I. J. Computer Network and Information Security, **2024, 3, 84-98**
Published Online on June 8, 2024 by MECS Press (http://www.mecs-press.org/)
DOI: 10.5815/ijcnis.2024.03.07

A Novel Privacy Preservation Scheme by Matrix Factorized Deep Autoencoder

Pooja Choudhary*
Department of Computer Science and Application, Kurukshetra University, Kurukshetra, India
E-mail: poojachoudhary3108202@gmail.com
*Corresponding author

Kanwal Garg
Department of Computer Science and Application, Kurukshetra University, Kurukshetra, India
E-mail: gargkanwal@kuk.ac.in

Received: 05 May 2023; Revised: 22 August 2023; Accepted: 10 October 2023; Published: 08 June 2024

Abstract: Data transport entails substantial security to avoid unauthorized snooping as data mining yields important and quite often sensitive information that must be and can be secured using one of the myriad Data Privacy Preservation methods. This study aspires to provide new knowledge to the study of protecting personal information. The key contributions of the work are an imputation method for filling in missing data before learning item profiles and the optimization of the Deep Auto-encoded NMF with a customizable learning rate. We used Bayesian inference to assess imputation for data with 13%, 26%, and 52% missing at random. By correcting any inherent biases, the results of decomposition problems may be enhanced. As the statistical analysis tool, MAPE is used. The proposed approach is evaluated on the Wiki dataset and the traffic dataset, against state-of-the-art techniques including BATF, BGCP, BCPF, and modified PARAFAC, all of which use a Bayesian Gaussian tensor factorization. Using this approach, the MAPE index is decreased for data which avails privacy safeguards than its corresponding original forms.

Index Terms: Privacy Preservation, Matrix Factorization, Autoencoder, Deep Learning.

1. Introduction

During the transmission of data between variant parties, ensuring security is a crucial step so that third parties don't get insight into the contents of the original communication. The information thus gained via data mining's output is both sensitive and significant and must be hidden using one of many techniques collectively recognized as the methods of Data Privacy Preservation (DPP). Users may feel more secure about sharing their personally identifiable information (PII) with service providers and applications thanks to this technology while still enabling businesses to take advantage of data-driven strategies. Due in large part to a new family of technologies that have arisen over the last few years, these privacy-centred solutions bring together the rising need for user-level data with growing requests for consumer privacy, dispelling the myth that data-driven marketing and privacy are incompatible.

Differential privacy may guarantee individuality in a system for disseminating fresh open data [1]. Unrestrained expansion in surveillance, health care, and finance creates a lot of data, including sensitive information that might become public if not cleared. Federated learning algorithms may safeguard private data and categorize non-private data by privately learning from a privacy filter (PF) across numerous sensitive datasets [2]. PFs and ALS-based NMF designs may protect data before release. A data release mechanism that meets differential privacy and rigorous protection of worker locations is quite functionally dexterous when used with a partitioning method based on worker density and non-uniform worker distribution. A Geocast region selection method for task assignment that considers worker travel distance and notification number adds to the potentiality of achieving the goal [3]. Matrix factorization preserves data patterns and facilitates vector machine classification gauge accuracy after data distortion [4]. An anonymous protection model and re-publication method [5] protect sensitive set-valued datasets with internal and external updating during re-publishing scenarios and sequential anonymous publication. Local differential privacy can protect even unknown graph community features [6]. Generative adversarial network (GAN) models may study the underlying distribution of a dataset and randomly build realistic samples according to its anticipated distribution to limit information leakage, especially when employed on private or sensitive data. Gradient-pruning and well-controlled noise provide GANobfuscator with the much-entailed differential privacy [7]. The method ensures differential privacy, high-quality data, and realistic

privacy budgets. A differentially private kernel k-means approach generates realistic synthetic samples that perform arbitrary counting queries and accurately represent massive datasets with strong privacy guarantees and high usability [8]. MAC [9] anonymizes the allocated social network better than AC and retains data uniqueness without suppressing or generalizing values. A hybrid method based on nonnegative matrix factorization and random perturbation technologies may be used to build the recommendation system [10].

Healthcare data processing, sharing, and privacy can be improved via deep neural networks. Privacy-preserving IoT time-series data access is provided via user-customized deep autoencoder objective functions [11]. GAN-based Private Aggregation of Teacher Ensembles (PATE) architecture PATE-GAN [12] is yet another efficient methodology. High-dimensional data may be protected using Mondrian-based k-anonymity on Deep Neural Network (DNN) architecture [13]. Standard anonymization fails to eliminate PII and balance privacy-data utility. GBDT training [14] uses training data gradients and leaf node cutting to reduce sensitivity constraints and improve noise allocation. Standard k-anonymity-based privacy-preserving algorithms cannot safeguard large-scale trajectory data, and frequent recalculation increases processing cost and trajectory availability [15]. MSA generalization correlation assaults on 1:M records may be defeated by "(p,l) - Angelization" [16]. Matrix decomposition and Kronecker product features provide flexible randomized response techniques [17]. BIRCH [18] secures sensitive social media material. MDACF tree data release study [19]. ConTPL [20] transforms differentially private streaming data release to automatically limit Temporal Privacy Leakage (TPL). Shallow NMF-based network community discovery optimizes efficiently but is tougher to train owing to its massive factor matrices [21]. Data distortion using NMF to safeguard driver location privacy and grouping drivers by position, relative distances, directions, and timestamps may overcome VANET performance reduction due to privacy breaches [22]. PATE [23] black-boxes discontinuous dataset models like user records from different subsets. The technique averages 84% on generic datasets and 93.8% on specific ones [24]. However, some of the challenges and problems that are to be addressed are that,

- with the increasing volume of data, especially in IoT and healthcare applications, ensuring privacy in high-dimensional data becomes a complex task. Techniques to effectively preserve privacy in such datasets are required.
- While autoencoders offer a method for data compression and feature learning without relying on labels, their architecture can be complex, especially in the context of matrix factorization-based autoencoders. Ensuring the scalability and efficiency of these models is crucial.
- However, in privacy-preserving can be challenging, especially when it involves sensitive attributes.

Hence, the application of deep neural networks and privacy-preserving techniques hold great promise in enhancing healthcare data processing, sharing, and privacy. However, challenges related to high-dimensional data, complex autoencoder architectures, and handling sensitive missing attributes need to be effectively tackled to ensure the continued success of privacy-preserving solutions in data-driven applications.

The methodologies have been thoroughly analyzed during the paper. The details of the analysis could be summarized to some extent in section 2. as the review of the studied literature.

1.1. Motivation & Contributions

The fundamental cornerstone of privacy-preserving data mining is to provide algorithms for transforming the source data in such a manner that sensitive information is protected even after mining has been completed. Data with a high temporal and spatial granularity and a high degree of correlation between their samples calls for an analytical approach that is both flexible and extensible. By selectively changing just the sensitive parts of data, this operation may be carried out without affecting the desired or non-sensitive data. One common method for creating such a model is replacing sensitive data with data that is almost identical to the non-sensitive data [25-26]. There are myriad ways of achieving the objectives – autoencoders being one of the pronounced methodologies among them. Autoencoders are a type of neural network that is used for data compression and feature learning. They can be used for data privacy preservation because they can learn to reconstruct an input data point from a lower-dimensional representation, known as the encoding, without relying on any labels or supervision. A way in which autoencoders can be used for data privacy preservation is by training an autoencoder on sensitive data and then only sharing the encoder part of the network with others [27]. The encoding produced by the encoder can be used as a substitute for the original data, while the decoder, which is not shared, can be used to reconstruct the data if necessary. This allows the sensitive data to be used in downstream tasks without exposing the raw data to others.

There are numerous registered advantages of a matrix factorization-based autoencoder scilicet. Its straightforward architecture makes it less time-consuming and resource-intensive to train. Particularly helpful in cases when either training data or computing resources are scarce [28]. Further justification can be that autoencoders built on top of matrix factorization tend to be more resistant to data noise and outliers. This may be helpful for protecting privacy since it makes it harder for unauthorized parties to decode the encoded representation and get private data. Finally, when attempting to grasp the correlations between various aspects in the data, matrix factorization-based autoencoders may be more interpretable than their deep neural network counterparts. The primary contributions of this study encompass:

- The development of an imputation technique to address missing data prior to the acquisition of item profiles.

- The enhancement of the Deep Auto-encoded Non-negative Matrix Factorization (NMF) through the incorporation of a customizable learning rate.
- Bayesian inference is also employed to evaluate the imputation technique for datasets with missing values at random, specifically with missingness rates of 13%, 26%, and 52%.

