Analysis of Physical Activity Propagation in a Health Social Network

NhatHai Phan
University of Oregon, USA
haiphan@cs.uoregon.edu

Dejing Dou
University of Oregon, USA
dou@cs.uoregon.edu

Xiao Xiao
University of Oregon, USA
xiaox@uoregon.edu

Brigitte Piniewski
PeaceHealth Laboratories
BPiniewski@peacehealthlabs.org

David Kil
HealthMantic, Inc
david.kil@healthmantic.com

ABSTRACT

Modeling physical activity propagation, such as the activity level and intensity, is the key to prevent the cascades of obesity, and help spread wellness and healthy behavior in a social network. However, there has been lacking of scientific and quantitative study to elucidate how social communication may deliver physical activity interventions. In this work we introduce a **C**ommunity-level **P**hysical Activity **P**ropagation (CPP) model to analyze physical activity propagation and social influence at different granularities (i.e., individual level and community level). CPP is a novel model which is inspired by the well-known Independent Cascade and Community-level Social Influence models. Given a social network, we utilize a hierarchical approach to detect a set of communities and their reciprocal influence strength of physical activities. CPP provides a powerful tool to discover, summarize, and investigate influence patterns of physical activities in a health social network. The detail experimental evaluation shows not only the effectiveness of our approach but also the correlation of the detected communities with various health outcome measures (i.e., both existing ones and our novel measure, named *Wellness score*, which is a combination of lifestyle parameters, biometrics, and biomarkers). Our promising results potentially pave a way for knowledge discovery in health social networks.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications—*Data Mining*

General Terms

Theory; Algorithms; Experimentation

Keywords

Physical activity propagation; health social network

CIKM '14 November 03 - 07 2014, Shanghai, China
Copyright 2014 ACM 978-1-4503-2598-1/14/11$15.00.
http://dx.doi.org/10.1145/2661829.2662025.

1. INTRODUCTION

Regular physical activity reduces the risk of developing cardiovascular decease, diabetes, obesity, osteoporosis, some cancers, and other chronic conditions [15]. Public health goal standards recommend adults to participate in at least 30 minutes of moderate-intensity physical activity on 5 or more days a week [16]. However, less than 50% of the adult population meets these standards in many industrialized countries [1, 15]. Thus finding the effective population-based intervention strategies to propagate the physical activity is a key challenge.

The exploiting of Internet and the success of online social networks hold promise for wide-scale promotion of physical activity behavior change. In many developed countries, Internet access is greater than 63% and keeps increasing [5]. The Internet is identified as an important source of health information and may thus be an appropriate delivery for health behavior interventions [10]. Since 2000, a wide range of studies evaluating Internet-delivered health behavior interventions have been reported. Over half of them have been reported positive behavioral outcomes [9, 17, 18, 27]. Recently, online social networks can help people to interact and participate various physical activities and thus could better promote and spread physical activities with affordable cost. However, there has been lacking of scientific and quantitative study to elucidate how social network may contribute to physical activity propagation.

Besides online social network, 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 2010-2011 as a collaboration between PeaceHealth Laboratories, SK Telecom Americas and University of Oregon to record daily physical activities, social activities (i.e., text messages, social games, meetup events, competitions, etc.), biomarkers, and biometric measures (i.e., cholesterol, triglyceride, BMI, etc.) for a group of 254 individuals who formed a health social network. Physical activities are reported via a mobile device carried by each user. All users enroll an online social network application allowing them make friend and communicate each other. Biomarkers and biometric measures are recorded via monthly medical tests performed at our laboratories on each user. The fundamental problems this study seeks to answer,

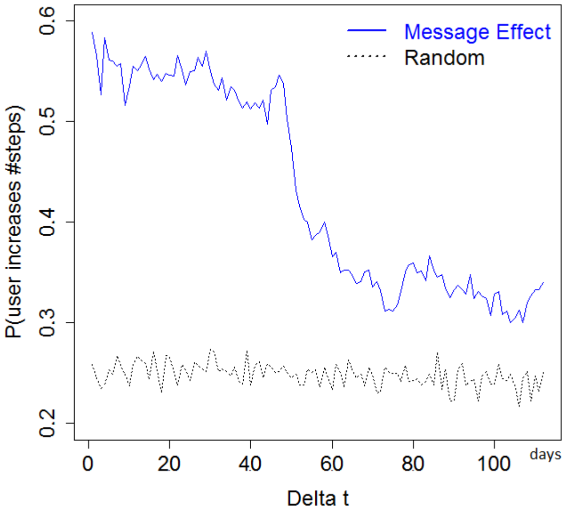


Figure 1: Probability of a message becomes effective to propagate physical activities.

which are also the key in understanding the determinants of healthy behavior propagation, are as follows:

1. Can social communication affect the physical activity propagation?
2. How can we leverage the social interaction to understand the physical activity propagation?
3. How can we understand the propagation process with different granularities?
4. Can we clarify the effect of physical activity propagation to health outcome measures?

