Differential Privacy for Image Publication

Liyue Fan
liyue.fan@uncc.edu

ABSTRACT

Rapidly generated image data can be shared to advance research and benefit various communities. However, sensitive individual information captured in the data, such as license plates and identities, may inflict privacy concerns when sharing image data with untrusted parties. Existing image privacy solutions rely on standard obfuscation, e.g., pixelization and blurring, which may lead to re-identification. In this study, we explore how differential privacy can enhance standard image obfuscation. Specifically, we extend the standard differential privacy notion to image data, which protects individuals, objects, and/or their features. We develop differentially private methods for two popular image obfuscation techniques, i.e., pixelization and Gaussian blur. Empirical evaluation with real-world datasets demonstrates the utility and privacy of our methods. Furthermore, our methods are shown to effectively reduce the success rate of CNN based re-identification attacks.

The abstract extends our previously published conference paper [4], with a newly developed `DP-Blur` method and additional empirical results. Discussions are also provided for future work on differential privacy for image publication.

ACM Reference Format:
Liyue Fan. 2019. Differential Privacy for Image Publication. In *Proceedings of TPDP '19: Theory and Practice of Differential Privacy (TPDP '19).* ACM, New York, NY, USA, 4 pages. https://doi.org/XXX

1 INTRODUCTION

Image data is largely shared nowadays for research and services, e.g., surveillance and energy saving, or for social purposes, e.g., online social networks. The sanitization of sensitive content in image data, e.g., faces and texts, has largely relied on standard image obfuscation techniques, such as pixelization (also referred to as Mosaicing) and blurring. However, recent development and applications of machine learning have rendered such obfuscation techniques ineffective for privacy preservation [6, 9]. Specifically, Hill et al. [6] have demonstrated that Hidden Markov Models (HMM) can learn to decode redacted documents. Moreover, McPherson et al. [9] have shown that Convolutional Neural Networks (CNN) are highly adaptable to standard obfuscation, which can re-identify up to 96% of pixelized faces. Therefore, we are in need of image obfuscation methods that can provide rigorous privacy guarantees.

The *goal* of this study is to ensure a rigorous notion, i.e., *differential privacy* [2], for image data sharing. Our work differs from the differentially private synthetic data generation approach, such as [14], by protecting sensitive content in each individual image. As a result, our work is applicable to a wider range of settings, e.g., when data owners have small amount of data or correlated image data. The specific contributions[1] are as follows:

(1) To extend the standard differential privacy notion to image data, we propose the m-neighborhood notion, which allows for the protection of any sensitive information represented by up to m pixels.

(2) We develop two differentially private methods, i.e., `DP-Pix` and `DP-Blur`, which enhance the privacy of widely used image obfuscation techniques, i.e., pixelization and Gaussian blur, respectively.

(3) We empirically evaluate the differentially private methods with real-world datasets. Two utility metrics are adopted to measure the absolute error and the perceptual quality. We show that our private methods yields comparable output to the non-private obfuscation.

(4) We simulate the re-identification attacks via deep learning and the results show that the differentially private methods significantly reduces the re-identification risk, even with low privacy requirements, i.e., $\epsilon \geq 0.1$ and $m = 16$.

2 RELATED WORK

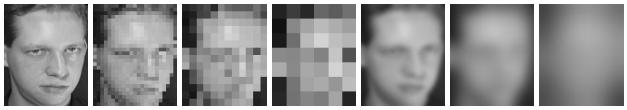
Image Obfuscation. Two popular image obfuscation techniques are *pixelization* and *blurring*. Pixelization [6] can be achieved by superposing a rectangular grid over the original image and averaging the color values of the pixels within each grid cell. On the other hand, blurring, i.e., Gaussian blur, removes details from an image by convolving the 2D Gaussian distribution function with the image. YouTube provides its own face blur implementation [13] for video uploads. McPherson et al. [9] studied pixelization and YouTube face blur and concluded the obfuscated images using those methods can be re-identified. Given sufficient training data, Convolutional Neural Networks (CNN) can effectively learn the association between the obfuscation and the underlying identity.

