Contagious

MODELING THE SPREAD OF BEHAVIOURS, PERCEPTIONS AND EMOTIONS IN SOCIAL NETWORKS

ERIC FERNANDES DE MELLO ARAÚJO

Contagious: Modeling the Spread of Behaviours, Perceptions and Emotions in Social Networks

Eric Fernandes de Mello Araújo

VU UNIVERSITY AMSTERDAM

Thesis Reading Committee:

prof.dr. Tibor Bosse	Radboud University, The Netherlands *Behavioural Science Institute*
prof.dr. Andreas Flache	University of Groningen *Department of Sociology*
prof.dr. Peter Kerkhof	Vrije Universiteit Amsterdam *Faculty of Social Sciences*
dr. Nicola Perra	University of Greenwich Business School *Centre for Business Network Analysis*
prof.dr. Hajo Reijers	Vrije Universiteit Amsterdam *Department of Computer Science* Eindhoven University of Technology *Department of Mathematics and Computer Science*

This research is financed by the Brazilian Science without Borders (SwB) program, through the Coordination for the Improvement of Higher Education Personnel, CAPES (reference 13538-13-6).

SIKS Dissertation Series No. 2018-22
The research reported in this thesis has been carried out under the auspices of SIKS, the Dutch Research School for Information and Knowledge Systems.

Cover design by Eric Araújo. The convex hulls' graphic in the background of the cover was plotted in R. The code for the graphic was created by Marcus Volz, who kindly gave permission for its use in this cover.

ISBN: 978-94-028-1117-9

VRIJE UNIVERSITEIT

CONTAGIOUS: MODELING THE SPREAD OF BEHAVIOURS, PERCEPTIONS AND EMOTIONS IN SOCIAL NETWORKS

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad Doctor of Philosophy
aan de Vrije Universiteit Amsterdam,
op gezag van de rector magnificus
prof.dr. V. Subramaniam,
in het openbaar te verdedigen
ten overstaan van de promotiecommissie
van de Exacte Wetenschappen
op woensdag 5 september 2018 om 9.45 uur
in de aula van de universiteit,
De Boelelaan 1105

door

Eric Fernandes de Mello Araújo

geboren te Belo Horizonte, Brazilië

promotoren:	prof.dr. J. Treur
	prof.dr. A. van Halteren
copromotor:	dr. M.C.A. Klein

Summary

It has been shown that social interactions can affect a person's behaviours, perceptions and emotions in many ways, from contagion of emotions to obesity spread. The quantification of the spread of behaviours, perceptions and emotions is a task that requires an understanding of the social contagion phenomenon, as well as good methods for data collection. Social contagion stands for the effect caused by our relationships, on our own identity, character, decisions, opinions, positioning, emotions, etc. It is a process that happens unconsciously and naturally throughout our entire lives.

Fortunately, the data provided by the new web technologies recently created, combined with the advances in devices such as smartphones, physical activity trackers and other sensors, are a good source for investigation into how people affect each other, and how their connections are shaping their behaviours, perceptions and emotions. On the other hand, understanding human behaviour in order to model and predict future states can be considered a very complex task, as it requires a multidisciplinary approach and very strong methods to validate the whole process.

This thesis aims to understand, model and predict different sorts of behaviours, perceptions and emotions through cognitive models and social contagion in social networks. The models developed here can be applied broadly, for example from the promotion of a healthy lifestyle to the reactions to web media posts.

We firstly explore a social contagion model that accounts for the spread of behaviours, perceptions and emotions in social networks. The model uses differential equations and a temporal-causal approach to describe the different scenarios studied. The model is then used for validation attempts in different data sets. The data sets used here contain physical activity behaviour information of different groups of people and the social network of the individuals participating in the experiments. Besides validating the model, we also try to simulate possible interventions and verify what are the side effects of changing the network states. We also use statistical analyses to explain changes in the behaviour of different groups of people, i.e. connected and non connected individuals.

Many tasks were necessary to create a realistic representation of the social contagion effect in the data collected. We used parameter tuning in many cases to define the traits of the individuals in the network, or to adjust speed factors, thresholds and other characteristics required for the chosen models. In addition, social network analyses were performed in a few studies to understand the dynamics of the social networks where the spread was happening.

After exploring the social contagion model for physical activity behaviour, we show that it is possible to extend the mathematical model to other scenarios where the social contagion phenomenon is relevant. The first context is the spread of messages in disaster situations. This scenario is built on a context of how people receive a notice about some ongoing disaster, and how the sender and the means affect the credibility of the message transmitted. The second context is in a web media (Twitter) interaction with political posts, and how the posts affect the positioning of a person.

This thesis provides a significant contribution to the state of the art on social contagion modeling and on behavioural informatics studies. It also presents methods useful to tackle the challenges of data collection, analysis and fine tuning of models for the spread of behaviour in social networks. Therefore, we believe that many aspects of it can be derived from this work in potential applications aiming to improve the lifestyle of different groups of people by understanding, modeling and the simulation of temporal-causal network models.

Samenvatting

Het is aangetoond dat sociale interacties op vele manieren invloed kan hebben op iemands gedrag, percepties en emoties, van de besmetting van emoties tot de verspreiding van obesitas. De kwantificering van de verspreiding van gedrag, percepties en emoties is een taak dat een goed begrip van het fenomeen sociale besmetting vereist, evenals goede methoden voor het verzamelen van gegevens. Sociale besmetting staat voor het effect dat veroorzaakt wordt door onze relaties, onze eigen identiteit, karakter, beslissingen, meningen, positionering, emoties, etc. Het is een proces dat zich onbewust en natuurlijk gedurende ons hele leven afspeelt.

Gelukkig is de data van nieuwe web technologieën die onlangs zijn gemaakt, gecombineerd met de vooruitgang op het gebied van apparaten zoals smartphones, trackings voor fysieke activiteit en andere sensoren, een goede bron voor een onderzoek naar de manier waarop mensen elkaar beïnvloeden en hoe hun relaties vormgeven aan hun gedrag, percepties en emoties. Aan de andere kant kan het begrijpen van menselijk gedrag om toekomstige staten te modelleren en te voorspellen als een zeer complexe taak worden beschouwd, omdat het een multidisciplinaire aanpak en zeer sterke methoden vereist om het hele proces te valideren.

Deze scriptie heeft tot doel om verschillende soorten gedrag, percepties en emoties te begrijpen, te modelleren en te voorspellen door middel van cognitieve modellen en sociale besmetting in sociale netwerken. De hier ontwikkelde modellen kunnen breed worden toegepast, bijvoorbeeld van het promoten van een gezonde levensstijl tot de reacties op webmediaberichten.

We verkennen eerst een sociaal besmettingsmodel dat verantwoordelijk is voor de verspreiding van gedrag, percepties en emoties in sociale netwerken. Het model maakt gebruik van differentiaalvergelijkingen en een temporeel-causale benadering om de verschillende bestudeerde scenario's te beschrijven. Het model wordt vervolgens gebruikt voor validatiepogingen in verschillende datasets. De datasets die worden gebruikt, bevatten informatie over de lichaamsbeweging van verschillende groepen mensen en het sociale netwerk van de individuen die deelnemen aan de experimenten. Naast het valideren van het model, proberen we ook mogelijke interventies te simuleren en na te gaan wat de bijwerkingen zijn bij het veranderen van de netwerkstaten. We gebruiken ook statistische analyses om veranderingen in het gedrag van verschillende groepen mensen, dat wil zeggen verbonden en niet-verbonden personen, te verklaren.

Er waren veel taken nodig om een realistisch beeld te krijgen van het sociale besmettingseffect in de verzamelde gegevens. We hebben in veel gevallen parameterafstemming gebruikt om de eigenschappen van de personen in het netwerk te definiëren

of om snelheidsfactoren, drempelwaarden en andere kenmerken aan te passen die voor de gekozen modellen zijn vereist. Daarnaast werden sociale netwerkanalyses uitgevoerd in enkele studies om de dynamiek te begrijpen van de sociale netwerken waar de verspreiding plaatsvond.

Na het verkennen van het sociale besmettingsmodel voor lichaamsbeweging, laten we zien dat het mogelijk is om het wiskundige model uit te breiden naar andere scenario's waarin het fenomeen van sociale besmetting relevant is. De eerste context is de verspreiding van berichten in rampsituaties. Dit scenario is gebaseerd op een context van hoe mensen een melding ontvangen over een aantal lopende rampen en hoe de afzender en de middelen de geloofwaardigheid van het verzonden bericht beïnvloeden. De tweede context is in een interactie via sociale media (Twitter) met politieke berichten, en hoe de berichten van invloed zijn op de positionering van een persoon.

Deze scriptie levert een belangrijke bijdrage aan de stand van de techniek op het gebied van sociale besmettingsmodellering en op gedragsinformatiestudies. Het bevat ook methoden die nuttig zijn om de uitdagingen aan te gaan van gegevensverzameling, analyse en de afstemming van het model voor de verspreiding van gedrag in sociale netwerken. Daarom zijn wij van mening dat vele aspecten ervan kunnen worden afgeleid van dit werk in potentiële toepassingen die gericht zijn op het verbeteren van de levensstijl van verschillende groepen mensen door inzicht in, modellering en de simulatie van temporeel-causale netwerkmodellen.

Sumário

Sabe-se que as interações sociais podem afetar os comportamentos, percepções e emoções das pessoas de diferentes formas, do contágio de emoções até a propagação de obesidade. A quantificação da propagação de comportamentos, percepções e emoções é uma tarefa que requer uma compreensão do fenômeno de contágio social, assim como bons métodos para a coleta de dados. Contágio social é definido como o efeito causado por nossos relacionamentos em nossa própria identidade, caráter, decisões, opiniões, posicionamentos, emoções, etc. É um processo que ocorre inconscientemente e de forma natural ao longo de nossas vidas.

Felizmente, os dados disponíveis graças às novas tecnologias recentemente criadas, em combinação com os avanços nos dispositivos móveis como smartphones, rastreadores de atividade física e outros sensores são uma boa fonte para investigações sobre como as pessoas se afetam, e como as suas conexões moldam seu comportamento, percepções e emoções. Por outro lado, entender o comportamento humano visando modelar e prever estados futuros pode ser considerado como uma tarefa muito complexa, uma vez que requer uma abordagem multidisciplinar e métodos bem definidos para validar todo o processo.

Esta tese tem como objetivo entender, modelar e prever diferentes tipos de comportamentos, percepções e emoções através de modelos cognitivos e contágio social em redes sociais. Os modelos desenvolvidos aqui podem ser aplicados amplamente, por exemplo, para a promoção de estilo de vida saudável até reações a postagens em mídia sociais.

Inicialmente o modelo de contágio social é explorado. Este modelo é responsável por definir a propagação de comportamentos, percepções e emoções em redes sociais. O modelo usa equações diferenciais e uma abordagem temporal-causal para descrever os diferentes cenários estudados. O modelo é então usado para tentativas de validação em diferentes conjuntos de dados. Os conjuntos de dados usados aqui contem informações sobre o comportamento relacionado à atividade física de diferentes grupos de pessoas, bem como a rede social dos indivíduos participantes dos experimentos. Além de validar o modelo, tenta-se simular possíveis intervenções e verificar quais são os efeitos colaterais da alteração dos estados da rede. Também são utilizadas análises estatísticas para explicar as mudanças no comportamento de diferentes grupos de pessoas, por exemplo indivíduos conectados ou não-conectados.

Várias tarefas foram necessárias para criar uma representação realista do efeito de contágio social. O uso de algoritmos para ajuste de parâmetros (parameter tuning) foi usado em alguns casos para definir as características dos indivíduos na rede, ou

para ajustar os fatores de velocidade, os limiares e outras características relevantes para o modelo escolhido. Ademais, análise de redes sociais foram feitas em alguns dos estudos para entender a dinâmica das redes sociais aonde a propagação estava ocorrendo.

Após explorar o modelo de contágio social para o comportamento de atividade física, é demonstrado que o modelo matemático é extensível a outros cenários aonde o fenômeno de contágio social é predominante. O primeiro contexto é na propagação de mensagens em situações de desastre. Este cenário é construído em um contexto onde as pessoas recebem a informação de um desastre em ação, e como o emissor da mensagem e os meios de comunicação afetam a credibilidade do conteúdo transmitido. O segundo contexto é o da interação em uma mídia social (Twitter) com postagens sobre política, e como essas postagens afetam o posicionamento de uma pessoa.

Esta tese contém contribuição significativa para o estado da arte em modelagem de contágio social e em estudos de informática comportamental. Também apresenta métodos úteis para atacar os desafios de coleta de dados, análise e ajuste de parâmetros em modelos de propagação de comportamento em redes sociais. Desta forma, acredita-se que vários dos aspectos do trabalho aqui apresentado podem ser derivados para outras aplicações visando melhorar a qualidade de vida de diferentes grupos de pessoas por meio da compreensão, modelagem e simulação de modelos de redes temporais-causais.

Acknowledgements

The acknowledgments section is one of the few places in a scientific work that allows the author to express a little more emotion, or even to hint at their poetic side. The story of this work began around 33 years ago, and would require a whole new book to record the account in full. This PhD is more than just a collection of papers, it is a dream being realized. It is the peak of a long journey of learning, challenges, victories and losses that started with my family in my home country of Brazil. In order to not bore the reader too much, I'll try to make a long story short (well, at least I can say: I tried!). I'm sorry if it is too long for some of you. It's just that it means a lot to me to have this chance to say how important each of these people are in my life, and how they have helped me to get to this point.

Nothing would have happened if it wasn't for the humble people early on in my life, who didn't have an easy life, and who struggled and sacrificed a lot in order to sustain my studies. I have a hero in the figure of my dad (*pai*) Enos, who, besides telling the worst jokes ever, is the one who always pushed me towards my limits. He believed in me in the moments when I thought I'd reached my limits, encouraging me to persevere. He also taught me that hard work and honesty is always the best path to take. My mom (*mãe*) Sandra gave me the books to read as well as a passion for devouring thick books, and she challenged me to dream high. She always surrounded me with compassion and a sensitive heart. My second mother, Ritinha, stepped in a bit later, when I began my teenage years. She taught me discipline and the importance of planning well in life. With each of you in my life, I have been able to climb step-by-step in my academic career, and always received encouragement and so have had the stamina to move forward when needed. To you all, I give my huge gratitude, admiration and love.

To the two lovely people with whom I spent much of my childhood, my sisters, Bruna and Samantha, you are the best sisters I could ever wish for. Strong women, who dare to live out their lives fully in what they seek to accomplish. I find so much encouragement from you, that it would take the writing of a book to express it all! You are both incredible moms, who, despite all the adversities in your lives, have yet thrived to be where you are today. I love you with all my heart, and dedicate this victory to you, especially considering your putting up with such an annoying older brother for so many years. You are both really brave!

As if doing a PhD wasn't adventurous enough, during the second year of my time in the Netherlands, I met this most beautiful lady in the community of Christ Church South (an international Anglican congregation in Amsterdam), and Amy joined me to journey together in life. You are the person with the most beautiful heart I've ever met in my life, with such an incredible love and compassion. You are so

understanding and the best support I could possibly have when the pressure was mounting. Amy, my lovely wife, I dedicate this work to you, all the 'bloed, zweet en tranen'. You made my way to accomplish this victory smoother, always being patient and kind. I love you more than I can say.

I cannot express how much gratitude I have for Amy's family. You embraced me when I joined the family around 2 years ago. I am so thankful to God for allowing me to be part of your beautiful family. I've learned so much from you. I love you! Dan, thanks for translating my abstract to Dutch. You have a really promising career ahead. Count on me for anything you need! Mandy and Alan, you both inspire me so much and in so many ways. You are also like parents to me, and I consider every word you say to me. Thanks for receiving me as a son, and for being so warm and kind during these past years. Thanks to Kev and Tonya for all the support as well, and for making my first trip to London so memorable and enjoyable.

I have the utmost respect, admiration and gratitude for my supervisors. Jan, thank you so much for teaching me how to do serious research and how to conduct myself in the academic context. I just feel so proud for having the chance to learn from you and to grow academically with your model. Michel, thanks for your always optimistic feedback and motivation, especially when I'd enter your office in despair not knowing what to do. Your patience and kindness are remarkable, and were very important in key moments of the PhD. Your input to my work, the discussions about possible research questions and new approaches to the many challenges, were essential to permit me to deliver this work with a better quality. Dank je wel! Aart, thanks for the enthusiasm in our discussions about the million things that we could do, and for providing resources and input that greatly improved the quality of my work. I really appreciate learning from your experience, and I hope to use this knowledge to boost my research in the near future.

The remarks and feedback given by the reading committee were also very constructive and insightful to improve the quality of this work. Therefore, thank you Hajo Reijers, Tibor Bosse, Andreas Flache, Nicola Perra, and Peter Kerkhof for taking precious time to read, evaluate and provide valuable comments on my work.

In Brasil (sorry, it's not a typo, it will always be 'Brasil' in my heart...), I've had so many people support me with their friendship and advice, with their love and wisdom, with prayers and kind words at important moments. I must not forget to thank my uncle Evandro, for being a rock throughout my life, spending many hours over the years talking and sharing about my conflicts, my dreams, my problems, and my heart and soul. I love you, and you are part of this. My grandma Maria, who prays for me every single day. So many times I passed by your window and heard you praying with such a devotion that really fires up my heart and strengthens my faith! To my granddad José, your strong character and sweet smile are always a welcome picture in my heart when I'm missing you for being so far away. To my friends who have always cheered me up in the sad times of my life, and celebrated with me in the good times. Amanda, Armando, Araci, Kelly, Tiago Fumaça, Breno Fumacinha, Moisés, Leyber, Eduardo and Natalie van den Berg, Jamilly, Francisval and Ismênia, you are amazing friends for life, and you've made my life so much better in these past years. I love you! A big thanks to L'Abri Brasil. L'Abri is a place where I am able to really find space to calm down my mental and emotional storms, and I just feel so at home there. Thanks Vanessa, Guilherme and Rodolfo for providing such an environment where I can nurture my intellect, my faith and my character.

This section is quite long already. I tried to keep it short. I'm afraid I failed, because there are more people to say thank you to!

I'd like to extend my gratitude to the UFLA (Federal University of Lavras). The effort of the direction of the University (especially prof. Scolforo and prof. Édila) on providing a substitute to cover my lectures whilst I was doing my PhD, and the support of the chief of the Computer Science Department, Renato, were fundamental in the process of my being able to be fully dedicated to the studies and research during these past years.

Christ Church Amsterdam South has been a very special place to me during these years. This is where I met my wife, and where I have found friends for life, who would do anything to serve my family and to show their love for us. I thank you Benjie and Leigh Ann for your sincere friendship. You guys rock! Thank you Cristina and Andre, for all the wine tasting we did together, for the long talks and for receiving me as part of your family. Thanks Alan and Natalie, for your very kind hearts, for Alan's brilliant sermons which would refresh my faith and my hope, and help me to locate myself in the midst of the messiness of life. Thanks Natalie for your sweet heart and for providing Amy with such a beautiful friendship that helps us both to grow together. Thank you Joy Buswell for being so kind and for strengthening my belief in humanity through the example of your life. To all the members of this lovely community, my gratitude and love to you! You have made my life in Amsterdam so much richer!

When I first arrived in the Netherlands, I was warmly welcomed and well taken care of by the ISMA group (International Student Meetingpoint Amsterdam). Francina, Pieter, Eveline, Jenn, Wil and Stefan were a huge blessing in my life. Thanks Francine for being so kind and enthusiastic about ISMA, and for allowing me to be a part of it for a few years. Pieter and Eveline, I just love you so much, my Dutch family! I will never forget your help taking me to the 'Dorcas' (second hand furniture store), bringing me there by car, and stopping the car en route just so I could take a photo of a windmill (I had never seen one before!). I'm grateful for Pieter's horrible jokes, for the moments shared in your house, for your presence at my wedding as witnesses, for the wedding speech, and for all the joy I (and later Amy also) shared with your beautiful family. All my love to you, to Pim and Pascal for the many times allowing me to feel at home in a strange country. Jenn, my dear, I want to say that you are a true example of mercy and perseverance. I've learned to admire you through these 4 years of friendship, and I'm really thankful for your listening ear, your time and your support in so many moments. I had so much fun serving ISMA with you. Stefan! You helped me fix my first IKEA furniture! So much has happened since then. Thanks for taking me to the beach to take some photos, and for giving me rides to various IFES events. Your friendship is precious to me, and I really appreciate all you've done to help me adapt in the Netherlands. May God bless you and your beautiful family! Thank you all ISMA/IFES people!

Soon after arriving in The Netherlands I met a group of Brazilians who became very good friends helping me with my *saudade* (there is no translation for this word... sorry) of Brasil. Cláudia, Gabi do Frevo, Gabi Manezinha da Ilha, Lenin, Daniel and Debora, Fernando, Monica and Clara, Veruska, Ana Clara Mathias, and Ermerson. Thank you all for so many good times together!

We have a great research group, formerly known as ASR (Agent Systems Research) and recently named BI (Behavioural Informatics Group), the new name which, besides being much fancier, reflects more accurately the kind of research I love to do. As you will see in the pages of this work, I wasn't alone in the quest to unveil the mysteries of human behaviour. I'd like to say thanks to Adnan and Julienka, with whom I shared most of the work done here. I've grown so much working with you, and I thank you for being patient enough to work with me despite my limitations and flaws. Bojan, my dear, it was such a pleasure to write some papers with you (a few to go still...). Thanks for always being so enthusiastic and for cheering me up to keep the good work going! Never forget to restart your Jupyter notebooks! That's precious! I also want to say thank you to the undergraduate students that I had the chance to supervise and who delivered work beyond what was expected from them. Anita, Rukshar, Annelore and Dave, thanks for allowing me to guide you in your final projects, and thanks for believing that we would make an international paper out of it. It has also been a real privilege to work with the team in Nijmegen, especially Kris, Thabo, Moniek, Crystal and Laura. My utmost gratitude for the provision of the precious data and the opportunity to work in collaboration with you. Thanks to the BI group team for all the exchanges, be they scientific, or simply going to grab a coffee on the 4th floor (and sometimes borrowing coffee from the maths department...). Amin, Altaf, Daniel and Lenin (the brs), Sahand, Fawad, Fakhra, Ali, Dilhan, Laura and Romy (may the girl power be with you two!), Natalie. Thank you!

I'd also like to say thanks to Charlotte my office buddy on Wednesdays. Wednesdays became my favourite day of the work week because Charlotte would always be a good companion for having encouraging talks and motivating each other's productivity. Thank you for all the advice, for providing me with a list of the best conferences, and for sharing information that only people who have already passed through this whole PhD process can give.

A special thanks to Mojca, our secretary. Since I started my work you have always been prompt to help and are super efficient and professional. I'd like to say thank you for your service, for the abundance of chocolates, the "Bom dia"s and the "Tudo bem?"s. It was always nice to step by your office. I will miss our chats!

Even though I have a long list of people to whom I am very grateful, I may have forgotten some of you. If that's your case, just please forgive me and also feel a part of this project as well, and celebrate with me!

Last, but not least, I want to praise Jesus Christ, for His example of life, for His teachings, and that through faith in Him, I can persevere in seeking the good in humanity. His example encourages me to thrive in the academic career, and keeps me strong in my intent to serve humanity with wisdom, justice, love and compassion.