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 his 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 timestamp t (i.e., user might or might not receive a message at a random time t). It is clear that with $\Delta t = 1$ *day* the probability a user increasing his total number of steps is up to 0.58 and significantly larger than 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.

Motivated by the evidence, our goal in this paper is to understand the dynamics of physical activity propagation via social communication channels at both individual level and community level. More in concrete: 1) we aim to evaluate the probability of physical activity propagations for every social communication edge. The estimated probabilities can be used in many applications (i.e., propagation prediction, health behavior interventions, etc); 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 which provides data analysts with a compact, and yet meaningful, view of patterns of influence and activity diffusion over health social networks. Members in the same community tend to play the same role in the propagation process.

To achieve this goal, we are inspired by the well-known Independent Cascade (IC) model [7] and the Community-level Social Influence (CSI) model [12] to fit a health social network. In our health social network, users are strongly encouraged to communicate each other. The correlation between effective messages and ineffective messages does not truly represent the user-user influence relationship. Therefore, existing models (e.g., CSI) cannot extract meaningful community structures. To overcome this issue we propose a new model called **C**ommunity-level **P**hysical Activity **P**ropagation (CPP) in which effective messages are combined with a user's responsibility to infer the probability of physical activity propagations in a health social network. 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. In order to clarify the effect of activity propagation to health outcome, we analyze the correlation between detected communities not only with existing health outcome measures (i.e., biometrics, BMI, average number of steps, BMI slope) but also with a novel measure, named *Wellness score*, which is modeled as a combination of lifestyle parameters, biometrics, and biomarkers.

The main contributions of this paper are as follows:

1. We introduce the Community-level Physical Activity Propagation (CPP) model, which is inspired by the ideas of IC and CSI models.
2. Given a set of disjoint communities, we devise an Expectation-Maximization algorithm to effectively learn the strength of their pairwise influence relationships. Then we utilize a greedy algorithm which explores a given hierarchical partitioning of the network. Our approach results in a community structure that guarantees a good balance between the accuracy in describing identified propagation activities and a compact representation of the influence relationships.
3. We propose a novel health outcome measure, named *Wellness score*, which is a combination of lifestyle parameters, biometrics, and biomarkers towards a mimic percentile user ranking.
4. Through a comprehensive experiment on the YesiWell social network, we show the effectiveness of our approach. Our discovery potentially paves a way for knowledge discovery and data mining in health social networks (e.g., physical activity interventions).

The rest of the paper is organized as follows. In Sec. 2, we formally define the problem tackled in this paper and explain the technical detail of our model. The experimental evaluation is in Sec. 3. We briefly review related prior art in Sec. 4 and conclude the paper with a summary of our major findings and future research directions in Sec. 5.

2. COMMUNITY-LEVEL PHYSICAL ACTIVITY PROPAGATION MODEL

We first give a definition of a single trace of physical activity propagations and review the fundamental independent cascade propagation (IC) model [7] in Sec. 2.1. Then we introduce CPP model (Sec. 2.2). Finally, we present our parameter learning process and model selection in Sec. 2.3.

2.1 Preliminaries and the Independent Cascade (IC) Model Review

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 & 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) [7]. However, the process in health social networks is usually slower than that. Following [11], we circumvent this problem by adopting a *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, $U_1 \cap U_2 \cap \ldots \cap U_n = \emptyset$; 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 $p(v, u)$ which represents the strength of such influence relationship. 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, 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 Independent Cascade (IC) model proposed by Kempe et al. [7] 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. [21] 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 CPP model to shift the modeling of influence strength from node-to-node to community-to-community. In our community-based model, all vertices which belong to the same cluster are assumed to have identical influence probabilities towards other clusters.

2.2 The CPP 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 paper. Let $I_{\alpha,u}$ be the set of user u's neighbors that potentially influence u's activation in the trace α:

$$I_{\alpha,u} = \{v | (v, u) \in E, \ \textit{if } u \in U_i \textit{ then } v \in U_{i-1}\} \quad (1)$$

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:

$$\log L(D|p) = \sum_{\alpha \in D} \log L_\alpha(p) \quad (2)$$

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 $1 - p(v, u)$. We define $\gamma_{\alpha,v,u}$ as users' responsibility which represents the probability that in trace α. The activation of u was due to the success of the activation trial performed by v. The traces are assumed to be i.i.d. By using $\gamma_{\alpha,v,u}$, we can define the likelihood of the observed propagation as follows:

$$L_\alpha(p) = \prod_{u \in V} \prod_{v \in I_{\alpha,u}} p(v, u)^{\gamma_{\alpha,v,u}} \big(1 - p(v, u)\big)^{1-\gamma_{\alpha,v,u}} \quad (3)$$

Note that social communication is very important to keep people following health intervention programs. Consequently we encourage social communications, i.e., message sending. Thus users may receive many messages but we only consider successful arcs of physical activity influence in Eq.3.

