A Group-Based Personalized Model for Image Privacy Classification and Labeling

Haoti Zhong
Dept. of Electrical Eng.
Pennsylvania State University
hzz133@psu.edu

Anna Squicciarini
Information Sciences and Technology
Pennsylvania State University
acs20@psu.edu

David Miller
Dept. of Electrical Eng.
Pennsylvania State University
djmiller@engr.psu.edu

Cornelia Caragea
Department of Computer Science
University of North Texas
ccaragea@unt.edu

Abstract

We address machine prediction of an individual's label (private or public) for a given image. This problem is difficult due to user subjectivity and inadequate labeled examples to train individual, personalized models. It is also time and space consuming to train a classifier for each user. We propose a Group-Based Personalized Model for image privacy classification in online social media sites, which learns a set of archetypical privacy models (groups), and associates a given user with one of these groups. Our system can be used to provide accurate "early warnings" with respect to a user's privacy awareness level.

1 Introduction

There has been increasing interest in understanding users' privacy attitudes, especially in social network contexts [Lampinen *et al.*, 2011; Wang *et al.*, 2011b; Sheehan, 2002; Mazzia *et al.*, 2012]. Decisions related to image privacy have been under scrutiny, due to the potential negative effect of sharing an image with unintended audiences [Wang *et al.*, 2011a]. To date, however, research has mainly focused either on "universal" privacy prediction models, at one extreme, or on personalized models, for generic content types (e.g. text, media content, etc.) at the other extreme[Mazzia *et al.*, 2012]. Both bodies of work fail to address some concerns. First, every image is unique and carries different degrees and types of personal information with it. Further, generic privacy patterns do not necessarily reflect an individual user's unique sharing comfort levels. For instance, consider Figure 1. Two users may agree on what is "public" for non-personal content (e.g. landscapes or scenes from public places), but may differ greatly on images with kids, partial nudity, partying, and other life events. Ultimately, whether an online image should be private or not is subjective, affected by one's personality, experience, and degree of privacy awareness [Xu *et al.*, 2012; Wang *et al.*, 2011b]. Yet, several studies also suggest it is possible to predict how users will treat certain content, given their level of privacy awareness and personal taste [Sheehan, 2002; Spyromitros-Xioufis *et al.*, 2016].

A technical challenge that personalized models need to overcome is the amount of data necessary to train accurate models. Even *if* a large amount of training data is available, it is both time and space consuming (e.g., in the cloud) to train and store models for every user. But sufficient personalized training resources may anyway not be available. Thus we argue new approaches are needed, leveraging common "archetypical" privacy patterns across subsets of users, with statistical strength in learning thus borrowed across users.

We propose a stochastic Group-Based Personalized Model (GBPM) for image privacy classification in online social media sites. We introduce the concept of privacy groups, which model a subset of users, and treat (exclusive) group membership as a latent variable for each user. Our model is "personalized" in that each (test) user probabilistically associates with each of the groups, based on any labeled examples supplied by the user and the user's demographics info. The user's posterior probability that any given image is private is thus a user-specific (personalized) average of the privacy posteriors under each of the groups.

Our model performs comparatively well regardless of the amount of data used for training, consistently outperforming several baselines, including the obvious approach of training a separate personalized model for each user. Our experiments show that, on a dataset of 114 users and about 3,400 image labelings, our model achieves an overall accuracy measure of 79.31% when a few (15) images are used to infer group associations for each (test) user. We also achieve a promising overall average accuracy of 62.2% even when test users provide *no* labeled data, but merely some personal profile data (thus addressing the Cold Start Problem[Schein *et al.*, 2002]).

Furthermore, because our model groups users according to their privacy labeling patterns, we can analyse whether groups of users who display similar behavior with respect to image privacy are also similar in demographics and personal characteristics. Our results indicate that group dynamics are complex, and while our groups cluster users well with respect to privacy, no group appears to be well-described by a single set of demographics. Rather, certain profiles, although different, (e.g. an elder Asian woman and a Hindu young man) could exhibit similar privacy patterns. To the best of our knowledge, this is the first image privacy recommendation model able to guide users' privacy preferences even in the case of *no* or very limited labeled data for new users.We also note that the proposed model could easily be extended to address more general recommendation systems, with minor changes. This is because recommendation systems are typically also highly subjective. Users patterns can be exploited by our model (e.g. a users background and purchase history could help us group users for similar recommendations).

2 Related Works

A number of recent studies have analyzed sharing patterns and social discovery in image sharing sites like Flickr [Lampinen *et al.*, 2011; Yu *et al.*, ; Wang *et al.*, 2011a; Bonneau and Preibusch, 2010]. Among other interesting findings, scholars have determined that images are often used for self and social disclosure.

