Wellness Representation of Users in Social Media: Towards Joint Modelling of Heterogeneity and Temporality

Mohammad Akbari, Xia Hu, *Member, IEEE,* Fei Wang, *Senior Member, IEEE,* Tat-Seng Chua

Abstract—The increasing popularity of social media has encouraged health consumers to share, explore, and validate health and wellness information on social networks, which provide a rich repository of Patient Generated Wellness Data (PGWD). While data-driven healthcare has attracted a lot of attention from academia and industry for improving care delivery through personalized healthcare, limited research has been done on harvesting and utilizing PGWD available on social networks. Recently, representation learning has been widely used in many applications to learn low-dimensional embedding of users. However, existing approaches for representation learning are not directly applicable to PGWD due to its domain nature as characterized by longitudinality, incompleteness, and sparsity of observed data as well as heterogeneity of the patient population. To tackle these problems, we propose an approach which directly learns the embedding from longitudinal data of users, instead of vector-based representation. In particular, we simultaneously learn a low-dimensional latent space as well as the temporal evolution of users in the wellness space. The proposed method takes into account two types of wellness prior knowledge: (1) temporal progression of wellness attributes; and (2) heterogeneity of wellness attributes in the patient population. Our approach scales well to large datasets using parallel stochastic gradient descent. We conduct extensive experiments to evaluate our framework at tackling three major tasks in wellness domain: attribute prediction, success prediction and community detection. Experimental results on two real-world datasets demonstrate the ability of our approach in learning effective user representations.

Index Terms—Latent Space Learning, User's Wellness, Social Networks.

1 Introduction

With the advent of Web 2.0, we have witnessed the revolutionary changes in many disciplines brought by an explosion of user-generated contents; and health is of no exception [2], [7], [21], [44]. In such a context, millions of users increasingly utilize social networks such as Twitter and Instagram to share their wellness data and to fulfil their health demands. For example, diabetic patients not only share about events happening around them but also frequently post about their current health conditions, medication uses, and outcomes of medications. They frequently post the latest values of their blood glucose, diet, and exercises on Twitter using "*#diabetes*" and "*#BGnow*" hashtags [2], [21]. Effective mining of patient generated wellness data (PGWD) can provide actionable insights into the wellness of individuals as well as collaborative behaviour of communities. While data-driven approaches are increasingly used for personalized healthcare [11], [12], [22], [48], as an important and distinct data source, understanding PGWD available on social networks presents great opportunities to improve care delivery.

- *M. Akbari and T.-S. Chua are with School for Integrative Sciences and Engineering and School of Computing, National University of Singapore, Singapore 117417.*
  E-mail: akbari@u.nus.edu, chuats@comp.nus.edu.sg
- *X. Hu is with Texas A&M University.*
  E-mail: hu@cse.tamu.edu
- *F. Wang is with Department of Healthcare Policy and Research, Weill Cornell Medicine.*
  E-mail: few2001@med.cornell.edu

Manuscript received April 10, 2016; revised xx xx, xxxx.

Despite its value and significance, PGWD on social networks has not been fully utilized due to the following challenges. (1) *Longitudinality*. Wellness data are longitudinal per se, which means multiple measurements or repeated events are available for each subject [22], [48], [53]. For example, Hemoglobin A1c (HbA1c) test might be done several times per year for diabetic patients. The longitudinal nature of the problem provides a matrix of wellness data describing patient at different time points [40], [48], [53]. This is quite different from standard machine learning representation where we have a static vector of features. In such a context, time dimension plays an essential role. (2) *Noisiness and Incompleteness*. Social media is a highly varied and informal media; arising from various background and intention of users [39]. Moreover, missing data is an intrinsic nature of PGWD since patients do not persistently report their wellness data. In most cases, users are not sufficiently keen to expose the event or they self-censor the content due to privacy concerns [7], [21]. This means that the absence of a wellness event in PGWD does not always mean that the event did not happen [33]. (3) *Heterogeneity*. An intrinsic characteristic of the wellness domain is heterogeneity of the patient population according to their health conditions; meaning that wellness attributes and events related to each user can be highly different from the others [27]. For instance, even though diabetic users often share similar characteristics, they are still different from each other based on demographic attributes (e.g., age and gender), type of disease (e.g., Type I Diabetes, Type II Diabetes, Gestational Diabetes, etc.), and many other behavioral and

genetic factors. Even though patient stratification is a well-established approach in health informatics [42], this kind of disease-specific context has not been fully investigated in many wellness models such as re-admission prediction [12], disease progression modelling [43], [52], risk prediction [41]; and the assumption of a homogenous cohort does not hold in the population. How to share information among homogenous population while simultaneously avoid interactions between heterogeneous populations is still an open problem in wellness modelling.

Representation learning, also called latent feature learning, has been widely used as an effective tool for many machine learning and data mining tasks to derive an effective latent space from original data [17], [45], [49], [50]. The key idea of representation learning is to seek a low-dimensional embedding of data instances while preserving different discriminative factors of variation behind the data. Recently, factorization based methods have been attracting a lot of interests in modeling user behaviors and interests due to its ability to alleviate data sparsity [14], [17], [49], [50]. For example, MaxMF [45] is developed to represent each user with a set of latent factors representing his/her different latent interests. Zhao et al. [49] incorporated social connections into latent space to improve the performance of recommendation. Seen from the personalization aspect, Zhao et al. [50] proposed a personalized feature projection method that employs users' projection matrices and items' factors to solve the one-class recommendation problem. However, existing representation learning approaches have been designed for attribute-value data and cannot be directly applied to longitudinal data due to the following factors. First, many existing methods assume that data instances are fully observed and construct a model from original data treating missing values as zero, which is clearly violated in longitudinal data [35], [48], [53]. Second, traditional representation learning methods assume that the data instances are independent and identically distributed (*i.i.d*), which is clearly violated in longitudinal wellness data. Several decades of research in health science states that longitudinal wellness data are strongly related along temporal dimension and wellness attributes progress smoothly over time instead of sudden changes in consecutive time points [22], [48].

To deal with the challenges raised by the distinct PGWD, in this paper, we aim to learn wellness representation of users from social media. Our framework, in contrast to conventional models, determines the wellness latent space directly from users' longitudinal data, instead of attribute-value data, by considering two types of domain priors, namely the heterogeneity in data space and temporal contingency of wellness concepts. In particular, the proposed approach decomposes longitudinal data into two components: wellness latent space, and temporal representation of users. To effectively handle data heterogeneity, the learned wellness latent space is comprised of two sub-spaces, i.e., the shared and personalized latent spaces. The learned temporal representation is constrained to model the temporal progression of wellness attributes and simultaneously tackle the problems arising from missing data values. The proposed framework has been extensively examined through several machine learning tasks to evaluate its effectiveness in user embedding.

The main contributions of this paper are as follows:

- We propose a representation learning approach for longitudinal wellness data available in social networks. Specifically, we decompose longitudinal PGWD into wellness latent space and the temporal progression of users in that space.
- We exploit consistency within homogenous population as well as distinction between heterogeneous population to learn a shared and personalized latent space for embedding users.
- We incorporate the temporal progression prior of wellness data in the learning process to tackle the problems arising from missing and sparsity of data.

The remainder of this paper is organized as follows. In Section 2, we formally define the problem we study. We next model longitudinal representation learning as an optimization problem in Section 3 followed by the algorithmic details in Section 4. In Section 5, we report experimental results followed by related work in Section 6. Finally, we summarize the paper and outline the future work in Section 7.

2 Problem Statement

The task at hand is to embed users in a latent space with lower-dimensionality while preserving important discriminative features amongst users. Intuitively, we aim to learn a projection of users' longitudinal data into a sparse latent space. The learnt embedding demonstrates discriminative features of users which can be used to solve different machine learning tasks such as attribute prediction, success prediction and patient clustering. To learn an effective latent representation, we simultaneously incorporate prior knowledge, such as temporality of wellness features and heterogeneity of users, in the learning process. In this section, we first present the notations and then formally define the problem of representation learning of longitudinal data.

2.1 Problem Formulation

Notation. We use boldface uppercase letters (e.g., $\mathbf{A}$) to denote matrices, boldface lowercase letters (e.g., $\mathbf{a}$) to denote vectors, and lowercase letters (e.g., a) to denote scalars. The entry at the i-th row and j-th column of a matrix $\mathbf{A}$ is denoted as $\mathbf{A}_{(i,j)}$. $\mathbf{A}_{(i*)}$ and $\mathbf{A}_{(*j)}$ denote the i-th row and j-th column of a matrix $\mathbf{A}$, respectively. Meanwhile, subscript, i.e., $\mathbf{A}_i$ is used to denote the i-th item in the set of items $\mathcal{A}$. $\|\mathbf{A}\|_1$ is the ℓ_1-norm and $\|\mathbf{A}\|_F^2$ is the Frobenius norm of matrix $\mathbf{A}$. Specifically, $\|\mathbf{A}\|_1 = \sum_{i=1}^{m} \|\mathbf{A}_{(i*)}\|_1$ and $\|\mathbf{A}\|_F^2 = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} \mathbf{A}_{(i,j)}^2}$.

Let $\mathcal{U} = \{\mathbf{U}_1, \mathbf{U}_2, \dots \mathbf{U}_n\}$ denote a set of n users' longitudinal information. Each user longitudinal information $\mathbf{U}_i$ is denoted by $\mathbf{U}_i \in \mathbb{R}^{f \times t}$, where f is the number of different wellness events and features[1] and t is the length of observation window in which we measure the events. Note that the user's longitudinal data is a matrix where $\mathbf{U}_{i(j,k)}$

1. In this text, we use wellness feature (e.g., blood glucose, hypertension) and wellness events (onset of asthma attack, hyperglycemia) interchangeably.