1.2. Literature Review

Security is entailed even in streamlined data transfer between parties. Myriad Data Privacy Preservation (DPP) methods can be used to conceal data mining's crucial yet sensitive output. This technology enables transferring PII with service providers and applications safer and allows organizations to employ data-driven initiatives. These privacy-focused technologies reconcile data-driven marketing with customer privacy. More particularly patient data circulation requires demands considerable vigilance and caution.

Deep neural network pairs are designed to train on synthetic SPRINT trial participants' data but with differential privacy. Deep neural networks are capable in elevating data sharing, clinical data analysis as well as participant privacy. A framework for releasing fresh open data can also safeguard individuals' uniqueness via differential privacy [1], which may be customized to varied utility cases, from time-series to continuous and discrete data production. Unrestrained expansion in surveillance, health care, and finance, generate a lot of data and include sensitive material that might become public if not adequately sanitized. Privately learning from a privacy filter (PF) across many sensitive datasets can facilitate a federated learning algorithm that protects private data and classifies non-private data accurately [2]. A PF and a new non-negative matrix factorization (NMF) design that employs alternating least squares (ALS) can potentially protect data before publication. Only utility representations are shared by all PFs to train a federated learning model. A data release mechanism that satisfies differential privacy and rigorous protection of worker locations, when amalgamated with a partitioning method based on worker density and non-uniform worker distribution can fulfill the objective working along with a Geocast region selection method for task assignment that considers worker travel distance and notification number [3]. Privacy-preserving data mining accuracy on distorted datasets [4] uses matrix factorization retains data pattern and allows vector machine classification to evaluate accuracy preservation after data distortion by different methodologies. Based on dynamic traditional relational data re-publishing research, an anonymous protecting model and re-publication algorithm [5] protects sensitive information of set-valued dataset with both internal and external update during re-publishing scenario and sequential anonymous publication.

Community structure information and synthetic social network data are released under edge probability reconstruction limits using a local differential privacy strategy for social network publishing. An uncertain graph is injected into local differential privacy to safeguard user community attributes [6]. Generative adversarial network (GAN) model can be employed to examine the underlying distribution of a dataset and randomly create realistic samples according to their projected distribution to limit information leakage, especially under GANs deployed on private or sensitive data. GANobfuscator, [7] a differentially private GAN, uses gradient-pruning and precisely tuned noise to provide differential privacy under GANs. Differential privacy, high-quality data, and realistic privacy budgets are guaranteed by the method. Data can be grouped into k sets and sent to distinct generative neural networks using a unique differentially private kernel k-means approach, which creates realistic synthetic samples that consistently conduct arbitrary counting queries and properly represent huge datasets with strong privacy guarantees and significant usability [8]. Modified Anatomy-based Clustering (MAC) [9] anonymizes the ascribed social network better than AC and retains data originality by neither suppressing or generalizing values. It reduces Total Loss second-best and retains community structure best. On a related note, the recommendation system can be implemented using a hybrid method based on nonnegative matrix factorization and random perturbation technology that protects user private data and is not sensitive to the number of neighbours k and t [10]. This method explains local characteristics, yields reliable recommendations, lets the server swiftly acquire recommendation data, and protects user privacy.

A user-customized deep autoencoder objective function offers a privacy-preserving sensing framework for IoT time-series data access [11]. Replacement Autoencoder transforms discriminative data characteristics into non-sensitive inference features. On a related note, PATE-GAN is a GAN-based Private Aggregation of Teacher Ensembles (PATE) architecture having comparable dexterity [12]. Where, standard anonymization fails to remove PII and addresses the privacy-data utility trade-off, a Mondrian-based k-anonymity hinged Deep Neural Network (DNN) architecture has the aptitude to protect high-dimensional data privacy [13]. A new GBDT training method [14] adaptively regulates training data gradients and leaf node clipping to shrink sensitivity bounds and enhance noise allocation. Standard k-anonymity-based privacy-preserving techniques fail to protect large-scale trajectory data and frequent recalculation of the current trajectory data increases processing cost and trajectory availability [15]. Privacy under (k, δ) security criteria can be reinforced using time division, the quasi-identifier, and clustering using the generalization function. Even a well formalized and analyzed MSA generalization correlation assault, a unique attack on 1:M records utilizing MSAs, can be tackled through a privacy-preserving data publishing mechanism called "(p,l) - Angelization" that outperforms its respective global analogues.

Matrix decomposition and Kronecker product features provide flexible randomized response strategy frameworks. Reconstruction strategies include optimal flipping perturbation reconstruction [17]. Balanced Iterative Reducing and Clustering utilizing Hierarchies (BIRCH) [18] can efficiently safeguard sensitive social media material. The study highlights MDACF tree data release. In a different vein, foolability is apparently sensitive to recollection overfitting and this generates difficulties of deceiving systems that produce fresh synthetic data. DP–Foolability implies GAM–

Foolability [19]. Temporal correlations' impact on privacy loss from continually created data is underappreciated. ConTPL [20] automatically turns a differentially private streaming data release mechanism into one bounding Temporal Privacy Leakage (TPL) within a predefined level to tackle it. Shallow NMF-based network community discovery is identified and implemented as a feature representation learning inspired deep autoencoder-like NMF (DANMF) [21] which facilitates efficient optimization, but its considerably big factor matrices make its training tougher than shallow NMF-based techniques. Novel innovative learning algorithms with enhanced privacy cost analysis under differential privacy are nifty in training deep neural networks with non-convex objectives within a minimal privacy budget at an acceptable cost in software complexity, training efficiency, and model quality. Data distortion using Nonnegative Matrix Factorization (NMF) to safeguard driver location privacy and grouping drivers by position, relative distances and directions, and timestamps can overcome VANET performance reduction due by privacy breach concerns [22]. Model training data storage leaks sensitive data. PATE [23] inspired method, scaled to larger projects with great usefulness and strong privacy (ε <1.0), black-boxes numerous discontinuous dataset models, such as user records from different subsets. The model averages 84% accuracy on generic datasets and 93.8% on specific datasets [24]. Dispersed singular value decomposition (DSVD-autoencoder) [24] lets autoencoders train in distributed situations without sharing raw data, protecting privacy. It's scalable, noniterative, hyperparameter-free, and convergence-free, and thus, can handle enormous data sets with data privacy restrictions, minimize training time from hours to seconds while keeping accuracy, and function as a lightweight model that may benefit an Internet of Things network of low-capacity sensors. Participants' anonymity is protected during GAN model training also seen in [25]. Differentially Private Conditional GAN (DP-CGAN) training framework based on a new clipping and perturbation strategy, generates synthetic data and labels and also clips the gradients of discriminator loss on real and fake data separately [26]. Similar works are seen in [27-28].

It sums up two sets of gradients, adds Gaussian noise to the sum, and leverages the recently introduced Rényi differential privacy (RDP) accountant to track the privacy budget. The technique improves model performance, preserves training dataset privacy, and yields aesthetically and empirically good results on the MNIST dataset using a single-digit epsilon parameter in differential privacy. Using simulated UK primary care data to demonstrate the design, imply CVD research may employ synthetic primary care data. One-class classification overcomes privacy problems in anomaly or novelty detection.

1.3. Paper Organization

The rest of this article is divided into the following sections. In Section II, the general technique is provided in a structured overview. The complete proposed methodology is provided in Section III. The results and comparison with other cutting-edge techniques are presented in Section IV. Section V presents the conclusion as well as the future scope.

2. Preliminaries

2.1. Standard NMF

The approach of non-negative matrix factorization (NMF) is extensively utilitarian in the processing of large-scale data, including pictures and audio signals. NMF closely demonstrates a low-dimensional representation of a data set. The most significant advantage of NMF is the augmentation of the comprehension of deconstructed submatrices while still allowing high-dimensional information to reduce redundant pieces and lower-dimensional spaces. In order to find patterns in large-scale matrix data that may be explained by analytical functions, NMF enforced non-negativity restrictions on the element matrix. It splits a positive matrix into a base matrix and a coefficient matrix, both of which are also positive. This essentially projects that the results of its breakdown have a rather precise physical meaning, something that is no longer useful in other matrix decomposition techniques. The descriptions of the frequently used notations encountered henceforth are briefly demonstrated through the Table 1.

