetheless, sensitive information can often be *inferred* through various statistical methods based solely on *quasi-identifiers* (such as date of birth), seemingly "non-sensitive" and "non-identifying" information [2][3][4]. Our work explores the consequences of systematically misrepresenting these quasi-identifiers, even given it is then likely that the service's results may be less accurate.

II. Prior work

Much of the academic literature on preserving data privacy[1] focuses on the *statistical database problem* [7]: given a database provider collects and aggregates data regarding subject individuals, how may the contents of the database be revealed to third parties while preserving the subject's privacy interests? This literature is concerned with both measuring privacy when data is released to third parties and methods to preserve it. Papers often treat both measures and methods together but it is useful to consider them separately.

A. Measuring privacy in databases

Given a particular subject's attribute record (*microdata* in the literature) is already part of a larger database, several privacy measures have been defined:

- *k-anonymity* guarantees that a record will not be distinguishable from $k-1$ records in the database which share the same sensitive attributes. However, this method is weak when all records share sensitive values; although individuals cannot be linked to particular records, they are known to be members of a group sharing some sensitive attribute.
- *l-diversity* extends k-anonymity by also requiring at least l "well-represented" values for the sensitive attribute to introduce inter-group diversity [8].
- *t-closeness* addresses l-diversity's weakness in skewed distributions by additionally requiring that the distance between the distribution of a sensitive attribute in an equivalence class and the distribution of the attribute in the entire dataset is not greater than t [9].
- *Differential privacy* defines a criteria for minimizing the risk of joining a statistical database; the addition or removal of the subject's attributes should not substantially change the outcome of data analysis [10].

A significant limitation to all of these measures is accounting for the likelihood that an attacker could have information

[1]For an overview of privacy in personalized services see [5][6].

other than that contained in the database itself. Bounding this *background knowledge* problem is considered in [11].

B. Methods for sanitizing databases

Database privacy-preserving methods, which operate *after* individuals' private information has been collected, have been proposed to prevent the re-identification of individuals in anonymized datasets, and/or their association with sensitive information. These *sanitization* methods include:

- *Value suppression and generalization* replaces a value with a less specific but semantically consistent one [4].
- *Randomization* adds noise to records in an attempt to constrain an attacker to deriving only aggregate data distributions [12].
- *Data swapping* exchanges sensitive values between records to mask their true values [13]

Database sanitization affects the utility of the database for those parties who use it for future queries [14] [15]. We consider the related (but not equivalent) problem of POS output accuracy after input generalization in section IV-D.

C. Data privacy without databases: ex-ante methods

Of particular relevance to this paper is work to preserving privacy before the subject discloses their attributes. Terms such as *perturbation* and *obfuscation* have been used to describe transforming input data to reduce the exposure of sensitive data through input quality degradation, namely reducing the level of detail of the provided information. Examples of this research can be found in the contexts of geo-location [16][17][18], private web search [19][20][21], recommendation systems [22][23], and social networks [24]. This literature can be contrasted with that related to noise injection and pollution techniques, less relevant here, such as when decoy traffic is created by confusing user queries with false ones [25] or with other users' queries [26].

In our work, we prefer the term "generalization" as used in the database privacy literature because our goal is not merely to confuse a potential attacker, but also to preserve the utility of the personalized service. We say *input* generalization to make clear that generalization is performed *ex-ante*, that is, *before* the sensitive data are supplied to the service.

III. *Ex-ante* Input Generalization

Our work is concerned with *forward privacy*, that is, with minimizing the risk of future re-identification and exploitation of potentially sensitive data based upon information shared today but which may (today) be considered neither sensitive nor identifying. The notion of forward privacy is inspired by *forward secrecy* [27], a desirable cryptographic property which prevents the use of compromised keys to decrypt messages encrypted before the compromise.

Toward this end we introduce *input generalization*, wherein a less precise version of users' private information is provided *ex-ante* to a POS, that is, *before* individuals have disclosed or released (e.g., from medical records) their private information. Input generalization allows an individual to "blend in" to the crowd of current and future POS users to minimize the risk of future re-identification and misuse of personal data. Reduced precision relative to a service's input requirements is accomplished by systematically adding noise to a subset of the data disclosed by the user to the service.

Input generalization differs from the methods discussed in section II-B. Those methods are applied *ex-post*, whereas in our work the transformation is applied *ex-ante* at the point of initial disclosure. Because input generalization operates ex-ante, users need not reveal their true data. Input generalization thus goes beyond the scope of existing databases by accounting for the universe of individuals who share the relevant attributes for a problem domain. For example, in a service for cancer patients, the universe might include all current and future cancer patients. The goal of forward privacy is precisely to *transcend* specific services and databases, and to mitigate background knowledge attacks which some of the methods described earlier fail to address.

Input generalization is a perturbation method in the manner discussed in section II-C, where users define the desired level of privacy [19] through inaccuracy and imprecision [20]. However, compared to other ex-ante approaches, it considers general attribute sets (rather than a single attribute of a particular type), attributes that are explicitly requested by the service provider, and by design trades-off result accuracy for privacy.

A common, simple *ex-ante* method to assure forward privacy is to refuse to provide sensitive data in the first place. Alternatively, one can provide data to parties judged sufficiently trustworthy, and accept the seemingly inevitable failures.[2] Neither approach recognizes the tension between privacy and personalized service output quality and so does not attempt to reconcile them, whereas input generalization is focused precisely on this tension. An analytic treatment of the privacy/utility tradeoff is in [21].

Rather than simply excluding or allowing disclosure of sensitive information, POS users should be empowered to disclose information as they see fit, for purposes they judge useful, and with the discretion to provide only the information necessary to obtain the desired service, possibly adjusting as needed to gain quality at the cost of disclosure.