Several mechanisms to protect user-uploaded images have been proposed recently[Ra *et al.*, 2013; Backes *et al.*, 2014; Tierney *et al.*, 2013; Klemperer *et al.*, 2012; Zerr *et al.*, 2012; Tran *et al.*, 2016]. These mostly rely on the premise that users make privacy decisions typically consistent with socially accepted norms of what is sensitive and what is not [Sheehan, 2002]. Accordingly, the authors focus on finding "universal" features and a universal model for image privacy detection, based on the Scale-invariant feature transform (SIFT), color histogram deep learinig (DL) features, and text features from tags and captions. However, these methods treat all users identically. Consistently, they find that a single SVM trained on an ensemble of features performs fairly well. However, we demonstrate experimentally that these approaches lose accuracy by not accounting for subjectivity in privacy. Also note that one primary baseline we compare with (universal SVM) follows the method[Zerr *et al.*, 2012]. This "universal" SVM has much lower accuracy compared with our proposed model.

Recently, [Spyromitros-Xioufis *et al.*, 2016] explored personalized privacy models using DL features. They separately trained a logistic regression model for every user using labeled data provided by the user plus some additional labeled data from a set of other users. However this approach required many labeled examples from each user to be most effective. Our approach is a compromise between these strategies, more flexible than the single model approach, less personalized than a "pure personalized" approach, but borrowing statistical strength from users in the same (discovered) group to reduce labeled data requirements. Moreover, besides providing privacy predictions, our approach discovers multiple groups given a set of personalized labelings/ratings.

3 Method

We assume P_u training users, with user p providing N_p images and a corresponding binary vector of labels $\mathbf{b_p}$, indicating private or public for each image. For a private image, only spatially localized regions may contain sensitive content, with other (e.g. background) regions not really influencing a privacy decision. For modeling purposes, we process each image into L overlapping patches or regions. Image patches are chosen randomly, with the top-left corner's position chosen based on a uniform distribution, and then with the bottom-right corner chosen uniformly conditioned on the remaining size of the image. Each patch's width and length are restricted to be at least 20 pixels, to make sure each patch contains enough visual information.

3.1 Group-Based Personalized Model (GBPM) for Users' Privacy Decisions

Our model assumes M *types* of private content (M *a priori* unknown), with the characteristics of these types also *a priori* unknown, e.g., they could represent bare skin, children's faces, religious, or political content. Private content may be included in one or more regions of a given image. Accordingly, an image is assumed private if and only if it contains at least one region with private content.

Consistent with recent work on users' privacy awareness and online behavioral studies [Wang *et al.*, 2011b], we assume that a user's privacy decisions follow one of a finite set of patterns, i.e. we define *privacy groups*. These patterns (e.g. younger users are comfortable sharing images with bare skin) may result in different privacy decisions for the same image by different groups, according to a group's preferences. These groups need to be discovered through our learning mechanism. Supposing for now that there are K groups, we introduce a 0-1 latent variable V_{pk}, which indicates whether user p belongs to group k, where $\sum_{k=1}^{K} V_{pk} = 1$. V_{pk} (for both training and test users) will be inferred from a user's available labelings and from their personal profile data.

As mentioned above, images are processed into L patches. We parametrize the probability that patch l for image i (provided by user p) is private to user p, assuming he belongs to group k, as:

$$P[C_{pilk} = 1 | \mathbf{x}_{pil}] = \frac{\sum_{m=1}^{M} e^{w_k^{(m)T} \mathbf{x}_{pil} + w_{km0}}}{1 + \sum_{m=1}^{M} e^{w_k^{(m)T} \mathbf{x}_{pil} + w_{km0}}}, \quad (1)$$

i.e. as a generalized logistic regression function, where C_{pilk} is the class label for the l-th patch of image i provided by user p, assuming he is in group k, $\mathbf{x}_{pil}$ is the D dimensional feature vector for this patch(deep learning extracted, as described later), $w_k^{(m)}$ is a vector of weights for the m-th type of private content in group k and w_{km0} is a bias. We choose the *maximum* private probability among all patches as the probability that the image is private(w.r.t user p, assuming he is from group k), *i.e.*:

$$P[C_{pik} = 1 | \mathbf{X}_{pi}] = \max\{P[C_{pilk} = 1 | \mathbf{x}_{pil}] : l = 1, \cdots, L\}, \quad (2)$$

where C_{pik} is the class label for image i and $\mathbf{X}_{pi}$ is the collection of patch feature vectors for image i.