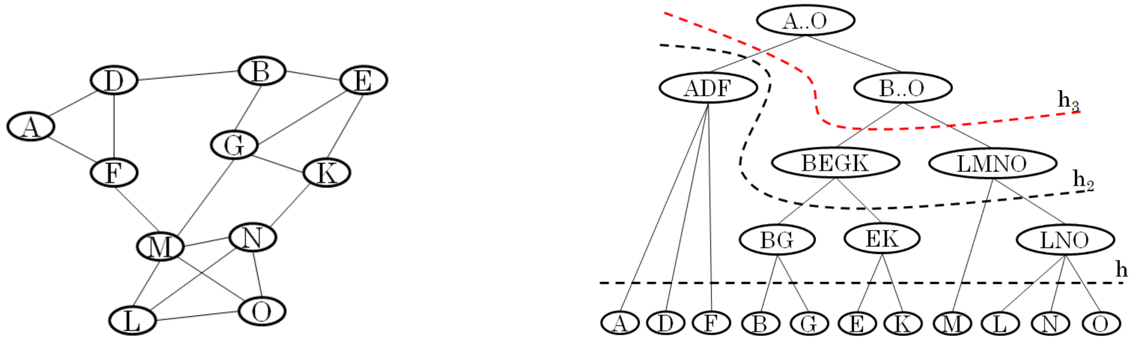
To shift the influence strength estimation from node-to-node to community-to-community in the CPP model, we use a hierarchical decomposition H of the network G. In detail, H is a *tree* with the network G as a root r, the nodes in V as leaves, and an arbitrary number of internal nodes (i.e., between 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 CPP model, all vertices that belong to the same cluster are assumed to have identical influence probabilities towards other clusters. Given a 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:

$$p_h(v, u) = \hat{p_h}(c(v), c(u)) \quad (4)$$

In the next section, we will show that we can 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 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 2 and 3 (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 to use a *model selection function* f that takes into account both likelihood and the complexity of the model.

For instance, Figures 2 and 3 respectively illustrate an example of input and output for our problem, i.e., a CPP



A network G of physical activity propagations

Hierarchical decomposition H of the G

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

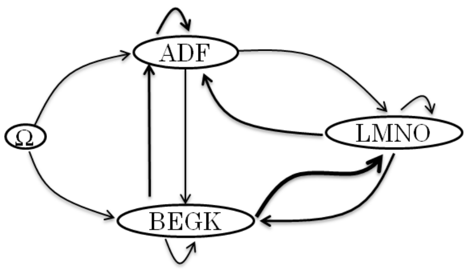


Figure 3: A possible detected community structure resulted from the input of Figure 2 and corresponding to the cut h_3. The edge thickness represents the strength of the influence.

model. The cut h_1 corresponds to the leaf level model where each single node of the social graph constitutes a state of the CPP model. Essentially this is the maximum likelihood cut that corresponds to the idea of standard independent cascade model [7] (i.e., individual level). Two other cuts are also presented, where 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 the example.

Then we can formally define the model learning problem addressed in this paper. 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. *CPP 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*

$$h^* = \arg\min_{h\in C_H} f(L(D|\hat{p_h}), h) \tag{5}$$

It is interesting to note that the two extreme cases outlined above, i.e., uniform probability, or all links have a different probability can be modeled in our approach. Indeed, 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).

2.3 Learning inter-Community Influence & Model Selection

In this section, we propose an expectation-maximization (EM) approach for estimating the pairwise influence strength among the clusters of nodes, i.e., the parameters of the CPP 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. Remind 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 [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 do not know which one. As discussed before, by using users' responsibilities $\gamma_{\alpha,v,u}$ we can define the complete expectation log likelihood of the observed propagation as follows:

$$Q(\hat{p_h}, \hat{p_h}^{previous}) = \log\Big(\prod_{\alpha\in D}\prod_{u\in V}\prod_{v\in I_{\alpha,u}} \hat{p_h}(c(v), c(u))^{\gamma_{\alpha,v,u}} \tag{6}$$

$$\big(1 - \hat{p_h}(c(v), c(u))\big)^{1-\gamma_{\alpha,v,u}}\Big)$$

where $\hat{p_h}^{previous}$ means the probability of the previous partition. Assume that we have an estimate of every $\gamma_{\alpha,v,u}$, we can determine the $\hat{p_h}$ which maximizes Eq.6 by solving $\frac{\partial Q(\hat{p_h}, \hat{p_h}^{previous})}{\partial \hat{p_h}(x,y)} = 0$ for all pair of clusters $x, y \in \mathcal{P}_h$. This gives the following estimate of $\hat{p_h}(x, y)$.

$$\hat{p_h}(x, y) = \frac{\sum\limits_{\alpha\in D}\sum\limits_{u\in C(y)}\sum\limits_{v\in I_{\alpha,u}\cap C(x)} \gamma_{\alpha,v,u}}{\sum\limits_{\alpha\in D}\sum\limits_{u\in C(y)}\sum\limits_{v\in C(x)} \mathbb{I}(v \in I_{\alpha,u})} \tag{7}$$

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

$$\gamma_{\alpha,v,u} = \frac{\hat{p}_o(v,u)}{\sum_{z\in I_{\alpha,u}} \hat{p}_o(z,u)} \tag{8}$$

We can summarize our learning method 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: $\hat{p}_o(v,u) = \frac{\sum_{\alpha\in D}\gamma_{\alpha,v,u}}{\sum_{\alpha\in D}\mathbb{I}(v\in I_{\alpha,u})}$. Repeats the two following steps until convergence.

   step 1 - Estimate each successful probability $\hat{p}_o$.

   step 2 - Update each influence responsibility $\gamma_{\alpha,v,u}$ by using the Eq.8.