A. Trading-off Utility for Privacy

Generally, POSs are oblivious and insensitive to nuanced privacy concerns, offering only Hobson's choice: answer all questions or do without service. Some answers are collected in order to personalize the service as offered. However inputs are also collected or precision is demanded for the provider's purposes such as advertising, statistical analysis, implementation convenience, or outright resale.

[2]This is the "you have zero privacy anyway" viewpoint.

Our central hypothesis is that it is often unnecessary to disclose sensitive information to obtain satisfactory output results from a POS, and that input generalization can provide forward privacy in such cases. That is, depending on the context and the nature of the disclosed data, as well as the subjective value attributed to the service, the gain in privacy through undisclosed or partially disclosed information can outweigh potentially reduced service utility (i.e., accuracy). In section IV below we offer a model for input generalization and its analysis. In the section V that follows, we provide several case studies that support this hypothesis.

IV. Details of Input Generalization

A. Model

There are multiple *principals* (persons, agents, companies, etc.) who have some interest in the operation of a given POS. We can distinguish the *user*—i.e., the "subject principal"—from potentially many "secondary principals", including the operators of the POS, and other entities who might access the service or its data for whatever reason. The user supplies his or her personal data *ex-ante* (that is, at input time) in the form of *attributes* to the POS for the purpose of *personalizing* the service. The service maintains a set of *ex-post* (that is, taking place any time after the user's input) functions relevant to the various principals. We refer to the *main* function as the one that services the user's request. This function is presumably personalized with respect to some of the user's inputs. Other *auxiliary ex-post* functions may satisfy the needs of the non-subject principals; for example, demographic information may be provided to advertisers.

In input generalization a less precise version of users' private attributes is provided as input to the POS by adding noise to the actual attribute values as desired.[3] Our method deals with *quasi-identifiers* rather than *direct identifiers* such as name, address, or social security number. Quasi-identifiers are items of personal information, such as residential zip code and date of birth, which by themselves cannot be used to identify an individual, but may, in combination with other information, be used to infer individual identity.[4]

The foundational assumption of the present model is that anyone aside from the user (i.e., the subject principal) should be viewed as a potential attacker, *including* the (non-subject) principals who operate the service itself. That is, it is the very fact that the user has supplied personal data to the service *ex-ante* that leads to threats to privacy.[5] We therefore expressly advocate that the user supply *perturbed* ("generalized") data to the service.[6]

B. Measuring Acceptable Accuracy Degradation

In input generalization a less precise version of users' private attributes is provided by adding noise to the actual attribute values, for example, substituting selected attributes with either a broader containing class (e.g., a neighborhood is substituted by its containing city), a similar item in the same class (e.g., zip code 92617 is substituted with the nearby 92606), or a less precise item (e.g., the date of birth March 10th, 2000 is substituted with a less specific version such as March, 2000).

Of course, input generalization is likely to change the results returned by the service. Therefore, determining the extent to which this perturbation acceptably degrades the service's output accuracy is of central concern. For example, often a modest perturbation in a user's date of birth may not affect the outcome of the service, but privacy may improve significantly. Input generalization needs to be *knowledge-based*, meaning the degree and type of input perturbation is based on domain knowledge and on the sensitivity of a particular online service to its inputs.

As input generalization allows an individual to "blend in" to the population representative of the universe of potential users, generalization needs to account for uniform and non-uniform data distributions. For example, most populations are roughly uniform with respect to birthday, but not so with respect to zip code where there is a greater probability that a person lives in a higher populated zip code than in a scarcely-populated one.

The goal of blending suggests *weighted randomization*, where attribute values have different weights corresponding to the probability of being chosen as a substitute for the real values. For example, weighting would take account of the population distribution among substitute zip codes.

We note that there are limits to the privacy gained by adding noise: with enough data an attacker may achieve his goals despite the added noise [29].

C. Quantifying input perturbation

We use the term *generalization distance* for the degree of perturbation introduced by input generalization. For attributes in metric spaces, generalization distance may be straightforwardly calculated as Euclidean distance. Attribute domains which are not naturally modeled as metric spaces

[3]This approach depends upon users' willingness to provide perturbed values for certain form fields [28].

[4]Users should share their direct identifiers appropriately—relying on law, technology, boycott, or prevarication as needed. We do not consider privacy of direct identifiers to be a solved problem in practice, but for the purposes of this work we assume that other mechanisms are employed to protect them.

[5]These attributes may be explicitly supplied by the subject user (e.g., date-of-birth), or they may be implicitly obtained (e.g., browser version), in which case the generalization would need to be performed in software.

[6]To the extent that input generalization violates the assumptions of input validity made by the non-subject principals, the method is likely to have downstream implications for those non-subject principals' estimates of the validity of the service's database, and therefore may reduce the utility of user-provided information for these secondary purposes. Of course, this is usually of little concern to the user, so long as the main function returns personalized results that he or she finds useful. Our approach prioritizes the users' right to protect their own information as opposed to blindly trusting service providers; it is ultimately the user's choice whether or not to disclose information truthfully.

might still accommodate a distance metric. For example, Joslyn et al. model a gene ontology as a tree or graph where distance between nodes can be measured [30].

Generalization distance is a measure of induced input error (noise) which, from the POS' perspective is a loss of functionality affecting at least output accuracy; this is quantified in section IV-D. From the subject's perspective, this is a potential gain of privacy; the quantification of that gain is the topic of section IV-E.

D. Quantifying output accuracy

As a tradeoff for increased privacy, input generalization may reduce the accuracy of a service's output. Accuracy refers to how close a value is to the correct one. Generally the accuracy of the output will be inversely proportional to the generalization distance of the input. As is the case with input accuracy, output accuracy may be measured in a number of ways.

1) Accuracy in metric spaces: For services providing results that can be interpreted as a vector in a metric space, the accuracy may be given as a normalized distance between the original and generalized output.