Differential Privacy. Differential privacy [2] has become the state-of-the-art privacy paradigm for sanitizing statistical databases. While it provides rigorous privacy guarantees for each individual data record in a database, it is challenging to apply the standard notion to non-aggregated data. Several variants of the privacy notion have been proposed. For instance, event-level privacy [3] aims to protect the presence of individual events in one person's data when releasing aggregated data. Local privacy [1] enables answering aggregate queries without a trusted data curator. Recently, differential privacy has been applied to generative adversarial networks (GAN) for synthetic data generation, such as in [14]. Such an approach protects the presence or absence of any training image, but it is hard to apply when each data owner wishes to publish a small number of images, or highly correlated images, e.g., a video sequence. Beside our work [4], there have not been any studies on differential

[1] A prior version of the study can be found in [4]

TPDP '19, November 11, 2019, London, UK

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/XXX



(a) orig (b) b=4 (c) b=8 (d) b=16 (e) k=15 (f) k=45 (g) k=99

Figure 1: A sample AT&T [12] image, its pixelization (b-d) with different b values, and Gaussian blur (e-g) with different k values.

privacy in individual-level image publication or its mitigation effect of CNN based re-identification attacks.

3 METHOD

3.1 Preliminaries

Problem Setting. A data owner wishes to share one or more images with a wide range of untrusted recipients, e.g., researchers or the greater public. For privacy protection, the data owner must sanitize sensitive content in the image data prior to its publication. Here we focus on *grayscale* images: an image I is regarded as an M x N matrix with integer values between 0 and 255 (0 is black and 255 is white).

Image Pixelization and Gaussian Blur. The pixelization technique renders a source image using larger blocks. It is achieved by partitioning the image using a two-dimensional grid, and the average pixel value is released for each grid cell. We adopt a "square" grid where the pixel width is equal to the pixel height in the grid cells, i.e., each cell contains b x b pixels. In general, a smaller b value yields better visual quality, as is shown in Figure 1b, 1c, 1d.

The Gaussian blur smooths a source image by removing detail and noise. It is achieved by convolving the image with a two-dimensional Gaussian kernel. For each pixel, a k x k neighborhood around the pixel is considered and the gaussian weighted average is released. In general, a larger k value yields lower visual quality by over smoothing the image, as is shown in Figure 1e, 1f, 1g.

3.2 Image Differential Privacy

We adapt the standard Differential Privacy [2] definition, which operates in statistical databases, to image data.

Definition 1. *[ϵ-Image Differential Privacy] A randomized mechanism $\mathcal{A}$ gives ϵ-differential privacy if for any neighboring images I_1 and I_2, and for any possible output $\widetilde{I} \in Range(\mathcal{A})$,*

$$Pr[\mathcal{A}(I_1) = \widetilde{I}] \leq e^{\epsilon} \times Pr[\mathcal{A}(I_2) = \widetilde{I}] \quad (1)$$

where the probability is taken over the randomness of $\mathcal{A}$.

The parameter ϵ specifies the degree of privacy offered by $\mathcal{A}$, i.e., a smaller ϵ implies stronger privacy and vice versa. The concept of "neighboring images" is the key to the differential privacy notion, which should clearly define the private content under the protection of differential privacy. In this paper, we propose the following notion of image neighborhood.

Definition 2. *[m-Neighborhood] Two images I_1 and I_2 are neighboring images if they have the same dimension and they differ by at most m pixels.*

Allowing up to m pixels to differ enables us to protect the *presence or absence* of any object, text, or person, represented by those pixels

(a) PETS [7] (b) Example neighbor (c) AT&T [12]

Figure 2: Example neighboring images and representative sensitive information

in an image. For instance, each red rectangle in Figure 2a illustrates sensitive information which can be represented by $\sim$ 360 pixels, such as a pedestrian, a van, an object on grass, and a signage. One example neighboring image is shown in Figure 2b, differing only at the left-most pedestrian. By differential privacy, an adversary cannot distinguish between any pair of neighboring images by observing the output image. The privacy of the pedestrian, and any other sensitive information represented by at most m pixels, can thus be protected. The m-Neighborhood notion can also be applied to protect *features* of an object or person. For instance, the rectangle in Figure 2c contains $\sim$ 120 pixels and encloses the area of the eyes which is reportedly the optimal feature for face recognition tasks [11]. When adopting the above definition, the data owner can choose an appropriate m value in order to customize the level of privacy protection, i.e., achieving indistinguishability in a smaller or larger range of neighboring images.

3.3 Differentially Private Methods

The notion of image differential privacy can be applied to enhance the privacy of standard image obfuscation, such as pixelization and blurring. We propose two methods below.

`DP-Pix`. In a nutshell, the algorithm first performs pixelization on an input image, i.e., by computing the average pixel value for each grid cell, and applies perturbation to the pixelized image. Intuitively, the global sensitivity of pixelization for each b x b grid cell is $\frac{255m}{b^2}$. The corresponding Laplace mechanism [2] is adopted to draw random noises, in order to achieve ϵ-differential privacy.