2. After obtaining $\gamma_{\alpha,v,u}$, keep $\gamma_{\alpha,v,u}$ fixed for different partitions $\mathcal{P}_h$, and update $\hat{p_h}(x,y)$ according to the Eq.7.

We have already presented our learning method to maximize the log likelihood $L(D|p_h)$ at individual 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. We need thus an approach to address the trade-off between model accuracy and model complexity. In this work, we utilize the *Bayesian Information Criterion* (BIC) [22] as a selection function f in the Eq.5. In statistics, the BIC is a criterion for model selection among a finite set of models.

$$BIC = -2\log L(D|p_h) + |h|\log(|D|) \tag{9}$$

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

Finally, we can evaluate different cuts $h \in C_H$ of the hierarchical decomposition of the network. Next, we utilize the heuristic bottom-up greedy algorithm proposed in [12] 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.

3. EXPERIMENTS

The CPP model generalizes the presentation of physical activity propagations in health social networks. In the following we will describe how a CPP model can be exploited for different purposes including data understanding, and characterization of physical activity propagation flow. Furthermore it can be used to categorize users based on influence behaviors and health outcomes. We use the real world user behavior data and the corresponding social network to empirically validate the effectiveness of the CPP model. We first elaborate on the experiment configurations on the data set, and health outcome evaluation metrics. Then, we introduce the experimental results and how we can utilize our discovery in different applications.

3.1 Experiment Configuration and Health Outcome Metrics

Human Physical Activity Dataset. The YesiWell study is conducted in 2010-2011 as 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 in ten months from October 2010 to August 2011.

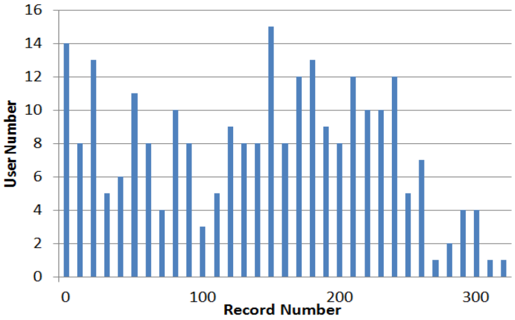


Figure 4: Distribution of the record number and user number.

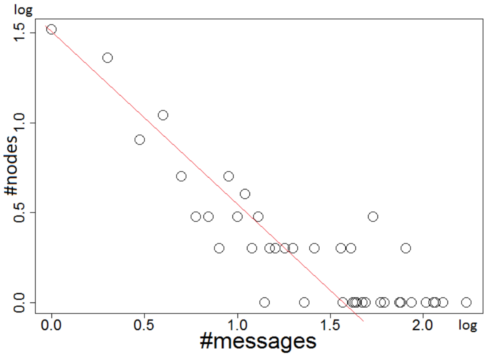


Figure 5: The number of inbox messages and the number of users distribution.

The initial physical activity data, collected by a special electronic equipment for each user, includes information of the number of walking and running steps. Since in the dataset, some users' daily records are missing, we show the basic analysis on the distribution of physical activity record numbers in Figure 4. In the 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 (i.e., users must send (resp., receive) messages to (resp., from) other users). Finally, we have 123 users for experiments. Figure 5 illustrates the distribution of the number of inbox messages and the number of users in our data. It clearly follows Power law distribution.

Body Mass Index (BMI) is a measure for human body shape based on an individual's mass and height, $BMI = \frac{mass(kg)}{(height(m))^2}$. The BMI is used in a wide variety of contexts as a simple method to assess how much an individual's body weight departs from what is normal or desirable for a person of his or her height. Indeed, BMI provides a simple numeric measure of a person's thickness or thinness, allowing health professionals to discuss overweight and underweight problems more objectively with their patients. The current value settings are as follows: a BMI of 18.5 to 25 may indicate optimal weight, a BMI lower than 18.5 suggests the person is underweight, a number above 25 may indicate the person is overweight, a number above 30 suggests the person is obese.

Wellness Score. The medical establishment has acknowledged major shortcomings of BMI. BMI depends upon weight and the square of height but it ignores basic scaling laws whereby mass increases to the 3rd power of linear dimensions. Hence, larger individuals, even if they had exactly the same body shape and relative composition, always have a larger BMI. Also, its assumptions about the distribution between lean mass and adipose tissue are somehow inexact [14, 25]. Thus, to enrich the health outcome and to rank user's health, we further propose a novel measure called *Wellness score*. In 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 the 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, etc.) 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, 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 [8].