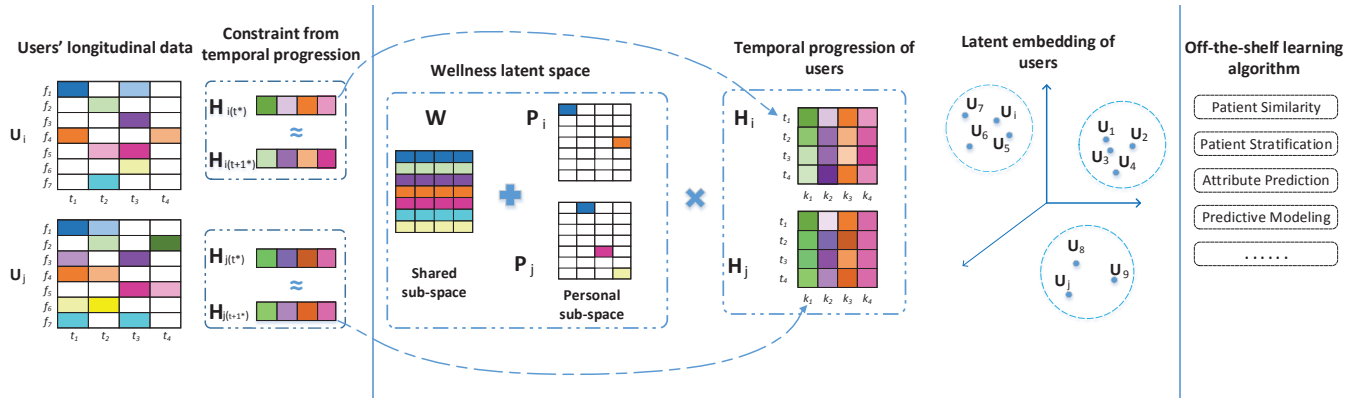


Fig. 1. The conceptual view of the proposed framework for representation learning of longitudinal data from social networks. The wellness latent space is comprised of two sub-spaces: shared and personal latent space. The final representation of each user, i.e., $\mathbf{H}_i$, embeds the user in the latent space while each row is his/her representation at one time point, where different colors show distinct features and color intensity shows relative weight of the feature.

represents the measurement value of wellness event j at time point k for user i.

We want to learn a low-rank representation of users in $\mathcal{U}$ so that if two users u and v have similar wellness data, their representation would be closer. We assume that the longitudinal data can be factorized to two components: a latent space representing wellness concepts and the temporal progression of each user in the latent space, as shown in Figure 1. The factorization process is capable of reconstructing the user data matrix on observed values. In general, a user's longitudinal representation is formally defined as a matrix $\mathbf{H}_i$, where each row of the matrix, i.e., $\mathbf{H}_{i(j*)}$, represents the user wellness state at time point j.

With the notation above, we formally define the longitudinal user representation problem as: *Given a set of users' longitudinal information $\mathcal{U}$, we aim to learn a model as follows,*

$$f : \mathcal{U} \rightarrow \{\mathbf{W}_i, \mathbf{H}_i\}, \tag{1}$$

which can compute ***wellness latent space*** $\mathbf{W}_i \in \mathbb{R}^{f \times k}$ *and* ***temporal progression*** *of each user in wellness latent space, i.e.,* $\mathbf{H}_i \in \mathbb{R}^{t \times k}$.

The final representation of each user, i.e., $\mathbf{H}_i$, precisely embeds the user in wellness latent space while each row is his/her representation at one time point.

3 Factorization of Longitudinal Data

As mentioned, PGWD includes two major aspects: wellness aspect and temporal aspect. Constructing an effective representation requires to subtly decompose these two components from each other. The key hypothesis behind longitudinal data factorization is that user's data matrix can be decomposed into two factors: (1) wellness latent space, and (2) the temporal onset of wellness events over the observation window, i.e., time dimension.

3.1 Preliminaries

Previous studies have shown that the wellness features can be projected to a latent space with a lower dimensionality; resulting in a dense representation of the original features [53]. This factorization process is capable of reconstructing the observed entries of original matrix, i.e., patient longitudinal wellness data. Inspired by these research findings, we utilized nonnegative matrix factorization (NMF) to decompose patient data matrix into two low rank matrices which are capable of approximately reconstructing the observed matrix. NMF is a matrix factorization algorithm that factorizes the non-negative data matrix into two positive matrices [18]. Assume that $\mathbf{U}_i \in \mathbb{R}^{f \times t}$ represents the data matrix for patient i, the aim of factorization is to decompose $\mathbf{U}_i$ into to non-negative matrices $\mathbf{W}_i \in \mathbb{R}^{f \times k}$ and $\mathbf{H}_i \in \mathbb{R}^{t \times k}$, whose product provide a good approximation of $\mathbf{U}_i$, i.e., $\mathbf{U}_i \approx \mathbf{W}_i \mathbf{H}_i^T$, where k is a pre-specified parameter denoting the dimension of reduced space. For instance, in topic modeling, k represents the number of topics while it denotes the number of desired latent dimensions in feature learning. Formally, NMF aims to minimize the following objective function,

$$\min_{\mathbf{W}_i, \mathbf{H}_i} \|\mathbf{U}_i - \mathbf{W}_i \mathbf{H}_i^T\|_F^2 \quad \text{s.t.} \quad \mathbf{W}_i \geq 0, \mathbf{H}_i \geq 0, \tag{2}$$

where $\mathbf{W}_i$ is called the *wellness basis matrix* and $\mathbf{H}_i$ is the *temporal progression matrix*. Intuitively, $\mathbf{H}_i$ represents how wellness dimensions evolve over time for the given user. In other words, it demonstrates how the user's wellness is going to improve, stable, or worsen as time passes. As the above objective function is not jointly convex in $\mathbf{W}_i$ and $\mathbf{H}_i$, finding the global minima is infeasible [18]. Therefore, alternating minimization is iteratively utilized to find a local minima. The iterative update rules are as follows,

$$\mathbf{W}_i \leftarrow \mathbf{W}_i \odot \frac{\mathbf{U}_i \mathbf{H}_i}{\mathbf{W}_i \mathbf{H}_i^T \mathbf{H}_i}, \quad \mathbf{H}_i \leftarrow \mathbf{H}_i \odot \frac{\mathbf{U}_i^T \mathbf{W}_i}{\mathbf{H}_i \mathbf{W}_i^T \mathbf{W}_i}. \tag{3}$$

where $\odot$ and the division symbol in this matrix context denote element-wise multiplication and division. Note that the above setting is different from standard matrix factorization where $\mathbf{U}_i$ represents an item-feature matrix constructed from the whole dataset.

3.2 Shared Wellness Space for Homogenous Cohort

Factorization of user's longitudinal data provides an intuitive decomposition of data matrix of a given user into wellness latent features and their temporal progression over time. However, decomposing wellness data of each user

in isolation may not provide effective representation due to the excessive sparsity in data. Besides, comparing latent spaces of different users would be a challenging task since the factorization process may extract diverse latent features fitted on each user data. Therefore, extracting a common latent space from the entire collection of data is normally preferred. The hypothesis behind collective latent space learning is that the wellness latent space extracted from different data instances, in our case users, should admit the same underlying structure, corresponding to higher-level latent features constructed from the combination of lower level features. At the same time, the temporal progression of these wellness latent features can vary from user to user depending on user's attributes, behaviors and so on. Mathematically, it can be formulated as the following objective function,

$$\min_{\mathbf{W},\mathbf{H}_i} J_{SLS} = \frac{1}{2n}\sum_{i=1}^{n}\|\mathbf{U}_i - \mathbf{W}\mathbf{H}_i^T\|_F^2 + \frac{\lambda_1}{2}(\|\mathbf{W}\|_F^2 + \frac{1}{n}\sum_{i=1}^{n}\|\mathbf{H}_i\|_F^2) \quad \text{s.t.} \quad \mathbf{W} \geq 0, \mathbf{H}_i \geq 0, \tag{4}$$

where the first term factorizes users' longitudinal data, where the second and third terms control the complexity of models. Here, $\mathbf{W}$ is to compute the shared wellness latent space among all patients.

The above objective function assumes that all patients share the same wellness space and learns a unique mapping $\mathbf{W}$ from the original feature space to wellness latent space. With sharing of the latent space among all patients, we indeed transfer knowledge among the patient cohorts, which is attractive especially when the available information for each patient is limited and the cohort is homogenous [29], [33]. Sharing also reduces the effect of noise since the latent space is derived from a large amount of data.

3.3 Personalized Wellness Space for Heterogeneous Cohort

Even though learning a common latent space from dataset is an intuitive and well-established tradition in machine learning, its performance is highly varied in real applications since it assumes a rigid consensus in dataset; i.e., all the data instances need to follow a specific latent space [29]. This is, however, impossible in real situations since patients can be divided into different cohorts with different characteristics. For example, diabetic users can be divided into three major patient groups: type I, type II, and gestational diabetics based only on disease type, where each group holds different characteristics [11], [27], [35]. This suggests that we need a personalized feature learning framework to deal with heterogeneity in data space.