Contents

I Foundations 1

1 Introduction 3
1.1 Motivation . . . 3
1.2 Background . . . 4
1.2.1 Data Analysis . . . 4
1.2.2 Social contagion . . . 6
1.2.3 Modeling . . . 7
1.2.4 Health and Social networks . . . 10
1.3 Research questions . . . 11
1.4 Research Methodology . . . 12
1.4.1 Systematic review . . . 12
1.4.2 Data collection, processing and analysis . . . 13
1.4.3 Social network analysis . . . 14
1.4.4 Computational modeling . . . 14
1.5 Structure of this thesis . . . 15
1.5.1 Part I: Foundations . . . 15
1.5.2 Part II: Data Analytics in Networks . . . 16
1.5.3 Part III: Using Social Contagion Models for Explaining Physical Activity . . . 16
1.5.4 Part IV: Use of Contagion Models for Perceptions and Emotions 17
1.5.5 Part V: Discussion and evaluation . . . 17

2 Refinement of a Temporal-Causal Network Model 23
2.1 Introduction . . . 24
2.2 Emotion absorption: the temporal-causal network model . . . 24
2.3 Analysis of the absorption model . . . 27
2.3.1 Analysis of the original model . . . 28
2.3.2 Mathematical analysis of the problem . . . 29
2.4 Alternative ways to model the speed factor η_A . . . 30
2.5 Mathematical Analysis . . . 32
2.6 Results . . . 34
2.7 Conclusion . . . 36

II Data Analytics in Networks 39

3 Applying machine learning algorithms for deriving personality traits in social network 41
3.1 Introduction . . . 42
3.2 Related work . . . 42
3.3 Social Contagion Model . . . 43

3.4 Methods . . . 44
3.5 Data . . . 45
3.6 Results . . . 46
3.6.1 Tuning the weights of the questionnaire . . . 46
3.6.2 Tuning the traits for each participant . . . 47
3.7 Discussion and conclusion . . . 47

4 Analysis of Social Contagion in a Social Network 53
4.1 Introduction . . . 54
4.2 Social Contagion Model . . . 55
4.3 Experimental setup . . . 56
4.3.1 Physical Activity Level . . . 57
4.3.2 Tie Characteristics . . . 57
4.3.3 Personality Traits . . . 58
4.3.4 Network Structure . . . 59
4.3.5 Data Preparation & Analysis . . . 60
4.4 Results . . . 61
4.5 Discussion . . . 63
4.6 Conclusions . . . 64

5 The effect of a community in a health promotion program 69
5.1 Introduction . . . 70
5.1.1 Motivation . . . 70
5.1.2 Aim . . . 70
5.1.3 Related work . . . 71
5.2 Methods . . . 72
5.2.1 Data collection . . . 72
5.2.2 Data selection . . . 72
5.3 Results . . . 79
5.3.1 Visual comparison . . . 79
5.3.2 Multiple linear regression model . . . 79
5.3.3 Linear mixed model . . . 81
5.4 Discussion . . . 84
5.4.1 Principal Results . . . 84
5.4.2 Interpretation . . . 85
5.4.3 Conclusions . . . 86

6 Social Connection Dynamics in a Health Promotion Network 91
6.1 Introduction . . . 92
6.2 Dynamical Social Network Analysis . . . 92
6.3 Methods . . . 93
6.3.1 Data Set and Data Selection . . . 93
6.3.2 Social Network Analysis . . . 94
6.4 Results . . . 95
6.4.1 Nodes, edges and degree distribution . . . 96
6.4.2 Largest component and other components . . . 97
6.4.3 Centrality measurements . . . 98
6.4.4 Homophily . . . 100
6.4.5 Identifying influential participants . . . 101
6.5 Conclusions . . . 102

III Using Social Contagion Models for Explaining Physical Activity 107

7 Explaining changes in physical activity through a computational model of social contagion 109
7.1 Introduction . . . 110
7.2 Background . . . 111
7.3 Methods . . . 112
7.3.1 Data collection . . . 112
7.3.2 Data preprocessing . . . 112
7.3.3 Model simulations . . . 113
7.3.4 Analyses . . . 115
7.4 Results . . . 115
7.5 Conclusions . . . 116

8 Using Simulations for Exploring Interventions in Social Networks 121
8.1 Introduction . . . 122
8.2 Background . . . 123
8.3 Methods . . . 124
8.3.1 The data . . . 125
8.3.2 Model implementation . . . 125
8.3.3 Building the network . . . 127
8.3.4 Agents characteristics . . . 129
8.3.5 Parameter tuning . . . 131
8.3.6 Strategies for selecting targets for intervention . . . 131
8.4 Results . . . 134
8.4.1 Tuning of threshold and change factor . . . 134
8.4.2 Exploring different strategies . . . 136
8.5 Discussion . . . 138
8.6 Conclusion . . . 139

IV Use of Contagion Models for Perceptions and Emotions 145

9 A Temporal-Causal Model for Spread of Messages in Disasters 147
9.1 Introduction . . . 148
9.2 Social Contagion and Behaviour in Disaster Situations . . . 148
9.3 Agent-based model . . . 149
9.4 Data generation and parameter tuning . . . 152
9.4.1 Experimental datasets . . . 152
9.4.2 Parameter tuning methods . . . 154
9.5 Simulations and Results . . . 155
9.5.1 Scenario 1: Receiving a tensed and negative phone call . . . 155
9.5.2 Scenario 2: Receiving a happy and positive text message . . . 156
9.5.3 Scenario 3: Outputs as inputs . . . 158
9.5.4 Mathematical analysis of the model . . . 159
9.6 Discussions and future works . . . 160

10 A Computational Cognitive Model for Political Positioning and Reactions in Web Media 165
10.1 Introduction . . . 166
10.2 Psychological aspects of Political Behaviour . . . 166

10.2.1 Political Psychology . . . 167
10.2.2 Political Neuroscience . . . 169
10.3 The Effect of Mood and Persuasion on Interpretation of Messages . . 170
10.4 Agent-based model for political positioning on web media . . . 171
10.5 Simulations . . . 176
10.6 Discussion and future work . . . 180

11 Detecting Dutch political tweets **187**
11.1 Introduction . . . 188
11.2 Related work . . . 189
11.3 Collecting and processing the tweets . . . 190
11.3.1 Collecting data . . . 190
11.3.2 Cleaning the data . . . 191
11.3.3 Tagging the tweets . . . 192
11.3.4 Rule-based method . . . 192
11.4 Structure of the classifier . . . 193
11.5 Results . . . 196
11.5.1 Rule-based method . . . 196
11.5.2 Supervised learning . . . 196
11.6 Discussion . . . 198

V Discussion and Evaluation **203**

12 Discussions and Conclusion **205**
12.1 Discussion of the contributions of this thesis . . . 205
12.2 Research questions . . . 209
12.2.1 Research question 1 . . . 209
12.2.2 Research question 2 . . . 211
12.2.3 Research question 3 . . . 212
12.3 Limitations of this work . . . 213
12.4 Future Work . . . 214
12.5 Contributions and chapters overview . . . 216

Part I

Foundations

"Then two wonders happened at the same moment. One was that the voice was suddenly joined by other voices; more voices than you could possibly count. They were in harmony with it, but far higher up the scale: cold, tingling, silvery voices. The second wonder was that the blackness overhead, all at once, was blazing with stars. They didn't come out gently one by one, as they do on a summer evening. One moment there had been nothing but darkness; next moment a thousand, thousand points of light leaped out – single stars, constellations, and planets, brighter and bigger than any in our world. There were no clouds. The new stars and the new voices began at exactly the same time. If you had seen and heard it, as Digory did, you would have felt quite certain that it was the stars themselves which were singing, and that it was the First Voice, the deep one, which had made them appear and made them sing."

C.S. Lewis, *"The Magician's Nephew" (1955)*[1]

1 Introduction

> *"The future is something which everyone reaches at the rate of sixty minutes an hour, whatever he does, whoever he is."*
>
> — **C.S. Lewis**
> "The Screwtape Letters" (1942)[1]

1.1 Motivation

More than ever before, we are connected. In the European Union, 63% of Internet users between the age of 16 and 74 used social media in 2016. If you look at the younger generation between the age of 16 and 24, 9 in 10 Internet users participated in social media according to Eurostat[2]. Some effects of the adoption of social media usage has been reflected in recent events with global interest. In the United States of America, the social media based Obama campaigns of 2008 and 2012 dragged millions of people to debate and participate in political life. In 2011, the Arab Spring shed light on the impact of social media for civic acts.

The increasing use of social media and the impact it's causing in society has attracted the interest of scientists from many different fields, for example from the social sciences, psychology, neuroscience, computer science, data science, etc. Some of these studies are interested in making predictions, such as who is going to win the next elections or how the market of shares is going to behave next week. Other studies are interested in evaluating the impact of interventions for the promotion of behaviours in groups, e.g. increasing the amount of physical activity or motivating a healthy food intake habit. All these processes exemplified above are connected by networks, and require more than social network analysis alone to explain the phenomenon behind the cases listed above.

The exchange of ideas, beliefs, sentiments and behaviour in networks can be better understood by social contagion theory. Social contagion can be observed in people's interactions and has been confirmed by recent neuro-cognitive discoveries [26, 27]. These discoveries shed light on the mechanisms of the brain that can help explain why and how people connected in a social network affect each other through their interactions. The social and health sciences have also shown consistently that relationships can shape people's mentality and/or behaviour [15, 38].

[1] THE SCREWTAPE LETTERS by CS Lewis © copyright CS Lewis Pte Ltd 1942.
[2] http://ec.europa.eu/eurostat/en/web/products-eurostat-news/-/DDN-20170713-1

This thesis intends to expand the knowledge about social contagion and come up with realistic and trustworthy explanations that can be simulated or verified with real data concerning the spread of physical activity behaviour, political positioning and panic in disaster situations. This is done by creating new ways of learning about real life through a network-oriented perspective. The knowledge developed here can be applied to find potential applications to assist people and provide a better quality of life. The work presented here requires a multidisciplinary framework where computational models are used to develop realistic analyses of real life situations. This is done by unfolding possible ways to understand, explain and predict individual habits that impact and build group behaviour in different scenarios.

Therefore, this thesis aims to understand, model and predict different sorts of behaviour through cognitive models and social contagion in social networks. The models developed here can be applied broadly from promoting healthy lifestyle to reactions to web media posts. Data analysis is also used for understanding how the physical activity of a group of adults and a group of children evolve over time.

The following section of this chapter introduces the main concepts and basis for the presented work. Section 1.3 presents the research questions that have guided this work. Section 1.4 presents the methodology applied in order to develop the research and obtain the results. Section 1.5 presents the structure of the other chapters of this thesis.

1.2 Background

The concepts presented below are important to contextualize and enrich the understanding of the following chapters of this thesis. On section 1.2.1 data analysis methods and its applicability to this research is discussed. Section 1.2.2 will explain the theory behind social contagion and how it can be used to explain reality within many contexts. Section 1.2.3 will present the modeling techniques used throughout the thesis to represent scenarios involving the social relations and cognitive aspects of humans. Lastly, Section 1.2.4 will explain the relationship between research on social networks and research on health. It is also shown how this research contributes to the task of predicting the effect of interventions on people's health and understanding the dynamics of social contagion regarding physical activities.

1.2.1 Data Analysis

Recently, much attention has been given to quantitative research due to the increasing amount of data and the development of powerful computational tools for data analysis. Besides statistics, the ground for data analysis has become stronger with the advances in computer science. The popularity of social media websites, combined with the increase of mobile device usage and online communication has facilitated the gathering of people's behaviour data. Learning from personal data has become less intrusive as people leave their "fingerprints" when using social media, or tracking devices, like physical activity devices or smartphones freely.

The use of these big databases of personal interactions and information has for many years been of interest to computer science, social sciences and health sciences. Even though the amount of data is abundant, the quality of the data can present serious limitations to the exploration and analysis processes [31]. To find patterns in data it is required that the data is manipulated, processed, cleaned and crunched. The task of explaining data or validating theoretical models presents a number of limitations that can be challenging for any data analyst to overcome, i.e. missing data, low periodicity of the data collection process, etc.

Besides the limitations imposed by the social media websites regarding the amount of data that someone can collect (if they can), other problems can emerge when scientists are creating their own environment for data collection. To collect longitudinal data, for instance, regularity is required on the data collection and very detailed information about the dynamics of the traits and relationships within the population observed. At the same time, it is a very hard task to track subjects in a very depicted way, as the quality of the data collection implies more intrusion and subsequently higher taxes of participants dropping out of the experiment, or feeling violated in their privacy.

Ethical aspects are also very relevant when developing research on data analysis. Charles [12] gives some good examples on how data analysis can be used to shape consumers' habits and its side effects. Besides increasing the profit of companies, cases such as the mega company Target is one of the examples on how not to use data analysis. Target collected data on their consumers regarding their grocery shopping habits through the data provided by membership cards and credit cards for more than a decade. The data analysis department developed algorithms to detect if the costumer was single or married, if they have kids at home, or even if they were moving house, or pregnant. The pregnancy algorithm would trigger vouchers for items related to the baby care for the mother to be. In one situation, the father of a teenager girl started receiving vouchers for pregnancy items after his daughter used his credit card for shopping, bringing up a very embarrassing (if not unethical) situation for the company.

In this thesis we use a couple of data sets related to physical activity and other behaviours of heterogeneous groups of people. All the data sets were collected respecting the privacy of the participants and following research methods that will be further explained in Section 1.4.

This current section presented the data analysis background and relevance for part of the data oriented research done and shown in this thesis. The knowledge of data analysis techniques and its limitations were important to define the methods used in Chapters 3, 4, 5, 6, 7, 8 and 11. The techniques developed in data analysis and its methods are combined with the network-oriented models throughout this thesis to bring up knowledge on how can we understand, model and predict human behaviour changes through social contagion models.

1.2.2 Social contagion

Social contagion is the concept that individuals tend to follow the same ideas, sentiments and/or behaviour as those with whom they communicate, their social network. This theory assumes that people don't need to have the intention or awareness to affect others. The process happens involuntarily [44]. In this sense, the relationships, or ties, are very important when studying social contagion, as it is through the connections that people are influenced in their attitudes.

The assumption that social contagion is real in our daily lives led many scientists to research how it can be understood within a large range of fields, such as the spread of obesity [14], smoking behaviour [13], alcohol consumption [29], happiness [21], depression [42], divorce [34] and new products diffusion [28].

The recent discovery of the mirror neurons and the role they play in our brains also gave a bold and solid basis to scientifically claim that we have a biological mechanism that explains why we cry when watching a drama or reading a novel, or why we reproduce violence when raised in a violent environment [26]. These neurons are important for the imitation learning system. That is, humans can learn from imitating others, or from actions done by other individuals. Therefore, mirror neurons mechanisms provide the ground to model and validate scenarios of people interacting and exchanging emotions, opinions, information and letting themselves be affected by others. Other mechanisms are also important to understand how people react when receiving information of others' actions. The suppression mechanism, for instance, helps us to avoid shameful situations, like firing back when your boss is angry about some mistake made. To combine all the mechanisms of the brain without losing the knowledge from the other mechanisms, many scientists chose to apply network-oriented models, as they are equipped for loops and potential cross-relations between different parts of the human cognitive system. These models can become naturally complex, just as the brain is.

Most of the work in this thesis is based on the understanding and modeling of social contagion. The current knowledge about mirror neurons, social contagion, neurological responses for threatening situations is extensively applied. This research brings all these topics together, trying to understand how people behave when interacting on web media (Chapter 10), or in a disaster situation (Chapter 9). These tools are also used for modeling physical activity behaviour spread, both in adults and kids, attempting to validate the social contagion model proposed to predict potential improvement or degradation of people's lifestyles (Chapters 4, 5 and 7). A cognitive model for political positioning presented in Chapter 10 brings knowledge from neuropolitics and psychology together to understand how political opinion is shaped and affected by the burst of daily information received though social media.

In Chapter 2 a contagion model is proposed based on the previous work of Bosse et al. [9]. This model is the basis for the other simulations throughout the thesis.

1.2.3 Modeling

Modeling has been present since the first registries of humans from the stone age onwards. There are registries of new models created and used to improve every-day life by the time of 2.000 BC in at least three cultures, including the Egyptians and the Indians [45]. The art of making representations of reality, or copying it was further transferred to the mathematical field with the representation of numbers using bones, up to our present time, where computational models are used to simulate all sorts of real scenarios.

Computational models can be understood as models that use computers as the tool to be simulated. The power provided by the new computational machines created in recent years permit complex and massive simulations to be performed in a feasible amount of time. Helbing [24] adds that computational models can complement classical research methods in the social sciences providing a tool to test if theoretical models and frameworks explain observed phenomena and provide reliable explanation of the reality. Computational models and simulations are very useful for testing hypothesis, performing analyses of many sorts of scenarios, and for prediction. Usually these models are accompanied by real-life data for validation purposes. Furthermore, computational models can be useful in guiding data collection, revealing dynamic analogies, demonstrating trade-offs, decision support and many other reasons [20].

This thesis is focused mainly on creating and testing models that can explain human behaviour, perceptions and emotions concerning many aspects, from physical activity to political opinions. Different modeling approaches are also used according to the context where it is applied, from individual cognition characteristics to a group's healthy lifestyle increase. Therefore, here we explain the importance of these modeling approaches and what the key concepts are for each of them.

Agent-Based Model (ABM) is a modeling method in which agents and their interactions with each other or the environment are accessible as a program or in a physical structure, like a robot. In ABM, the agents can be representing people, animals, government agencies, groups of cells or even countries, and have their own characteristics and actions [18, 39]. The agents can also be entities that have no representation in real life, with the purpose to gather information from the environment, for instance.

ABM is a very good method to use especially in social sciences and behaviour theory tests. In these cases, the behaviour of the entities and dynamics of the interactions are formalized by algorithms containing equations and decision rules, making the modeling of the behaviour more flexible [24]. The flexibility provided by this method allows simulations to provide heterogeneity, where each agent can have its own characteristics and behaviour, and the possibility of stochastic scenarios where random events can be simulated. The flexibility of ABMs can be extended to use it with other kinds of models [19, 40, 47]. ABMs have been extensively used in simulations of the spread of diseases [40], policies for food production [46], emotions [8, 10, 48], social influence and opinion formation [33, 50], etc.

Complex systems is a new science that still lacks a solid definition. Latora et al. [32] roughly defines complex systems as “a system made by a large number of single units (individuals, components or agents) interacting in such a way that the behaviour of the system is not a simple combination of the behaviours of the single units”. The main characteristics of complex systems is the lack of a central control. That is, the overall behaviour of the system is not derivable from the individual entities within it. In this case, the interactions between the components of the system are extremely relevant to better understanding what is happening in the whole complex system [51]. The complex systems are used for modeling and prediction in animals’ behaviour, economics, politics, neuroscience, etc. [3, 5, 32, 51]. Over the years there has been an effort on understanding the characteristics of the individuals in the many contexts, as well as how they interact with each other. The net of interactions and how the individuals affect each other is named as the backbone of the complex system. When the backbone is described in terms of nodes and edges we have the complex network of the complex system [32]. The **complex networks** are studied within a new science discipline, the network science. It is basically the study of social interactions, social web media, the neurons in the human brain, and any other systems composed by a large amount of interconnected nodes through very complex arrangements. It is a multidisciplinary field, where the challenges are mainly related to the understanding of the evolution of connections, or in the spread of information throughout the system. As the networks are context-based, the knowledge from this field can be applied to studies in diverse fields, e.g. from virus and biology to global political patterns within countries [32].

The complex network tools are very useful for behavioural observations and analysis. The advances of social network analysis and complex systems brought ideal tools to evaluate these scenarios beyond the pure statistical analysis as in the past.

Understanding the dynamics of complex networks provides knowledge about the many features that can describe social relationships, such as who befriends whom and why. It is also possible to predict potential individual problems, like loneliness, or burnout, based on the amount (or the lack) of interaction with other people at work, at home or with friends. Collective behaviour can also be studied using these tools, like reactions in emergency situations, or the spread of ideas. It is possible to understand what the means are where ideas are more efficiently diffused, like political ideas, or simple information about facts happening in the world.

Ultimately, complex networks are not a completely new discovery, but a new way of looking at the world and interpreting it. It is a change in perspective that allows scientists to build more realistic views of what we call reality, and to provide tools that allow us to read this reality within new boundaries. The contexts studied in this thesis are very suitable for a complex network representation, as it is going to be shown throughout the following chapters. The spread of behaviours, perceptions and opinions requires knowledge about the individuals in a group and the connections between them.

Within the context of the complex networks, Treur [54] presents network-oriented modeling using temporal-causal models as a very suitable alternative to represent the complexity of human and social processes. Science has built more information on cognitive, affective and social neuroscience fields, permitting that more complexity can be added to the modeling of the mechanisms of the brain [54]. Network-oriented modeling can be inspired by neurological and social events. This paradigm intends to bring together processes that are traditionally studied separately, like cognitive and emotion processes, or individual and collective phenomena. It is known that many human processes involve sub-processes that run in parallel and/or in cyclic to the main process. When using network-oriented modeling, it is possible to represent all the sub-processes without isolating or ignoring part of them. It also permits that a time dimension can be incorporated to the model, which can be useful for timing the processes and generating more realistic simulations.

For this sort of modeling, states (nodes) and connections (edges) are the main actors. States contain activation levels that will determine when the state is triggered or not. The combining functions of the states reflect the aggregated impact caused by the other states connected to a specific state. The connections determine causal relations between the states, and influence the volume of information transmitted between states depending on its weight. The use of network-oriented modeling with **temporal-causal networks** incorporates the semantics of the network necessary to account for the dynamics of the states and connections whereas providing a temporal dimension to the simulations [53]. Even though two different models can have the same isomorph graph representation for the nodes and edges, as exemplified by Treur [53], more information is needed to provide the dynamics of the process being modeled. The information includes how the connections between the nodes are interpreted, or what does a node mean. Without the semantics of the mechanisms present in the structure of the network, we could consider it as a simple graph theory problem. The semantics permit that the dynamics of the states can reproduce diffusion or contagion within a network. The same way, the semantics provide dynamics to the network structure, permitting that adaptive or evolving networks can be represented as such. All there dynamics are connected through causal relations, as they represent the chain of interactions between nodes and/or edges. The perspective of changes in the phenomena studied in this thesis depends also on a temporal structure as each state (or node) affect each other over time. To incorporate the dynamics and temporal aspects of such a modeling approach, differential equations are used, and the network generated with the semantics of these network characteristics are called **temporal-causal network** [53].

This thesis presents works that can be categorized as ABMs, complex systems and network-oriented modeling with temporal-causal networks. All the tools provided by these areas are very useful when modeling how social contagion changes the characteristics of a group of individuals connected to each other. Most of our studied scenarios contain people in networks where their physical activities are being shared (i.e. Chapters 3 to 8). Some of the scenarios are related to the spread of information, like messages in disasters (i.e. Chapter 9). Some of the systems simulated describe the cognitive aspects and the brain of humans when interacting with information in web media (i.e. Chapter 10). In each case, the states (nodes) are defined according to the context where the model is built, together with the ties between the states. The areas explored by this research are neuroscience, psychology and health sciences.

1.2.4 Health and Social networks

Much of this thesis is dedicated to explore how a person's healthy lifestyle is affected by their connections in a network. It also tries to predict intervention targets that can enhance the overall physical activity of a group. The issue of the spread of physical activity throughout networks is a relevant topic that addresses the use of social networks to promote health in a population.

Social network studies have existed for decades, but have received more interest and publicity in recent years, due to the growth of the Internet and other communication devices that use network as a core feature [6, 11, 16, 17, 57]. Social network studies have also been important to explain behaviours that were unexplainable using the classical theories of behaviour, like why one person quits smoking, while another person doesn't [13].

Almost every health topic can be described as a social network [55]. Examples include HIV/STDs transmission networks [37, 43], drug addiction [56], smoking [2, 13], suicide [4, 41] and obesity [14, 15]. Many other studies provide models that can account for health issues in many aspects, from the spread of diseases to the diffusion of good preventive information, from depression and loneliness tendencies to suicide ideation.

Including the relations as part of the model is important to fill the gaps that other studies in behaviour lack. As social beings, we each have different traits, character and opinions. But we are also largely influenced by our relationships when it comes to how we behave or what we believe. Knowing who is a friend of whom, and how much time is spent between people can tell a lot about many aspects that could not be addressed by simply knowing the characteristics of someone. It is known and much study has shown that there is a tendency for people to befriend others who are like themselves. This is called *homophily* [16, 36, 52]. This fact takes us to one of the biggest challenges when using social network models to explain people's behaviours: while random sampling was acceptable in the classical behavioural science, to address the effects of the network on individuals requires a more careful and strict approach.

Random sampling can remove the entire net of relationships of the individual, requiring that other methods to collect network relations are applied, like ego centric data collection techniques [16, 25]. Chapter 4 used this technique to address the social network of a group of bachelor students, starting from one individual's contacts.

Besides the difficulties in data collection, social network models require that the connections present a reliable metric, where you can differentiate acquaintances from spouses, and spouses from best friends. The frequency of contact is also relevant, as it is the dynamics of the network. Not only the individuals change, but also the connections.

Other prominent studies on connections of people is related to potential interventions in order to change people's behaviour. This kind of work is becoming increasingly important, as the health situation of the population is a big issue for governments and society in general. We attack this problem in Chapter 8 when trying to predict which kids are the most ideal to apply interventions in a network of students from schools in the Netherlands. For this purpose, it is relevant to know who are the more isolated individuals, who are the bridges, who are the most connected, or the ones who are more centralized and verify which selection of nodes delivers better results.

This thesis dedicates a big part for understanding the spread of health behaviour throughout social networks. The aspects of data collection, analysis and difficulties in this sort of research are addressed in Chapters 4, 5, 7 and 8.

1.3 Research questions

The aim of this research is to understand how behaviour, perceptions and emotions are spread in social networks on many levels, from social interactions to the cognitive system and its reaction towards the external world. We use artificial intelligence methods combined with data analysis and neuroscience. Ultimately, we want to address the following question:

How can we create and validate computational models that explain social influence and contagion in social networks?

This overall question can be subdivided in the sub-questions below.

1. How can we design and use temporal-causal models based on networks to better understand and describe social contagion taking into account personal characteristics?

The first question consists of a process of depicting a phenomenon (the social contagion) in many parts. The first part consists of the design of social science experiments to collect social network data that can provide us with a temporal-causal structure, both for the states of the individuals and for the connections between them. The second part consists of treating the data collected, running the model and adjusting the parameters to find a better fit between the designed model and the empirical data. The third part is based on the reflection on the important aspects that describe the studied phenomenon. It involves adapting the models to better describe reality, or propose new methods that incorporate other disciplines (i.e. social sciences) in the process of building up traits of individuals and the net.

2. How can we predict changes in behaviour using the relationships and data related to physical activity and how can we measure social contagion in a social network regarding people's PAL?

Besides describing the dynamics of interactions in a social network, predicting the future of the structure of the states and connections is also very important. In this work we used the Physical Activity Levels (PALs) as the diffusion unit of behaviour. In this sense, we are interested in verifying and predicting how the PALs are affected based on the social connections in a network. To achieve the results we wanted, a combination between differential equations and agent-based modeling in complex systems were used. The quantification of the amount of social contagion is still a very open research question, and we therefore explore it in order to shed light in this matter.

From the group's perspective, we also want to know if connectivity is a factor that can describe the increase (or deterioration) on PAL of people. And if that's the case, we want to identify influential nodes in the network in order to promote healthy lifestyle and optimize the spread of behaviour.

3. What are potential applications for modeling behaviour in social networks and how can we apply the knowledge of temporal-causal network modeling using different contexts and methodologies?

The third question is related to the extension of this approach to other contexts, such as the description of phenomenons at a cognitive level. Here we want to investigate if the method is useful for other contexts besides the PAL. Thus, we have to understand internal (cognitive) reactions for people in various situations, e.g. in disasters, or when interacting with web media. We also want to investigate the cognition behind political opinions using temporal-causal models based on networks. As part of this process, we also need some tools to help in the data collection, such as classifiers for political content.

1.4 Research Methodology

This section explains the methods used throughout this thesis to achieve the understanding to answer the questions presented in the section above.

1.4.1 Systematic review

As stated by Akerjordet and Severinsson [1], "a systematic review involves the identification, selection, critical analysis and written description of existing information". Multidisciplinary research demands a huge effort on reading and understanding the state of the art of many phenomena involved in any sort of simulation or analysis. Due to the many disciplines involved throughout this work, systematic review was used constantly in order to understand concepts and developments from the fields of neuroscience and psychology.

The models created for political positioning (Chapter 10) for sharing behaviour on web media (Chapter 11) and for disaster situations (Chapter 9) required extensive research on many different disciplines. We used a large body of literature to build models that describe the cognition of people interacting with web environments or a person's reaction in a specific context of disaster information spread. The literature shed light on the knowledge about the main mechanisms behind these scenarios and how much is known about the brain functioning when it comes to these topics.

The studies on health and physical activity also demanded a large amount of work to understand the methods to quantify and interpret lifestyle of people. A good amount of systematic review was done especially when trying to understand the dynamics of changes in the behaviour of children in Chapter 8, and also when developing measurements for physical activity level among young adults in Chapter 4.

Other studies related to social sciences were also relevant to assess personal traits of participants in experiments created in Chapter 3. For this reason, we invested significant time in building knowledge on which methods are used and how efficient they are when handling social experiments.