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.

Experiment Setting. Our proposed model (source code[1]) requires input as a hierarchical decomposition of the network. Following [12], we obtain this hierarchy by recursively partitioning the underlying network using METIS [6], which reportedly provides high quality partitions. Finally, the delay threshold Δt and the time window w are respectively set to a day and a week. We ran our experiments on a Intel i7 2.8 GHz processor and 4 GB memory.

3.2 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 network of physical activity propagations for our dataset. The node size is the average number of steps for all users in their community. While the edge width is proportional to the probability of physical activity influences. The shapes will be described later. Note that we only consider the arcs which have probabilities larger than 0.25. It is very interesting since the network is almost acyclic, and this suggests a clear directionality pattern in the flow of physical activities. Moreover, with the CPP model we are able to categorize the eight detected communities into three kinds of group based on their influence behavior as follows:

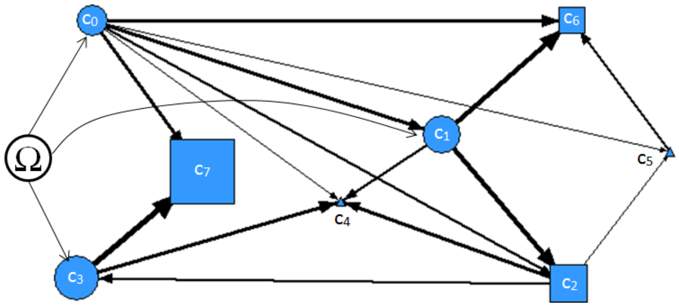


Figure 6: Detected community structure in our health social network data.

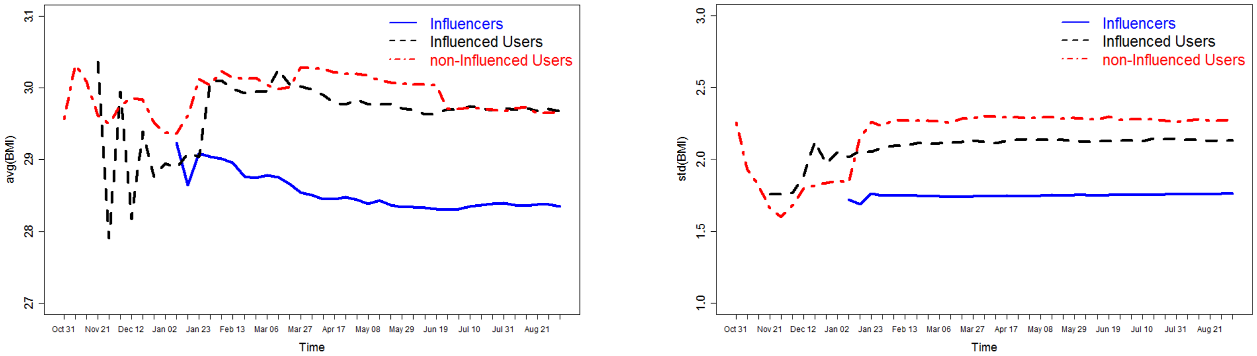
1) Influencer - This group can be seen as *circle nodes* in Figure 6. Indeed, these nodes have the strongest influence probability to deliver physical activities to other users in other communities. In addition, they almost do not receive physical activity delivering from other communities.
2) Influenced users - This group can be seen as *rectangle nodes* in Figure 6. These nodes are easy to be influenced by influencers (i.e., circle nodes) since they receive the physical activity delivering with high propagation probabilities. Moreover, the average number of steps of these nodes are quite large, even larger than the influencer nodes. These influenced users sometimes try to deliver physical activities to other communities but not much.
3) Non-Influenced users - This group can be seen as *triangle nodes* in Figure 6. These nodes are very hard 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 *non-influenced nodes* is very small compared with the other mentioned kinds of nodes.

Essentially, the effectiveness of our approach can be validated by exploiting the differences among the three user categories in terms of behaviors, life styles, 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, #steps, Wellness score) over time for the three groups. Note that in the next experiments, all the users in the same category will be gathered together and thus we will have only three groups of users instead of the eight detected communities.

BMI. Figure 7 illustrates the average and the standard deviation of BMI for the three groups (i.e., influencers, influenced users, and non-influenced users). Interestingly, the influencer group has 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 probabilities to other groups. Meanwhile, the non-influenced users have almost the highest average and standard deviation of BMI. Even they have quite similar BMI values with the influenced user group at the beginning.

Physical activity record number. Figure 8 illustrates the average number of steps for the three groups over time.

[1] ix.cs.uoregon.edu/~haiphan/Publications/CPP.rar



(a) Average BMI

(b) BMI standard deviation

Figure 7: Average and standard deviation of BMI for the three user categories.

