Part III

Using Social Contagion Models for Explaining Physical Activity

“We must all show great constancy,” Caspian was saying. “A dragon has just flown over the tree-tops and lighted on the beach. Yes, I am afraid it is between us and the ship. And arrows are no use against dragons. And they’re not at all afraid of fire.”

“With your Majesty’s leave-” began Reepicheep.

“No, Reepicheep,” said the King very firmly, “you are not to attempt a single combat with it.”

C.S. Lewis, The Voyage of the Dawn Treader (1952)\textsuperscript{1}

\textsuperscript{1}THE VOYAGE OF THE DAWN TREADER by CS Lewis © copyright CS Lewis Pte Ltd 1952.
Explaining changes in physical activity through a computational model of social contagion

“If you think of this world as a place simply intended for our happiness, you will find it quite intolerable: think of it as a place for training and correction, it’s not so bad.”

— C.S. Lewis
“God in the Dock”

Abstract

Social processes play a key role in health behaviour. Understanding the underlying mechanisms of such processes is important when designing health interventions with a social component. In this work, we apply a computational model of social contagion to a data set of 2,472 users of a physical activity promotion program. We compare this model’s predictions to the predictions of a simple linear model that has been derived by a regression analysis. The results show that the social contagion model performs better at describing the pattern seen in the empirical data than the linear model, indicating that some of the dynamics of the physical activity levels in the network can be explained by social contagion processes.

2GOD IN THE DOCK by CS Lewis © copyright CS Lewis Pte Ltd 1970.
7.1 Introduction

Physical inactivity is a major worldwide concern, as it can lead to many long-term health risks [6, 7]. These risks can be reduced if an adult fulfills the requirement (according to recommendations of the WHO and other public health organizations) of at least 150 minutes of moderate or 75 minutes of vigorous intensity physical activity per week, or a combination of both [8, 19]. An active lifestyle not only improves a person's physical health, but it also has positive effects on mental health [13].

If used in innovative ways, eHealth and mHealth hold great potential to steer physical activity promotion programs in the right direction and let greater numbers of people benefit from it. However, this requires the right choices about the way in which technology is embedded in these programs. For example, simply using a wearable device alone will not suffice to achieve sustainable behaviour change [14]. To maintain new behaviour for a longer period of time, other important ingredients are needed, e.g. evidence-based techniques such as goal setting and timely feedback, and a supportive social environment.

Social processes play a key role in health behaviour. It has been shown that people become more successful in maintaining a healthy lifestyle when they function within their social context [18, 20]. In addition, the social environment enables people to compare their physical activity achievements with their peers or to seek social support from them. Within online social networks, this is commonly implemented via leader boards with achievements, building on the theory of social comparison [17]. Overall, in the context of health promotion programs, social processes can provide a leveraging mechanism to achieve and maintain a healthy lifestyle. Understanding these mechanisms is therefore important.

In this paper, we use a data set about health behaviour in a social context to understand the underlying social processes. It is a continuation of earlier work on this subject [10, 11]. In [11], a large data set of an online physical activity promotion program was used to compare the physical activity levels of people who are part of an online social network with those who did not opt to join the network. One of the conclusions was that participants who are part of an online community have significantly higher activity levels and a higher increase in activity compared to participants who chose not to become part of the community. However, this did not answer the question what kind of social phenomenon was causing the higher activity levels.

In this work, we try to answer the question whether the increase in physical activity can be explained by social contagion [5]. Our main hypothesis is that the higher activity levels of the community users can be partially explained by social contagion and partially by the effect of the health promotion program. The research question is addressed by comparing the activity data of the participants with two types of predictions: (1) based on a simple linear model that captures the effect of participating in the program and the online community, and (2) based on a model of social contagion combined with the linear model.
7.2 Background

Because a majority of the adults in the Western world does not meet the guidelines for physical activity, public health professionals are aiming at population-wide interventions. Since decades, the area of preventive medicine is investigating how people can be stimulated to be more physically active [15]. More recently, the smartphone has been discovered as tool for measuring and influencing physical activity [3]. Many of these technology-mediated interventions use some kind of social influence. A specific appearance of social influence is the phenomenon of social contagion [5]. It has been shown that people can influence each other via their social networks up to three degrees of distance. Although these claims have been criticized [16], one could imagine that people transitively influence each other via social relations.