2) Accuracy for unordered result sets: Consider an output result to be a set of objects $R = \{r_1, r_2, ..., r_n\}$ with no significance assigned to the order of elements. Let T be the set of elements corresponding to the true output, and G the set corresponding to the generalized output. We refer to false negative results as $F_n = T \setminus G$ (elements in T but not in G) and false positive results as $F_p = G \setminus T$ (elements in G but not in T). We note that:

$$|T \cap G| = |T| - |F_n| = |G| - |F_p|$$

and:

$$|T \cup G| = |T| + |F_p| = |G| + |F_n|$$

Given this, we can quantify accuracy using the Tversky (similarity) index s [31]. Setting Tversky's parameters $\alpha = \beta = 1$ is the most general approach, weighting false positives and false negatives equally. For $T \cup G \neq \emptyset$:

$$s = \frac{|T \cap G|}{|T \cap G| + \alpha |F_p| + \beta |F_n|} = \frac{|T \cap G|}{|T \cup G|} \tag{1}$$

In the best case, the sets are equivalent ($T = G$) and $s = 1$, including the special case where both sets are empty. In the worst case, where $s = 0$, the (non-empty) sets have no elements in common ($T \cap G = \emptyset$, where $T \cup G \neq \emptyset$).

3) Accuracy for ordered result sets: If the order of result set is significant—a model that may be more appropriate for evaluating the results from services such as search engines and recommendation systems—an *order-sensitive non-conjoint intersection measure* is required. We apply a function composition technique accounting for order based on [32]. In this case we call the first n elements of the set R_n, and define s_n to be Tversky's s (as given above) operating on the first n-element prefix of R, and define $k = max(|T|, |G|)$. Our weighted order-sensitive overlap measure σ is then the average overlap of all the prefixes:

$$\sigma = \frac{1}{k} \sum_{n=1}^{k} s_n$$

As with s, σ ranges between 0 (complete disagreement) and 1 (perfect match). This calculation has the property of exponentially weighting early elements over late elements.

E. The blending factor: quantifying gain in privacy

Input generalization creates the perturbed attributes A' by replacing each selected attribute name-value pair $\langle a, v \rangle$ in A with a corresponding $\langle a, v' \rangle$ given $v' \cong v$.

The privacy gain achieved by input generalization, which we call the *blending factor* (β), is the ratio of the portion of the population U matching the attributes after generalization A' versus the original attributes A. β expressed in the context of of k-anonymity would be the ratio k'/k between k before and k' after generalization and suppression.

The treatment of the blending factor that follows is in the spirit of Sánchez et.al.'s [20] *Profile Exposure Level (PEL)* but, firstly we measure improvement in privacy rather than remaining exposure, and secondly we segregate the difficult-to-obtain joint distribution data into a separate error term.

Define X as the elements of A that are not to be generalized (hence unchanged in A'), G as elements from A which are to be generalized, and G' as the attribute set corresponding to G as generalized in A':

$$X = A \cap A' \qquad G = A \setminus X \qquad G' = A' \setminus X$$

We can represent the proportion of the population U that is selected by a set of attributes:

$$\frac{\Pr(\cap a' : a' \in A')}{\Pr(\cap a : a \in A)} = \frac{\Pr(A')}{\Pr(A)} = \frac{\Pr(G' \cap X)}{\Pr(G \cap X)} = \beta$$

Since A' is a more general selector of elements in U than A is, $\beta \geq 1$.

We can then express β as a ratio of conditional probabilities given the ungeneralized attributes X:

$$\beta = \frac{\Pr(G'|X) \Pr(X)}{\Pr(G|X) \Pr(X)} = \frac{\Pr(G'|X)}{\Pr(G|X)} \tag{2}$$

If G (the attributes to be generalized) is statistically independent of X (the non-generalized attributes), then:

$$\beta = \frac{\Pr(G')}{\Pr(G)} \tag{3}$$

Further, if the generalized attributes themselves are statistically independent of one another (such as birth date and gender), then by the multiplication rule:

$$\beta = \frac{\prod_{g' \in G'} \Pr(g')}{\prod_{g \in G} \Pr(g)} \tag{4}$$

Statistical dependence between attributes does not necessarily affect β's value. Returning to the conditional probability formulation in equation 2, we can use Bayes' Rule so that β is conditional upon G:

$$\beta = \frac{\Pr(G'|X)}{\Pr(G|X)} = \frac{\Pr(X|G')\Pr(G')}{\Pr(X)} / \frac{\Pr(X|G)\Pr(G)}{\Pr(X)}$$

which we can rearrange as (contrast with equation 3):

$$\beta = \frac{\Pr(X|G')}{\Pr(X|G)} \cdot \frac{\Pr(G')}{\Pr(G)}$$

Often the first term can be estimated as being close to 1, bypassing the difficulty of determining joint distributions. For example, consider the proportion of the people born between Nov 1 and Nov 10 who like the TV series *House*, versus the proportion of people born on exactly Nov 5 who like *House*. This estimation requires only the judgment of *relative* statistical independence when generalizing an attribute, rather than independence itself or the distribution of the attribute in the population. In other words, what matters is not the degree of attribute dependency, but that they are equally dependent before and after generalization.

The blending factor resulting from a seemingly modest input generalization can be significant: Two orders of magnitude result merely from generalizing a required date-of-birth input to a 100-day range. But very large blending factors may be required to reduce the risk of re-identification in the face of 1) redundant *over-identification* among attributes, 2) background information, and 3) non-perturbed attributes.[7]

F. Estimating confidence in the accuracy of the output

Because input generalization may affect a service's output, it is useful to have a confidence estimate of the (un)certainty of the service's output accuracy with respect to the perturbed inputs. This estimate will be a function of both the *salience* and the *amplitude* of the perturbed attribute. Salience denotes the relevance of a given attribute on the service's functionality. Not all attributes are likely to have the same salience. For example, we can assert with high probability that a user's zip code has little effect on a service that provides movie recommendations, but that it will have a substantial impact on a service that recommends restaurants. Amplitude refers to the magnitude of the perturbation. Inducing a small perturbation (say, of meters) in the user's location will not substantially change the list of recommended restaurants, but a more significant perturbation (say, of miles) may significantly degrade such recommendations [33]. Given these knowledge-based judgments, a model of the specific type of service in a domain should be used to estimate the confidence in the generalization output. These models should be constructed based on an analysis of which attributes inarguably have an impact on the service's outcome (and to what degree), versus those that do not.

