Part II

Data Analytics in Networks

"And more," said Queen Lucy, "for it will not go out of my mind that if we pass this post and lantern either we shall find strange adventures or else some great change in our fortunes."

"Madam," said King Edmund, "the like foreboding stirreth in my heart also."

"And in mine, fair brother," said King Peter.

"And in mine too," said Queen Susan. "Wherefore by my counsel we shall lightly return to our horses and follow this White Stag no further."

"Madam," said King Peter, "therein I pray thee to have me excused. For never since we four were Kings and Queens in Narnia have we set our hands to any high matter, as battles, quests, feats of arms, acts of justice, and the like, and then given over; but always what we have taken in hand, the same we have achieved."

"Sister," said Queen Lucy, "my royal brother speaks rightly. And it seems to me we should be shamed if for any fearing or foreboding we turned back from following so noble a beast as now we have in chase."

"And so say I," said King Edmund. "And I have such desire to find the signification of this thing that I would not by my own good will turn back for the richest jewel in all Narnia and all the islands."

"Then in the name of Aslan," said Queen Susan, "if ye will all have it so, let us go on and take the adventure that shall fall to us."

C.S. Lewis, *The Lion, The Witch, and The Wardrobe (1950)*[1]

[1] THE LION, THE WITCH AND THE WARDROBE by CS Lewis © copyright CS Lewis Pte Ltd 1950.

3

Applying machine learning algorithms for deriving personality traits in social network[1]

> *"You can't get a cup of tea big enough or a book long enough to suit me."*
>
> — **C.S. Lewis to Walter Hooper**
> "Of Other Worlds"[2]

Abstract

Social and cognitive sciences' knowledge about social behaviour and social networks combined with the new computational machine learning techniques can facilitate the creation of better models. We propose and evaluate a new methodology for finding personality traits of young adults involved in a network using hyper optimization algorithms. We used a social contagion model for the spread of behaviour (measured by the physical activity level) among the participants. A part of the Big-5 questionnaire was used to gather information about people regarding their traits of openness and expressiveness. Then we try to fine tune the model using machine learning algorithms. The fine tuning of questions from an intake questionnaire can be very useful in validating a model. The accuracy delivered by machine learning pure algorithms is shown to be better, but the inclusion of data related to people's traits is beneficial in defining their characteristics.

[1] Published as: **Eric F. M. Araújo**, Bojan Simoski, and Michel Klein. 2018. Applying machine learning algorithms for deriving personality traits in social network. In Proceedings of ACM SAC Conference, Pau, France, April 9-13, 2018 (SAC'18).

[2] OF OTHER WORLDS by CS Lewis © copyright CS Lewis Pte Ltd 1966.

3.1 Introduction

Recent research has revealed that social contagion is one of the main factors that shape people's opinions, beliefs and behaviour [8, 16]. Questionnaire-based self-reports are frequently used for measuring or quantifying cognitive and personality characteristics as traits of people, mood state, character aspects, etc. Although self-reports are well accepted as a standard part of social sciences methodology, it is known that bias can be a factor. At least 48 types of biases have been identified [7].

In this paper, we introduce and explore a new method for finding the traits of people in a social network. To show how the proposed method works, we use data collected in an experiment where the physical activity level of 25 people was measured for one month. The participants were asked to fill in questionnaires regarding their personal traits and their social network (friendship and level of contact with other participants). A social contagion model is used as a predictor for the future states of the nodes in the network of the participants. We use machine learning techniques to find the best weights for each question in the questionnaire. We also use parameter tuning algorithms to find the traits directly without the questionnaires. The proposed method could be used to evaluate the questionnaires or as an alternative method for collecting personal information such as traits.

The following section presents related works and a brief description of our previous work with this data set. Section 3.3 explains the social contagion model used for simulations. Section 3.4 presents the new methodology proposed, while Section 4.3.5 describes the data. Section 4.4 contains the results of the case study. Finally, in Section 4.5 we discuss the usefulness of our approach and potential future improvements.

3.2 Related work

There are many kinds of behaviour models in social networks. Studies have addressed this topic in social sciences for many years [5, 13], and most of them aim to find correlations through statistical analysis as outcomes from the observed data. Most of the works that combine human behaviour and machine learning algorithms categorize behaviours in order to predict future states. Wang [19] has used machine learning tools to detect Twitter accounts that don't behave like humans, i.e. detect bots. Ellis et al. [10] classify activities related to the GPS information collected.

