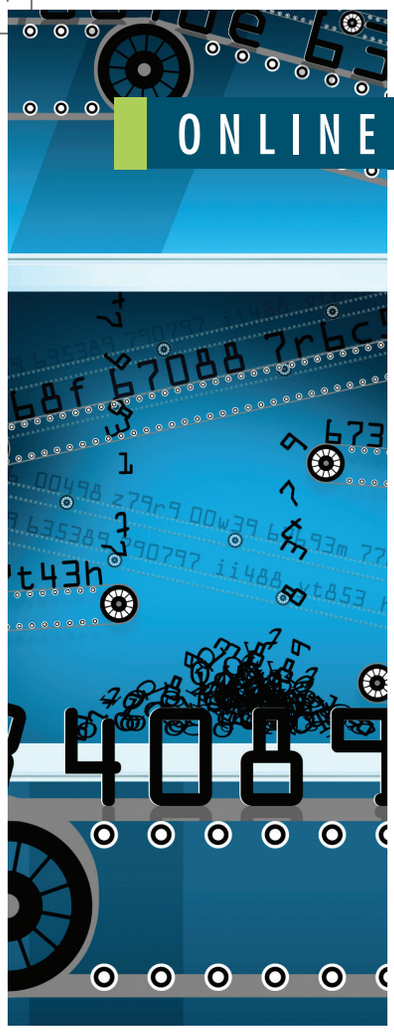

Topic-Aware Physical Activity Propagation in a Health Social Network

Nhathai Phan and Javid Ebrahimi, *University of Oregon*

Dave Kil, *Civitas Learning*

Brigitte Piniewski, *PeaceHealth Laboratories*

Dejing Dou, *University of Oregon*

Modeling physical activity propagation, such as physical exercise level and intensity, is the key to preventing the conduct that can lead to obesity; it can also help spread wellness behavior in a social network.

Regular physical activity reduces the risk of developing cardiovascular disease, diabetes, obesity, osteoporosis, some cancers, and other chronic conditions.[1] Public health gold standards recommend that adults participate in at least 30 minutes of moderate-intensity physical activity five or more days per week.[2] However, less than 50 percent of the adult population meets these standards in most industrialized countries.[1,3] Therefore, finding effective intervention strategies to propagate physical activity is a core challenge.

The Internet is an important source of health information and could thus be an appropriate delivery mechanism.[4] Since 2000, a wide range of studies evaluating Internet-delivered health interventions has reported positive behavioral outcomes.[5,6] In particular, the widespread popularity of online social networks holds promise for wide-scale promotion of physical activity behavior changes. In addition, recent advances in mobile technology provide new opportunities to support healthy behaviors through lifestyle monitoring and online communities. Utilizing these technologies, we conducted a project in 2011 called YesiWell in collaboration with PeaceHealth Laboratories, SK Telecom Americas, and the University of Oregon to record daily physical activities, social activities (text messages, social games, competitions, and so on), biomarkers, and biometric measures (cholesterol, triglycerides, body mass index [BMI], and so on) for a group of 254 individuals. The users enrolled in an online social network application, allowing them to become friends and communicate with each other, and they carried mobile devices that reported their physical activities.

Our goal in this article is to further this work and understand the dynamics of

1541-1672/16/$33.00 © 2016 IEEE
Published by the IEEE Computer Society

Related Work in Online Social Networks

Since 2000, more than 15 studies[1] have evaluated website-delivered intervention to improve physical activity, a little over half of which reported positive behavioral outcomes. However, the intervention effects were short-lived, and there was limited evidence of maintenance of physical activity changes.

In recent years, social influence and the phenomenon of influence-driven propagations in social networks have received considerable attention. One of the key issues in this area is to identify a set of influential users in a given social network. Domingos and Richardson[2] approach the problem with Markov random fields, whereas Kempe and colleagues[3] frame influence maximization as a discrete optimization problem. Another line of study focuses on learning the influence probabilities on every edge of a social network, given an observed log of propagations over it.[4]

Many tasks in machine learning and data mining involve finding simple and interpretable models that, nonetheless, provide a good fit to observed data. In graph summarization, the objective is to provide a coarse representation of a graph for further analysis. Tian and colleagues[5] consider algorithms to build graph summaries based on node attributes, whereas Navlakha and colleagues[6] use the minimum description length principle[7] to find good structural summaries of graphs. Mehmood and colleagues[8] introduce a hierarchical approach to summarize patterns of influence in a network by detecting communities and their reciprocal influence strength.