Since confidence scores are domain and service specific, these can also be determined by performing sensitivity analyses for the different input attributes. Specifically, perturbations can be performed repeatedly for each attribute (perhaps using different perturbation magnitudes in the case of metric attributes, such as distance) to quantify by observation the impact on the output. Based upon these measurements, probabilities of change in the output can be derived. Assuming attribute independence, a combined probability or overall confidence score can then be provided based on attribute-specific sensitivity analyses.

For those attributes whose difference cannot be measured, but which maintain a hierarchical or containment relationship, the magnitude of the error induced must be estimated based on domain-specific scales.

V. Case studies: Clinical trial matching

Although input generalization can be applied in almost any domain, we have chosen to demonstrate and analyze it through case studies in the healthcare domain. User-disclosed data in healthcare is critical and sensitive. Furthermore online healthcare POSs often request a wide range of personal data.

Clinical trial matching services (*CMTS*s) are particularly interesting examples of POSs. These services match a patient to clinical trials based on the patient's medical information. *Inclusion* and *exclusion criteria* describe the necessary qualifications to participate in a trial. Inclusion criteria are those attributes that qualify a patient to participate in a trial (e.g., patient must have stage III lung cancer). Exclusion criteria are those attributes that *dis*qualify the patient (e.g., the patient must not have previously had a platinum-based chemotherapy drug). This space is large and complex; according to ClinicalTrials.gov [34] there are presently about 130,000 ongoing clinical trials, 35,000 of which are specific to cancer. A typical trial, for example, studies the effectiveness of the combination of the experimental drug MK2206 and the approved drug Paclitaxel on patients with advanced or metastatic breast cancer; the trial excludes patients who have already been treated with Paclitaxel.

We studied two real CTMSs for cancer patients: EmergingMed [35] and BreastCancerTrials [36]. We observed the extent to which perturbing the user's input affects the results. Both services are well known and widely used. The former includes searching trials for different types of cancer, while the later is specific to breast cancer. Both require the patient to enter a large amount of personal data such as zip code, birth date, cancer type and stage, and current and past treatments.

For the case study, we used a sample patient database of fourteen fictitious patients with plausible melanoma, breast, colorectal, and lung cancer profiles created by a clinical

[7] Our intuition is that the notion of over-identification justifies creating variants of record counting -based privacy metrics (such as k-anonymity) to allow for fractional values between 0 and 1.

researcher expert in clinical trials for cancer patients. We consider our sample size sufficient to investigate our input generalization approach due to our knowledge of the salient attributes in CTMSs, acquired through discussion with experts in the field. As an example of relevant nuance in this area, consider that a patient's date of birth by itself is usually not relevant, but it places the patient within an age cohort to determine whether pediatric, adult, or geriatric trials should be provided. Similarly, the patient's zip code is typically relevant only in that it helps determine if the patient could commute to the trial location.

Input generalization methods

CTMSs typically collect five to ten quasi-identifying attributes in the course of supplying service. We chose to generalize three: zip code, birth date, and prescription drugs.

Zip Code: The patients' zip code is replaced by a random one within a fixed radius of the original. We normalized the selected areas from which we drew these random zip codes to obtain approximately the same number (approx. 80).

Date of birth: The patient's date of birth is replaced by a random month and year within a range of ± 5 years.

Prescription drugs: Prescription drugs are hierarchically arranged, where drug brands are grouped within generic drugs, and the latter are grouped by therapeutic drug classes. We replace prescribed drugs—given a brand or a generic drug as input—by alternative ones within the same therapeutic class. For example, if the patient was treated with Eloxatin, it might be replaced by another drug in the "alkylating agents" class, such as Platinol. The rationale for this generalization is knowledge-based—clinical trials are usually more concerned with a patient having been previously treated with a member of a drug class, rather than with a specific drug.

Test 1: Same service, different diagnosis

In this first test, our goal is to assess the accuracy of the output when the input for patients' zip code, date of birth, and (in some cases) prescription drugs is generalized. We use EmergingMed for this analysis, and applied input generalization to example cases of melanoma, colorectal, and lung cancer.

We measured the output accuracy s (formula 1 in section IV-D2), which compares the results for both the real and generalized inputs. We note that the order of results returned from EmergingMed has no significance. We calculated the blending factor (β, formula 4 in IV-E), assuming relative statistical independence between attributes, given modest perturbations. Our results (table I) show that this POS is mostly insensitive to zip code and birth date substitutions chosen as specified in section V. Patient 7 is the exception, where the same clinical trial was provided but in an alternative location 35 miles away from the correct one.

In most cases, generalizing prescription drugs—the replacement of one drug by other in the same therapeutic class—did not have a significant impact on the results. We can observe, however, a degradation of the result for patients 4 and 9 where a larger number of salient attributes—in this case drug names—have been perturbed. Note that not all drugs have been generalized, but the real drug is maintained when there is no viable alternative in the available options.

Test 2: Same service and diagnosis, different generalizations

The goal of this test (on EmergingMed) is to assess whether approximately the same results are obtained when different drug substitutions are made for the same true value. For example, if the prescribed drug is Platinol (generic: Cisplatin, an alkylating agent), it can be replaced by either Neosar or Eloxatin—both alkylating agents as well. In this test we generalized only prescribed drugs.

Our results (summarized in table II) show that the service is mostly insensitive to the selection of a specific drug within a given therapeutic class, the worst case being 90.4% accuracy in the fourth generalization for patient 9.