Inspired by the notion of "dirty models" in machine learning for handling heterogeneous high-dimensional data [16], [17], we assume that individual's wellness latent space can be slightly deviated from the shared space extracted from the whole population. Mathematically, we consider the following learning model,

$$\min_{\mathbf{W},\mathbf{H}_i,\mathbf{P}_i} J_{PLS} = \frac{1}{2n}\sum_{i=1}^{n}\|\mathbf{U}_i - (\mathbf{W}+\mathbf{P}_i)\mathbf{H}_i^T\|_F^2 + \frac{\lambda_1}{2}(\|\mathbf{W}\|_F^2 + \frac{1}{n}\sum_{i=1}^{n}\|\mathbf{H}_i\|_F^2) + \frac{\lambda_2}{n}\sum_{i=1}^{n}\|\mathbf{P}_i\|_1 \quad \text{s.t.} \quad \mathbf{W} \geq 0, \mathbf{H}_i \geq 0, \mathbf{P}_i \geq 0, \tag{5}$$

where the latent space is estimated by the summation of two parameters $\mathbf{W}$ and $\mathbf{P}_i$. The first part of Eq. (5) learns three sets of parameters: (1) $\mathbf{W}$ is the shared latent space for all users inferred from the entire dataset; (2) $\mathbf{P}_i$ is to model heterogeneity in data space, i.e., the personalized feature space; and (3) $\mathbf{H}_i$ demonstrates the temporal evolution of each individual in the latent space. By imposing different regularizations for each parameter, we can fit an effective personalized learning model. The above formulation includes two set of regularizers; the second term, i.e., $(\|\mathbf{W}\|_F^2 + \frac{1}{n}\sum_{i=1}^{n}\|\mathbf{H}_i\|_F^2)$, controls the generalization performance of the model to avoid overfitting and the third term (ℓ_1-norm) leads to a sparse model. By imposing a ℓ_1 norm over $\mathbf{P}_i$, we indeed learn a sparse model for the personalized latent space, enforcing the wellness features of individuals slightly deviate from the shared features extracted from the whole population. It is worth noting that the aforementioned model extends the concept of dirty model to longitudinal data [16].

From clinical aspects, the proposed model is closely related to precision medicine [11], [24], where medical treatments are tailored to individual patients based on their detailed genetic and clinical profiles as well as lifestyle factors. By learning personalized latent space, i.e. $\mathbf{P}_i$, our model follows precision medicine paradigm through modeling distinct characteristics of individuals. Our model also considers disease principle paradigm by providing a computational model with the shared feature space, i.e., $\mathbf{W}$, where disease treatment and prevention are learned from the entire population. This also presents significance in treating patients with missing values.

3.4 Modeling Temporal Information

Recall that wellness attributes smoothly evolve over time. The temporal progression of wellness attributes suggests that these values gradually changes over time [22], [48]. Further, modelling the temporal evolution of wellness attributes can effectively reduce the noise and sparsity of the wellness data through imputation of missing values as pointed by [35], [48]. As each row of the temporal progression matrix $\mathbf{H}_{i(j*)}$ indicates the wellness representation of user i in time point j, we hence penalize the sudden changes of wellness attributes between neighbouring time points. Specifically, the temporal progression of wellness attributes can be mathematically modelled as,

$$\mathcal{R}_{temporal} = \frac{1}{2n}\sum_{i=1}^{n}\sum_{j=1}^{t-1}\|\mathbf{H}_{i(j*)} - \mathbf{H}_{i(j+1*)}\|^2, \tag{6}$$

where $\mathbf{H}_{i(j*)}$ denotes the wellness representation of user i at time point j. To facilitate the optimization of the temporal

progression term, Eq.(6) can be restated in an equivalent form as follows,

$$\begin{aligned}\mathcal{R}_{temporal} &= \frac{1}{2n}\sum_{i=1}^{n}\sum_{j=1}^{t-1}\|\mathbf{H}_{i(j*)} - \mathbf{H}_{i(j+1*)}\|^2 \\ &= \frac{1}{2n}\sum_{i=1}^{n}\|\mathbf{H}_i\mathbf{R}_i\|_F^2,\end{aligned} \tag{7}$$

where $\mathbf{R}_i \in \mathbb{R}^{t\times t-1}$ is the temporal smoothness indicator and is precalculated by the following definition,

$$\mathbf{R}_{i(j,k)} = \begin{cases} 1 & \text{if } j = k; \\ -1 & \text{if } j = k+1 \\ 0 & \text{otherwise .} \end{cases} \tag{8}$$

It is worth noting that temporal smoothness constraints have been also used in the problem of sound source separation to increase the robustness of separation [26], [37]. Inspired by [37], Eq.(7) constrains that the wellness representation of the given user at two consecutive time points to be close to each other.

4 Algorithm Details

The optimization framework, which integrates prior information into representation, is defined as follows,

$$J_{Space} + \alpha\mathcal{R}_{temporal}, \tag{9}$$

where the first term, i.e., J_{Space}, denotes the objective function for learning latent space, i.e. Eq.(4) and Eq.(5) for homogenous and heterogenous settings, respectively.

In this section, we introduce an efficient algorithm to solve the optimization problems and discuss its time complexity. Note that the optimization problem of homogenous setting is a special case of the heterogenous setting. Therefore, we only provide the algorithm for heterogenous setting. Here, by substituting Eq.(5) in the above equation, we have the following cost function,

$$\begin{aligned}\min_{\mathbf{W},\mathbf{H}_i,\mathbf{P}_i} \mathcal{O} = &\frac{1}{2n}\sum_{i=1}^{n}\|\mathbf{U}_i - (\mathbf{W}+\mathbf{P}_i)\mathbf{H}_i^T\|_F^2 \\ &+ \frac{\alpha}{2n}\sum_{i=1}^{n}\|\mathbf{H}_i\mathbf{R}_i\|_F^2 + \frac{\lambda_1}{2}(\|\mathbf{W}\|_F^2 + \frac{1}{n}\sum_{i=1}^{n}\|\mathbf{H}_i\|_F^2) \\ &+ \frac{\lambda_2}{n}\sum_{i=1}^{n}\|\mathbf{P}_i\|_1 \\ &\text{s.t.} \quad \mathbf{W}\geq 0, \mathbf{H}_i \geq 0, \mathbf{P}_i \geq 0,\end{aligned} \tag{10}$$

where α, λ_1, and λ_2 are regularizers to control the trade-off between different components.

4.1 Optimization Algorithm

We adopt an alternating optimization strategy to find the optimal values for model parameters. Specifically, we alternatively update $\mathbf{W}$, $\mathbf{H}_i$, and $\mathbf{P}_i$ to minimize the objective function while keeping the others fixed. To enforce the non-negativity constraints, we need to incorporate Lagrange multipliers. Let Λ_w, Λ_{pi}, and Λ_{hi} be the Lagrange matrices for constraints $\mathbf{W} \geq 0$, $\mathbf{P}_i \geq 0$, and $\mathbf{H}_i \geq 0$, respectively. The Lagrange $\mathcal{L}$ is:

$$\mathcal{L} = \mathcal{O} + Tr(\Lambda_w\mathbf{W}) + \sum_{i=1}^{n}(Tr(\Lambda_{pi}\mathbf{P}_i) + Tr(\Lambda_{hi}\mathbf{H}_i)). \tag{11}$$

4.1.1 Optimizing $\mathbf{W}$

By fixing $\mathbf{H}_i$ and $\mathbf{P}_i$, we can rewrite the objective function as follows,

$$\begin{aligned}\min_{\mathbf{W}}\mathcal{L} = &\frac{1}{2n}\sum_{i=1}^{n}\|\mathbf{U}_i - (\mathbf{W}+\mathbf{P}_i)\mathbf{H}_i^T\|_F^2 + \frac{\lambda_1}{2}\|\mathbf{W}\|_F^2 \\ &+ Tr(\Lambda_w\mathbf{W}) + C,\end{aligned} \tag{12}$$

where C is constant with respect to $\mathbf{W}$. Taking the derivative with respect to $\mathbf{W}$, we have,

$$\frac{\partial\mathcal{L}}{\partial\mathbf{W}} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{P}_i\mathbf{H}_i^T\mathbf{H}_i - \mathbf{U}_i\mathbf{H}_i) + \frac{1}{n}\sum_{i=1}^{n}\mathbf{W}\mathbf{H}_i^T\mathbf{H}_i + \lambda_1\mathbf{W} + \Lambda_w. \tag{13}$$

Using the Karush-Kuhn-Tucker (KKT) complementary condition that $\Lambda_{w(i,j)}\mathbf{W}_{(i,j)} = 0$, we have the following update rule for $\mathbf{W}$,

$$\mathbf{W} \leftarrow \mathbf{W} \odot \frac{\sum_{i=1}^{n}\mathbf{U}_i\mathbf{H}_i - \sum_{i=1}^{n}\mathbf{P}_i\mathbf{H}_i\mathbf{H}_i^T}{\sum_{i=1}^{n}\mathbf{W}\mathbf{H}_i^T\mathbf{H}_i + n\lambda_1\mathbf{W}}. \tag{14}$$

4.1.2 Optimizing $\mathbf{P}_i$

By ignoring terms that are independent of $\mathbf{P}_i$ in Eq.(11), the objective function boils down to:

$$\min_{\mathbf{P}_i}\mathcal{L} = \frac{1}{2n}\|\mathbf{U}_i - (\mathbf{W}+\mathbf{P}_i)\mathbf{H}_i^T\|_F^2 + \frac{\lambda_2}{n}\|\mathbf{P}_i\|_1 + Tr(\Lambda_{pi}\mathbf{P}_i). \tag{15}$$

The above objective function is non-smooth since it is the composition of a smooth term and a non-smooth term, i.e., ℓ_1 penalty, and gradient descent method is not available for solving the formulation. Inspired by [6], [25], we utilize the accelerated proximal method (APM) to solve its equivalent smooth reformulation. APM has been excessively utilized in data mining and machine learning communities [6], [15] due to its optimal convergence rate among all first-order techniques and its ability of dealing with large-scale non-smooth optimization problems. Note that in this paper, we focus on discussing the key concepts of APM, i.e, the proximal operator and its efficient computation; the detailed description of APM can be found in [25].