References

1. C. Vandelanotte et al., "Website-Delivered Physical Activity Interventions: A Review of the Literature," *Am. J. Preventive Medicine*, vol. 33, no. 1, 2007, pp. 54–64.
2. P. Domingos and M. Richardson, "Mining the Network Value of Customers," *Proc. Knowledge Discovery in Databases*, 2001, pp. 57–66.
3. D. Kempe, J. Kleinberg, and E. Tardos, "Maximizing the Spread of Influence through a Social Network," *Proc. Knowledge Discovery in Databases*, 2003, pp. 137–146.
4. A. Goyal, F. Bonchi, and L.V.S. Lakshmanan, "Learning Influence Probabilities in Social Networks," *Proc. Web Search and Data Mining* , 2010, pp. 241–250.
5. Y. Tian, R. Hankins, and J. Patel, "Efficient Aggregation for Graph Summarization," *Proc. Special Interest Group on Management of Data*, 2008, pp. 567–580.
6. S. Navlakha, R. Rastogi, and N. Shrivastava, "Graph Summarization with Bounded Error," *Proc. Special Interest Group on Management of Data*, 2008, pp. 419–432.
7. J. Rissanen, "A Universal Prior for Integers and Estimation by Minimum Description Length," *Annals Statistics*, vol. 14, no. 5, 1983, pp. 416–431.
8. Y. Mehmood et al., "CSI: Community-Level Social Influence Analysis," *Proc. European Conf. Machine Learning Principles and Practice of Knowledge Discover in Databases*, 2013, pp. 48–63.

physical activity propagation via social communication channels at both the individual and community levels. More concretely, we aim to evaluate the probability of physical activity propagations for every social communication edge and devise a graph summarization paradigm to analyze physical activity propagation and social influence. We want 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 this goal, we were inspired by the well-known Independent Cascade (IC) model,[7] the Community-level Social Influence (CSI) model,[8] and the Physical Activity Propagation (CPP)[9] model (see the sidebar for "Related Work in Online Social Networks"). In this article, we extend our previous work by taking into account the content of social communication instead of a binary status (message sent or not sent) between two users. A message could belong to different topics and have different correlations with individuals' social influences. To address this issue, we propose combining the number of messages, their topics, and the effects of individuals 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 approach, the Topic-aware Community-level Physical Activity Propagation (TaCPP) model, is designed to capture the social influences of messages in the YesiWell study. To clarify the effect of activity propagation on health outcome, we analyze the correlation between detected communities and health outcome measures[10] through a comprehensive experiment on the YesiWell social network.

TaCPP Model

To understand how our model works, we first need to explain how to identify a single trace when user v influences another user u by sending a message. Assume that at time t, user v sends message m to user u; given a Δt, v is considered 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 time stamp (that is, $t + 1$),[7] but the process in health social networks is usually slower than that. Following other research,[8,9] we circumvent this problem by using time window w to define a single trace as follows: given a chain of users $\alpha = \{U_1, \ldots, U_n\}$ such that U_i is a set of users,

$U_1 \cap U_2 \cap \ldots \cap U_n = \varnothing$; α is called a single trace if $\forall i \in [1, n-1]$, and $\forall u \in U_{i+1}$ 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 (that is, v is a potential influencer for u) and is associated with a probability $p(v, u)$, which represents the strength of such influence in relationships. Let $D = \{\alpha_1, \ldots, \alpha_r\}$ 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, and Ω makes an attempt to activate every vertex $v \in V$, succeeding with probability $p(\Omega, v)$. At subsequent time steps, when node v becomes active, it makes one attempt at influencing each inactive neighbor u, which receives a message from v with probability $p(v, u)$. Multiple nodes can try to independently activate the same node at the same time.

We start by introducing the likelihood of a single trace a when expressed as a function of single-edge probability, which is useful for defining the problem that we tackle in this article. Let $I_{\alpha,u}$ be the set of user u's neighbors that potentially influence u's activation in trace α:

$$I^+_{\alpha,u} = \left\{v | (v, u) \in E, \text{ iff } u \in U_i \text{ then } v \in U_{i-1}\right\}. \quad (1)$$

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

$$I^-_{\alpha,u} = \left\{v | (v, u) \in E, \text{ iff } v \in U_{i-1} \text{ then } u \notin U_i\right\}. \quad (2)$$

Let p: $V \times V \rightarrow [0, 1]$ 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

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

Each $v \in I^+_{\alpha,u}$, where v succeeds in activating u on the considered trace α with probability $p(v, u)$ and fails with probability $1 - p(v, u)$. Message content is crucial to understanding users' physical activities. Given a set of topics K, each message could be related to a topic $k \in K$. In time window w, user v can send m messages in topic k to another user u, denoted $m_{k,v,u}$. Following other work,[8,9] we define $\gamma_{\alpha,v,u,k}$ as user responsibility, which represents the probability that in trace α, the activation of u was due to v's successful activation trial on topic k. The traces are assumed to be independent and identically distributed (i.i.d.). By using $\gamma_{\alpha,v,u,k}$, we can define the likelihood of the observed propagation as follows:

$$L_\alpha(p) = \prod_{u \in V}\left[1 - \prod_{v \in I^+_{\alpha,u}}\left(1 - p(v,u)^{\frac{\sum_{k \in K} m_{k,v,u}\gamma_{\alpha,v,u,k}}{Z(\alpha,v,u)}}\right)\right] \times \left[\prod_{v \in I^-_{\alpha,u}}\left(1 - p(v,u)^{1 - \frac{\sum_{k \in K} m_{k,v,u}\gamma_{\alpha,v,u,k}}{Z(\alpha,v,u)}}\right)\right], \quad (4)$$