		Original values				**Generalized values**							
		zip code	dob	prescriptions	T	zip code	dob	prescriptions	G	F_P	F_N	accuracy	β
melano.	Patient 1	94025	Jan 1955	Vemurafenib, Dacarbazine	20	95056	Jun 1951	Vemurafenib, Carmustine	20	0	0	100%	1.7×10^5
	Patient 2	10016	Jan 1965	Ipilimumab, Dacarbazine	3	11222	Aug 1960	Ipilimumab, Temozolomide	3	0	0	100%	1.7×10^5
	Patient 3	60601	Jan 1975		3	60202	Nov 1977		3	0	0	100%	9.6×10^3
colorectal	Patient 4	94025	Jan 1955	Bevacizumab, Erlotinib, Fluorouracil, Oxaliplatin	284	95129	Jul 1957	Bevacizumab, Gefitinib, Fluoracil, Cisplatin	283	2	3	98.3%	8.6×10^5
	Patient 5	10016	Jan 1955	Fluorouracil, Oxaliplatin	3	10024	Jan 1958	Capecitabine, Cisplatin	3	0	0	100%	1.4×10^6
	Patient 6	60601	Jan 1975		3	60621	Feb 1972		3	0	0	100%	9.6×10^3
lung	Patient 7	94025	Jan 1955		21	94544	May 1950		21	1	1	90.9%	9.6×10^3
	Patient 8	10016	Jan 1965	Carboplatin, Oxaliplatin	3	10105	Oct 1967	Lomustine, Cyclophosphamide	3	0	0	100%	3.1×10^6
	Patient 9	60601	Jan 1975	Cetuximab, Carboplatin, Oxaliplatin	250	60706	Sep 1980	Gefitinib, Cisplatin, Cyclophosphamide	248	2	4	97.6%	1.5×10^7

T is the true output cardinality, G the generalized output cardinality, F_P the number of false positives, and F_N the number of false negatives.

Table I
TEST 1

	Original values		Generalization 1				Generalization 2				Generalization 3				Generalization 4			
	prescriptions	T	prescriptions	F_P	F_N	acc.	prescriptions	F_P	F_N	acc.	prescriptions	F_P	F_N	acc.	prescriptions	F_P	F_N	acc.
Patient 1 melanoma	Vemurafenib, Dacarbazine	20	Vemurafenib, Oxaliplatin	0	0	100%	Vemurafenib, Carboplatin	0	0	100%	Vemurafenib, Cisplatin	0	0	100%	Vemurafenib, Temozolomide	0	0	100%
Patient 4 colorectal	Bevacizumab, Erlotinib, Fluorouracil, Oxaliplatin	257	Bevacizumab, Erlotinib, Fluorouracil, Cisplatin	1	2	98.8%	Bevacizumab, Cetuximab, Fluorouracil, Cisplatin	2	4	97.7%	Bevacizumab, Cetuximab, Fluorouracil, Oxiplatin	1	2	98.8%	Bevacizumab, Gefitinib, Fluorouracil, Oxiplatin	0	0	100%
Patient 9 lung	Cetuximab, Carboplatin, Oxaliplatin	219	Gefitinib, Lomustine, Cyclophosphamide	0	11	95%	Erlotinib, Lomustine, Cisplatin	2	3	97.7%	Cetuximab, Cisplatin, Oxaliplatin	0	0	100%	Erlotinib, Caboplatin, Cisplatin	20	3	90.4%
Patient 12 breast	Trastuzumab, Doxorubicin Liposomal	18	Trastuzumab, Epirubicin	0	0	100%	Trastuzumab, Mitoxantrone	0	0	100%	Lapatinib, Epirubicin	1	0	94.7%	Lapatinib, Mitoxantrone	1	0	94.7%

Table II
TEST 2

		Original values				Generalized values								
	pat.	zip code	dob	prescriptions	T	zip code	dob	prescriptions	G	F_P	F_N	acc.	β
EmergingMed	10	94025	Jan 1955		16	94402	Mar 1956		16	0	0	100%	9.6×10^3
	11	10016	Jan 1975		46	11101	Sep 1963		46	0	0	100%	9.6×10^3
	12	60601	Jan 1975	Trastuzumab, Doxorubicin Liposomal	18	60153	Apr 1973	Lapatinib, Mitomycin	19	1	0	94.7%	6.7×10^5
	13	80202	Nov 1945	Bevacizumab, Capecitabine, Carboplatin, Letrozole, Raloxifene, investigational	196	80045	Feb 1940	Bevacizumab, Cisplatin, Methotrexate, Anastrozole, Toremifene, investigational	195	1	2	98.5%	3.9×10^7
	14	32034	Mar 1973	Trastuzumab, Zoledronic acid, Cisplatin, Doxorubicin, Oxaliplatin, Raloxifene, Tamoxifen, investigational	28	31520	Jul 1970	Lapatinib, Zoledronic acid, Carboplatin, Cyclophosphamide, Epirubicin, Goserelin, Toremifene, investigational	27	1	2	89.7%	1.4×10^{10}
BreastCancerTrials	10	940	1955		44	944	1956		44	0	0	100%	2.9×10^2
	11	100	1975		43	111	1973		43	1	1	95.5%	2.5×10^2
	12	606	1975	Trastuzumab, Doxorubicin Liposomal	68	601	1973	Lapatinib, Mitoxantrone	69	3	2	93%	1.2×10^4
	13	802	1945	Bevacizumab, Capecitabine, Carboplatin, Letrozole, Raloxifene	75	800	1940	Bevacizumab, Methotrexate, Cisplatin, Anastrozole, Toremifene	79	8	4	85.5%	9.7×10^5
	14	320	1973	Trastuzumab, Zoledronic acid, Cisplatin, Doxorubicin, Raloxifene, Tamoxifen	54	315	1970	Lapatinib, Pamidronate, Cyclophosphamide, Epirubicin, Raloxifene, Toremifene	52	3	5	86%	1.2×10^7