Nonnegative Matrix Factorization (NMF)

NMF makes use of a feature matrix X that is not negative and has feature variables wherein columns represent the nodes' social connections.

$$\mathrm{X} = [\mathrm{x}_{1,}, \mathrm{x}_{2,} \dots\dots\dots \mathrm{x}_{\mathrm{N}}]^{\mathrm{T}} \in \mathbb{R}^{\mathrm{N} \times \mathrm{k}} \tag{1}$$

Where $x_i \in \mathbb{R}^k (i = 1, \dots . N)$ are the attributes that a signal node should have. The formula often used to determine NMF is as follows:

$$X \approx HW \tag{2}$$

There are two main terms for describing data scilicet $H \in \mathbb{R}^{N \times d}$ along with $W \in \mathbb{R}^{d \times k}$. The decomposition is effectuated in such a manner that the reconstructed matrix is almost indistinguishable from the original X. The rows of H are the decomposition (or encoding) coefficients, while the columns of W are typically referred to as the basis vector [29]. The initial characteristics are generated by linearly combining these basis vectors. On a similar vein, certain

decomposition techniques, such as principal component analysis, allow for the elimination of components, but NMF only allows for additive actions, making it a more logical and practical breakdown technique. Since this is a nonlinear optimization problem, the minimum value of the following objective function can be possibly found through:

$$\min_{W,H} E(X||HW) = \frac{1}{2}\|X - HW\|_F^2 \quad s.t. H \geq 0, W \geq 0 \tag{3}$$

Where $\|.\|_F$ refers to the Frobenius Norm.

Table.1 Frequently used Notations

Notations	Explanation
$X = [x_1, x_2, \ldots\ldots\ldots x_N]^T$	Input vector with k features and N is the total number of nodes in the DTN network
$X \approx H\,W$	$H \in \mathbb{R}^{N\times d}$ basis vector and $W \in \mathbb{R}^{d\times k}$ are decomposition (or encoding) coefficients with d decomposed
$\overline{X} \approx WW^TX$	Reconstructed Matrix
$\mathcal{X}\epsilon\mathbb{R}^{I\times J\times K}$	Third order tensor matrix used in the data imputation
N	No. of elemental disintegrations of $\mathcal{X}$, each of rank 1
$\emptyset_D, \theta_J, \eta_F$	three-dimensional tensor bias factors
μ	mean of the three bias variables
C	Represents the communities in the data X
p	Number of layers in deep autoencoder
$\mathcal{L}_E$	A unified loss function
$\lambda tr(W_p L {W_p}^T)$	The graph regularizer with λ as regularization parameter, tr is the trace function, and L is the graph Laplacian matrix
$\mathcal{L} = \mathcal{L}_D + \mathcal{L}_E + \lambda\mathcal{L}_{Reg}$	Complete loss function
H_i, W_i, W_p	Mapping Matrix (H_i), feature matrix (W_i) and the community membership matrix (W_p)
Θ	Lagrangian Multiplier
θ_{i-1}	Variable to represent the set of feature matrix $H_1H_2 \ldots\ldots H_{i-1}$
ξ_{i+1}	Variable to represent the set of feature matrix $H_{i+1\ldots\ldots\ldots}H_{p-1}H_p$
y_i , z_i	Accounts for missing values in the calculation of MAPE

The local optimum solution to this issue may be discovered using the multiplicative update technique and the update criteria given below.

$$H_{\{i,j\in N\}} \leftarrow H_{ij}\frac{(W^TX)_{ij}}{(W^TWH)_{ij}} \tag{4}$$

$$W_{\{i,j\in N\}} \leftarrow W_{ij}\frac{(X^TW)_{ij}}{(H^TWW^T)_{ij}} \tag{5}$$

2.2. NMF Autoencoder

The projection of the original data onto the basis matrix to produce the coefficient matrix is the encoding operation in an autoencoder. Decoding is analogous to data reconstruction from the coefficient matrix. For NMF, the most straightforward formulation is as a one-layer autoencoder:

$$\textit{Encoder: } H = W^TX \tag{6}$$

$$\textit{Decoder: } \overline{X} = HW \tag{7}$$

$$\textit{Autoencoder: } \overline{X} = WW^TX \tag{8}$$

Here, the reconstructed matrix is represented by $\overline{X}$. The conception of the aforementioned process is exemplified through the following illustration:

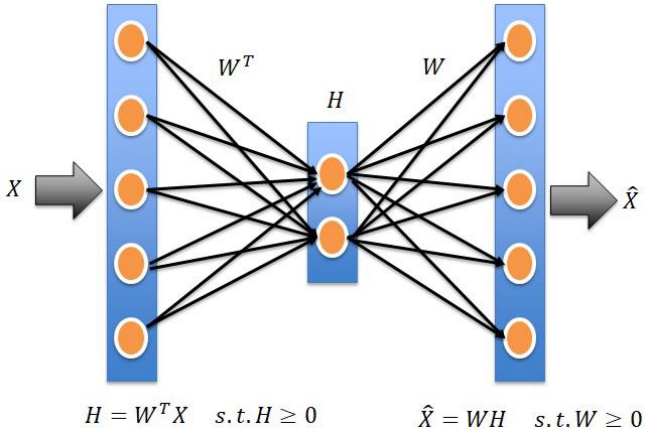

Fig.1. A single layer NMF autoencoder

3. Imputation Based Privacy Preservation

3.1. Proposed Methodology

As part of the paper's endeavor to advance knowledge in this area, cardinal contribution involves two parts of the study: the Deep Auto-encoded NMF optimization with a variable learning rate, and an imputation approach designed to fill in the gaps of the unobserved data before learning the item profiles. As the backbone of this study, tensor matrices are used, which are a special kind of multidimensional matrix. Tensor factorization may be accomplished in a number of ways, and several methods are available [30-31]. However, the PARAFAC method not only delivers faster convergence, but also more stable tensor decomposition. Decomposing the tensor into a rank one matrix yields the matrix. Some studies have used tensor decomposition to impute high-dimensional data. This only occurred in a few instances. Unpredictable fluctuations, or noise, in the data values are not yet accounted for by the PARAFAC decomposition. Outliers are extreme deviations from the norm. In this work, a revised version of the regularized PARAFAC algorithm is presented. The tensor is split into three parts before bias matrices are constructed for each component. These two measures are used to ensure the software does not become stuck on a local optimal solution.

3.2. Imputation

As was seen previously, one of the most constructive areas of study is the secure extraction of crucial latent features from the data matrix. Inspired by this finding, the NMF is modified into a powerful and efficient data representation method, to include differential privacy (DP) [32-33]. This is not an easy process, and most current methods are not transferable to NMF due to the following issues: Since the DP-added sounds might be negative, the non-negative of the learnt latent profiles is no longer guaranteed. Second, NMF is iterative, and errors seem to accumulate over time. Even though an adaptive sensitivity is applied in the differentially private MF process, items with sparse ratings are more susceptible to the disturbances introduced by DP, leading to learnt item profiles that deviate significantly from their ideal solution. It doesn't matter where the noise is introduced into, this problem occurs with other differentially private MF techniques as well. To get around this issue, an imputation technique is developed to fill in the blanks of the unobserved data before learning the item profiles. Then this is used to offer a method for protecting users' anonymity using DANMF and imputation.

To proceed with the procedure, the Parallel Factor Analysis (PARAFAC) technique is used to factorize the tensor matrix. Third-order tensor $\mathcal{X}\epsilon\mathbb{R}^{I\times J\times K}$ data is used in this work's data imputation, and the PARFAC decomposition for high-rank ($R > 10$), may be shown as in equation 9 [34].

$$\mathcal{X}_{itj} \approx [\![D, J, F]\!] := \sum_{r=1}^{R} D_{ir} J_{tr} F_{jr} \tag{9}$$

Since the primary concern involves the decomposition up to rank $R = 1$, it instinctively renders the fact that equation 11 holds true for sufficiently big R [35]. For example, the tensor matrix may be broken down into N elements, each of rank 1.

$$\mathcal{X}\epsilon\, a^{(1)} \circ a^{(2)} \circ a^{(3)} \circ \ldots . a^{(N)} \tag{10}$$

The cross product is denoted by $\circ$. In terms of tensors, $\mathcal{X}\epsilon\, a \circ b \circ c$ representation for the third order. The proof that equation 9 may be decomposed to a tensor of first order is as follows:

$$\mathcal{L}(\mathcal{X}; D, J, F) = \frac{1}{2}\|\mathcal{X} - [D, J, F]\|_F^2 \tag{11}$$