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

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

To shift the influence strength estimation from node-to-node to community-to-community in the TaCPP model, we use a hierarchical decomposition H of the network G. In detail, H is a tree with network G as root r, the nodes in V as leaves, and an arbitrary number of internal nodes (that is, between root r and 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 root r to v. Therefore, by removing all 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 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 node $u \in V$ belongs to partition $\mathcal{P}_h$. In the TaCPP model, all vertices that belong to the same cluster are assumed to have identical influence probabilities toward other clusters. Given a probability function $\widehat{p_h} : \mathcal{P}_h \times \mathcal{P}_h \rightarrow [0,1]$. that assigns a probability between any two clusters of the partition $\mathcal{P}_h$, we define

$$p_h(v, u) = \widehat{p_h}\left(c(v), c(u)\right). \quad (6)$$

In the next section, we'll see that we can find $\widehat{p_h}$ by using an expectation maximization (EM) algorithm. But for the moment, let's assume that $\widehat{p_h}$ is induced by h in a deterministic function because our aim is to identify our problem in terms of finding 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 (that is, the individual level). Reducing the number of pairwise influence probabilities the model uses can only result in a lower likelihood, but the model complexity can be simplified, which

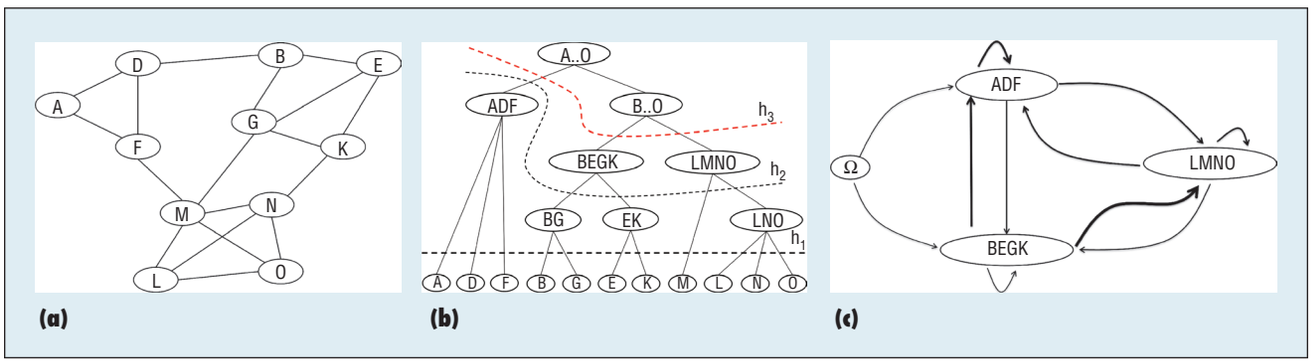

Figure 1. Input and output for the Topic-aware Community-level Physical Activity Propagation (TaCPP) model: (a) input, graph *G* of physical activity propagations (each undirected edge is considered as the corresponding two directed arcs); (b) hierarchy *H* generated by applying hierarchical clustering algorithms on *G*; and (c) output, a possible detected community structure resulted from Figure 1b and corresponding to cut h_3. Edge thickness represents the influence's strength.

is why we propose using a model selection function f that takes into account both likelihood and model complexity.

Figure 1 illustrates an example of the TaCPP model's input and output. Cut h_1 corresponds to the leaf-level model, where each single node of the social graph constitutes a state of the model. Essentially, this is the maximum likelihood cut that would correspond to the idea of a standard independent cascade model (that is, the individual level).[7] Two other cuts are also presented, where h_2 corresponds to clustering $\{\{A, D, F\}, \{B, G\}, \{E, K\}, \{M\}, \{L, N, O\}\}$ and cut h_3 results in our model in Figure 1b, which is the best option according to the model selection function f in this example.

Next, we need to formally define the model learning problem. Note that network G and hierarchy H remain fixed. Model complexity is only affected by cut $h \in C_H$.

Definition 1: TaCPP model learning. Given 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

$$h^* = \arg\min_{h \in C_H} f\left(L\left(D \mid \widehat{p_h}\right), h\right). \tag{7}$$

Intercommunity Influence and Model Selection

We propose an EM approach for estimating pairwise influence strength among node clusters—that is, the parameters of the TaCPP model. As presented earlier, we assume that the clusters in a partition $\mathcal{P}_{\hat{h}}$ 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; let $C(x) \subseteq V$ denote the set of vertices that belong to cluster $x \in \mathcal{P}_{\hat{h}}$.

