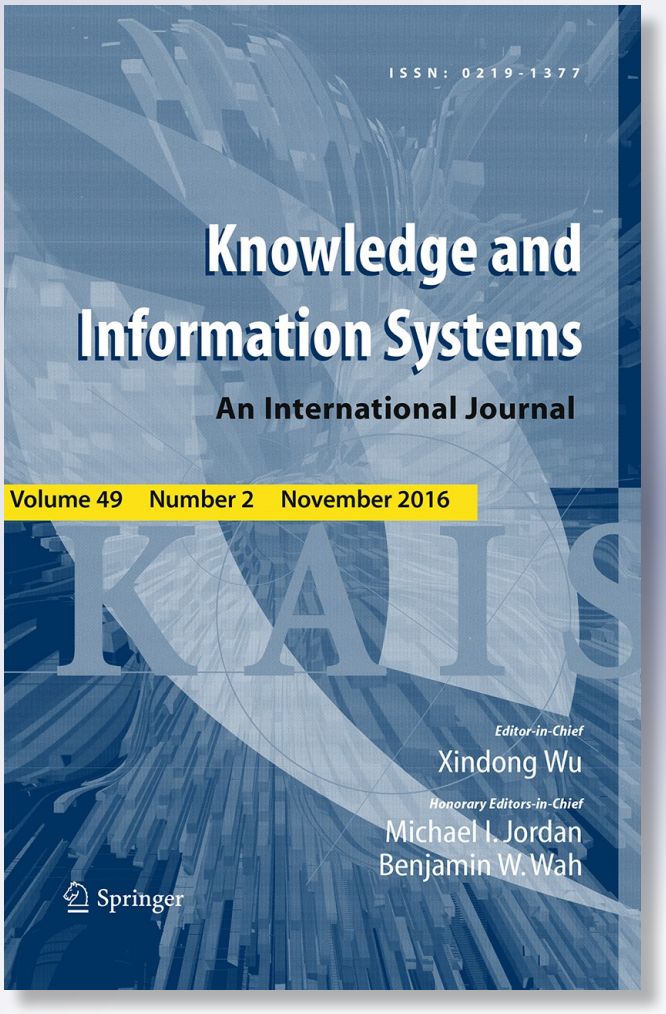
Dynamic socialized Gaussian process models for human behavior prediction in a health social network

Yelong Shen, NhatHai Phan, Xiao Xiao, Ruoming Jin, Junfeng Sun, Brigitte Piniewski, David Kil & Dejing Dou

Knowledge and Information Systems
An International Journal

ISSN 0219-1377
Volume 49
Number 2

Knowl Inf Syst (2016) 49:455-479
DOI 10.1007/s10115-015-0910-z



Springer

Knowl Inf Syst (2016) 49:455–479
DOI 10.1007/s10115-015-0910-z

REGULAR PAPER

Dynamic socialized Gaussian process models for human behavior prediction in a health social network

Yelong Shen[1] · NhatHai Phan[2] · Xiao Xiao[2] · Ruoming Jin[1] · Junfeng Sun[3] · Brigitte Piniewski[4] · David Kil[5] · Dejing Dou[2]

Received: 21 November 2014 / Revised: 21 October 2015 / Accepted: 14 December 2015 /
Published online: 31 December 2015

Abstract Modeling and predicting human behaviors, such as the level and intensity of physical activity, is a key to preventing the cascade of obesity and helping spread healthy behaviors in a social network. In our conference paper, we have developed a social influence model, named socialized Gaussian process (SGP), for socialized human behavior modeling. Instead of explicitly modeling social influence as individuals' behaviors influenced by their friends' previous behaviors, SGP models the dynamic social correlation as the result of social influence. The SGP model naturally incorporates personal behavior factor and social correlation factor (i.e., the homophily principle: Friends tend to perform similar behaviors) into a unified model. And it models the social influence factor (i.e., an individual's behavior can be affected by his/her friends) implicitly in dynamic social correlation schemes. The detailed experimental evaluation has shown the SGP model achieves better prediction accuracy compared with most of baseline methods. However, a Socialized Random Forest model may perform better at the beginning compared with the SGP model. One of the main reasons is the dynamic social correlation function is purely based on the users' sequential behaviors without considering other physical activity-related features. To address this issue, we further propose a novel "*multi-feature SGP model*" (mfSGP) which improves the SGP model by using multiple physical activity-related features in the dynamic social correlation learning. Extensive experimental results illustrate that the mfSGP model clearly outperforms all other models in terms of prediction accuracy and running time.

Yelong Shen and NhatHai Phan are co-first authors.

✉ Dejing Dou
dou@cs.uoregon.edu

1 Computer Science Department, Kent State University, Kent, OH, USA

2 Computer and Information Science, University of Oregon, Eugene, OR 97403, USA

3 National Institutes of Health, Bethesda, MD, USA

4 PeaceHealth Laboratories, Springfield, OR, USA

5 HealthMantic Inc, Austin, TX, USA

Keywords Socialized Gaussian process · Dynamic social correlation · Health social network

1 Introduction

The prevalence of obesity and overweight is increasingly becoming a major public health problem, within the USA and throughout the world. Recent studies have shown obesity can spread over the social network [5], bearing similarity to the diffusion of innovation [8] and word-of-mouth effects in marketing [11]. Though the common belief is that the spreading comes from the social and cultural influence of poor health habits such as lack of exercise and fast food consumption, there has been a lack of scientific and quantitative study to elucidate how social relationships may contribute to macro-level human behaviors.

Recent advances in mobile technology and online social networks provide new opportunities to support healthy behaviors through lifestyle monitoring and online communities. Mobile devices can track and record the walking/jogging/running distance and intensity of individuals; online social networks can help people to interact and participate in various physical activities and exercise. Utilizing this technology [17], the recent YesiWell study, conducted in 2010–2011 as a collaboration among several health laboratories and universities, recorded the social network and daily physical activities for a group of 254 individuals for a period of 10 months. The fundamental problems this study seeks to answer, which are also the key in understanding the determinants of human behaviors, are as follows:

1. *Could social affiliations affect individual behaviors?*
2. *How can we leverage social networks to help predict individuals' behaviors?*

1.1 Challenges

It is nontrivial to determine how much impact social influence could have on one's behavior. Human behaviors are constantly changing, for a variety of reasons. These can include the following factors:

- *Personal factor*: e.g., one person did not do any exercise on every Thursday this term because he took lots of courses on Thursdays.
- *Social correlation factor*: e.g., individuals tend to do sports together with their friends, e.g., playing basketball together, rather than playing it alone.
- *Social influence factor*: e.g., one person does not like sports at the beginning, but he or she will try to play basketball, because most of his or her friends play everyday.
- *Random external factor*: e.g., external factors and events that happen suddenly can make people behave randomly and unpredictably.

There have been rather extensive studies in social influence, which can be largely classified into two categories:

- *Contagion-based social influence model*: This model assumes that an individual (node) would follow his or her friends' past behavior. More precisely, it assumes that the behavior, itself, has certain infectiousness to some degree. Contagion-based social influence has been studied extensively in virtual marketing [4], information spreading [23] etc. Another feature of contagion-based social influence is that once an individual (node) is infected, its state will not be changed (nor influenced by others).
- *Latent state-based social influence model*: This model extracts the latent state for each node and then builds a unified factor graph model to capture the dynamics of the latent

state for each node. That is to say, the individual's state at some future time would be conditionally dependent on his or her friends' previous states [21,22].

Obviously, the contagion-based social influence model cannot tackle the problem in our case, because an individual's physical activities would dynamically change over time. The latent state-based social influence model fits our problem, but there are some technical challenges. First, it is hard to determine for how long the friends' past behaviors could affect his/her future behavior. Second, the latent state-based social influence model often assumes that influence factors caused by different friends are independent (factors factorization would help reduce the model complexity). However, it might not be true in practice. For example, if a person is deciding whether to play basketball or not, and only one of his friends suggests that he play basketball, the influence would be rather weak. If three or four of his friends suggest that he play basketball, it could significantly affect his decision. Third, theories in sociology and psychology [7] show that social influence could lead people to engage behaviors similar to their friends. Social influence could also be considered as one source of social correlation. Therefore, if a social influence model could not lead to the increasing of social correlation, it would be rather confusing. However, the latent state-based social influence model is not guarantee to increase social correlation.

1.2 Our contribution

In our conference paper [18], we proposed a social influence model, named socialized Gaussian process (SGP), for socialized human behavior modeling. In the SGP model, we have devised an implicit social influence approach. Instead of explicitly modeling social influence as an individual's behavior influenced by his friends' previous behaviors, we model the dynamic social correlation which we assume as a result of social influence. The proposed SGP model naturally incorporates personal behavior factor and social correlation factor into a unified model, and models the social influence factor implicitly in dynamic social correlation schemes. The detailed experimental evaluation has shown the SGP model achieves better prediction accuracy compared with most of baseline methods. However, a Socialized Random Forest [3] model may perform better at the beginning compared with the SGP model. One of the main reasons is the dynamic social correlation function is purely based on the users' sequential behaviors without considering other physical activity-related features. To address this issue, in this paper, we further propose a novel "*multi-feature SGP model*" (mfSGP). Compared with the SGP model, the dynamic social correlation schemes have been improved by leveraging multiple personal behavior-related features. In addition, we further develop an intuitive method to capture the dynamic of social correlation, and we devise an online prediction/inference scheme for the mfSGP model to predict individuals' future behaviors by learning from historical records, which include users' behaviors and other related features, and from social network information. The purpose of the online prediction evaluation is not to build a perfect predictor of human behaviors. Rather, we aim to evaluate whether the modeling assumptions of mfSGP are reasonable and to what degree the social affiliations could affect an individual's behaviors.