The Frobenius norm, represented by $||\ .\ ||_F^2$. This repeatedly solves the non-convex optimization problem by

$$\min_{D,J,F} \mathcal{L}(\mathcal{X}; D, J, F) \tag{12}$$

Because of the non-uniform scaling of factors, Equation 12 has a conflict with local optimum solutions [36]. To fix this, Paatero [37] proposed including a Tikhonov regularization factor in Equation 12. This effectively eliminated the need for so many local solutions in Equation 11 and brought it up to date as

$$\mathcal{L}_P(\mathcal{X}; D, J, F) = \frac{1}{2}\|\mathcal{X} - [D, J, F]\|_F^2 + \frac{\lambda}{2}(\|D\|_F^2 + \|J\|_F^2 + \|F\|_F^2) \tag{13}$$

$\lambda > 1$ is a control attribute for λ the regulating parameter. China Traffic Data is a dataset that has been utilized in this research. There are three components to this tensor data set: road segments, time slots, and elapsed times. However, it is not always the case that the combined rating for user and content is mapped in the latent space using Equation 13. Irrespective of how well the data is characterized, the imputation might be off if the temporal dependence on the data is not considered. Equation 3 has to account for this biasing with respect to both the user and the material. Addition to Eq. 13 is seen in Eq. 14. In this context, the PARAFAC factorization has been referred to as "augmented PARAFAC."

$$\mathcal{L}_A(\mathcal{X}; D, J, F) = \frac{1}{2}\left\|\mathcal{X} - \mu - \emptyset_D - \theta_J - \eta_F - [D, J, F]\right\|_F^2 + \frac{\lambda}{2}(\|D\|_F^2 + \|J\|_F^2 + \|F\|_F^2 + \|\emptyset_D\|_F^2 + \left\|\theta_J\right\|_F^2 + \|\eta_F\|_F^2) \tag{14}$$

Here's how to put it in its simplest form:

$$\begin{aligned}\mathcal{L}_A(\mathcal{X}; D, J, F) = \frac{1}{2}\Big(\|\chi\|_F^2 + \|\mu\|_F^2 + \|\phi_D\|_F^2 + \left\|\theta_J\right\|_F^2 + \|\eta_F\|_F^2 + \|[D, J, F]\|_F^2\Big) - \langle(\chi, \mu)\rangle - \langle(\chi, \phi_D)\rangle - \\ \langle\left(\chi, \theta_J\right)\rangle - \langle(\chi, \eta_F)\rangle - \langle(\chi, [D, J, F])\rangle + \langle(\mu, \phi_D)\rangle + \langle\left(\mu, \theta_J\right)\rangle + \langle(\mu, \eta_F)\rangle + \langle(\mu, [D, J, F])\rangle + \\ \langle\left(\phi_D, \theta_J\right)\rangle + \langle(\phi_D, \eta_F)\rangle + \langle(\phi_D, [D, J, F])\rangle + \langle\left(\theta_J, \eta_F\right)\rangle + \langle(\theta_J, [D, J, F])\rangle + \\ \langle(\eta_F, [D, J, F])\rangle + \frac{\lambda}{2}(\|D\|_F^2 + \|J\|_F^2 + \|F\|_F^2 + \|\emptyset_D\|_F^2 + \left\|\theta_J\right\|_F^2 + \|\eta_F\|_F^2)\end{aligned} \tag{15}$$

Where $\emptyset_D, \theta_J, \eta_F$ represents the three-dimensional tensor bias factors, the value μ is the mean of the three variables. The biases represent the degree to which each variable deviates from the mean μ. The proposed tensor $\mathcal{X}$ is shown in figure 2. The tensor factor matrix may alternatively be written as $\widehat{\mathcal{X}}_{i,j,t} = \mu + \emptyset_D + \theta_J + \eta_F + [D, J, F]$.

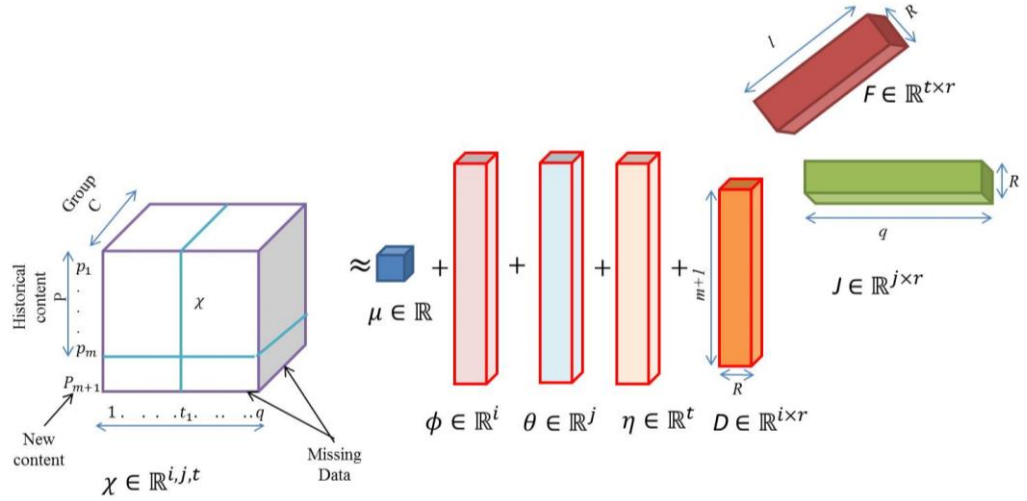

Fig.2. Tensors of the third order are proposed to be factorized using an improved version of the PARAFAC factorization

The Alternating least square technique and, Gradient descent optimization, are used to find the minimum in Equation 15 in the proposed modified-PARAFAC, which is an iterative procedure. In contrast, these strategies tend to become trapped in local minima and have a slow rate of convergence. Fast and globally optimized Adam optimization [38] is critical to the development of deep learning. Adam optimization has been used to factor the tensor $\mathcal{X}$. Additionally, the bias is changed repeatedly in each stochastic step. A random number between 0 and 0.1 is used to set this as the starting point. Even though these biases' starting point of input is completely arbitrary, for the respective purposes it has been fixed at $b \epsilon [0,0.1]$, where b is a representation of the biases. In each iteration of the Adam optimization process, the biases are changed. Adam optimization has an objective function given by Equation 14, and the error from that is utilized to adjust the bias. The error is iteratively minimized by multiplying the error from equation 14 by a regularization value

that decreases exponentially. The chaotic disturbance in error is multiplied by a less random component, which is determined by a logistic map, in order to produce an additive error term. The revised biases are as follows:

$$\emptyset_{D,t} = \emptyset_{D,t-1} + (\mathcal{X}_{itj} - \widehat{\mathcal{X}}_{i,j,t}) * x_n * e^{-\alpha\frac{t}{T}} \quad (16)$$

Here, the factor deviation is $\left(\mathcal{X}_{itj} - \widehat{\mathcal{X}}_{i,j,t}\right)$ along with the constant $\alpha$of value 20. In the Adam optimization, t and T stand for the current iteration and the maximum iteration, respectively. The parameter x_n that is altered by logistic mapping is found to be $x_n = rx_n(1 - x_n), n = 0,1,2,3 \ldots .$ The r represents a scalar input to the system as $[0,4]$ and $x_n \epsilon$ $[0,1]$. In this manner, both θ_J and η_F will be updated along with.

Algorithm 1 depicts the pseudo code for the revision of biases.

Algorithm 1: Iterative updates to the tensor biases are implemented in a pseudocode

Input: $\mathcal{X}$, $[D, J, F]$
Output: $\emptyset_D$, θ_J, η_F, μ

1. Initialize the $\emptyset_D$, θ_J, η_F with uniform random numbers in between 0 and 0.1
2. Calculate the average $\mu = \frac{1}{\Omega}\sum_{(i,j,t)\epsilon\Omega} x_{i,j,t}$
3. While not converge do
4. For $i\epsilon(1, n1)$
5. Update $\emptyset_{D,t} = \emptyset_{D,t-1} + (\mathcal{X}_{itj} - \widehat{\mathcal{X}}_{i,j,t}) * x_n * e^{-\alpha\frac{t}{T}}$
6. end for
7. for j$\epsilon(1, n2)$
8. Update $\theta_{J,t} = \theta_{D,t-1} + (\mathcal{X}_{itj} - \widehat{\mathcal{X}}_{i,j,t}) * x_n * e^{-\alpha\frac{t}{T}}$
9. end for
10. for t$\epsilon(1, n3)$
11. Update $\eta_{F,t} = \eta_{D,t-1} + (\mathcal{X}_{itj} - \widehat{\mathcal{X}}_{i,j,t}) * x_n * e^{-\alpha\frac{t}{T}}$
12. end for
13. end while loop

3.3. DANMF

Lee and Seung [39] introduced the NMF, a family of algorithms in multivariate analysis and linear algebra where a matrix X is factorized into (usually) two matrices W and H with the property that all three matrices have no negative elements, to cover cases where the data matrix H is not strictly non-negative. As a consequence, NMF is able to learn new, lower-dimensional characteristics from the data, all of which have a natural non-negativity that makes it easy to evaluate the resultant matrices, where H represents soft membership indications for each data point.