Some works are interested in defining the traits of people instead of only classifying the behaviour. Durupinar et al. [9] have used adjectives to define the traits of people in a crowd in order to simulate heterogeneous subgroups acting in particular circumstances. The tuning of the simulation is done manually and does not involve any sort of computational method.

Alam et al. [1] have used different ML classification algorithms to recognize Big-5 personality traits in Facebook's social network using the text from the status of users and self-reports. Even though the approach used can be compared to ours, they are not concerned about weighting the questions to find out which of them are relevant or not. Araújo et al. used a network-oriented social contagion model in order to predict if the PAL of each person is going to improve or deteriorate with an accuracy of above 80% using two traits from the Big-5 inventory. This work does not address the question about the relevance of the questions in the intake questionnaires.

We propose a new approach for defining the traits of people using hyper optimization using a data set of people's behaviour and a social contagion model. We couldn't find any similar approach in literature, as most of the works define the traits based on manual operations and formulas created without any automation or parameter tuning algorithms.

3.3 Social Contagion Model

We use a model of social contagion to relate observable (physical activity) behaviour with personality traits. Social contagion is a phenomenon that concerns the attitudes, beliefs and behaviours being spread among people [14]. A computational model was designed by Bosse et al. [6] to interpret and model group emotion spreading over time among work team members. The model is not restricted to emotions. It is also applicable in explaining behaviour contagion. Araújo et al. [3] used the model to predict the change in physical activity where PAL is the spreading factor in the network.

We use the social contagion model for the spreading of physical activity behaviour as the internal state for the nodes in a network. Each person has an internal state $q \in [0, 1]$ that affects the internal states of other persons in the network, q_i. Differently from [3], we use q_i as the weekly mean of the PAL of person i instead of the daily PAL. This is done to decrease the fluctuation and picks in the values caused by weekends or unusual days of intense activities that are not part of the routine of the person.

Three factors affect the contagion process in the network, namely the *expressiveness* of the sender (B) ϵ_B, the *openness* of the receiver (A) δ_A, and the connection strength between sender B and receiver A, w_{BA}. The *expressiveness* determines the strength by which the internal state of person B is expressed to the other members of the network. The *openness* of the receiver A, gives an indication to what extent the person is open to be influenced by other members, while the connection strength describes how strong the relationship between B and A is. These parameters are numerically represented as real numbers between 0 and 1. The overall contagion strength is the total contagion of all connections towards person A, and is calculated as shown in equation 3.1.

$$\gamma_A^* = \sum_{B \neq A} \epsilon_B w_{BA} \delta_A \tag{3.1}$$

.

The value of γ_A^* is used as a speed factor for changes in the model dividing it for the number of neighbors of node A,

$$\gamma_A = \gamma_A^* / num\_neighbors_A$$

following the modification proposed in [2]. The aggregated impact is the real amount of behavioural influence person A receives and it is calculated as shown in equation 3.2, where the proportional weight of the contagion for each node B to A is given as $\omega_{BA} = \frac{\epsilon_B w_{BA}}{\sum_{C \neq A} \epsilon_C w_{CA}}$.

$$\mathbf{aggimpact}_A = \sum_{B \neq A} \omega_{BA} q_B \tag{3.2}$$

Then finally, new state of person A in time $t + \Delta t$ is given as $\Delta q_A(t + \Delta t) = q_A(t) + \gamma_A(\mathbf{aggimpact}_A(t) - q_A(t))\Delta t$.

3.4 Methods

We use machine learning techniques to find the best weights for each question in a questionnaire-based social experiment. We also explore whether the same parameter tuning approach can be used to find the traits directly; we evaluate this by comparing the resulting traits with the ones provided by the questionnaires.

The *Hyperopt* library for Python [4] was used to calculate the best parameters for the following scenarios:

(1) Find the weights for each of the 18 questions from the questionnaire and the speed factor for the model simulation;

(2) Find the personality traits of openness and expressiveness for each of the 20 participants and the speed factor for the model simulation.