To sum, there are three main contributions in the paper:

– In the SGP and mfSGP models, human personal factor and social correlation factor are incorporated into a unified socialized Gaussian process model, in which the observed human behaviors can be interpreted from two aspects: Individuals tend to follow their own behavior patterns (i.e., personal factor), and individuals would like to be coordinated with their friends (i.e., social correlation factor).

– We propose a novel dynamic social correlation scheme to implicitly model the social influence (i.e., social influence factor). We develop an intuitive method to capture the dynamics of social correlation. In the mfSGP model, multiple personal behavior-related features are integrated into our social correlation scheme.
– Experimental results show that the novel mfSGP model achieves the best performance, i.e., the highest prediction accuracy, compared to baseline methods and the SGP model, which further demonstrates the advantage of the new model.

The rest of this paper is organized as follows. In Sect. 2, we define the problem of social human behavior prediction and formally introduce the personal, social correlation, and social influence factors. In Sect. 3, we introduce the socialized Gaussian process models for human behavior modeling. In Sect. 4, we perform a detailed experimental evaluation. In Sect. 5, we review the related work and provide further discussion. We conclude our work in Sect. 6.

2 Problem definition and socialized human behavior prediction

In this study, we consider two major data sources: (1) The social network $\mathcal{G} = (V, E)$, where $(i, j) \in E$ indicates users u_i and u_j are friends. Here, we consider that the friend relationship is mutual and thus $\mathcal{G}$ is an undirected graph. (2) Time series $\mathcal{X}$ based on users' behaviors, where user u_i's sequential behavior is denoted as $\mathcal{X}^i = (x_1^i, x_2^i, \ldots x_T^i)$ with $x_t^i \in (-1, 0, 1)$. Note, $x_t^i = 1$ indicates user u_i does sports at day t, $x_t^i = -1$ indicates user u_i does not do sports at day t, while $x_t^i = 0$ indicates user u_i's record is missing at day t. Then, we defined the *socialized human behavior prediction problem* as follows:

Given the social network $\mathcal{G}$ and individuals' past behaviors until day t, $\mathcal{X}_{1\ldots t} = (X_{1\ldots t}^1, X_{1\ldots t}^2, \ldots, X_{1\ldots t}^N)$, where $X_{1\ldots t}^i = (x_1^i, x_2^i, \ldots x_t^i)$ and N is the number of users in the social network, socialized human behavior prediction problem is to predict the individual's behaviors at day $t + 1$, i.e., $\mathcal{X}_{t+1}$.

Figure 1 illustrates the snapshots of user behaviors and social network at different points in time.

As illustrated in the figure, an individual's future behaviors can be predicted from three aspects. First, **Personal behavior pattern:** personal behavior pattern is computed based on individuals' past behaviors, i.e., calculating the autocorrelation function. Since most individuals have their own regular behavior cycles, an individual's historical behavior records can be leveraged for predicting his or her future behaviors. Second, **Social Correlation:** social correlation indicates that individuals tend to perform the same behaviors as their friends.

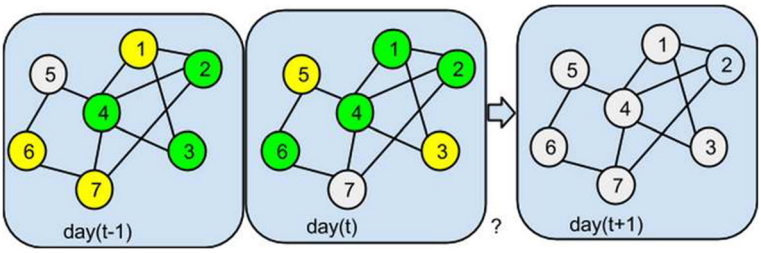


Fig. 1 Snapshots of user behaviors and social network. *Green circles* indicate time points during which the user exercises. *Yellow circles* indicate time points during which the user does not exercise. *White circles* indicate that the user's behavior is unknown (color figure online)

Third, **Social Influence:** social influence indicates that an individual's behavior can also be influenced by his or her friends' past behaviors. In general, the socialized human behavior prediction problem could be formulated as:

$$(X_t^1, X_t^2, \ldots, X_t^N) = \mathcal{X}_t \sim p(\mathcal{X}_t | \mathcal{X}_{1:t-1}). \tag{1}$$

Considering that there are three types of conditional independent factors: Personal factor, Social Correlation factor, and Social Influence factor which affect individual's future behaviors, the formulation 1 can be rewritten as follows:

$$p(\mathcal{X}_t | \mathcal{X}_{1:t-1}) \propto \prod_{i=1..N} p(X_t^i | X_{1\ldots t-1}^i; \theta_i) \prod_{i=1\ldots N} p(X_t^i | X_{1\ldots t-1}^i, X_{1\ldots t-1}^{F(i)}; \phi_i) \prod_{(i,j)\in\mathcal{G}} p(X_t^i, X_t^j; \Omega), \tag{2}$$

where notation $F(i)$ represents the neighbors of user u_i in the social network. The posterior probability has three types of factor functions, corresponding to the intuitions: (1) In particular personal factor $p(X_t^i | X_{1\ldots t-1}^i; \theta_i)$ represents the posterior probability of u_i's behavior X_t^i at time t given the past behavior records $X_{1\ldots t-1}^i$, where θ_i is the parameter. (2) Social correlation factor $p(X_t^i, X_t^j; \Omega)$ reflects the correlation between pair of friends' behaviors at time t, where Ω is the social correlation parameter. (3) Social influence factor $p(X_t^i | X_{1\ldots t-1}^i, X_{1\ldots t-1}^{F(i)}; \phi_i)$ indicates friends' influence on u_i's behavior at time t, where ϕ_i is the influence parameter.

Then we introduce a latent space $\mathcal{S}(S_t^i \in \mathcal{S})$, where S_t^i assumes to be the sufficient statistics for $X_{1..t}^i$. Thus, the Eq. 2 can be rewritten as follows,

$$p(\mathcal{X}_t | \mathcal{X}_{1:t-1}) \propto \int_{S_{t-1}, S_t} \prod_{i=1\ldots N} p(X_{1\ldots t-1}^i | S_{t-1}^i) p(X_t^i | S_t^i) \prod_{i=1\ldots N} p(S_t^i | S_{t-1}^i; \theta_i) \prod_{i=1\ldots N} p(S_t^i | S_{t-1}^i, S_{t-1}^{F(i)}; \phi_i) \prod_{(i,j)\in\mathcal{G}} p(S_t^i, S_t^j; \Omega). \tag{3}$$

It isolates the last three terms in the Eq. 3 denoted by $p(S_t | S_{t-1})$, which represents the probability of S_t given S_{t-1}. By incorporating the three factor functions together, the posterior distribution can be naturally modeled in a conditional random field,

$$p(S_t | S_{t-1}) = \frac{1}{Z} exp\Big(\sum_{i=1\ldots N} \theta_i f(S_t^i, S_{t-1}^i) + \sum_{i=1\ldots N} \phi_i g(S_t^i, S_{t-1}^i, S_{t-1}^{F(i)}) + \sum_{(i,j)\in\mathcal{G}} \Omega_{ij} k(S_t^i = S_t^j) \Big), \tag{4}$$

where Z is the partition function (normalization factor). The feature functions f, g, k correspond to the personal factor, social influence factor, and social correlation factor. Suppose the latent sufficient statistics S_t^i get value from space $\mathcal{S}$, which has K discrete states. Then, the model parameters for personal factor, social influence factor, and social correlation factor are $\theta_i \in R^{K*K}$, $\phi_i \in R^{K^{|F(i)|+2}}$, and $\Omega_{ij} \in R$. Therefore, the bottleneck of the general influence model is that the number of parameters for social influence factor is very huge ($O(K^{|F(i)|+2})$), which makes the model intractable.

Another possible solution is to model joint social influence factor as coupled influence factors:

$$\phi_i g(S_t^i, S_{t-1}^i, S_{t-1}^{F(i)}) = \sum_{j\in F(i)} \phi_{ij} g'(S_t^i, S_{t-1}^i, S_{t-1}^j). \tag{5}$$

As a result the number of parameters for social influence factors ϕ_{ij} would be reduced in $O(K^3)$. Although the coupled social influence factor model could largely decrease the number of model parameters (model complexity), it fails to capture the important property in social influence: "the whole is larger than the sum of its parts." For example, one day a person is making a decision about whether to play basketball or not. If only one of his friends suggests that they play basketball together, the influence would be rather weak. However, if three or four of his friends suggest that he play basketball with them, it would significantly affect his decision. Therefore, the coupled social influence model would be inaccurate in practice.

Since the general social influence model is intractable due to the huge number of social influence parameters, the huge number of parameters would also lead to the increasing risk of overfitting. For instance, suppose there are two users u_A and u_B, u_A plays table tennis, and u_B plays basketball. After they befriend each other, u_A starts swimming. In this case, instead of viewing u_A's behavior is influenced by u_B it is better to say that u_A's behavior might be caused by other unknown factors.

In terms of social influence, many studies [1,7] show that social influence can be viewed as the source of social correlation. Therefore, it is not necessary to model social influence factor, as that an individual's behavior is explicitly influenced by his or her friend's past behaviors. Instead, social influence factor can be implicitly incorporated by modeling the dynamic of social correlation.