In [2] (based on [4]), a temporal-causal computational model is presented that describes how the mutual absorption of emotions in a social network affects the emotions of the individuals. This model was used for the study that is described in this paper. Our assumption is that physical activity behaviour is influenced by internal states like motivation, attitudes and goals, and that those spread in a similar way as described in the model of emotion contagion.

The model proposed by Araújo and Treur [2] describes how internal state $q_A$ of person $A$ affects the internal states of other persons $B_i$. This process is determined by the strength by which the state is expressed ($c_A$), the openness of the receiver ($\delta_B$) and the strength of the channel between them ($\alpha_{AB}$). Together, these factors determine the connection weight $\omega_{AB}$. Thus, the impact $\text{impact}_{AB}(t)$ of the state of person $A$ on the state of person $B$ is:

$$\text{impact}_{AB}(t) = \omega_{AB}q_A$$

(7.1)

The aggregated impact $\text{aggimpact}_B(t)$ at time $t$ of the states $q_{A_i}$ of all connected persons on state $q_B$ is modelled as a scaled sum. From this it follows that $\text{aggimpact}_B(t)$ is calculated as a weighted average of all the impacts of the different connections of a person:

$$\text{aggimpact}_B(t) = \sum_{A_i \neq B} w_{AB}q_{A_i}(t)$$

(7.2)

with $w_{AB}$ chosen in such a way that it is proportional to $\omega_{AB}$ and the sum of all weights is 1. The new state for each person in the network is calculated by integrating some factor $\eta$ of the aggregated impact:

$$\text{contagion\_effect}(t) = \eta_A[\text{aggimpact}_B(t) - q_B(t)]$$

(7.3)

$$q_B(t + \Delta t) = q_B(t) + \text{contagion\_effect}(t)\Delta t$$

(7.4)
For the purpose of this study, we assumed that all people have the same expressiveness and openness, and that all connections were of the same strength. This was done out of necessity, as our data set does not contain specific information about these factors. The model's parameters for openness, expressiveness and channel strength were thus set to a default value of 0.5.

7.3 Methods

This section describes how the data was collected and preprocessed, as well as what types of analyses were run.

7.3.1 Data collection

The data originates from a physical activity promotion program in which participants are asked to wear an activity monitor that measures physical activity level (PAL) using an accelerometer. Based on the activity data that is repeatedly uploaded by the participants, the program stimulates them towards a more active lifestyle by gradually increasing the weekly activity targets over a 12-week activity plan. The baseline for this activity plan is established in an initial assessment week. After completing a plan, participants can choose to take another 12-week activity plan or decide to remain at the level of their last completed plan.

After the initial assessment week, participants also get access to a dashboard with information about energy expenditure (calories burnt) and their achievements relative to a weekly goal. The program provides an opt-in online community that allows participants to establish connections and to compare achievements. Each participant in the community will see how their achievements rank compared to other participants with whom they are connected. Community participants see the ranking within their own network each time they upload data from their activity monitor. The network structure and some social network analyses are discussed in [1].

7.3.2 Data preprocessing

The original data set contains data for 52,788 users. Since the aim of this paper is to demonstrate the influence of social contagion on people's physical activity levels, we are only interested in the 5,041 users who opted in for the online community of the program.

First, any participants that joined the program for testing purposes or users with missing information, such as gender or body mass index (BMI), were removed from the data set, as well as participants that didn't have a start date for their first plan. The resulting data set contains participants for whom valid physical activity data is available. The network was further pruned by removing connections that were initiated by one participant, but never confirmed by the other participant.
As the online community feature was not part of the program until April 28\textsuperscript{th} 2010, all data before that date was disregarded. Community data was available until August 6\textsuperscript{th} 2010, but the PAL data was incomplete for the last couple of days. This can be explained by the fact that some users did not upload their data for those days yet. Therefore, only the data up to July 28\textsuperscript{th} 2010 was considered, resulting in a data selection that spanned a period of 91 days.

Within this period of 91 days, only active and connected participants were included in the current analysis. In other words, any users who entered the program, but did not join the online community, or users that dropped out of the program before this period started, were removed from the data set. This data cleaning process leaves us with 2,472 relevant nodes in the period between April 28\textsuperscript{th} 2010 and July 28\textsuperscript{th} 2010.

Although the primary unit of physical activity in the data set is the PAL, users see percentages of their goal achieved rather than the PAL itself on their online dashboard. The ranking with connected users on is also based on this relative performance. Therefore, our analyses are also based on the ratios of goals achieved, i.e. the current PAL divided over the target PAL.