1.4.2 Data collection, processing and analysis

Many governments, companies and researchers became intrigued with the usefulness of the data provided by users not too long after the emergence of the Internet in the 1990's. The increasing number of websites and the boom of social media increased on a large scale the amount of data available, and this was mainly behavioural and relational data. Even a simple email can contain contextual information and provide data for tracking some kinds of behaviour or relations between sender and receiver. A discussion over a post on Facebook involves many agents that can be connected as 'friends' or not, as well as showing the tone and reactions to every response.

Besides from all the online data provided on the Internet, offline data has also increased with the miniaturization of the sensors and electronic devices [22]. Activity trackers can follow people without any need for connection throughout the day, providing a good amount of information about how active people are.

In some parts of the work in this thesis we used Fitbit One devices as a way to collect data from participants. In other parts of the work shown in this thesis, other devices were used, always aiming to gather data of people in social relations. This data went through the process of data analysis, described below.

To collect relational data for part of our data set we used questionnaires such as the Big Five Inventory, and based our methods on previous works from social sciences. When the connections happen in a virtual environment (i.e. the Internet), we used the data from the logs of ties built over time. The information of the connections is very important to build the social networks, as they are the means through which the spread of behaviour happens.

The data analysis required for this thesis involves the use of statistical methods to evaluate differences between groups of people connected or not connected (i.e. Chapter 5), or linear regression for the matter of comparing with other models and finding trends on the individuals' PAL data (i.e. in Chapters 4, 7 and 8).

The pre-processing and cleaning of data is also very important when analyzing data. In many of the following chapters the reader will learn how the data was selected to draw conclusions about the behaviour of the participants in the experiments. The choices about how to select the data was done with the utmost care to not lose information and at the same time bring a reliable interpretation of the phenomena studied.

Most of the work was done using Python and Jupyter notebooks [30]. Python is a very powerful tool as it permits that data can be read, cleaned, processed, and statistically analyzed without having to change the environment. Python also provides tools to save results, plot graphics, generate agent-based modeling codes, etc. [35]. The Jupyter Notebooks facilitate the visualization of the code and the results. It also permits the graphics to be extracted straight from the interface of the code. Some of the Chapters will provide the link to the notebooks that were used for the data analysis, which also facilitates the verification and questioning of the methods used. We also used R for some of the statistical analysis, such as the multiple linear regression model and the linear mixed model of Chapter 5.

1.4.3 Social network analysis

As discussed in Section 1.2.4, 'social networks' are how we name the representation of actors and interactions in a formal way. In other words, social networks are the way relational data is stored or represented. Differently from the classic sets of data, which are focused on the attributes of people, the study of social networks requires a new set of techniques that incorporate the relational characteristics of the data [16, 49]. This is where Social Network Analysis (SNA) becomes important. According to Crossley et al. [16], SNA is "a set of interconnected concepts, theories and techniques developed for the most part within a relatively cohesive, interdisciplinary research 'network', devoted to gathering and analysis of relational data".

The two essential elements of social networks are a set of nodes and a set (or sets) of ties. Both elements can contain specific attributes, and the presence of multiple sets of ties can be useful to represent what is called multilayer networks. Some concepts are also important for the SNA. The connected components will tell us which subset of nodes are linked by a path, and the centrality can tell us about the level of centrality of each node in the network. Centrality can be measured in terms of degree, paths, eigen vectors and other metrics that can describe structural characteristics of the nodes. For a good understanding of these metrics and the concepts behind SNA, see [7], [25] or [49].

In this work, SNA is present in several chapters as ways to better understand the structure of the networks where social contagion is going to be evaluated. In Chapter 4 we evaluate the structure of a network of young adults from the same undergraduate class collected using a snow-ball method [23]. In Chapter 6 we evaluate the structure of a relational data set of individuals in a health promotion program in order to understand the dynamics of the creation of ties and the evolution of the whole network. In Chapter 8 we apply SNA to understand the structure of the network of children in Dutch schools for a study on health and physical activity propagation through the ties of the kids.

1.4.4 Computational modeling

Section 1.2.3 explained the modeling concepts used in this thesis. Chapter 2 covers the explanations of the core of the contagion model used in many of the chapters in this thesis. The models presented here are based on a network-oriented approach, meaning that the mathematical representation of the events are done using nodes as agents or states, and the ties represent the relationship or the influence between two edges. Some of the studies are in the context of health, and some are related to behaviour and cognition, as explained below.

We applied the model presented by Bosse et al. [9] for the spread of physical activity in networks in several chapters:

- In Chapter 3: to find the traits of the agents in the models using parameter tuning techniques.
- In Chapter 4: we try to verify if the model can predict the increase or degradation of PAL for a group of young adults from the same class.
- In Chapter 7: we try to fit the model to account of a big network of participants in a health promotion program.
- In Chapter 8: to predict behaviour change among children from Dutch schools.

The concepts of network-oriented model are also applied to different contexts in Part IV of this thesis. In Chapter 9 we define a model to describe people's reactions when receiving messages related to disaster situations from a neuro-psychological perspective. Chapter 10 addresses the political positioning on web media through a cognitive model based on psychological and neuroscientific discoveries about the brain.

1.5 Structure of this thesis

This Section presents the organization of this thesis. As a cumulative research, the chapters can be read individually or as a whole body of work.

1.5.1 Part I: Foundations

The first part of the thesis aims to give the background and introduce the reader to the topics of social contagion, social networks and complex networks. It is expected that the reader can grasp the fundamental theories that guide the other chapters and that provide the basis for the whole work.

Chapter 1 (current chapter) presents the literature review and the questions that are addressed in this work. It also describes the methods used to build up the contributions of the thesis.

Chapter 2 presents the adaptations proposed in the absorption model of emotions previously created. The goal of this chapter is to present the mathematical structure of the contagion model and the improvements proposed in order to deliver more stable results when applying it within the different contexts.

1.5.2 Part II: Data Analytics in Networks

Part II is dedicated to the studies regarding physical activities in networks. For this part, we are interested in understanding how physical activity levels (PAL) are spread in networks and how we can quantify the PAL and predict future states based on social contagion models.

Chapter 3 is about a method proposal to derive personality traits based on empirical data and social contagion. The method includes machine learning algorithms to fine tune the parameters and contrast them with methods from social sciences based on intake questionnaires.

Chapter 4 presents a social experiment involving social contagion of PAL in young adults. The work contains the data collection, analysis and study of the applicability of the social contagion model on this data set.

Chapter 5 shows the results of a data analysis of a data set with originally around 5.000 participants in a health promotion program. The group was connected through a social network where they could visualize their friends' PALs and compare themselves with their peers. For this work we evaluate if individuals that chose to participate in the social network present better improvement in their PAL (in comparison to those who opted out) in the following two scenarios: (1) people who are connected in the network from day 1 of the program, and (2) people who are willing to be in the community program. This work gives attention to one of the methodological issues when selecting data.

Chapter 6 is a Social Network Analysis (SNA) of the data set from Chapter 5. In this work we are interested in understanding how the dynamics of the connections happen over time using SNA tools, like centrality measurements and correlations.

1.5.3 Part III: Using Social Contagion Models for Explaining Physical Activity

This part is dedicated to propose ways of tuning and validating the social contagion model in different contexts. By validating we mean that we try to use the social contagion model to explain the data collected in two different sets.

Chapter 7 will use the same data set from Chapters 5 and 6. In this chapter we want to fit the data about the PALs of the individuals and their connections to our social contagion model shown in Chapter 2. Chapter 7 also presents all the decisions regarding the cleaning of the data and the process of generating the simulations together with the results obtained.

Chapter 8 applies the same strategy using an alternative model to explain the changes in PAL in a network of children from schools in the Netherlands. The model used here is also based on social contagion, but with a different mathematical framework from our social contagion model.

1.5.4 Part IV: Use of Contagion Models for Perceptions and Emotions

This part contains other works in contexts other than PAL that can also rely on social contagion phenomena. The main goal of this part is to show that the approach applied for the PAL is expandable to many possible investigations, including neurological and cognitive explorations.

Chapter 9 presents a temporal-causal model for the spread of messages in disasters. It is based on cognitive knowledge about responses from the brain and human interactions when in a context of the spread of information about disasters.

Chapter 10 presents a cognitive model for political positioning of people when interacting on social media websites.

Chapter 11 presents the construction of a classifier of messages for the political model presented in Chapter 10. A machine learning method is proposed to classify social media messages (tweets) between political or non-political.

1.5.5 Part V: Discussion and evaluation

Part V of this thesis aims to discuss the contributions of this work and explain how the research questions presented in Chapter 1 were answered throughout the chapters in this thesis. This discussion and evaluation is present in **Chapter 12**.

Bibliography

[1] Kristin Akerjordet and Elisabeth Severinsson. „Emotional intelligence: a review of the literature with specific focus on empirical and epistemological perspectives“. In: *Journal of clinical nursing* 16.8 (2007), pp. 1405–1416 (cit. on p. 12).

[2] Cheryl Alexander, Marina Piazza, Debra Mekos, and Thomas Valente. „Peers, schools, and adolescent cigarette smoking“. In: *Journal of adolescent health* 29.1 (2001), pp. 22–30 (cit. on p. 10).

[3] Alain Barrat, Marc Barthelemy, and Alessandro Vespignani. *Dynamical processes on complex networks*. Cambridge university press, 2008 (cit. on p. 8).

[4] Peter S Bearman and James Moody. „Suicide and friendships among American adolescents“. In: *American journal of public health* 94.1 (2004), pp. 89–95 (cit. on p. 10).

[5] Nino Boccara. *Modeling complex systems*. Springer Science & Business Media, 2010 (cit. on p. 8).

[6] Jeremy Boissevain. *Friends of friends: Networks, manipulators and coalitions*. St. Martin's Press, 1974 (cit. on p. 10).

[7] Stephen P Borgatti, Martin G Everett, and Jeffrey C Johnson. *Analyzing social networks*. Sage, 2018 (cit. on p. 14).

[8] Tibor Bosse, Rob Duell, Zulfiqar Ali Memon, Jan Treur, and C Natalie Van Der Wal. „Multi-Agent Model For Mutual Absorption Of Emotions.“ In: *ECMS* 2009 (2009), pp. 212–218 (cit. on p. 7).

[9] 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), pp. 48–67 (cit. on pp. 6, 15).

[10] 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. 7).

[11] Ronald S Burt. „Autonomy in a social topology“. In: *American Journal of Sociology* 85.4 (1980), pp. 892–925 (cit. on p. 10).

[12] Duhigg Charles. *The Power of Habit*. Random House Publishing Group, 2012 (cit. on p. 5).

[13] Nicholas A Christakis and James H Fowler. „The collective dynamics of smoking in a large social network“. In: *New England journal of medicine* 358.21 (2008), pp. 2249–2258 (cit. on pp. 6, 10).

[14] Nicholas A Christakis and James H Fowler. „The spread of obesity in a large social network over 32 years“. In: *New England journal of medicine* 357.4 (2007), pp. 370–379 (cit. on pp. 6, 10).

[15] Ethan Cohen-Cole and Jason M Fletcher. „Is obesity contagious? Social networks vs. environmental factors in the obesity epidemic“. In: *Journal of health economics* 27.5 (2008), pp. 1382–1387 (cit. on pp. 3, 10).

[16] Nick Crossley, Elisa Bellotti, Gemma Edwards, et al. *Social network analysis for ego-nets: Social network analysis for actor-centred networks*. Sage, 2015 (cit. on pp. 10, 14).

[17] Alain Degenne and Michel Forsé. *Introducing social networks*. Sage, 1999 (cit. on p. 10).

[18] Joshua M Epstein. *Generative social science: Studies in agent-based computational modeling*. Princeton University Press, 2006 (cit. on p. 7).

[19] Joshua M Epstein. „Modelling to contain pandemics“. In: *Nature* 460.7256 (2009), p. 687 (cit. on p. 7).

[20] Joshua M Epstein. „Why model?“ In: *Journal of Artificial Societies and Social Simulation* 11.4 (2008), p. 12 (cit. on p. 7).

[21] James H Fowler and Nicholas A Christakis. „Dynamic spread of happiness in a large social network: longitudinal analysis over 20 years in the Framingham Heart Study“. In: *Bmj* 337 (2008), a2338 (cit. on p. 6).

[22] Bruno Gonçalves and Nicola Perra. *Social phenomena: From data analysis to models*. Springer, 2015 (cit. on p. 13).

[23] Leo A Goodman. „Snowball sampling“. In: *The annals of mathematical statistics* (1961), pp. 148–170 (cit. on p. 14).

[24] Dirk Helbing. „Agent-based modeling“. In: *Social self-organization*. Springer, 2012, pp. 25–70 (cit. on p. 7).

[25] Marina Hennig, Ulrik Brandes, Jürgen Pfeffer, and Ines Mergel. *Studying social networks: A guide to empirical research*. Campus Verlag, 2012 (cit. on pp. 10, 14).

[26] Marco Iacoboni. *Mirroring people: The new science of how we connect with others*. Farrar, Straus and Giroux, 2009 (cit. on pp. 3, 6).

[27] Marco Iacoboni and John C Mazziotta. „Mirror neuron system: basic findings and clinical applications“. In: *Annals of neurology* 62.3 (2007), pp. 213–218 (cit. on p. 3).

[28] Raghuram Iyengar, Christophe Van den Bulte, and Thomas W Valente. „Opinion leadership and social contagion in new product diffusion“. In: *Marketing Science* 30.2 (2011), pp. 195–212 (cit. on p. 6).

[29] Rosenquist J, Murabito J, Fowler JH, and Christakis NA. „The spread of alcohol consumption behavior in a large social network“. In: *Annals of Internal Medicine* 152.7 (2010), pp. 426–433. eprint: `/data/journals/aim/20203/0000605-201004060-00007.pdf` (cit. on p. 6).

[30] Thomas Kluyver, Benjamin Ragan-Kelley, Fernando Pérez, et al. „Jupyter Notebooks-a publishing format for reproducible computational workflows.“ In: *ELPUB*. 2016, pp. 87–90 (cit. on p. 14).

[31] Gueorgi Kossinets. „Effects of missing data in social networks“. In: *Social networks* 28.3 (2006), pp. 247–268 (cit. on p. 5).

[32] Vito Latora, Vincenzo Nicosia, and Giovanni Russo. *Complex networks: principles, methods and applications*. Cambridge University Press, 2017 (cit. on p. 8).

[33] Jan Lorenz, Heiko Rauhut, Frank Schweitzer, and Dirk Helbing. „How social influence can undermine the wisdom of crowd effect“. In: *Proceedings of the National Academy of Sciences* 108.22 (2011), pp. 9020–9025 (cit. on p. 7).

[34] Rose McDermott, James H Fowler, and Nicholas A Christakis. „Breaking up is hard to do, unless everyone else is doing it too: Social network effects on divorce in a longitudinal sample“. In: *Social Forces* 92.2 (2013), pp. 491–519 (cit. on p. 6).

[35] Wes McKinney. *Python for data analysis: Data wrangling with Pandas, NumPy, and IPython*. " O'Reilly Media, Inc.", 2012 (cit. on p. 14).

[36] Miller McPherson, Lynn Smith-Lovin, and James M Cook. „Birds of a feather: Homophily in social networks“. In: *Annual review of sociology* 27.1 (2001), pp. 415–444 (cit. on p. 10).

[37] Martina Morris. *Network epidemiology: A handbook for survey design and data collection*. Oxford University Press on Demand, 2004 (cit. on p. 10).

[38] 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. 3).

[39] Mario Paolucci. „N. Gilbert and KG Troitzch, Simulation for the Social Scientist“. In: *Journal of Management & Governance* 12.2 (2008), pp. 225–231 (cit. on p. 7).

[40] Jon Parker and Joshua M Epstein. „A distributed platform for global-scale agent-based models of disease transmission“. In: *ACM Transactions on Modeling and Computer Simulation (TOMACS)* 22.1 (2011), p. 2 (cit. on p. 7).

[41] Bernice A Pescosolido and Sharon Georgianna. „Durkheim, suicide, and religion: Toward a network theory of suicide“. In: *American Sociological Review* (1989), pp. 33–48 (cit. on p. 10).

[42] J Niels Rosenquist, James H Fowler, and Nicholas A Christakis. „Social network determinants of depression“. In: *Molecular psychiatry* 16.3 (2011), p. 273 (cit. on p. 6).

[43] Richard B Rothenberg, John J Potterat, Donald E Woodhouse, et al. „Social network dynamics and HIV transmission“. In: *Aids* 12.12 (1998), pp. 1529–1536 (cit. on p. 10).

[44] Clifford W Scherer and Hichang Cho. „A social network contagion theory of risk perception“. In: *Risk analysis* 23.2 (2003), pp. 261–267 (cit. on p. 6).

[45] Hermann Schichl. „Models and the History of Modeling“. In: *Modeling Languages in Mathematical Optimization*. Ed. by Josef Kallrath. Boston, MA: Springer US, 2004, pp. 25–36 (cit. on p. 7).

[46] Pepijn Schreinemachers and Thomas Berger. „An agent-based simulation model of human–environment interactions in agricultural systems“. In: *Environmental Modelling & Software* 26.7 (2011), pp. 845–859 (cit. on p. 7).

[47] Frank Schweitzer. „Brownian agents and active particles. On the emergence of complex behavior in the natural and social sciences“. In: *Springer Series in Synergetics. Berlin, Germany, Springer* (2003) (cit. on p. 7).

[48] Frank Schweitzer and David Garcia. „An agent-based model of collective emotions in online communities“. In: *The European Physical Journal B* 77.4 (2010), pp. 533–545 (cit. on p. 7).

[49] John Scott. *Social network analysis*. Sage, 2017 (cit. on p. 14).

[50] Yilun Shang. „An agent based model for opinion dynamics with random confidence threshold“. In: *Communications in Nonlinear Science and Numerical Simulation* 19.10 (2014), pp. 3766–3777 (cit. on p. 7).

[51] Robert Siegfried. *Modeling and simulation of complex systems: A framework for efficient agent-based modeling and simulation*. Springer, 2014 (cit. on p. 8).

[52] Jeffrey A Smith, Miller McPherson, and Lynn Smith-Lovin. „Social distance in the United States: Sex, race, religion, age, and education homophily among confidants, 1985 to 2004“. In: *American Sociological Review* 79.3 (2014), pp. 432–456 (cit. on p. 10).

[53] J Treur. „The Ins and Outs of Network-Oriented Modeling: from Biological Networks and Mental Networks to Social Networks and Beyond (Keynote Lecture)“. In: (2018) (cit. on p. 9).

[54] Jan Treur. „Network-Oriented Modeling and Its Conceptual Foundations“. In: *Network-Oriented Modeling: Addressing Complexity of Cognitive, Affective and Social Interactions*. Cham: Springer International Publishing, 2016, pp. 3–33 (cit. on pp. 8, 9).

[55] Thomas W Valente. *Social networks and health: Models, methods, and applications*. Vol. 1. Oxford University Press New York, 2010 (cit. on p. 10).

[56] Thomas W Valente, Peggy Gallaher, and Michele Mouttapa. „Using social networks to understand and prevent substance use: a transdisciplinary perspective“. In: *Substance use & misuse* 39.10-12 (2004), pp. 1685–1712 (cit. on p. 10).

[57] Thomas W Valente, Anamara Ritt-Olson, Alan Stacy, et al. „Peer acceleration: effects of a social network tailored substance abuse prevention program among high-risk adolescents“. In: *Addiction* 102.11 (2007), pp. 1804–1815 (cit. on p. 10).

2

Analysis and Refinement of a Temporal-Causal Network Model for Absorption of Emotions[1]

"Without the aid of trained emotions the intellect is powerless against the animal organism."

— C.S. Lewis
"The Abolition of Man"[2]

Abstract

An earlier proposed temporal-causal network model for mutual absorption of emotions aims to model emotion contagion in networks using characteristics such as traits of openness and expressiveness of the members of the network, and the strengths of the connections between them. The speed factors describing how fast emotional states change, were modeled based on these characteristics according to a fixed dependence relation. In this paper, particular implications of this choice are analyzed. Based on this analysis, a refinement of the model is proposed, offering alternative ways of modeling speed factors. This refinement is also analyzed and evaluated.

[1]Published as: **Fernandes de Mello Araújo E.**, Treur J. (2016) *Analysis and Refinement of a Temporal-Causal Network Model for Absorption of Emotions*. In: Nguyen NT., Iliadis L., Manolopoulos Y., Trawiński B. (eds) Computational Collective Intelligence. ICCCI 2016. Lecture Notes in Computer Science, vol 9875. Springer, Cham

2.1 Introduction

The social phenomenon called emotion contagion indicates the process by which emotions of a person are affected by emotions of other persons when they are interacting in a social network. This concept has a foundation in neurological findings on mirror neurons [6], and can be used to understand emotions, for example in situations where decisions can be affected by the emotional state of a person. This can occur in urgent situations, when events with a short duration can create disturbances in decisions, but also in processes with longer durations, like mood and depression, commitment with work, et cetera.

Different computational models have been proposed to model emotion contagion. Among them are temporal-causal network models [8] such as the absorption model, introduced in [4] and the amplification model introduced in [3]. The current paper focuses on the absorption model. In this model emotion contagion is modeled using personal characteristics (or traits) such as openness (how a person is open to be influenced by others) and expressiveness (how a person expresses him or herself in the social network), and the strength of the connection between persons. This paper presents an analysis and refinement of the absorption model, in particular by considering multiple options for the way in which the speed factor is modeled. In the original absorption model, a fixed dependence relation is used for the speed factor, describing how the speed factor relates to the traits and connections in the network. In the proposed refined absorption model, in addition two alternative ways are offered that relate the speed factor in different ways to these network characteristics. The effects and improvements that are obtained from these alternative options are analyzed and evaluated as well. This work also shows a more in depth mathematical analysis to better understand convergence and stability in the model. These analyses show that the presented temporal-causal network model is trustworthy and can be very useful to understand different contexts of emotions in social networks.

The paper has the following structure: Section 2.2 will explain the original absorption model in detail. In Section 2.3 an analysis is made in particular concerning the speed factor in the model. Section 2.4 presents two possible alternative ways to model the speed factor. A scaled approach and an advanced logistic function approach are the options explored in this section. Section 2.5 presents mathematical analysis of the model regarding monotonicity and equilibria. Section 2.6 presents results using the new approach for the model, and Section 2.7 presents the conclusions and future works.

2.2 Emotion absorption: the temporal-causal network model

In this section, the computational model for mutual absorption of emotions is presented [2, 4]. This model has been developed as a temporal-causal network model; see [8]. First, the most important concepts used in the model are explained, both in terms of a conceptual representation and a numerical representation. The section concludes with examples of applications of this computational model of emotion contagion.

The model distinguishes some characteristics of persons and the connections between them, represented by parameters. These characteristics affect emotion contagion in the network. The model describes how internal emotion states q_A of persons A affect each other. However, internal states do not affect each other in a direct manner. First, they have to be expressed, after which they can be observed by another person, and in turn such an observation can affect the internal state of this other person. So, internal emotion states q_A affect each other by *contagion as a three-step process*, for which each step has its own characteristics (indicated by ϵ_B, α_{BA}, δ_A, respectively):

ϵ_B	from internal emotion state q_B of B to expressed emotion by B
α_{BA}	from expressed emotion by B to observed emotion by A
δ_A	from observed emotion by A to internal emotion state q_A of A

The characteristic for the extent to which a person B expresses him or herself within the network is captured by the concept of *expressiveness*, modeled by parameter ϵ_B. Similarly, the characteristic for the extent to which a person A is open to be influenced is represented by the *openness*, modeled by parameter δ_A. The strength of the relation between two people in the network is described by the *channel strength*, modeled by parameter α_{BA}. They are formalized by the numerical representations ϵ_B, α_{BA}, and δ_A as real numbers between 0 and 1.

Based on the steps above, the overall contagion process is modeled in terms of the *connection weight* ω_{BA} from sender B to receiver A. This represents the resulting influence of the internal emotion state of sender B on the internal emotion state of receiver A and depends on the three parameters above as shown in equation 2.1.

$$\omega_{BA} = \epsilon_B \alpha_{BA} \delta_A \tag{2.1}$$

In the model this ω_{BA} is used to determine the strength of the impact from the emotion state of B to the emotion state of A at some time point t, as shown in equation 2.2.

$$\mathbf{impact}_{BA}(t) = \omega_{BA} q_B(t) \tag{2.2}$$

In equation 2.2 $q_B(t)$ is the emotion level of B at time t. The overall contagion strength ω_A to q_A represents the total effect from all nodes that are connected to emotion state q_A of person A; it is modeled as in equation 2.3.

$$\omega_A = \sum_{B \neq A} \omega_{BA} \tag{2.3}$$

The aggregated impact $\mathbf{aggimpact}_A(t)$ at time t of all connected emotion states q_{B_i} on emotion state q_A is modeled by a *scaled sum function* (see [8]) $\mathbf{ssum}_{\omega_A}(\dots)$ with the overall connection strength ω_A as scaling factor, as shown in equation 2.4.

$$\begin{aligned}
\mathbf{aggimpact}_A(t) &= \mathbf{ssum}_{\omega_A}(\mathbf{impact}_{B_1A}(t), \dots, \mathbf{impact}_{B_k,A}(t)) \\
&= (\mathbf{impact}_{B_1,A}(t) + \dots + \mathbf{impact}_{B_kA}(t))/\omega_A \\
&= (\omega_{B_1A}q_{B_1}(t) + \dots + \omega_{B_kA}q_{B_k}(t))/\omega_A \\
&= (\omega_{B_1A}/\omega_A)q_{B_1}(t) + \dots + (\omega_{B_kA}/\omega_A)q_{B_k}(t)
\end{aligned} \tag{2.4}$$

From this it follows that $\mathbf{aggimpact}_A(t)$ is calculated as a weighted average of the emotion levels of the connected states q_B as in 2.5.

$$\mathbf{aggimpact}_A(t) = \sum_{B \neq A} w_{BA} q_B(t) \tag{2.5}$$

where

$$w_{BA} = \frac{\omega_{BA}}{\omega_A} = \frac{\epsilon_B \alpha_{BA} \delta_A}{\sum_{C \neq A} \epsilon_C \alpha_{CA} \delta_A} = \frac{\epsilon_B \alpha_{BA}}{\sum_{C \neq A} \epsilon_B \alpha_{CA}} \tag{2.6}$$

The sum of these weights is 1. The dynamics for the contagion for this temporal-causal network model (see also [8]) is described in equation 2.7.

$$\Delta q_A(t + \Delta t) = q_A(t) + \eta_A[\mathbf{aggimpact}_A(t) - q_A(t)]\Delta t \tag{2.7}$$

We denote with η_A the speed factor of A, which is chosen $\eta_A = \omega_A$ here. Sometimes the aggregated impact $\mathbf{aggimpact}_A(t)$ on A is denoted by the shorter notation $q^*_A(t)$.