Under an independent and identically distributed training data assumption, we can express the incomplete data log likelihood function over all the training images of all users, where

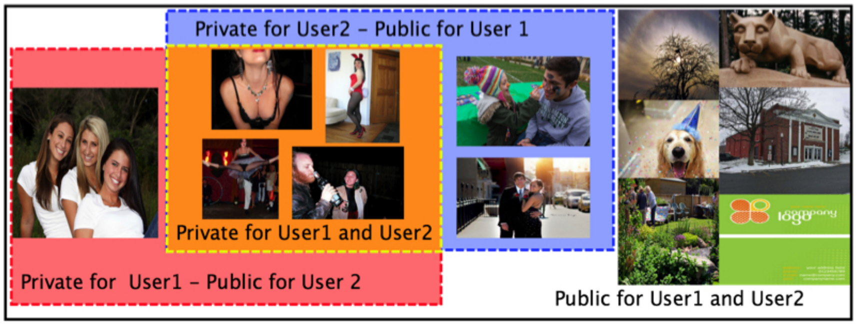
Private for User2 – Public for User 1
Private for User1 and User2
Private for User1 – Public for User 2
Public for User1 and User2

Figure 1: User 1 labeled images in the blue dash box as private while User 2 labeled images private in the red dash box. Orange-boxed images are images labeled private by both, and images outside both boxes are labeled public by both users.

b_{pi} is the binary class label for image i(with 1 indicating the private class , and 0 otherwise), by

$$\sum_{p=1}^{P_u} log(\sum_{k=1}^{K} \frac{1}{K} \prod_{i=1}^{N_p} P[C_{pik} = 1|\mathbf{X}_{pi}]^{b_{pi}} (1 - P[C_{pik} = 1|\mathbf{X}_{pi}])^{1-b_{pi}}) \tag{3}$$

Because the incomplete data likelihood is difficult to optimize directly, we instead invoke Expectation-Maximization (EM). We first define the *complete* data log likelihood:

$$\sum_{p=1}^{P_u} \sum_{i=1}^{N_p} \sum_{k=1}^{K} V_{pk}\{b_{pi} \log P[C_{pik} = 1|\mathbf{X}_{pi}] + (1 - b_{pi}) \log(1 - P[C_{pik} = 1|\mathbf{X}_{pi}])\}, \tag{4}$$

where V_{pk} is the *hidden data* within the EM framework. The E-step and M-step are applied alternately until the log-likelihood function (3) is locally maximized. Specifically, in the E-step, we calculate the expectation of the latent variables given the observed data:

$$E[V_{pk}|\mathbf{X_p}, \mathbf{b_p}] = P[k|\mathbf{X_p}, \mathbf{b_p}] \propto exp(\sum_{i=1}^{N_p} \{b_{pi} \log P[C_{pik} = 1|\mathbf{X}_{pi}] + (1 - b_{pi}) \log(1 - P[C_{pik} = 1|\mathbf{X}_{pi}])\}) \tag{5}$$

In the M-step, we maximize the expected value of the complete data log likelihood over the weight vector model parameters, with the expected latent variables held fixed. Since a closed form solution does not exist, we used gradient ascent. Backtracking line search[Nocedal and Wright, 2006] was used for choosing the step size automatically. To deal with the max function (2), we used the pseudo-gradient [Teow and Loe, 1997], a mathematically sound method based on Fourier convergence analysis of side-derivatives to derive an approximate gradient for max-min error functions.Thus, we can obtain the approximate gradient of the log-likelihood $\nabla_{w_k^{m'}} \log \mathcal{L}$, with respect to each weight vector w_k^m, as:

$$\sum_{p=1}^{P_u} \sum_{i=1}^{N_p} \sum_{k=1}^{K} \frac{P[k|\mathbf{X_p}, \mathbf{b_p}]}{(N(\Theta(X_{pi})))} \sum_{l' \in \Theta(X_{pi})} [(\frac{(-b_{pi})}{P[C_{pi} = 1|\mathbf{X}_{pi}]} + \frac{(1 - b_{pi})}{(1 - P[C_{pi} = 1|\mathbf{X}_{pi}])}) (\frac{[\mathbf{x}_{pil'}, 1] e^{-w_k^{(m')^T} \mathbf{x}_{pil'} + w_{km'0}}}{(1 + \sum_{m=1}^{M} e^{-w_k^{(m)^T} \mathbf{x}_{pil'} + w_{km0}})^2})], \tag{6}$$

where $|N(\Theta(X_{pi}))| = |\{l' \in 1, \cdots, L|P[C_{pil'} = 1|x_{pil'}]\|| = \max|\{P[C_{pil} = 1|\mathbf{x}_{pil}] : l = 1, \cdots, L\}|$.