7.3.3 Model simulations

Previous work has shown that the combination of participating in the program and joining the online community is associated with a small but significant average increase in PAL [11]. The objective of the current work was to demonstrate whether the dynamics of users’ physical activity levels can be (partially) explained by social contagion. Therefore, we compared the predictive performance of two different models: (1) a simple linear model, that describes the effect of the program on community members; and (2) a combined model, that captures the social contagion process and incorporates the known linear increase as well.

**Scenario 1: Simple linear model.** The simple linear model describes the effect of the physical activity promotion program and the online community on the users’ physical activity levels. Previous analyses have shown that this effect is an average PAL increase of 0.0005821 per day [11]. These analyses were based on a subset of users from the same data set, with all users being in their first plan and member of the community. The increase in PAL translates to an increase in energy expenditure of 1.05 kCal for an average male with a basal metabolic rate (BMR) of 1800 kCal/day [12].

To translate this increase in PAL to the unit predicted by the model (i.e., the goal achieved), the simple linear model adds a daily increase of 0.0005821 divided by the current target PAL to the user’s goal achieved, as shown in Equation 7.5 and Equation 7.6.

\[
 linear\_effect(t) = \frac{0.0005821}{target\_pal(t)} \tag{7.5}
\]
Scenario 2: Combined social contagion model. The combined social contagion model describes the linear increase in PAL as well, but combines it with the model of social contagion that captures the dynamics between the nodes in the network, as summarized in Equation 7.7, where \( \text{contagion\_effect}(t) \) denotes the social contagion effect as described in Section 7.2, Equation 7.3. In this case, the state \( q \) represents the percentage of goal achieved. By enriching the social contagion model with the daily increase in PAL (as in the simple linear model), we account for the demonstrated stimulating effect of the program and the community, and thereby nullify a possible disadvantage on the social contagion model.

\[
\text{goal\_achieved}(t + \Delta t) = \text{goal\_achieved}(t) + \text{contagion\_effect}(t) + \text{linear\_effect}(t) \tag{7.7}
\]

As mentioned in Section 7.3.2, the analyses were based on the predictions of the goal achieved, i.e. the proportion of the target PAL achieved by the user, rather than the user’s current PAL. Additionally, the model predictions were done for users in their first plan. Of the 2,472 relevant users identified in Section 7.3.2, 1,939 were participating in their first plan for at least part of the time period under consideration. The reason behind this choice is that users in their first plan are most comparable to the general population: they have just entered the program, and therefore have no prior knowledge of or experience with the plans or other parts of the intervention. Also, it is likely that people in their first plan have the highest adherence rates and interact more with the program, which makes them a more interesting population as well. However, users who have not yet started or already completed their first plan can still influence users in their first plan through social contagion. Therefore, they are considered by the social contagion model, but only as input of the contagion process towards the users under consideration (i.e., users in their first plan).

To run the models, the initial values have to be determined. For all users for whom a target PAL is not available (i.e., users who are in their assessment week and have yet to start their first plan), the initial goal achieved value was based on the average PAL of their assessment week and their first target PAL. For all users with a target PAL, the initial goal achieved was calculated by dividing the average PAL for one week before the start date of the simulations (i.e., April 28\textsuperscript{th} 2010) by the current target PAL. If for some reason, no data was available for that week, the initial goal achieved was based on the average PAL in the month prior to the start date of the simulations.

In the social contagion model, we used the initial goal achieved values of the simulated nodes as described above, and the empirical data from the surrounding nodes as input to the contagion process. This choice was motivated by the fact that we were only interested in simulating the effect of the behaviour of users on users in their first plan, rather than simulating the behaviour of those other users as well.
7.3.4 Analyses

To evaluate the accuracy of the two models, we first calculated their average predictions for the approximately 1,939 users in their first plan in the data set, as well as the average goal achieved values based on the empirical data. Based on these values, we tested whether there is a significant difference in the magnitude of the errors of the two models with a Mann Whitney U test. In addition, we determined the correlations of both models’ predictions to the empirical data by means of Mann Kendall tests.

7.4 Results

As explained in Section 7.3.2, after thorough preprocessing of the data, 2,472 relevant users remained in the period between April 28th and July 28th 2010.

Following the procedures described in Section 7.3.3, the two models were run on the initial data. Figure 7.1 provides an impression of the predicted goal achieved values for the 1,939 users in their first plan by the two models. The simulation of the linear model shows a steady increase in the goal achieved. The combined model shows the effect of the contagion between the users, in combination with the steady increase. Any interruptions of the lines in either plot are caused by users entering the program or community, or by users dropping out of the program.