`DP-Blur`. The initial design of the algorithm first applies Laplace perturbation to each pixel and then performs Gaussian blur, in order to produce a "smooth" output image. The global sensitivity of direct pixel-wise operation is $255m$, which induces larger noises compared to the pixelization based approach. To reduce the effect of the noise, we first run `DP-Pix` on the input image, using a small cell width b_0. We then upsample the private pixelization to the original size and perform Gaussian blur.

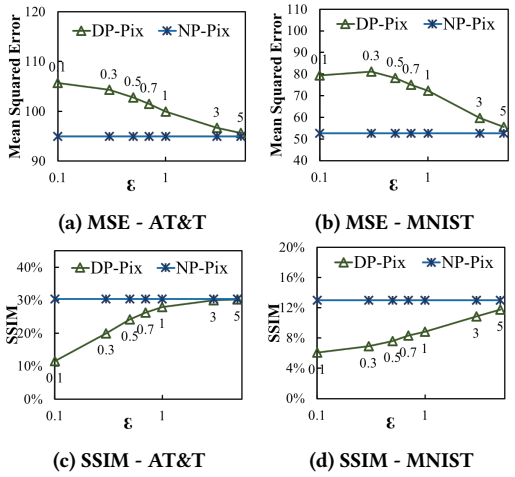
Theorem 1. `DP-Pix` and `DP-Blur` satisfy ϵ-differential privacy.

The proof of the theorem is straight-forward. The proof for `DP-Pix` can be found in [4] and the result of `DP-Blur` follows.

4 RESULTS

Datasets: In this abstract, we present results with two widely used datasets in the computer vision[2], including the re-identification attacks via deep learning [9]: *AT&T* [12] database of faces which

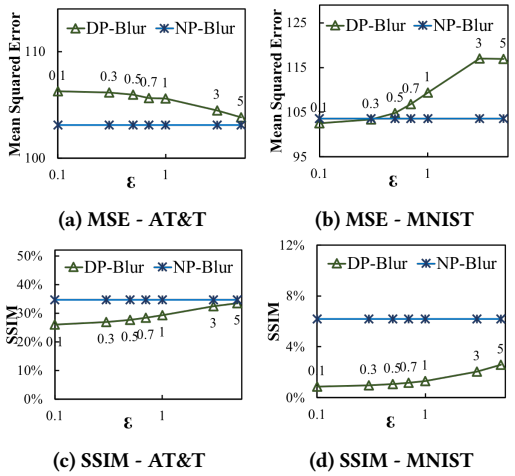
[2]more datasets are adopted in [4]



(a) MSE - AT&T (b) MSE - MNIST

(c) SSIM - AT&T (d) SSIM - MNIST

Figure 3: Utility of *pixelization* versus privacy ϵ



(a) MSE - AT&T (b) MSE - MNIST

(c) SSIM - AT&T (d) SSIM - MNIST

Figure 4: Utility of *blurring* versus privacy ϵ

contains 400 grayscale images of 40 individuals with 92 x 112 resolution; and *MNIST* [8] which contains 60,000 grayscale images of handwritten digits with 28 x 28 resolution.

Setup: We prototyped our method in Python with OpenCV. The parameters take the following values, unless specified otherwise: $b_0 = 4$, $b = 16$, $m = 16$, $k = 99$, $\epsilon = 0.5$. The *utility* of the obfuscated image is measured by: (1) standard Mean Square Error (MSE) and (2) a widely used perceptual quality measure, i.e., Structural Similarity (SSIM) [15], the range of which is from 0 to 1 with 1 being the highest quality. Both utility metrics are defined between the input image and the obfuscated image. In each experiment, the average result is reported in each dataset. As a reference for utility, we also included the standard pixelization and Gaussian blur, i.e., `NP-Pix` and `NP-Blur`, which are parameterized with the same b and k values. A complete set of experiments can be found in [4].