According to the discrete-time independent cascade model,[7] given a single trace α, at least one of user $v \in I^+_{\alpha,u}$ was successful to deliver physical activities to user u independently, but we don't know which one. As discussed earlier, through user responsibilities $\gamma_{\alpha,v,u,k}$, we can define the complete expectation log likelihood of the observed propagation as follows:

$$Q\left(\widehat{p_h}, \widehat{p_h}^{\text{old}}\right) = \sum_{\alpha \in D} \sum_{u \in V} \left\{ \sum_{v \in I^+_{\alpha,u}} \left[\frac{\sum_{k \in K} m_{k,v,u} \gamma_{\alpha,v,u,k}}{Z(\alpha, v, u)} \log \widehat{p_h}(c(v), c(u)) + \left(1 - \frac{\sum_{k \in K} m_{k,v,u} \gamma_{\alpha,v,u,k}}{Z(\alpha, v, u)}\right) \log\left(1 - \widehat{p_h}(c(v), c(u))\right) \right] + \sum_{v \in I^-_{\alpha,u}} \log\left(1 - \widehat{p_h}(c(v), c(u))\right) \right\}, \tag{8}$$

where $\widehat{p_h}^{\text{old}}$ means the probability of the previous partition. Assuming that we have an estimate of every $\gamma_{\alpha,v,u,k}$, we can determine the $\widehat{p_h}$ that maximizes Equation 8 by solving $\left(\partial Q\left(\widehat{p_h}, \widehat{p_h}^{\text{old}}\right) \Big/ \partial \widehat{p_h}(x, y)\right) = 0$ for all pairs of clusters $x, y \in \mathcal{P}_{\hat{h}}$. This gives the following estimate of $\widehat{p_h}(x, y)$:

$$\widehat{P_h}(x, y) = \frac{1}{S_{x,y}} \sum_{\alpha \in D} \sum_{u \in C(y)} \sum_{v \in I^+_{\alpha,u} \cap C(x)} \sum_{k \in K} m_{k,v,u} \gamma_{\alpha,v,u,k}, \tag{9}$$

where

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

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 $\mathcal{P}$. Indeed, if v is believed to influence u on topic k in the trace α, this belief shouldn't change for different ways of clustering the two nodes. Therefore, we estimate $\gamma_{\alpha,v,u,k}$ from the model where every $u \in V$ belongs to its own cluster, which results in simplified estimates that only depend on network structure. By denoting this model as

$\hat{p}_o$, we obtain the following estimation of $\gamma_{\alpha,v,u,k}$:

$$\gamma_{\alpha,v,u,k} = \frac{m_{k,v,u}\widehat{p_o}(v,u)}{\sum_{z\in I^+_{\alpha,u}\cup I^-_{\alpha,u}}\sum_{k\in K} m_{k,z,u}\widehat{p_o}(z,u)}. \quad (11)$$

Our learning method for the TaCPP model is as follows:

- Apply topic modeling methods[11] to assign topics to every message m.
- Identify all possible traces $\alpha \in D$ following the definition of single trace presented earlier.
- 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

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

  Repeat the two following steps until convergence: one, estimate each successful probability $\hat{p}_o$, and two, update each influence responsibility $\gamma_{\alpha,v,u,k}$ by using Equation 11.
- Apply hierarchical clustering on $G = (V, E)$ to generate the hierarchy H. Each arc $(v, u) \in E$ represents an influence relationship $\hat{p}_o(v,u)$.
- After obtaining $\gamma_{\alpha,v,u,k}$, keep $\gamma_{\alpha,v,u,k}$ fixed for different partitions $\mathcal{P}_h$. Next, we utilize a heuristic bottom-up greedy algorithm to report the best solution found as output given the hierarchical decomposition H. In each iteration, the algorithm finds the two best communities to merge and update the model so that the selection function $f\left(L\left(D \mid \widehat{p_h}\right), h\right)$ in Equation 7 is minimized.

The probability between two clusters x and y in any partition $\mathcal{P}_h$, denoted $\widehat{p_h}(x,y)$, is computed according to Equation 9. 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 with the best value. Then, we can construct the community-level physical activity propagation network, such as in Figure 1c.

We already presented our learning method to maximize the log likelihood $L(D|p_h)$ at the individual level and gave a partition $\mathcal{P}_h$ to minimize the selection function $f(L(D \mid \widehat{p_h}), h)$. 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 tradeoff between model accuracy and model complexity. In this work, we use the *Bayesian Information Criterion* (BIC)[12] as a selection function f in Equation 7. In statistics, the BIC is a criterion for model selection among a finite set of models:

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

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 network's hierarchical decomposition.

Evaluating our objective function is computationally intensive because it involves re-estimating model parameters and computing the likelihood of D given those parameters. This might be too slow to be useful in practice. To speed up the algorithm,[8] we apply the following observation: merging two communities x and y, which exhibit exactly the same influence probabilities with all other communities z, doesn't affect the likelihood of D at all. In real contexts, such precise communities x and y rarely exist, but we can still find a merge where x and y are as similar as possible. To avoid computing the entire objective function for every possible merge, we find the merge that's the best in terms of the following similarity function, which respects the above condition:

$$\mathrm{sim}(x,y) = \sum_z \left(p(x,z)p(y,z) + p(z,x)p(z,y)\right). \quad (13)$$

The fifth step of our procedure, in each iteration, finds the best merge using Equation 13 and updates the model given this.