In both scenarios the algorithm runs the contagion model described in Section 3.3 and aims to minimize the difference between the simulated physical activity level of the participants and the actual physical activity level in the experimental data. We also used a simulated annealing algorithm to try a different approach rather than the grid search provided by Hyperopt library. Unfortunately due to the big number of dimensions, this didn't yield useful results.

Figure 3.1 (left) shows the framework for method (1), tuning the weights of the questions. Given an initial weights set, it uses the questionnaires to calculate the openness and expressiveness for each person. The same weights for each of the questions are used for all the participants. After that, the model calculates the simulated change in physical activity behaviour over time. The error is the sum of the squared differences between empirical and simulation data. The learning algorithm then makes adjustments to the weights and a new simulation is performed. Figure 3.1 (right) shows the framework for the process of finding the best parameters for each participant for method (2). In this case the two personality traits (openness and expressiveness) are tuned directly.

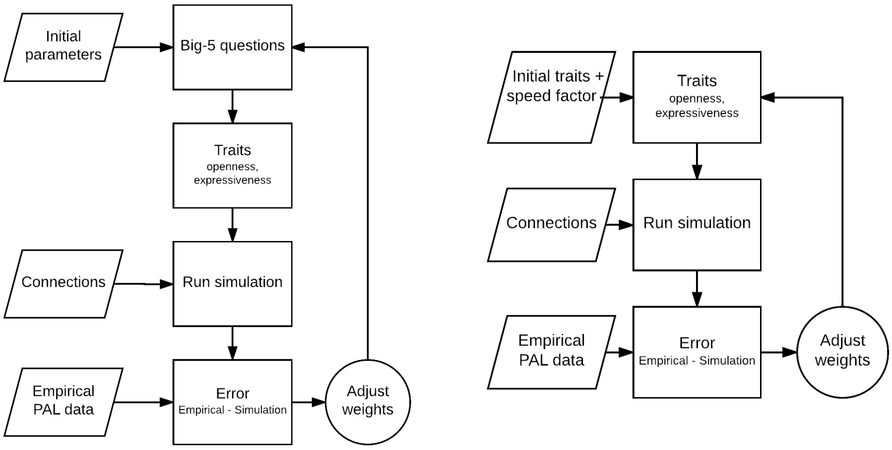

Fig. 3.1: Methodology for tuning the questionnaire questions weights (left). Methodology for tuning the traits for each participant (right).

3.5 Data

The data used is collected from an experiment described in [3]. The aim of the original experiment was to compare changes in the PAL of a group of young adults over a range of 30 days. The data was collected through questionnaires about the participants' personality traits and the kind of relationship that they have with each other. An adapted Big-5 Questionnaire [11, 15] was used to find traits of the participants, namely expressiveness and openness. The questions could be answered as totally disagree (0), disagree (0.25), neutral (0.5), agree (0.75) and totally agree (1). 6 of the questions were related to openness to new experiences, 6 to extraversion and 6 to agreeableness. We associate the dimension openness to new experiences to the trait openness, and the dimensions extraversion and agreeableness to the trait expressiveness.

The network consisted of 25 participants. They were asked to wear a Fitbit One device from 11/05/2015 until 09/06/2015. This device traced their activities and recorded the number of steps walked/ran daily, as well as the light, moderate and very active minutes over the day. Fitbit One is an activity tracker for measuring physical activity and has been used in many works with good reliability [17, 18]. Five participants provided less than 25 days of data, and were removed from the experiment.

The intake questionnaires were administered at the beginning of the experiment in order to collect (1) personal characteristics, (2) the level of friendship with other people, and (3) frequency of contact with other participants in person or in virtual environments like social media. Questionnaire (1) was used to define the traits of the participants, namely openness and expressiveness. This questionnaire is based on the Big-5 Inventory containing 18 questions related to the dimensions of openness, agreeableness and extraversion. Questionnaires (2) and (3) were used to calculate the strength of the connections between all the participants.

To calculate the PAL for the participants in the experiment, we used the Metabolic Equivalent of Task (MET) as a basis [12]. MET is the energy spent while performing physical activities. 1 MET is equivalent to the energy spent while seated at rest. Fitbit categorizes the daily active minutes of each user to lightly active, fairly active and very active minutes, based on the MET value associated to each physical activity that the user performs. We are calculating the daily PAL value for each participant as $PAL = 2 \times (lightly\_active) + 4 \times (moderately\_active) + 8 \times (very\_active)$. The PAL values divided by 1.500, which we consider as a maximum daily PAL value.