3 Dynamic social correlation and socialized Gaussian process models

In this section, we propose an intuitive dynamic social correlation model to capture the evolution of social correlation over time, and also socialized Gaussian process models to incorporate both personal factor and social correlation factor into a unified model. To enhance the smoothness of the paper, all the notations are summarized in Table 1. First, we give the formulation of the general (implicit) social influence model based on dynamic social correlation as follows:

$$p(X_t|X_{1:t-1};\Theta,\Omega) \propto \int_{S_t,S_{t-1}} p(X_{1:t-1}|S_{t-1})p(X_t|S_t)p(S_t|S_{t-1};\Theta,\Omega) \tag{6}$$

where

$$p(S_t|S_{t-1};\Theta,\Omega) = \frac{1}{Z}exp\left(\sum_{i=1..N}\theta_i f(S_t^i, S_{t-1}^i) + \sum_{(i,j)\in\mathcal{G}}\Omega_{ij,t}k(S_t^i = S_t^j)\right). \tag{7}$$

$$\Omega_t \sim p(\Omega_t|\Omega_{t-1}, S_{t-1}). \tag{8}$$

In the general dynamic social correlation-based social influence model, we do not explicitly incorporate social influence factor in Eq. 7. However, the influence factor is implicitly captured by modeling the dynamic of social correlation in Eq. 8. The parameters Ω_t for social correlation factor at time t can be viewed as a weighted graph, where $\Omega_{ij,t}$ is the weight of the edge (i, j) that measures the correlation between u_i and u_j at time t.

3.1 Dynamic social correlation

In this subsection, we introduce an intuitive evolution model for capturing the dynamic social correlation. Assuming the social correlation parameter $\Omega_{ij,t}$ for u_i and u_j at time t depends

Table 1 Notations

Notation	Description
$\mathcal{G} = (V, E)$	The social network
V	The set of vertices
E	The set of edges
T	The number of timestamps in the training set
N	The number of users in the social network
$F(i)$	The neighbors of user u_i
$\mathcal{X}^i = (x_1^i, x_2^i, ...x_T^i)$	User u_i's sequential behavior
$\mathcal{X}_{1...t} = (X_{1...t}^1, X_{1...t}^2, \ldots, X_{1...t}^N)$	Individuals' past behaviors until day t
θ_i	The personal factor parameter
Ω	The social correlation parameter
ϕ_i	The influence parameter
$\Omega_{ij,t}$	Social correlation parameter for u_i and u_j at time t
$\mathcal{S}(S_t^i \in \mathcal{S})$	A latent space
K	The number of discrete states in latent space $\mathcal{S}$
Φ	The cumulate Gaussian function
K^i	The covariance (kernel) matrix
$K_{t,t'}^i = C(S_t^i, S_{t'}^i, \theta_i)$	The covariance function
$cycle_t^i$	The autocorrelation function (ACF)
$F = \{\mathcal{F}_1, \ldots, \mathcal{F}_n\}$	A set of behavior-related features

on its previous value $\Omega_{ij,t-1}$ and also $\Omega_{iF(i),t-1}$, $\Omega_{jF(j),t-1}$, where $F(i)$ and $F(j)$ are the friends set for u_i and u_j.

$$\Omega_{ij,t} = q(\Omega_{ij,t-1}, \Omega_{iF(i),t-1}, \Omega_{jF(j),t-1}). \tag{9}$$

Intuitively, there are two type of effects proposed to characterize the evolution of social correlation:

Flock Effect: People tend to have the similar behaviors with the one whose behavior is largely correlated with others. For example, User u_A's behavior is similar with a large number of other users. Then, user u_B could also tend to follow u_A with high probability.

Clustering Effect: People tend to have similar behaviors with another if they have: (a) a large number of friends in common, and (b) highly correlated friends in common.

Therefore, by incorporating the two effects, we adopt a linear regression approach for estimating Ω_t. Take $\Omega_{ij,t}$ for example,

$$\Omega_{ij,t} = \alpha_1\Omega_{ij,t-1} + \alpha_2\overline{\Omega}_{i,t-1} + \alpha_2\overline{\Omega}_{j,t-1} + \alpha_3\widetilde{\Omega}_{ij,t-1} + \alpha_4 log(|F(i) \cap F(j)|). \tag{10}$$

Where $\overline{\Omega}_{i,t-1}$ and $\overline{\Omega}_{j,t-1}$ are the corresponding average social correlation coefficient for u_i and u_j, respectively. $\widetilde{\Omega}_{ij,t-1}$ indicates the average social correlation coefficient between

u_i, u_j and their common friends. $|F(i) \cap F(j)|$ is the number of common friends for u_i and u_j.

$$\begin{aligned} \overline{\Omega}_{i,t-1} &= \frac{1}{|F(i)|} \sum_{k \in F(i)} \Omega_{ik,t-1}, \\ \overline{\Omega}_{j,t-1} &= \frac{1}{|F(j)|} \sum_{k \in F(j)} \Omega_{jk,t-1}. \end{aligned} \tag{11}$$

$$\widetilde{\Omega}_{ij,t-1} = \frac{1}{|F(i) \cap F(j)|} \sum_{k \in F(i) \cap F(j)} (\Omega_{ik,t-1} + \Omega_{jk,t-1}) \tag{12}$$

Therefore, in the proposed dynamic social correlation model, there are only four parameters to implicitly characterize the social influence process.

3.2 Gaussian process models

Notice that so far we have not provided the details about how to obtain sufficient statistics S_t, how to estimate the emission probability $p(X_{1:t-1}|S_{t-1})$ and $p(X_t|S_t)$, and how to specify the dimension of latent space $S(S_t^i \in S)$.

However, if latent space $\mathcal{S}$ is defined with K-dimensional discrete states, S_t would be in K^N (N is the number of users in social network). The proposed model is still intractable, since there is no closed form to integral out S_t and S_{t-1} in the Eq. 6.

Therefore, a nonparametric Bayesian approach, Gaussian process, is employed to make the model tractable. Here latent space $\mathcal{S}$ is defined in a continuous real value space R. Then define the emission probability $p(X_t|S_t)$ as follows:

$$p(X_t|S_t) = \prod_{i=1..N} \Phi(X_t^i * S_t^i) \tag{13}$$

where the probit function Φ is the cumulative Gaussian function, $\Phi(x) = \int_{-\infty}^{x} \mathcal{N}(\tau|0, 1)d\tau.$, and $\Phi(x) + \Phi(-x) = 1$.

Since S_{t-1}^i is a single real value in R, and since it cannot be a sufficient statistic for an individual's past behaviors, we rewrite the Eqs. 6 and 7, by replacing S_{t-1}^i with $S_{1:t-1}^i$

$$p(X_t|X_{1:t-1}; \Theta, \Omega) = \int_{S_t, S_{1:t-1}} p(X_{1:t-1}|S_{1:t-1}) p(X_t|S_t) p(S_t|S_{1:t-1}; \Theta, \Omega) \tag{14}$$

$$p(S_t|S_{1:t-1}; \Theta, \Omega) = \frac{1}{Z} exp\left(\sum_{i=1..N} f(S_t^i, S_{1:t-1}^i, \theta_i) + \sum_{(i,j) \in \mathcal{G}} -\Omega_{ij,t} (S_t^i - S_t^j)^2 \right). \tag{15}$$

where the feature function $f(S_t^i, S_{1:t-1}^i, \theta_i)$ can be explicitly given by the Gaussian process model,

$$f(S_t^i, S_{1:t-1}^i, \theta_i) = -\begin{pmatrix} S_{1:t-1}^i \\ S_t^i \end{pmatrix}^T \begin{pmatrix} K_{1:t-1}^i & K_{(1:t-1),t}^i \\ K_{(1:t-1),t}^i{}^T & K_t^i \end{pmatrix} \begin{pmatrix} S_{1:t-1}^i \\ S_t^i \end{pmatrix} \tag{16}$$

where K^i is the covariance (kernel) matrix, with its element $K_{t,t'}^i$ defined by covariance function $K_{t,t'}^i = C(S_t^i, S_{t'}^i, \theta_i)$. Supposing $S_{1..t}^i$ is the stationary process, the covariance function is defined by the individual's behavior autocorrelation kernel,

$$C(S_t^i, S_{t'}^i, \theta_i) = \theta_1^i exp(mod((t-t'), cycle^i(t)) * \theta_3^i) + \theta_2^i exp((t-t')/cycle^i(t) * \theta_4^i) + \theta_5^i * I(t=t'). \quad (17)$$

where $cycle_t^i$ is computed by autocorrelation function (ACF),

$$cycle^i(t) = argmax_\tau \Psi(\tau) = \frac{1}{t-\tau} \sum_{n=1..t-\tau} X_n^i * X_{n+\tau}^i \quad (18)$$

Therefore, by replacing the term $f(S_t^i, S_{1:t-1}^i, \theta_i)$ in Eq. 15 with the formulation 16, we can obtain that the closed form of conditional probability over S_t given $S_{1:t-1}$ is a multivariate Gaussian distribution. (Here, we omit the mathematical details).

$$p(S_t|S_{1:t-1}; \Theta, \Omega) = \mathcal{N}(S_t|\mu_t, \nu_t) \quad (19)$$

where the mean and covariance μ_t, ν_t are given as follows,

$$\begin{aligned} \mu_t &= (\mathcal{V}_t + \mathcal{L}_t)^{-1}\mathcal{V}_t\mathcal{U}_t \qquad \nu_t = (\mathcal{V}_t + \mathcal{L}_t)^{-1} \\ \mathcal{U}_t &= (K_{(1:t-1),t}^i{}^T K_{t-1}^i{}^{-1}(S_{1:t-1}^i))_{i=1..n}^T \\ \mathcal{V}_t &= diag\{K_t^i - K_{(1:t-1),t}^i{}^T K_{t-1}^i{}^{-1} K_{(1:t-1),t}^i\}_{i=1..n} \\ \mathcal{L}_{t,ii} &= \sum_{j=1..n} \Omega_{ij,t} \qquad \mathcal{L}_{t,ij} = -\Omega_{ij,t} \end{aligned} \quad (20)$$

We also developed online prediction/inference phases to verify the effectiveness of the proposed socialized Gaussian process model. In the online inference phase, the posterior probability of latent state S_t in the model is estimated given the observation of individuals' behavior X_t at time t. In the online prediction phase, we estimate the conditional probability of X_t given the past individuals' latent state $S_{1...t-1}$. In Fig. 2, the graphical model for the socialized Gaussian process is given, and the online prediction/inference phases are also illustrated. The first step is to calculate the conditional probability of S_t given S_{t-1}. The second step is the online predicting phase. The third step is the online inference phase, which updates S_t by estimating the posterior probability of S_t given observation X_t.