Table III
TEST 3

		Original values		Input generalization				Random substitution				
		prescriptions	T	prescriptions	G	F_P	F_N	acc.	prescriptions	R	F_P	F_N	acc.
	Patient 1 (melanoma)	Vemurafenib, Dacarbazine	20	Vemurafenib, Carmustine	20	0	0	100%	Vemurafenib, Tamoxifen	20	0	0	100%
EmergingMed	Patient 4 (colorectal)	Bevacizumab, Erlotinib, Fluorouracil, Oxaliplatin	284	Bevacizumab, Gefitinib, Fluoracil, Cisplatin	283	2	3	98.3%	Bevacizumab, Irinotecan, Fluorouracil, Raltitrexed	285	4	3	97.6%
	Patient 9 (lung)	Cetuximab, Carboplatin, Oxaliplatin	250	Gefitinib, Cisplatin, Cyclophosphamide	248	2	4	97.6%	Trastuzumab, Gemcitabine, Vinorelbine	238	5	17	91.4%
	Patient 13 (breast)	Bevacizumab, Capecitabine, Carboplatin, Letrozole, Raloxifene, investigational	196	Bevacizumab, Cisplatin, Methotrexate, Anastrozole, Toremifene, investigational	195	0	1	99.5%	Bevacizumab, Erlotinib, Zoledronic acid, Goserelin, Anastrozole, investigational	193	3	6	95.5%
BCT.org	Patient 14 (breast)	Trastuzumab, Zoledronic acid, Cisplatin, Doxorubicin, Raloxifene, Tamoxifen	54	Lapatinib, Pamidronate, Cyclophosphamide, Epirubicin, Raloxifene, Toremifine	52	3	5	86%	Ixabepilone, Methotrexate, Topotecan, Anastrozole, Raloxifene, Leuprolide	56	14	12	61.8%

R is the cardinality of the set of trials result of random substitutions.

Table IV
TEST 4

	true value	generalized	clinical trial	F_N	F_P	justification
Patient 7 (lung)	94025	94544	Carboplatin and Paclitaxel With or Without Bevacizumab and/or Cetuximab in Treating Patients With Stage IV or Recurrent Non-Small Cell Lung Cancer (NCT00946712)	✗	✗	provided the same trial at a closer location (Castro Valley instead of MountainView)
Patient 13 (breast)	Nov 1945	Feb 1940	A Pharmacokinetic Study of Trabectedin in Patients With Advanced Malignancies and Hepatic Dysfunction (NCT01273493)	✗		eligibility is 18 years to 70 years old, and generalized patient is 72 years old.
Patient 4 (colorectal)	Bevacizumab, Erlotinib, Fluorouracil, Oxaliplatin	Bevacizumab, Gefitinib, Fluoracil, Cisplatin	GDC-0980 in Combination With a Fluoropyrimidine, Oxaliplatin, and Bevacizumab in Patients With Advanced Solid Tumors (NCT01332604)		✗	excludes studies for patients which have received Oxaliplatin-based therapy within 1 year of initiation of study treatment

Table V
EXAMPLES OF INCLUSION OR EXCLUSION CRITERIA

Test 3: Different services, same diagnosis

This test uses two clinical trial matching services (EmergingMed and BreastCancerTrials) to assess whether the accuracy of our results is consistent across services, or if the accuracy obtained through input generalization depends on the specific service. Although the information requested is slightly different, the attributes we are interested in generalizing are consistent across the selected services.

For this test we generalized zip code, date of birth, and prescription drugs for breast cancer patients. Note that BreastCancerTrials requires only the first three digits of the zip code, and only the year of one's birth. (It is more sensitive from the outset towards users' privacy.)

Overall, the generalized output of BreastCancerTrials is less accurate than that of EmergingMed (arithmetic mean of 96.6% vs. 92%). There are several plausible reasons for this difference, including differences in the matching algorithms, and the difference of detail in the input. A more definitive explanation would require knowledge of the details of the systems' implementation and database contents.

As observed in table III, the results of our input generalization method are mostly favorable, with an accuracy arithmetic mean of 94.3%. The worst case is for patient 13 in BreastCancerTrials with an overall accuracy of 85.5%, but in this case six input values have been generalized.

Test 4: Generalization vs. random substitution

In this test we ask whether non-knowledge-based generalization—in this case random substitution—significantly reduces output accuracy, as expected. For example, Decarbazine, a chemotherapy drug, is replaced by Tamoxifen, a hormonal therapy. Our results are summarized in table IV.

Surprisingly, the difference between the generalized output and the one yielded by random substitution in the input offered to EmergingMed is insignificant—the difference between their accuracy arithmetic means is only 2.7%. On the other hand, a significant difference is observed when random substitutions are offered to BreastCancerTrials—a difference of 24.2%. We can conclude that BreastCancerTrials is more sensitive to the input compared to EmergingMed, corroborating the difference in accuracy observed in test 3.

Discussion

Based on our observations, age and zip code have relatively small effect on the results in these CTMSs. Even when a zip code is substituted with one hundreds of miles away, the clinical trials are the same ones in most cases, although an alternative location is sometimes offered. These services are, however, more sensitive to drug generalization.

In tests 1 and 3 a significant blending factor is obtained through input generalization, maintaining 100% accuracy in many cases. For example, zip code, date of birth, and one drug were generalized for melanoma patient 1 from the pool of approximately 80 zip codes, a range of 10 years, and a set of 18 alkylating agents. Assuming that population is distributed uniformly across zip codes and there is an equal probability of taking any drug in the alkylating agents class, the blending factor β, i.e., privacy, increases by a factor of 120 (months) $\times$ 80 (zip codes) $\times$ 18 (alkylating agents) = 172,800 (or $\sim 10^5$). This is a notable result given that no accuracy was sacrificed. Of course more attributes could be perturbed, and by a greater degree, if there is higher tolerance for output inaccuracy.