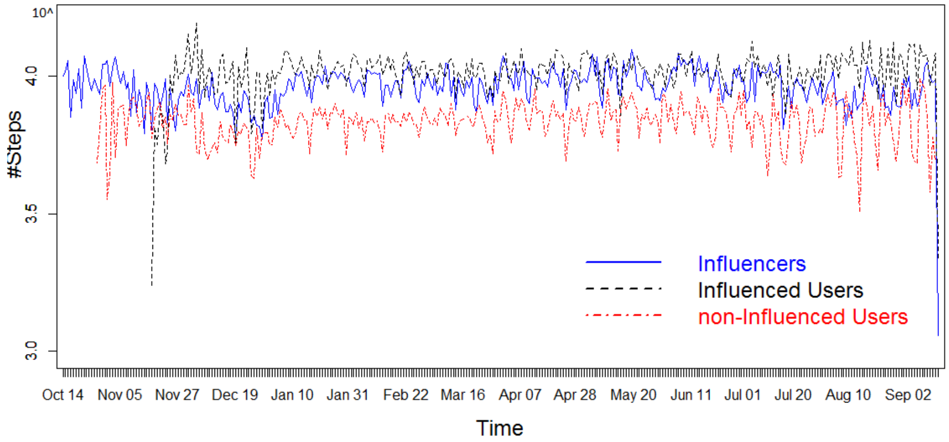


Figure 8: Average steps for all users in the three kinds of community, i.e., Influencer, Influenced users, and non-Influenced users. (Best view in color)

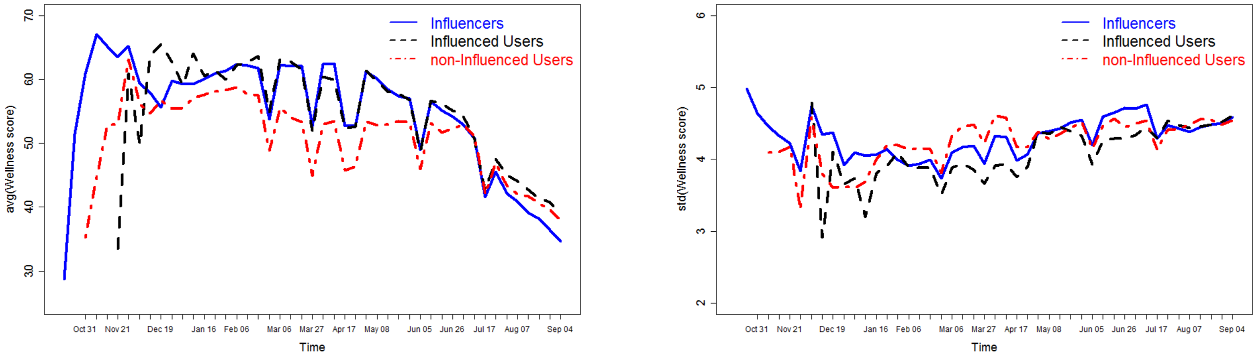
We can see that the influencer group not only has the best BMI values but also is stable in doing practices day by day (i.e., a good life style) from the beginning to the end of the study. Together with the CPP model results, it clarifies the activity delivering role of the influencer group. Regarding the influenced user group, they did less 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 the influencer group until the end of the program. With the CPP model results, we can say that the influencer group has been successful to deliver physical activities to the influenced user group.

Regarding the non-influenced user group, there is no big change in their physical activity behaviors. They have the lowest activity performance and it usually fluctuates in the whole program lifetime. It is only a short period (i.e., January to March, 2011) within that they have a quite stable (but the lowest) activity performance. So, we can say that it is hard to improve the practice behavior of non-influenced user group via social communications.

Wellness score. We have illustrated the correlation between the CPP model results and health outcome measures such as BMI and the exercise activity record number independently above. However, these individual measure cannot reflect the actual user health status which is a complex combination of a user lifestyle, biometrics, and biomarkers. Our proposed wellness score is a such metric. 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 user group has a low score at the beginning but after that they had increased their scores to be one of the highest ones. Meanwhile, the non-influenced user group has the lowest score even they has a better starting point compared with the influenced user group.

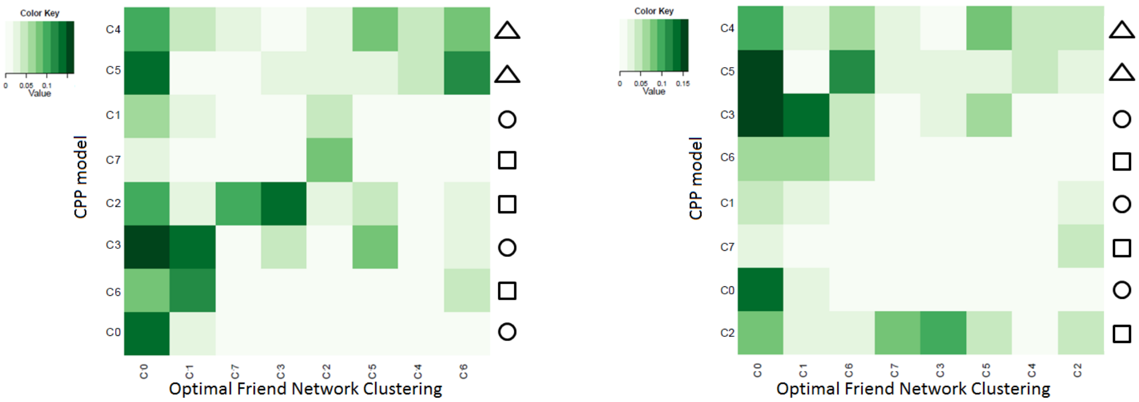
Community consistency. Interestingly, in Figure 7b and Figure 9b, the standard deviations of the BMI and Wellness score are quite small (i.e., from 1.5 to 2.5 for the BMI standard deviation, and from 3 to 5 for the Wellness score standard deviation). Furthermore they are quite stable (i.e., no big changes) for all the three user groups. Therefore, 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.