3.2 Modeling the User's Profile

By itself, the algorithm in Section 3.1 can learn groups with strong intra-group privacy agreement. However, this gives no way to infer privacy decisions for *new* users who have labeled no images. This motivates us to include user profile information in our GBPM model, which may be correlated with interpersonal characteristics[Xu *et al.*, 2012], e.g religion belief, age, and gender. We extend our model to use profile data $\mathbf{y}_p$ of every user p. $\mathbf{y}_p$ is a 30-dimensional binary vector[1], with each entry representing the value of a particular profile attribute. Based on this, we can define prior probabilities on group membership:

$$prior_{pk} = \frac{exp(\beta_k^t \mathbf{y}_p + \beta_{k0})}{\sum_{k'=1}^{K} exp(\beta_{k'}^t \mathbf{y}_p + \beta_{k0'})} \tag{7}$$

Thus our complete data log likelihood function now becomes:

$$\sum_{p=1}^{P_u} \sum_{k=1}^{K} V_{pk}(\log(prior_{pk}) + \sum_{i=1}^{N_p} \{b_{pi} \log(P[C_{pik} = 1|\mathbf{X}_{pi}]) + (1 - b_{pi}) \log(1 - P[C_{pik} = 1|\mathbf{X}_{pi}])\}). \tag{8}$$

[1] There are 7 demographic variables. The total cardinality of these 7 variables is 30, e.g. for ethnicity there are 5 possible values. Thus, we create a 30-dimensional binary vector, with each binary entry taking on value 1 when the associated demographic variable takes on a particular value, and zero otherwise.

EM is still applied to learn the model, where:

$$\begin{aligned} & P[k|\mathbf{X_p}, \mathbf{b_p}] \propto \\ & exp(\log(prior_{pk}) + \sum_{i=1}^{N_p} \{b_{pi} \log P[C_{pi} = 1|\mathbf{X}_{pi}] \\ & + (1 - b_{pi}) \log(1 - P[C_{pi} = 1|\mathbf{X}_{pi}])\}). \end{aligned} \quad (9)$$

Even if new (test) users provide *no* labeled examples, our model can still use the learned prior probabilities to infer their group memberships, based on their profile data.[2]

3.3 Prediction

Given that a new user p has already labeled some images, we can compute $P[k|\mathbf{X_p}, \mathbf{b_p}]$ according to the E-step equ (9). Then for a new image i, we can compute the *a posteriori* privacy probability for user p as:

$$P[C_{pi}|\mathbf{X_p}, \mathbf{b_p}] = \sum_{k=1}^{K} P[k|\mathbf{X_p}, \mathbf{b_p}] P[C_{pik} = 1|\mathbf{X}_{pi}]. \quad (10)$$

Currently, we threshold this probability at 0.5 to make a decision. However, the threshold could be chosen to achieve any desired true positive/false alarm tradeoff.

4 Experimental Validation

4.1 Dataset

We collected our own dataset for testing purposes as follows. The imageset was taken from the Picalert study, a collection of images with varying degrees of sensitivity [Zerr *et al.*, 2012]. We randomly sampled 2,700 images, and split them evenly into 90 subsets such that each subset has 30 images. In order to learn and test users' differences in terms of privacy choices, each subset was assigned to two unique Mechanical Turk workers, who were recruited using the Turk platform. Each worker's hit included two tasks: 1. Complete a survey demographics and Social Network usage practice, and 2. Label as private or public 30 images provided by us. The first subtask specifically requested: gender, age range, education level, ethnicity, religious belief, social network access frequency (expressed on a 5-item frequency scale) and frequency of posting on a social network.

In total, 114 valid user responses were collected and 3420 labels in total (2496 public labels and 924 private labels). In 28.24% of the cases, images were labeled inconsistently by the users who evaluated them. In terms of demographics, 59.29% of the respondents were male, the average age was in the range of 25-35, and the average respondent has a college degree (46.9%). The majority of the population is white (69%), followed by Asian, African American, Hispanic or Latino and other ethnicities. 53 participants are Christian (43.6%), 2 are Buddhist, 10 are Hindu, 33 do not affiliate with any religion, and the remaining respondents believe other religions. Finally, on average, 77.67% of the respondents access and use social network sites frequently, about once a week. On average, users post content once a week - similar to common usage patterns in the social media population [Correa *et al.*, 2010].