This temporal-causal network model of emotion contagion has been investigated further and applied in a number of studies. For example, it was applied to predict the emotion levels of team members, in order to maintain emotional balance within a team [5]. When the teams' emotion level was found to become deficient, the model, which was embedded in an ambient agent, provided support to the team by proposing the team leader to give his employees a pep talk [5]. The pep talk is an example of an intervention strategy. Another study experimented with simulations of changes in the social network structure in order to guide the contagion process in a certain direction [7]. Yet another study used the model to predict changes on Physical Activity levels of a group of friends/acquaintances from the same course, applying the model to behaviour contagion [1].

2.3 Analysis of the absorption model

The absorption model is based on two main assumptions, one of which addresses the *level* of the emotions and the other one the *speed* of change of the emotion levels:

1. The emotion level $q_A(t)$ of a person A is affected linearly by the weighted average $\sum_{B \neq A} w_{BA} q_B(t)$ of the emotion levels of the connected persons B.

2. For each person A the speed of change η_A of his or her emotion level linearly depends on the overall connection strength ω_A within the network: $\eta_A = \omega_A$.

Roughly spoken, assumption 1 entails that the members of the network adapt to an average emotion level in the network. As a consequence, the emotion levels will converge to a common emotion level, which is between the minimal and maximal initial emotion levels of the connected members (see Section 2.3.1 for example simulations showing this, and Section 2.5 for a mathematical proof). This is in contrast to, for example, the amplification model introduced in [3] where assumption 1 is not taken as a point as departure, and as a result emotion contagion spirals can be modeled that reach levels higher (or lower) than any of the initial levels.

The assumption 2 states that the more connected members in the network, the higher the speed of change will be, in a proportional manner. In the current paper, assumption 1 is kept, but assumption 2 is critically analyzed in more depth and loosened in order to create room for alternatives. This second assumption (2) is an answer for the more open question:

How does the speed of change of the emotion level depend on the network structure and size?

Specific variants of this question are the following. If a person has more connections to members with a given average emotion level, will he or she adapt faster to this average emotion level? Has the number of relations in real life effect on your speed of change for adapting to them? If a person has more friends, will his or her emotions be affected faster than the emotions of another person with fewer friends? If so, to which extent? Is this relation linear or proportional, or is it inherently nonlinear? Is this increase of the speed going on indefinitely, or is there some bound for it?

In [3] these questions were answered in a most simple, linear, proportional manner, as expressed by assumption 2 above. However, it is doubtful whether this most simple linear option is the most plausible option for realistic networks.

The initial studies of the absorption model itself in [3] already highlighted two constraints: (a) In dynamic property P3 in Section 5 (referring to Theorem 5 in Section 4) in [3] it is stated that for some initial emotion values the emotion values eventually can run out of their boundaries 0 or 1. Also, (b) ω_A is a cumulative value based on the number of connections and their weights; when this number increases, because of the assumption 2, $\eta_A = \omega_A$, also the speed factor η_A increases in a proportional manner, without any limitation. For (b) Bosse et al. [3] used adaptations to the choice of the step size Δt to control that the model stays within the boundaries. That can work well for a few nodes, but this entails that all the time a smaller value for Δt has to be chosen, when the number of nodes becomes larger. This is possible, but not very practical.

The hypothesis is that problem (a) relates to the strongly increasing value for the speed factor η_A for larger networks entailed by the choice of taking it equal to ω_A. Some experiments were run for analysis keeping the same characteristics of the experiment done by Bosse et al. [3] but with more nodes. The idea is to analyze if the choice for ω_A as a speed factor η_A indeed is the bottleneck for issues (a) and (b) of the absorption model.

2.3.1 Analysis of the original model

The same simulations found at [3] were run again, with more nodes added to the scenario in order to better understand how the model works, and what alternatives are possible. Seven scenarios were created according to the Appendix A of the paper (`http://www.few.vu.nl/~tbosse/emotion/ECMS09.pdf`). The scenarios are:

1. All members have $\omega = 1$ – fully open channels (1a)
2. All members have $\omega > 0$ – big openness for all (1b)
3. All members have $\omega > 0$ – small openness for all (1c)
4. All members have $\omega = 0$ – no changes on emotional levels (2)
5. One member has $\omega = 0$ ($\delta = 0$) (3)
6. Only one member has $\omega \neq 0$ (all other members have $\delta = 0$) (4)
7. One member has $\omega = 0$ ($\delta = 0$ and $\epsilon = 0$) (5)

Below a brief analysis of the effects on scenario 1a is made, showing what happens when the number of nodes is increased. The results of the other scenarios can be found at Appendix A [3]. The first scenario (1a) considers the maximal contagion that can happen. For that, all the parameters (expressiveness, openness and channel strengths) are set to 1. Moreover, Δt is set to 0.1. Figure 2.1 shows the differences between graphics for different numbers of members, 3, 9, and 18 nodes. As the initial values for the emotion levels for the tests have been generated at random, different convergence points, according to the average of the initial values are shown in the graphics. The convergence value for all the nodes that emerges is an average of the initial emotion levels, as shown in [3].

As all parameters are equal to 1, the speed factor η_A for each member A is the in-degree of the nodes minus 1: $\eta_A = n - 1$, with n = number of nodes. As the speed factorη_A determines the next emotion level for all the nodes (equation 2.7),

[3]The file can also be found at `http://www.few.vu.nl/~efo600/iccci16/ICCCI16_A.pdf`

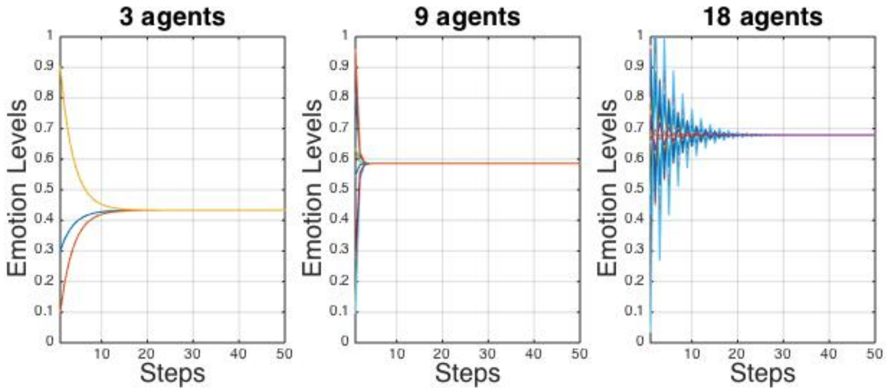


Fig. 2.1: Full channel connections for 3 to 18 members.

the emotion levels converge faster, and at some network size (after 12 members) oscillation in the emotional levels occurs due to the sudden changes caused by the high speed factor. Note that to see this effect Δt was not decreased, what normally would be a measure taken; it was kept at 0.1. Such a decrease would be possible; however, decreasing Δt with the size of the network indefinitely is neither practical nor desirable.

2.3.2 Mathematical analysis of the problem

This section addresses mathematical analysis of the problem concerning the oscillation of the emotion levels. The sudden changes will be explained showing the effect of the increase in the value of the choice of the speed factor $\eta_A = \omega_A$ for larger networks.

The equation for ω_A is the sum of all the connection strengths generated by the nodes in-connected to q_A. The strength by which the emotion from each B is received by A is calculated by $\omega_{BA} = \epsilon_B \alpha_{BA} \delta_A$, as seen before. If all nodes have ϵ and δ higher than zero, and if the network is fully connected (in other words, there is no $\delta_{BA} = 0$ to any pair B, A), and the number of neighbors in the network increases, the value of A also increases proportionally and by the assumption (2) $\eta_A = \omega_A$ the same holds for η_A.

While the speed factor increases, the changes from $q_A(t)$ at the next emotion level $q_A(\Delta t + t)$ become more sudden, and less realistic. As a matter of illustration, imagine a fully connected network (all the channel strengths α_{BA} equal to 1), where every member has openness and expressiveness equals 0.5. So, in that case

$$\omega_{BA} = \epsilon_B \alpha_{BA} \delta_A = 0.5 \times 1 \times 0.5 = 0.25 \tag{2.8}$$

Therefore, $\omega_{BA} = 0.25$ is the connection strength between the emotion states q_A for each of the members that are connected. So, if n is the number of nodes, ω_A for any of the fully connected network will be $(n-1) \times 0.25$. If the number of nodes is increased around 10 times, the speed factor $\eta_A = \omega_A$ will be increased 10 times as well. For an increase to 1000 nodes, the speed factor will be 100 times bigger.

Figure 2.2 (left graph) shows that this increase follows a linear tendency. As it may be doubted that this indefinite increase of η_A is realistic, in Section 4 alternative options for the speed factor η_A will be considered, with patterns corresponding to the other graphs in Figure 2.2, where there is some bound in the increase of the speed factor.

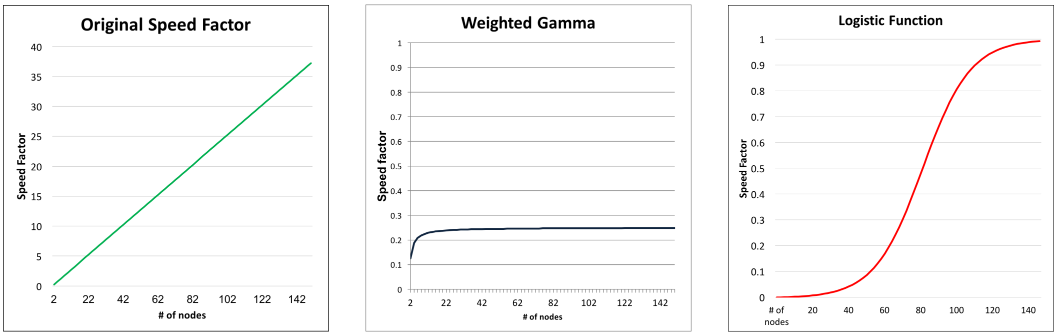


Fig. 2.2: Three types of relations for the speed factor depending on number of nodes:
left: η_A increasing linearly as the number of members increases
middle: η_A increases up to a limit, due to a weighted speed factor
right: η_A increases according to an advanced logistic function

So assuming the speed factor $\eta_A = \omega_A$ (assumption 2 of the original absorption model) causes an unbalanced behaviour of the model. In order to avoid abrupt changes, usually the value of Δt is made smaller and smaller for larger networks. This approach is conceptually and practically inadequate as the speed factor refers to the velocity of the changes in the emotions, whereas Δt refers to the time steps of the model, and has no relation with the speed factor η_A. That incompatibility is a reason to consider alternative answers on the main question concerning speed factors, as discussed next.

2.4 Alternative ways to model the speed factor η_A

In this section two alternative ways of modeling the speed factor η_A are explored. In both cases the speed factor increases with the size of the network, but stays under a certain bound, according to patterns as shown in Figure 2.2 middle and right graphs. The first option is by modeling the speed factor as a *scaled* ω_A, with scaling factor the number n of nodes in the network:

$$\eta_A = \frac{\omega_A}{n} \tag{2.9}$$

This option avoids the effect caused by the increasing in the number of members. In this case, and using the same network as in Figure 2.2, it can be seen that the value for η_A converges to the ω_{BA} which is the same for all nodes in this scenario. Figure 2.2 middle graph shows the new situation when the above scaled model is used for η_A.

Using this option will assure that the speed factor η_A has boundaries defined according to the following mathematical analysis. For the new calculation for η_A, the following holds for all A:

$$0 \leq \eta_A \leq 1 \tag{2.10}$$

This can be verified as follows. It holds

$$\sum_{B \neq A} \epsilon_B \alpha_{BA} \delta_A \leq n - 1 \tag{2.11}$$

as each of the terms of this sum is less or equal 1. Therefore

$$\eta_A = \frac{\omega_A}{n} = \frac{\sum_{B \neq A} \epsilon_B \alpha_{BA} \delta_A}{n} \leq \frac{n-1}{n} < 1 \tag{2.12}$$

In other words the speed factor η_A is now bounded by 1. Note that by a slight modification this bound can be set to any number η by multiplying this by an extra parameter η (the same holds for the second alternative discussed below):

$$\eta_A = \eta \frac{\omega_A}{n} \tag{2.13}$$

A second alternative is to use an advanced logistic function in order to gradually increase the speed with network size but keep the values for the speed factors within some bound. The advanced logistic function has a S shape, or sigmoid curve, and is described by the equation 2.14.

$$\textbf{alogistic}_{\sigma,\tau} = \left[\frac{1}{1 + e^{-\sigma(\eta_A - \tau)}} - \frac{1}{1 + e^{\sigma\tau}} \right] (1 + \exp^{-\sigma\tau}) \tag{2.14}$$

Here σ is the steepness and τ is the threshold value. The values for σ and τ can be chosen according to a person's traits. While some persons respond gradually to the increasing influence of people to whom they are connected, other persons may respond by flare-ups. For the situation of a person that responds mode linearly to the increasing on their cumulative ω_A, a low steepness value such as $\sigma = 0.3$ can be chosen, and, for example, $\tau = 20$. The results for this situation can be seen at Figure 2.2, right graph.

As can be seen, the logistic function also keeps the speed factor values between 0 and 1, and if the parameters of the function are well adjusted, the equation can give more realistic outcomes.

2.5 Mathematical Analysis

This section presents some of the results of a mathematical analysis for the model after the changes at the speed factor calculation.

Definition 1. *A network is called strongly connected if for every two nodes A and B there is a directed path from A to B and vice versa.*

Lemma 1. *Let a temporal-causal network model be given based on scaled sum functions for states q_A:*

$$\mathbf{d}q_A/\mathbf{d}t = \eta_A \left[\frac{\sum_{B \neq A} \omega_{BA} q_B}{\omega_A - q_A}\right]$$

Then the following hold.

(a) If for some state q_A at time t for all states q_B connected toward q_A it holds $q_B(t) \geq q_A(t)$, then $q_A(t)$ is increasing at t: $\mathbf{d}q_A(t)/\mathbf{d}t \geq 0$; if for all states B connected toward A it holds $q_B(t) \leq q_A(t)$, then $X_A(t)$ is decreasing at t: $\mathbf{d}q_A(t)/\mathbf{d}t \leq 0$.

(b) If for all states q_B connected toward q_A it holds $q_B(t) \geq q_A(t)$, and at least one state X_B connected toward q_A exists with $q_C(t) > q_A(t)$ then $q_A(t)$ is strictly increasing at t: $\mathbf{d}q_A(t)/\mathbf{d}t > 0$. If for all states X_B connected toward q_A it holds $q_B(t) \leq q_A(t)$, and at least one state q_B connected toward q_A exists with $q_C(t) < q_A(t)$ then $q_A(t)$ is strictly decreasing at t: $\mathbf{d}q_A(t)/\mathbf{d}t < 0$.

Proof of Lemma 1. *From the differential equation for $q_A(t)$*

$$\begin{aligned}
\mathbf{d}q_A/\mathbf{q}_t &= \eta_A \left[\frac{\sum_{B \neq A} \omega_{BA} q_B}{\omega_A} - q_A\right] \qquad (2.15)\\
&= \eta_A \left[\frac{\sum_{B \neq A} \omega_{BA} q_B - \omega_A q_A}{\omega_A}\right]\\
&= \eta_A \left[\frac{\sum_{B \neq A} \omega_{BA} q_B - \sum_{B \neq A} \omega_{BA} q_A}{\omega_A}\right]\\
&= \eta_A \left[\frac{\sum_{B \neq A} \omega_{BA} (q_B - q_A)}{\omega_A}\right]
\end{aligned}$$

it follows that $\mathbf{d}q_A(t)/\mathbf{d}_t \geq 0$, so $q_A(t)$ is increasing at t. Similar for decreasing.

For (b) it follows that $\mathbf{d}q_A(t)/\mathbf{d}_t > 0$, so $q_A(t)$ is strictly increasing. Similar for decreasing. ■

Theorem 1 (convergence to one value)**.** *Let a strongly connected temporal-causal network model be given based on scaled sum functions for the states q_A*

$$\mathbf{d}q_A/\mathbf{d}_t = \eta_A \left[\frac{\sum_{B \neq A} \omega_{BA} q_B}{\omega_A - q_A}\right] \qquad (2.16)$$

and with equilibrium values $\underline{q}_A$. Then for all A and B the equilibrium values $\underline{q}_A$ and $\underline{X}_B$ are equal: $\underline{q}_A = \underline{q}_B$. Moreover, this equilibrium state is attracting.

Proof of Theorem 1. *Take a state q_A with highest value $\underline{q}_A$. Then for all states q_C it holds $\underline{q}_C \leq \underline{q}_A$. Suppose for some state q_B connected toward q_A it holds $\underline{q}_B < \underline{q}_A$. Take a time point t and assume $q_C(t) = \underline{q}_C$ for all states q_C.*

Now apply Lemma 1(b) to state q_A. It follows that $\mathbf{d}q_A(t)/\mathbf{d}_t < 0$, so $q_A(t)$ is not in equilibrium for this value $\underline{q}_A$. This contradicts that this $\underline{q}_A$ is an equilibrium value for q_A. Therefore the assumption that for some state q_B connected toward q_A it holds $\underline{q}_B < \underline{X}_A$ cannot be true.

This shows that $\underline{q}_B = \underline{q}_A$ for all states connected towards q_A. Now this argument can be repeated for all states connected toward q_A. By iteration every other state in the network is reached, due to the strong connectivity assumption; it follows that all other states in the temporal causal network model have the same equilibrium value as q_A. From Lemma 1(b) it follows that such an equilibrium state is attracting: if for any state the value is deviating it will move to the equilibrium value. ■

Proposition 1 (Monotonicity Conditions)**.** *It follows that:*

(a) If $q^*_A(t) \leq q_A(t)$ then $q_A(t)$ is monotonically decreasing; it is strictly decreasing when $q^*_A(t) < q_A(t)$.
(b) If $q^*_B(t) \geq q_B(t)$ then $q_B(t)$ is monotonically increasing; it is strictly increasing when $q^*_B(t) > q_B(t)$.

Proof of Proposition 1. *This follows from the differential equation*

$$\mathbf{d}q_C(t)/\mathbf{d}t = \eta_C(q^*_C(t) - q_c(t)) \tag{2.17}$$

and the fact that $0 \leq q_C(t) \leq 1$ and $0 \leq q^*_C(t) \leq 1$. ■

Lemma 2. *Suppose all ω_{CD} are nonzero. Then for an equilibrium the following hold:*

(a) $q^*_A = 0$ if and only if $q_C = 0$ for all $C \neq A$
(b) $q^*_B = 1$ if and only if $q_C = 1$ for all $C \neq B$

Proof of Lemma 2. *It follows that*

(a) From

$$q^*_A = \sum_{C \in G \backslash \{A\}} w_{CA} q_C = 0 \tag{2.18}$$

and the fact that all terms are nonnegative it follows that $w_{CA}q_C = 0$ for all $C \neq A$ and conversely.

(b) From

$$q^*_B = \sum_{C \in G \backslash \{B\}} w_{CB} q_C = 1 \tag{2.19}$$

and the fact that

$$\sum_{C \in G \backslash \{B\}} w_{CB} = 1 \tag{2.20}$$

it follows that $q_C = 1$ for all $C \neq B$ and conversely. ■

Lemma 3. *For an equilibrium for any member the following hold:*

(a) If $q_A = 0$ then $q^*_A = 0$
(b) If $q_B = 1$ then $q^*_B = 1$

Proof of Lemma 3. *It follows that*

(a) From

$$q^*_A - q_A = 0 \tag{2.21}$$

with $q_A = 0$ it follows

$$q^*_A = 0 \tag{2.22}$$

(b) From

$$q^*_B - q_B = 0 \tag{2.23}$$

with $q_B = 1$ it follows

$$q^*_B = 1 \tag{2.24}$$

■

Proposition 2. *Suppose some A is given and all w_{BA} are nonzero. Then for an equilibrium the following hold:*

(a) If $q_A = 0$ then $q_C = 0$ for all C.
(b) If $q_B = 1$ then $q_C = 1$ for all C.

Proof of Proposition 2. *This immediately follows from Lemma 2 and Lemma 3.* ■

Proposition 3. *Suppose all w_{DC} are nonzero. Then for an equilibrium it holds*

(a) If $q_A = 0$ for some A then $q_C = 0$ for all $C \in G$.
(b) If $q_B = 1$ for some B then $q_C = 0$ for all $C \in G$.

Proof of Proposition 3. *This immediately follows from Proposition 2.* ■

2.6 Results

Using the alternative models for the speed factors η_A and comparing them with the outcomes for the original model, it turns out that the oscillation is not present anymore in any of the new approaches. For scenario 1(a), Figure 2.3, it is possible to observe that for 3 members the logistic function delays the convergence point, especially because the logistic function will give a lower value when ω_A is lower.

For scenario 3, Figure 2.4, it is clear how the use of both the scaled and or logistic model for the speed factor corrects the awkward slopes from the original model without needing any change on the time step Δt used. As noticed at scenario 1(a), for fewer nodes, the logistic function still keeps the convergence point later. This can be handled at the logistic function itself through steepness and threshold adjustments.

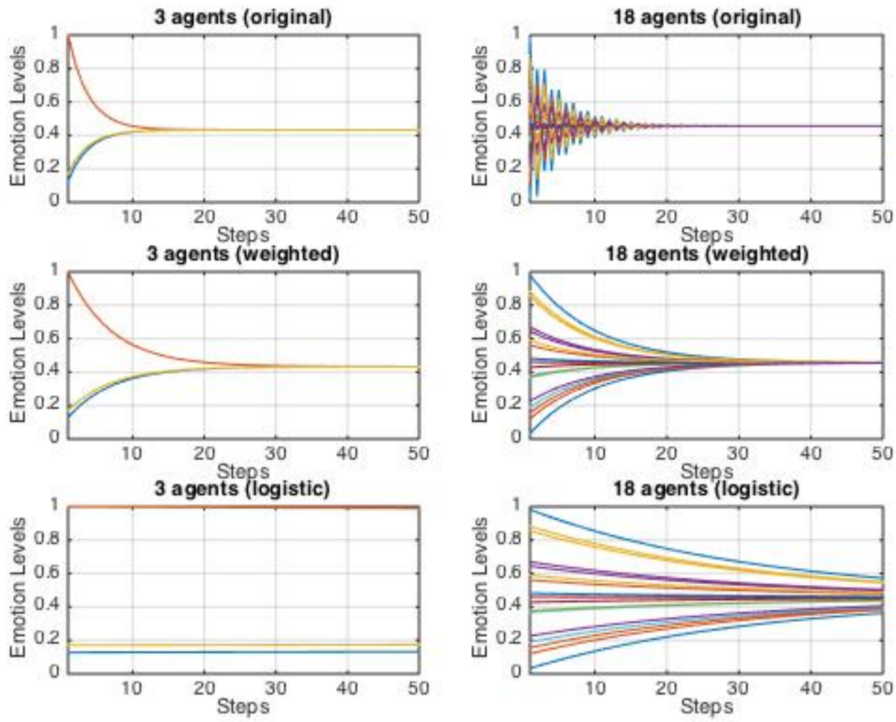


Fig. 2.3: Comparison for scenario 1(a) between the 3 speed factors

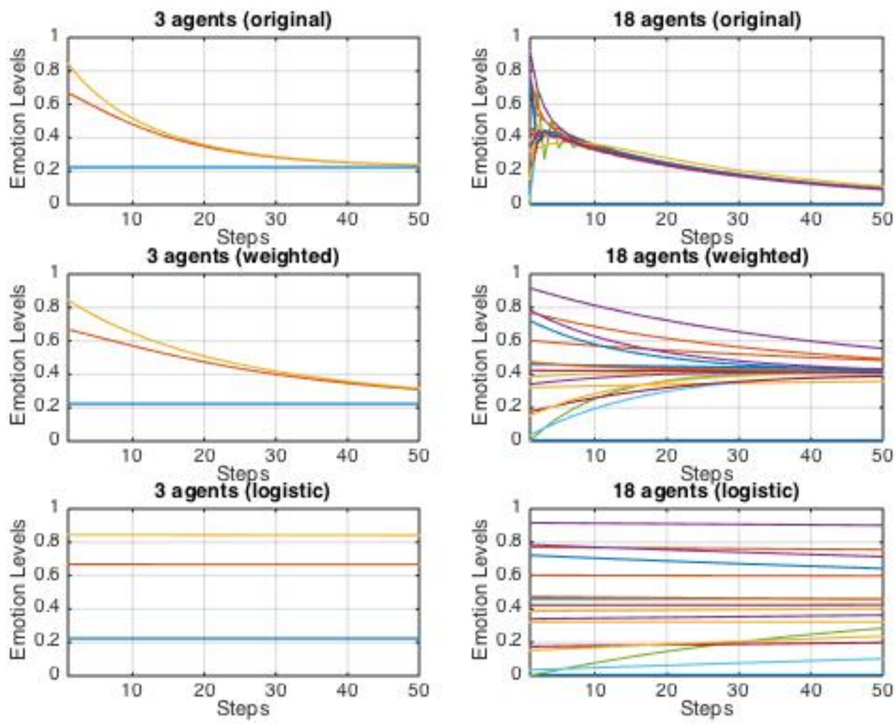


Fig. 2.4: Scenario 3 and the different speed factors used to calculate the emotion levels

More results and analysis for the other scenarios and graphics can be found at Appendix B[4]. As it can be observed, the issue with the oscillation is fixed for all the scenarios explored using weighted or logistic function as a speed factor for the model.

2.7 Conclusion

Mathematical models are used in order to mimic the real world. Regarding the temporal-causal network model for absorption of emotions in a network introduced by Duell et al. [5] it has become clear that the assumption made about the speed factor isn't perfect, and gives room to alternatives. Two of such alternatives were explored here: a scaled model and an advanced logistic model. The expressiveness, openness to changes, and the strength of links still play a role in modeling the speed of the change of the emotion level. By these alternative models the speed can be well regulated between boundaries and do not lead to sudden changes that conflict with our understanding of emotional evolution over time. Limiting the value of speed factor η_A between 0 and 1 creates a stable slope in the emotion changes in networks, what brings the model closer to what it is expected to do.

A mathematical analysis also shows some of the features of the model. Part of the analysis explains characteristics of the model such as convergence and stability.