It's conceivable that the complicated hierarchical and structural information is encoded in the mapping W between this new representation according to H and the original characteristics X. Take, for instance, the challenge of associating faces with their owners; besides the subject's likeness, a face photograph also conveys information about the subject's attitude and expression. A superior higher-level feature representation H, as shown in Figure 3, might be achieved, according to this line of thinking, by further factorizing this mapping W in such a manner that each factor provides an additional layer of abstraction decreasing the dimensionality of the representation. Using the NMF idea applied to a multi-layer structure that can learn hidden representations of the original data, a unique Deep Autoencoding technique has been offered in this study. Figure 3 shows how expanding the model to a deep one enables this model to learn new representations of the original data that continue to have a more than relevant meaning according to the many latent qualities of the used dataset.

In this excerpt, DANMF's utility has been demonstrated in establishing relationships that expose a previously concealed structure. $G_t = \{N, ST\}$, Adjacency matrix-based graph assembly whose nodes represent social ties. Conspicuously, there is a community in every network. C indicates the group of cliques, i$C = \{C_i || C_j \neq \emptyset, C_i \neq C_j 1 \leq i, j \leq m\}$, where m is the total number of distinct communities. Discontinuous social connection identification is required to avoid overlap. $C_i \cap C_j = \emptyset, i \neq j$. Non-negative matrices represented by $H \in \mathbb{R}^{N\times d}$ and $W \in \mathbb{R}^{d\times k}$ respectively become available after breaking down. Relationships between communities are shown for each row in H and additionally, the columns in W indicate the many communities to which each node belongs. The connection that a group has with its origins $[H]_{im}[W]_{mj}$ involving the interlinkage $[\widehat{ST}]_{ij} = \sum_{l=1}^{m}[H]_{il}[W]_{lj}$ of the nodes i and j due of their shared involvement in their local areas. Figure 4 depicts the structural differences between NMF and the deep autoencoder. The count of layers in the deep autoencoder aids with feature extraction and delay.

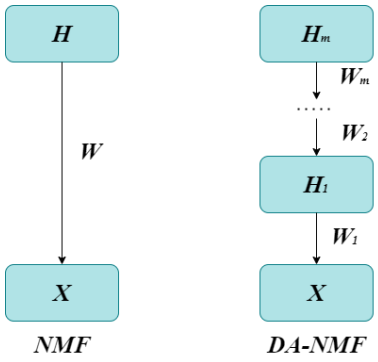

Fig.3. Expanding the NMF to a deep model

The optimized objective function, defined by Eq.3, is the sketched objective. All of the data gathered from projects how the individual nodes are connected to the larger communities. Implementing the abstraction layers, DANMF's implicit low-level hidden feature extraction takes into aspects$H$. Therefore, X can be assimilated into$p+1$ composite nonnegative factor matrices as:

$$X \approx H_1 H_2 \dots\dots\dots H_p W_p \tag{17}$$

where $W_p \in \mathbb{R}^{m\times N}$,$H_i \in \mathbb{R}^{r_{i-1}\times r_i}(1 \le i \le p)$, coupled with constraints over $N = r_0 \ge r_1 \ge \cdots\dots\dots \ge r_{p-1} \ge r_p = m$. The breakdown of Eq. 2 into its p constituent layers is as follows:

$$\begin{aligned} W_{p-1} &\approx H_p W_p, \\ &; \\ &; \\ W_2 &\approx H_3 \dots\dots H_p W_p \\ W_1 &\approx H_2 \dots\dots H_p W_p \end{aligned} \tag{18}$$

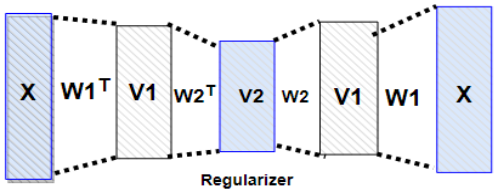

Fig.4. Implicit structure of deep ANMF networks

Since the constraint of nonnegative is maintained on W_i all intermediate representations, the similarities between nodes at different levels of granularity may be captured accurately. The structure of deep layers allows for more accurate social-link identification. Learning to factor matrices is accomplished through the goal function shown below. Optimization of Equation 19 is carried out to draw out the latent features.

$$\min_{H_i W_p} \mathcal{L}_D = \left\| X - H_1 H_2 \dots\dots\dots H_p W_p \right\|_F^2 \tag{19}$$

$s.t. H_i \ge 0, W_p \ge 0, \forall_i = 1,2, \dots\dots p$

The autoencoder's original network may be decoded and reconstructed using Eqs. 3 and 19 respectively. Inclusion of the encoder component is required for optimal results. An ideal membership matrix for a group depends on two pillars: H mapping to get as close as possible to the original network design, and$H$ 2) Direct projection of the initial social relationship into the area of community membership through $U = H^T X$. A unified loss function is developed for the encoder and decoder to guide the learning process and enable the nodes to arrive at the correct community membership. This is stored using the encoder's own objective function.

$$\min_{H_i W_p} \mathcal{L}_E = \left\| W_p - H_p^T \dots\dots\dots H_1^T H_2^T X \right\|_F^2 \tag{20}$$

$s.t. H_i \ge 0, W_p \ge 0, \forall_i = 1,2, \dots\dots p$

The Deep ANMF unified goal function is as follows:

$$\min_{H_i W_p} \mathcal{L} = \mathcal{L}_D + \mathcal{L}_E + \lambda \mathcal{L}_{Reg} = \left\| X - H_1 H_2 \dots\dots\dots H_p W_p \right\|_F^2 + \left\| W_p - H_p^T \dots\dots\dots H_1^T H_2^T X \right\|_F^2 + \lambda tr\left(W_p L {W_p}^T\right) \quad (21)$$

$s.t. H_i \geq 0, W_p \geq 0, \forall_i = 1,2, \dots\dots p$

Here, $\lambda tr\left(W_p L {W_p}^T\right)$, the graph regularizer having λ as assimilation feature, tr is the trace function, whereas L is the graph Laplacian matrix. It was learned by generalized approach $L - D - X$ (with D diagonal matrix formed of the $X's$ elements accumulated in the process).

The next step is to optimize the layers. During pretraining, the factor matrices H_i and W_i have approximations made for each layer. The training of the model may be completed more quickly thanks to pre-training. In the training process, the adjacency matrix is initially deconstructed as (collective connection X) by $X \approx H_1 W_1 \|X - H_1 W_1\|_F^2 + \|W_1 - H_1^T X\|_F^2$ $W_1 \in \mathbb{R}^{r_1 \times m}$ $H_1 \in \mathbb{R}^{m \times r_1} W_1$ $W_1 \approx H_1 W_2$ $\|W_1 - H_2 W_2\|_F^2 + \|W_2 - H_2^T W_1\|_F^2$ $W_1 \in \mathbb{R}^{r_1 \times m}$ $H_1 \in \mathbb{R}^{m \times r_1}$. This procedure is performed as many as necessary to ensure that all layers have been exposed to the pre-training phase. The layers are then fine-tuned by minimizing Equation 21 in a cyclic fashion.

