Topic-Aware Physical Activity Propagation with Temporal Dynamics in a Health Social Network

NHATHAI PHAN and JAVID EBRAHIMI, University of Oregon
DAVID KIL, HealthMantic Inc.
BRIGITTE PINIEWSKI, PeaceHealth Laboratories
DEJING DOU, University of Oregon

Modeling physical activity propagation, such as activity level and intensity, is a key to preventing obesity from cascading through communities, and to helping spread wellness and healthy behavior in a social network. However, there have not been enough scientific and quantitative studies to elucidate how social communication may deliver physical activity interventions. In this work, we introduce a novel model named **T**opic-aware **C**ommunity-level **P**hysical Activity Propagation with **T**emporal Dynamics (TCPT) to analyze physical activity propagation and social influence at different granularities (i.e., individual level and community level). Given a social network, the TCPT model first integrates the correlations between the content of social communication, social influences, and temporal dynamics. Then, a hierarchical approach is utilized to detect a set of communities and their reciprocal influence strength of physical activities. The experimental evaluation shows not only the effectiveness of our approach but also the correlation of the detected communities with various health outcome measures. Our promising results pave a way for knowledge discovery in health social networks.

CCS Concepts: • **Information systems → Data mining**; • **Human-centered computing → Social network analysis**; • **Applied computing → Health informatics;**

Additional Key Words and Phrases: Physical activity propagation, health social network

ACM Reference Format:
Nhathai Phan, Javid Ebrahimi, David Kil, Brigitte Piniewski, and Dejing Dou. 2016. Topic-aware physical activity propagation with temporal dynamics in a health social network. ACM Trans. Intell. Syst. Technol. 8, 1, Article 2 (August 2016), 20 pages.
DOI: http://dx.doi.org/10.1145/2873066

1. INTRODUCTION

Regular physical activity reduces the risk of developing cardiovascular disease, diabetes, obesity, osteoporosis, some cancers, and other chronic conditions [U.S. Department of Health & Human Services 1996]. Public health goal standards recommend that adults participate in at least 30min of moderate-intensity physical activity 5 or

This work is supported by NIH grant R01GM103309 (PI Dou). The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the NIH, or the U.S. Government.

Dejing Dou is the corresponding author.

Authors' addresses: N. Phan, J. Ebrahimi, and D. Dou, Computer and Information Sciences Department, University of Oregon, 1202 University of Oregon, Eugene, OR 97403, USA; emails: {haiphan, javid, dou}@cs.uoregon.edu; D. Kil, HealthMantic, Inc, 6715 Windrift Way, Austin, TX 78745, USA; email: david.kil@healthmantic.com; B. Piniewski, PeaceHealth Laboratories, 1115 SE 164th Avenue, Vancouver, WA 98683, USA; email: BPiniewski@peacehealthlabs.org.

DOI: http://dx.doi.org/10.1145/2873066

more days a week [Pate et al. 1995]. However, less than 50% of the adult population meets these standards in many industrialized countries [Bauman et al. 2003; U.S. Department of Health & Human Services 1996]. Thus, finding effective population-based intervention strategies to propagate physical activity is a key challenge.

The Internet is identified as an important source of health information, thus may be an appropriate delivery mechanism for health behavior interventions [Internet World Stats 2016; Marshall et al. 2005]. Widespread success of online social networks holds promise for wide-scale promotion of changes in physical activity behavior. Since 2000, a wide range of studies evaluating Internet-delivered health interventions have been reported. Over half have reported positive behavioral outcomes [Marcus et al. 2000; Vandelanotte et al. 2007]. Online social networks can help people interact and participate in various physical activities and can better promote and spread physical activities with affordable cost. However, there has been a lack of scientific and quantitative study to elucidate how social networks may contribute to physical activity propagation.

Along with online social networks, recent advances in mobile technology 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 an individual. Utilizing these technologies, our recent study, named YesiWell, conducted in 2010 to 2011 as a collaboration between PeaceHealth Laboratories, SK Telecom Americas, and the University of Oregon, recorded daily physical activities, social activities (i.e., text messages, social games, meetup events, competitions, and so on), biomarkers, and biometric measures (i.e., cholesterol, triglyceride, body mass index, and the like) for a group of 254 individuals who formed a health social network. Physical activities were reported via a mobile device carried by each user. All users enrolled in an online social network application, allowing them "friend" and communicate with each other. Biomarkers and biometric measures were recorded via daily/weekly/monthly medical tests performed at home (individually) or at the laboratories. The fundamental problems that this study seeks to answer, which are also the key in understanding the determinants of healthy behavior propagation, are as follows:

1. Does social communication affect physical activity propagation?
2. How can we leverage social communication to understand and model physical activity propagation?

For the first question, to illustrate that social communication can deliver physical activity, we have performed a simple statistical analysis on our health social network. Assume that a user u receives a message m at timestamp t from another user; we compare the total number of walking and running steps of u in the future period $[t, t + \Delta t]$ with the past period $[t - \Delta t, t]$. If u increases total number of steps, then m is considered as an effective message. The solid line in Figure 1 illustrates the probability of a message becoming effective; meanwhile the dashed line shows the probability of users increasing total number of steps when randomly choosing a timestamp t (i.e., the user might or might not receive a message at a random time t). It is clear that with $\Delta t = 1$ *day*, the probability of a user increasing one's total number of steps after receiving a message is up to 0.58 and significantly larger than the 0.26 of random t. This phenomenon remains when Δt increases to 50 days before dropping down. This evidence strengthens our belief that social communications in health social networks can help propagate physical activities.

Our goal in this article is to understand the dynamics of physical activity propagation via social communication channels at both the individual level and the community level. More concretely: (1) We aim to evaluate the probability of physical activity propagations for every social communication edge. The estimated probabilities can be used

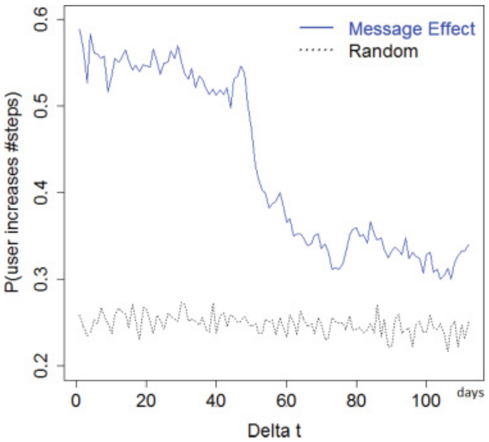


Fig. 1. Probability that a message becomes effective in propagating physical activities.

in many applications (i.e., propagation prediction, health behavior interventions, and so on); (2) we then devise a graph summarization paradigm for the analysis of physical activity propagation and social influence. In fact, we aim to find an abstraction of the propagation process that provides data analysts with a compact, yet meaningful, view of patterns of influence and activity diffusion over health social networks.

To achieve these goals, we are inspired by the well-known Independent Cascade (IC) model [Kempe et al. 2003], the Community-level Social Influence (CSI) model [Mehmood et al. 2013], and our previous Physical Activity Propagation model (CPP) [Phan et al. 2014] to fit a health social network. In this article, we mainly extend our previous work, the CPP model, by taking into account the content of social communication instead of a binary status (i.e., sent or did not send messages) between two users. In essence, a message could belong to different topics, since people discuss different things, for example, events, competitions, and physical activities. In different traces, different topics could have different correlations with individuals' social influences. To address this issue, we propose combining the number of messages, topics of messages, and individual effects into a hierarchical clustering algorithm to infer the probability of physical activity propagations at different granularities. Regarding our discovered structure, a community is identified by *a set of communicated nodes* that share a *similar physical activity influence tendency* over nodes belonging to other communities. Our proposed model is called the **T**opic-aware **C**ommunity-level **P**hysical Activity Propagation with **T**emporal Dynamics (TCPT) model, which is designed to capture the social influences for different topics and temporal dynamics of the messages in the YesiWell study. In order to clarify the effect of activity propagation upon health outcomes, we analyze the correlation between detected communities and existing health-outcome measures such as body mass index (BMI), average number of steps, Wellness score [Kil et al. 2012], and the like.

The main contributions of this paper are as follows:

1. We introduce the TCPT model, which is inspired by the ideas of CPP, IC, and CSI models.
2. Through a comprehensive experiment on the YesiWell social network, we show the effectiveness of our approaches. Our discovery potentially paves a way for knowledge discovery and data mining in health social networks, for example, physical activity interventions.

The rest of the article is organized as follows. We briefly review related prior work in Section 2. In Section 3, we formally define the problem tackled in this article and explain the technical detail of our models. The experimental evaluation is in Section 4. In Section 5, we present our conclusions, with a summary of our major findings and future research directions.