4.1 Impact of ϵ

We present the tradeoff between utility and privacy by varying the parameter ϵ in Figure 3 and 4. Lower ϵ value ensures stronger privacy, and yields lower utility. When increasing ϵ, the differentially private methods show a lower MSE and a higher SSIM, approaching the utility of the non-private baselines. This is due to a lower perturbation error required by differential privacy. An exception is observed for *MNIST* dataset, where the MSE of DP methods increases initially, e.g., $0.1 \leq \epsilon \leq 0.3$ in Figure 3b and $0.1 \leq \epsilon \leq 3$ in Figure 4b. The reason is that *MNIST* depicts white (255) digits on a black (0) background. Such extreme pixel values are not significantly affected by very large Laplace noises (when ϵ is small), as the perturbed pixel values are truncated to the range of [0, 255]. The SSIM measure is shown to be more robust than MSE, exhibiting a consistently increasing trend as ϵ increases in Figure 3d and 4d.

4.2 CNN Re-identification Attacks

While differential privacy provides a rigorous indistinguishability guarantee, we conducted the following study to understand whether the differentially private methods can mitigate known, practical privacy risks. We simulated the CNN based attacks proposed in [9] such that the attack is *adapted* to each obfuscation method. Specifically, assume the adversary has access to the training set obfuscated by a given method, and the label of each training image, i.e., individual identity (1-40) and digits (0-9). The goal of the adversary is re-identify the testing set obfuscated by the same given method, i.e., predicting the label for each obfuscated testing image. We reported the attack results in Table 1.

The DP methods are applied with various ϵ values and compared with the non-private methods and a random baseline, which randomly picks a label. As can be seen, the non-private baselines inflict high re-identification rates and the differentially private methods significantly reduce the attack success. Especially, the `DP-Blur` method produces obfuscated images that are hard to re-identify. For the *AT&T* dataset, `DP-Blur` achieves the lowest the re-identification rate, making such attacks harder than random guessing. As for the *MNIST* dataset, the re-identification rate is less sensitive to the privacy parameter ϵ since the images are dominated by black and white pixels and at a lower resolution. For images obfuscated by `DP-Blur` with $\epsilon = 0.1$, the attack CNN model simply classifies every instance as "1", which is the majority class in *MNIST*.

4.3 Qualitative Utility

Sample *AT&T* images obfuscated under the default parameter setting are provided in Table 2. As can be seen, the differentially private methods inflict minor visual quality loss, compared to non-private obfuscation techniques. Recall that `NP-Pix` and `NP-Blur` lead to high re-identification risks in Table 1 and differentially private methods significantly reduces the risk. We conclude that image differential privacy can enhance the privacy of standard obfuscation, without compromising the quality of the obfuscated image.

5 CONCLUSION AND DISCUSSIONS

We have presented two differentially private methods for image obfuscation, which was the first attempt at extending differential privacy to individual-level image publication. We proposed the m-neighborhood notion to define the indistinguishability requirement, i.e., roughly the same output for any images differing at up to m pixels. We empirically evaluated the utility of differentially private

Dataset	Random	NP-Pix	DP-Pix ($b = 16$)			NP-Blur	DP-Blur ($k = 99$)		
	-	$b = 16$	$\epsilon = 0.1$	0.5	1	$k = 99$	$\epsilon = 0.1$	0.5	1
AT&T	2.50	96.25	3.75	43.75	77.50	88.75	**1.25**	7.50	17.50
MNIST	**10.00**	52.13	16.41	21.51	22.95	76.35	11.35	11.75	13.43

Table 1: Accuracy (in %) of CNN Re-Identification Attacks

Table 2: Qualitative utility of differentially private obfuscation: each column represents one obfuscation method.

methods with multiple real-world image datasets, and showed that our methods yield similar utility to that of the non-private obfuscation techniques. In addition, CNN based re-identification attacks were simulated and the results showed that differentially private methods significantly reduce the attack success even at low privacy requirements, i.e., $\epsilon \geq 0.5$ and $m = 16$. Therefore, we concluded that our methods are simple yet powerful.

As a new endeavor, a number of directions can be explored for future work on differential privacy for image publication:

1. Although our results show that differential privacy can enhance the privacy of standard image obfuscation, it would be interesting to study the potential benefit of differential privacy in *utility*. Intuitively, relaxations of ϵ-differential privacy can be applied, such as (ϵ, δ)-DP and Rényi DP [10]. Furthermore, we may consider the application of differentially private methods in image super-resolution [16], e.g., the low-resolution image can be protected under `DP-Pix`. Moreover, we should explore whether differential privacy can be achieved independent of standard obfuscation, such as pixelization and blurring, in order to overcome the limitations on utility imposed by those techniques. Our recent work [5] initiates the study of this possibility.