4.2 Deep Learning for Feature Extraction

Our image features are deep learning(DL) features, constructed as follows. We use the pre-trained eight-layer implementation of Convolutional Neural Networks implemented by Caffee [Jia *et al.*, 2014], which is a benchmark standard for image classification and object detection tasks [Razavian *et al.*, 2014]. The first five layers of this network extract features by convolution with a set of image filters. Because we are not interested in object identification, we disregard the 8th layer and treat the 4096-dimensional output of the 7th layer of the network as a feature vector, describing high-level features of each image or image patch but not objects or categories. Note that prior to feeding an image or an image patch into the Deep Learning Network, it is re-sized to 224*224. In this way, a patch can be represented as a 4096 dimensioned DL feature vector regardless of its size.

4.3 Hyper-parameter Selection and Initialization

In GBPM, three hyper-parameters need to be chosen – the number of private content types M, the number of patches L and the number of user groups K; we use nested cross validation (CV) to choose M and K. We first divide the dataset into 10 (outer) folds, and use 9 of these folds for training-plus-validation, with the last fold used for testing. To calculate the optimized hyper-parameter, we further split the collection of nine training-plus-validation fold samples, again using 5-fold cross validation, with four of these (inner) folds used for training and one for validation. The search grid for K is chosen from 4 to 7 with search step of 1, and M is chosen over a range from 20-50 with a search step of 5, to maximize the average (inner) validation fold CV accuracy. We found that M=40 and $K = 6$ fit best for this dataset. Larger M and K may be found for larger datasets (with more users). L was chosen to be the minimum number such that the patches cover 90% of the image support. Thus, $L = 100$.

Since Equ (1) is non-convex, initialization plays an important role since otherwise EM may converge to a poor local maximum. In our study, we initialized our model based on seven "seed" users with distinguished backgrounds and asked them to label 60 common images (which were also included in the training users' set for baseline methods we evaluated). We then trained the GBPM for each of them by setting the group number $K = 1$. Finally, we combined these initial pre-trained models and randomly generated the profile data weights to instantiate our initial model parameters.

4.4 Baselines

We validate the accuracy of our model with 10-fold cross validation by randomly splitting all 114 users into 10 user folds, using 9 user folds for training and 1 for testing. For the testing user fold we further split the labeled data provided by each test user into two halves. We use the first half to infer the test user's group memberships for our model (this is achieved by applying only the E-step using these test user labeled examples). The second half labeled data is used for

[2] It is possible to add these examples to the labeled training set and retrain the model for the new user. This approach entails much greater system complexity, and was not investigated in this paper.

Classifier	**Overall Accuracy**	TNR	TPR	F1-measure	SD of Overall Accuracy
Universal SVM	53.51%	72.13%	34.88%	0.4702	0.1797
PSVM	59.53%	80.22%	38.83%	0.5233	0.1717
IPSVM	56.84%	79.51%	34.16%	0.4778	0.1803
IPPSVM	59.42%	74.29%	44.55%	0.5570	0.1918
GBPM without Profile	65.03%	81.97%	48.09%	0.6062	0.1922
GBPM	79.31%	85.38%	73.23%	0.7883	0.1934

Table 1: Classification results using Deep Learning features.

calculating accuracy, comparing all baselines' and GBPM's prediction results with the test user labels (which are treated as ground-truth for accuracy evaluation purposes). We compare our model's accuracy with baselines when each test user provides 30 labeled images, half used to evaluate testing accuracy. For every baseline, we also measure the Standard Deviation (SD) of the accuracy across users, so as to check whether the performance is stable over all users.

We first compared our model mentioned in Section 3.1, where no profile data is used, with our final GBPM, with profile information used. We trained our models using 9 folds of training users' data, and infer the group type latent variable using only the first half labeled data from a test user. As shown in Table 1, we obtained an average TPR of 48.09% and TNR of 81.97% when profile data is not used. The GBPM model (with profile info) greatly improves the overall accuracy. We also tested several baselines as a comparison to our model; these baselines all use as input features the 4096 DL features obtained by feeding the whole image into the DL network and extracting the layer 7 outputs. One approach is to use all 9 fold training users' data to train a Universal SVM for all users, assuming all users have similar taste in privacy. We used an RBF kernel SVM as the classifier; nested cross validation was used to get the optimal hyper-parameters of the SVMs. At the other extreme, we train a separate SVM for every user. We call this second baseline Personalized SVM (PSVM). Since in our dataset each user labeled 30 images, we used half of these data to train and the rest to test without using other users' data. Due to the small amount of training data, linear SVM was used here to avoid overfitting[3].