Different relaxations are used to generate the updated matrices. Each variable in Eq. 21 foreshadows a component of the mapping matrix H_1 ($i \leq i < p$) are taken to be unchanging parameters. The lowered order of objectives is:

$$\min_{H_i} \mathcal{L}(H_i) = \mathcal{L}_D + \mathcal{L}_E + \lambda \mathcal{L}_{Reg} = \left\| X - \theta_{i-1} H_i \xi_{i+1} W_p \right\|_F^2 + \left\| W_p - {\xi_{i+1}}^T {H_i}^T {\theta_{i-1}}^T X \right\|_F^2 \quad (22)$$

$s.t. H_i \geq 0$

Here, $\theta_{i-1} = H_1 H_2 \dots . H_{i-1}$ and $\xi_{i+1} = H_{i+1} \dots\dots\dots H_{p-1} H_p$. Assuming $i = 1$, thus $\theta_0 = I$, the identity matrix. In a similar vein, when $i = p$, the $\xi_{p+1} = I$. By imposing the nonnegative restrictions on H_i the Lagrangian multiplier matrix Θ_i in equation. 22 is solved, and the associated objective function is obtained:

$$\min_{H_i \Theta_i} \mathcal{L}\left(H_i, \Theta_i\right) = \left\| X - \theta_{i-1} H_i \xi_{i+1} W_p \right\|_F^2 + \left\| W_p - {\xi_{i+1}}^T {H_i}^T {\theta_{i-1}}^T X \right\|_F^2 - tr\left(\Theta_i {H_i}^T\right) \quad (23)$$

Further attenuation can be done as

$$\min_{H_i \Theta_i} \mathcal{L}\left(H_i, \Theta_i\right) =$$
$$tr\left(\mathrm{X}^T \mathrm{X} + {W_p}^T W_p - 4 \mathrm{X}^T \theta_{i-1} H_i \xi_{i+1} W_p + {W_p}^T {\xi_{i+1}}^T {H_i}^T {\theta_{i-1}}^T \theta_{i-1} H_i \xi_{i+1} W_p + \mathrm{X}^T \theta_{i-1} H_i \xi_{i+1} {\xi_{i+1}}^T {H_i}^T {\theta_{i-1}}^T X - \Theta_i {H_i}^T\right) \quad (24)$$

The partial derivative in Eq. 24 has been updated to 0. That lower sum is

$$\Theta_i = -4 {\theta_{i-1}}^T X {W_p}^T {\xi_{i+1}}^T + 2\, \pi_i \quad (25)$$

The growth of π_i is

$$\pi_i = {\theta_{i-1}}^T \theta_{i-1} H_i \xi_{i+1} W_p {W_p}^T {\xi_{i+1}}^T + {\theta_{i-1}}^T \mathrm{X}^T \mathrm{X} \theta_{i-1} H_i \xi_{i+1} {\xi_{i+1}}^T \quad (26)$$

The Karush-Kuhn-Tucker complimentary slackness condition is a modification of Eqn. 26:

$$\Theta_i \odot H_i = \left(-4 {\theta_{i-1}}^T X {W_p}^T {\xi_{i+1}}^T + 2\pi_i\right) \odot H_i = 0 \quad (27)$$

where $\odot$ represents the operator for the element-wise product. Fixed-point representations require that Equation 26 is true at convergence. This is the result received after applying the update rule:

$$H_i \leftarrow H_i \odot \frac{2\, {\theta_{i-1}}^T X {W_p}^T {\xi_{i+1}}^T}{\pi_i} \quad (28)$$

Next up is the process of establishing updated rules for the community membership matrix W_p. The procedure is the same as in the first example, with the difference that the variable W_p is now fixed, and Equation 21 is simplified to read as:

$$\min_{W_p} \mathcal{L}(W_p) = \left\|X - \theta_p W_p\right\|_F^2 + \left\|W_p - {\theta_p}^T X\right\|_F^2 + \lambda tr\left(W_p L {W_p}^T\right) \tag{29}$$

$s.t. W_p \geq 0$

Algorithm 2: Advanced methods for optimizing and updating the ANMF

Input: X (Adjacency matrix) ,λ
Regularization parameter and layers (r_i) **Pre-training Process** $W_1, H_1 \leftarrow Punning\ (X, r_1)$ **For** $\boldsymbol{i} = 2: p$ $W_i, H_i \leftarrow Punning\ (X, r_i)$ **End** **Fine Training Process** **While (**not converged**) do** **For** i=1: p $\theta_{i-1} \leftarrow \prod_{q=1}^{i-1} H_q\ (\theta_o \leftarrow I)$ $\xi_{i-1} \leftarrow \prod_{q=1}^{i-1} H_q\ (\xi_{p+1} \leftarrow I)$ Update H_i using Equation.17 $\theta_i \leftarrow \xi_{i-1} H_i$ Update W_i using Equation.19 ($i < p$) or Equation.21 ($i = p$) **End** **End**
Output: Mapping Matrix (H_i), feature matrix (W_i) and the community membership matrix (W_p)

then diminishing W_p as

$$W_p \leftarrow W_p \odot \frac{2{\theta_p}^T X + \lambda W_p X}{{\theta_p}^T \theta_p W_p + W_p + \lambda W_p D} \tag{30}$$

You may see a table of equation notations (Table 1) here. W_i $(1 \leq i \leq p)$ represents the rule for optimizing feature matrices as the last step. W_i has no bearing on the objective and is thus unnecessary (Equation. 21). This enables the masked characteristics of each layer to be reclaimed.W_iThe augmentation for W_i involves

$$\min_{W_i} \mathcal{L}(W_i) = \|X - \theta_i W_i\|_F^2 + \left\|W_i - {\theta_i}^T X\right\|_F^2 \tag{31}$$

$s.t. W_i \geq 0$

In the new version,

$$W_i \leftarrow W_i \odot \frac{2{\theta_i}^T X}{{\theta_i}^T \theta_i W_i + W_i} \tag{32}$$

The complete method is described in detail in Algorithm 3.1.

4. Results and Evaluations

4.1. Dataset

A. Traffic Dataset

The Communication Commission of Guangzhou Municipality, China, published the Traffic dataset utilized in this study. This dataset includes information on vehicle speeds on 214 different stretches of road. Included are data on urban traffic for the 61 days beginning August 1, 2016 and ending September 30, 2016. Observation is supplied for each road segment with a daily speed value of 144, and the time frame for each observation is 10 minutes. This dataset is preferred in our study due to its relevance, official source, comprehensive and granular data, local insights, longitudinal analysis potential, and integration possibilities.

B. Wiki Dataset

The 2866 image-text pairings in this dataset were culled from featured articles on Wikipedia. Only 2173 of the possible image-text combinations are utilized during training, with the remaining pairs being used for testing. This information can be classified into 10 different labels. Labels include "art and architecture," "biology," "geography and locales," "history," "literature," "media," "music," "royalty and nobility," "sport and warfare," and "other." The 10-dim Linear Discriminant Analysis is used to extract text features, while the Convolutional Neural Network is used to extract picture features. This dataset was primarily chosen for its large size, diverse topics, credibility, natural language format, broad coverage, and multilingual support.

4.2. Data Privacy Preservation Evaluation

A. Traffic Speed Prediction

In much of the prior research, the missing scenario involves arbitrarily omitting elements from the original tensor. Allowing $\Omega = \{(i,j,k)|S_{ijk} \cdot x_{ijk} \neq 0\}, \Omega \subset \Omega_1$, and ℓ to be a tensor with random entries 0 and 1, $\mathcal{X} \Leftarrow \ell * \mathcal{X}$ can be derived. However, the missing information in this scenario (i.e., missing elements) may not reflect the missing information in the actual world; for instance, the speed data for a certain section of road on a certain day may be totally absent. Therefore, a particular fiber-like random missing is taken into consideration in the experiment; specifically, if given a matrix S$\in \mathbb{R}^{n_1 \times n_2}$ with missing values between 0 and 1, tensor $\mathcal{X} \Leftarrow \ell * \mathcal{X}$ can be calculated were $\ell(i,j,k) = S(i,j)$ for all $i \in \{1,\dots,n_1\}$, $j \in \{1,\dots,n_2\}$ and $k \in \{1,\dots,n_3\}$. These missing values are imputed as discussed in section 3.2 followed by the proposed privacy preservation algorithm. Thus, privacy preserved data is used to predict the speed at several time intervals. The mean absolute percent error (MAPE) and root mean square error is used to evaluate the prediction difference. The time horizon of 2, 4, 6, 12, 18, 24, 30, 36, 42, 48, 54 are used for the prediction. 40% missing data is imputed before privacy preservation. Figure 5 shows the comparison between the proposed privacy preserved data's prediction accuracy and without privacy preserved prediction. With the increase in time horizon, the prediction accuracy reduces. The DANMF privacy preserved data has higher prediction accuracy and lower MAPE values than without privacy preserved data prediction process. Though the RMSE for both cases are approximately same.