APM maintains two sequences of variables: a feasible solution sequence $\{\mathbf{P}_i^j\}$ and a searching point sequence $\{\mathbf{S}^j\}$, where the superscript, i.e., j, shows the index in the sequence. We denote the smooth and non-smooth part of the objective function $\mathcal{L}$ by $f(.)$ and $g(.)$. APM reformulates the optimization problem by a proximal operator which is formally defined as,

$$\mathbf{P}_i^{j+1} = \arg\min_{\mathbf{P}_i^j}\mathcal{M}_{\gamma^j,\mathbf{S}^j}(\mathbf{P}_i^j), \tag{16}$$

where,

$$\mathcal{M}_{\gamma^j,\mathbf{S}^j}(\mathbf{P}_i) = f(\mathbf{S}^j) + \langle\nabla f(\mathbf{S}^j), \mathbf{P}_i^j - \mathbf{S}^j\rangle + \frac{\gamma^j}{2}\|\mathbf{P}_i^j - \mathbf{S}^j\|_F^2, \tag{17}$$

where $\mathbf{S}^j$ is computed based on the past solutions by $\mathbf{S}^j = \mathbf{P}_i^j + \tau^j(\mathbf{P}_i^j - \mathbf{P}_i^{j-1})$ and $\nabla f(\mathbf{S}^j)$ denotes the derivatives of the smooth component $f(.)$ in the objective function, i.e., Eq.(15), at the search point $\mathbf{S}^j$. The parameter γ^j is the step size and is determined by line search according to Armijo-Goldstein rule. By ignoring terms that are independent of $\mathbf{P}_i^j$ the objective function boils down to:

$$\mathbf{P}_i^{j+1} = \arg\min_{\mathbf{P}_i^j} \|\mathbf{P}_i^j - \mathbf{Q}^j\|_F^2, \tag{18}$$

where $\mathbf{Q}^j = \mathbf{S}^j - \frac{1}{\gamma^j}\nabla f(\mathbf{S}^j)$ and indeed the solution of $\mathbf{P}_i^j$ is the Euclidian projection of $\mathbf{Q}^j$ onto convex set of constraints [25]. Here, $\nabla f(\mathbf{S}^j)$ denotes the gradient of the smooth component $f(.)$ in Eq.(15) at $\mathbf{S}^j$, which is defined as:

$$\nabla f(\mathbf{P}_i) = \frac{1}{n}(\mathbf{W}\mathbf{H}_i^T\mathbf{H}_i + \mathbf{P}_i\mathbf{H}_i^T\mathbf{H}_i - \mathbf{U}_i\mathbf{H}_i). \tag{19}$$

4.1.3 Optimizing $\mathbf{H}_i$

To minimize the cost function with respect to $\mathbf{H}_i$, we first fix $\mathbf{W}$ and $\mathbf{P}_i$, and then compute the derivative with respect to $\mathbf{H}_i$ as follows,

$$\frac{\partial \mathcal{L}}{\partial \mathbf{H}_i} = \frac{1}{n}[-\mathbf{U}_i\mathbf{P}_i - \mathbf{U}_i^T\mathbf{W} + \mathbf{H}_i(\mathbf{W}+\mathbf{P}_i)^T(\mathbf{W}+\mathbf{P}_i)] + (\frac{\lambda_1}{n}\mathbf{I} + \frac{\alpha}{n}\mathbf{R}_i\mathbf{R}_i^T)\mathbf{H}_i + \Lambda_{hi}, \tag{20}$$

where $\mathbf{I}$ denotes the identity matrix with correct dimensions. Using the Karush-Kuhn-Tucker (KKT) complementary condition that $\Lambda_{hi(m,n)}\mathbf{H}_{i(m,n)} = 0$, we have the following update rule for $\mathbf{H}_i$,

$$\mathbf{H}_i \leftarrow \mathbf{H}_i \odot \frac{\mathbf{U}_i^T\mathbf{P}_i + \mathbf{U}_i^T\mathbf{W}}{\mathbf{H}_i(\mathbf{W}+\mathbf{P}_i)^T(\mathbf{W}+\mathbf{P}_i) + (\lambda_1\mathbf{I} + \alpha\mathbf{R}_i\mathbf{R}_i^T)\mathbf{H}_i}. \tag{21}$$

4.2 Computational Complexity

We now analyze the time complexity of our learning framework using big O notation. The learning algorithm includes three main steps for optimizing three set of variables, i.e. $\mathbf{W}$, $\mathbf{P}_i$, and $\mathbf{H}_i$. In update rule for $\mathbf{W}$, the time complexity is $O(nkft)$, where n is the number of users, k is the dimension of latent space, f is the dimension of original feature space, and t is the length of the observation window. The main computational time for $\mathbf{P}_i$ is to compute the derivation of smooth part of objective function, i.e., Eq.(19), which is $O(ftk)$. As we need to update $\mathbf{P}_i$ for all samples, in our case each user, the total computational time is in order of $O(nkft)$. The computation for $\mathbf{H}_i$ is similar to $\mathbf{P}_i$ with time complexity of $O(nkft)$. If we need q iteration for updating the values of variables, the time complexity of the final algorithm is in order of $O(qnkft)$. As t denotes the length of observation window and it is in the size of few hundred, which is a small constant, in our experiment it is a six months period and $t = 25$, the final complexity can be approximated by $O(qnkft) \approx O(qnkf)$, making PLS a linear representation learning algorithm. We empirically verified this in our experiments, as the actual running time of our framework was similar to running plain NMF on all longitudinal data matrices.

TABLE 1
The list of seed hashtags and twitter support groups used for collecting twitter user pool.

Hashtags		Support Groups	
#Dibetes	#Bgnow	@AmDiabetesAssn	@WDD
#Diabetic	#T1D	@DiabeticConnect	@DiabetesUK
#type2diabetes	#T2D	@diabetesdaily	@NDEP
#diabeteschat	#Doc	@DiabetesMine	@citiesdiabetes
#LivingwithDiabetes	#Dblog	@DiabetesHealth	@diabeteshf

5 Experiments

In this section, we conduct extensive experiments to evaluate the effectiveness of the proposed representation learning of users from social networks in both homogenous and heterogeneous settings. We used our approach in two real-world datasets to accomplish different tasks, which show superiority of our proposed approach over the state-of-the-art baseline methods.

5.1 Experimental Settings

5.1.1 Datasets

Diabetes Dataset. We evaluated our approaches on a real-world dataset containing social postings of diabetic users about diabetes and their associated symptoms, medications, and activities. To construct the dataset, we first gathered a set of users who actively utilized diabetes related hashtags like "#diabetes" and "#bgnow" or follow diabetes support groups in Twitter microblogging service. Table 1 shows the list of hashtags and twitter support groups which were used for collecting candidate twitter users.

We next crawled the twitter profile of these users using Twitter API and selected the users who explicitly mention diabetes as an interest in their Twitter profile, resulting into $14,108$ different candidate user accounts. To construct ground truth labels, we utilized an automatic approach, inspired by similar efforts in computational social science [21], based on users who self-declared their disease information. We used expressions like "I am (Type—T) $(1|2)$ diabetic" to extract disease type for each user based on his/her profile information [2]. Disease type here refers to the major types of diabetes and includes three categories: Type I diabetes, Type II diabetes, and Others. We merged all the other non-common diabetes types as one category[3]. Table 2 shows the statistics of our dataset. As you can see, we could extract the health attributes of more than 50 percent of users ($7,474$ Twitter accounts) based on their self-declared information in their profiles, which we will use for the evaluation of our framework. Table 3 shows some example profiles from our collected dataset and their associated regular expression and ground truth labels[4].

To evaluate the reliability of the automatic annotation approach used in constructing the ground truth labels, we conducted a crowd-sourcing experiment in which we asked

2. We followed a bootstrapping approach similar to [36] to ensure the coverage and diversity of used patterns, where all extracted patterns are manually verified to ensure accuracy.

3. In our dataset, there are three non-common diabetes types: gestational diabetes, diabetes LADA (Type 1.5), and diabetes insipidus.

4. Due to user privacy concerns, some words/sentences may be different from original version.

TABLE 2
Statistics of the Diabetes Dataset

# of Users		14,108
# of Tweets		11,491,036
Disease Type	Diabetes Type I	4,194
	Diabetes Type II	2,477
	Others	803

TABLE 3
Sample profiles from our diabetes dataset

Husband. Dad. I've diagnosed as Type 1 diabetic since DATE. On a journey ...	I *diagnose* Type (1—2) diabetic	Type 1
I LOVE LIFE!! I am type 2 diabetic and take insulin	I * Type (1—2) diabetic	Type 2
Writer, avid reader, ...; live with T1 diabetes, ...	* with (T1—T2) diabetes	Type 1

three volunteers to annotate a randomly selected subset of the dataset containing $1,000$ users. The annotators were trained with short tutorials and a set of typical examples before the labelling process. We asked them to carefully read biographies of the users and annotate them based on disease type to three categories of type 1, type 2, and other diabetes. A majority voting scheme among the annotators was adopted to alleviate any ambiguity and inconsistency. The manually constructed ground truth was compared with its rule-based generated counterpart. The inter-agreement between annotators and rule-based approach was 0.882 with Cohen's Kappa metric, which demonstrates a substantial agreement between annotators and the rule-based approach.

BG Dataset. This is a public dataset of diabetic users who actively share their wellness information on Twitter. They not only post about their lifestyle information and activities such as their diet, activities, and emotional states but also share their health information in terms of medical events and measurements like their blood glucose value, HbA1c test results and hypoglycaemia/hyperglycaemia onset. This dataset was first used for extracting personal wellness events from social media posts of users [2]. Each user in the dataset has been labelled with a "successful", and "unsuccessful" tags showing that he managed to maintain an on-target blood glucose value for the pertaining week or failed to do so, respectively. We used this dataset to evaluate the effectiveness of our method in predicting the wellness states of users (such as the blood glucose value) based on the longitudinal wellness data of users on social media. This is important since wellness states are highly dependent on historical values, i.e. temporally dependent, showing that we need to consider longitudinal information of user's wellness instead of merely considering current state. Table 4 shows the statistics of this dataset.