A few questions were given to the participants to verify (a) what kind of relationship they have with each other, (b) the frequency of their contact in real life or in private conversation through social media, and (c) the frequency of their contact in groups on social media including seeing posts by the others. The three questionnaires were normalized so the total would be in a range between 0 and 1. The overall connection strength was calculated as $connections = 5 \times (a) + 3 \times (b) + (c)$

3.6 Results

3.6.1 Tuning the weights of the questionnaire

For the optimization of the weights we have used a grid search algorithm. The algorithm variates the weights of each of the 18 questions between 0 and 1 in order to find the a set of weights that results in the most accurate simulation of our empirical data. Running the grid search algorithm for 82.000 iterations and all parameters we get a minimum error of 5.79.

When running the model with all the weights equal to 1 and a speed factor of 1, the error between the model and the empirical data is 5.97. The parameter tuning thus results in a model prediction that is closer to the empirical data. The small difference can be explained by the fact that the best traits found through the grid search are very close to the values of the traits when they are calculated using equal weights. Appendix A (`https://goo.gl/iNwbRG`) shows the traits obtained from the best fit set of weights found in comparison with the traits that are calculated based on equal weights for the questions (the second and fourth column for openness, and the fifth and seventh column for expressiveness). The average difference between the hyperopt algorithm and equal weights approach is 0.0993 for openness and 0.0955 for expressiveness, with low standard deviations (0.0835 for openness and 0.0442 for expressiveness). Although the traits resulting from the hyperopt algorithm are quite similar to the ones calculated with equal weights, they are very different than the standard values of 1.

To investigate whether this is only accidentally the case for the optimal solution or whether the algorithm actually converges towards useful weights, we selected the 100 solutions that resulted in the lowest error. Figure 3.2 shows the mean of the 100 best predictions for each of the weights and the standard deviation. It shows that the results are very stable.

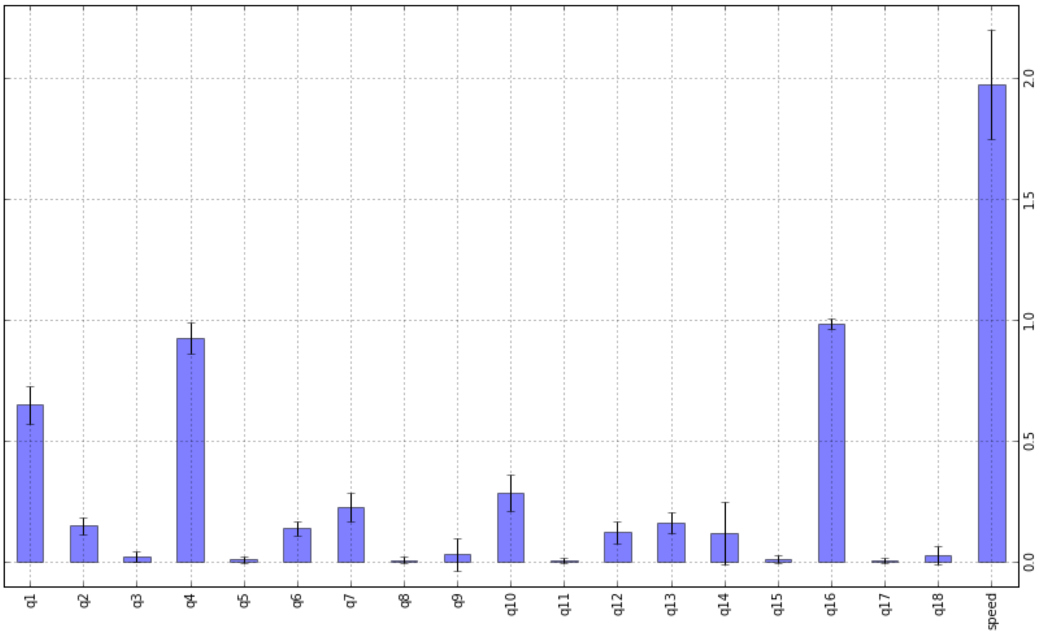

Fig. 3.2: Mean and standard deviation for the 100 best solutions using the grid search.