Future work can be done to investigate how these alternative models for the speed factor affect the results of previous research, and how they can be combined with the model for emotion contagion spirals from [4].

[4] This appendix can also be accessed at `http://www.few.vu.nl/~efo600/iccci16/ICCCI16_B.pdf`

Bibliography

[1] 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 p. 26).

[2] Tibor Bosse, Rob Duell, Zulfiqar A Memon, Jan Treur, and C Natalie van der Wal. „Agent-based modeling of emotion contagion in groups“. In: *Cognitive Computation* 7.1 (2015), pp. 111–136 (cit. on p. 24).

[3] Tibor Bosse, Rob Duell, Zulfiqar Ali Memon, Jan Treur, and C Natalie Van Der Wal. „Multi-Agent Model For Mutual Absorption Of Emotions.“ In: *ECMS* 2009 (2009), pp. 212–218 (cit. on pp. 24, 27, 28).

[4] 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), pp. 48–67 (cit. on pp. 24, 36).

[5] Rob Duell, Zulfiqar A Memon, Jan Treur, and C Natalie Van Der Wal. „An ambient agent model for group emotion support“. In: *Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on*. IEEE. 2009, pp. 1–8 (cit. on pp. 26, 36).

[6] Marco Iacoboni. *Mirroring people: The new science of how we connect with others*. Farrar, Straus and Giroux, 2009 (cit. on p. 24).

[7] Michel Klein, Adnan Manzoor, Julia Mollee, and Jan Treur. „Effect of changes in the structure of a social network on emotion contagion“. In: *Proceedings of the 2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)-Volume 03*. IEEE Computer Society. 2014, pp. 270–277 (cit. on p. 26).

[8] Jan Treur. „Dynamic modeling based on a temporal–causal network modeling approach“. In: *Biologically Inspired Cognitive Architectures* 16 (2016), pp. 131–168 (cit. on pp. 24, 26).

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]

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 $\omega_{BA} = \frac{\epsilon_B w_{BA}}{\sum_{C \neq A} \epsilon_C w_{CA}}$.

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

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

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.

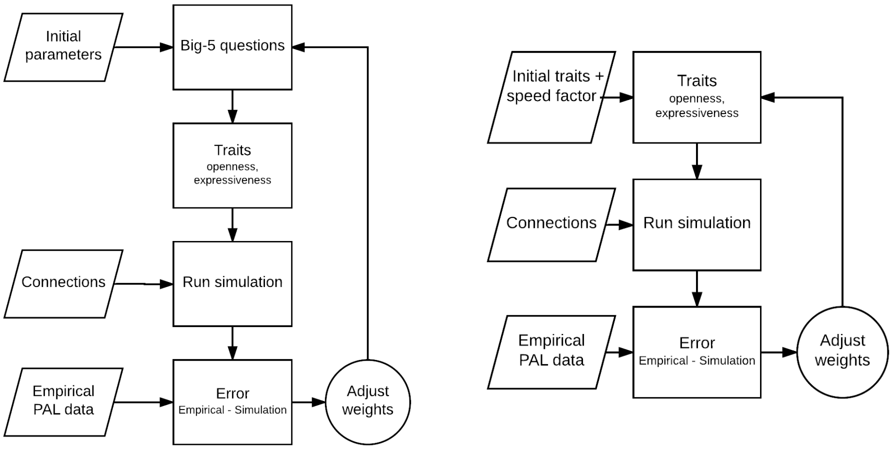


Fig. 3.1: Methodology for tuning the questionnaire questions weights (left). Methodology for tuning the traits for each participant (right).

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.

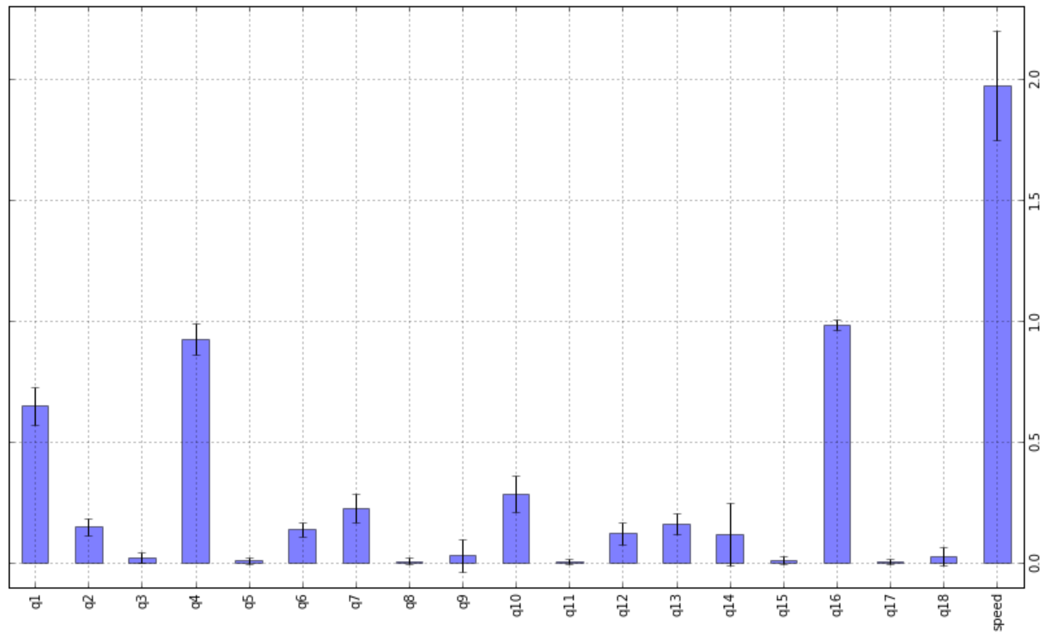


Fig. 3.2: Mean and standard deviation for the 100 best solutions using the grid search.

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

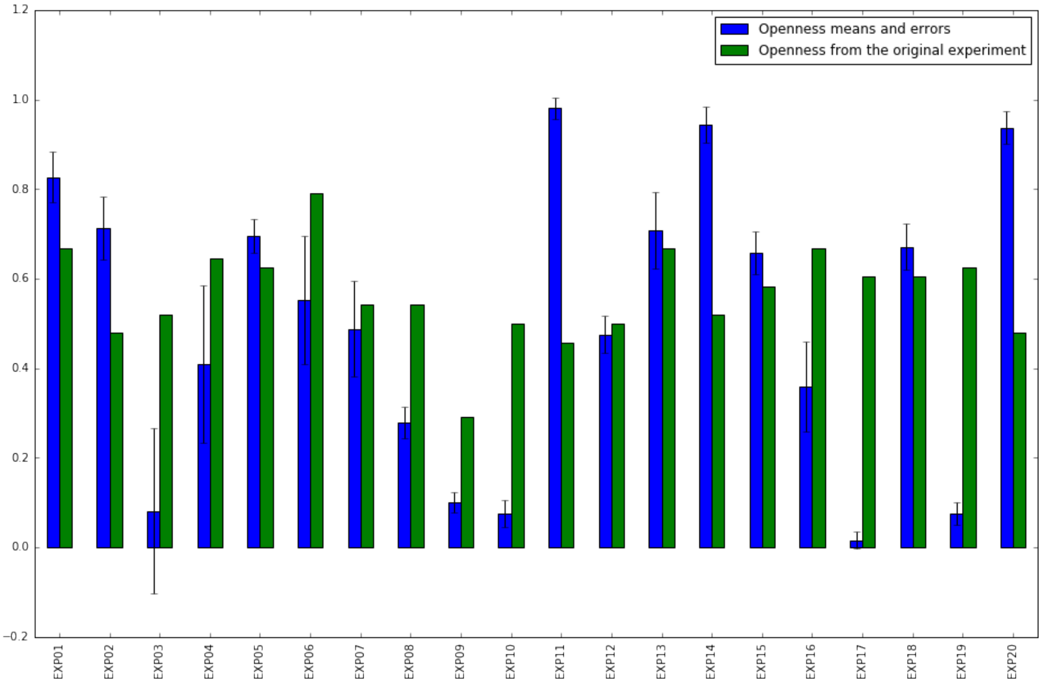


Fig. 3.3: Mean and error for the 100 best solutions using the grid search for openness (blue bars) and the solutions obtained with the questionnaires alone (green bars).

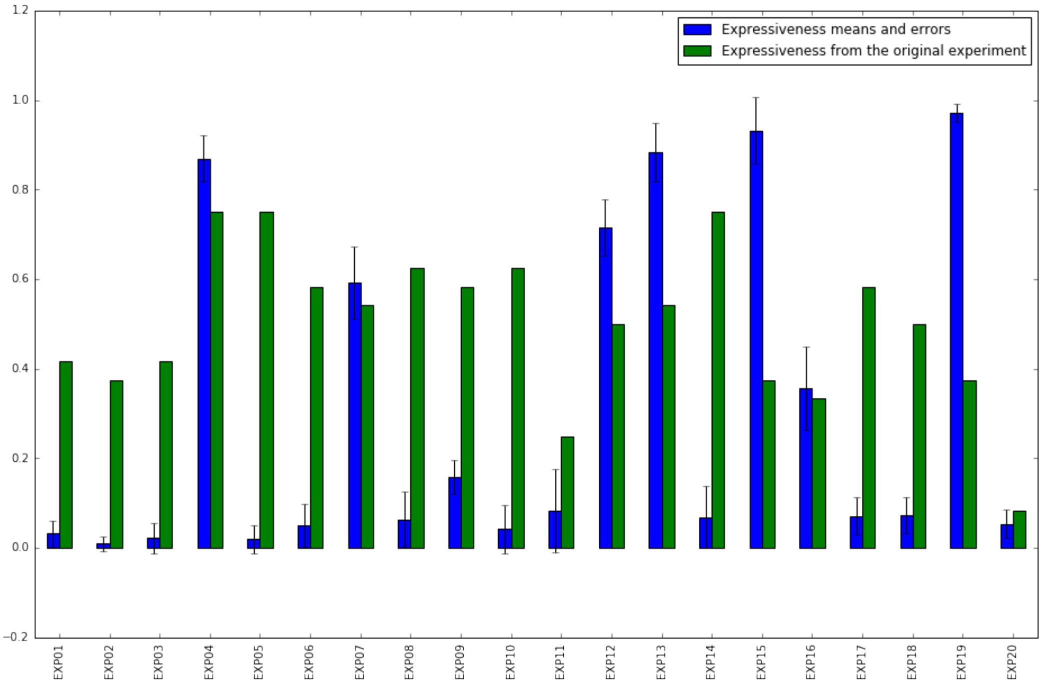


Fig. 3.4: Mean and error for the 100 best solutions using the grid search for expressiveness (blue bars) and the solutions obtained with the questionnaires alone (green bars).

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).

4

Analysis and Evaluation of Social Contagion of Physical Activity in a Group of Young Adults[1]

> *"Friendship arises out of mere Companionship when two or more of the companions discover that they have in common some insight or interest or even taste which the others do not share and which, till that moment, each believed to be his own unique treasure (or burden). The typical expression of opening Friendship would be something like, 'What? You too? I thought I was the only one'."*
>
> — **C.S. Lewis**
> "The Four Loves"[2]

Abstract

It is known that opinions, attitudes and emotions spread through social networks. Several of these cognitions influence behavioural choices. Therefore, it is assumed that the level of physical activity of a person is influenced by the activity levels of the people in its social network. We have performed an experiment with 20 participants between 19 and 28 years old, measuring their physical activity levels for 30 days, in order to observe if there is a contagion effect due to the relationships in the social network. Using our social contagion model, we investigated if people will become more or less active according to the contacts with their peers within the network. Our model correctly predicts the direction of the change (increasing or decreasing) in 80% up to 87% of the cases investigated.

[1]Published as: **Eric F. M. Araújo**, Anita V. T. T. Tran, Julia S. Mollee, and Michel C. A. Klein. 2015. Analysis and Evaluation of Social Contagion of Physical Activity in a Group of Young Adults. In Proceedings of the ASE BigData & SocialInformatics 2015 (ASE BD&SI '15). ACM, New York, NY, USA, Article 31, 6 pages. DOI: https://doi.org/10.1145/2818869.2818922

4.1 Introduction

Physical inactivity is a major public health challenge in the developed world and is recognized as a global epidemic [1]. Insufficient physical activity is a risk factor for cardiovascular diseases and other conditions. The amount of physical activity of a person is usually represented by the Physical Activity Level (PAL). It refers to “any bodily movement produced by skeletal muscles that results in energy expenditure” [5]. The global recommendation for daily exercise is an accumulated 30 minutes of moderate intensity activity, such as cycling, brisk walking or swimming, in segments of at least 10 minutes per activity [15]. Research has shown that a large part of the Western population does not meet these guidelines [9]. Sports medicine and public health constituencies also acknowledged a concern about the deleterious health consequences of insufficient physical activity [7].

It is known that social influences play a key role in lifestyles and are fundamental to whomever wants to maintain healthy behaviour Several aspects underlying a lifestyle, such as emotions, opinions and behaviours, can spread through a social network, in a process called “social contagion”. Social contagion theories explain how one’s social network influences these aspects and how this social environment can provide support in changing them [14].

This research builds on the belief that the social environment can be used as an unobtrusive, even unconscious and therefore suitable way of supporting people to become more physically active [2, 11]. To develop practical applications of lifestyle interventions based on social influence, it is important to have a thorough comprehension of the dynamics underlying the social contagion process. To contribute to this understanding, we have performed an experiment in which we compared the predictions of a model that describes social contagion in a community [3, 4, 6] with the actual change in physical activity level. Our assumption is that the model can be applied to describe the spread of behaviour, considering the willingness to be more active is led by the emotions, attitudes and motives of each person. In the experiment, we constructed a graph of the social network of a group of young adults between 19 and 28 years old, using the strength of the relations between the participants. In addition, we assessed the important characteristics of each participant, such as their openness and expressiveness. For all participants, we collected the PAL data during a period of 30 days. The change of the physical activity per participant was compared with the change predicted by the model.

The paper is organized as follows: in Section 4.2, the social contagion model is explained in more detail. Section 4.3 describes the setup of the conducted experiment. In Section 4.4, we present the results obtained, and we discuss the results in Section 4.5. In Section 4.6, we conclude our explorations and discuss some ideas about possible future work.

4.2 Social Contagion Model

In this section, we briefly summarize the computational model of social contagion and explain how it is used to predict the change in the physical activity level [3, 4, 6].

In the context of this research, the factor that is assumed to spread through the social network is the *physical activity level* of the people in the network. The extent to which people express themselves, which affects the strength of their influence on others, is captured by the concept of *expressiveness*. Similarly, the extent to which people are open to receive influence is represented by the *openness*. The strength of the relation between two people in the network is described by the *connection strength*. These concepts form the key parameters of the contagion model, see Table 4.1. They are formalized as real numbers between 0 and 1.

Tab. 4.1: Parameters for personal and social characteristics.

Physical activity level of person A	q_A
Expressiveness of person A	ϵ_A
Openness of person A	δ_A
Connection strength between sender B and receiver A	α_{BA}

The contagion process is modeled in terms of the contagion strength γ_{BA} between sender B and receiver A. This contagion strength represents the influence of sender B on receiver A. The contagion strength depends on three aspects of the relationship: the expressiveness of person B, ϵ_B, the openness of node A, δ_A, and the connection strength between person B and person A, α_{BA}. The contagion strength between sender B and receiver A is calculated as in (4.1).

$$\gamma_{BA} = \epsilon_B \alpha_{BA} \delta_A \tag{4.1}$$

The overall contagion strength γ_A represents the total contagion strength of all connections of person A in the network. It is computed as in (4.2).

$$\gamma_A = \sum_{B \neq A} \gamma_{BA} \tag{4.2}$$

The proportional weight of the contagion from sender B to receiver A is computed as in (4.3).

$$\omega_{BA} = \epsilon_B \alpha_{BA} / \sum_{C \neq A} \epsilon_C \alpha_{CA} \tag{4.3}$$

The aggregated impact q_A* of all connections of person A is calculated by means of a weighted average as in (4.4).

$$q_A^* = \sum_{B \neq A} \omega_{BA} q_B \quad (4.4)$$

The set of differential equations for the contagion of the physical activity level is formed by (4.5) for all persons N in the network.

$$\Delta q_N(t + \Delta t) = q_N(t) + \gamma_N(q_N^*(t) - q_N(t))\Delta t \quad (4.5)$$

This computational model of social contagion has been used in several studies. For example, it was applied to predict the emotion levels of team members, in order to maintain emotional balance within the team [6]. If the team's emotion level was found to become deficient, the model, which was embedded in an ambient agent, provided support to the team by proposing the team leader to give his employees a pep talk [6]. Another study experimented with simulations of changes in the social network structure in order to guide the contagion process in a certain direction [10].

In the current research, the contagion model that was developed and simulated on emotions is applied on the contagion of physical activity, which to our knowledge has not been done in earlier research. The involvement of the main aspects, such as expressiveness, openness and connection strength are based on the model of social contagion. A detailed method for determining these parameter values for the computational model is described in Section 4.3.

4.3 Experimental setup

The goal of the experiment is to compare the actual change in the activity level of people in a network with the change predicted by the computational model explained in Section 4.2. To do so, an empirical experiment with people that were part of a social network was conducted for 30 days. Characteristics of the persons and their relations were gathered via a questionnaire and objective data about their physical activity was collected with an electronic activity monitor.

The network consisted of 25 participants, all between the age of 19 and 28. The participants were recruited from one person's social network, thus, every participant has at least one connection to another node in the network. Five of the 25 participants provided less than 25 days of useful data and were taken off the experiment, which left a number of 20 participants.

As we use a stable network with relationships that have been established before the start of the experiment, we expected the changes of the physical activity level would be small. This is due to the fact that no external trigger was introduced, e.g., a support system or an encouragement program for doing more activities. Nevertheless, the fact that people are participating in this experiment could intensify their awareness of others' physical activity levels. Therefore, changes could still occur in a smaller ratio.

At the start of the experiment, an intake questionnaire was administered using an online survey software tool. Via this questionnaire, information was obtained about the participants': (1) physical activity level, (2) personal characteristics, (3) level of friendship with the other participants, and (4) frequency of contact with other participants. This information was used to determine the values for the parameters of the computational model.

During the period of 30 days, the participants wore an activity monitor (Fitbit One[3]) that kept track of their daily physical activity. In addition to the data obtained from the Fitbit, short questionnaires were used regularly to collect additional data about their exercise.

The data about the participants' characteristics and their relations is used as basis for the parameters values in the model; the activity data is used as initial input for the model simulations and to compare the outcome of the simulations with. In the next sections, we explain the specific steps that were taken to convert the collected data to numerical values that are suitable for the computational model.

4.3.1 Physical Activity Level

The participants' physical activity data that was collected by the Fitbit activity monitor was automatically stored in the Fitbit servers. After the period of 30 days, this data was exported from the participants' personal Fitbit accounts. The exported data consists of the number of steps taken per day, the number of minutes that the participant was fairly or highly active per day and the number of floors climbed per day. These numbers were divided by the number of recommended steps, fairly/highly active minutes and floors, which are 10,000, 30 and 10 respectively. Days that contained less than 1,500 steps were considered as days that the participant (partly) forgot to wear the Fitbit, so these days were discarded.

Finally, weights were assigned to each aspect, according to their importance. The number of steps taken gives the best estimation of the amount of physical activity and is therefore the most important. The number of fairly/highly active minutes is chosen to be slightly more important than the number of floors, because meeting the recommendation of 30 active minutes per day contributes more to a physically active lifestyle than climbing 10 floors. Therefore, the PAL is calculated as in (4.6).

$$PAL = \left(\frac{steps}{10000} \times 0.7\right) + \left(\frac{am}{30} \times 0.2\right) + \left(\frac{f}{10} \times 0.1\right), \tag{4.6}$$

where am is the active minutes and f is the number of floors climbed.

4.3.2 Tie Characteristics

The strength of connections between people is a combination of the amount of time, emotional intensity, intimacy, frequency of contact and reciprocal services [12]. In this model, a distinction between different types of interaction was added because of

[3] http://www.fitbit.com/one

the assumption that contact in real life and one-to-one communication (also through private chats) both contribute to a higher level of contagion than contact that takes place in group conversations through smartphones or social media and by observing someone's public posts on social media. Therefore, a combination of the level of friendship and the frequency and type of interaction is used to calculate a value for the strength of the connection.

The parameter used to represent the tie strength from node B to node A is the *connection strength* (αBA). This parameter was operationalized by a combination of the type of relation and the frequency of contact. These aspects were measured through questions included in the intake questionnaire. The levels of friendship were measured using a scale on which each participant rated all other network members as: unknown (0.0), acquaintance (0.2), good acquaintance (0.4), friend (0.6), best friend (0.8) and partner (1.0). In addition, two questions concerning the frequency of interaction were included, distinguished by type. The participants gave an estimation, only for the participants who they stated to be connected to in the previous question, of how often they interact with them. These two questions about contact in real life or in private conversations and contact in groups or on social media were answered by the following scale, using the accompanied assigned values: less than once a month (0.0), 1-2 times a month (0.2), once a week (0.4), 2-5 times a week (0.6), once a day (0.8) and more than once a day (1.0). The formula for the tie strength is shown in (4.7).

$$\alpha_{BA} = (fl \times 0.6) + (crl \times 0.25) + (cg \times 0.15), \tag{4.7}$$

where *fl* is the friendship level, *crl* is the amount of contact in real life (i.e., private conversations), and *cg* is contact in groups and social media.

4.3.3 Personality Traits

Personality traits of a person were measured by statements that give an indication of the expressiveness and openness of this person. We formulated a number of statements based on the aspects extraversion, openness to new experience and agreeableness from the Big Five Inventory [8] that were taken as a measure of the values of the participants' expressiveness and openness. The statements that were used to assess these values are listed in Table 4.2. For each domain, three out of six statements were reversed. When using questionnaires with only positive (or negative) sentences, the subjects may be biased in a positive (or negative) way.

Participants were asked to assess how strongly they agreed or disagreed with each statement. A value was assigned to each answer as follows: strongly disagree (0.0), disagree (0.25), neutral (0.5), agree (0.75) and strongly agree (1.0). Some statements were reversed and therefore, the score was subtracted from 1 to obtain the right score for these answers.

We assumed that the expressiveness in our model depends on the Extraversion domain from the Big Five Inventory, and that the Openness and Agreeableness can be used for the degree of openness of the receiver in our model. Thus, statements 1 to 6 determined the value for expressiveness, which is the average of the calculated

Tab. 4.2: Statements measuring expressiveness and openness. (Reversed statements are marked by *.)

Extraversion
1. I keep my feelings and thoughts to myself.*
2. I am assertive.
3. I am outgoing and enthusiastic.
4. I think carefully before I speak.*
5. I am shy and do not like to be the center of attention.*
6. I often post things on social media (Facebook and Instagram).
Openness to new experience
7. I have a strong opinion.*
8. I am interested in what others do and think.
9. My decisions are thoughtful.*
10. I rather remain in my current safe habits and environment than trying and exploring new things.*
11. I am open for suggestions, ideas and opinions of others.
12. I am a curious person.
Agreeableness
13. I have a rigid personality.*
14. I am easily influenced by what others think or do.
15. I have a strong feeling of empathy.
16. I am difficult to persuade by other people.*
17. I have a distant personality towards others.*
18. I feel sorry for other people very quickly.

score over these six questions. The average value of statements 7 to 18 represents the overall value for openness.

4.3.4 Network Structure

Figure 4.1 shows the social network of the participants. Each node represents a participant in the experiment, and the arrows are the direction of the connection, according to the questionnaires filled out by the participants. The thickness of the edges represents the weight of the connection. An overview of the structural characteristics of the network can be found in Table 4.3.

The degrees of the nodes in the graph follow a normal distribution, according to the Shapiro-Wilk test (p = 0.723, H_0 = normal distribution), Anderson-Darling test (p = 0.585, H_0 = normal distribution), and Jarque-Bera test (p = 0.629, H_0 = normal distribution).

The network forms one connected component, which is due to the nature of the selection of the participants. Despite the fact that all of the participants are from the same class, the density of the network is not too high (47.9% of the possible edges existing).

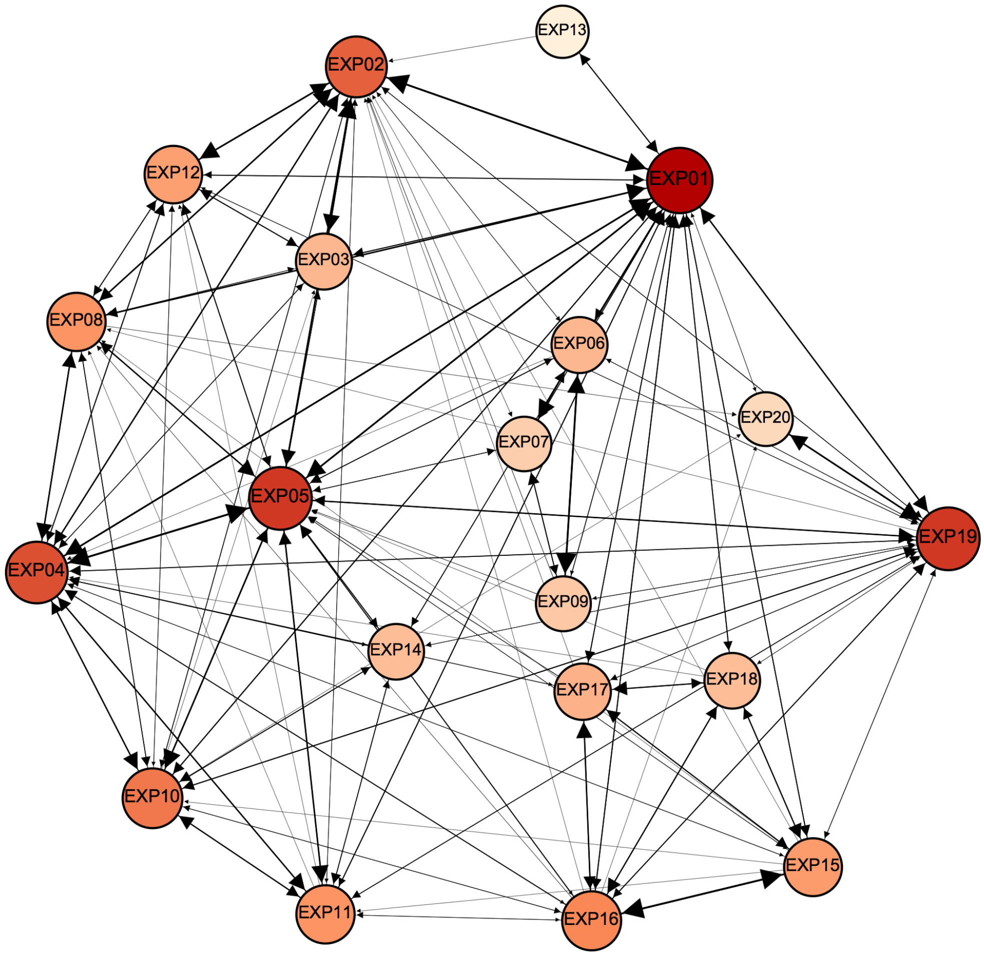


Fig. 4.1: Network structure of the experiment.