2. An extension of this work is to obfuscate sensitive content in videos, where each video can be considered a sequence of images, i.e., frames. For instance, a face will be captured in each frame in the sequence, which requires obfuscation consistently. Due to the composibility, applying differentially private obfuscation to each frame quickly reduces the privacy guarantee. A first-cut solution is to apply the DP method to the first frame only and replace the face region in each subsequent frame using the same DP obfuscation. More utility-friendly solutions can be developed, given specific video analysis tasks.

3. We will also discuss how the work on differential privacy for image data could benefit other communities and applications. For instance, if differentially private image data can be published in real time with sufficient accuracy, surveillance and event detection can be performed to monitor home safety and patient mobility disorders. Another area of interest is the interplay of differential privacy and security in multimedia information systems: the methods proposed in this work introduce irreversible perturbation noise, which makes it hard to recover the original visual data when needed, e.g., for forensic applications. Last but not least, we should explore whether differential privacy tools can help with the behavioral research community which relies on visual recordings of participants, as participants may have personalized privacy needs that current visual privacy solutions fail to address.

REFERENCES

[1] J. C. Duchi, M. I. Jordan, and M. J. Wainwright. 2013. Local Privacy and Statistical Minimax Rates. In *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*. 429–438. https://doi.org/10.1109/FOCS.2013.53

[2] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. 2006. *Calibrating Noise to Sensitivity in Private Data Analysis.* Springer Berlin Heidelberg, Berlin, Heidelberg, 265–284. https://doi.org/10.1007/11681878_14

[3] Cynthia Dwork, Moni Naor, Toniann Pitassi, and Guy N Rothblum. 2010. Differential privacy under continual observation. In *Proceedings of the forty-second ACM symposium on Theory of computing*. ACM, 715–724.

[4] Liyue Fan. 2018. Image Pixelization with Differential Privacy. In *Data and Applications Security and Privacy XXXII*, Florian Kerschbaum and Stefano Paraboschi (Eds.). Springer International Publishing, Cham, 148–162.

[5] Liyue Fan. 2019. Practical Image Obfuscation with Provable Privacy. In *2019 IEEE International Conference on Multimedia and Expo, ICME 2019*. To Appear.

[6] Steven Hill, Zhimin Zhou, Lawrence Saul, and Hovav Shacham. 2016. On the (in) effectiveness of mosaicing and blurring as tools for document redaction. *Proceedings on Privacy Enhancing Technologies* 2016, 4 (2016), 403–417.

[7] Laura Leal-Taixé, Anton Milan, Ian Reid, Stefan Roth, and Konrad Schindler. 2015. Motchallenge 2015: Towards a benchmark for multi-target tracking. *arXiv preprint arXiv:1504.01942* (2015).

[8] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. 1998. Gradient-based learning applied to document recognition. *Proc. IEEE* 86, 11 (1998), 2278–2324.

[9] Richard McPherson, Reza Shokri, and Vitaly Shmatikov. 2016. Defeating Image Obfuscation with Deep Learning. http://arxiv.org/abs/1609.00408. *CoRR* abs/1609.00408 (2016).

[10] Ilya Mironov. 2017. Rényi Differential Privacy. In *30th IEEE Computer Security Foundations Symposium, CSF 2017, Santa Barbara, CA, USA, August 21-25, 2017*. 263–275. https://doi.org/10.1109/CSF.2017.11

[11] Matthew F Peterson and Miguel P Eckstein. 2012. Looking just below the eyes is optimal across face recognition tasks. *Proceedings of the National Academy of Sciences* 109, 48 (2012), E3314–E3323.

[12] F. S. Samaria and A. C. Harter. 1994. Parameterisation of a stochastic model for human face identification. In *Proceedings of 1994 IEEE Workshop on Applications of Computer Vision*. 138–142. https://doi.org/10.1109/ACV.1994.341300

[13] Ryan Stevens and Ian Pudney. 2017. Blur select faces with the updated Blur Faces tool. https://youtube-eng.googleblog.com/2017/08/blur-select-faces-with-updated-blur.html [Online; posted 21-August-2017].

[14] Reihaneh Torkzadehmahani, Peter Kairouz, and Benedict Paten. 2019. DP-CGAN: Differentially Private Synthetic Data and Label Generation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.

[15] Zhou Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing* 13, 4 (April 2004), 600–612. https://doi.org/10.1109/TIP.2003.819861

[16] J. Yang, J. Wright, T. S. Huang, and Y. Ma. 2010. Image Super-Resolution Via Sparse Representation. *IEEE Transactions on Image Processing* 19, 11 (Nov 2010), 2861–2873. https://doi.org/10.1109/TIP.2010.2050625