2. RELATED WORK

Regular physical activity decreases the risk of developing cardiovascular disease, diabetes, obesity, osteoporosis, some cancers, and other chronic conditions. Website-delivered physical activity interventions have the potential to overcome many of the barriers associated with traditional face-to-face exercise counseling or group-based physical-activity programs. An Internet user can seek advice at any time, any place, and often at a lower cost compared with other delivery modalities [Ritterband et al. 2003]. In 2000, a set of articles that identified the potential of interactive health communications, including Internet and website-delivered interventions, for improving health behaviors were published [Marcus et al. 2000]. Since then, over fifteen studies [Vandelanotte et al. 2007] evaluating a website-delivered intervention to improve physical activity have been reported. Better outcomes were identified when interventions had more than five contacts with participants and when the time to follow-up was short ($\leq$3 months; 60% positive outcomes), compared to medium-term (3–6 months, 50%) and long-term ($\geq$6 months, 40%) follow-up. A little over half of the controlled trials of website-delivered physical activity interventions have reported positive behavioral outcomes. However, intervention effects were short-lived, and there was limited evidence of maintenance of physical-activity changes. Although the website-delivered approaches reported positive results, research is needed to identify elements that can improve behavioral outcomes. Social networks have potential for being adopted for this purpose, since they take advantage of natural social relationships to deliver healthy behaviors. Furthermore, social networks can be a life-long environment, thus the retention of participants could be improved.

Social influence and the phenomenon of influence-driven propagations in social networks have received considerable attention in recent years. One of the key issues in this area is to identify a set of influential users in a given social network. Domingos and Richardson [2001] approach the problem with Markov random fields, while Kempe et al. [2003] frame influence maximization as a discrete optimization problem. Another line of study has focused on the problem of learning the influence probabilities on every edge of a social network, given an observed log of propagations over this network [Goyal et al. 2010; Saito et al. 2008]. In addition, many tasks in machine learning and data mining involve finding simple and interpretable models that nonetheless provide a good fit with observed data. In graph summarization, the objective is to provide a coarse representation of a graph for further analysis. Tian et al. [2008] and Zhang et al. [2010] consider algorithms to build graph summaries based on node attributes, while Navlakha et al. [2008] use the Minimum Description Length (MDL) principle [Rissanen 1983] to find good structural summaries of graphs. Mehmood et al. [2013], introduce a hierarchical approach to summarize patterns of influence in a network by detecting communities and their reciprocal influence strength. In 2007, Christakis and Fowler began to publish their research, focusing on the spread of obesity and other human behavior and health status, such as alcoholism, stress, and smoking in social networks [Christakis and Fowler 2007; Fowler and Christakis 2009; Mednick et al. 2010; Rosenquist et al. 2010]. They found that the directional nature of the effects of friendships is especially important with regard to the inter-personal induction of obesity, because friends do not simultaneously become obese as a result of contemporaneous exposures to unobserved factors.

In our most recent work on the TaCPP model [Phan et al. 2016], we integrate topics of messages but not the temporal dynamics into the propagation model. In this article, we also extend our experiments to general online social networks, for example, Yahoo! Meme, to discover more meaningful observations.

3. TOPIC-AWARE COMMUNITY-LEVEL PHYSICAL ACTIVITY PROPAGATION WITH TEMPORAL DYNAMICS MODEL

3.1. Preliminaries

We first explain how to identify a single trace when a user v influences another user u by sending a message. Assume that at time t, user v sends a message m to user u; given a Δt, v is called to *activate* u at time t if the total number of (walking and running) steps of u in $[t, t+\Delta t]$ is larger than or equal to the total number of steps of u in the past period $[t-\Delta t, t]$. Normally, the influence can be further propagated if u successfully *activates* other users at the next timestamp (i.e., $t+1$) [Kempe et al. 2003]; however, the process in health social networks is usually slower than that. Following Mathioudakis et al. [2011], Phan et al. [2014], and Mehmood et al. [2013], we circumvent this problem by using a **time window** w to define a single trace as follows: Given a chain of users $\alpha = \{U_1, \dots, U_n\}$ such that U_i is a set of users, $U_1 \cap U_2 \cap \dots \cap U_n = \emptyset$; α is called a single trace if $\forall i \in [1, n-1], \forall u \in U_{i+1}$ is activated by some user $u' \in U_i$ such that $t_\alpha(u) \in [t_\alpha(u'), t_\alpha(u') + w]$, where $t_\alpha(u)$ is the *activation time* of u in α. In real cases, U_1 can be a user instead of a set of users.

Let $G = (V, E)$ denote a directed network, where V is the set of vertices and $E \subseteq V \times V$ denotes a set of directed arcs. Each arc $(v, u) \in E$ represents an influence relationship (i.e., v is a potential influencer for u) and it is associated with a probability $p(v, u)$ that represents the strength of this influence relationship. Let $D = \{\alpha_1, \dots, \alpha_r\}$ denote a log of observed propagation traces over G. We assume that each propagation trace in D is initiated by a special node $\Omega \notin V$, which models a source of influence that is external to the network. More specifically, we have $t_\alpha(\Omega) < t(v)$ for each $\alpha \in D$ and $v \in V$. Time unfolds in discrete steps. At time $t = 0$, all vertices in V are inactive; Ω makes an attempt to activate every vertex $v \in V$ and succeeds with probability $p(\Omega, v)$. At subsequent time steps, when a node v becomes active, it makes one attempt at influencing each inactive neighbor u, who receives a message from v, with probability $p(v, u)$. Multiple nodes may try to independently activate the same node at the same time.

There are different ways to evaluate the function p. The IC model proposed by Kempe et al. [2003] can be instantiated with an arbitrary choice of p. They use a uniform probability q in their experiments; that is, $p(v, u) = q$ for all $(v, u) \in E$. On the other hand, Saito et al. [2008] estimate a separate probability $p(v, u)$ for every $(v, u) \in E$ from a set of observed traces. These two approaches can be viewed as opposite ends of a complexity scale. Using a single parameter results in a simple, but potentially low accuracy, model while estimating a different probability for each arc might provide a good fit, but at the price of risking to overfit. Next, we introduce our TCPT model to shift the modeling of influence strength from node-to-node to community-to-community. In our community-based model, all vertices that belong to the same cluster are assumed to have identical influence probabilities toward other clusters.

3.2. The TCPT Model

We start by introducing the likelihood of a single-trace α when expressed as a function of single-edge probability. This is useful to define the problem that we tackle in this article. In our health social network, there is a constraint, that is, user v *"must"* send a message m to user u at time $t_{m,\alpha}(v, u)$ in order for $t_{m,\alpha}(v, u) \in [t_\alpha(v), t_\alpha(v) + w]$ to be considered to be trying to activate user u in the trace α. Let $I_{\alpha,u}$ be the set of user u's neighbors that potentially influence u's activation in the trace α:

$$I^+_{\alpha,u} = \{v|(v,u) \in E, \exists m : t_{m,\alpha}(v,u) \in [t_\alpha(v), t_\alpha(v) + w], \textit{if and only if } u \in U_i \textit{ then } v \in U_{i-1}\}. \tag{1}$$

Similarly, we define the set of user u's neighbors who clearly failed in influencing u's activation in the trace α:

$$I^-_{\alpha,u} = \{v|(v,u) \in E, \exists m : t_{m,\alpha}(v,u) \in [t_\alpha(v), t_\alpha(v) + w], \textit{ if and only if } v \in U_{i-1} \textit{ then } u \notin U_i\}. \tag{2}$$

Let $p : V \times V \to [0, 1]$ denote a function that maps every pair of nodes to a probability. The log likelihood of the traces in D given p can be defined as

$$\log L(D|p) = \sum_{\alpha \in D} \log L_\alpha(p). \tag{3}$$

Each $v \in I^+_{\alpha,u}$, v succeeds in activating u on the considered trace α with probability $p(v,u)$ and fails with probability $1 - p(v,u)$. In addition, the content of messages is crucial to understanding physical activities of users. Given a set of topics K, each message could be related to a topic $k \in K$. In a time window w, a user v can send m messages in topic k to another user u, denoted $m_{k,v,u}$. Following Mehmood et al. [2013] and Phan et al. [2014], we define $\gamma_{\alpha,v,u,k}$ as users' effect, which represents the probability that in trace α, the activation of u was due to the success of the activation trial performed by v on topic k. In addition, inspired by Gomez-Rodriguez et al. [2014] and Gomez-Rodriguez et al. [2011], we incorporate $\gamma_{\alpha,v,u,k}$ into the temporal dynamics of the propagation process. The main idea is this: the faster the propagation process, the stronger the propagation probability. The traces are assumed to be i.i.d. By using $\gamma_{\alpha,v,u,k}$, we can define the likelihood of the observed propagation as follows:

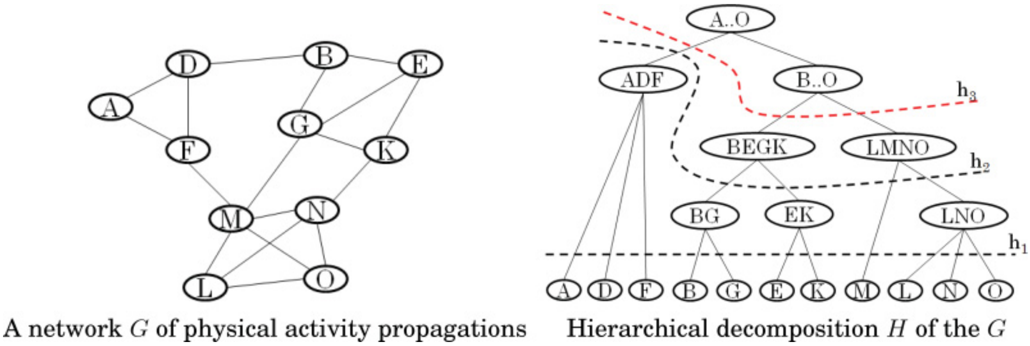
$$\begin{aligned} L_\alpha(p) \;=\; & \prod_{u \in V} \left[1 - \prod_{v \in I^+_{\alpha,u}} \left(1 - p(v,u)^{\frac{\sum_{k \in K} m_{k,v,u} \times \gamma_{\alpha,v,u,k} e^{-\gamma_{\alpha,v,u,k}(t_{m,\alpha}(v,u) - t_\alpha(v))}}{Z(\alpha,v,u)}} \right) \right] \\ & \times \left[\prod_{v \in I^-_{\alpha,u}} \left(1 - p(v,u)^{1 - \frac{\sum_{k \in K} m_{k,v,u} \times \gamma_{\alpha,v,u,k} e^{-\gamma_{\alpha,v,u,k}(t_{m,\alpha}(v,u) - t_\alpha(v))}}{Z(\alpha,v,u)}} \right) \right], \end{aligned} \tag{4}$$

where $Z(\alpha, v, u)$ is a normalization function that can be defined as follows:

$$Z(\alpha, v, u) = \sum_{v \in I^+_{\alpha,u} \cup I^-_{\alpha,u}} \sum_{k \in K} m_{k,v,u} \times \gamma_{\alpha,v,u,k} e^{-\gamma_{\alpha,v,u,k}(t_{m,\alpha}(v,u) - t_\alpha(v))}. \tag{5}$$

To shift the influence strength estimation from node-to-node to community-to-community in the TCPT model, we use a hierarchical decomposition H of the network G. More specifically, H is a *tree* with the network G as a root r, the nodes in V as leaves, and an arbitrary number of internal nodes (i.e., between the root r and the leaves $u \in V$). A cut h of H is a set of edges of H, so that, for every $v \in V$, one and only one edge $e \in h$ belongs to the path from the root r to v. Therefore, by removing all the edges in h from H, we disconnect every $v \in V$ from r.

Let C_H denote the set of all possible cuts of H. Each $h \in C_H$ results in a partition $\mathcal{P}_h$ of the network G, so that all vertices in V that are below the same edge $e \in h$ in H belong to the same cluster $c_e \subseteq V$. Let $c(u)$ denote the cluster to which the node $u \in V$ belongs to the partition $\mathcal{P}_h$. In the TCPT model, all vertices that belong to the same cluster are assumed to have identical influence probabilities on other clusters. Given a



A network G of physical activity propagations Hierarchical decomposition H of the G

Fig. 2. An example of input for the TCPT model: a graph G of physical activity propagations (each undirected edge is considered as the corresponding two directed arcs), a hierarchy H.

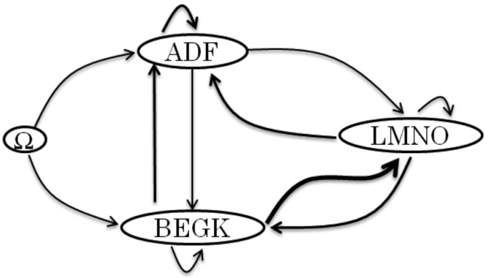


Fig. 3. A possible detected community structure resulted from the input of Figure 2 and corresponding to the cut h_3. The edge thickness represents the strength of the influence.

probability function $\hat{p}_h : \mathcal{P}_h \times \mathcal{P}_h \to [0, 1]$ that assigns a probability between any two clusters of the partition $\mathcal{P}_h$, we define

$$p_h(v, u) = \hat{p}_h(c(v), c(u)). \tag{6}$$

In the next section, we will find $\hat{p}_h$ using an expectation maximization (EM) algorithm. For the moment, we can assume that $\hat{p}_h$ is induced by h in a deterministic function since our aim is to find an optimal cut $h^* \in C_H$. In fact, a straightforward solution is the cut at the leaf level of H that maximizes the likelihood defined in Equations (3) and (4) (i.e., individual level). Reducing the number of pairwise influence probabilities used by the model can only result in a lower likelihood, but the model complexity can be simplified. That is the reason why we propose using a *model selection function* f that takes into account both likelihood and the complexity of the model.

For instance, Figures 2 and 3 illustrate an example of input and output for our approach, that is, a TCPT model. The cut h_1 corresponds to the leaf level model, in which each single node of the social graph constitutes a state of the TCPT model. This is the maximum likelihood cut that would correspond to the idea of the standard independent cascade model [Kempe et al. 2003] (i.e., the individual level). Two other cuts are also presented, in which h_2 corresponds to the clustering $\{\{A, D, F\}, \{B, G\}, \{E, K\}, \{M\}, \{L, N, O\}\}$ and the cut h_3 results in our model in Figure 3, which is the *best* model according to the model selection function f in this example.

Then, we can formally define the model-learning problem addressed in this article. Note that the network G and the hierarchy H remain fixed. The model complexity is only affected by the cut $h \in C_H$.

Definition 1 (*TCPT Model Learning*). Given a network $G = (V, E)$, a set of propagation traces D across G, a hierarchical partitioning H of G, and a model selection function f, find the optimal cut of H defined as

$$h^{*} = \arg \min_{h \in C_H} f(L(D|\hat{p}_h), h), \tag{7}$$

where $\hat{p}_h : \mathcal{P}_h \times \mathcal{P}_h \rightarrow [0, 1]$ is a probability function that assigns a probability between any two clusters of the partition $\mathcal{P}_h$.

It is interesting to note that the two extreme cases outlined earlier, that is, uniform probability, or all links have a different probability, can be modeled in our approach. The cut h_1 in Figure 2 places all vertices of G in separate clusters, which corresponds to the most complex model with a separate influence probability on every edge. The cuts h_2 and h_3 induce models with a lower granularity (i.e., community level). Finally, if there is no cut, then all vertices are in the same cluster, which results in the simplest possible model with a constant $p(v, u)$ for each edge (v, u).

3.3. Learning Intercommunity Influence and Model Selection

In this section, we propose an EM approach for estimating the pairwise influence strength among the clusters of nodes, that is, the parameters of the TCPT model. As presented before, we assume that the clusters in a partition $\mathcal{P}_h$ have been induced by a cut h of a given hierarchical decomposition H of G. However, the EM method presented in this section can be applied to an arbitrary disjoint partition of V. Remember that $c(u)$ denotes the cluster to which u belongs, and let $C(x) \subseteq V$ denote the set of vertices that belong to cluster $x \in \mathcal{P}_h$.