Until now, we can conclude that there are significant differences in terms of behaviors, lifestyles, biometrics, and biomarkers between the three user groups. Indeed, the CPP model offers us an effective tool to discover the flow of physical activity propagations. Base on that we can easily exploit unrevealed influence patterns and distinguish users in terms of physical activity delivering. Moreover, the detected communities are internally consistent. It is very useful for many



(a) Average Wellness score (b) Wellness score standard deviation

Figure 9: Average and standard deviation of Wellness score for the three user categories.



(a) #steps (b) Wellness score

Figure 10: CPP model vs social link based on health outcome. The markers correspond to the three user categories in Figure 6.

other tasks such as activity propagation prediction. Consequently, the CPP model has a strong correlation with health outcomes that is very meaningful for us to design physical activity interventions through health social networks.

The CPP model vs social link clustering. The output of the CPP model can be graphically represented to analyze the influence probability between two communities and social link relationships. An effective way is plotting the corresponding heat-maps, as shown in Figure 10. In these figures, we plot the Jaccard similarity in terms of number of steps and wellness score between the CPP model and obtained clusters by clustering the social network links. Note that the clustering algorithm maximizes the high correlation within-cluster and low between-cluster. Given two clusters A and B, the Jaccard similarity is computed as follows:

$$J(A, B, steps) = \frac{\sum_{u \in A \cap B} u.steps}{\sum_{u \in A \cup B} u.steps} \tag{10}$$

where $u.steps$ is the total number of steps reported by u. We use the similar equation for $J(A, B, wellness\ score)$.

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

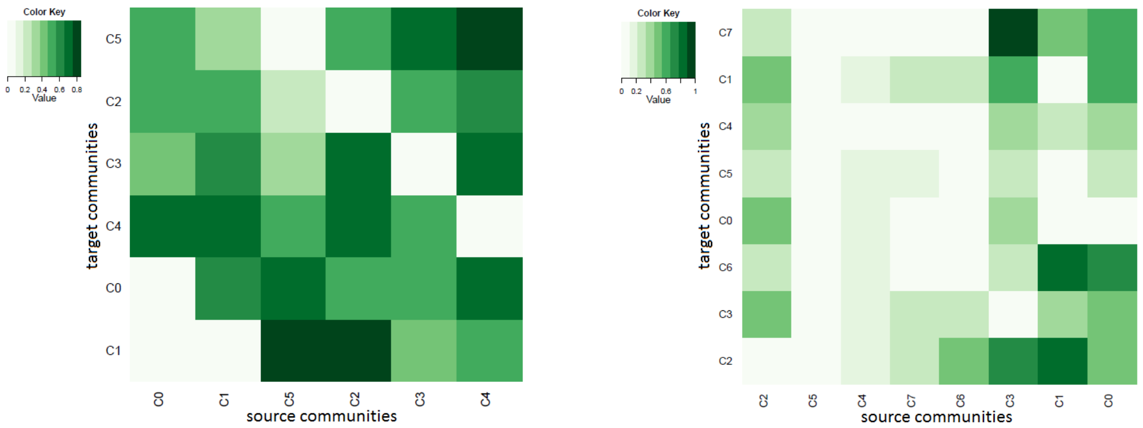
Comparison of the CPP model and the CSI model [12]. To highlight the effectiveness of our CPP model, we further compare our results with a CSI model. Indeed, we applied both model selection functions MDL [19] and BIC proposed in a CSI model. The former function generates only one community while we observe 6 communities from the latter function. In Figure 11, we plot the intensity of the influence probability between two communities observed from the CSI model (BIC model selection function) and the CPP model. In the CPP model, it is clear to see the influence role of the communities c_0, c_1, and c_3 while c_7, c_6 and c_2 receive strong influence probabilities. Furthermore, c_4 and c_5 do not contribute much to the process.

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

4. RELATED WORK

4.1 Physical Activity Intervention Approaches

Regular physical activities decrease the risk of developing cardiovascular disease, diabetes, obesity, osteoporosis, some cancers, and other chronic conditions. Thus, find-



(a) CSI model

(b) CPP model

Figure 11: CPP and CSI models on our health social network data.

ing effective population-based intervention strategies to promote physical activities is a key challenge. 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 [20].