![Fig. 7.1: Predictions of the simple linear model (left) and the combined model (right).](image)

After averaging the model predictions, as well as the empirical data, for all users in their first plan per day, the graph in Figure 7.2 was obtained. It shows the average predictions of the linear model (green) and the combined model (blue), and the empirical data (red). The sharp troughs in the empirical data mark the Sundays, when physical activity levels on average are substantially lower.

Figure 7.2 already gives the impression that the combined model is much closer to the empirical data than the linear model. Indeed, the mean absolute error (MAE) of the linear model is 0.02212, whereas the mean absolute error of the combined model is 0.01321. A Mann-Whitney U test shows that the difference between the errors of the two models is significant, $p < 0.001$.

Besides comparing the size of the errors, we also investigated whether the predicted lines were correlated with the empirical data. A Mann Kendall test shows that the
Fig. 7.2: Average predictions of the two models (green: linear, blue: combined), and the empirical data (red).

The linear model is significantly correlated with the empirical data, although negatively ($\tau = -0.46227$, $p < 0.001$). The combined model is also significantly correlated, but in this case positively ($\tau = 0.53895$, $p < 0.001$).

Tab. 7.1: Model evaluations

<table>
<thead>
<tr>
<th></th>
<th>Absolute Error</th>
<th>Kendall's correlation test</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>St. Dev.</td>
</tr>
<tr>
<td>Linear Model</td>
<td>0.02212</td>
<td>0.01378</td>
</tr>
<tr>
<td>Combined Model</td>
<td>0.01321</td>
<td>0.00855</td>
</tr>
</tbody>
</table>

7.5 Conclusions

The results described in Section 7.4 show that the combined model, which integrates the social contagion model with a steady linear increase in PAL, is indeed better able to capture the dynamics of the physical activity levels in our data set than the linear model. Its predictions show a significant positive correlation with the empirical data. Additionally, the errors of the combined model’s predictions are significantly smaller than those of the linear model.

One of the main strengths of this work is its foundation on a large set of empirical data covering several months. Careful and extensive preprocessing of the empirical data was conducted to ensure data that is sensible for the simulated models. For example, we dynamically removed connections to users who practically dropped out of the program (but were still in the system), to prevent their (missing) data from affecting the results.
Another strength of our work is that we compared the performance of the model we were mainly interested in to an informed linear model. That way, we do not impose a disadvantage on the baseline model, thus increasing the chances of superiority of our more complex model. However, it is interesting to see that the empirical data shows a development that is actually opposite to the direction of the linear increase model. One possible explanation for this observation could be that the linear increase was found after aligning the data by the day in the program rather than the calendar date. The pattern in the current data set is then caused by users in different phases of the first plan entering and leaving the program over time (e.g., because their first plan is finished halfway the period that we selected). A second possible explanation is that the linear model describes an increase in PAL, whereas it is transformed and applied to the progress towards the target PAL in this work. A third possible explanation is that the linear model was based on a different subset of the same data set, so maybe the subset analyzed in this work does not show an average increase in PAL.

One of the limitations of this work is its restricted generalizability. As all analyses were based on data collected in the context of a physical activity promotion program (see also Section 7.3.1), the results cannot directly be transferred to the general population. However, by choosing to focus on people who are exposed to the program for the first time, we have tried to minimize that discrepancy.

Another limitation is that the social contagion model only considers the online community as the network through which the behaviour spreads, although contagion also takes place on different levels and in different contexts. Additionally, we did not take into account whose data is actually shown on the user’s dashboard: all connections were treated equally, whereas the performance of friends may not be shown on the dashboard when the difference was too big (e.g., more than 10 position difference). Future work could reveal whether limiting the contagion model to only the connected users who are visible on the dashboard improves the performance of the model. A further limitation is that we used default values of 0.5 for the parameters (for expressiveness, channel strength and openness) in the combined model. In future work, we could investigate whether using calibrated values would yield better results. It is also possible to experiment with models that incorporate the principle of non-linearity in behaviour change, e.g. by exploiting thresholds for effects [9].

Up to our knowledge, we present the first analysis of the ability of a computational model of social contagion to capture the pattern of physical activity levels in a community over time. The results show that the enriched social contagion model performs better at describing the pattern in the empirical data than the linear model, indicating that some of the dynamics of the physical activity levels in the network can be explained by social contagion processes. This is vital information for designers of health interventions with a social component, as such models can then be used to maximize the benefits of social influence processes.