According to the discrete-time independent cascade model [Kempe et al. 2003], given a single trace α, at least one of user $v \in I^{+}_{\alpha,u}$ was successful in delivering physical activities to user u independently, but we do not know which one. As discussed before, by using users' effects $\gamma_{\alpha,v,u,k}$, we can define the complete expectation log likelihood of the observed propagation as follows:

$$\begin{aligned} &Q(\hat{p}_h, \hat{p}_h{}^{old}) \\ &= \sum_{\alpha \in D} \sum_{u \in V} \Bigg\{ \sum_{v \in I^{+}_{\alpha,u}} \Bigg[\frac{\sum_{k \in K} m_{k,v,u} \times \gamma_{\alpha,v,u,k} e^{-\gamma_{\alpha,v,u,k}(t_{m,\alpha}(v,u) - t_{\alpha}(v))}}{Z(\alpha, v, u)} \log \hat{p}_h(c(v), c(u)) \\ &\quad + \Bigg(1 - \frac{\sum_{k \in K} m_{k,v,u} \times \gamma_{\alpha,v,u,k} e^{-\gamma_{\alpha,v,u,k}(t_{m,\alpha}(v,u) - t_{\alpha}(v))}}{Z(\alpha, v, u)}\Bigg) \log(1 - \hat{p}_h(c(v), c(u)))\Bigg] \\ &\quad + \sum_{v \in I^{-}_{\alpha,u}} \log(1 - \hat{p}_h(c(v), c(u)))\Bigg\}, \end{aligned} \tag{8}$$

where $\hat{p}_h{}^{old}$ means the probability of the previous partition. Assume that we have an estimate of every $\gamma_{\alpha,v,u,k}$; we can determine the $\hat{p}_h$ that maximizes Equation (8) by solving $\frac{\partial Q(\hat{p}_h, \hat{p}_h{}^{old})}{\partial \hat{p}_h(x,y)} = 0$ for all pairs of clusters $x, y \in \mathcal{P}_h$. This gives the following estimate of $\hat{p}_h$(x, y):

$$\hat{p}_h(x, y) = \frac{1}{S_{x,y}} \sum_{\alpha \in D} \sum_{u \in C(y)} \sum_{v \in I^{+}_{\alpha,u} \cap C(x)} \sum_{k \in K} m_{k,v,u} \times \gamma_{\alpha,v,u,k} e^{-\gamma_{\alpha,v,u,k}(t_{m,\alpha}(v,u) - t_{\alpha}(v))}, \tag{9}$$

where

$$S_{x,y} = \sum_{u\in C(y)}\sum_{k\in K}\sum_{z\in(I^+_{\alpha,u}\cup I^-_{\alpha,u})\cap C(x)} m_{k,z,u} \times \gamma_{\alpha,z,u,k} e^{-\gamma_{\alpha,z,u,k}(t_{m,\alpha}(z,u)-t_\alpha(z))}. \tag{10}$$

Next, we need to provide an estimate for every $\gamma_{\alpha,v,u,k}$. We do this based on the assumption that the probability distributions $\gamma_{\alpha,v,u,k}$ are independent of the partition $\mathcal{P}$. If v is believed to influence u on topic k in the trace α, this belief should not change across different ways of clustering the two nodes. Therefore, we estimate $\gamma_{\alpha,v,u,k}$ from the model in which every $u \in V$ belongs to its own cluster, since this results in simplified estimates that only depend on the network structure. By denoting this model as $\hat{p}_o$, we obtain the following estimation of $\gamma_{\alpha,v,u,k}$:

$$\gamma_{\alpha,v,u,k} = \frac{m_{k,v,u}\hat{p}_o(v,u)}{\sum_{z\in I^+_{\alpha,u}\cup I^-_{\alpha,u}}\sum_{k\in K} m_{k,z,u}\hat{p}_o(z,u)}. \tag{11}$$

Our learning method for the TCPT model is as follows:

1. Run the EM algorithm without imposing a clustering structure to estimate $\hat{p}_o(v,u)$ for all arcs $(v,u) \in E$. Note that the estimate of $\hat{p}_o(v,u)$ is:

$$\hat{p}_o(v,u) = \sum_{\alpha\in D}\frac{\sum_{k\in K} m_{k,v,u} \times \gamma_{\alpha,v,u,k} e^{-\gamma_{\alpha,v,u,k}(t_{m,\alpha}(v,u)-t_\alpha(v))}}{\sum_{z\in I^+_{\alpha,u}\cup I^-_{\alpha,u}}\sum_{k\in K} m_{k,z,u} \times \gamma_{\alpha,z,u,k} e^{-\gamma_{\alpha,z,u,k}(t_{m,\alpha}(z,u)-t_\alpha(z))}}.$$

   Repeat the two following steps until convergence.
   Step 1: Estimate each successful probability $\hat{p}_o$.
   Step 2: Update each influence effect $\gamma_{\alpha,v,u,k}$ by using Equation (11).
2. After obtaining $\gamma_{\alpha,v,u,k}$, keep $\gamma_{\alpha,v,u,k}$ fixed for different partitions $\mathcal{P}_h$, and update $\hat{p}_h(x,y)$ according to Equation (9).

We have already presented our learning method to maximize the log likelihood $L(D|p_h)$ at the individual level and given a partition $\mathcal{P}_h$. Recall that the log likelihood is maximized for the cut h that places every node in its own cluster. Thus, we need an approach to address the trade-off between model accuracy and model complexity. In this work, we utilize the *Bayesian Information Criterion* (BIC) [Schwarz 1978] as a selection function f in Equation (7). In statistics, the BIC is a criterion for model selection among a finite set of models:

$$BIC = -2\log L(D|p_h) + |h|\log(|D|), \tag{12}$$

where $|h|$ is the number of intercommunity influences $\hat{p}_o(x\ ,y)$ that we need to estimate and $|D|$ is the number of traces in D.

Finally, we can evaluate different cuts $h \in C_H$ of the hierarchical decomposition of the network. We utilize the heuristic bottom-up greedy algorithm proposed in [Mehmood et al. 2013] to report the best solution found as output given the hierarchical decomposition H. In each iteration, the algorithm finds out the two best communities to merge and to update the model. The resulting cut as well as the corresponding parameters are stored in the set C. Once the algorithm reaches H's root, it evaluates the objective function for every cut in C and returns the one having the best value.

4. EXPERIMENTS

In this section, we use the real-world YesiWell data and the corresponding social network to empirically validate the effectiveness of our proposed model. We first elaborate on the experiment configurations on the dataset, then introduce the experimental results.

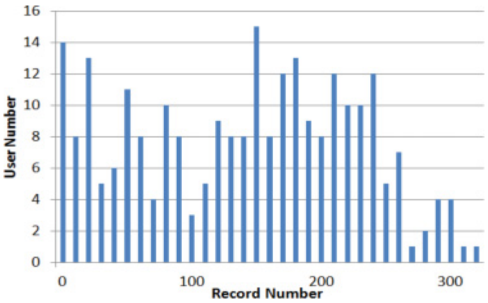


Fig. 4. Distribution of the record number and user number in YesiWell data.

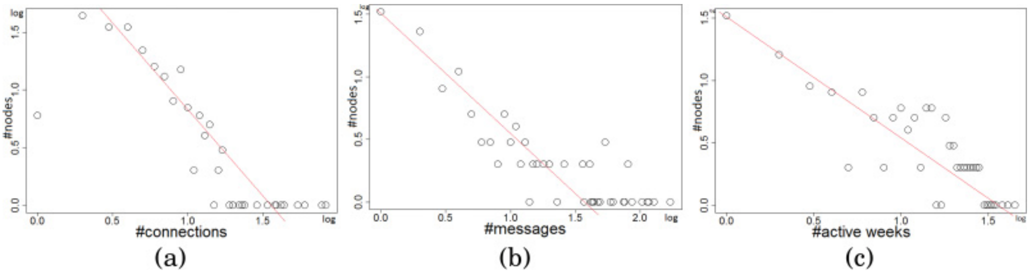


Fig. 5. The distributions of friend connections (a), inbox messages (b), and active users (c) in the YesiWell study.

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

The initial physical-activity data, collected from each user via special electronic equipment for each user, includes information on the number of walking and running steps. Since some users' daily records are missing, we show the basic analysis on the distribution of physical-activity record numbers in Figure 4. In Figure 4, there are 14 users with their daily physical-activity record number smaller than 10, and 8 users with their record number larger than 10 but smaller than 20. Thus, to clean the data, we filtered the users whose daily physical-activity record number is smaller than 80. In addition, we only consider users who contribute to the social communication, that is, users must send (resp., receive) messages to (resp., from) other users. Finally, we have 123 users with 2,766 inbox messages for our experiments. Figure 5 illustrates the distributions of friend connections and the number of received messages in the health social network. They clearly follow the Power law distribution. In addition, 90% of 254 users are non-Hispanic White or Latino origin. They might or might not know each other before the study. A total of 254 overweight and obese individuals volunteered to join the study. Therefore, they are not under any pressure to do more or less exercise.