To balance between these two extremes, we also trained a global RBF-SVM with 9-folds training users' data; then for every new user, 15 labeled data (first half) was added to the training dataset and (personalized) retraining was applied. We call this baseline Incremental Personalized SVM (IPSVM). Further, we also designed a 4th baseline, Incrementally Personalized Profile-SVM (IPPSVM), which is the same as IPSVM except that it uses the user's profile data as additional features. This approach uses exactly the same features and data resources as GBPM.

When provided 15 test user labeled images to infer group memberships, GBPM substantially outperforms all these baselines, as shown in Table 1. Our results confirm on the one hand that the assumption of one model for all users - per the Universal-SVM model - is too simple. A pure personalized model (PSVM), on the other hand, while better than the universal SVM, still leads to poor performance, due to lack of sufficient training data. Finally, even with IPPSVM, which offers a compromise between these two strategies, we did not observe any significant performance improvement. We believe this is due to the fact that IPPSVM "indiscriminately" adds (global) labeled examples to train a personalized SVM – many of these are from users with different privacy tolerances than the current user, and thus are not helpful in achieving more accurate personalization. Also note that as reported in Table 1, our model and all baselines' SDs are similar, which shows our approach is as "stable" as the baselines.

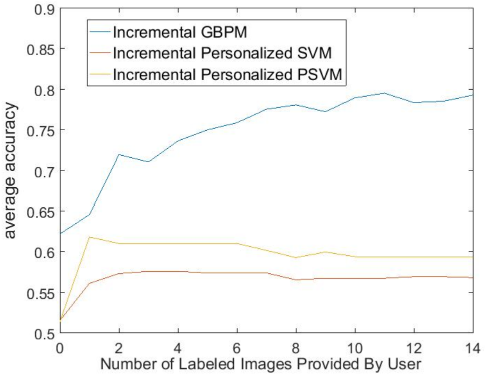

Figure 2: Accuracy of our model, IPSVM and IPPSVM versus the number of test user provided labeled images.

Impact of Different Amounts of Test User Labels We varied the number of test user-provided labeled examples from none to 15 images to assess the effect of adding labeled test user data. Note that when no images are labeled by the user (0 user-provided labels), this becomes an instance of the Cold-Start problem. In Figure 2, we compare our model with the IPSVM and IPPSVM baselines since both methods, similar to ours, exploit the test user-provided data (IPSVM and IPPSVM use them for Incremental retraining purpose, whereas our model (only) uses them to infer a better group association for a new user using the E-step (without updating the model)). As shown in Figure 2, GBPM achieves a higher average accuracy than these two baselines, both in the Cold-Start case and as labeled images are added. Also as expected, with more labeled test user data, the average accuracy of our model increases. In general, the improvement is greater at the beginning and slows down as more images are added. However, GBPM is seen to benefit much more from test user examples than the baselines.

Complexity The training step for our model has a rela-

[3] Due to insufficient number of training images, CV was not used here; thus we directly used the (margin slackness) hyper-parameter learned by the universal SVM.

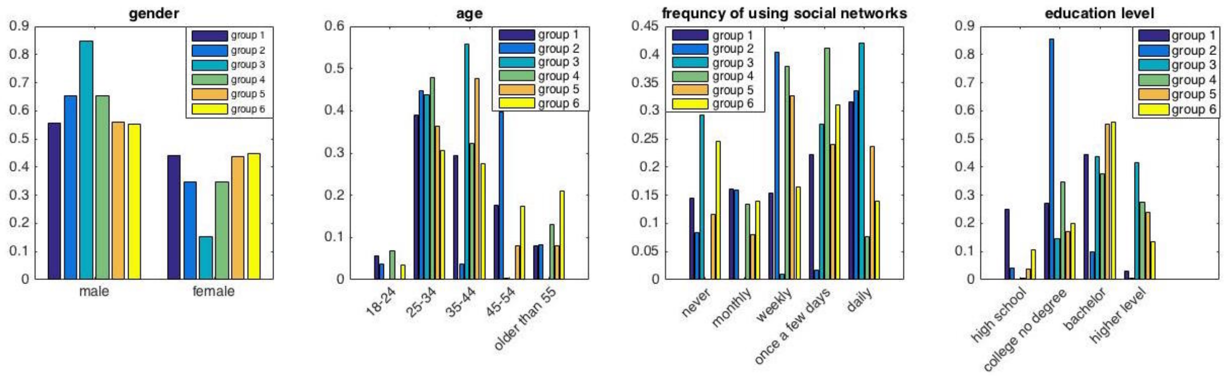

Figure 3: User Type distribution examples for four attributes: gender, age, frequency of social network usage, education level. As shown, all attribute values are distributed across the groups.