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 [9, 17, 18]. Since then, over fifteen studies [27] evaluating a website-delivered intervention to improve physical activities that used the Internet or e-mail have been reported. Improvement in physical activities was reported in eight. 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. Indeed, 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. The maintenance of change and the engagement and retention of participants; larger and more representative study samples are also needed. Indeed, social network has this potential for being adopted since it take the advantage of the nature of social relationships to deliver healthy behavior. Furthermore, social network could be a long-life environment and thus the retention of participants could be naturally improved. Though we are in a long way to reach the goal, our proposed model and discovery is the foundation for further researches since it offers us a powerful tool to understand the physical activity propagation on a health social network.

4.2 Social Influence and Information Propagation Models

Social influence and the phenomenon of influence-driven propagations in social networks have received considerable attention in the 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 [3] approach the problem with Markov random fields, while Kempe et al. [7] 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 [4, 21, 24, 28]. In addition, 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 et al. [26] and Zhang et al. [29] consider algorithms to build graph summaries based on node attributes, while Navlakha et al. [13] use Minimum Description Length principle (MDL) [19] to find good structural summaries of graphs. In [12], Mehmood et al. introduce a hierarchical approach to summarize patterns of influence in a network, by detecting communities and their reciprocal influence strength.

5. CONCLUSIONS AND FUTURE WORK

In this paper we introduce a hierarchical approach to analyze the physical activity propagation through social communications at the community level (which also can be applied to individual level). Our proposed CPP model offers a more compact representation of the network of propagations. Furthermore it 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 the three meaningful observations: 1) social networks have great potential to propagate physical activities via social communications, 2) the propagation network found in a health social network by the CPP model is almost acyclic, and 3) the physical activity-based influence behavior has a strong correlation to health outcome measures such as BMI, lifestyles, and our proposed Wellness score.

Since online social networks have been exploited in recent years, our first observation paves an early brick on a new, promising, and perhaps most effective way to propagate physical activities to wide population. While the second observation offers interesting insights, 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. The third observation might be exploited to categorize users or to predict user macro-activities based on their influence behaviors [23].

In the near future, we are going to clarify the correlation between the physical activity propagation via social communications and a corresponding friend network. Indeed, homophily principle is important to deliver healthy behavior on health social networks [2]. Therefore, by discovering the correlation between homophily effect and social communications, we could have a complete picture. As a result we will be able to build up better human behavior predictive models and physical activity intervention approaches.

6. ACKNOWLEDGMENTS

This work is supported by the NIH grant R01GM103309. 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.

7. REFERENCES

[1] A. Bauman, T. Armstrong, J. Davies, N. Owen, W. Brown, B. Bellew, and P. Vita. 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, 2003.

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

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

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

[5] http://www.internetworldstats.com/stats.htm.

[6] G. Karypis and V. Kumar. A fast and high quality multilevel scheme for partitioning irregular graphs. *SIAM J. Sci. Comput.*, 20(1):359–392, 1998.

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

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

[9] B. Marcus, C. Nigg, D. Riebe, and L. Forsyth. Interactive communication strategies: implications for population-based physical activity promotion. *American Journal of Preventive Medicine*, 19(2):121–6, 2000.

[10] A. Marshall, E. Eakin, E. Leslie, and N. Owen. 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, 2005.

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

[12] Y. Mehmood, N. Barbieri, F. Bonchi, and A. Ukkonen. Csi: Community-level social influence analysis. In *Proceedings of ECML-PKDD'13*, pages 48–63, 2013.

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

[14] N. I. of Health. Aim for a healthy weight: Assess your risk. *National Institutes of Health*, 2007.

[15] U. D. of Health and H. Services. 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*, 1996.

[16] R. Pate, M. Pratt, S. Blair, and et al. Physical activity and public health. a recommendation from the centers for disease control and prevention and the american college of sports medicine. *JAMA*, 273(5):402–7, 1995.

[17] K. Patrick. Information technology and the future of preventive medicine: potential, pitfalls, and policy. *American Journal of Preventive Medicine*, 19(2):132–5, 2000.

[18] J. Prochaska, M. Zabinski, K. Calfas, J. Sallis, and K. Patrick. Pace+: interactive communication technology for behavior change in clinical settings. *American Journal of Preventive Medicine*, 19(2):127–31, 2000.

[19] J. Rissanen. A universal prior for integers and estimation by minimum description length. *The annals of statistics*, 14(5):416–431, 1983.

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

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

[22] G. Schwarz. Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464, 1978.

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

[24] J. Tang, J. Sun, C. Wang, and Z. Yang. Social influence analysis in large-scale networks. In *Proceedings of KDD'09*, pages 807–816, 2009.

[25] R. Taylor. Use of body mass index for monitoring growth and obesity. *Paediatrics & Child Health*, 15(5):258, 2010.

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

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

[28] R. Xiang, J. Neville, and M. Rogati. Modeling relationship strength in online social networks. In *Proceedings of WWW'10*, pages 981–990, 2010.

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