Experiment Setting. Our proposed model requires input as a hierarchical decomposition of the network. Following Mehmood et al. [2013], we obtain this hierarchy by recursively partitioning the underlying network using METIS [Karypis and Kumar 1998], which reportedly provides high-quality partitions. The delay threshold Δt and the time window w are set to a day and a week, respectively. Finally, we performed the *Latent Dirichlet Allocation* model (LDA) [Blei et al. 2003] on text messages in the

Table I. Topic Description of the Messages in YesiWell Data

Technical	Physical Activity	General	Program-Social Activity
hpod	day	weight	competition
step	step	don	find
today	work	food	week
day	walk	good	don
computer	week	life	program
time	back	work	goal
goal	run	love	david

YesiWell dataset to extract the underlying topics in users' messages. We found 4 coherent major topics in the messages that are *technical* related, *physical-activity* related, *program-social activity* related, and one overlapping topic, which is called the *general* topic. More specifically, after removing the stop words, each message is treated as an item in the LDA model. The hyper-parameters α and β are set to 0.5 and 0.01, respectively. By performing Gibbs sampling for inference, we can eventually find the mixture proportion for each message [Heinrich 2004]. For more clarification on how we distinguish topics, we mention some of the keywords in each topic (Table I). In the technical topic, we have words such as hpod, upload, data, computer, and recorded. "Hpod" means health pod, which was designed to record activities and to serve as both a storage mechanism and a gateway for other health data, such as blood pressure, weight, body fat, and pulse rate, captured through other Bluetooth-enabled devices. Many of these messages were questions about technical difficulties that the users faced with the device. In the physical-activity topic, words such as work, walk, walking, running, workout, and treadmill are mentioned. One of the main features of the program is its social aspect. For instance, patients in different teams join competitions on their physical activity during a period of time. Because of words such as program, goals, progress, join, competitions, team, friends, and weeks, we call the fourth topic the program-social related topic. This topic includes messages in which users are discussing their progress in the prior weeks or results of the competitions. Note that "step" appears in both Technical and Physical Activity because (1) The users have some issues with the mobile device, "hpod," which tracks the number of walking/jogging/running "steps." Users send messages to ask how to use their hpods to record their steps, how the hpods will record their steps, and so on. When they experience technical difficulties with their hpods, they may lose their record numbers, for example, the data regarding the number of steps they have walked or run that day might be lost. They also send messages to other people to learn how to recover their data. Losing records of steps or not being able to record steps due to technical problems has direct effects on their "goals," as those goals are generally a target number of steps that they aspire to achieve in a given time interval. Therefore, the word "goal" also appears in the Technical topic. The keywords in the third topic are a bit diverse to be assigned a distinct topic. Thus, we call it the general topic. We ran our experiments on an Intel i7 2.8GHz processor with 4GB memory.

4.1. Experimental Results

An effective way of summarizing influence relationships in the network is to consider the community-level influence propagation network. In Figure 6, we show the networks of physical-activity propagations detected by the TCPT model for our dataset. The node size is the average number of steps for all users in their community. The arrow head size is proportional to the probability of physical-activity influence. The shapes will be described later. Note that we consider only the arcs that have probabilities larger than

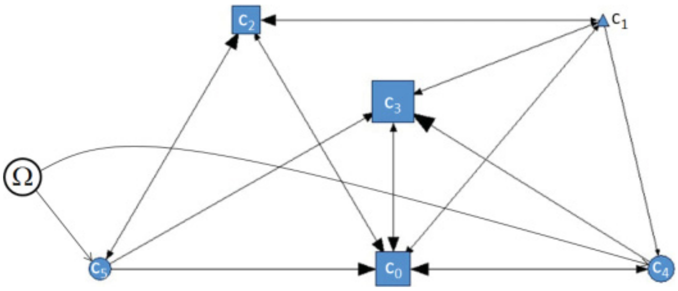


Fig. 6. Detected community structure in YesiWell data.

0.25. It is very interesting, since the network is almost acyclic; this suggests a clear directionality pattern in the flow of physical activities. Moreover, with the models, we are able to categorize the detected communities into three kinds of groups based on their influence behaviors, as follows:

(1) Influencer: This group can be seen as *circle nodes* in Figure 6. These nodes have the strongest influence probability to propagate physical activities to other users in other communities. In addition, they receive almost no physical activity propagation from other communities.
(2) Influenced users: This group can be seen as *rectangle nodes* in Figure 6. These nodes are easily influenced by influencers (i.e., by circle nodes) since they receive the physical-activity propagation with high propagation probabilities. Moreover, the average number of steps of these nodes is quite large, even larger than the influencer nodes. These influenced users sometimes try to propagate physical activities to other communities, but not much.
(3) Noninfluenced users: This group can be seen as *triangle nodes* in Figure 6. It is very hard for these nodes to be influenced, since they receive very small probabilities of physical activity propagations from other groups. In addition, the average number of steps of the *noninfluenced nodes* is small compared with the other mentioned kinds of nodes.

The effectiveness of our approach can be validated by exploring the differences among these three user categories in terms of behaviors, lifestyles, and health outcomes to explain why they have such physical-activity propagation behaviors. We will illustrate the varying of health outcome measures (i.e., BMI, number of steps, Wellness score [Kil et al. 2012]) over time for the three groups. Note that in the next experiments, all the users in the same category will be gathered together; thus, we will have only three groups of users instead of the six detected communities represented in Figure 6.

Physical activity record number. Figure 7 illustrates the average number of steps for the three groups over time. We can see that users in the influencer group not only have the best average BMI value, but also are stable in doing exercises day by day (i.e., a good lifestyle) from the beginning to the end of the study. This clarifies the activity-delivering role of the influencer group. Regarding the influenced user group, they performed fewer physical activities at the beginning (i.e., at the middle of November, 2010), but after that, they had rapidly increased activities, even more than the influencer group. Interestingly, their activity performance is stabilized, along with that of the influencer group, until the end of the program. It appears clear that the influencer group has been successful in propagating physical activities to the influenced user group.

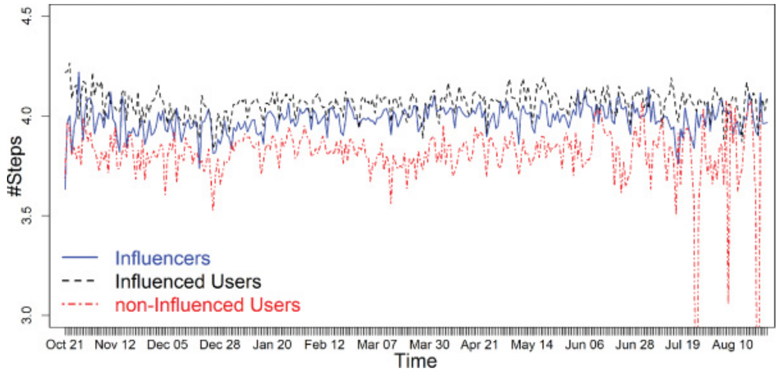
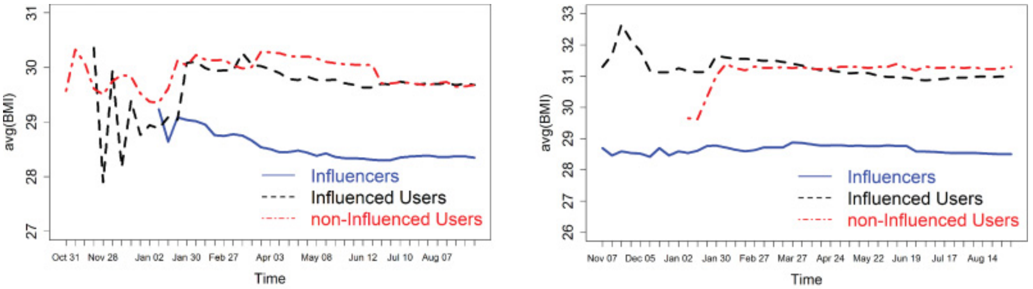


Fig. 7. Average steps for all users in the three kinds of communities: influencer, influenced users, and noninfluenced users.



(a) CPP model [Phan et al. 2014]

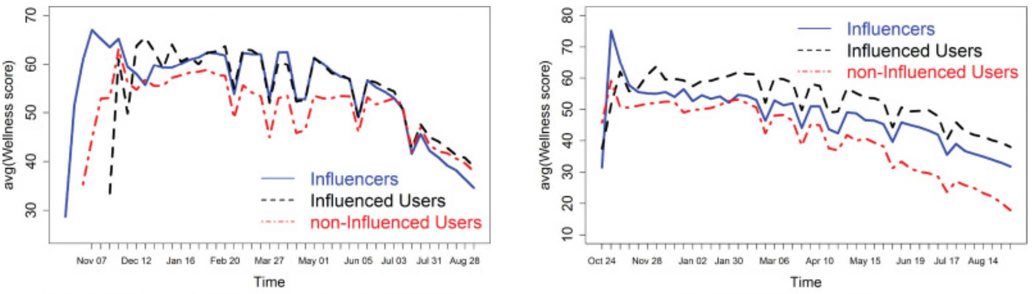
(b) TCPT model

Fig. 8. Average BMI for the three user categories.