Online Prediction: In the online prediction phase, socialized Gaussian process model predicts the individuals' future behavior X_t at time t by estimating the conditional probability of X_t given $S_{1...t-1}$, i.e.,

$$p(X_t|S_{1:t-1}; \Theta, \Omega, \alpha) \propto \int_{S_t} p(S_t|S_{1:t-1}) \prod_{i=1..N} \Phi(X_t^i S_t^i) \propto \prod_{i=1..N} \Phi\left(\frac{X_t^i \mu_t^i}{\sqrt{\nu_t^i}}\right) \quad (21)$$

where μ_t^i and ν_t^i are given in the Eq. 20.

Online Inference: In the online inference phase, the posterior probability over S_t given the observation of individuals' behavior X_t is estimated,

$$p(S_t|X_t, S_{1:t-1}; \Theta, \Omega) \propto p(S_t|S_{1:t-1}) \prod_{i=1..N} \Phi(X_t^i S_t^i) \quad (22)$$

Unfortunately, the posterior is non-Gaussian. In practice, the first two moments of S_t are often used to construct a Gaussian approximation. Here, our approximation is based on a variational approach known as adaptive density filter (ADF) [20]. Given a posterior distribution for S_t, ADF finds a Gaussian approximation that matches the first two moments of S_t. Specifically,

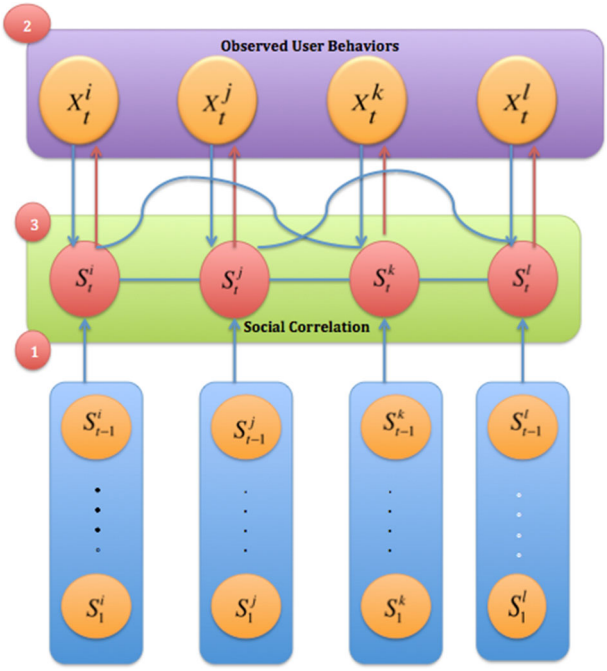


Fig. 2 Social influence Gaussian Process for human behavior prediction

let $\mathcal{N}(S_t; \mu_t^*, v_t^*)$ be the target Gaussian, whose parameters μ_t^*, v_t^* are chosen to minimize the Kullback–Leibler divergence:

$$minKL\left(\prod_{i=1..N} \Phi(X_t^i S_t^i)\mathcal{N}(S_t; \mu_t, v_t)||\mathcal{N}(S_t; \mu_t^*, v_t^*)\right) \tag{23}$$

The optimization problem can be solved by moment matching up to the second order, yielding:

$$\mu_t^{*i} = \mu_t^i + X_t^i \sqrt{v_t^i}\mathcal{V}\left(\frac{\mu_t^i X_t^i}{\sqrt{v_t^i}}\right). \tag{24}$$

$$v_t^{*i} = v_t^i\left[1 - \mathcal{W}\left(\frac{\mu_t^i X_t^i}{\sqrt{v_t^i}}\right)\right]. \tag{25}$$

with $\mathcal{V}(x) = \frac{\mathcal{N}(x;0,1)}{\Phi(x)}$, $\mathcal{W}(x) = \mathcal{V}(x)(\mathcal{V}(x) + x)$.

Model Parameters: The model parameters $\Lambda = \{\Theta, \Omega, \alpha\}$, where Θ in the definition of covariance matrix in Eq. 17, are hyper-parameters for personal factor modeling, α is defined in Eq. 10 to capture the dynamic of social correlation, and Ω is defined in Eq. 15, to reflect the correlation of a pair of friends. In terms of parameters Θ, which serve as hyper-parameters in Gaussian process, we adopt a cross-validation method to determine Θ instead of performing learning strategy to infer hyper-parameters Θ. We found that setting the hyper-parameters manually could also achieve comparable performance. For social correlation parameters Ω,

we employ a prediction/updating scheme to dynamically update the Ω online. In the prediction phase, Ω_t is estimated by performing the dynamic social correlation process. In the updating phase, Ω_t is updated according to the following formulation:

$$\Omega_{ij,t} = \frac{1}{T} exp\left(- \sum_{m=t-T\ldots t} (X_m^i - X_m^j)^2 \right). \tag{26}$$

where T is the number of timestamps in the training set. Then parameters α are also dynamically updated by solving the linear regression model which maps Ω_{t-} to Ω_t.

Multi-Feature Social Correlation: In Eq. 26, the social correlation is inversely proportional to the square distance of binomial behaviors between two users i and j. The assumption behind Eq. 26 is that two users who have more similar behaviors will have a higher social correlation value. This function works well; however, only considering binomial behaviors might not be enough. Intuitively, two users can have the same behavior in the future if they share similar behavior-related features, e.g., the number of competitions, the number of goals, in the past. Let us denote a set of such features as $F = \{\mathcal{F}_1, \ldots, \mathcal{F}_n\}$. To consider F into the social correlation $\Omega_{ij,t}$, Eq. 26 has been modified as follows:

$$\begin{aligned} \Omega_{ij,t} = \frac{1}{T} exp\Big[& - \sum_{m=t-T\ldots t} (X_m^i - X_m^j)^2 \\ & - \sum_{\mathcal{F}\in F} \left(1 - cosine(\mathcal{F}^i_{\{t-T,\ldots,t\}}, \mathcal{F}^j_{\{t-T,\ldots,t\}})\right)\Big]. \end{aligned} \tag{27}$$

where F is a set of users' behavior-related features $\mathcal{F}$, $\mathcal{F}^i_{\{t-T,\ldots,t\}}$ is sequence of feature $\mathcal{F}$'s values from timestamp $t-T$ to timestamp t given user i, and $cosine(\cdot)$ is cosine similarity function. The parameters α are dynamically updated by solving the linear regression model which maps Ω_{t-} to Ω_t.

Note that the new Eq. 27 is the main difference between the novel multi-feature socialized Gaussian Process (mfSGP) model and the original socialized Gaussian process (SGP) model [18].

3.3 Overall online prediction/inference algorithm

The overall algorithm for online prediction/inference for socialized Gaussian process model is outlined in Algorithm 1. The time complexity of the algorithm is $O(TN^3)$ (N is the number of users in the social network), where $O(N^3)$ is the time complexity for computing matrix inversion. Even though the social network we collected is not very large, it is still acceptable. Here we omit the time complexity for updating α, since currently there are lots

Algorithm 1 Online Prediction/Inference Alg($\mathcal{X}_{1\ldots T}, \mathcal{G}, \Lambda$)

Require: $\mathcal{X}_{1..T}$:individuals' time series behavior records; $\mathcal{G}$: social network; Λ: model hyperparameters;
1: Loopy $t = (1 \ldots T)$
2: Perform the **dynamic social correlation** process according to Eq. 10.
3: Estimate conditional probability of S_t given S_{t-1} according to Eq. 19.
4: Online predict individuals' behavior X_t according to Eq. 21.
5: Estimate posterior probability of S_t given observation X_t according to Eq. 22.
6: Update Ω_t according to Eq. 27.
7: Update α by Least Squares Method.
8: END.

of mature technologies to solve the large-scale linear regression problem. In Algorithm 1, we do not consider the starting time problem, since there is no previous information that can be leveraged for prediction at the beginning of time $t = 1$. Therefore, in practices (experiments) we start the online prediction/inference algorithm from a middle time point.

4 Experiments

In this section, we use the real-world YesiWell data described in [16] and the corresponding social network to empirically validate and compare the effectiveness of the novel multi-feature socialized Gaussian process (mfSGP) model and the SGP model [18]. We first elaborate on the experimental configurations on the data set, evaluation metrics, and baselines. Then, we introduce the experimental results on individuals' behavior prediction. Finally, we conduct experiments via statistical data analysis for dynamic social correlation, and how it could help improve prediction accuracy. We provide several baseline models including two Random Forest models to compare the results with the mfSGP and SGP models.

4.1 Experiments configuration

Human Physical Activities Dataset. The YesiWell study was conducted in 2010–2011 as a collaboration between several health laboratories and universities to help people develop and maintain active lifestyles and lose weight. The dataset was collected from 254 users, including personal information, social network activities, and their daily physical activities for 10 months, from October 2010 to August 2011.