5.1.2 Extraction of Longitudinal Wellness Descriptions

Feature extraction is an important aspect in our approach since it determines the original representation of data. To comprehensively represent user's wellness, inspired by studies in clinical text mining [2], [3], [47], we extracted three kinds of features as follows.

TABLE 4
Statistics of the BG dataset

# of Users	1,174
# of Tweets	1,060,105
# Successful Users	436
# Unsuccessful Users	738

1) ***RxNorm description***. Medication information is one of the most important types of wellness data. It is critical for healthcare safety and quality as well as for prognostic modeling [54]. To extract medication information, we utilized the approach proposed in [47] which utilizes semantic parser and domain knowledge to accurately extract medication information, i.e. medication names and signatures, from free texts and was commonly used as medication representation in literature. We utilized the widely-used tf-idf weighting scheme to construct feature vectors representing users.

2) ***UMLS description***. We also used a widely-used knowledge-based system called MetaMap to assign Unified Medical Language System (UMLS) Meta-thesaurus semantic concepts to user's social posts [3]. We collected all MetaMap's finding in the dataset and used their gold standard medical concepts as features. Along with the analogy of bag-of-words, we constructed a Bag-of-Concepts (BoC) in medical terminology and represent each user in the resulting space. The final BoC contains $5,370$ distinct concepts. Similar to RxNorm description, we utilized the tf-idf weighting scheme to compute the feature vectors of users.

3) ***Personal Wellness Events***. Personal wellness events are defined as events that are directly related to wellness of individuals; providing a summary of users' lifestyle and wellness such as diet event, medication use, and hospitalization [2]. Patients frequently post these events in their social accounts. We utilized the approach proposed in [2] to extract personal wellness events from users' published messages on Twitter. This will provide a high level description of user's wellness state; containing 14 distinct dimensions.

To construct the longitudinal wellness matrices of users, we utilized social media posts of users. We need to select a granularity level in time dimension and extract the information according to the selected granularity. We observed that the daily granularity is too sparse with more than 0.95% of users reluctant to report information daily. We thus constructed the users' longitudinal data at the weekly granularity. As we collected the data for six months, from May to October 2015, we constructed 25 time points for the entire period [5].

5.1.3 Evaluation Tasks and Metrics

To demonstrate the effectiveness of the proposed representation learning approach, we implicitly evaluated its performance in two commonly-used machine learning settings: supervised and unsupervised learning. The hypothesis behind implicit evaluation is that a good representation will improve the performance of the selected tasks as compared to other baselines.

5. We did not consider the first week of May and the last week of October because the data was partially crawled.

Attribute Prediction. Attribute detection was widely applied in user profiling to infer latent attributes of users such as age and gender prediction, education and occupation detection, and politic party detection [8], [10]. As inferring wellness attributes is a critical step in many downstream applications like recommendation [46], we hence proposed to predict wellness attributes of users using information from social media. We evaluated the performance of learning representation in predicting disease type which is a major wellness attribute of users. We utilized linear support vector machine (SVM) as a supervised classification approach, where in all experiments we used libSVM standard setting to be fair in comparison. To evaluate our approach, we utilized diabetes dataset with 10-fold cross validation and reported the performance in terms of precision, recall, and the area under the receiver operating characteristic curve (AUC). As the AUC metric is naturally defined for binary classification, we applied the one versus other classes setting and reported the average values. Due to the imbalance nature of the dataset, the average AUC provides a good explanation of the effectiveness of the proposed method [30].

Success Prediction. Success prediction is the task of predicting whether a specific user can successfully maintain his/her health indicators in a suggested range. For example, a diabetic patient who can successfully control his blood glucose value in the healthy range would be categorized as a successful patient, otherwise an unsuccessful patient. Due to its importance in wellness domain [44], we evaluated our feature learning framework in predicting users' success in managing their diabetes, i.e., maintaining their blood glucose value in the healthy range. Here, we considered the success prediction as a binary classification problem and utilized blood glucose dataset to evaluate our problem. Concretely, similar to the previous task, we train a linear SVM to predict the success of user in managing his blood glucose values.

Patient Clustering. We also evaluated our representation learning approach under the clustering task. Compared to classification, clustering is totally unsupervised and heavily relies on the learned features and similarity measure. Patient clustering is an inevitable need in healthcare domain as medical experts often need to deep dive on features discriminating patients subgroups. Patient stratification is a widely known approach in the wellness domain [34]. We adopted the commonly used cosine similarity for clustering of users in the learned latent space. We compared the performance of different approaches in terms of accuracy and normalized mutual information (NMI) on diabetes dataset. Overall, this experiment is to verify the robustness of the proposed approach in unsupervised machine learning setting.

5.1.4 Learning with Longitudinal Data

Standard machine learning techniques and baselines work on vector data, where we have a feature vector for each data point. However, our model learns longitudinal representation for users (i.e., $\mathbf{P}_i$ and $\mathbf{H}_i$). Hence, inspired by [53], we transformed the latent representation of users into a feature vector. To extract the feature vectors from learned representations, we derived features by averaging the latent features along the time dimension within a given observation window (25 weeks). In other words, for each feature, we averaged distinct values of the feature in different time points, resulting in a single value. Therefore, the size of final feature vectors would be equal to the number of latent dimensions extracted. To be fair in comparison, we used similar experimental setting for all baselines as mentioned in previous section.

TABLE 5
Performance of attribute and success prediction

	All	LapScore	Spec	NDFS	SLS	PLS
Disease Type Prediction						
Prec	42.31	44.71	41.50	46.32	53.02	59.34
Recall	42.66	46.11	44.82	43.71	48.21	54.20
AUC	63.05	64.47	62.35	67.33	69.85	72.15
Success Prediction						
Prec	62.21	67.34	64.08	68.82	71.33	74.12
Recall	67.45	66.72	64.31	65.01	68.20	68.75
AUC	64.10	61.20	61.40	68.95	72.21	76.80

5.2 On Performance Comparison

To the best of our knowledge, we are the first to study feature learning of the longitudinal data in social media. To demonstrate the effectiveness of representation learning approaches, we compared our learned features with those of other state-of-the-art unsupervised feature learning methods, while keeping the classification and clustering scheme fixed. We compared the following baseline methods:

- **ALL**. All original features are adopted for each user.
- **LapScore**. Laplacian score evaluates feature importance by its ability to preserve the local manifold structure of data [13].
- **Spec**. Features are selected by spectral analysis. This approach can be considered as an extension of Laplacian score method [51].
- **NDFS**. Nonnegative discriminate unsupervised feature selection via joint nonnegative spectral analysis and $\ell_{2,1}$-norm regularization [19].
- **Shared Latent Space (SLS)**. Users are embedded into shared latent space of Eq.(4).
- **Personal Latent Space (PLS)**. Each user's is represented using personalized latent space learned from Eq.(5); modelling both temporality and heterogeneity.

We followed previous research studies to tune the parameters for all baseline methods [13], [19]. The neighborhood size has been fixed to 5 for **LapScore** and **NDFS**, as suggested to be the best in [13], [19]. There are some regularization parameters for **NDFS**, and **LapScore**, which were set based on the experiments from the original papers. **SLS**, and **PLS** have three different regularizer parameters α, λ_1, and λ_2. In the experiments, we empirically set $K = 200$, $\alpha = 0.1$, $\lambda_1 = 10$, and $\lambda_2 = 0.4$ using grid search and reported the performance of the models using 10-fold cross validation. More details about the effects of these parameters on the proposed framework will be discussed in Sections 5.3 and 5.4.

We evaluated the predictive performance of the proposed framework in supervised setting using attribute prediction and success prediction experiments. The performance of attribute prediction and success prediction is

TABLE 6
Performance of users clustering

	All	LapScore	Spec	NDFS	SLS	PLS
ACC	51.32	56.10	52.84	54.88	56.11	58.01
NMI	0.0224	0.0227	0.0233	0.0240	0.0272	0.0287

presented in Table 5 in terms of precision, recall, and AUC. From the Table, we can observe the following points: (1) Feature selection is important as well as effective. The selected features can not only reduce the computational time of the algorithm [51] but more importantly can improve the final prediction performance, where all the feature learning approaches outperform **ALL** baseline. (2) **LapScore** and **Spec** have a neck to neck performance with a slight improvement by **LapScore** which is consistent with the results reported in past research efforts [19], [51]. (3) **NDFS** often outperforms both **LapScore** and **Spec** which is attributed to the feature selection process in **NDFS**. **LapScore** and **Spec** analyze features individually which may overlook possible correlation between distinct features, as reported in [19], while **NDFS** considers feature correlation. (4) **SLS** and **PLS** consistently outperform other baseline methods on both tasks. For example, **PLS** approximately gained up to 6% and 3% relative improvement in terms of precision in attribute prediction and success prediction, respectively. The reason is probably because **SLS** and **PLS** takes advantages of temporal correlation between feature values to mitigate problems arising from data sparsity and missing values. However, all baseline methods assume the *i.i.d* assumption, which is not valid in the wellness domain [48]. Moreover, **PLS** outperforms **SLS** most of the time, which shows the importance of modeling heterogeneity in data space, as reported by past efforts [17], [22]. Overall, these observations support the fact that joint learning features and modeling domain prior knowledge would achieve the best performance [22], [35].

We also evaluated our method under unsupervised setting, i.e., clustering. Table 6 summarizes the result of clustering users in learned latent space in terms of accuracy and NMI. The results are similar to that for supervised setting, i.e., classification. (1) **SLS** and **PLS** approaches outperform all the baseline methods in terms of accuracy and NMI, which demonstrates the importance of modeling temporal progression of wellness features as well as feature learning. The reason is probably because vector-based representation cannot capture the context around each user due to excessive sparsity of data, noisy information in social media, and inability to model temporal evolution of user. (2) **PLS** can effectively improve the performance with relative improvement of 2% over **SLS**, in terms of accuracy. This improvement is attributed to the effectiveness of modeling heterogeneity of patient populations, i.e., different subpopulations in patients, which is modeled in **PLS** while **SLS** assumes a homogeneous cohort of patients. Overall, the proposed method of *joint modeling temporality of wellness features and heterogeneity of user space* can outperform other baselines and achieve the state-of-the-art performance. This result is consistent with several past research in multi-feature machine learning where dirty models are used to model heterogeneity in samples [22], [35].