Experiments

We used the real-world YesiWell data and its corresponding social network to empirically validate the effectiveness of our proposed models. The YesiWell dataset, collected from 254 users, includes personal information, a social network, and daily physical activities over 10 months from October 2010 to August 2011. The initial physical activity data, collected by a special electronic device worn by each user, includes information about the number of walking and running steps in each 15-minute interval. Because some users' daily records are missing, we filtered those users whose daily physical activity record number is smaller than 80. In total, we ended up with approximately 7 million data points of physical exercise and 21,205 biomarker and biometric measurements. We only considered users who contributed to social communication—those who sent or received messages to or from other users. Ultimately, we had 123 users with 2,766 inbox messages for experiments.

Experiment Setting

Our proposed model (www.dropbox.com/s/3avaoe0hqdbiwnw/TaCPP.

Table 1. Topic description keywords of the messages in YesiWell data.

Technical	Physical activity	General	Program-social activity
hpod	day	weight	competition
steps	steps	don	find
today	work	food	weeks
days	walking	good	don
computer	walk	life	program
time	week	work	goals
goal	back	love	david

rar?dl=0) requires input as a hierarchical decomposition of the network. Following other work,[8] we obtain this hierarchy by recursively partitioning the underlying network using METIS,[13] which reportedly provides high-quality partitions. We set delay threshold Δt and time window w to a day and a week, respectively. Finally, we performed the *Latent Dirichlet Allocation* (LDA)[11] model on text messages in the YesiWell dataset to extract the underlying topics in users' messages. We found four coherent major topics in the messages: technique, physical activity, program-social activity, and an overlapping topic called general. Table 1 gives more clarification on how we distinguish topics via keywords in each topic.

Experimental Results

An effective way of summarizing influence relationships in the network is to consider the community-level influence propagation network. Figure 2 shows the networks of physical activity propagations detected by the TaCPP model for our dataset. Node size is the average number of steps for all users in a community. Arrowhead size is proportional to the probability of physical activity influence; we describe the shapes later. Note that we only consider the arcs that have probabilities larger than 0.25, which is very interesting because the network is almost acyclic, suggesting a clear directionality pattern in the flow of physical activities. With the models, we can categorize the detected communities into three kinds of groups based on their influence behavior as follows:

- *Influencer* (circle nodes in Figure 2). Indeed, these nodes have the strongest influence probability to deliver physical activities to other users in other communities. In addition, they receive almost no physical activity delivered from other communities.
- *Influenced users* (rectangle nodes in Figure 2). These nodes are easily influenced by influencers (circle nodes) because they receive the delivering of physical activity with high propagation probabilities. Moreover, the average number of steps taken by these nodes is quite large, even larger than influencer nodes. These influenced users sometimes try to deliver physical activities to other communities but not with a lot of strength.
- *Noninfluenced users* (triangle nodes in Figure 2). It's very hard for these nodes to be influenced because 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.

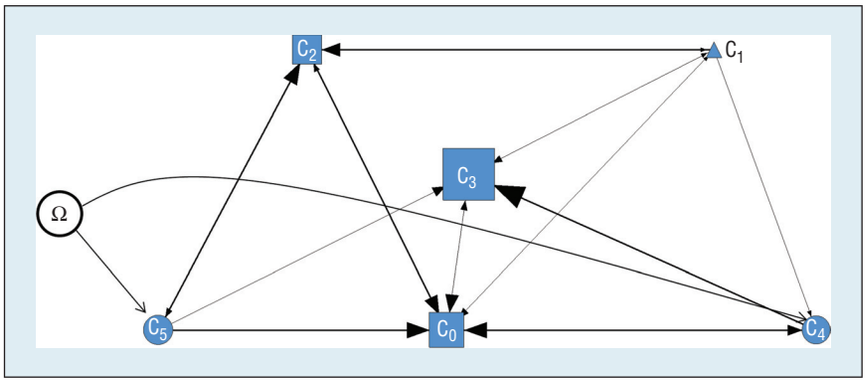

Figure 2. Detected community structure in YesiWell data. Node size is the average number of steps for all users in a community, and arrowhead size is proportional to the probability of physical activity influence.

Our approach's effectiveness can be validated by exploring the differences among these three user categories in terms of behaviors, life styles, and health outcomes to explain their physical activity propagation behaviors. Note that in these next experiments, all users in the same category were gathered together, thus we have only three groups of users instead of the six detected communities in Figure 2.

Physical activity record number. Figure 3 illustrates the average number of steps for the three groups over time. We can see that the influencer group not only has the best average BMI value among the groups, but its members are also stable in doing exercises day by day (that is, they exhibit a good, healthy life style) from the