The initial physical activity data, collected by special electronic equipment for each user, are the numbers of steps and running distances. In our study, we transform the physical activities into the value $\{1, -1\}$, which indicates whether the user exercises or not. In the YesiWell dataset, there are a total of 40 weeks of records of individuals' daily activities. In the dataset, some users' daily records are missing. Therefore, we first give the basic analysis on the distribution of physical activity records in Fig. 3. As shown in Fig. 3, there are 14 users with their daily physical activities record number smaller than 10, and eight users with their records number larger than 10 but smaller than 20, etc. Therefore, to clean the data,

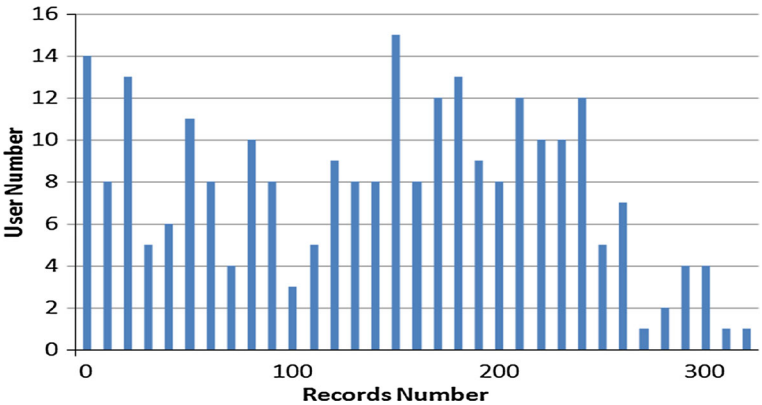


Fig. 3 Distribution of the record number and user number

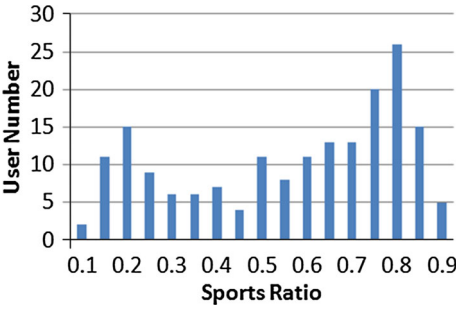


Fig. 4 Distribution of the sports ratio and user number

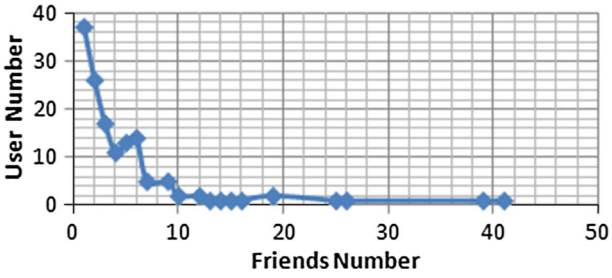


Fig. 5 Distribution of friends number in the YesiWell social network

we filtered the users whose daily physical activity record numbers are smaller than 80. We only use the remaining 185 users for experiments. The general statistics of the users' daily physical activities data are shown in Fig. 4, which shows the distribution of sports ratio[1] and user number. For instance, 15 users have done exercises in nearly 20 percent of their daily physical activity records.

The dataset contains 684 connections in a social network that consists of 185 individuals. On average each individual has four friends in the social network. In Fig. 5, it shows the distribution of the number of friends in the social network. We also use the daily number of competitions, the number of goals, and the number of steps at the users' behavior-related features, i.e., F in Eq. 26. The dataset contains 3100 competitions and 1229 goals.

Evaluation metric. To verify the effectiveness of the proposed SGP model and the novel mfSPG model, we conduct the experiments by predicting the individual's future activities according to their past behaviors and social network information. In the experiment, we select two weeks as the unit for prediction, i.e., leveraging the previous 10 weeks' daily records to predict the 11th and 12th weeks' behaviors of users. We use the metric **accuracy** to measure the prediction quality between week t and $t+1$.

$$accuracy = \frac{\sum_{i=1\ldots N} \sum_{d\in\{t,t+1\}} I((X_d^i \neq 0) = \widetilde{X}_d^i)}{\sum_{i=1\ldots N} \sum_{d\in\{t,t+1\}} I(X_d^i \neq 0)} \tag{28}$$

where X_d^i is the true user activity at day d for u_i, and $\widetilde{X}_d^i$ denotes the predicted value. $(X_d^i \neq 0)$ indicates the physical activity record is not missing. I is the indication function, where $I(X) = 1$ when X is true; otherwise, $I(X) = 0$. N is the number of users.

Baseline and Comparison Models. Our proposed models are compared with several baseline methods for individual behavior prediction. Typically, the baseline methods are

[1] Sports ratio is defined as the percentage of days a user has done exercise.

divided into two categories: personalized behavior prediction method and socialized behavior prediction method. Personalized method only leverages an individual's personal past behavior records for future behavior prediction. Socialized methods not only use personal behavior records, but also incorporate friends' past behavior for prediction. Specifically, the following eight prediction models are compared together:

Logistical Autoregression: Logistical autoregression (LAR) [2] utilizes logistical regression method to leverage historical activities for predicting future behaviors, i.e., $p(\widetilde{X}^i_d = 1) = logit(\alpha_0 + \sum_{m=1...W} \alpha_m X^i_{d-m})$, where m is the lag length for autoregression (set to be 9 for all users), which is determined by cross-validation. Results reported in all experimental results are based on the parameter configurations that produce the best results.

Personalized Gaussian Process: Personalized Gaussian process (PGP) model does not incorporate the social network information. It purely utilizes personal historical records for prediction. The observation of an individual's behavior is discrete, especially binary. Therefore, binary Gaussian process is used in our experiments [13]. In the setting of the personalized Gaussian process, the covariance (kernel) matrix is defined in Eq. 17. The hyperparameters Θ in PGP are determined by cross-validation.

Socialized Logistical Autoregression: Socialized autoregression (SLAR) method borrows the idea from [15], which combines the social influence and autoregression model together. SLAR models the social influence explicitly by incorporating friends' historical behaviors into the unified regression model, i.e.,

$$p(\widetilde{X}^i_d = 1) = logit\left(\alpha_0 + \sum_{m=1...W1} \alpha_m X^i_{d-m} + \sum_{f \in F(i)} \pi^i_f \sum_{m=1...W2} \beta_m X^f_{d-m}\right) \quad (29)$$

where parameter π^i_f indicates what degree friend u_f's past behavior could effect u_i's behavior. $W1$ and $W2$ are the two lag lengths determined by cross-validation.

Behavior Pattern Search: Behavior pattern search (BPS) is an ad hoc method for personalized behavior prediction. The main idea of BPS is searching the most similar behavior pattern from past behavior logs for predicting future behavior, i.e.,

$$\begin{aligned} &\widetilde{X}^i_d = X^i_{d-p}, where \\ &p = argmax_{p=1...W1} \sum_{m=1...W2} X^i_{d-m} X^i_{d-m-p} \end{aligned} \quad (30)$$

where $W1$ is the size of the search window, and $W2$ is the size of the window pattern. In our experiments, $W1$ and $W2$ are set to 60 and 15, respectively, toward achieving the best performance.

Random Forest: Random Forest (RF) model [3] incorporates each user's historic behaviors and makes predictions of his or her future behavior based on his or her past behaviors. It looks through each user's behavior records and learns a random forest. It combines the learned random forest with the user's records of the next timestamp to predict the user's future behavior. The model is updated with the data for each iteration.

Socialized Random Forest: Socialized Random Forest (SRF) model incorporates not only the personalized historical behaviors but also the user's social network information. When creating random forest trees, it combines the user's physical behavior and also his or her friends' physical behaviors at each time period as a set of different features. Then the model picks the whichever one has the best information gain and splits the tree together.

Socialized Gaussian Process: Socialized Gaussian process (SGP) model [18] incorporates both personalized historical behavior records and also dynamic social correlation

Table 2 Prediction accuracy comparison with different models (T7–T38)

Weeks	LAR	PGP	SLAR	BPS	RF	SGP	SRF	mfSGP
T7–T8	0.56122	0.62448	0.55918	0.53877	0.68571	0.71836	0.73244	**0.76531**
T9–T10	0.60404	0.63872	0.62716	0.5881	0.63873	0.64595	**0.69717**	0.68642
T11–T12	0.64591	0.64177	0.62102	0.5988	0.70393	0.67731	**0.70731**	0.70539
T13–T14	0.67729	0.71812	0.69023	0.6344	0.71833	0.74402	0.74808	**0.76295**
T15–T16	0.69497	0.71923	0.71490	0.6603	0.71833	0.74870	0.74982	**0.76603**
T17–T18	0.72879	0.74734	0.7526	0.70053	0.74812	0.75088	0.74988	**0.76855**
T19–T20	0.73877	0.73518	0.7441	0.6921	0.73454	0.76211	0.75896	**0.78007**
T21–T22	0.72760	0.74680	0.7239	0.7029	0.74755	0.75411	0.75666	**0.77239**
T23–T24	0.72310	0.73526	0.71562	0.6604	0.73727	0.74555	0.75535	**0.76427**
T25–T26	0.71372	0.71567	0.70001	0.6679	0.71691	0.73417	0.72741	**0.75365**
T27–T28	0.72745	0.76089	0.7173	0.693	0.74983	0.77001	0.75322	**0.79027**
T29–T30	0.72866	0.71553	0.73304	0.6969	**0.76973**	0.74288	0.74002	0.76477
T31–T32	0.78965	0.79083	0.7755	0.7403	0.72259	0.80846	0.74575	**0.83196**
T33–T34	0.74007	0.75416	0.7387	0.7055	0.76347	0.76056	0.76382	**0.78617**
T35–T36	0.71059	0.73913	0.7105	0.6752	0.73051	0.74048	0.73894	**0.76766**
T37–T38	0.74616	0.76490	0.7427	0.712	0.74383	0.76320	0.75723	**0.79727**

The number in bold is the highest prediction accuracy among all models in each test period

information for future behavior prediction. In the SGP model, parameters α, in Eq. 10 are estimated by solving the linear regression problem. The hyperparameters Θ are determined by validation like in PGP.