TABLE 7
Effectiveness evaluation of each involved component in our proposed models(df: degree of freedom).

	Precision	Recall	P-value	t-stat	df
PLS	74.12	68.75	-	-	-
SLS	71.33	68.20	3.4e-3	-3.47	9
PLS-noTP	64.02	58.91	1.7e-3	-3.39	9
SLS-noTP	62.37	56.09	2.6e-4	-5.23	9

5.3 On the Effect of Temporal Information

We are now interested in figuring out the effectiveness of different components in our proposed model. In particular, we compared the performance of incorporating temporal smoothness of wellness features in our model. i.e., $\mathcal{R}_{temporal}$. We hence conducted experiments to comparatively validate the following experimental settings:

- **PLS**. Our proposed framework which models both heterogeneity and temporality, i.e., Eq.(5).
- **SLS**. Our proposed framework which models temporality with homogenous assumption, i.e., Eq.(4).
- **PLS-noTP**. A variant of **PLS** without considering the temporal smoothness by setting $\alpha = 0$.
- **SLS-noTP**. A variant of **SLS** without considering the temporal smoothness by setting $\alpha = 0$.

We only reported the results for the success prediction task since similar observations have been made for the other tasks. The results of component-wise analysis are reported in Table 7. From the table, the following observations can be made: (1) **SLS-noTP** achieves the worst results. This can be explained by the fact that it neither models the temporal smoothness in wellness features, nor considers the heterogeneity in the patient population. These results highlight the importance of joint modeling the temporality of wellness features and heterogeneity of the patient population. (2) **SLS** and **PLS** consistently outperforms their counterparts **SLS-noTP** and **PLS-noTP**, which significantly supports the importance of modeling temporality of wellness features. This result has also been reported in modeling disease progression based on patient's EHR [48], [54]. (3) **PLS** is superior to the others; demonstrating that all components in our proposed model is indispensable.

It is worth noting that we also conducted a significance test based on the precision of success perdition task. In particular, we performed paired t-test between our **PLS** model and the other baseline methods based on 10-fold cross validation and the results shows that the improvements of our proposed model are statistically significant (p-values are smaller than 0.01).

5.4 On Parameter Sensitivity

We also studied the parameter sensitivity of our proposed method. Our model holds two sets of parameters: (1) the latent space dimension, i.e., k; and (2) the regularizers α, λ_1, and λ_2 in Eq.(5). We first evaluated the sensitivity of the proposed approach to the dimension of the latent space and then examined the effects of other parameters in combination with latent space to see how the parameters affect the learned latent space. We only performed parameter study for attribute prediction and clustering tasks to save space.

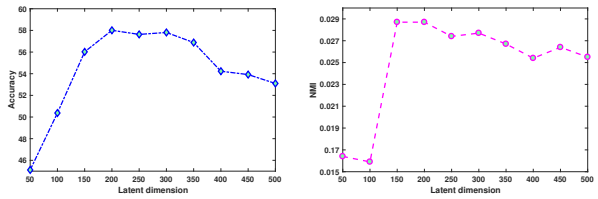


Fig. 2. Effect of latent space dimension on the performance of patient clustering task. Small values of latent dimension result into limited discrimination power, a large values yield overfitting.

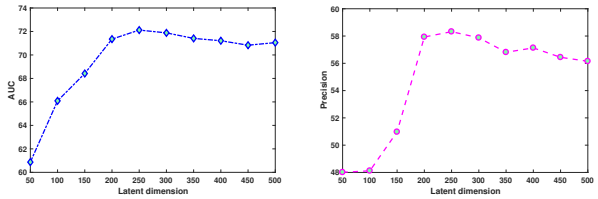


Fig. 3. Effect of latent space dimension on attribute prediction task. Small values of latent dimension result into limited discrimination power, a large values yield overfitting.

We first varied the dimension of latent space k in the range of $\{50, 100, 150, \ldots, 500\}$ while fixing the other parameters, i.e., α, λ_1, and λ_2. Figure 2 illustrates the clustering performance in terms of accuracy and NMI. The clustering performance is the best when the number of latent dimensions is around 200. Figure 3 shows the performance of attribute prediction in terms of AUC and precision. Similarly, the prediction performance first increases, reaches its peak and then gradually decreases. The results show that when the number of latent dimensions is too small, the model is unable to find a good representation. In contrast, a large latent dimension tends to overfit the data which leads to loss in performance. It is worth noting that how to determine the number of features is still an open problem in data mining [19].

To assess the effect of parameter λ_1 which controls the complexity of the model, we varied λ_1 as $\{0.001, 0.01, \ldots, 100\}$ while fixing λ_2, and α. Figure 4(a) and Figure 5(a) show the sensitivity of our framework with respect to various values of λ_1, and k for clustering and attribute prediction tasks, respectively. As shown in the figures, with the increase of λ_1, the performance rises rapidly and then keeps stable between the range of 1 to 10. A high value of λ_1 controls the effects of noise; making the model more robust. The results also demonstrate that the performance is more sensitive to the number of latent dimensions than λ_1.

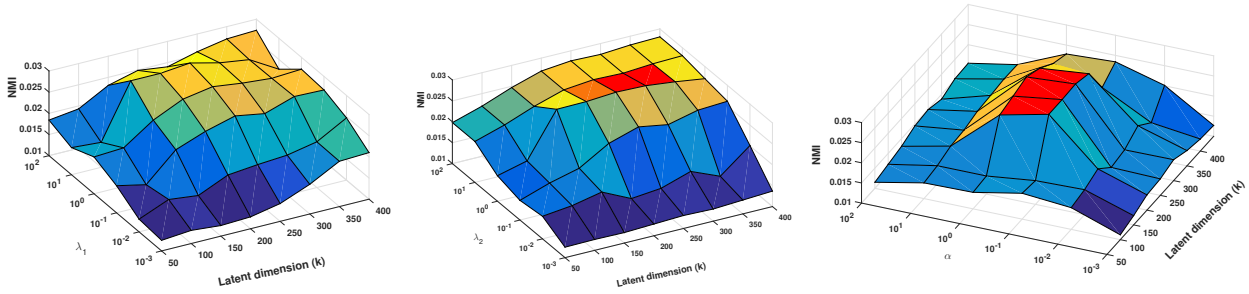
We also studied the effect of parameter λ_2 which controls the personalization aspects of feature learning; making the model more robust in heterogeneous data. Similarly, we changed λ_2 in the range of $\{0.001, 0.01, \ldots, 100\}$ while making the other parameters fixed. The results are shown in Figure 4(b) and Figure 5(b) for clustering and attribute prediction tasks, respectively. It can be seen that the performance of our model significantly improved when λ_2 varies between 1 and 10, verifying that modeling heterogeneity in the patient population is vital in wellness domain.

We finally investigated the trade-off between temporal smoothness of wellness features and latent space dimension by varying α in $\{0.001, 0.01, \ldots, 100\}$ as presented in Figure 4(c) and Figure 5(c). As shown in the Figures, in most cases, the performance first increases, reaches its peak and then gradually decreases. The best performance was achieved when α is around 0.1. These observations suggest the importance of modeling both temporal smoothness of wellness features as well as feature learning.

5.5 On Explanation of the Latent Space

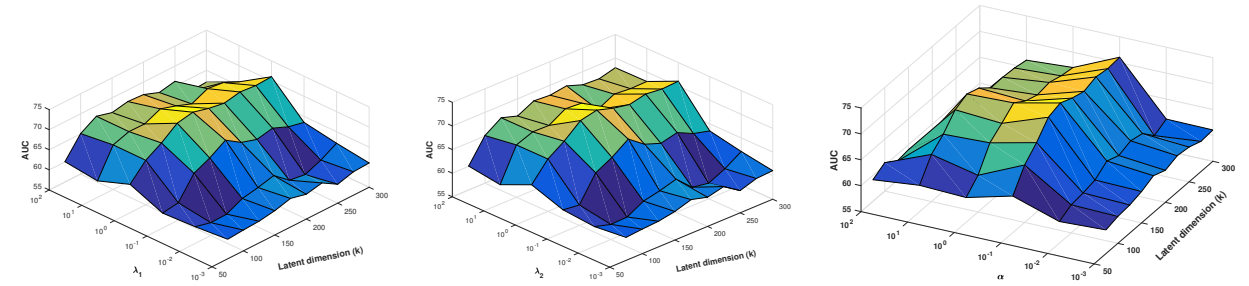
While macro-level quantitative evaluation is useful, it is also instructive to examine the actual results to better understand the latent space learnt by the proposed model. To accomplish this, in this section, we examine the latent features learnt from the data. This provides us a tangible understanding of the output of the proposed model. In the SLS model, W is the latent space with low-dimensionality, while H_i demonstrates the progression of i-th user in the latent space. Thus, we are able to represent the learnt embedding by mapping them into the original features. To accomplish this, we first normalized the weights of the columns in W such that the sum of each column is equal to 1. We then ranked the features for each latent dimension according to their normalized weights and found representative original features for each latent dimension. Table 8 shows several examples of latent dimensions learnt with the list of highly ranked features for each of the latent dimension. For the sake of readability, we manually named each row with a representative wellness concept. The results demonstrate that the correlated and relevant features are grouped together to form a latent dimension. For example the first dimension shows a group of features related to diabetes medication, e.g. Insulin (0.188), Injection (0.185), and Novolog (0.137). Similarly, the third dimension contains features related to the symptoms and comorbidities of diabetes in which Heart disease (0.141) and Cardiovascular (0.131) are top features. Overall, this results verify that the latent dimensions reveals different wellness aspect of users from their longitudinal social media data.