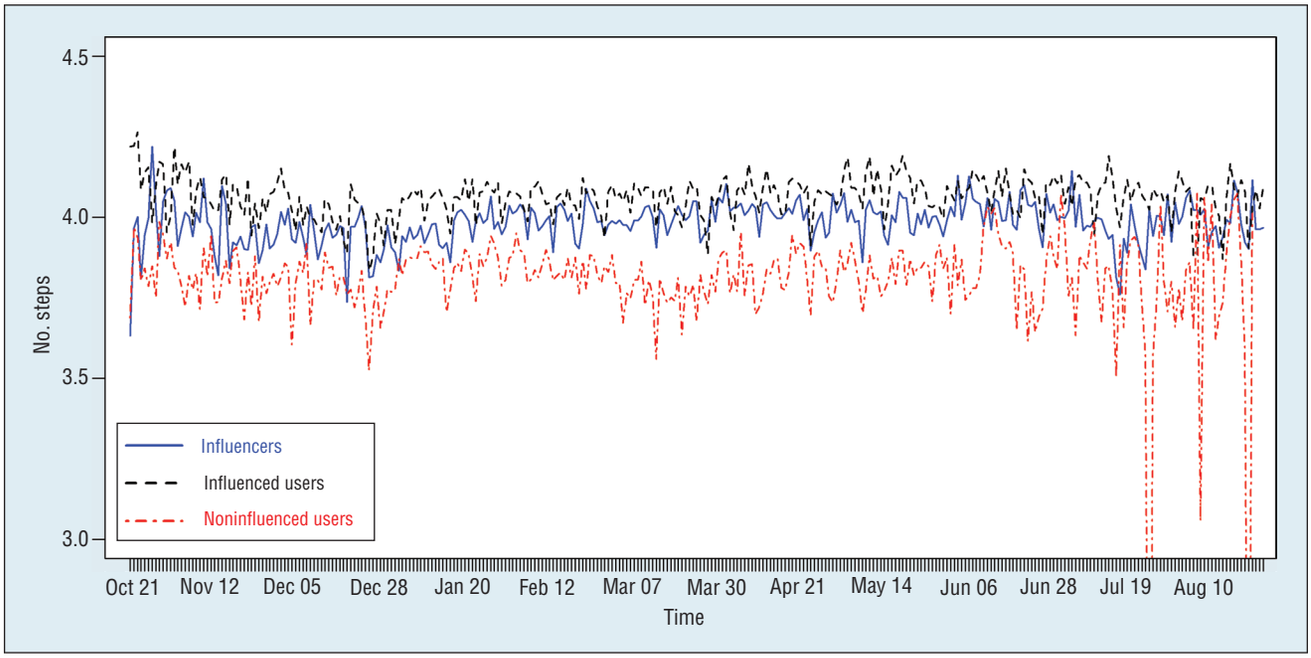

Figure 3. Average steps for all users in the three kinds of communities: influencer, influenced users, and noninfluenced users. (best viewed in color). It appears that the influencer group is successful at delivering physical activities to the influenced user group.

beginning to the end of the study. This clarifies the influencer group's activity-delivering role. Regarding the influenced user group, its members performed fewer physical activities at the beginning (middle of November 2010), but after that, they rapidly increased their activities, even more than the influencer group. Interestingly, their activity performance stabilized, along with that of the influencer group, until the end of the program. Clearly, it appears that the influencer group is successful at delivering physical activities to the influenced user group.

BMI. Figures 4a and 4b illustrate the average and the standard deviation of BMI for the three groups. Interestingly, the influencer group had average and standard deviation of BMI significantly lower than the other two groups. Because one of the goals of participants who enrolled in this study was to reduce their BMIs, the influencer group could potentially be an external motivation, which is one reason why the influencer group had strong influence probabilities on other groups. In addition, in Figure 4b, we can recognize that influenced users had higher BMIs than noninfluenced users in the beginning, but they eventually reduced their BMIs to be better than noninfluenced users. Meanwhile, noninfluenced users had almost the highest average and standard deviation of BMI (Figures 4a through 4d). Eventually, they had quite similar, or even better, BMI values than the influenced user group at the beginning.

Wellness score. Individual measures don't reflect the actual user health status, which is a complex combination of a user's life style, biometrics, and biomarkers. Our proposed wellness score[10] is such a metric; Figures 4e and 4f illustrate it for the three user groups. Clearly, the influencer group always had a high wellness score, but the influenced user group had a big change in its scores. In fact, the influenced user group had a low score at the beginning, but after that, it increased its scores to be among the highest. Meanwhile, the noninfluenced user group had the lowest score, despite a better starting point than the influenced user group.

TaCPP versus CPP. Our previous CPP model[9] could only distinguish the influencers in Figure 4a and the noninfluenced users in Figure 4e; it's difficult to clarify the behaviors of other user categories in this model. Fortunately, TaCPP produces a better community structure that offers a more insightful pattern of user influences. Indeed, it's very easy to discriminate the three user categories via their behaviors in Figures 4b and 4f, compared with the ones in Figures 4a and 4e. In addition, the communities detected by the TaCPP 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] for the TaCPP model and [1.5, 2.5] and [3, 5] in the CPP model.