4.3.5 Data Preparation & Analysis

This study aims to investigate whether the computational model of social contagion is able to predict the change in the PALs of the participants correctly. In order to derive the direction of change of the PAL of each person, the trendlines for each participant were calculated. They were used to determine whether there was an increase or decrease in their physical activity during the experimental period. This is later compared with the direction of change predicted by the model.

The computational model needs an initial value for the activity level of each of the participants to start the simulation with. Therefore, the initial value is highly important and needs to be chosen carefully. As there is a relatively large variation in the PAL values for different days, it would be unrealistic to use only the PAL from one single day as the starting point for the model. Figure 4.2 shows the PAL values for one participant during the 30 days of the experiment. Since trendlines of the real data were calculated and the value of the trendline at the start provide a good aggregation of the first days of activities, we used the first point of the trendline

Tab. 4.3: Structural characteristics of the network.

Measure	Value
Sample Size	20
Edges	182
Minimum Degree	3.0
Maximum Degree	38.0
Mean Degree	18.2
Std. Deviation Degree	9.082
Graph density	0.479

for each participant as the initial value for the model. The model was then used to simulate the activity level of all participants for 30 days.

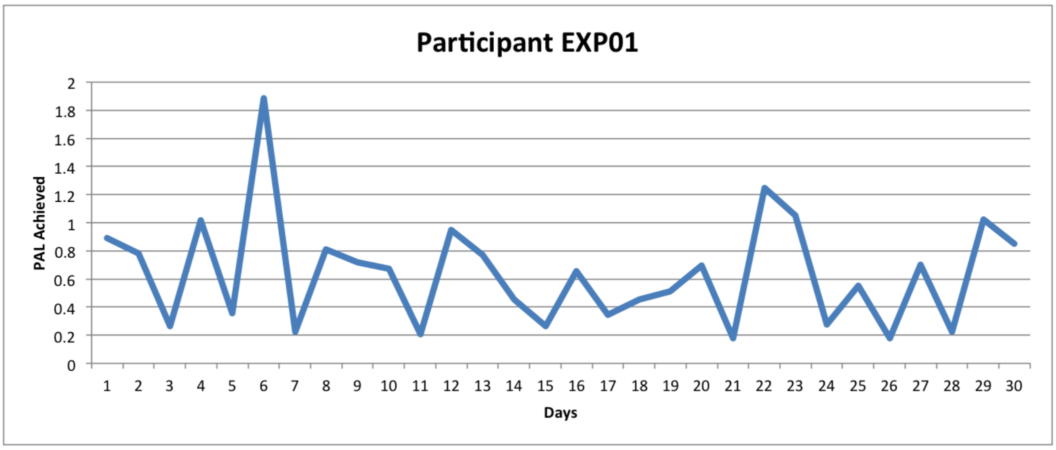


Fig. 4.2: Physical Activity Level achievement of participant EXP01 during 30 days.

The computational model of contagion [3, 4, 6] was implemented in Python, which performed calculations following the formulas discussed in Section 2. It takes three matrices as input for computing the change of PAL over time for each person. The first matrix contains the initial PALs for each participants, the second one contains values of the connection strengths in the network, and the third one contains values for each person's expressiveness and openness. The values in the latter two matrices were the same for all days, as it was assumed that the network structure and the personality characteristics did not change in the period of 30 days.

Before the model was run for the first set of experiments, some tuning was performed. To achieve a realistic speed of change, all contagion strengths were divided by 10, which resulted in realistic levels of contagion over time and in plausible values for changes in the levels of physical activity.

4.4 Results

The trendlines of the physical activity levels (based on the real data) can be seen in Figure 4.3(a). Mann-Kendall tests showed that the trendlines are not significant in

75% of the cases. That fits in our assumption that big changes should not occur in a relatively short period of time within a stable network.

Figure 4.3(b) shows the simulation results for each participant, using a simulation period of 30 days and with parameter values calculated as described in Section 4.3.2 and Section 4.3.3.

Table 4.4 shows that in 80% of the cases, the slopes of the trendlines have the same direction (increasing or decreasing) as the model predicts. It shows that, for stable networks and for a short period of time, the model can predict the direction of the change of the activity level of people with high precision.

As not all of the trendlines presented a clear slope, we took off the slopes with a tau value less than 0.03, i.e. the lines that were almost flat. This value is obtained by a Mann-Kendall test for analyzing the significance of the trendlines for the experiment. The Mann-Kendall test statistically assesses if there is a monotonic upward or downward trend over time of the variable under analysis. In our case, this variable is the PAL of each person. The test has the initial assumption that there is no monotonic trend as its null hypothesis (H_0). If the null hypothesis is rejected, then we have a reasonable indication that there is a trend. The tau value is the variable that assesses the slope of the trend, in case H_0 is rejected. Table 4.4 shows an 87% accuracy after taking out the trendlines that were almost flat.

Tab. 4.4: Comparison of line tendency between contagion model and real data[a] and comparison of line tendency between contagion model and real data after removing trendlines with a slope less than 0.03 [b]. Correct predictions[c] are shown in the last row.

	Model	Real data	Matches[a]	Matches[b]
EXP01	Up	Down	No	–
EXP02	Down	Down	Yes	Yes
EXP03	Up	Up	Yes	Yes
EXP04	Up	Up	Yes	Yes
EXP05	Up	Up	Yes	Yes
EXP06	Up	Up	Yes	Yes
EXP07	Up	Up	Yes	–
EXP08	Down	Down	Yes	Yes
EXP09	Up	Up	Yes	Yes
EXP10	Down	Up	No	No
EXP11	Down	Down	Yes	Yes
EXP12	Up	Up	Yes	Yes
EXP13	Up	Down	No	–
EXP14	Down	Down	Yes	Yes
EXP15	Down	Down	Yes	–
EXP16	Down	Down	Yes	Yes
EXP17	Down	Up	No	No
EXP18	Up	Up	Yes	Yes
EXP19	Up	Up	Yes	Yes
EXP20	Down	Down	Yes	No
Total matches[c]			16 (80%)	13 (87%)

Tab. 4.5: Mean Squared Errors of prediction lines

Method	MSE
Trendline	0.3613
Extrapolation	6.8992
Contagion Model	0.4272

Another way to determine the adequacy of the model is to compare the error of the model predictions with the error of the trendlines. However, this comparison is not fair, as the trendlines consider all the data points in the period, fitting the best linear regression to the entire data set, while the model simulations only use one initial value (based on the start value of the trendline). To make the predictions of a regression line comparable to the model, we created linear graphs using only the first 7 days of data, so it uses the same amount of input data as the model.

Table 4.5 shows the mean squared error for three situations: trendlines based on all data, linear graphs extrapolated from the first 7 days and the predictions of the contagion model. The values of the lines or model predictions in the three cases were compared with the actual data, the differences were squared and the average of these errors was calculated.

4.5 Discussion

Considering that our network is stable, with no formations of new connections nor additional support systems attached to it, we would not expect major changes in the daily level of activities of the participants. Still, we would expect some changes to happen due to the continuous social contagion effect between people from the same class and the increased awareness because of their participation in the experiment.

Figure 4.3(b) shows that the activity levels in the model simulations tend to converge to an average after some time. This is a consequence of the fact that the simulations assume that neither the network nor the personal characteristics of the participants change.

There is a large difference between the mean squared errors shown in Table 4.5 for the model predictions and the extrapolation of the first 7 days. Extrapolating the first 7 days of physical activity and creating new trendlines for each participant increased the mean squared error to 6.8992. This shows that our model results in a far better prediction than linear regression using a similar amount of information. The large error for the extrapolation can be explained by the high variation in the PALs of the participants per day, which makes it difficult to predict the trend based on a few days only. If we take a longer period of analysis, we can reduce the error for the statistical trendline.

In this experiment, we assumed that social contagion was the only factor that influences change in physical activity level. However, this is a clear simplification. Factors like weather and daily duties also have an effect on a person's changes in activity. Similarly, there are also other factors than stated in this research that influence the contagion, such as group norms, physical closeness, age and many more. This makes that our model does not provide a complete description of the contagion process. Related to this, the structure of a social network itself is usually not independent of the characteristics of the participants. This effect is called homophily, which describes the tendency of people to connect with others that have the same lifestyle [13].

4.6 Conclusions

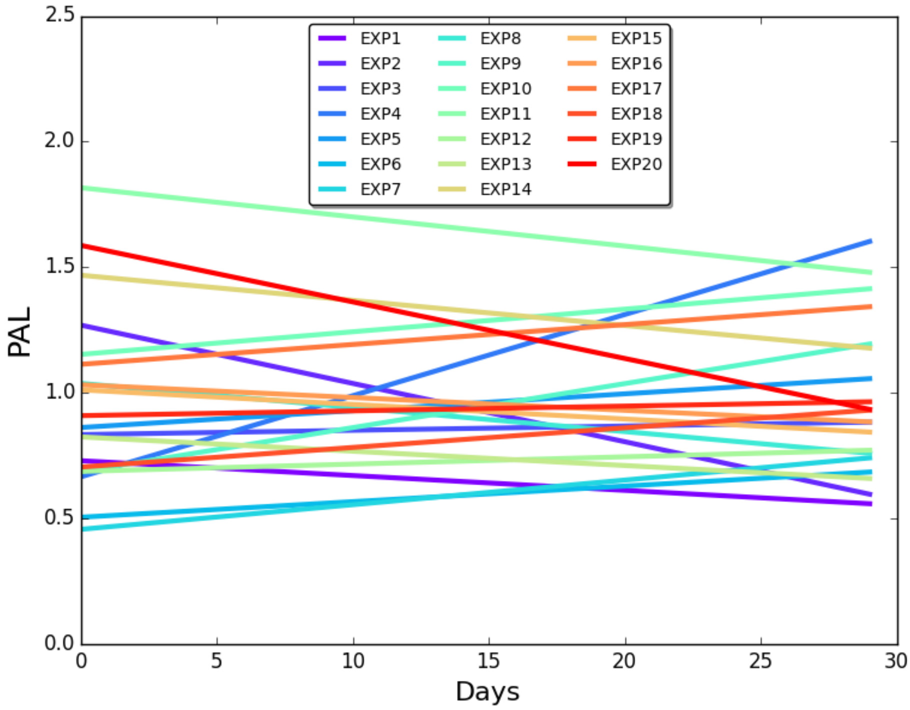
The results of the experiment show changes in physical activity levels of all members of the social network, in smaller or bigger ratios. The computational model of contagion predicted in 80% of cases correctly whether the PAL increases or decreases, which is the vast majority of the participants, and it grows to 87% if only people with a clear change in activity level are considered. We can conclude that the model has a good accuracy in predicting the tendency of the physical activity levels in a small network of 20 people, in a relatively short period of time.

The current experiment aims to understand the behaviour of a stable network and answers the question whether it is possible to use the model to predict the direction of change of activity. These outcomes provide valuable insights for further explorations. To be able to draw conclusions about more dynamic situations, additional experiments should be done. These experiments should last longer, have a higher number of participants, and include changes in the network, e.g. changes in connections or applying interventions in order to influence people's behaviour using external triggers.

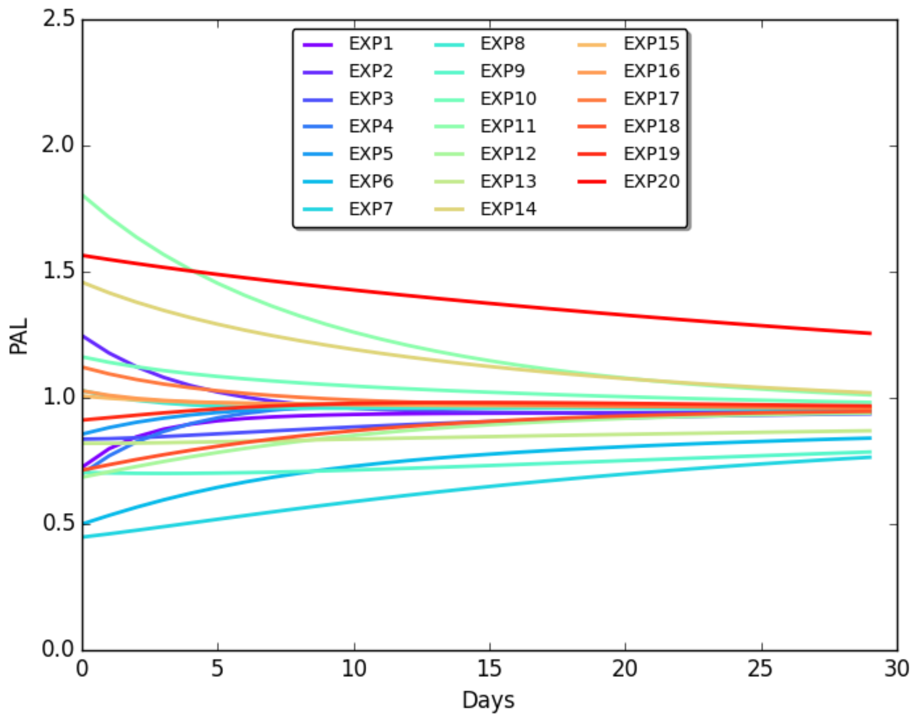
Future work should also investigate other strategies for parameter tuning and compare the resulting parameters with the values obtained from the questionnaires.

Acknowledgements

The authors would like to thank prof. dr. Jan Treur for his valuable input and feedback on the manuscript.



(a) Trendlines for each participant during 30 days.



(b) Simulation results of the social contagion model.

Fig. 4.3: Graphs with trendlines of real data and simulation results of the model.

Bibliography

[1] Steven Allender, Gill Cowburn, and Charlie Foster. „Understanding participation in sport and physical activity among children and adults: a review of qualitative studies“. In: *Health Education Research* 21.6 (2006), pp. 826–835. eprint: `http://her.oxfordjournals.org/content/21/6/826.full.pdf+html` (cit. on p. 54).

[2] Sinan Aral and Dylan Walker. „Identifying Influential and Susceptible Individuals in Social Networks: Evidence from a Randomized Experiment“. In: *Proceedings of WISC* (2010), pp. 1–7 (cit. on p. 54).

[3] Tibor Bosse, Rob Duell, Zulfiqar A Memon, Jan Treur, and C Natalie 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*. Springer, 2009, pp. 48–67 (cit. on pp. 54, 55, 61).

[4] Tibor Bosse, Rob Duell, Zulfiqar A Memon, Jan Treur, and C Natalie van der Wal. „Agent-Based Modeling of Emotion Contagion in Groups“. In: *Cognitive Computation* 7.1 (2015), pp. 111–136 (cit. on pp. 54, 55, 61).

[5] Carl J Caspersen, Kenneth E Powell, and Gregory M Christenson. „Physical activity, exercise, and physical fitness: definitions and distinctions for health-related research.“ In: *Public health reports* 100.2 (1985), p. 126 (cit. on p. 54).

[6] Rob Duell, Zulfiqar Memon, Jan Treur, C Natalie Van Der Wal, et al. „An ambient agent model for group emotion support“. In: *Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on*. IEEE. 2009, pp. 1–8 (cit. on pp. 54–56, 61).

[7] David W Dunstan, Bethany Howard, Genevieve N Healy, and Neville Owen. „Too much sitting–a health hazard“. In: *Diabetes research and clinical practice* 97.3 (2012), pp. 368–376 (cit. on p. 54).

[8] 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. 58).

[9] Pedro C Hallal, Lars Bo Andersen, Fiona C Bull, et al. „Global physical activity levels: surveillance progress, pitfalls, and prospects“. In: *The lancet* 380.9838 (2012), pp. 247–257 (cit. on p. 54).

[10] Michel Klein, Adnan Manzoor, Julia Mollee, and Jan Treur. „Effect of Changes in the Structure of a Social Network on Emotion Contagion“. In: *Web Intelligence (WI) and Intelligent Agent Technologies (IAT), 2014 IEEE/WIC/ACM International Joint Conferences on*. Vol. 3. IEEE. 2014, pp. 270–277 (cit. on p. 56).

[11] Institute of Medicine (US). Committee on Health and Practice. *Health and behavior: The interplay of biological, behavioral, and societal influences*. National Academies Press, 2001 (cit. on p. 54).

[12] André Escórcio Soares and Miguel Pereira Lopes. „Social networks and psychological safety: A model of contagion“. In: *Journal of Industrial Engineering and Management* 7.5 (2014), pp. 995–1012 (cit. on p. 57).

[13] Thomas W Valente. *Social networks and health: Models, methods, and applications*. Oxford University Press, 2010 (cit. on p. 64).

[14] Ward Van Breda, Jan Treur, and Arlette Van Wissen. „Analysis and support of lifestyle via emotions using social media“. In: *Social Informatics*. Springer, 2012, pp. 275–291 (cit. on p. 54).

[15] Corneel Vandelanotte, Kym M Spathonis, Elizabeth G Eakin, and Neville Owen. „Website-delivered physical activity interventions: A review of the literature“. In: *American journal of preventive medicine* 33.1 (2007), pp. 54–64 (cit. on p. 54).

The effect of a community in a health promotion program[1]

5

> *"It is easy to acknowledge, but almost impossible to realize for long, that we are mirrors whose brightness, if we are bright, is wholly derived from the sun that shines upon us."*
>
> — **C.S. Lewis**
> "The Four Loves"[2]

Abstract

In this paper we study the effect of being part of an online social network on the change in physical activity level (PAL) during a health promotion program.
Our aim is to understand the difference between the effect of being part of an online community and the self-selection effect.
We compare the effect on change in physical activity of people that are part of a community during the health intervention with the effect on people that are not part of such a community but will become a member later.
Our analysis confirms that people who partake in an online community have a higher level of PAL. It also shows that being part of a network during the intervention has no significant effect on the change in PAL. People who are willing to become part of a community have a significantly higher change in physical activity than people who do not become a member of such a community.
The fact that people in an online community are more active seems to be mainly caused by self-selection. There is no positive effect visible of the network during the intervention. The willingness to become a member is a good predictor for the increase in physical activity.

[1]Extended work from the paper published as: A. Manzoor, J. S. Mollee, **E. F. M. Araújo**, A. T. V. Halteren and M. C. A. Klein, *Online Sharing of Physical Activity: Does It Accelerate the Impact of a Health Promotion Program?*, 2016 IEEE International Conferences on Big Data and Cloud Computing (BDCloud), Social Computing and Networking (SocialCom), Sustainable Computing and Communications (SustainCom) (BDCloud-SocialCom-SustainCom), Atlanta, GA, 2016, pp. 201-208.

5.1 Introduction

5.1.1 Motivation

Engaging in sufficient physical activity has many positive effects on physical and mental health [5, 7], while low levels of physical activity have been associated with increased risk of cardiovascular diseases, cancer, diabetes, and mental illness [14]. Despite this, a large proportion of the Western population does not meet the guidelines of being moderately to vigorously active for at least 30 minutes on at least five days a week [12]. Therefore, physical activity promotion is a priority in most Western countries and many (online) intervention programs exist. It is important to understand which elements of these physical activity programs are effective or could potentially accelerate the impact of these health promotion programs.

5.1.2 Aim

Previous research has suggested that the physical activity of people is positively influenced when they are connected in an online social network that allows them to share their results [10]. Our first aim is to verify whether this is visible in a large scale dataset of people participating in a health promotion program. To do so, we compare the change in physical activity level (PAL) of participants that were connected to others in an online community with participants that were involved in the same promotion program but were not connected to others in an online community. Secondly, we want to investigate whether the change is caused by the actual sharing of the results, or by a self-selection bias, e.g. because people that opt-in for an online community are more motivated or otherwise better prepared for a positive change in physical activity. We examine this by looking at the physical activity of people that are not connected during the health promotion program. We compare the change in physical activity between people that will later become part of a community and people that will never opt-in for such a community.

In short, the two hypotheses we investigate are: (a) people who are connected are influenced by their peers in positive ways, and (b) people who are willing to become part of an online community where they share information about their PAL will have a bigger improvement on their PAL than people who do not want to become part of such a community.

The remainder of this paper is organized in the following way. Next, we discuss some related work about the influence of online social networks and (mobile) healthy lifestyle interventions on behaviour. Section 5.2 presents the data, its characteristics and our methods for processing it. In Section 5.3, we provide statistical analyses to answer our research questions. We conclude with our main findings and directions for future work in Section 5.4.

5.1.3 Related work

Previous analyses [10] show that there is a positive relation between being part of the online community of a physical activity intervention and the physical activity level of participants. The online community therefore matters. It has also been shown that the number of contacts in the online community does not have a significant effect on the physical activity level, while network density even has a significant, negative effect. On the other hand, adding online community features to an Internet-mediated walking program did not increase average daily step counts, but did reduce participant attrition [22].

Online social interaction plays an important role in forming or adapting some kind of behaviour based on the peer's behaviour. Studies have shown recently that online social networks are equally responsible (as offline networks) in the diffusion of one's emotions to another [6]. It is often difficult to adopt new behaviour and adhere to it, but it has been shown that close social circles (such as family, friends, and co-workers) are helpful in sustaining a healthy lifestyle [15, 26]. In [25], the role of online social interactions is discussed in the context of developing and maintaining a healthy lifestyle, e.g. an ambient system can continuously monitor and help people to alter their social ties in order to sustain healthy behaviour. Having an infrastructure like a social network already available, social network interventions could be designed to leverage the full potential of a social network [13], for example in case of a health behaviour change program.

With the rise of mobile technology, there has also been a steep increase in the number of healthy lifestyle interventions that are available through a smartphone. As of May 2016, the number of apps in the Health & Fitness category has grown to 67,552 for the Google Play Store [2] and 68,248 for the iTunes App Store [21]. A systematic review of apps that promote physical activity has shown that even though most apps apply only a few behaviour change techniques [1, 17], a majority of these apps (approximately 58%) provided a form of social support or social change [17]. This was done, for example, through providing chat possibilities among users or through enabling a link to an external virtual social network, where users could share their goals or achievements [17].

It is widely believed that mobile technology can be a useful tool to promote physical activity among a large part of the population. Firstly, average smart phone ownership numbers are high: 68% in the United States, with higher numbers among young adults (86% in ages 18-29 years and 83% in ages 30-49 years) [19], and 80% in the Netherlands [8] in the third quarter of 2015. This means that interventions designed for smartphones can theoretically reach a large number of people. Secondly, mobile interventions are always accessible to the user, and also allow for continuous monitoring and (if applicable) feedback. In addition, similar to interventions delivered over the Internet, mobile interventions can reduce stigma and lower the barrier for people to address their (health) issues [9]. In combination with the relatively high number of apps that enable social support or social change, these advantages of mobile interventions imply that smartphone apps are a very suitable means to guide social influence for behaviour change.

5.2 Methods

This section describes the data set that is used for the analysis. In Section 5.2.1, we describe the process of data collection and the resulting data. Section 5.2.2 describes the way in which we processed the data to select suitable subsets, and some of the structural characteristics of the selected social network components are presented in Section 5.2.2.

5.2.1 Data collection

The data set used for the analysis contains the physical activity levels of people who were part of an online physical activity promotion program. The original data set contains data of around 50,000 people, from where some were selected as intervention and control groups for the scenarios studied in this paper.

The promotion program has three different phases. The first phase is a one-week assessment period that is used to evaluate the user's activity level during their daily routine. The assessment is followed by the second phase: a 12-week plan that aims to gradually increase the user's activity level towards a specified end goal. The goal is determined based on the physical activity reported during the assessment week. After the plan, the members of the program can opt to start a new 12-week plan to further increase their activity level or simply continue with the activity goal set during the last week of their program. This constitutes the third phase.

The activity promotion program provides an online community and joining this social network is optional for the users. Each member of the community can connect to other users (i.e., become online friends), exchange messages and see the relative achievements of themselves and their connections (which are only visible after the participants confirm their connection). Around 5,000 people in the data set opted to join the online community at some point in time during their usage of the system.

The participants in the program wear an activity monitor device that measures their physical activity level (PAL). When they register to the program via the website, the participants fill in their gender, age, and nationality. In addition, the data set contains information about the date that people start the program, the company they work at (if the program is offered via a company), and their friendship connections with other participants. In order to ensure anonymity of the participants, their age was omitted from the data set before the analysis.

5.2.2 Data selection

This section presents the data selection process for the two scenarios studied in this work. Each scenario compares two subgroups of individuals, namely intervention and control groups. *Scenario 1* compares individuals that are effectively connected to at least one other participant in the community since day one of the person's program (intervention) with individuals who are never connected to anyone (control). *Scenario 2* compares individuals who joined the community for sharing their PALs in the social network provided by the program (intervention) with individuals that opted out of the community (control). In scenario 2 the participants are willing to join the online network, and will eventually get connected to other participants throughout the program.

The data is represented in two files: a GEXF (Graph Exchange XML Format) file representing the network structure of the community; and a CSV file with the PAL values of all participants, and their characteristics, such as gender, BMI (Body Mass Index), corporation and country. In the GEXF file, every new connection (edge) is included when a friendship request is accepted. Connections where the friendship request was not accepted were removed from the graph to not influence the results.

We first describe the data selection process for both scenario's. Then, we provide a structural analysis of the social networks that play a role in scenario 2.

Scenario 1: individuals connected in the online network

The first scenario compares individuals who had created at least one friendship (connection) at day 1 of their plan. The "connected" individuals are treated as the intervention group. A selection of individuals that opted out of the community in the program with similar demographic characteristics are selected to compose the control group.

We only selected participants from both groups that (a) have PAL data for all 84 days of the program, (b) did not drop out from the experiment. People from the intervention group should have a degree bigger or equal to 1 when their first plan starts. All the individuals from both groups started their plans within a week of difference. The requirements for the selection of the individuals from the intervention group resulted in 40 individuals.