To have a deeper understanding of the effect of latent representation on the defined tasks, we also inspected the top parameters for the prediction of Diabetes Type II class. Our analysis revealed that the first two latent dimensions in Table 8 are prominent features for the prediction of Diabetes Type II. This is indeed reasonable as the top features, i.e. terms, in "Medication" latent dimension are related to Type II Diabetes treatments. Similarly, important features in the second latent dimension, i.e. "Symptoms and Medication of Type II", are also deeply relevant to Diabetes Type II. It is worth noting that inspection of the tweets pertaining to this class also demonstrates features relating to Diabetes Type II. Two sample tweets in this class are "Doctor today have to redo my insulin pump numbers to high. right side of back still swollen another month of muscle relaxers" and "#Bloodsugars 5.5 mmol/L Dinner 50g cabs, 475 kcal. 7 units of #NovoRapid". Overall, the results validated that learning a good latent representation contributes to the per-



(a) (b) (c)

Fig. 4. The effect of different regularizers on the performance of clustering task. Overall, latent dimension is an important factor in learning good representation. Besides, finding the best values for hyperparameters results into learning an effective latent space.



(a) (b) (c)

Fig. 5. The effect of different regularizers on the performance of attribute prediction task. Overall, latent dimension is an important factor in learning good representation. Besides, finding the best values for hyperparameters results into learning an effective latent space.

TABLE 8
Prominent feature for the Shared Latent Space and their weights.

ID	Name	Top features (weight)
01	Medication	Diabetes(0.201), Glucose(0.191), Insulin(0.188), Injection(0.185), Novolog(0.137), Humalog(0.012), Hypoglycemia(0.012), Victoza(0.012)
02	Symptoms and Medication of Type II	Metformin(0.221), Sugar(0.151), Glucophase(0.146), mg(0.087), Weight loss(0.041), Diarrhea(0.034), Pain(0.034), Actos(0.027)
03	Comorbidities	Heart disease(0.141), Cardiovascular(0.131), Surgery(0.129), Diabetes(0.121), Hypertension(0.052), Ischemic(0.019), Pain(0.018), Respiratory(0.018)

formance of the prediction task. It is worth noting that 39% of tweets hold features corresponding to RxNorm, which indicates that patients often discuss their medication online. Besides, 26% and 16.5% of the collected tweets are relevant to UMLS and Personal Wellness Events, respectively. This is attributed to the fact that patients leverage the power of social media to receive and provide social support from peers with similar wellness conditions [1], [2].

6 Related Work

Representation learning, or latent feature learning, is a popular approach for discovering low-dimensional structure from high dimensional data. We are interested in factorization based models which aim to find a low rank decomposition of original space approximately recovering the original space including sparse coding, Singular Value Decomposition (SVD), Principal Component Analysis (PCA), Weighted Matrix Factorization (WMF), and so on [5], [20], [38]. Recently, latent factor decomposition has been attracting much interest to alleviate data sparsity in recommendation task where user-item matrix is used to model user interests and intentions [4], [14], [17], [49], [50]. For example, Cai et al. [4] proposed a graph regularized NMF (GNMF) approach which employs the geometrical information of data space in factorization process. Similarly, semi-supervised GNMF (SGNMF) incorporates label information into the graph construction [23]. NMF has also been applied onto multi-view data, where a shared latent factor is inferred from different views [14], [29], [33]. For instance, joint NMF has been applied to multi-view clustering of Web 2.0 items by decoupling the learnt latent factors inferred from different views [14].

Personalization of latent factor modelling was first explored in [32], where a joint personal and social latent factor (PSLF) has been utilized for social recommendation. Similarly, Pan et al. [28] aggregated the features of a group of related users to reduce the uncertainty of the selected training instances. Zhao et al. [49] leveraged social connections to improve the performance of one-class recommendation. Lately, they proposed a personalized feature projection method that employs users' projection matrices and items' factors to solve one-class recommendation problem [50]. Similarly, Rendle et al. [31] proposed a next-basket recom-

mendation approach based on personalized Markov chains over sequential set data. By introducing personalization over the transition graph of each user, they employed both advantages of Markov chain and Matrix Factorization.

Multivariate time series has also applied for the task of supervised learning in healthcare domain. For instance, multivariate time-series with multi-task Gaussian process was used for predicting the severity of illnesses in ICU [9]. Similarly, several multi-task learning has also been utilized for progression modeling of chronic diseases. The existing approaches in this domain can be group to three main categories of: mono-modal mono-task learning, multi-task learning, and multimodal analysis. For example Nie et al. [26] proposed a regression model, named adaptive multimodal multi-task learning (aM2L), to predict the progression of chronic diseases such as Alzheimer's Disease. Their proposed model incorporated three types of prior knowledge into the learning process: 1) modality agreement; 2) adaptive modality weighting; and 3) temporal progression.

Most of the existing approaches for latent factor learning have been designed for vector-based representation to embed users (or items) in a low dimensional space. They will fail to provide effective representation if applied to longitudinal wellness data. Furthermore, existing feature learning techniques assume that data items are *i.i.d.*, which is clearly violated in longitudinal data. Moreover, most of these approaches fail to model heterogeneity in data space or model temporal dependency as a regularized multi-task learning framework but overlook heterogeneity in data space. Our aim is to learn a latent representation directly from longitudinal data where temporality and heterogeneity of data are jointly modeled.

In the area of data-driven health care, phenotyping has been applied to Electronic Health Records (EHRs) to predict the onset of congestive heart failure (CHF) and end stage renal disease (ESRD) by learning a general model [53]. Our framework, however, is different from their approach since we simultaneously model the shared latent space between homogenous populations to transfer knowledge among homogenous population as well as learn personalized latent space for each user to learn individual-based features. Their framework either considers a shared space or an individual latent space, which can be considered as a special case of our formulation, i.e., **SLS**. Similarly, Wang et al. [42] proposed a clustering-based approach to model the heterogeneity in the patient population, where the shared latent space is learnt for each group of users. It is worth noting that multi-task learning paradigm was also used for investigating EHRs, where they mostly assume the task are homogenous and learn task models simultaneously [27], [54].

7 Conclusions and Future Work

In this paper, we introduced a novel representation learning approach for longitudinal wellness data. The proposed method jointly models the temporal progression of wellness attributes as well as the heterogeneity in the patient populations. In particular, we factorized user's longitudinal data into two components, namely, the latent space representation and user temporal evolution in the space. The latent space is comprised of two sub-spaces: shared latent space and personalized latent space, which permits to exploit both consistency within homogenous cohorts as well as difference amongst heterogeneous cohorts to share an effective representation. Extensive experiments on two real-world datasets and different learning tasks in wellness domain verified the potential ability of the proposed framework in learning a good user embedding.

This study demonstrates the importance of feature learning approaches intrinsically designed for longitudinal data. Different extensions of this work are currently being investigated. The first is to utilize the social context around users in a collaborative learning approach. As social media users are linked to each other, incorporation of network-centric information is a promising direction. Moreover, users mostly utilize multiple social networks, thus integration of user descriptions from multiple social networks would be a promising research direction. The application of learned space in different wellness problems can be another promising direction.

Acknowledgments

The work is, in part, supported by NSF IIS-1657196. The work of Fei Wang is partially supported by NSF IIS-1650723. The views and conclusions contained in this paper are those of the authors and should not be interpreted as representing any funding agencies.

References

[1] M. Akbari and T.-S. Chua. Leveraging behavioral factorization and prior knowledge for community discovery and profiling. In *ACM WSDM*, 2017.
[2] M. Akbari, X. Hu, L. Nie, and T.-S. Chua. From tweets to wellness: Wellness event detection from twitter streams. In *AAAI*, 2016.
[3] A. R. Aronson. Effective mapping of biomedical text to the umls metathesaurus: the metamap program. In *AMIA Symposium*, 2001.
[4] D. Cai, X. He, J. Han, and T. S. Huang. Graph regularized nonnegative matrix factorization for data representation. *PAMI*, 2011.
[5] C. Chen, D. Li, Y. Zhao, Q. Lv, and L. Shang. Wemarec: Accurate and scalable recommendation through weighted and ensemble matrix approximation. In *SIGIR*, 2015.
[6] X. Chen, W. Pan, J. T. Kwok, and J. G. Carbonell. Accelerated gradient method for multi-task sparse learning problem. In *ICDM*, 2009.
[7] M. De Choudhury, M. R. Morris, and R. W. White. Seeking and sharing health information online: comparing search engines and social media. In *SIGCHI*, 2014.
[8] A. Farseev, L. Nie, M. Akbari, and T.-S. Chua. Harvesting multiple sources for user profile learning: a big data study. In *ICMR*, 2015.
[9] M. Ghassemi, M. A. Pimentel, T. Naumann, T. Brennan, D. A. Clifton, P. Szolovits, and M. Feng. A multivariate timeseries modeling approach to severity of illness assessment and forecasting in icu with sparse, heterogeneous clinical data. In *AAAI*, 2015.
[10] S. Gottipati, M. Qiu, L. Yang, F. Zhu, and J. Jiang. Predicting user's political party using ideological stances. In *Social Informatics*. 2013.
[11] L. Groop. Genetics and neonatal diabetes: towards precision medicine. *The Lancet*, 2015.
[12] D. He, S. C. Mathews, A. N. Kalloo, and S. Hutfless. Mining high-dimensional administrative claims data to predict early hospital readmissions. *JAMIA*, 2014.
[13] X. He, D. Cai, and P. Niyogi. Laplacian score for feature selection. In *NIPS*, 2005.
[14] X. He, M.-Y. Kan, P. Xie, and X. Chen. Comment-based multi-view clustering of web 2.0 items. In *WWW*, 2014.
[15] X. Hu, L. Tang, J. Tang, and H. Liu. Exploiting social relations for sentiment analysis in microblogging. In *WSDM*, 2013.
[16] A. Jalali, S. Sanghavi, C. Ruan, and P. K. Ravikumar. A dirty model for multi-task learning. In *NIPS*, 2010.