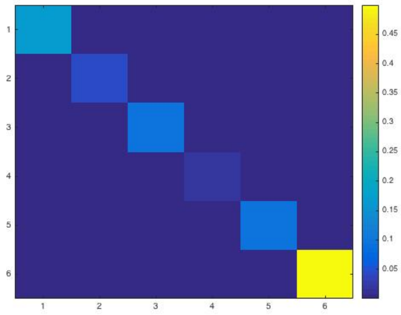

Figure 4: User Group Similarity Matrix

tively high complexity, due to the high dimensional feature space and gradient-ascent based M-step. Precisely, it takes O($K * M * P * N_p * L * D$) operations to compute the E-step and O($K * M^2 * P * N_p * L * D$) to calculate each approximate gradient of the log-likelihood for the M-step, where D is the largest order (4096). However, note that GBPM's training occurs only *once* for the user population, compared with training a model for every user as for all other personal baselines. Also, our model needs O($K * N_p * L * M * D$) time complexity to infer the group membership of a test user, with N_p being the number of labeled images provided by a new user p, and 0(1) space complexity to save this information. By contrast, even PSVM, which required the least training time, needs at least O($N_p^2 * D$) complexity to train and O(N_p) space complexity to save the model for (every) new user. That is, while our training complexity primarily grows with the number of groups (which depends on the number of *training* users), complexity for a personalized SVM grows with the number of *test* users, which in general is much greater.

4.5 User Clustering Analysis

Since GBPM groups users according to their privacy behavior, it is interesting to see whether users with similar privacy patterns share common profile attributes. Using inferred indicator variables V_{pk} we can calculate the group distribution for each of the profile attributes. Recall that, for our 114 users, we have determined 6 groups (from 2 to 54 members per group, average 19 and SD 18.60), using 7 distinct profile attributes and their labeled image data. In Figure 3 we report the distributions of four representative profile attributes: gender, age, education level and frequency of social networks usage. Viewed individually (based on their group conditioned histograms), individual attributes are not very group-discriminating. Moreover, there does not appear to be even a single *vector* of values (over all attributes) that is group-defining, or group-characteristic.

Another approach to explore how distinct the user groups are is to check the overlap of combinations of user profile attributes among groups. We calculate a similarity matrix (see Figure 4), where M_{ij} represents the percentage of similar profile data in group j compared with the data in group i. We define a pair of user profiles similar if two profiles have two or less different attribute values (28% difference). As shown, the distinction between different groups is significant, since the values on non-diagonal elements are quite small. We also observe that the diagonal elements, which show the similarity of profile data inside of a group, are not very high, indicating that users in a group do not share identical attributes. As mentioned before, people with similar level of privacy awareness are not always identical with respect to demographics and interpersonal characteristics. As privacy is highly dependent on complex dimensions beyond social and demographic background, such as exposure to privacy outcries and personal experiences [Xu *et al.*, 2012], two highly similarly profiled users may exhibit very different privacy patterns.

5 Future Work

First, we can extend our model to address multi-group decision making in other contexts, e.g. Supreme Court Decisions or recommendation systems. Second, it would be interesting to extend the privacy problem to the multiple levels of privacy case [Squicciarini *et al.*, 2014]. Third, our learning objective function is non-convex; alternative training could solve this problem more efficiently. Finally, model retraining to include each new user's labeled examples could be investigated.

Acknowledgments

Portions of Dr. Squicciarini and Caragea's work were supported by National Science Foundation grant 1421776 and 1421970, respectively.

References

[Backes *et al.*, 2014] Michael Backes, Sebastian Gerling, Stefan Lorenz, and Stephan Lukas. X-pire 2.0: A user-controlled expiration date and copy protection mechanism. In *Proceedings of the 29th Annual ACM Symposium on Applied Computing*, pages 1633–1640. ACM, 2014.

[Bonneau and Preibusch, 2010] Joseph Bonneau and Sören Preibusch. The privacy jungle: On the market for data protection in social networks. In *Economics of information security and privacy*, pages 121–167. Springer, 2010.

[Correa *et al.*, 2010] Teresa Correa, Amber Willard Hinsley, and Homero Gil De Zuniga. Who interacts on the web?: The intersection of users personality and social media use. *Computers in Human Behavior*, 26(2):247–253, 2010.

[Jia *et al.*, 2014] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. In *Proc. of the ACM International Conference on Multimedia*, pages 675–678. ACM, 2014.

[Klemperer *et al.*, 2012] Peter Klemperer, Yuan Liang, Michelle Mazurek, Manya Sleeper, Blase Ur, Lujo Bauer, Lorrie Faith Cranor, Nitin Gupta, and Michael Reiter. Tag, you can see it!: Using tags for access control in photo sharing. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 377–386. ACM, 2012.