The control group has the same number of individuals as the intervention group. Initially, 77 participants were selected with PAL data for all 84 days in the plan and an initial start date within a week of difference to the intervention group. Then we paired 40 nodes from this selection with the 40 nodes from the intervention group. For doing so, we selected nodes that match their peer with the following criteria: (a) the initial mean PAL is in a range of $[-15\%, 15\%]$ compared to the connected peer; (b) same gender; (c) same first plan dates; (d) work at the same corporation.

The characteristics of the control and intervention groups can be seen in Table 5.1. The two groups present very similar characteristics, and therefore are very well balanced. The BMI of the individuals in the control group is not present in the data set. The same holds for the country where the person lives. The mean *before plan PAL* is calculated in the week before the person starts its program, the assessment week. It is the average of the daily PAL in the 7 days before the start of the program.

Tab. 5.1: Meta-Properties of Intervention and Control Groups of Scenario 1

	Connected (intervention)	Not connected (control)
Number of Participants	40	40
Dropouts	0	0
Start (earliest)	03/05/2010	03/05/2010
Start (latest)	10/05/2010	10/05/2010
Gender ratio (M/F:%)	M: 42.5%, F: 57.5%	M: 45.0%, F: 55.0%
Average BMI	27.10	Not available
Mean before plan PAL	1.6767	1.6265

Scenario 2: willingness to be part of a community data set

For the second scenario, we select people that will become part of a community later, but do no have any connections during their first plan (the first 84 days of the intervention program). For this experiment, we also balance the nodes in the intervention and control group very carefully. In addition to the criteria used before (initial PAL, gender, company, start date), we also have to balance the *type of community* that people will partake in. Our aim is that the structural characteristics of networks that people will be part of are comparable.

To do so, we take the complete network graph containing all the connections and split it in connected components. A connected component (or just component) of an undirected graph is a subgraph in which it is possible to find a path linking any two nodes. The component can't be connected to any other node in the super graph. In order to extract the components, we use Python's NetworkX library [11], which is based on the community detection algorithm Tarjan's algorithm with Nuutila's modifications [18, 24].

The total number of communities (network components) is 395. One of them is a large community with 3,926 participants; the second largest community consists of 42 participants. Figure 5.1 shows an overview of the number of participants in each of the components. The components are ordered by size, and the largest component (of 3,962 participants) is left out.

Since there can be multiple consecutive plans (i.e., periods of twelve weeks in which people are stimulated to increase their activity level), the PAL values used in the analysis represent the first 12-week plan, in order to ensure fair comparisons.

For the analysis we only selected the network components with (1) a limited number of participants with missing data and (2) a maximum difference of four months between the plan start dates of the members of the component. We discarded the largest component, as the earliest and latest start dates are three years apart. The second largest component with 42 nodes is not included in the analysis because a lot of data is missing for that component.

This selection process yielded 10 of such connected components, consisting of 109 individuals in total. We left out 25 individuals for whom PAL values were missing for

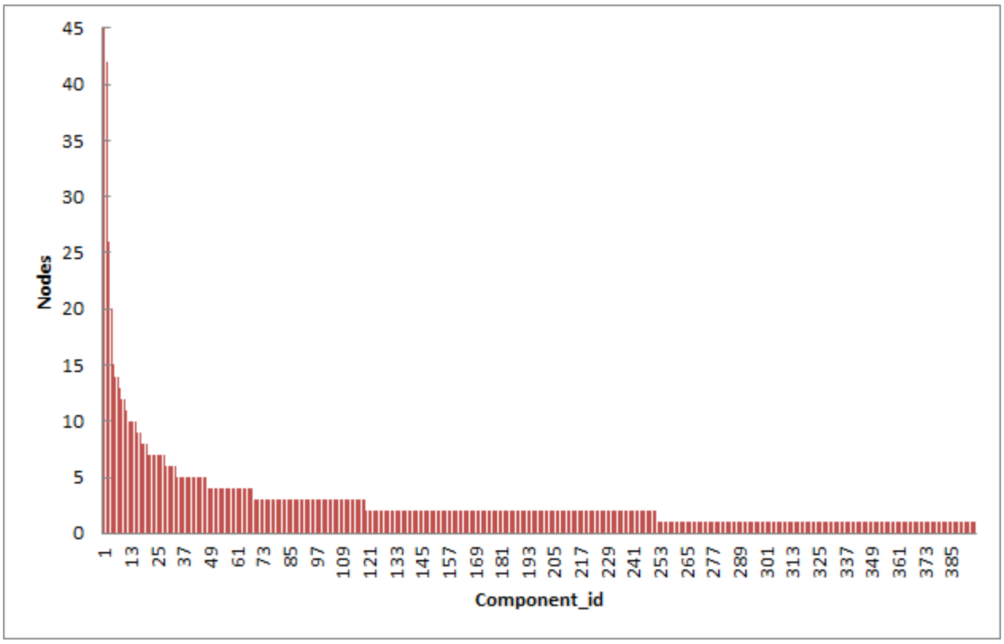


Fig. 5.1: Number of nodes in each of the components

one or more weeks, for instance because they dropped out of the program. Eventually, this resulted in 84 individuals in the intervention group for this scenario.

For the control group, we selected a set of individuals who never opted in for the community, but who are otherwise similar to the participants in each of the components in the intervention group. We balanced the data with respect to the following characteristics: (a) the participants work in the same companies, (b) their plan earliest and latest start and date are similar to the corresponding component, and (c) their gender ratio is also similar to the matching component. As the number of non-community individuals is much larger than the number of individuals within a community, we randomly selected a set of around five times the size of the number of people in the community component with corresponding characteristics, resulting in a set of 501 people. For this data set, we also avoided including individuals with missing data, i.e. individuals who dropped out of the program or who had missing PAL data for one or more weeks. In total, this resulted in a set of 498 participants. Based on the selection of the individuals in both groups, the PAL values are extracted for the two subsets.

A summary of the data used in scenario 2 is given in Table 5.2 and Table 5.3. In Table 5.2, each row shows several meta-data characteristics of the selected components of the network. 'Component' is the ID of the network component, and 'Number of Participants' represents the total number of participants in the component. It can be seen that the size of the selected components varies between 7 and 26 participants. 'Dropouts' shows the number of people that were omitted from the component, because at least one week of PAL data was missing. The 'Start (earliest)' and 'Start (latest)' columns show the earliest or the latest date on which people in a component started their first plan. As can be seen, in this scenario the differences between the start dates are much larger than for scenario 1.

The last column shows the number of individuals with similar characteristics identified in the non-community data set and the number of them that were randomly selected for the control group. As mentioned earlier, the size of the control group is about five times the intervention group size. For example, for component A (consisting of 26 participants), 130 participants were randomly selected from a set of 2,735 individuals with similar characteristics. However, for some components, we could not find enough individuals with similar characteristics for the non-community data subset. For example, only six individuals were found for the non-community data subset corresponding to component G.

Table 5.3 illustrates different characteristics of people in the components, such as their nationality. The 'Country' column shows that the participants in each of the communities are from the same country, namely Germany, the Netherlands or the United States of America. 'Number of Corporations' shows whether all people in a certain component work in the same or different organizations. It is possible that people in a community work in different organizations, like in components C, H and J. In rest of the components, the participants all work in the same company. The column 'Gender Ratio' provides information about the ratio of male and female participants in each of the communities. 'Average BMI' represents the average BMI for each of the components.

Tab. 5.2: Meta-Properties of Selected Components (Intervention Group) of Scenario 2

Component	Number of Participants	Dropouts	Start (earliest)	Start (latest)	Number of Participants Non-Community
A	26	0	25/01/2010	22/03/2010	130 / 2,735
B	15	4	15/02/2010	26/04/2010	70 / 838
C	13	4	18/05/2009	17/05/2010	65 / 178
D	9	1	16/05/2009	20/07/2009	45 / 74
E	9	3	25/01/2010	12/04/2010	45 / 2,839
F	8	0	25/05/2009	19/04/2010	40 / 608
G	8	6	19/04/2010	21/06/2010	6 / 6
H	7	6	15/02/2010	07/06/2010	30 / 35
I	7	1	22/02/2010	22/03/2010	35 / 358
J	7	0	02/03/2009	27/07/2009	35 / 335

Structural analysis of the components of scenario 2

As described above, we selected 10 components from the community for the analysis in scenario 2, ranging from 7 to 26 participants each and with different configurations. The difference in the structural characteristics between the components can be seen in Figure 5.2. Social network analyses were done on the components in order to understand the structure of the connections.

The components are mostly sparse networks with a low average degree and low clustering coefficient, meaning that the neighbors of each node are not well connected among themselves. Because of the nature of the online friendship connections, all connections in the network are bidirectional.

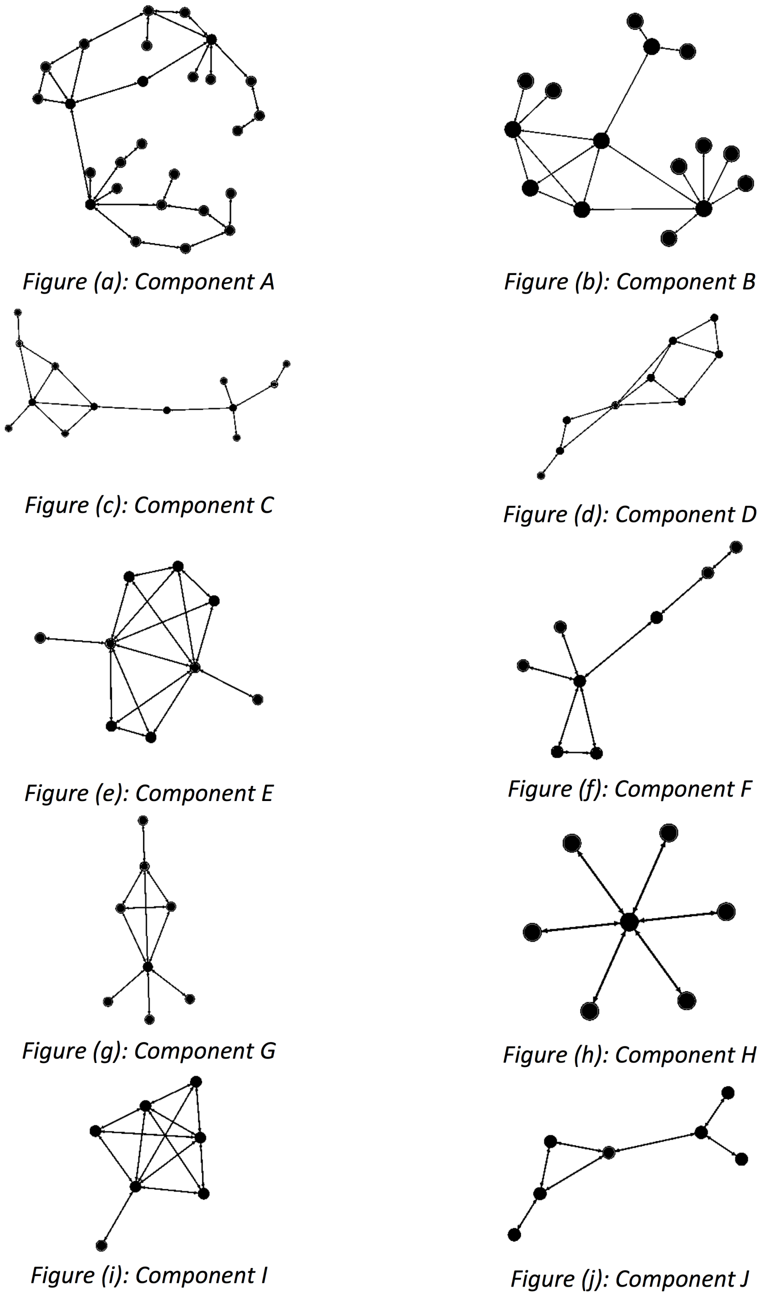


Fig. 5.2: The network components used for the community analysis in scenario 2.

Tab. 5.3: Characteristics of Participants in Selected Components (Intervention Group) of Scenario 2

Component	Country	Number of Companies	Gender Ratio (M/F:%)	Average BMI
A	DE	1	M:88.5, F:11.5	25.85
B	NL	1	M:100.0, F:0.0	25.74
C	US	3	M:47, F:53	24.37
D	US	1	M:89, F:11	31.76
E	DE	1	M:34, F:66	23.97
F	US	1	M:87, F:13	30.78
G	NL	1	M:58, F:42	25.60
H	NL	4	M:100.0, F:0.0	32.10
I	NL	1	M:86, F:14	28.05
J	DE	3	M:86, F:14	25.65

Details of two components will be given to further illustrate the data. Component I has the highest average density and clustering coefficient, both more than 60%. It also presents a small diameter, which means that the nodes are very well connected, and are very close to each other. In this network, the degree of the nodes ranges from 2 to 12. One of the nodes with the highest degree is connected to all the other nodes in the network, having an important role for the social influences in this component.

Component E is also a well-connected component with a small network diameter, as in most social networks in real life. This component has an average density and clustering coefficient of around 50%, which makes the network well connected, but not very dense. Two nodes have only one connection, and the rest of the network presents a very good clustering coefficient. Table 5.4 shows the detailed characteristics of this component.

Tab. 5.4: Detailed Characteristics of Component E

Number of Nodes	9
Edges	32
Average Degree	3.556
Average Path Length	1.583
Network Diameter	3
Density	0.444
Average Clustering Coefficient	0.622
Country	Germany

The pictures shown in Figure 5.2 show the final clusters after the end of the first plan for all of the nodes in the network. A more detailed study about the dynamics of the connections is presented by Mello Araújo et al. [16] for this same data set. Mello Araújo et al. [16] present a very detailed study about the evolution of the giant component in the network, showing how the number of edges grow over time, changing the centrality and other measurements that would help identifying leaders (or potential influential nodes) in the network.

5.3 Results

In this section we describe an analysis of the change in the physical activity levels of individuals participating in a health promotion program, according to the scenario's described before. We first present a visual comparison of the trend lines, and in the remaining subsections we provide a more detailed statistical analysis.

5.3.1 Visual comparison

Our first analysis is based on a visual comparison of the differences between the two contrast groups for both scenarios 1 and 2.

For Scenario 1 (connected individuals), the average PAL values for both groups over the program is shown in Figure 5.3. It can be seen that (a) connected people are more active at the beginning, as their mean PAL is higher than the group of non connected people, and (b) both groups have a similar trend line showing a degradation of the PAL over time.

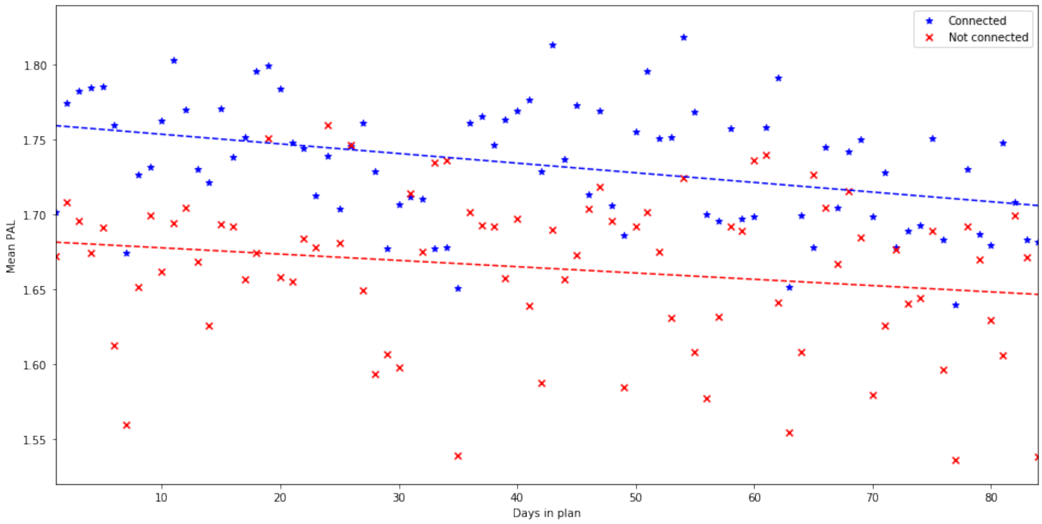


Fig. 5.3: Physical activity levels (PALs) of connected vs. not connected participants during 84 days of the intervention, including linear trendlines.

For scenario 2 the average PAL values for both groups during twelve weeks (84 days) are shown in Figure 5.4. The figure illustrates that community people are more active, since their average PAL is higher than the average PAL of the non-community participants. It also shows that the linear trend line of both groups has a different slope: the average of individuals that eventually become part of a community shows an increase of PA, while the trend of the other group is slightly negative.

5.3.2 Multiple linear regression model

For a more thorough analysis, we use statistical methods. A multiple linear regression model is performed to predict the average PAL at the end of the program (i.e., the last three weeks) based on the group to which the person belongs and the average

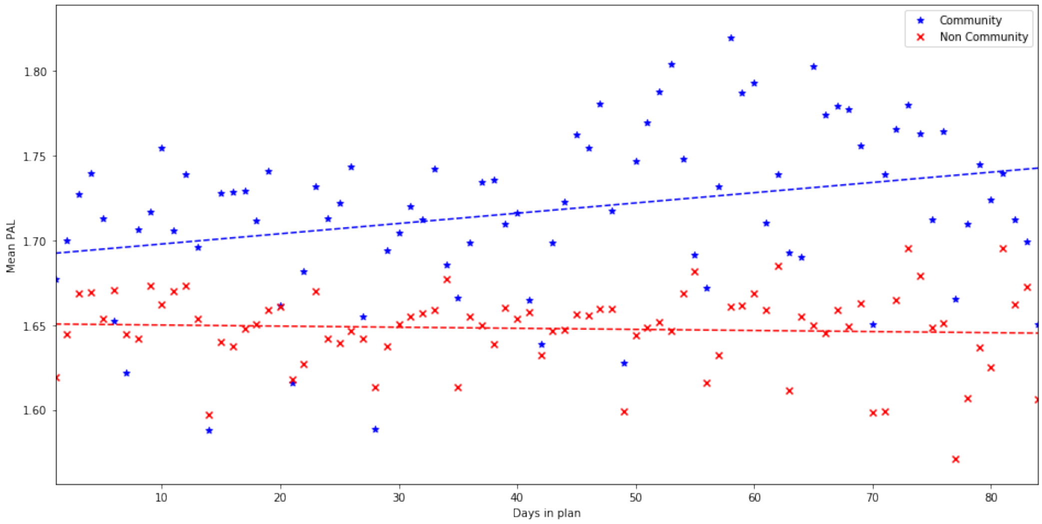


Fig. 5.4: Physical activity levels (PALs) of community vs. non-community participants during 84 days of the intervention, including linear trendlines.

initial PAL (i.e. at the start of the program). For the average PAL at the start of the program, we consider the second and third weeks. The first week is left out, because this week is usually a bit atypical, presumably due to novelty effects of starting the program.

For scenario 1, a dummy variable (*Connected*) is coded with the value '1' if a person is connected to at least one other person in the network at day 1 in the program and '0' if the person is not connected at all.

The results are illustrated in Table 5.5. A significant regression model was found ($F(2,77) = 30.35$, $p < .001$). The model accounts for 44% of the variance in the PAL values of the participants at the end of the program, $R^2 = .4408$. For this scenario, the predictor variable *Start-PAL* is statistically significant ($p < 0.01$), but the predictor *Connected* is not. Therefore, this variable is removed from the model, as it does not explain differences between the two groups. Therefore, the final model shows that the predicted PAL for the last three weeks is based only on the initial PAL, equal to 0.286716 + 0.8081 * *Start-PAL*. In other words, the PAL by the end of the program is around 80.81% the initial PAL. This model shows that being connected is not a good indicator of changes in PAL after the health promotion program.

Tab. 5.5: Analysis using multiple linear regression for scenario 1

	Estimate	Std. Error	*p*-value	95% C.I.
(Intercept)	0.286716	0.180152	.116	[-0.07201158, 0.64544338]
Start-PAL	0.811268	0.105206	$< .001$	[0.60177511, 1.02076048]
Connected=1	-0.009441	0.051677	.856	[-0.11234235, 0.09346131]

For the analysis of the average difference between groups in scenario 2, a dummy variable (*Community*) is coded with the value '1' if a person will eventually become part of the online community and '0' if the person will never opt in for the community. The results are illustrated in Table 5.6. A significant regression model was found ($F(2,579) = 227$, $p < .001$). The model accounts for 44% of the variance in the PAL values of the participants at the end of the program, $R^2 = .4395$. Both predictor variables, *Start-PAL* and *Community*, are statistically significant, $p < .05$. The model shows that the predicted PAL for the last three weeks is equal to 0.23562 + 0.05061 * *Community* + 0.85041 * *Start-PAL*, where *Community* is 0 or 1. The model signifies that being member of a community is associated with an increase of approximately 0.05 in physical activity level.

Tab. 5.6: Analysis using multiple linear regression for scenario 2

	Estimate	Std. Error	*p*-value	95% C.I.
(Intercept)	0.23562	0.06743	<.001	[0.103183, 0.36806]
Start-PAL	0.85041	0.04091	<.001	[0.77005, 0.93076]
Community=1	0.05061	0.02327	.0300	[0.00490, 0.09631]

5.3.3 Linear mixed model

The regression model described in the previous section only compares the PAL at the start with the PAL at the end of the program. A linear mixed model can be used to take into account all days of data (except for the first week, as mentioned above). Since the data is longitudinal by nature, we follow the approach as outlined in [3] for the same scenarios presented in the previous section. A sample of the data for scenario 2 is shown in Table 5.7. Each row represents one day's PAL for an individual, and there are 77 rows (eleven weeks) for each individual. The column "Community" is set to 1 in case the participant decided to join the community program. For scenario 1 this column refers to participants with at least one connection at day 1 of the health promotion program, for scenario 2 it refers to people that will eventually partake in the community.

Tab. 5.7: Physical activity level data in long format for community and non-community individuals for scenario 2

Id	Time	PAL	Community
1	8	1.57800	1
1	9	1.85780	1
1	10	1.78080	1
.	.	.	.
.	.	.	.
582	82	1.5803	0
582	83	1.7658	0
582	84	1.4576	0

As explained in [3], we first conduct the test using a simple model based on the generalized least square method and later add random effects to the intercepts in the simple model to see if the two models differ significantly. For this purpose, R's NLME library is used [20]. Since we are primarily interested to see whether becoming a member of the intervention group makes a difference over time, the model includes an interaction term, i.e. a product of *Connected* (for scenario 1) or *Community* (for scenario 2) and *Time*.

Scenario 1

Table 5.8 shows the results for the simple model (no random effects) for scenario 1. The reference group in this scenario is the group of individuals that are connected to someone from day 1 of the program (*Connected*=1). The effect of being connected on the PAL is the estimator for this model. It tests if there is a difference on the PAL over time for the connected and not connected groups. As can be seen, this is not the case here, as the variable for connectedness does not present a significant estimate. Therefore, being connected is not a factor that can describe differences in the two groups.

Tab. 5.8: Analysis of scenario 1 using generalized linear regression

	Estimate	Std. Error	*t*-value	*p*-value
(Intercept)	1.7623347	0.014116961	124.83810	$< .001$
Connected=0	-0.0450159	0.019964398	-2.25481	.0242
Time	-0.0006828	0.000276326	-2.47116	.0135
Connected=0 : Time	-0.0002165	0.000390783	-0.55400	.5796

The analysis above does ignore the fact that the start PAL (i.e., the intercept) of each of the individual is very different. To account for this, we can add a mixed effect for this value. The results of this more advanced random intercept model are shown in Table 5.9. The results show that there are only some small differences in the standard error compared to the simple model. Similar to the model in Table 5.8, being connected to someone from the start of the program is taken as the reference group (*Connected*=1). The intercept therefore represents the predicted PAL scores for the people connected, and the estimated coefficient for *Connected*=0 indicates the difference between the predicted PAL for the people in the intervention group and the people in the control group. In this scenario, there is not a significant difference for the two groups when using the variable *Connected* to explain the changes in the PAL. The coefficient of *Time* indicates that for every unit of time, there is an decrease of -0.0006828 in the PAL for people in the connected group. The estimated coefficient for the interaction term represents the difference in the slope for the two groups. In other words, the interaction term tells us that the two groups (connected vs. not connected) show a significantly different change in PAL over a period of twelve weeks. Once more, for this scenario this is not the case, and the interaction term is not significant.

Tab. 5.9: Analysis using linear mixed effects modeling for scenario 1

	Estimate	Std. Error	df	*t*-value	*p*-value
(Intercept)	1.7623347	0.03968284	6078	44.41050	$< .001$
Connected=0	-0.0450159	0.05612001	78	-0.80214	.4249
Time	-0.0006828	0.00019633	6078	-3.47802	$< .001$
Connected=0 : Time	-0.0002165	0.00027765	6078	-0.77973	0.4356

The likelihood ratio test is often conducted to test the significance of predictor variables, i.e. to compare the fit of one model (with a reduced set of predictors variables) to the fit of another model (with a complete set of predictor variables).

Here, we also use this test to see which model provides a better fit for the empirical data. Model 1 is based on a generalized linear regression (Table 5.8) and model 2 is based on a linear mixed effects model (Table 5.9). The latter includes all the variables of model 1, plus an additional mixed effect for the individuals' intercepts. The results are shown in Table 5.10. The null hypothesis (stating that the between-subject variation in the intercept is equal to zero) is rejected, $\chi^2(1) = 3691.415$, $p < .001$. This tells us that adding a random effect for the individuals to the model is a significant improvement, therefore the mixed effect model provides better fit for the empirical data.

Tab. 5.10: Comparison of standard linear regression model with random intercept model for scenario 1

Model	df	logLik	Chi Sq.	p-value
standard (1)	5	-2131.4225		
advanced (2)	6	-197.5809	3691.415	$< .001$

Scenario 2

The results of scenario 2 for the simple model (without random effects) are shown in Table 5.11. Here, becoming part of the community is taken as the reference group (*Community*=1), in contrast to the model presented in Table 5.6. The estimates associated with the predictor variables indicate the effect of the program on the PAL. So, the interaction term tests whether the effect of the program on the PAL of the participants is different for people inside or outside the community group. The results show that this is indeed the case: people perform differently in the two groups. In this analysis, not being part of the community is again associated with a lower PAL value (with a difference of approximately 0.06).