[17] X. Jin, F. Zhuang, S. J. Pan, C. Du, P. Luo, and Q. He. Heterogeneous multi-task semantic feature learning for classification. In *CIKM*, 2015.
[18] D. D. Lee and H. S. Seung. Algorithms for non-negative matrix factorization. In *NIPS*, 2001.
[19] Z. Li, Y. Yang, J. Liu, X. Zhou, and H. Lu. Unsupervised feature selection using nonnegative spectral analysis. In *AAAI*, 2012.
[20] C. Lin, J.-M. Yang, R. Cai, X.-J. Wang, and W. Wang. Simultaneously modeling semantics and structure of threaded discussions: a sparse coding approach and its applications. In *SIGIR*, 2009.
[21] H. Lin, J. Jia, Q. Guo, Y. Xue, Q. Li, J. Huang, L. Cai, and L. Feng. User-level psychological stress detection from social media using deep neural network. In *ACM MM*, 2014.
[22] C. Liu, F. Wang, J. Hu, and H. Xiong. Temporal phenotyping from longitudinal electronic health records: A graph based framework. In *SIGKDD*, 2015.
[23] H. Liu, Z. Wu, X. Li, D. Cai, and T. S. Huang. Constrained nonnegative matrix factorization for image representation. *PAMI*, 2012.
[24] R. Mirnezami, J. Nicholson, and A. Darzi. Preparing for precision medicine. *NEJM*, 2012.
[25] Y. Nesterov. Introductory lectures on convex optimization: a basic course, 2004.
[26] L. Nie, L. Zhang, Y. Yang, M. Wang, R. Hong, and T.-S. Chua. Beyond doctors: Future health prediction from multimedia and multimodal observations. In *ACM MM*, 2015.
[27] N. Nori, H. Kashima, K. Yamashita, H. Ikai, and Y. Imanaka. Simultaneous modeling of multiple diseases for mortality prediction in acute hospital care. In *SIGKDD*, 2015.
[28] W. Pan and L. Chen. Gbpr: Group preference based bayesian personalized ranking for one-class collaborative filtering. In *IJCAI*, 2013.
[29] Y. Pan, T. Yao, T. Mei, H. Li, C.-W. Ngo, and Y. Rui. Click-through-based cross-view learning for image search. In *SIGIR*, 2014.
[30] D. M. Powers. Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation. 2011.
[31] S. Rendle, C. Freudenthaler, and L. Schmidt-Thieme. Factorizing personalized markov chains for next-basket recommendation. In *www*, 2010.
[32] Y. Shen and R. Jin. Learning personal+ social latent factor model for social recommendation. In *KDD*. ACM, 2012.
[33] X. Song, L. Nie, L. Zhang, M. Akbari, and T.-S. Chua. Multiple social network learning and its application in volunteerism tendency prediction. In *SIGIR*, 2015.
[34] J. Sun, F. Wang, J. Hu, and S. Edabollahi. Supervised patient similarity measure of heterogeneous patient records. *SIGKDD*, 2012.
[35] Z. Sun, F. Wang, and J. Hu. Linkage: An approach for comprehensive risk prediction for care management. In *SIGKDD*, 2015.
[36] M. Thelen and E. Riloff. A bootstrapping method for learning semantic lexicons using extraction pattern contexts. In *ACL*, 2002.
[37] T. Virtanen. Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria. *IEEE transactions on audio, speech, and language processing*, 2007.
[38] M. N. Volkovs and G. W. Yu. Effective latent models for binary feedback in recommender systems. In *SIGIR*, 2015.
[39] C. Wang, R. Raina, D. Fong, D. Zhou, J. Han, and G. Badros. Learning relevance from heterogeneous social network and its application in online targeting. In *SIGIR*, 2011.
[40] F. Wang, N. Lee, J. Hu, J. Sun, S. Ebadollahi, and A. F. Laine. A framework for mining signatures from event sequences and its applications in healthcare data. *PAMI*, 2013.
[41] F. Wang, P. Zhang, B. Qian, X. Wang, and I. Davidson. Clinical risk prediction with multilinear sparse logistic regression. In *SIGKDD*, 2014.
[42] F. Wang, J. Zhou, and J. Hu. Densitytransfer: A data driven approach for imputing electronic health records. In *ICPR*, 2014.
[43] X. Wang, D. Sontag, and F. Wang. Unsupervised learning of disease progression models. In *SIGKDD*, 2014.
[44] I. Weber and P. Achananuparp. Insights from machine-learned diet success prediction. In *PSB*, 2015.
[45] J. Weston, R. J. Weiss, and H. Yee. Nonlinear latent factorization by embedding multiple user interests. In *RecSys*, 2013.
[46] C. Wing and H. Yang. Fityou: integrating health profiles to real-time contextual suggestion. In *SIGIR*, 2014.
[47] H. Xu, S. P. Stenner, S. Doan, K. B. Johnson, L. R. Waitman, and J. C. Denny. Medex: a medication information extraction system for clinical narratives. *JAMIA*, 2010.
[48] T. Xu, J. Sun, and J. Bi. Longitudinal lasso: Jointly learning features and temporal contingency for outcome prediction. In *CIKM*, 2015.
[49] T. Zhao, J. McAuley, and I. King. Leveraging social connections to improve personalized ranking for collaborative filtering. In *CIKM*, 2014.
[50] T. Zhao, J. McAuley, and I. King. Improving latent factor models via personalized feature projection for one class recommendation. In *CIKM*, 2015.
[51] Z. Zhao and H. Liu. Spectral feature selection for supervised and unsupervised learning. In *ICML*, 2007.
[52] J. Zhou, J. Liu, V. A. Narayan, J. Ye, et al. Modeling disease progression via multi-task learning. *NeuroImage*, 2013.
[53] J. Zhou, F. Wang, J. Hu, and J. Ye. From micro to macro: data driven phenotyping by densification of longitudinal electronic medical records. In *SIGKDD*, 2014.
[54] L. Zhou, G. B. Melton, S. Parsons, and G. Hripcsak. A temporal constraint structure for extracting temporal information from clinical narrative. *Journal of biomedical informatics*, 2006.

Mohammad Akbari received the PhD degree in the NUS Graduate School for Integrative Sciences and Engineering, National University of Singapore. His research interests spans mainly in data mining and machine learning using large-scale datasets, with an emphasis on their applications in health informatics and social informatics. His research has been published in several major academic venues, including SIGIR, WSDM, ICMR, etc.

Xia Hu received the BS and MS degrees in computer science from Beihang University, China and the PhD degree in computer science and engineering from Arizona State University. He is currently an assistant professor at the Department of Computer Science and Engineering, Texas A&M University. His research interests are in data mining, social network analysis, machine learning, etc. As a result of his research work, he has published nearly 40 papers in several major academic venues, including WWW, SIGIR, KDD, WSDM, IJCAI, AAAI, CIKM, SDM, etc. One of his papers was selected in the Best Paper Shortlist in WSDM13. He is the recipient of the 2014 ASU's Presidents Award for Innovation, and Faculty Emeriti Fellowship. He has served on program committees for several major conferences such as IJCAI, SDM and ICWSM, and reviewed for multiple journals, including IEEE Transactions on Knowledge and Data Engineering, ACM Transactions on Information Systems and Neurocomputing. His research attracts wide range of external government and industry sponsors, including US National Science Foundation (NSF), ONR, AFOSR, Yahoo!, and Microsoft. He is a member of the IEEE.

Fei Wang is an Assistant Professor in Division of Health Informatics, Department of Healthcare Policy and Research, Weill Cornell Medicine, Cornell University. He got his PhD from Department of Automation, Tsinghua University in 2008. His major research interest is data mining, machine learning and their applications in health informatics. He has published around 190 papers on the top venues of related areas. His papers have received over 4,800 citations so far with an H-index 37. His (or his students') papers won best paper runner-up for ICDM 2016, best student paper for ICDM 2015, best research paper nomination for ICDM 2010, Marco Romani Best paper nomination in AMIA TBI 2014, and his paper was selected into the best paper finalist in SDM 2011 and 2015. He also won the Parkinson's Progression Markers' Initiative data challenge organized by Michael J. Fox Foundation. Dr. Wang is the vice chair of the KDD working group in AMIA. Dr. Wang is the track chair for Medinfo 2017 and program co-chair for ICHI 2015. Dr. Wang is an action editor of the journal Data Mining and Knowledge Discovery, an associate editor of Journal of Health Informatics Research and Smart Health, and an editorial board member of Pattern Recognition and International Journal of Big Data and Analytics in Healthcare.

Tat-Seng Chua is the KITHCT Chair Professor at the School of Computing, National University of Singapore. He was the Acting and Founding Dean of the School during 1998-2000. Dr Chua's main research interest is in multimedia information retrieval and social media analysis. He is the Director of a multi-million-dollar joint Center (named NExT) between NUS and Tsinghua University in China to develop technologies for live media search. The project will gather, mine, search and organize user-generated contents within the cities of Beijing and Singapore. Dr Chua is active in the international research community. He is the winner of the prestigious 2015 ACM SIGMM award for Outstanding Technical Contributions to Multimedia Computing, Communications and Applications. He is the Chair of steering committee of ACM International Conference on Multimedia Retrieval (ICMR) and Multimedia Modeling (MMM) conference series. Dr Chua was also the General Co-Chair of ACM Multimedia 2005, ACM CIVR (now ACM ICMR) 2005, ACM SIGIR 2008, and ACM Web Science 2015. He serves in the editorial boards of 4 international journals.