Even in the worst case, an accuracy of 85.5% was obtained for patient 13 (using BreastCancerTrial.org), but with a blending factor of (120 (months) $\times$ 80 (zip codes) $\times$ 8 (antimetabolites) $\times$ 18 (alkylating agents) $\times$ 4 (aromatase inhibitors) $\times$ 7 (hormones/antineoplastics)) / 40 (zip codes starting with 802) = 967,680 (or $\sim 10^6$). A more modest blending factor would still yield a significant gain in privacy with (potentially) increased result accuracy.

As expected, accuracy degrades with an increasing number of generalized (salient) attributes. There is an inverse relation between salience, perturbation amplitude, and number of generalized attributes and the services accuracy. Each perturbed attribute increases the probability of matching more trials' inclusion or exclusion criteria. Trials are excluded or included in the output because only patients within a specific age range, or who live within a given distance, or who have (or have not) taken a specific drug are qualified. Table V provides examples for each of these cases drawn from our tests, along with the reason why they have been included or excluded from the generalization output.

We are giving the same weight to false positives as to false negatives in calculating accuracy (by setting Tversky's α and β to 1 in the s computation of section IV-D). However, one might consider false negatives as more important in this case, given that it may not matter to obtain a few additional but irrelevant trials, whereas it may be critical not to miss any possible trials. This is the tradeoff between privacy and output accuracy that users need to make. Input generalization tools could help users make these sort of decisions; our metrics provide a means to parameterize such choices.

The foregoing case studies strengthen the claim that, at least in this domain, users often unnecessarily sacrifice their privacy by disclosing detailed personal, sensitive, information for the benefits of expected personalization, but that, at least in this domain, POSs are not sensitive to all the requested personal information. Specifically, these services are mostly insensitive to small changes in users' age and location, and only moderately sensitive to changes in drug values. One service used in our study exhibited insensitivity even to presumably relevant attributes such as cancer stage.

VI. Conclusions

> Lt. Kaffee: I want the truth!
> Col. Jessup: You can't handle the truth!
> — *A Few Good Men (1992)*

Input generalization promotes *forward privacy* by replacing private data with less precise values, thus minimizing the risk of future re-identification or unauthorized usage of this data. Input generalization applies privacy measures *ex-ante* in contrast to well known mechanisms concerned about privacy *ex-post* disclosure (e.g., differential privacy), and therefore outside the data owners' control. Generalization is implemented through the knowledge-based substitution of attribute values by a broader containing class, an equivalent in the same class, or a less accurate or precise version.

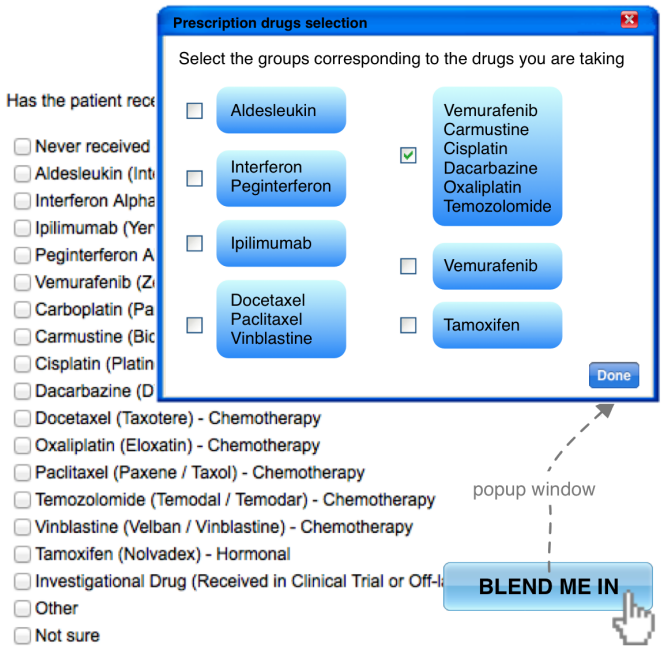

Figure 1. UI widget for drug generalization

We have demonstrated how one can quantify the improvement in privacy and reduction in accuracy provided by input generalization. The resulting degradation due to such generalization is contingent on the salience of the generalized attribute(s) with respect to the service, the amplitude of their perturbation, and the number of perturbed attributes. In our case studies these factors affected the accuracy in matching clinical trials' inclusion or exclusion criteria. However, the most notable result is that in many instances 100% accuracy was maintained despite *blending factors* (i.e., privacy gains) of five or six orders of magnitude.

Input generalization enables users to make the tradeoff between personal data disclosure and service accuracy by controlling the extent of disclosure. The method is applicable to any personalized online service where the tradeoff between privacy and accuracy depends upon the salience of user-provided attributes.

In conjunction with the case study described here we created a prototype *Blend me in* browser plug-in with interaction widgets for generalizing zip code, date-of-birth, and prescription drugs (the drug widget is shown in figure 1). For future work we envision a library of input generalization interaction widgets that would help users maximize forward privacy by permitting them to wisely choose a desired point in the trade-off between privacy and service utility.

Acknowledgments

We thank researchers at CollabRx Inc., especially Dr. Smruti Vidwans, for providing us with the hypothetical patient data used in our tests. We also thank Dr. Rachna Dhamija, Ian Fisher, Ted Goldstein, Eric Rescorla, and Dr. Richard Taylor for their valuable feedback. This work is financially supported by CommerceNet.

References

[1] M. J. Metzger, "Communication privacy management in electronic commerce," *Journal of Computer-Mediated Communication*, vol. 12, no. 2, pp. 335–361, 2007.

[2] C. Farkas, "The inference problem: a survey," *ACM SIGKDD Explorations Newsletter*, 2002.

[3] A. Narayanan and V. Shmatikov, "Myths and fallacies of 'personally identifiable information'," *Communications of the ACM*, vol. 53, no. 6, p. 24, 2010.