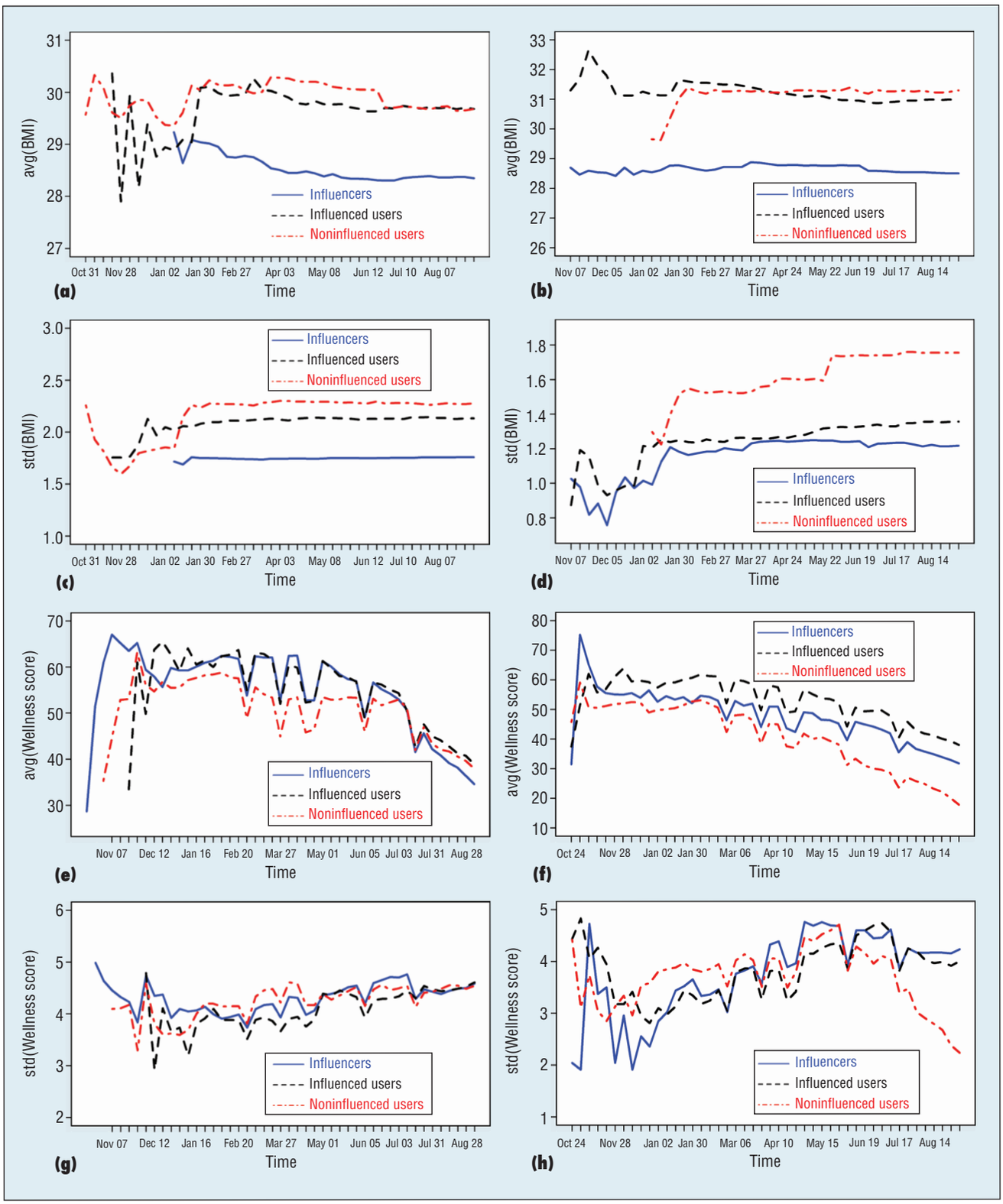

Figure 4. Health outcome measures for the three user categories. (a) Average BMI-CPP model; (b) average BMI-TaCPP model; (c) standard deviation of BMI-CPP model; (d) standard deviation of BMI-TaCPP model; (e) average wellness score-CPP model; (f) average wellness score-TaCPP model; (g) standard deviation of wellness score-CPP model; and (h) standard deviation of wellness score-TaCPP model.

The CPP and TaCPP models have strong correlations with health outcomes, which is very meaningful toward designing physical activity interventions through health social networks. But by incorporating message topics, the TaCPP model reveals a better community structure in terms of physical activity propagation, compared with the CPP model in the YesiWell social network.

Our proposed TaCPP model offers a more compact representation of propagation networks, and it can be easily plotted and exploited to understand and detect interesting properties in the information flow over a network. To clarify the sensitivity of our TaCPP model in topic modeling and hierarchical clustering, we apply different algorithms to assign topics to messages and generate different hierarchies H. Our domain experts labeled 2,766 messages in our data into 17 different topics: encouragement, fitness, follow-up, games, competition, personal, study protocol, tech, feedback, meetups, goal, social network, wellness meter, progress report, heckling, explanation, and invitation. In addition, we applied different agglomerative hierarchical clustering algorithms such as linking methods[14] (that is, the single, complete, weighted, and unweighted average linking methods), and methods that allow the cluster centers to be specified (that is, the median method[15] or centroid[16]). Our probabilistic inference method and all our novel observations haven't been affected by the clustering algorithms and this manual topic labeling. However, manually labeling messages by domain experts is impractical in real-world applications. Therefore, to scale the model to larger datasets, generative topic modeling methods are required. As long as we have an appropriate topic classification for messages and reasonable hierarchical decompositions, our probabilistic inference method and the final results won't be significantly affected. ■

THE AUTHORS

Nhathai Phan is a postdoctoral research associate in the Computer and Information Science Department at the University of Oregon. His interests include data mining, machine learning, human behavior modeling, online social network analysis, and spatiotemporal data mining. Phan received a PhD in computer science from the University of Montpellier 2, France. Contact him at haiphan@cs.uoregon.edu.