[Lampinen *et al.*, 2011] Airi Lampinen, Vilma Lehtinen, Asko Lehmuskallio, and Sakari Tamminen. We're in it together: interpersonal management of disclosure in social network services. In *Proc. of the SIGCHI conference on human factors in computing systems*, pages 3217–3226. ACM, 2011.

[Mazzia *et al.*, 2012] Alessandra Mazzia, Kristen LeFevre, and Eytan Adar. The pviz comprehension tool for social network privacy settings. In *Proceedings of the Eighth Symposium on Usable Privacy and Security*, SOUPS '12, pages 13:1–13:12, New York, NY, USA, 2012. ACM.

[Nocedal and Wright, 2006] Jorge Nocedal and Stephen Wright. *Numerical optimization*. Springer Science & Business Media, 2006.

[Ra *et al.*, 2013] Moo-Ryong Ra, Ramesh Govindan, and Antonio Ortega. P3: Toward privacy-preserving photo sharing. In *Presented as part of the 10th USENIX Symposium on Networked Systems Design and Implementation (NSDI 13)*, pages 515–528, 2013.

[Razavian *et al.*, 2014] Ali S Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. Cnn features off-the-shelf: an astounding baseline for recognition. In *Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 512–519. IEEE, 2014.

[Schein *et al.*, 2002] Andrew I Schein, Alexandrin Popescul, Lyle H Ungar, and David M Pennock. Methods and metrics for cold-start recommendations. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 253–260. ACM, 2002.

[Sheehan, 2002] Kim Bartel Sheehan. Toward a typology of internet users and online privacy concerns. *The Information Society*, 18(1):21–32, 2002.

[Spyromitros-Xioufis *et al.*, 2016] Eleftherios Spyromitros-Xioufis, Georgios Petkos, Symeon Papadopoulos, Rob Heyman, and Yiannis Kompatsiaris. Perceived versus actual predictability of personal information in social networks. In *International Conference on Internet Science*, pages 133–147. Springer, 2016.

[Squicciarini *et al.*, 2014] Anna C Squicciarini, Cornelia Caragea, and Rahul Balakavi. Analyzing images' privacy for the modern web. In *Proceedings of the 25th ACM conference on Hypertext and social media*, pages 136–147. ACM, 2014.

[Teow and Loe, 1997] Loo-Nin Teow and Kia-Fock Loe. An effective learning method for max-min neural networks. In *IJCAI*, pages 1134–1139, 1997.

[Tierney *et al.*, 2013] Matt Tierney, Ian Spiro, Christoph Bregler, and Lakshminarayanan Subramanian. Cryptagram: Photo privacy for online social media. In *Proceedings of the first ACM conference on Online social networks*, pages 75–88. ACM, 2013.

[Tran *et al.*, 2016] Lam Tran, Deguang Kong, Hongxia Jin, and Ji Liu. Privacy-cnh: A framework to detect photo privacy with convolutional neural network using hierarchical features. *AAAI 2016*, 2016.

[Wang *et al.*, 2011a] Yang Wang, Gregory Norcie, Saranga Komanduri, Alessandro Acquisti, Pedro Giovanni Leon, and Lorrie Faith Cranor. I regretted the minute i pressed share: A qualitative study of regrets on facebook. In *Proc. of the Seventh Symposium on Usable Privacy and Security*, page 10. ACM, 2011.

[Wang *et al.*, 2011b] Yang Wang, Gregory Norice, and Lorrie Faith Cranor. Who is concerned about what? a study of american, chinese and indian users privacy concerns on social network sites. In *International Conference on Trust and Trustworthy Computing*, pages 146–153. Springer, 2011.

[Xu *et al.*, 2012] Heng Xu, Hock-Hai Teo, Bernard CY Tan, and Ritu Agarwal. Research noteeffects of individual self-protection, industry self-regulation, and government regulation on privacy concerns: A study of location-based services. *Information Systems Research*, 23(4):1342–1363, 2012.

[Yu *et al.*,] J. Yu, D. Joshi, and J. Luo. Connecting people in photo-sharing sites by photo content and user annotations. In *Multimedia and Expo, 2009. ICME 2009. IEEE International Conference on*, pages 1464–1467.

[Zerr *et al.*, 2012] Sergej Zerr, Stefan Siersdorfer, Jonathon Hare, and Elena Demidova. Privacy-aware image classification and search. In *Proc. of the 35th international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '12, pages 35–44, 2012.