3.6.2 Tuning the traits for each participant

In the second phase we try to find the openness and expressiveness for each person directly by optimizing these values. Using the same hyperopt grid search algorithm we gathered 80.000 experiments and obtained **a minimum error of** 4.68. This value is better than using the grid search for the weights of the questions.

Appendix A (Available at `https://goo.gl/iNwbRG`) shows the best solutions for openness and expressiveness found by tuning the traits (third and sixth column) in comparison to the values found by tuning the weights and the equal weight approach. It shows that the values found by optimizing the personality traits are not as close to the values calculated based on the questionnaire. This is to be expected as a pure optimization algorithm will only be concerned about reducing the loss without considering the context.

To verify the variation of the best results found in the grid search, we also calculated the openness and expressiveness for the 100 best results obtained. The standard deviation for this scenario is bigger than the first scenario, as shown in Figures 3.3 (for openness) and 3.4 (for expressiveness). This shows that tuning the traits will give a larger range of values for the parameters with an approximate similar cost. The difference between the results obtained with the questionnaire and the best results obtained using a ML method are also very noticeable.

3.7 Discussion and conclusion

In this paper, we have explored two methods for deriving traits of people in a social experiment. The methods are evaluated with data from an experiment with 20

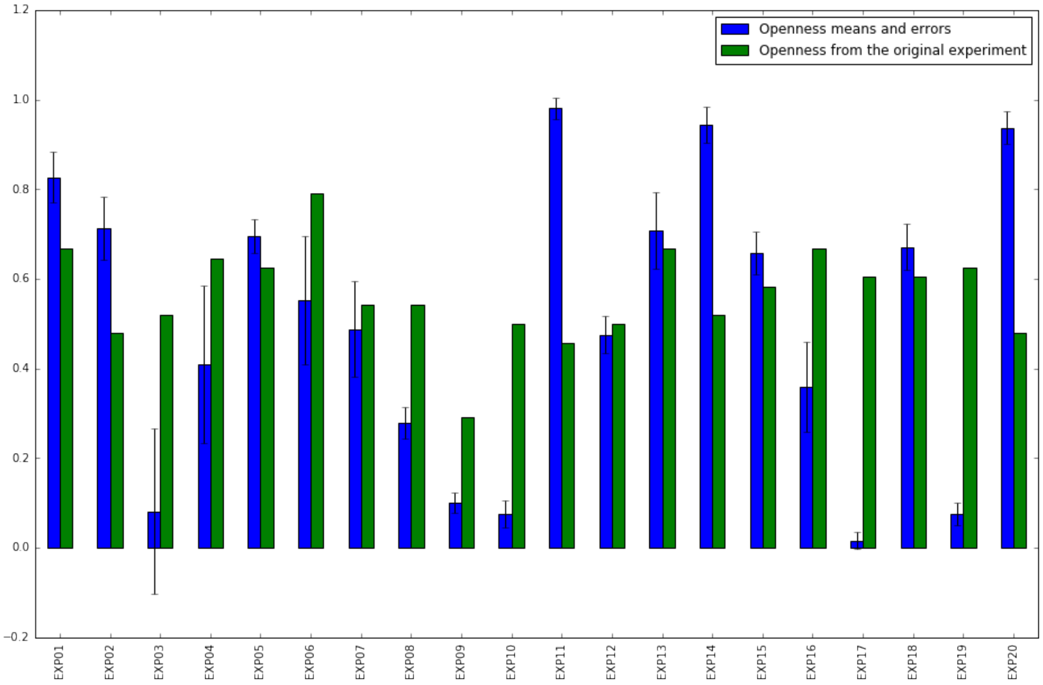

Fig. 3.3: Mean and error for the 100 best solutions using the grid search for openness (blue bars) and the solutions obtained with the questionnaires alone (green bars).