[4] L. Sweeney, "k-anonymity: A model for protecting privacy," *IEEE Security And Privacy*, vol. 10, no. 5, pp. 557–570, 2002.

[5] L. F. Cranor, "I didnt buy it for myself," in *Designing personalized user experiences in eCommerce*. Springer, 2004, pp. 57–73.

[6] A. Kobsa, "Privacy-enhanced personalization," *Communications of the ACM*, vol. 50, no. 8, pp. 24–33, 2007.

[7] S. Chawla, C. Dwork, F. McSherry, A. Smith, and H. Wee, "Toward privacy in public databases," in *Theory of Cryptography*. Springer, 2005, pp. 363–385.

[8] A. Machanavajjhala, D. Kifer, and J. Gehrke, "l-diversity: Privacy beyond k-anonymity," *ACM Transactions on Knowledge Discovery from Data*, pp. 1–36, 2007.

[9] N. Li and T. Li, "t-closeness: Privacy beyond k-anonymity and l-diversity," *Data Engineering*, no. 2, 2007.

[10] C. Dwork, "Differential privacy: A survey of results," *Theory and Applications of Models of Computation*, pp. 1–19, 2008.

[11] D. J. Martin, D. Kifer, A. Machanavajjhala, J. Gehrke, and J. Y. Halpern, "Worst-case background knowledge for privacy-preserving data publishing," in *Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference on*. IEEE, 2007, pp. 126–135.

[12] C. Aggarwal, "A General Survey of Privacy-Preserving Data Mining Models and Algorithms," *Privacy-preserving data mining*, 2008.

[13] S. Fienberg and J. McIntyre, "Data swapping: Variations on a theme by Dalenius and Reiss," *Privacy in statistical databases*, 2004.

[14] J. Brickell and V. Shmatikov, "The cost of privacy: destruction of data-mining utility in anonymized data publishing," in *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2008, pp. 70–78.

[15] T. Li and N. Li, "On the tradeoff between privacy and utility in data publishing," in *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2009, pp. 517–526.

[16] S. Ye, F. Wu, R. Pandey, and H. Chen, "Noise injection for search privacy protection," in *Computational Science and Engineering, 2009. CSE'09. International Conference on*, vol. 3. IEEE, 2009, pp. 1–8.

[17] M. Duckham and L. Kulik, "A formal model of obfuscation and negotiation for location privacy," in *Pervasive Computing*. Springer, 2005, pp. 152–170.

[18] J. Krumm, "A survey of computational location privacy," *Personal and Ubiquitous Computing*, vol. 13, no. 6, pp. 391–399, 2009.

[19] Y. Xu, K. Wang, B. Zhang, and Z. Chen, "Privacy-enhancing personalized web search," in *Proceedings of the 16th international conference on World Wide Web*. ACM, 2007, pp. 591–600.

[20] D. Sánchez, J. Castellà-Roca, and A. Viejo, "Knowledge-based scheme to create privacy-preserving but semantically-related queries for web search engines," *Information Sciences*, 2012.

[21] A. Krause and E. Horvitz, "A utility-theoretic approach to privacy in online services," *Journal of Artificial Intelligence Research*, vol. 39, pp. 633–662, 2010.

[22] J. Parra-Arnau, D. Rebollo-Monedero, and J. Forné, "A privacy-protecting architecture for collaborative filtering via forgery and suppression of ratings," in *Data Privacy Management and Autonomous Spontaneous Security*. Springer, 2012, pp. 42–57.

[23] H. Polat and W. Du, "Privacy-preserving collaborative filtering using randomized perturbation techniques," in *Data Mining, 2003. ICDM 2003. Third IEEE International Conference on*. IEEE, 2003, pp. 625–628.

[24] M. Hay, G. Miklau, D. Jensen, P. Weis, and S. Srivastava, "Anonymizing social networks," *Computer Science Department Faculty Publication Series*, p. 180, 2007.

[25] D. C. Howe and H. Nissenbaum, "Trackmenot: Resisting surveillance in web search," *Lessons from the Identity Trail: Anonymity, Privacy, and Identity in a Networked Society*, pp. 417–436, 2009.

[26] J. Castellà-Roca, A. Viejo, and J. Herrera-Joancomartí, "Preserving users privacy in web search engines," *Computer Communications*, vol. 32, no. 13, pp. 1541–1551, 2009.

[27] C. G. Gunther, "An identity-based key-exchange protocol," in *Advances in Cryptology EuroCrypt*, vol. 89, 1989, pp. 29–37.

[28] R. Agrawal and R. Srikant, "Privacy-preserving data mining," *ACM Sigmod Record*, vol. 29, no. 2, pp. 439–450, 2000.

[29] J. Domingo-Ferrer, F. Sebé, and J. Castella-Roca, "On the security of noise addition for privacy in statistical databases," in *Privacy in statistical databases*. Springer, 2004, pp. 149–161.

[30] C. A. Joslyn, S. M. Mniszewski, A. Fulmer, and G. Heaton, "The gene ontology categorizer," in *Intelligent Systems in Molecular Biology*. Oxford Univ Press, 2004, pp. 169–177.

[31] A. Tversky, "Features of similarity," *Psychological review*, vol. 84, no. 4, pp. 327–352, 1977.

[32] R. Fagin, R. Kumar, and D. Sivakumar, "Comparing top k lists," *SIAM Journal on Discrete Mathematics*, vol. 1, 2003.

[33] B. Gedik and L. Liu, "Location privacy in mobile systems: A personalized anonymization model," in *Distributed Computing Systems, 2005. ICDCS 2005. Proceedings. 25th IEEE International Conference on*. IEEE, 2005, pp. 620–629.

[34] "Clinicaltrials.gov," http://www.clinicaltrials.gov, accessed: 2012-06-18.

[35] "Emergingmed.com," http://emergingmed.com, accessed: 2012-06-18.

[36] "Breastcancertrials.org," http://www.breastcancertrials.org, accessed: 2012-06-18.