Multi-Feature Socialized Gaussian Process: Multi-feature socialized Gaussian process (mfSGP) model incorporates both personalized historical behavior records and multi-feature dynamic social correlation information for future behavior prediction. The hyperparameters Θ are determined by validation like in SGP.

4.2 Experiment results

In this subsection, we report the performance of different human behavior models for predicting an individual's future behaviors. The individual's behavior records are divided according to time series, e.g., $T1 - T8$ indicates the records from the first week to the eighth week. Therefore, we could evaluate the models in different time periods. As shown in the Tables 1 and 2, we compare accuracy across the eight human behavior prediction models: Logistical autoregression (LAR), personalized Gaussian process (PGP), socialized logistical autoregression (SLAR), behavior pattern search(BPS), Random Forest (RF), socialized Gaussian process (SGP), socialized Random Forest (SRF), and multi-feature socialized Gaussian process (mfSGP). The rows in the table indicate different time periods.

4.2.1 Prediction accuracy

As shown in Table 2 and Fig. 6, the multi-feature socialized Gaussian process model (mfSGP) outperforms the other baseline methods and the SGP model. In general, the performance of behavior pattern search method (BPS) appears to be the worst among all these methods. One

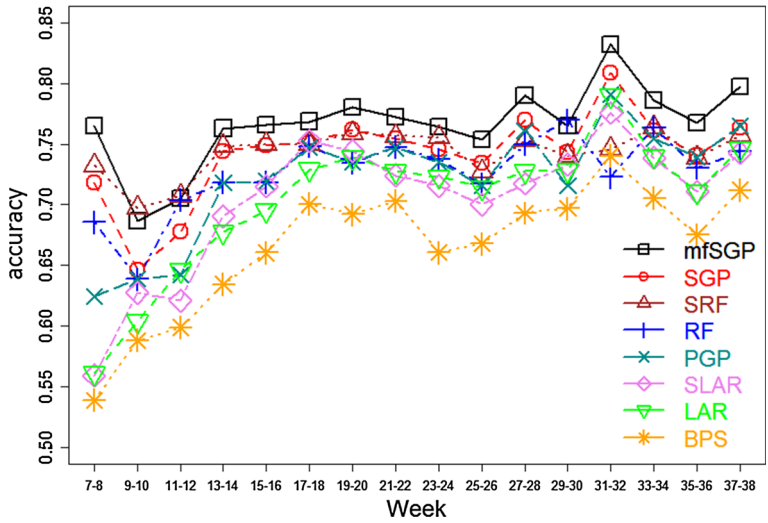


Fig. 6 Accuracy comparison of mfSGP and state-of-the-art human behavior prediction models

explanation is that it is sensitive to the noisy data. It is interesting to notice that the performance of logistical autoregression (LAR) is comparable with socialized logistical autoregression (SLAR) model which additionally incorporates friends' information into the model. Both the LAR and SLAR method get worse performance than the personalized Gaussian process (PGP) model. There are two reasons to explain why the SLAR method is unsuccessful in our experiments. First, incorporating additional parameters would increase the model complexity and also increase the risk of overfitting. Second, for humans' daily behavior, friends are often willing to participate in sports together at the same time, rather than following friends' previous behaviors.

In the experiments, BPS, LAR, and PGP are all personalized behavior prediction models. However, PGP significantly outperforms the other two methods by exploiting individuals' personal behavior records information. Multi-feature socialized Gaussian process (mfSGP) achieves further improvement based on SGP by incorporating multiple personal behavior-related features such as the number of goals, the number of competitions, and the number of steps into the dynamic social correlation information. In terms of accuracy, the mfSGP method improves the performance in average as high as 2.38, 2.47, 3.96, 4.59, 7.42, 7.80 and 16.13 % in contrast to SRF, SGP, RF, PGP, SLAR, LAR and BPS, respectively.

It is also interesting to compare SGP with SRF. Both are socialized behavior prediction models and perform better than the other five methods excluding the mfSGP model. At the beginning (i.e., from $T7$–$T8$ to $T15$–$T16$), SRF performs better than SGP (and other five models). When we have more training data, SGP becomes better than SRF and other baseline models from $T17$–$T18$ to the end (i.e., $T37$–$T38$). The only exceptions are $T21$–$T22$, $T23$–$T24$, and $T33$–$T34$, when SRF performs a little better than SGP. However, it is clear that the mfSGP model outperforms both SGP and SRF in most of the time periods.

Although in two short time periods, i.e., T9–T10 and T11–T12, the socialized Random Forest (SRF) model may be better than our proposed mfSGP model, it is hard to claim that the SRF model is generally better than our mfSGP model in short test period times. There are three specific observations that highlight that SRF is not generally preferable to our mfSGP model in short test period times: (1) The difference between the SRF model

Table 3 The mfSGP model vs the SRF model in short test time periods

	SRF	mfSGP
$T7$–$T8$	0.73244	0.76531
$T7$–$T10$	0.71480	0.72587
$T7$–$T12$	0.71231	0.71904

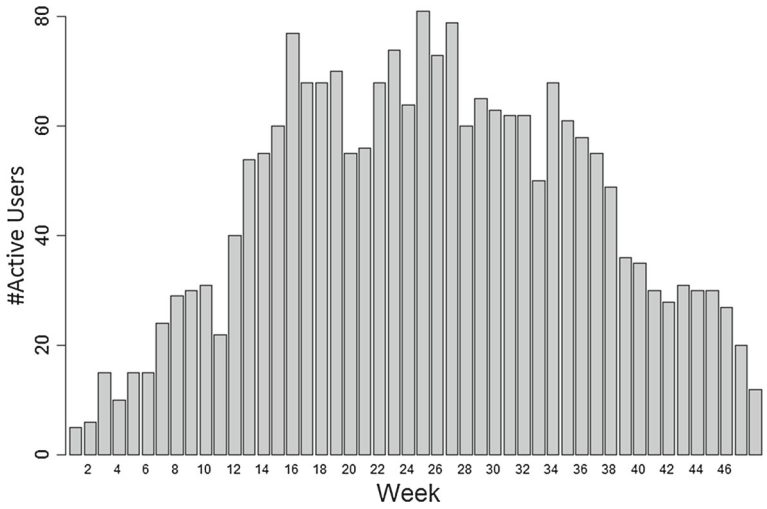


Fig. 7 The distributions of active users

Table 4 Paired t test (2-tail) results

	SGP	SRF	RF	PGP	SLAR	LAR	BPS
mfSGP	0.009	0.009	0.006	$1.36e-3$	$2.77e-4$	$2.02e-4$	$5.89e-8$

and our model in terms of accuracy is relatively small, i.e., $<1\,\%$ (Table 2); (2) We have compared our mfSGP model in different short test time periods within the first 2, 4, and 6 weeks, i.e., $T7$–$T8$, $T7$–$T10$, and $T7$–$T12$. In these three short periods, our mfSGP model outperforms the socialized Random Forest (SRF) model (Table 3); and (3) In other real-world applications, e.g., microarray-based cancer classification, some other model, such as support vector machine (SVM) [6], can outperform Random Forest (RF) [3] in general. However, it is not a big surprise if the SRF model performs better in few test cases [19]. It is similar to our study. Another possible reason is that, in the beginning of our study, our health social network had not been completely stable, since users needed time to warm up and engage to the program. For instance, the number of active users doing exercises in the first 12 weeks is significantly lower than other time periods in our study (Fig. 7). The most significant result is that the mfSGP model outperforms both SGP, SRF, and other models in most of the time periods.

Finally, to validate the statistical significance of our experiments, we perform the paired t test (2-tail) over the accuracy of the experiential result. As shown in Table 4, all the t

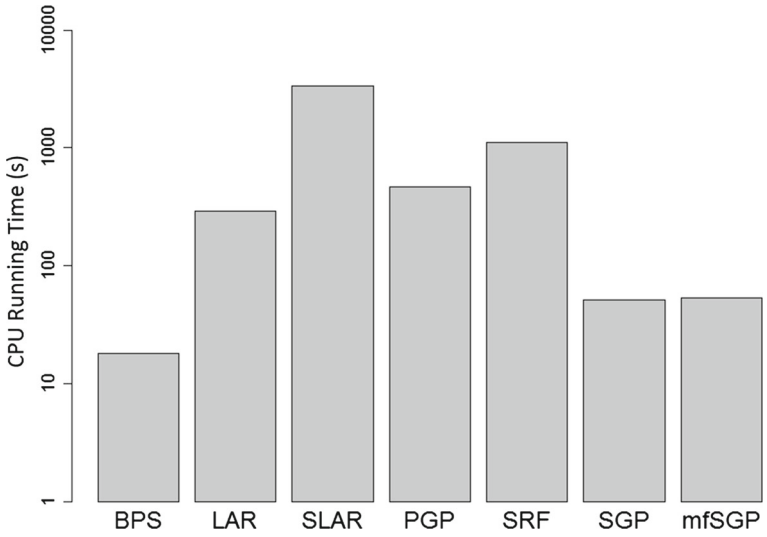


Fig. 8 Running time of mfSGP model and state-of-the-art human behavior prediction models

test results are <0.01, which means the improvements of mfSGP over other methods are statistically significant.