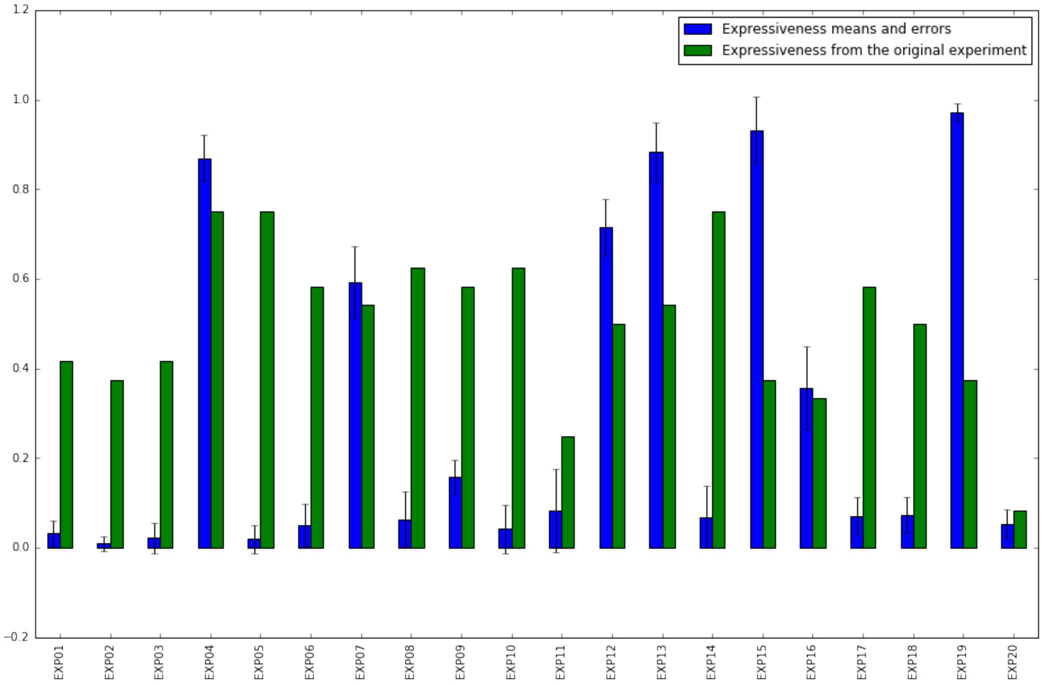

Fig. 3.4: Mean and error for the 100 best solutions using the grid search for expressiveness (blue bars) and the solutions obtained with the questionnaires alone (green bars).

young adults in a social network and their daily physical activity levels, combined with data from intake questionnaires used to gage their levels of openness and expressiveness and to quantify their relationships. A social contagion model based on differential equations is used to predict the PAL. The first method consisted in using a grid search algorithm to find the best *weights* for the 18 questions used in the self-report. The second method consisted of running a grid search algorithm to find the *personality traits* of each person directly.

The results of the first method are better than when using the original equal weights method. The results (see Appendix B at `https://goo.gl/iNwbRG`) show that 8 of the 18 questions provide little added value ($weight < 0.03$) for determining the traits of the persons. Moreover, only 5 questions seem particularly important ($weight > 0.2$). It is also shown that the results are stable for the 100 best solutions. Altogether, the outcome of this experiment suggests that our first method is a useful mechanism for optimizing complex questionnaires that are meant to reveal personality characteristics.

The results are less positive for the second method. Although the error is lower than using the grid search algorithm to find optimal values for the traits directly, the values are very different from the values derived from the questionnaire. Moreover, an analysis of the 100 best solutions show that the values are also quite diverse. This can be explained as a computationally optimal solution does not necessarily coincide with correct interpretation of the traits.

Our case study could still be improved in several ways. A bigger data set could provide stronger results. Other applications can use the same methodology, and new case studies could help to unfold other results and therefore improve the understanding of the advantages and limitations of these methods.

Bibliography

[1] F Alam, E A Stepanov, and Giuseppe Riccardi. „Personality traits recognition on social network-facebook“. In: *WCPR (ICWSM-13), Cambridge, MA, USA* (2013) (cit. on p. 42).

[2] Eric F. M. Araújo and Jan Treur. „Analysis and Refinement of a Temporal-Causal Network Model for Absorption of Emotions“. In: *International Conference on Computational Collective Intelligence*. Springer International Publishing. 2016 (cit. on p. 44).