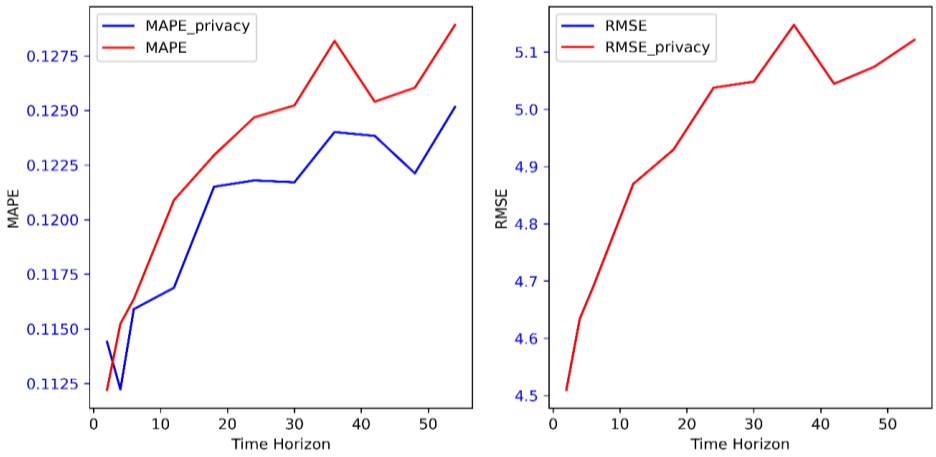

Fig.5. DANMF privacy preserved data and without privacy preserved traffic tensor data prediction comparison curve

B. Multimodal Data Analysis

For missing value creation, we used a random missing entries approach. We utilize rank 1 to complete the tensor data after having created missing values for 13%, 26%, and 52%. It is shown how the suggested decomposition solution with bias updating stacks up against the best existing methods. For randomly missing data, we evaluated imputation using Bayesian inference,[40]. Missing records in the Wiki dataset are also verified to ensure they are completely coincidental. Gradient descent (GD) and Adam optimization are also tested on the suggested method to find the best possible solution to equation 14. Mean Absolute Percent Error (MAPE) is used in the Statistical Analysis of Results. These indicators of performance are characterized by the assumptions that y_i and z_i are genuine missing values that may be estimated as:

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\frac{|y_i - z_i|}{y_i} \tag{33}$$

The suggested technique is assessed on the Wiki dataset and the traffic dataset, and the results are compared to those obtained using state-of-the-art methods such as BGCP [2], BCPF [22], BATF [38], which have a shared Bayesian Gaussian tensor factorization, and the modified PARAFAC versions. We use Bayesian optimization to fine-tune the CANDECOMP/PARAFAC weights. Table 2 compares the suggested data imputation strategy for traffic datasets with random missing items and it also shows that there is a noticeable improvement in performance in the case of a small

missing ratio, but this improvement is not sustained in the presence of a larger number of missing data. Superior performance is achieved as a result of the dynamic chaotic learning rate and the updated regularized tensor decomposition. The suggested approach is unaffected by variations in the missing rate.

Table 2. Analyzing the effects of randomly missing entries in a traffic dataset

		13% loss	26% loss	52% loss
Proposed Imputation		0.08265	0.08313	0.08304
Possible Scheme Variants	Modified PARAFAC-E-Biased	0.08313	0.08478	0.08914
	Modified PARAFAC-Adam optimization	0.08726	0.08926	0.09078
	Modified PARAFAC-GD Optimization	0.09204	0.09278	0.09291
Cutting-edge comparator schemes	BATF [38]	0.08252	0.08304	0.08413
	BGCP [2]	0.08213	0.08278	0.08313
	BCPF [22]	0.08326	0.08413	0.08526

For a given query of the Wiki dataset used, MAPE is computed for $R = 50$ results as well as for $R = all$, all these considered for without and then with privacy preservation to provide an overall picture of how well the suggested approach performs. We test the accuracy of both image-search and text-search operations. We estimate that in a multimodal searching framework, the user would be barely interested in more than 50 search results, hence we rank MAPE's results with all test outcomes as competitive. For this reason, we employ Linear Discriminant Analysis (LDA) to extract low-dimensional characteristics from the text. The paradigm of LDA is feature space dimension reduction. Data privacy preservation using DANMF is shown to reduce MAPE in fig. 5 by comparing non-privacy preserved data to privacy protected data. The results demonstrate that using the DANMF methodology for $R = 50$, image based searches encounters a higher MAPE index as compared to the text-based operations, while the converse is true for $R = all$. Further quite visible from the graph in fig. 6, using this methodology, ceteris paribus – the MAPE index is reduced privacy-preserved data as compared to the original forms.

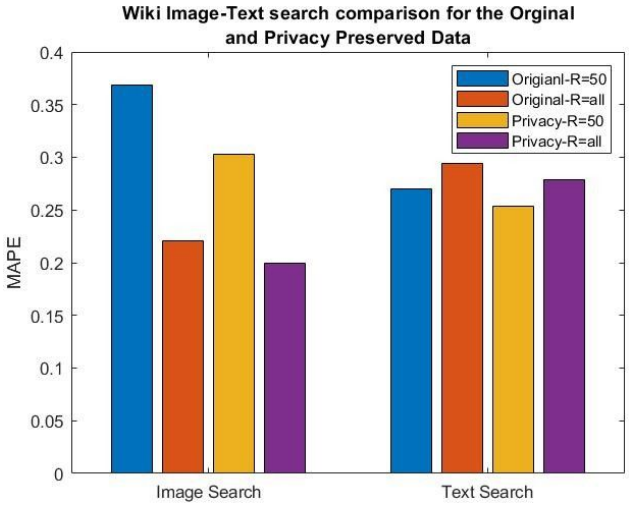

Fig.6. DANMF's comparison of data with and without privacy protections

5. Conclusions

Substantially correlated, fine-grained temporal and spatial data need an analytical approach that can grow as the data expands. Therefore, the foundation of privacy-preserving data mining is the provision of methods for modifying the source data in such a manner that sensitive information is secured even after mining is complete. The major goal of this research is to provide new insights to the relatively well-established study of data privacy protection. The primary contribution comes from two different components of the study: an imputation strategy to fill in the gaps of the unobserved data before learning the item profiles, and the optimization of the Deep Auto-encoded NMF at a variable learning rate. In this study, we used Bayesian inference to assess the efficacy of imputation for data that was missing at random for 13%, 26%, and 52%. For decomposition problems, the proposed solution with bias updating compares well to state-of-the-art approaches. The statistical analysis of results uses the Mean Absolute Percentage Error (MAPE) measure. The proposed approach is evaluated on the Wiki dataset and the traffic dataset by comparing the results to those produced by employing state-of-the-art techniques such as BATF, BGCP, BCPF, all of which share a Bayesian Gaussian tensor factorization, and

the modified PARAFAC versions. We consider MAPE to have competitive performance across the board. Results show that when utilizing the DANMF approach, for $R = 50$, image-based searches have a higher MAPE index than text-based operations, whereas the inverse is true when using $R = all$. In addition, everything else held unchanged, the MAPE index is less for data that has had its privacy protected via this approach than its respective original forms.

References

[1] Frigerio, L., Oliveira, A.S.D., Gomez, L. and Duverger, P., 2019, June. Differentially private generative adversarial networks for time series, continuous, and discrete open data. In IFIP International Conference on ICT Systems Security and Privacy Protection (pp. 151-164). Springer, Cham.

[2] Alsulaimawi, Z., 2020, September. A non-negative matrix factorization framework for privacy-preserving and federated learning. In 2020 IEEE 22nd International Workshop on Multimedia Signal Processing (MMSP) (pp. 1-6). IEEE.

[3] Yang, M., Zhu, T., Xiang, Y. and Zhou, W., 2018. Density-based location preservation for mobile crowdsensing with differential privacy. Ieee Access, 6, pp.14779-14789.

[4] Wang, J. and Zhang, J., 2007, May. Addressing accuracy issues in privacy preserving data mining through matrix factorization. In 2007 IEEE Intelligence and Security Informatics (pp. 217-220). IEEE.

[5] Wang, D., Wu, Y., Zhao, W. and Fu, L., 2019. A Model of Privacy Preserving in Dynamic Set-valued Data Re-publication. Journal of Internet Technology, 20(1), pp.147-156.

[6] Peng Liu, YuanXin Xu, Quan Jiang, Yuwei Tang, Yameng Guo, Li-e Wang,Xianxian Li, Local Differential Privacy for Social Network Publishing,Neurocomputing(2019), doi:https://doi.org/10.1016/j.neucom.2018.11.104

[7] Xu, C., Ren, J., Zhang, D., Zhang, Y., Qin, Z. and Ren, K., 2019. GANobfuscator: Mitigating information leakage under GAN via differential privacy. IEEE Transactions on Information Forensics and Security, 14(9), pp.2358-2371.

[8] Acs, G., Melis, L., Castelluccia, C. and De Cristofaro, E., 2018. Differentially private mixture of generative neural networks. IEEE Transactions on Knowledge and Data Engineering, 31(6), pp.1109-1121.

[9] Mohapatra, D. and Patra, M.R., 2019. Anonymization of attributed social graph using anatomy based clustering. Multimedia Tools and Applications, 78(18), pp.25455-25486.