Regarding the noninfluenced user group, there is no big change in their physical-activity behaviors. They have the lowest activity performance, which usually fluctuates throughout the whole program lifetime. It is only a short period (i.e., January to March, 2011) in which they have a quite stable (but the lowest) activity performance. Consequently, it is hard to improve the exercise behavior of the noninfluenced user group via social communications.

BMI. Figure 8 illustrates the average and standard deviation of BMI for the three groups (i.e., influencers, influenced users, and noninfluenced users). Interestingly, the influencer group has an average and standard deviation of BMI significantly lower than the other two groups. Since the purpose of participants who enrolled in this study is to reduce their BMIs, the influencer group can potentially be their external motivation. That is one of the reasons to explain why the influencer group has a strong influence probability upon other groups. In addition, in Figure 8(b) we can recognize that the influenced users have higher BMIs than the noninfluenced users at the beginning. However, they reduce their BMIs to be better than the noninfluenced users. Meanwhile, the noninfluenced users have almost the highest average and standard deviation of BMI (Figures 8 and 10). They even have quite similar or even better BMI values than the influenced user group at the beginning.

Wellness score. We have illustrated the correlation between the TCPT model results and health-outcome measures, such as BMI and the exercise activity record number independently discussed earlier. However, these individual measures cannot reflect the actual user health status, which is a complex combination of a user's lifestyle, biometrics, and biomarkers. Our proposed Wellness score [Kil et al. 2012] is a such metric. In



(a) CPP model [Phan et al. 2014]

(b) TCPT model

Fig. 9. Average Wellness score for the three user categories.

essence, Wellness score is a composite score of one's health based on lifestyle parameters, biometrics, and biomarkers. Lifestyle parameters encompass: physical activities measured in steps per minute, self-reported lifestyle parameters, the number of goals set and achieved, and social activities in terms of the size of and communications within one's social network, creation of and participation in competitions and social games, and public/private feed activities within our social network.

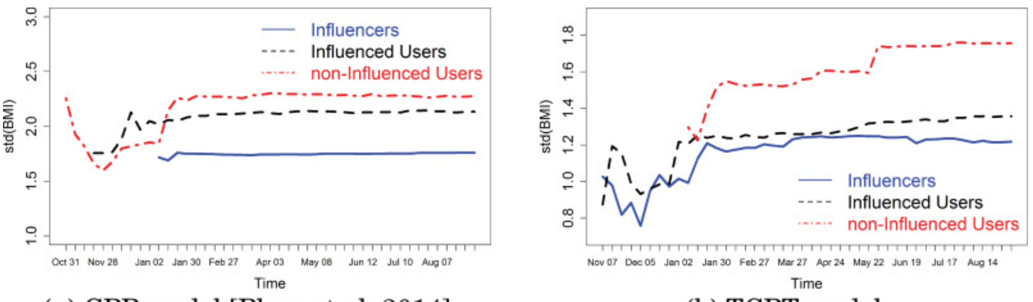
The biometric and biomarker component scores are based on a combination of utility functions (i.e., BMI vs. mortality, triglyceride/HDL vs. health risk, LDL vs. health risk, HbA1c vs. diabetes risk level, and so on) and correlation functions between BMI and biomarkers. In short, one's component risk score $y = \beta_1 U(BMI) + \beta_2 \rho_1 U(TG/HDL) + \beta_3 \rho_2 U(LDL) + \beta_4 \rho_3 U(HbA1c)$, where β is component weight, $U(.)$ is a specific utility function associated with the component in parentheses, and ρ is the correlation coefficient between BMI and the selected biomarker component. Lifestyle component score is based on a heuristic weighted combination of the number of steps per day, intensity of steps based on estimated speed, and various social activity–derived features highly associated with future weight loss [Kil et al. 2012].

Finally, raw Wellness scores are computed over multiple participants through Markov Chain Monte Carlo sampling in an attempt to remap the raw scores such that remapped scores mimic percentile ranking. For instance, a Wellness score of 90 means 90% ranking (i.e., top 10%). We also apply some boosting at the bottom, so that people do not become too discouraged when their scores are too low.

Figure 9 illustrates the Wellness score for the three user groups. It is quite clear that the influencer group always has a high Wellness score. In addition, the influenced user group has a big change in their scores. In fact, the influenced users have a low score at the beginning, but after that they had increased their scores to be among the highest ones. Meanwhile, the noninfluenced user group has the lowest score, even though they had a better starting point than the influenced user group.

Community consistency. Interestingly, in Figure 10 and Figure 11, the standard deviations of the BMI and Wellness score are quite small. In fact, for the CPP model [Phan et al. 2014], the ranges of BMI and Wellness score standard deviations of the detected communities are [1.5, 2.5] and [3, 5] (Figures 10(a), 11(a), resp.). The results of the TCPT model are even better since the ranges of BMI and Wellness-score standard deviations are reduced to [0.7, 1.7] and [2, 5]. Furthermore, they are quite stable (i.e., no big changes) for all three user groups. Thus, not only the health-outcome measures but also the lifestyles and physical-activity record numbers are quite consistent among the users in the same communities.

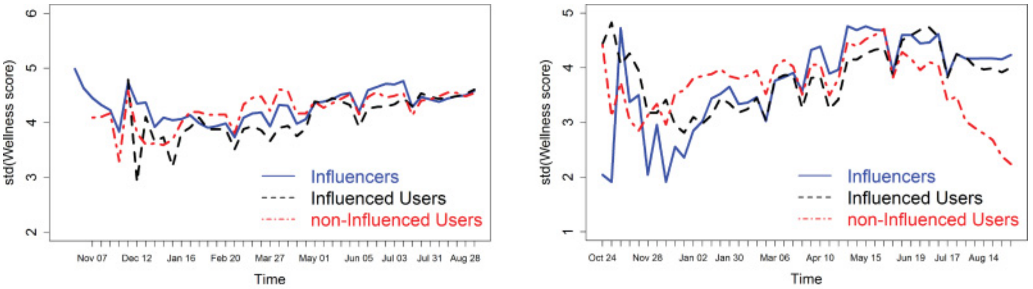
TCPT model versus CPP model. It is interesting to show that the communities detected by the TCPT model illustrate clearer behaviors. Regarding our previous CPP



(a) CPP model [Phan et al. 2014]

(b) TCPT model

Fig. 10. Standard deviation of BMI for the three user categories.



(a) CPP model [Phan et al. 2014]

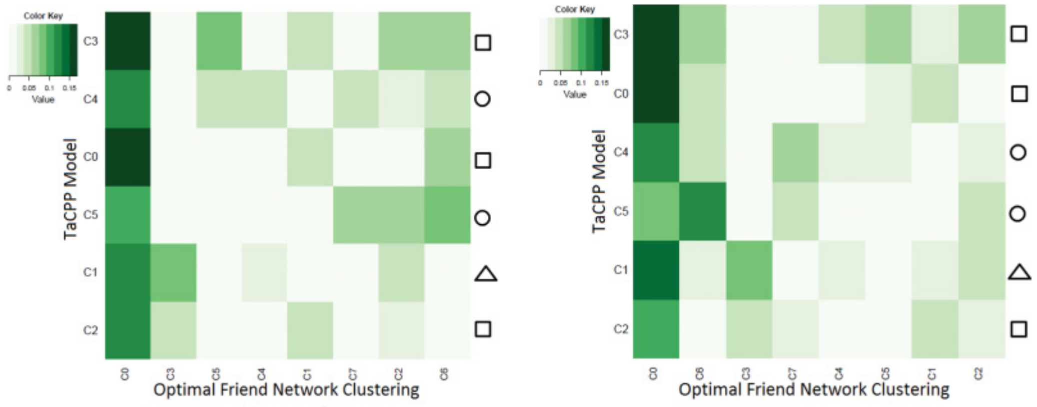
(b) TCPT model

Fig. 11. Standard deviation of Wellness score for the three user categories.

model [Phan et al. 2014], we only can distinguish the influencers in Figure 8(a) and the noninfluenced users in Figure 9(a). It is difficult to clarify the behaviors of the other user categories detected by the CPP model. Fortunately, the TCPT model results in a better community structure, which offers a more insightful pattern of user influences. It is very easy to delineate the three user categories via their behaviors in Figures 8(b) and 9(b) compared with the ones in Figures 8(a) and 9(a). In addition, the communities detected by the TCPT model are more consistent than the ones detected by the CPP model. The ranges of BMI and Wellness score standard deviations of the detected communities are [0.7, 1.7] and [2, 5], respectively, for the TCPT model. Meanwhile, the corresponding ranges for the CPP model are [1.5, 2.5] and [3, 5], respectively (Figures 10, 11).