4.2.2 Running time

In order to compare the algorithms in terms of running time, we compute average CPU processing time of all the algorithms. Each algorithm will be run 15 times, and the average running time will be reported. Figure 8 illustrates that our proposed SGP and mfSGP models achieve very competitive running time compared with other baseline approaches. The BPS model is fastest since it is a very simple ad hoc model and its accuracy is not good. The SGP and mfSGP models are similar to each other in terms of running time. The interesting point here is that the PGP model is much slower than the mfSGP and SGP models. In theoretical analysis, the time complexity of the PGP model is $O(N \times T^3 \times M)$. Meanwhile, the time complexity of the SGP and mfSGP models is $O(T \times N^3 \times M)$. M is the number of testing cases which is 14 in this case (Table 2; Fig. 6). In addition, $N = 185\ users$ and $T = 10(weeks) \times 7(days) = 70\ days$. Therefore, PGP must be faster than both SGP and mfSGP since $O(N \times T^3 \times M) < O(T \times N^3 \times M)$. However, the matrix of users $N \times N$ is much sparser than the matrix of times $T \times T$ in real-world cases. In fact, the number of entries in the matrix $N \times N$ we need to compute is the number of friend connections in the social network. In our health social network, the total number of friend connections is 1376, which is significantly smaller than the number of entries in the matrix of users $N \times N = 185 \times 185 = 34{,}225$. We only need to compute 1376 out of 34,225 entries. In fact, we can avoid 96 % of computations in total. It results in very efficient models, in terms of running time. The mfSGP model takes 53.2 s on average to accomplish all the learning and prediction tasks in the whole dataset. This is 8.74 times faster than the PGP model which takes 465.11 s in average to finish its tasks.

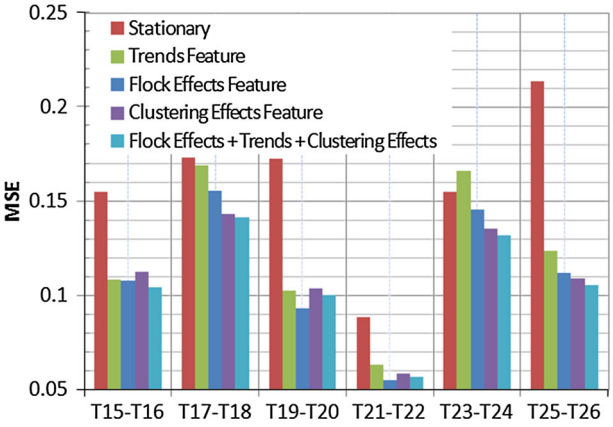


Fig. 9 Dynamic social correlation: comparison with MSE on different features

4.3 Statistical data analysis for dynamic social correlation

In the proposed mfSGP model, social influence is implicitly modeled by assuming that social influence would lead to the evolution of social correlation. Therefore, in this subsection, we perform a detailed analysis on dynamic social correlation. To capture the dynamic of social correlation, we utilize the mean square error (MSE) to measure the difference of social correlations at two time points, i.e., social correlation Ω_t and $\Omega_{t'}$, their difference is measured:

$$MSE = \frac{1}{E} \sum_{(i,j)\in\mathcal{G}} (\Omega_{ij,t} - \Omega_{ij,t'})^2. \tag{31}$$

In our experiments, social correlation is pairwise based, and also time dependent. At time t, the social correlation is estimated as follows (where T is 7 in experiments):

$$\Omega_{ij,t} = exp\Big(- \frac{1}{2T} \sum_{d=t-T\ldots t+T} (X_d^i - X_d^j)^2 - \sum_{\mathcal{F}\in F} \big(1 - cosine(\mathcal{F}^i_{\{t-T,\ldots,t+T\}}, \mathcal{F}^j_{\{t-T,\ldots,t+T\}})\big)\Big). \tag{32}$$

Based on the metric of MSE, we can conduct the experiments to measure the accuracy of predicting future social correlation by our proposed dynamic social correlation model. In the paper, we select three types of features to capture the dynamic of social correlation along the time. They are: previous social correlation (to capture the trend of social correlation), clustering effects features, and flock effect features. We compared the performance of the three features individually with the unified model which joined all of them together. In Fig. 9, we give the comparison with MSE on different features. "Stationary model" assumes that social correlation is static. Thus, the MSE in "stationary model" is the difference of social correlations in the two time periods, i.e., MSE at time $T15 - T16$ in "stationary model" is the difference of social correlation at time $T13 - T14$ and at time $T15 - T16$. The other four models in Fig. 9 leverage their corresponding features to predict the dynamic of social correlation. Generally, the best model (feature) could achieve the lowest MSE. From Fig. 9, we can observe that the Trends feature can achieve pretty good performance, sometimes. Yet, in at least one other instance, its performance was even less effective that than of the

Table 5 Prediction accuracy comparison with different dynamic social correlation feature (T15–T26)

Model	T15–T16	T17–T18	T19–T20	T21–T22	T23–T24	T25–T26
Stationary	0.7524	0.7583	0.7628	0.7431	0.7473	0.7462
Trends	0.7556	0.7591	0.7634	0.7595	0.744	0.7514
Clustering	0.7568	0.7605	0.7660	0.7595	0.744	0.7514
Flock	0.7547	0.7623	0.7606	0.7594	0.7599	0.7530
Unified	0.7566	0.7612	0.7628	0.7601	0.7621	0.7544

stationary model. Both the flock effects feature and the clustering effects feature have more stable performance than the Trends feature, and their performances are comparable. Among all of the methods, the joint feature model is slightly better than the others.

In addition, we leverage several different dynamic social correlation models for predicting individuals' activities. The performance comparison is shown in Table 5. The performance of the stationary model is as expected. Generally, the stationary model consistently performs the worst among all of the dynamic social correlation models, since it does not consider the evolution of social correlation, whereas the other three models are comparable with each other. The unified model could achieve only a slightly better performance than the other two methods. Therefore, based on experiments, the conclusion is that modeling the dynamic of social correlation more accurately would lead to further improvement in the proposed mfSGP model.

5 Related work and discussion

Social influence has been extensively studied in sociology [5], marketing [11], and psychology [12]. Empirical work on social influence shows that innovation [8], obesity [5] and news [9] can spread through social networks. In the literature, there are two different social influence models; namely, linear threshold model (LT) and independent cascade model (IC) [10] proposed to explain the diffusion process in social network. Therefore, both LT and IC models can also be viewed as contagion-based social influence models, which assume that an individual would follow his or her friends' past behaviors, and once the individual is infected, his or her behavior would not be influenced by others. Therefore, contagion-based social influence model is not suitable when the individual's behavior is dynamically changing along the time. Recently, social influence has been extended to community-level [14,16]. In [16], Phan et al. introduce a Community-level Physical Activity Propagation (CPP) model which is a graph summarization paradigm for the analysis of physical activity propagation and social influence. The CPP model reveals that social influences can propagate physical activity in health social networks at both user-level and community-level. A community is identified by a set of communicated nodes that share a similar physical activity influence tendency over nodes belonging to other communities.

Another possibility of modeling social influence is the latent state social influence model (LSSI), which considers the friends' past behaviors as an influence factor to potentially affect the individual's future behaviors [21]. However, this approach has some technical challenges. First, it is hard to determine how long an individual's friends' past behaviors could affect his or her future behavior. Second, in the LSSI model, it often models joint social influence factor

as coupled influence factors to reduce the model complexity. The assumption of a coupled approach is that social factors caused by different friends are conditionally independent. However, this assumption fails to capture the property "the whole is larger than the sum of parts" in social influence. Third, the LSSI approach modeling social influence explicitly would largely increase the risk of overfitting. Studies from both sociology and psychology suggest people consider the social influence as the source of social correlation [12]. However, the LSSI model does not promise to increase social correlation in the social influence process. Thus, it may introduce lots of "false" social influence into the model.

The short version of this paper [18] was published at IEEE International Conference on Data Mining (ICDM'12). Compared with the socialized Gaussian process model (SGP) in [18], in this paper, the dynamic social correlation schemes have been improved by leveraging multiple personal behavior-related features. Furthermore, we introduce two new baseline methods, Random Forest (RF) and social Random Forest (SRF) methods, for comparison with the new approach. The experiment results demonstrate that through incorporating the dynamic social correlation factors with multiple personal behavior-related features, our method can achieve the best prediction accuracy.

Our proposed models, i.e., SGP and mfSGP, have been designed to work with mutual friend relationship social networks. However, the models can be easily extended from mutual friend networks to non-mutual friend networks which are very popular in real-world applications, e.g., the social networks of Twitter, and Facebook. To address this point, users i and j in the social correlation function, i.e., Equations 10, 11, 12, and 26, means user i "follows" user j. A "follows" relationship can be considered as a non-mutual friend connection. Therefore, our proposed models can be easily applied to both mutual and non-mutual friend networks. In addition, the condition $(i, j) \in \mathcal{G}$ in Equations 2, 3, 4, 7, and 15 is considered as user i "follows" user j. This modification allows us to directly apply the proposed SGP and mfSGP models on non-mutual friend networks such as Twitter and Facebook.