[10] Li, T., Wang, Y., Ren, Y., Ren, Y., Qian, Q. and Gong, X., 2022. Nonnegative matrix factorization-based privacy-preserving collaborative filtering on cloud computing. Transactions on Emerging Telecommunications Technologies, 33(6), p.e3914.

[11] Malekzadeh, M., Clegg, R.G. and Haddadi, H., 2017. Replacement autoencoder: A privacy-preserving algorithm for sensory data analysis. arXiv preprint arXiv:1710.06564.

[12] Jordon, J., Yoon, J. and Van Der Schaar, M., 2018, September. PATE-GAN: Generating synthetic data with differential privacy guarantees. In International conference on learning representations.

[13] Andrew, J., Karthikeyan, J. and Jebastin, J., 2019, March. Privacy preserving big data publication on cloud using Mondrian anonymization techniques and deep neural networks. In 2019 5th international conference on advanced computing & communication systems (ICACCS) (pp. 722-727). IEEE.

[14] Li, Q., Wu, Z., Wen, Z. and He, B., 2020, April. Privacy-preserving gradient boosting decision trees. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 34, No. 01, pp. 784-791).

[15] Li, S., Shen, H., Sang, Y. and Tian, H., 2020. An efficient method for privacy-preserving trajectory data publishing based on data partitioning. The Journal of Supercomputing, 76(7), pp.5276-5300.

[16] Kanwal, T., Shaukat, S.A.A., Anjum, A., Choo, K.K.R., Khan, A., Ahmad, N., Ahmad, M. and Khan, S.U., 2019. Privacy-preserving model and generalization correlation attacks for 1: M data with multiple sensitive attributes. Information Sciences, 488, pp.238-256.

[17] Liu, C., Chen, S., Zhou, S., Guan, J. and Ma, Y., 2021. A general framework for privacy-preserving of data publication based on randomized response techniques. Information Systems, 96, p.101648.

[18] Wang, Z., Myles, P. and Tucker, A., 2019, June. Generating and evaluating synthetic UK primary care data: preserving data utility & patient privacy. In 2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS) (pp. 126-131). IEEE.

[19] Zhang, J., Zhao, B., Song, G., Ni, L. and Yu, J., 2019. Maximum delay anonymous clustering feature tree based privacy-preserving data publishing in social networks. Procedia Computer Science, 147, pp.643-646.

[20] Bousquet, O., Livni, R. and Moran, S., 2019. Passing tests without memorizing: Two models for fooling discriminators.

[21] Ye, F., Chen, C. and Zheng, Z., 2018, October. Deep autoencoder-like nonnegative matrix factorization for community detection. In Proceedings of the 27th ACM international conference on information and knowledge management (pp. 1393-1402).

[22] Papernot, N., Abadi, M., Erlingsson, U., Goodfellow, I. and Talwar, K., 2016. Semi-supervised knowledge transfer for deep learning from private training data. arXiv preprint arXiv:1610.05755.

[23] Cao, Y., Xiong, L., Yoshikawa, M., Xiao, Y. and Zhang, S., 2018, August. ConTPL: controlling temporal privacy leakage in differentially private continuous data release. In Proceedings of the VLDB Endowment. International Conference on Very Large Data Bases (Vol. 11, No. 12, p. 2090). NIH Public Access.

[24] Papernot, N., Song, S., Mironov, I., Raghunathan, A., Talwar, K. and Erlingsson, Ú., 2018. Scalable private learning with pate. arXiv preprint arXiv:1802.08908.

[25] Beaulieu-Jones, B.K., Wu, Z.S., Williams, C., Lee, R., Bhavnani, S.P., Byrd, J.B. and Greene, C.S., 2019. Privacy-preserving generative deep neural networks support clinical data sharing. Circulation: Cardiovascular Quality and Outcomes, 12(7), p.e005122.

[26] Torkzadehmahani, R., Kairouz, P. and Paten, B., 2019. Dp-cgan: Differentially private synthetic data and label generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (pp. 0-0).

[27] Chathurdara Sri Nadith Pathirage, Jun Li, Ling Li, Hong Hao, Wanquan Liu, Pinghe Ni, Structural damage identification based on autoencoder neural networks and deep learning, Engineering Structures, Volume 172, 2018, Pages 13-28, ISSN 0141-0296, https://doi.org/10.1016/j.engstruct.2018.05.109.

[28] H. Li, C. He, Y. Zheng, X. Fei, Z. Hu and Y. Tang, "Boosting Nonnegative Matrix Factorization Based Community Detection with Graph Attention Auto-Encoder," in IEEE Transactions on Big Data, vol. 8, no. 4, pp. 968-981, 1 Aug. 2022, doi: 10.1109/TBDATA.2021.3103213.
[29] Lim, K. and Wang, X., 2015, April. Nonnegative matrix factorization based privacy preservation in vehicular communication. In SoutheastCon 2015 (pp. 1-2). IEEE.
[30] Fontenla-Romero, O., Pérez-Sánchez, B. and Guijarro-Berdiñas, B., 2021. DSVD-autoencoder: a scalable distributed privacy-preserving method for one-class classification. International Journal of Intelligent Systems, 36(1), pp.177-199.
[31] Alguliyev, R.M., Aliguliyev, R.M. and Abdullayeva, F.J., 2019. Privacy-preserving deep learning algorithm for big personal data analysis. Journal of Industrial Information Integration, 15, pp.1-14.
[32] Abadi, M., Chu, A., Goodfellow, I., McMahan, H.B., Mironov, I., Talwar, K. and Zhang, L., 2016, October. Deep learning with differential privacy. In Proceedings of the 2016 ACM SIGSAC conference on computer and communications security (pp. 308-318).
[33] Jayapradha, J., & Prakash, M. (2022). f-Slip: an efficient privacy-preserving data publishing framework for 1: M microdata with multiple sensitive attributes. Soft Computing, 26(23), 13019-13036.
[34] Koren, Yehuda, Robert Bell, and Chris Volinsky. "Matrix factorization techniques for recommender systems." Computer 8 (2009): 30-37.
[35] Minh X. Hoang, Xuan-Hong Dang, Xiang Wu, Zhenyu Yan, Ambuj K. Singh, "GPOP: Scalable Group-level Popularity Prediction for Online Content in Social Networks", Proceedings of the 26th International Conference on World Wide Web, 2017, pp 725-733.
[36] Chen, Xinyu, Zhaocheng He, and Jiawei Wang. "Spatial-temporal traffic speed patterns discovery and incomplete data recovery via SVD-combined tensor decomposition." Transportation research part C: emerging technologies 86 (2018): 59-77.
[37] Paatero, Pentti. "Construction and analysis of degenerate PARAFAC models." Journal of Chemometrics: A Journal of the Chemometrics Society 14, no. 3 (2000): 285-299.
[38] Kingma DP, Adam BJ., "A method for stochastic optimization: arXiv preprint arXiv:1412.6980. 2015, pp 1-15.
[39] Lee, Daniel D.; Sebastian, Seung, H. (1999). "Learning the parts of objects by non-negative matrix factorization" (PDF). Nature. 401 (6755): 788–791. Bibcode:1999Natur. 401..788L. doi:10.1038/44565. PMID 10548103. S2CID 4428232.
[40] Chen, Xinyu, Zhaocheng He, Yixian Chen, Yuhuan Lu, and Jiawei Wang. "Missing traffic data imputation and pattern discovery with a Bayesian augmented tensor factorization model." Transportation Research Part C: Emerging Technologies 104 (2019): 66-77

Authors' Profiles

Pooja Choudhary is pursuing Ph.D. in Computer Science & Applications from Department of Computer Science & Applications, Kurukshetra University, Kurukshetra. She completed her MCA from Department of Computer Science & Applications, Kurukshetra University, Kurukshetra. Her research area is Privacy Preservation in Data Mining.

Kanwal Garg is an Assistant Professor at Department of Computer Science & Applications, Kurukshetra University, Kurukshetra. He holds an experience of 18 years. He received his Ph.D. from GJU Science & Technology, Hisar. His area of research includes big data, Data Mining and Warehousing, Web Mining, Data Stream and OLAP cubes. He has published about 80 papers in National and International Journals.

How to cite this paper: Pooja Choudhary, Kanwal Garg, "A Novel Privacy Preservation Scheme by Matrix Factorized Deep Autoencoder", International Journal of Computer Network and Information Security(IJCNIS), Vol.16, No.3, pp.84-98, 2024. DOI:10.5815/ijcnis.2024.03.07