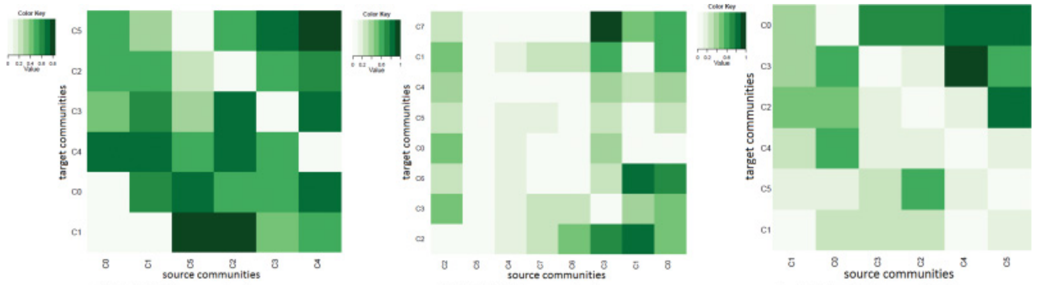
We can conclude that the CPP and TCPT models have strong correlations with health outcomes that is very meaningful toward the design of physical-activity interventions through health social networks. By incorporating the topics of the messages, the TCPT model reveals a better community structure in terms of physical-activity propagation compared with the CPP model in the YesiWell social network.

The TCPT model versus social link clustering. The outputs of the TCPT model can be graphically represented to analyze the influence probability between two communities and social link relationships. An effective way is to plot the corresponding heat maps, as shown in Figure 12. In these figures, we plot the Jaccard similarity in terms of number of steps and Wellness score between the TCPT model and obtained clusters by clustering the social network links using the METIS algorithm [Karypis and Kumar 1998]. Note that the clustering algorithm maximizes the high correlation within cluster and low between cluster. Given two clusters A and B, the Jaccard



(a) Number of steps (b) Wellness score

Fig. 12. TCPT model versus social link based on health outcome. The markers correspond to the three user categories in Figure 6(b): squares are influenced users; circles are influencers; triangles are noninfluenced users.



(a) CSI model (b) CPP model (b) TCPT model

Fig. 13. CSI, CPP, and TCPT models on our health social network data.

similarity is computed as follows:

$$J(A, B, steps) = \frac{\sum_{u \in A \cap B} u.steps}{\sum_{u \in A \cup B} u.steps}, \tag{13}$$

where *u.steps* is the total number of steps reported by user *u*. We use a similar equation for $J(A, B, Wellness\text{-}score)$.

In general, we discover almost no correlation between the TCPT model and the social-link clustering. Five out of eight detected communities in the TCPT model are found almost in the cluster 0, which is the densest cluster in our friend network. Thus, applying a normal clustering algorithm on social-network links cannot discover the communities obtained by the TCPT model.

TCPT model versus CPP [Phan et al. 2014] and CSI models [Mehmood et al. 2013]. To highlight the advantage of our TCPT model, we further compare our results with the CPP and CSI models. We applied both model selection functions MDL [Rissanen 1983] and BIC proposed in a CSI model. The former function generates only one community, while we observe 6 communities from the latter function. In Figure 13, we plot the intensity of the influence probability between two communities observed from the CSI model (BIC model selection function), the CPP model, and the TCPT model. In the CPP model (Figure 13(b)), the influence role of the communities c_0, c_1, and c_3 can be clearly seen, while c_7, c_6, and c_2 receive strong influence probabilities.

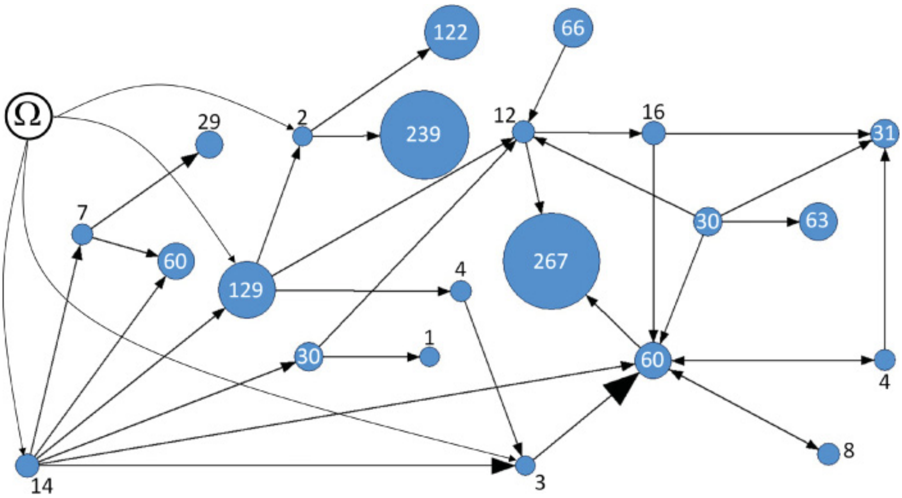


Fig. 14. The propagation network found in the Yahoo! Meme dataset.

Furthermore, c_4 and c_5 do not contribute much to the process. In Figure 13(c), the communities c_4 and c_5 have strong influence roles, while c_0, c_2, and c_3 receive strong influence probabilities. The community c_1 is not involved much in the influence process.

It is not clear how to distinguish the differences between the communities observed by the CSI model. Also, the probability range in the CSI model is [0, 0.7] smaller than the range in our TCPT model. The reason might be that our model is designed for a health social network, and we do not take into account users who clearly fail to influence others. In contrast, the CSI model does consider that.

YesiWell Health Social Network versus Yahoo! Meme. Our model is a little different from traditional information propagation models in online social networks such as Yahoo! Meme and Twitter. This is because the way we identify a user u who is trying to activate user v is different from traditional information propagation models in online social networks. In our health social network, there is a constraint, which is that user v *"must"* send a message m to user u at time $t_{m,\alpha}(v,u)$ in order for $t_{m,\alpha}(v,u) \in [t(v), t(v)+w]$ to be considered trying to activate user u in trace α. In traditional information propagation models in online social networks [Mehmood et al. 2013; Saito et al. 2008], user v does not need to send a message m to user u to be considered to be trying to activate v; user v only needs to be a friend of user u, or for u to be a follower of v [Mehmood et al. 2013; Saito et al. 2008]. By relaxing this constraint, our model can be applied in the general sense,that is, information propagation modeling in online social networks. Figure 14 illustrates the propagation network found in the Yahoo! Meme dataset, in which node size is proportional to community size, and arrow size is proportional to influence probability. Yahoo! Meme is a microblogging service, in which users can share different kinds of information, called "memes." Memes are shared on the main user's stream and a re-post button allows a user to display an item from another user's stream on one's personal stream. If the user v posts a meme that is later re-posted by the user u, we say that the meme propagates from v to u; thus, v is a potential influencer of u. The dataset contains 9,523 nodes, 759,369 links, 552,732 activations, and 9,578 traces.

We preprocessed the graph by pruning all the edges between communities that have influence probability less than 0.01. The propagation network is almost acyclic, which is consistent with state-of-the-art works [Mehmood et al. 2013]. In addition, the

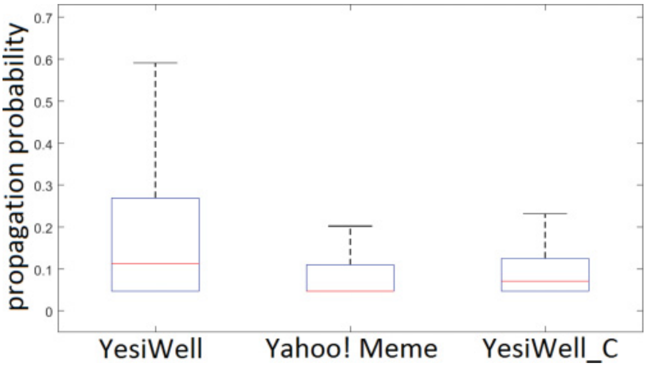


Fig. 15. The propagation rates found in the YesiWell health social network and Yahoo! Meme dataset.

propagation structures in our health social network and Yahoo! Meme are similar. Similar results are reported by Mehmood et al. [2013]. Even though the propagation structures are similar, the propagation rate in our health social network is greater than in Yahoo! Meme. Figure 15 illustrates the distributions of propagation probabilities among users in our health social network and Yahoo! Meme. To highlight the difference between our health social network and the Yahoo! Meme, we also relax the aforementioned constraint before applying our TCPT model on the YesiWell health social network, denoted YesiWell_C. We can see that the propagation rate in our health social network is greater than the propagation rate in Yahoo! Meme. This result is statistically significant, that is, p value = 1.5276e-08 performed by t-test. There is no statistically significant difference between the propagation rate in YesiWell_C and Yahoo! Meme. This illustrates that if we relax the constraint, then it is difficult to detect the propagation relationship among users in our health social network. The key reason is that, with the constraint relaxed, the propagation process is nolonger based on social communication. The consistency between this result and the analysis in Figure 1 strengthens our belief about the important role of social communication in physical-activity propagation in health social networks.