6 Conclusion and future work

In this paper, we have developed a novel social influence model, referred to as multi-feature socialized Gaussian process (mfSGP) for socialized human behavior modeling. It is an extension from the Socialized Gaussian Process (SGP) model in our conference paper [18]. In both SGP and mfSGP models, the dynamic social correlation is captured and modeled as the result of social influence. Thus, the SGP and mfSGP models naturally incorporate personal behavior factor and social correlation factor into a unified model, and model the social influence factor based on the dynamic social correlation. The detailed experimental evaluation has demonstrated the effectiveness and accuracy of the novel mfSGP model, which performs better than state-of-the-art human behavior prediction models including our previous SGP model. It is interesting that a socialized Random Forest model performs better than SGP at the beginning periods, but the new mfSGP model wins the competition by using multiple physical activity-related features in the dynamic social correlation learning. In our future work, we plan to expand the mfSGP model to study how the biometrics and biomarkers, such as BMI, HDL, LDL, and triglycerides, can be predicted through the physical activities. We also plan to study whether mfSGP can be adopted to model the general time series and network structure data, such as biological and/or economic systems.

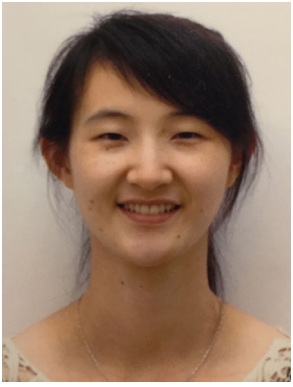
Acknowledgments This work is supported by the NIH grant R01GM103309. The authors appreciate the anonymous reviewers for their extensive and informative comments to help improve the paper. The authors also appreciate the contribution of Nafisa Chowdhury.

References

1. Anagnostopoulos A, Kumar R, Mahdian M (2008) Influence and correlation in social networks, in 'KDD'08', pp 7–15
2. Arnod BC, Robertson CA (1989) Autoregressive logistical regression. J Appl Probab 26:524–531
3. Breiman L (2001) Random forests. Mach learn 45(1):5–32
4. Chen W, Wang C, Wang Y (2010) Scalable influence maximization for prevalent viral marketing in large-scale social networks, in 'KDD'10', pp 1029–1038
5. Christakis NA, Fowler JH (2007) The spread of obesity in a large social network over 32 years. N Engl J Med 357(4):370–379
6. Cortes C, Vapnik V (1995) Support-vector networks. Mach Learn 20(3):273–297
7. Deutsch M, Gerard HB (1955) A study of normative and informational social influences upon individual judgment. J Abnorm Soc Psychol 51:629–636
8. Fichman RG, Kemerer CF (1999) The illusory diffusion of innovation: an examination of assimilation gaps. Inf Syst Res 10:255–275
9. Gomez Rodriguez M, Leskovec J, Krause A (2010) Inferring networks of diffusion and influence, in 'KDD'10', pp 1019–1028
10. Goyal A, Bonchi F, Lakshmanan LV (2010) Learning influence probabilities in social networks, in 'WSDM'10', pp 241–250
11. Huang J, Cheng X-Q, Shen H-W, Zhou T, Jin X (2012) Exploring social influence via posterior effect of word-of-mouth recommendations, in 'WSDM'12', pp 573–582
12. Kelman H (1958) Compliance, identification, and internalization: three processes of attitude change. J Confl Resolut 2:51–60
13. Kuss M, Rasmussen CE (2005) Assessing approximate inference for binary gaussian process classification. J Mach Learn Res 6:1679–1704
14. Mehmood Y, Barbieri N, Bonchi F, Ukkonen A (2013) Csi: Community-level social influence analysis, in 'ECML-PKDD'13', pp 48–63
15. Pan W, Cebrian M, Dong W, Kim T, Fowler J, Pentland A (2010) Modeling dynamical influence in human interaction patterns, Work abs/1009.0, 19. arxiv:1009.0240
16. Phan N, Dou D, Xiao X, Piniewski B, Kil D (2014) Analysis of physical activity propagation in a health social network, in 'CIKM'14', pp 1329–1338
17. Piniewski B, Muskens J, Estevez L, Carroll R, Cnossen R (2010) Empowering healthcare patients with smart technology. IEEE Comput 43(7):27–34
18. Shen Y, Jin R, Chowdhury NA, Dou D, Sun J, Piniewski B, Kil D (2012) Social gaussian process model for human behavior prediction in a health social network, In: ICDM12, pp 1110–1115
19. Statnikov A, Wang L, Aliferis C (2008) A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification. BMC Bioinform 9(1):319
20. Stern DH, Herbrich R, Graepel T (2009) Matchbox: large scale online bayesian recommendations. In 'WWW'09', pp 111–120
21. Tan C, Tang J, Sun J, Lin Q, Wang F (2010) Social action tracking via noise tolerant time-varying factor graphs. In 'KDD'10', pp 1049–1058
22. Tang J, Sun J, Wang C, Yang Z (2009) Social influence analysis in large-scale networks. In 'KDD'09', pp 807–816
23. Wang D, Wen Z, Tong H, Lin C-Y, Song C, Barabási A-L (2011) Information spreading in context. In 'WWW'11', pp 735–744

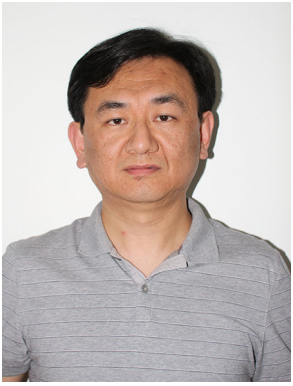
Yelong Shen is currently a research and software engineer in Microsoft Research, Redmond. He received his Ph.D. degree in Computer Science from Kent State University in Ohio in 2015, and Masters degree in Computer Science from BeiHang University, China in 2011. His main research interest includes deep learning, language understanding, and graph analysis.

NhatHai Phan is currently a Postdoctoral Research Associate at the Department of Computer and Information Science (CIS), University of Oregon. He received his Ph.D. degree in Computer Science from the French National Centre for Scientific Research (CNRS) - University Montpellier 2 (UM2) in 2013. His topics of interest focus on data science, machine learning, deep learning, and privacy and security, especially for health informatics, social network analysis, and spatiotemporal data mining. His contributions have been published in at least 24 publications at leading and top-tier venues, such as AAAI, ACM Multimedia, ACM CIKM, SDM, ECML-PKDD, ACM GIS, IEEE Intelligent Systems, KAIS.



Xiao Xiao is currently a software engineer at Poshmark, Menlo Park, CA. She received her M.S. degree in Computer Science from the University of Oregon in 2015. Her main research interest is utilizing machine learning algorithms to analyze user behavior in social networks.

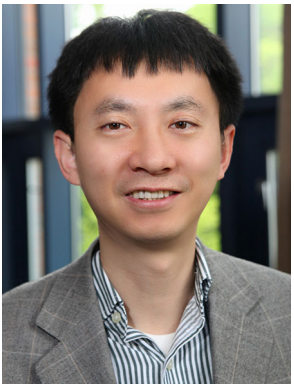
Ruoming Jin is currently an Associate Professor in the Department of Computer Science at Kent State University. He received his Ph.D. degree in Computer Science from the Ohio State University in 2005. His research interests include data mining (especially graph mining), complex network analysis, graph databases, data stream processing, and high performance computing. He has published over 70 technical papers and many appear in the top data mining and database conferences and journals, including SIGKDD, ICDM, SDM, SIGMOD, VLDB, ICDE, etc. Two of his papers received best paper nominations at the IEEE ICDM conference. He has served as senior PCs and PCs for various international data mining conferences and has co-chairs the four international workshops on mining multiple information sources. He is also a recipient of NSF CAREER award.



Junfeng Sun is currently a Mathematical Statistician at the Critical Care Medicine Department of Warren Grant Magnuson Clinical Center of National Institutes of Health. He received his Ph.D. in Biostatistics from Ohio State University in 2005. From 2005–2008, he was an Assistant Professor of Biostatistics in University of Nebraska Medical Center, and a Lead Statistician in Children's Oncology Group. From 2008–2009, he was an Assistant Professor of Biostatistics in Georgetown University.

Brigitte Piniewski is currently Chief Medical Officer of PeaceHealth Laboratories servicing Alaska, Washington and Oregon. She is also a longtime collaborator on clinical data projects with the University of Oregon and several other academic centers. With main research interest in technology-enabled community health optimization, Dr. Piniewski has led a number of projects advancing innovative self-care solutions. She has co-authored international technical reports as well as wireless health chapters and is best known for designing and implementing strategies for cross-generational collaboration. This work seeks to harness the digital skill surplus of our youth to advance the relative skill deficit of community members.

David Kil is currently Chief Data Scientist at CivitasLearning, building a family of SaaS connected analytics applications to help students succeed. Prior to CivitasLearning, he worked as Chief Science Officer at Humana and SKT Americas, building both enterprise and consumer health applications leveraging big data and networked sensors. He has also been building consumer health mobile apps. His research interests encompass all aspects of data processing, machine learning, complex event processing, nudging, and multi-level impact analysis. He has published over 35 papers in various journals and conferences, as well as a research monograph entitled "Pattern Recognition and Prediction with Applications to Signal Characterization by Springer-Verlag. He holds 13 US and international patents and is active as a reviewer for journals and grant agencies.



Dejing Dou is currently an Associate Professor in the Computer and Information Science Department at the University of Oregon. His research areas include ontologies, data mining, data integration, information extraction, and health informatics. Dejing Dou received his bachelor degree from Tsinghua University, China in 1996 and his Ph.D. degree from Yale University in 2004. Dejing Dou has published more than 60 research papers, some of which appear in prestigious conferences and journals like AAAI, KDD, ICDM, SDM, CIKM, ISWC, KAIS, JIIS and JoDS. His DEXA'15 paper received the best paper award. His KDD'07 paper was nominated for the best research paper award. He is on the Editorial Boards of Journal on Data Semantics and Journal of Intelligent Information Systems. Dejing Dou has received over $4.5 million PI research grants from the NSF and the NIH.