[3] Eric FM Araújo, Anita VTT Tran, Julia S Mollee, and Michel CA Klein. „Analysis and evaluation of social contagion of physical activity in a group of young adults“. In: *Proceedings of the ASE BigData & SocialInformatics 2015*. ACM. 2015, p. 31 (cit. on pp. 43, 45).

[4] James Bergstra, Brent Komer, Chris Eliasmith, Dan Yamins, and David D Cox. „Hyperopt: a python library for model selection and hyperparameter optimization“. In: *Computational Science & Discovery* 8.1 (2015), p. 014008 (cit. on p. 44).

[5] Lisa F Berkman and S Leonard Syme. „Social networks, host resistance, and mortality: a nine-year follow-up study of Alameda County residents“. In: *American journal of Epidemiology* 109.2 (1979), pp. 186–204 (cit. on p. 42).

[6] Tibor Bosse, Rob Duell, Zulfiqar Memon, Jan Treur, and C van der Wal. „A multi-agent model for emotion contagion spirals integrated within a supporting ambient agent model“. In: *Principles of practice in multi-agent systems* (2009) (cit. on p. 43).

[7] Bernard C. K. Choi and Anita W. P. Pak. „A catalog of biases in questionnaires“. In: *Preventing chronic disease* 2.1 (Jan. 2005), A13 (cit. on p. 42).

[8] Nicholas A Christakis and James H Fowler. „Social contagion theory: examining dynamic social networks and human behavior“. In: *Statistics in medicine* 32.4 (2013), pp. 556–577 (cit. on p. 42).

[9] Funda Durupinar, Jan Allbeck, Nuria Pelechano, and Norman Badler. „Creating Crowd Variation with the OCEAN Personality Model“. In: *Proceedings of the 7th International Joint Conference on Autonomous Agents and Multiagent Systems - Volume 3*. AAMAS ’08. Estoril, Portugal: International Foundation for Autonomous Agents and Multiagent Systems, 2008, pp. 1217–1220 (cit. on p. 42).

[10] Katherine Ellis, Suneeta Godbole, Simon Marshall, et al. „Identifying active travel behaviors in challenging environments using GPS, accelerometers, and machine learning algorithms“. In: *Frontiers in public health* 2 (2014) (cit. on p. 42).

[11] Samuel D Gosling, Peter J Rentfrow, and William B Swann. „A very brief measure of the Big-Five personality domains“. In: *Journal of Research in personality* 37.6 (2003), pp. 504–528 (cit. on p. 45).

[12] M Jette, K Sidney, and G Blümchen. „Metabolic equivalents (METS) in exercise testing, exercise prescription, and evaluation of functional capacity“. In: *Clinical cardiology* 13.8 (1990), pp. 555–565 (cit. on p. 46).

[13] Jean K Langlie. „Social networks, health beliefs, and preventive health behavior“. In: *Journal of health and social behavior* (1977), pp. 244–260 (cit. on p. 42).

[14] Paul Marsden. „Memetics and social contagion: Two sides of the same coin“. In: *Journal of Memetics-Evolutionary Models of Information Transmission* 2.2 (1998) (cit. on p. 43).

[15] Gerald Matthews, Ian J Deary, and Martha C Whiteman. *Personality traits*. Cambridge University Press, 2003 (cit. on p. 45).

[16] Julianna Pacheco. „The social contagion model: Exploring the role of public opinion on the diffusion of antismoking legislation across the American states“. In: *The Journal of Politics* 74.1 (2012), pp. 187–202 (cit. on p. 42).

[17] Serene S Paul, Anne Tiedemann, Leanne M Hassett, et al. „Validity of the Fitbit activity tracker for measuring steps in community-dwelling older adults“. In: *BMJ open sport & exercise medicine* 1.1 (2015), e000013 (cit. on p. 45).

[18] Judit Takacs, Courtney L Pollock, Jerrad R Guenther, et al. „Validation of the Fitbit One activity monitor device during treadmill walking“. In: *Journal of Science and Medicine in Sport* 17.5 (2014), pp. 496–500 (cit. on p. 45).

[19] Alex Hai Wang. „Detecting Spam Bots in Online Social Networking Sites: A Machine Learning Approach.“ In: *DBSec* 10 (2010), pp. 335–342 (cit. on p. 42).