5. CONCLUSIONS AND FUTURE WORK

In this article, we introduce a hierarchical approach to analyze physical-activity propagation through social communications at the community level (which also can be applied to the individual level). Our proposed CPP and TCPT models offer a more compact representation of the network of propagations, especially the novel TCPT model. Furthermore, our novel TCPT model can be easily plotted and exploited to understand and detect interesting properties in the information propagation flow over the network. Our empirical analysis over a real-world health social network emphasizes five meaningful observations: (1) Social networks have great potential to propagate physical activities via social communications; (2) the propagation networks found in a health social network by our models are almost acyclic; (3) the physical activity–based influence behavior has a strong correlation to health-outcome measures such as BMI, lifestyles, and our proposed Wellness score; (4) the propagation rate in our health social network is greater than the propagation rate in Yahoo! Meme; and (5) the TCPT model is more effective than the CPP model.

Since online social networks have been utilized productively in recent years, our first observation paves an early road toward a new, promising, and perhaps most effective way to propagate physical activities to a wide population. Meanwhile, our second observation offers interesting insights, as it shows the existence of a clear direction in the propagation of physical activities. That is useful for physical-activity intervention approaches to design more effective strategies. Our third observation can be applied to categorize users or to predict users' macro-activities based on their influence behaviors [Shen et al. 2012]. Our fourth observation illustrates the important role of understanding the information exchanged among the users in the social network. Capturing the information at a descriptive level potentially improves the effectiveness of the whole system. In the near future, we are going to clarify the correlation between the physical-activity propagation via social communications and a corresponding friend network. The homophily principle is important to propagate healthy behavior on health social networks [Christakis and Fowler 2007]. Therefore, by discovering the correlation between the homophily effect and social communications, we can have a complete picture. As a result, we will be able to build better human behavior-prediction models and physical-activity intervention approaches.

ACKNOWLEDGMENTS

We are grateful to Xiao Xiao and Rebeca Sacks for their contributions to data processing.

REFERENCES

A. Bauman, T. Armstrong, J. Davies, N. Owen, W. Brown, B. Bellew, and P. Vita. 2003. Trends in physical activity participation and the impact of integrated campaigns among Australian adults, 1997–99. *Australian and New Zealand Journal of Public Health* 27, 1, 76–9.

D. M. Blei, A. Y. Ng, and M. I. Jordan. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research* 3, 993–1022.

N. A. Christakis and J. H. Fowler. 2007. The spread of obesity in a large social network over 32 years. *New England Journal of Medicine* 357, 370–9.

P. Domingos and M. Richardson. 2001. Mining the network value of customers. In *Proceedings of KDD'01*. 57–66.

J. H. Fowler and N. A. Christakis. 2009. Dynamic spread of happiness in a large social network: longitudinal analysis of the Framingham Heart Study social network. *BMJ: British Medical Journal* 23–27.

M. Gomez-Rodriguez, D. Balduzzi, and B. Schölkopf. 2011. Uncovering the temporal dynamics of diffusion networks. In *ICML'11*.

M. Gomez-Rodriguez, J. Leskovec, D. Balduzzi, and B. Schölkopf. 2014. Uncovering the structure and temporal dynamics of information propagation. *Network Science* 2, 1, 26–65.

A. Goyal, F. Bonchi, and L. V. S. Lakshmanan. 2010. Learning influence probabilities in social networks. In *Proceedings of WSDM'10*. 241–250.

G. Heinrich. 2004. *Parameter Estimation for Text Analysis*. Technical Report. vsonix GmbH and University of Leipzig, Leipzig, Germany.

Internet World Stats. 2016. Internet Users in the World by Regions November 2015. Retrieved June 29, 2016 from http://www.internetworldstats.com/stats.htm.

G. Karypis and V. Kumar. 1998. A fast and high quality multilevel scheme for partitioning irregular graphs. *SIAM Journal on Scientific Computing* 20, 1, 359–392.

D. Kempe, J. M. Kleinberg, and E. Tardos. 2003. Maximizing the spread of influence through a social network. In *Proceedings of KDD'03*. 137–146.

D. Kil, F. Shin, B. Piniewski, J. Hahn, and K. Chan. 2012. Impacts of social health data on predicting weight loss and engagement. In *O'Reilly StrataRx Conference, San Francisco, CA*.

B. H. Marcus, C. R. Nigg, D. Riebe, and L. H. Forsyth. 2000. Interactive communication strategies: Implications for population-based physical activity promotion. *American Journal of Preventive Medicine* 19, 2, 121–6.

A. Marshall, E. G. Eakin, E. R. Leslie, and N. Owen. 2005. Exploring the feasibility and acceptability of using Internet technology to promote physical activity within a defined community. *Health Promotion Journal of Australia* 2005, 16, 82–4.

M. Mathioudakis, F. Bonchi, C. Castillo, A. Gionis, and A. Ukkonen. 2011. Sparsification of influence networks. In *Proceedings of KDD'11*. 529–537.

S. C. Mednick, N. A. Christakis, and J. H. Fowler. 2010. The spread of sleep loss influences drug use in adolescent social networks. *PloS One* 5, 3, e9775.

Y. Mehmood, Nicola Barbieri, F. Bonchi, and A. Ukkonen. 2013. CSI: Community-level social influence analysis. In *Proceedings of ECML-PKDD'13*. 48–63.

S. Navlakha, R. Rastogi, and N. Shrivastava. 2008. Graph summarization with bounded error. In *Proceedings of SIGMOD'08*. 419–432.

R. R. Pate, M. Pratt, S. N. Blair, et al. 1995. Physical activity and public health. A recommendation from the Centers for Disease Control and Prevention and the American College of Sports Medicine. *Journal of the American Medical Association* 273, 5, 402–7.

N. Phan, D. Dou, X. Xiao, B. Piniewski, and D. Kil. 2014. Analysis of physical activity propagation in a health social network. In *Proceedings of CIKM'14*. 1329–1338.

N. Phan, J. Ebrahimi, D. Kil, B. Piniewski, and D. Dou. 2016. Topic-aware physical activity propagation in a health social network. *IEEE Intelligent Systems* 31, 1, 5–14.

J. Rissanen. 1983. A universal prior for integers and estimation by minimum description length. *The Annals of Statistics* 14, 5, 416–431.

L.M. Ritterband, L. A. Gonder-Frederick, D. J. Cox, A. D. Clifton, R. W. West, and S. M. Borowitz. 2003. Internet interventions: In review, in use, and into the future. *Professional Psychology: Research and Practice* 34, 527–34.

J. N. Rosenquist, J. Murabito, J. H. Fowler, and N. A. Christakis. 2010. The spread of alcohol consumption behavior in a large social network. *Annals of Internal Medicine* 152, 7, 426–433.

K. Saito, R. Nakano, and M. Kimura. 2008. Prediction of information diffusion probabilities for independent cascade model. In *Proceedings of KES'08*. 67–75.

G. Schwarz. 1978. Estimating the dimension of a model. *The Annals of Statistics* 6, 2, 461–464.

Y. Shen, R. Jin, D. Dou, N. Chowdhury, J. Sun, B. Piniewski, and D. Kil. 2012. Socialized Gaussian process model for human behavior prediction in a health social network. In *Proceedings of ICDM'12*. 1110–1115.

Y. Tian, R. A. Hankins, and J. M. Patel. 2008. Efficient aggregation for graph summarization. In *Proceedings of SIGMOD'08*. 567–580.

U.S. Department of Health & Human Services. 1996. Physical activity and health: A report of the surgeon general. Atlanta GA: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion.

C. Vandelanotte, K. M. Spathonis, E. G. Eakin, and N. Owen. 2007. Website-delivered physical activity interventions: A review of the literature. *American Journal of Preventive Medicine* 33, 1, 54–64.

N. Zhang, Y. Tian, and J. M. Patel. 2010. Discovery-driven graph summarization. In *Proceedings of ICDE'10*. 880–891.

Received July 2015; revised November 2015; accepted December 2015