Javid Ebrahimi is a PhD candidate in the Computer and Information Science Department at the University of Oregon. His research interests are machine learning, natural language processing, and social networks. Ebrahimi received a BS in computer engineering from the University of Tehran, Iran. Contact him at javid@cs.uoregon.edu.

Dave Kil is the chief data scientist at Civitas Learning, overseeing the development of insight and action analytics. He holds 12 US and international patents. Kil received an MS in electrical engineering from Polytechnic University of New York and an MBA from Arizona State University. Contact him at david.kil@healthmantic.com.

Brigitte Piniewski is the chief medical officer at PeaceHealth Laboratories. Her research interests include collaborating with academic and technical experts to advance crowd-based approaches for producing evidence-based health intelligence at the pace of change. Piniewski also acts as vice chair of the Continua Health Alliance Wellness Solutions working group. Piniewski received an MD from the University of British Columbia. Contact her at BPiniewski@peacehealthlabs.org.

Dejing Dou is an associate professor in the Computer and Information Science Department at the University of Oregon, where he leads the Advanced Integration and Mining (AIM) Lab. His research areas include ontologies, data mining, data integration, information extraction, and health informatics. Dou received a PhD in artificial intelligence from Yale University. He's the principle investigator of NIH grant R01GM103309 and the corresponding author of this article. Contact him at dou@cs.uoregon.edu.

Acknowledgments

This work is supported by US National Institutes of Health (NIH) grant R01GM103309. We're grateful to Xiao Xiao, Rebeca Sacks, and Ellen Klowden for their contributions.

References

1. *Physical Activity and Health: A Report of the Surgeon General*, tech. report, US Dept. Health and Human Services, Centers for Disease Control and Prevention, Nat'l Ctr. Chronic Disease Prevention and Health Promotion, 1996.
2. R. Pate et al., "Physical Activity and Public Health: A Recommendation from the Centers for Disease Control and Prevention and the American College of Sports Medicine," *J. Am. Medical Assoc.*, vol. 273, no. 5, 1995, pp. 402–407.
3. A. Bauman et al., "Trends in Physical Activity Participation and the Impact of Integrated Campaigns Among Australian Adults, 1997–99," *Australian and New Zealand J. Public Health*, vol. 27, no. 1, 2003, pp. 76–79.
4. A. Marshall et al., "Exploring the Feasibility and Acceptability of Using Internet Technology to Promote Physical Activity within a Defined Community," *Health Promotion J. Australia*, vol. 2005, no. 16, 2005, pp. 82–84.
5. B. Marcus et al., "Interactive Communication Strategies: Implications for Population-Based Physical Activity Promotion," *Am. J. Preventive Medicine*, vol. 19, no. 2, 2000, pp. 121–126.
6. C. Vandelanotte et al., "Website-Delivered Physical Activity Interventions: A Review of the Literature," *Am. J. Preventive Medicine*, vol. 33, no. 1, 2007, pp. 54–64.
7. D. Kempe, J. Kleinberg, and E. Tardos, "Maximizing the Spread of Influence through a Social Network," *Proc. Knowledge Discovery in Databases*, 2003, pp. 137–146.

8. Y. Mehmood et al., "CSI: Community-Level Social Influence Analysis," *Proc. European Conf. Machine Learning Principles and Practice of Knowledge Discover in Databases*, 2013, pp. 48–63.
9. N. Phan et al., "Analysis of Physical Activity Propagation in a Health Social Network," *Proc. Conf. Information and Knowledge Management*, 2014, pp. 1329–1338.
10. D. Kil et al., "Impacts of Social Health Data on Predicting Weight Loss and Engagement," *O'Reilly StrataRx Conf.*, 2012; http://conferences.oreilly.com/strata/rx2012/public/schedule/detail/26120.
11. D. Blei, A. Ng, and M. Jordan, "Latent Dirichlet Allocation," *J. Machine Learning Research*, vol. 3, 2003, pp. 993–1022.
12. G. Schwarz, "Estimating the Dimension of a Model," *Annals Statistics*, vol. 6, no. 2, 1978, pp. 461–464.
13. G. Karypis and V. Kumar, "A Fast and High Quality Multilevel Scheme for Partitioning Irregular Graphs," *SIAM J. Scientific Computing*, vol. 20, no. 1, 1998, pp. 359–392.
14. F. Murtagh and P. Contreras, "Methods of Hierarchical Clustering," 2011; http://arxiv.org/abs/1105.0121.
15. J.C. Gower, "A Comparison of Some Methods of Cluster Analysis," *Biometrics*, vol. 23, no. 4, 1967, pp. 623–637.
16. P. Sneath and R. Sokal, *Numerical Taxonomy: The Principles and Practice of Numerical Classification*, Freeman, 1973.