Tab. 5.11: Analysis of scenario 2 using generalized linear regression

	Estimate	Std. Error	t-value	p-value
(Intercept)	1.6933389	0.009814656	172.53167	$< .001$
Community=0	-0.0586920	0.010610159	-5.53168	$< .001$
Time	0.0005821	0.000192112	3.02993	.0024
Community=0 : Time	-0.0006525	0.000207683	-3.14177	.0017

The results of the more advanced random intercept model are shown in Table 5.12. As done in scenario 1, we account for the fact that the start PAL (i.e., the intercept) of each of the individuals is different by adding a mixed effect for this value. The results show that there are only some small differences in the standard error compared to the simple model. Similar to the model in Table 5.11, being part of the community is taken as the reference group (*Community*=1). The intercept therefore represents the predicted PAL scores for the people in the community, and the estimated coefficient for *Community*=0 indicates the difference between the predicted PAL for the people in the community group and the people in the non-community group. The coefficient of *Time* indicates that for every unit of time, there is an increase of 0.0005821 in the PAL for people in the community group. The estimated coefficient for the interaction term represents the difference in the slope for the two groups. In other words, the interaction term tells us that the two groups (community vs. non-community) show a significantly different change in PAL over a period of twelve weeks.

Tab. 5.12: Analysis using linear mixed effects modeling for scenario 2

	Estimate	Std. Error	df	*t*-value	*p*-value
(Intercept)	1.6933389	0.02388	44230	70.901	<.001
Community=0	−0.0586920	0.02581	580	−2.273	.0234
Time	0.0005821	0.00015	44230	3.791	<.001
Community=0 : Time	−0.0006525	0.00016	44230	−3.931	<.001

To see which model provides a better fit for the empirical data, we also applied the likelihood ratio test. Model 1 is based on a generalized linear regression model (Table 5.11) and model 2 is based on a linear mixed effects model (Table 5.12). The latter includes all the variables of model 1, plus an additional mixed effect for the individuals' intercepts. The results are shown in Table 5.13. The null hypothesis (stating that the between-subject variation in the intercept is equal to zero) is rejected, $\chi^2(1) = 17{,}882.63$, $p < .001$. This tells us that adding a random effect for the individuals to the model is a significant improvement, therefore the mixed effect model provides a better fit for the empirical data, as in scenario 1.

Tab. 5.13: Comparison of standard linear regression model with random intercept model

Model	df	logLik	Chi Sq.	*p*-value
standard (1)	5	−15712.890		
advanced (2)	6	−6771.574	17882.63	<.001

5.4 Discussion

5.4.1 Principal Results

In this research we explore two different scenarios to investigate whether an intervention aiming at increasing physical activity presents different results for people who want to share their data with others through an online network. The first scenario consists of people who decided to participate in the community for sharing their behaviour and established connections from the beginning of the program. The second scenario consists of people who joined the community showing willingness to be part of a network, but weren't necessarily connected from the start of the program. Two statistical analyses were performed. In the first analysis, a significant linear regression model was found for both scenarios. Based on the adjusted R^2, we conclude that around 44% of the variance in the PAL values is explained by the multiple linear regression model in both cases. Nevertheless, in scenario 1 the variable that accounts for being connected isn't statistically significant. This means that, for this first analysis, being connected to other people in a physical activity promotion program is not a good predictor for the PAL in the end of the program, while the willingness of becoming part of a community is. In the second analysis, a linear mixed model was fitted on the whole data set (eleven weeks).

For scenario 1 there is no significant interaction between being connected and time. For the generalized linear regression being connected presents a significant difference ($p < 0.05$). This is not the case for the linear mixed effects analyses. That means that the PAL of people connected in the end of the program can't be explained by the fact that they are connected, but mainly by the time. For scenario 2 there is a significant difference between the increase in PAL of the two groups, even when a random factor for the (different value of the) start PAL is taken into account. It can thus be inferred from the results that, on average, people that participate in an online community at some time show a larger increase in activity level between the start and end of the program compared to people that will not participate in a community.

5.4.2 Interpretation

These results are a good starting point for understanding why participating in an online community has a positive effect on the physical activity of people. We have observed that in scenario 1 the people that are connected in the network follow the same trend as the control group. In scenario 2, the PAL of the participants that will become part of the community increases over time, while the PAL of the control group slightly decreases. Therefore, the intervention group (community people) seems to have a significant increase on their PAL when compared to the control group due to their *willingness* to become member of a community.

That is not the case for scenario 1, in which people in the intervention group are connected throughout the program. As shown in the results section, when people are connected, they start at a higher PAL, but present the same down trend observed in the intervention group. That is an indicator that being connected does not increase the PAL over time. A different social phenomena not included in our model can be the cause of it. For instance, it could be caused by the overall social contagion i.e. the process of influencing others (sometimes unconsciously) via a network of social relations [23]. The results can also be related to the density of the ego-network of the individual. It has been shown that the density of the links can also deprecate the performance of the individual for this same data set [10]. Further investigation is required to explain the fact that connected people do not have the same increase observed in the group of people willing to partake the program.

The visual representation in Figures 5.3 and 5.4 of the PAL during the period of the intervention shows more than just the slopes of the control and intervention groups. There is a regular pattern of peaks and dips in both groups. Since each participant always starts his/her plan on a Monday, the data is aligned per weekday. Our explanation is that the dips correspond to weekends, when people are less active on average. We can also see that the PAL of the people that are part of the control group does not increase at all for both scenarios, even though they participate in a physical activity promotion program. There is no obvious explanation for this observation, but it seems that the intervention is not effective in increasing the PAL for the participants who do not join the community (during or after the analyzed period of the program). It is possible, however, that the activity levels would have decreased without the intervention. Therefore, a comparison with a control group of people who do not participate in the program at all should reveal whether the intervention is effective in maintaining the PAL.

5.4.3 Conclusions

There is still a long way to go to define what the causes are of people's behaviour changes, especially concerning physical activities. This work collected data of people in an online community participating in a program to increase the amount of exercise they do every day. We were able to investigate the characteristics of the changes from the perspective of the network effects contrasting with participants that were not connected or even a part of a community environment throughout the program. We tried to ensure a fair comparison between the participants of the different groups by doing a very strict selection of the data, and being very specific about the characteristics needed to be selected.

From the selection of the data, two scenarios were built. The first scenario intends to verify if nodes who have connections in an online network provided together with the program causes changes in people's behaviour. We compare a group of 40 people selected with another 40 people who opted to not be in the community part of the program. Both groups present similarities regarding gender, date of plan start and average PAL on the assessment week before the plan starts. The participants in the intervention group have a network degree of at least 1 at the beginning of the program. These analyses showed that people in a network have a higher initial PAL than the people who are not connected at all. Nevertheless, both groups present a similar slope in the change of the PAL over the 12 weeks of program. No significant difference was found that would show an effect for the participants in a network.

In the second scenario, the willingness to participate in an online community in a physical activity promotion program was tested. The participants could choose to participate in an online community by sending and receiving invitations to befriend and share their achievements with others. In this case, we didn't verify if the nodes in the intervention group are indeed connected from day 1 of the program, but just if they opted in to be part of the community. We selected clusters of people that eventually would get connected to others. A criteria was used to select the components and later on the control group, containing people who opted out of the community but have similar characteristics with the nodes from the intervention group. For this scenario there is a significant difference between the two groups, indicating that opting in in the program is a good predictor for people who are going to have more success in their health promotion program.

Earlier work by Groenewegen et al. [10] has shown that people who opted in in a community for sharing their PALs in a health promotion program have a higher PAL than the participants who didn't opt in. These observations are also confirmed by our analysis in both scenarios studied. Also, we were able to conclude that the PAL of people that are willing to join a community shows an increase that is significantly greater compared to the other users. Since we balanced the data sets for possibly confounding factors like gender, time of the year and corporation, it is very likely that the fact that people are willing to become member of the community is the dominant factor that makes a difference for their increase in physical activity level.

To verify if the changes were happening due to the social contagion effect, caused by the exposure to others' PALs over time, we also checked if people who are connected from the start date of the plan present the same tendency, what was not shown in our analysis. We can conclude that the willingness to participate in an online social network for sharing activity data is associated with an increase in physical activity. However, since we also observed that those people are already more physically active at the start, part of the effect might be caused by the fact that people that are willing to participate in such a community are different from others, e.g. more motivated.

This data set has been studied from many perspectives so far. The structure of the network and the dynamics of the creation of the ties is evaluated by Mello Araújo et al. [16]. We also used an existing computational model of social contagion [4] combined with a linear increase on the community group to see whether this new model can explain and predict the changes on PAL. Our combined model showed better results when compared to two other approaches. More studies are needed. For instance, it is important to study the effect of other factors on the physical activity level, such as the community size and structure. That way, research can further uncover phenomena that are at the basis of the beneficial effects of online social networks in health promotion programs.

Acknowledgments: This research is supported by Philips and Technology Foundation STW, Nationaal Initiatief Hersenen en Cognitie NIHC under the partnership program Healthy Lifestyle Solutions.

Bibliography

[1] Charles Abraham and Susan Michie. „A taxonomy of behavior change techniques used in interventions." In: *Health psychology* 27.3 (2008), p. 379 (cit. on p. 71).

[2] AppBrain. May 2016 (cit. on p. 71).

[3] Paul D Bliese and Robert E Ployhart. „Growth modeling using random coefficient models: Model building, testing, and illustrations". In: *Organizational Research Methods* 5.4 (2002), pp. 362–387 (cit. on p. 81).

[4] 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. 87).

[5] Vicki S Conn, Adam R Hafdahl, and David R Mehr. „Interventions to increase physical activity among healthy adults: meta-analysis of outcomes". In: *American journal of public health* 101.4 (2011), pp. 751–758 (cit. on p. 70).

[6] Lorenzo Coviello, Yunkyu Sohn, Adam DI Kramer, et al. „Detecting emotional contagion in massive social networks". In: *PloS one* 9.3 (2014), e90315 (cit. on p. 71).

[7] Rochelle M Eime, Janet A Young, Jack T Harvey, Melanie J Charity, Warren R Payne, et al. „A systematic review of the psychological and social benefits of participation in sport for children and adolescents: informing development of a conceptual model of health through sport". In: *Int J Behav Nutr Phys Act* 10.98 (2013), p. 1 (cit. on p. 70).

[8] GfK. Dec. 2015 (cit. on p. 71).

[9] Frances Griffiths, Antje Lindenmeyer, John Powell, Pam Lowe, and Margaret Thorogood. „Why are health care interventions delivered over the internet? A systematic review of the published literature". In: *Journal of medical Internet research* 8.2 (2006) (cit. on p. 71).

[10] Maartje Groenewegen, Dimo Stoyanov, Dirk Deichmann, and Aart van Halteren. „Connecting with active people matters: the influence of an online community on physical activity behavior". In: *Social Informatics*. Springer, 2012, pp. 96–109 (cit. on pp. 70, 71, 85, 86).

[11] Aric A. Hagberg, Daniel A. Schult, and Pieter J. Swart. „Exploring network structure, dynamics, and function using NetworkX". In: *Proceedings of the 7th Python in Science Conference (SciPy2008)*. Pasadena, CA USA, Aug. 2008, pp. 11–15 (cit. on p. 74).

[12] William L Haskell, I-Min Lee, Russell R Pate, et al. „Physical activity and public health: updated recommendation for adults from the American College of Sports Medicine and the American Heart Association“. In: *Circulation* 116.9 (2007), p. 1081 (cit. on p. 70).

[13] Michel Klein, Adnan Manzoor, Julia Mollee, and Jan Treur. „Effect of changes in the structure of a social network on emotion contagion“. In: *Proceedings of the 2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)-Volume 03*. IEEE Computer Society. 2014, pp. 270–277 (cit. on p. 71).

[14] I-Min Lee, Eric J Shiroma, Felipe Lobelo, et al. „Effect of physical inactivity on major non-communicable diseases worldwide: an analysis of burden of disease and life expectancy“. In: *The lancet* 380.9838 (2012), pp. 219–229 (cit. on p. 70).

[15] Lorna Haughton McNeill, Matthew W Kreuter, and SV Subramanian. „Social environment and physical activity: a review of concepts and evidence“. In: *Social science & medicine* 63.4 (2006), pp. 1011–1022 (cit. on p. 71).

[16] Eric Fernandes de Mello Araújo, Michel Klein, and Aart van Halteren. „Social Connection Dynamics in a Health Promotion Network“. In: *International Workshop on Complex Networks and their Applications*. Springer. 2016, pp. 773–784 (cit. on pp. 78, 87).

[17] Anouk Middelweerd, Julia S Mollee, C Natalie van der Wal, Johannes Brug, and Saskia J te Velde. „Apps to promote physical activity among adults: a review and content analysis“. In: *Int J Behav Nutr Phys Act* 11.1 (2014), p. 97 (cit. on p. 71).

[18] Esko Nuutila and Eljas Soisalon-Soininen. „On finding the strongly connected components in a directed graph“. In: *Information Processing Letters* 49.1 (1994), pp. 9–14 (cit. on p. 74).

[19] PewResearchCenter. Oct. 2015 (cit. on p. 71).

[20] Jose Pinheiro, Douglas Bates, Saikat DebRoy, Deepayan Sarkar, and R Core Team. *nlme: Linear and Nonlinear Mixed Effects Models*. R package version 3.1-128. 2016 (cit. on p. 81).

[21] PocketGamer.biz. May 2016 (cit. on p. 71).

[22] CR Richardson, LR Buis, AW Janney, et al. „An Online Community Improves Adherence in an Internet-Mediated Walking Program. Part 1: Results of a Randomized Controlled Trial“. In: *J Med Internet Res* 12.4 (2010) (cit. on p. 71).

[23] Gerald Schoenewolf. „Emotional contagion: Behavioral induction in individuals and groups.“ In: *Modern Psychoanalysis* (1990) (cit. on p. 85).

[24] Robert Tarjan. „Depth-first search and linear graph algorithms“. In: *SIAM journal on computing* 1.2 (1972), pp. 146–160 (cit. on p. 74).

[25] Ward Van Breda, Jan Treur, and Arlette van Wissen. „Analysis and support of lifestyle via emotions using social media“. In: *Social Informatics*. Springer, 2012, pp. 275–291 (cit. on p. 71).

[26] Rick S Zimmerman and Catherine Connor. „Health promotion in context: the effects of significant others on health behavior change“. In: *Health Education & Behavior* 16.1 (1989), pp. 57–75 (cit. on p. 71).

6

Social Connection Dynamics in a Health Promotion Network[1]

> *"No doubt those who really founded modern science were usually those whose love of truth exceeded their love of power."*
>
> — **C.S. Lewis**
> "The Abolition of Man"[2]

Abstract

The influence of social connections on human behaviour has been demonstrated in many occasions. This paper presents the analysis of the dynamic properties of longitudinal (335 days) community data (n=3,375 participants) from an online health promotion program. The community data is unique as it describes how the network has evolved since its inception and because the information exchanged through the network was predominantly about the achievements of participants in the program and therefore influencing behaviour through social comparison. The analyses show that the largest component of the community network has characteristics of a small world network. The analyses also show that connections are formed according to a strong attachment preference according to the gender, and a weaker homophily for Body Mass Index. The presented analysis can serve as basis for creating novel interventions that influence physical activity behaviour through social connections.

[1] Published as: **de Mello Araújo E.F.**, Klein M., van Halteren A. (2017) *Social Connection Dynamics in a Health Promotion Network.* In: Cherifi H., Gaito S., Quattrociocchi W., Sala A. (eds) Complex Networks & Their Applications V. COMPLEX NETWORKS 2016. Studies in Computational Intelligence, vol 693. Springer, Cham

6.1 Introduction

Social Network Analysis (SNA) is a broad research area, with applications in many different disciplines, incorporating aspects of sociology, social psychology and anthropology [19]. SNA is useful for studying nodes' influences within a network, and how behaviours, opinions or sentiments are spread in social networks [3, 6]. The nodes with an important position can be used to find points of interventions to stop or to enhance the process under study [1, 2, 9, 11, 21].

However, many of the contributions in this field are based on static networks, without taking the time dimension into account. The dynamics of the network can reveal more about how the network evolves over time [5, 22].

In this paper, we investigate the dynamic properties of longitudinal (336 days) community data (n=3,375 participants) from an online health promotion program. This data set presents a network of people that share their physical activities and see others' activity levels. It is a data set specifically focusing on health promotion, in contrast with other research which is mostly using online social networks for general purposes, such as Facebook, Twitter, etc. [8, 15].

To build this data set, the participants wore an activity monitor device that tracks their physical activity level (PAL). They also had access to an online system where they could befriend other participants in order to share and see each others' PAL. The data sample used in this work was collected from 28/04/2010 until 30/03/2011. The analysis of the characteristics of this social network in a health promotion context provides a basis for answering the following questions:

1. How does the largest component of this specific social network develop over time?
2. Does this social network demonstrate the homophily phenomenon (concerning gender and BMI)?
3. Can we use the dynamic analysis of the network to determine influential nodes?

The paper is organized as follows. Section 6.2 discusses the dynamic aspects of social networks, and presents the concepts explored here. Section 6.3 explains the analysis performed, metrics used and the selection process. Section 6.4 shows the results of the analyses. Finally, Section 6.5 concludes the paper with a discussion of the consequences and the possible applications of the findings.

6.2 Dynamical Social Network Analysis

The dynamic aspects of social networks can be analyzed in two ways: (1) looking at the changes *inside* the network (changes in the nodes' attributes as opinions, beliefs, etc.), or (2) looking at the changes of the network itself (the topology of the network, the nodes' degrees, etc.). Dynamical networks are considered here as social networks where the topology changes over time due to new connections or new subjects inside the network.

Static measures of nodes' degrees, centrality, shortest paths, etc. of one fixed snapshot of the data are not sufficient to understand real networks that evolve over time. How new connections are made in or removed from the social network can to some extent be explained by these two phenomenons: homophily and preferential attachment ('more becomes more') [4, 14]. These concepts will be explored further in this work.

The dataset that we use is also used in [10]. In their work, the authors explore the internal states of the nodes and the correlations between the characteristics of the nodes for a shorter period (14 weeks). In [13], the same data set is the basis for a study on the differences between people inside and outside a community, showing how the community aspect plays a role in changing the physical activity level during an intervention. The current work is dedicated to the topological and structural aspects of the network and its connections over time.

6.3 Methods

This section explains the data collection and the data processing. The aim is to provide a clear understanding of how the data was collected, how the subset was selected and how the analysis was done.

6.3.1 Data Set and Data Selection

The data set is the result of an online physical activity promotion program, where the participants wore an activity monitor that tracks their physical activity level (PAL). The devices were synchronized with an online system, which also provided the possibility for them to join a community through connection requests. The participants could also participate in a health promotion program, and those who decided to do that were tagged in our data set with a 'start plan date'. The data used in this work spans 336 days, from 28/04/2010 until 30/03/2011.

As the decision to join the community was optional for the participants, around 10% of them decided to join the social network to exchange their information about the PAL tracked by their devices. In total there are almost 5,000 nodes that opted to join the online community at some moment during the experiment.

Due to changes in the system, some cleaning was necessary to keep the data set reliable for the analyses performed. From the originally 5,000 nodes and around 28,000 edges, we filtered nodes and edges according to the following characteristics:

- a) Nodes without 'start plan date' were removed;
- b) Nodes were included according to the date of their started plan;
- c) Nodes that dropped out the experiment (tagged with a value for 'dropout date') were taken off at the day when they quit the network and the program;

- d) Nodes without a value for BMI (Body Mass Index), gender and nodes in which all information was missing were taken out;

- e) Edges without 'start date' value were removed;

- f) Edges connected to excluded nodes were removed.

From a total of 28,418 edges, 3,802 edges didn't have information about the date of connection, because some requests for connections in the network were not approved from the receiving peer. As these edges are represented in two directions, 1,901 unique edges were discarded. From the 24,616 edges left, 12,047 are duplicated edges, i.e., node A connects to B, but the edge (B,A) already exists. As all connections are bidirectional, this is redundant data. So we have, in the end, a total of 12,569 edges representing connections that were formed during the experiment.

The data set originally contained 4,989 nodes. Of those, 1,614 nodes were not eligible because they do not have values for all the attributes needed for the analysis (i.e., gender, BMI and start plan date). The selected data set has 3,375 nodes left.

The nodes are only included in the network in the period between the start plan and the drop out date (for those that dropped out). After the node leaves the network, all its connections are deleted also. The impacts of the cleaning process are irrelevant, because the nodes and edges removed didn't participated in the program as demanded.

6.3.2 Social Network Analysis

The network measures that are calculated are [19]: (1) degree distribution; (2) average degree; (3) closeness centrality; (4) eigenvector centrality; (5) betweenness centrality; and (6) average shortest path . These aspects were analyzed for each day of the experiment.

Formula 6.1 shows the calculation for the **combined centrality**, a combination of the betweenness and closeness values:

$$Comb_C(i) = \frac{C_C(i) + C_B(i)}{2} \tag{6.1}$$

$C_C(i)$ and $C_B(i)$ are the closeness and betweenness centralities, respectively. This formula doesn't consider the balance between the two centrality measurements, and might be improved for future analysis. For our analysis it is correct to say that the Closeness centrality will influence more than the betweenness for having higher values in general.

Homophily is the tendency of nodes to create strong connections with others that are alike, have the same opinions, or share similar characteristics [14]. The homophily principle can be studied in two ways: the *social* homophily and the *value* homophily

[12, 20]. In this work, the *social* aspects (gender and BMI) are studied in depth, while the *value* aspects are left out of the analysis.

The homophily according to gender was calculated using the gender of the nodes' edges. These edges were categorized as follows:

Edge MM (EMM): a connection between two male nodes;

Edge MF (EMF): a connection between a male node and a female node;

Edge FF (EFF): a connection between two female nodes.

As the three categories are disjoint, the total number of edges equals to $EMM + EMF + EFF$. The homophily for female gender and male gender are given by equations 6.2 and 6.3, respectively.

$$Homophily_F = \frac{EFF}{EFF + EMF} \tag{6.2}$$

$$Homophily_M = \frac{EMM}{EMM + EMF} \tag{6.3}$$

To calculate homophily for the BMI, we considered nodes with BMI in the same range as equals. Two different thresholds were used: 5.0 and 6.5, which are the respective ranges for the group of Normal and Overweight BMI in the categorization according to Organization et al. [18].

The ratio between the nodes' edges with a small difference in BMI and the total number of edges yields the percentage that follows the homophily principle for the BMI. The equations follow the same principles of equations 6.2 and 6.3.

The **ego-network density** for the nodes is used to find important nodes. The density is calculated in two steps. First, the ego-network of all the nodes (including the observed node) is created using 1-step neighborhood. After this step, the density of the ego-network was calculated as: Ego-density $= \frac{|E|}{n(n-1)}$, where $|E|$ is the number of edges in this subgraph, and n is the number nodes.

6.4 Results

This section presents the results obtained from the social network analysis. The section is organized according to the questions from Section 6.1:

1. How does the largest component of this social network develop over time?

2. Does this social network demonstrate the homophily phenomenon (for gender and BMI) ?

3. Can we use the dynamic analysis of the network to determine influential nodes?

6.4.1 Nodes, edges and degree distribution

On day 98 of the experiment the number of nodes in the graph is stabilized at 2,996. The number of nodes in the largest component increases until the end of the experiment, due to new connections established among the nodes.

For the edges there is also a point of stabilization in the new connections around day 100. From that day onward there is a very small increase in the number of connections (around 8.2%). Most of the edges are in the largest component, as it is expected in a network that follows the Small World Network model.

The graph follows a Power-law distribution for the degrees of the nodes for all time steps. Figure 6.1 shows the degree distribution for the days 1, 100 and 336 in a log-log scale (for illustration[3]). The lower graphics show the coefficients for the linear regression of the correlation between the degree of the nodes and the number of nodes with certain degree.

As shown in the lower graphic, the p value is always significant for our data set, and the R-squared is close to 1, showing that the model explains very well the data, mainly after day 100.

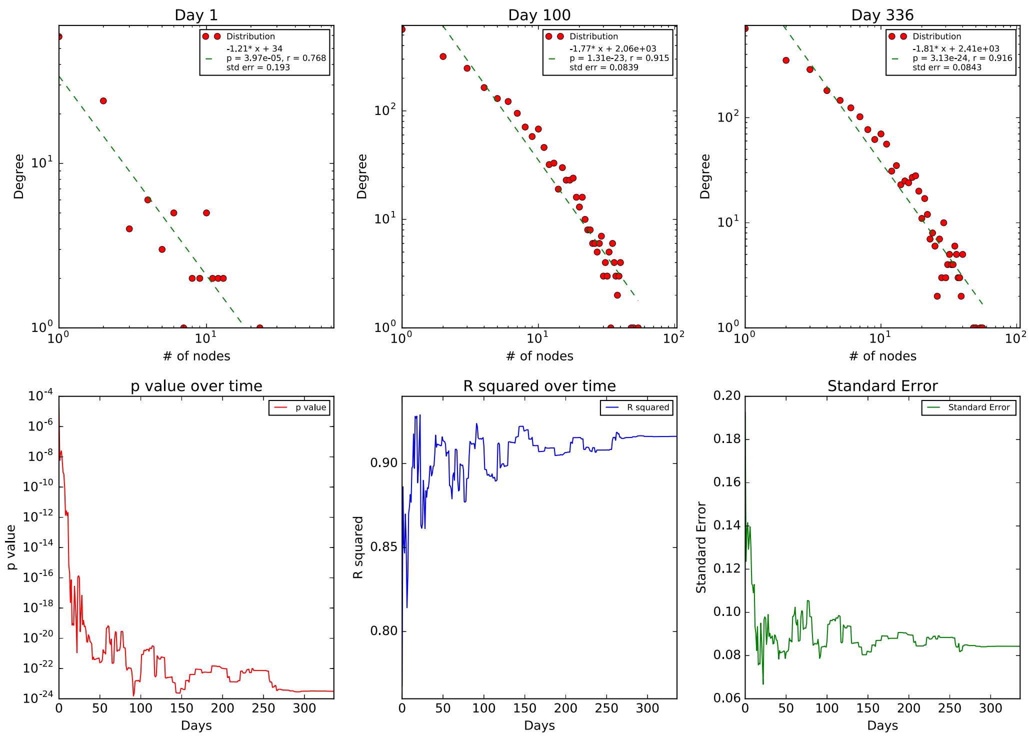


Fig. 6.1: Degree distribution in days 1, 100 and 336 (top) and p value for slope, R squared and standard error (bottom)

The 'more becomes more' principle is the assumption that nodes with higher degree have a higher chance of receiving more connections over time [16]. Figure 6.2 shows how the degrees of the nodes with the fewest connections (the 'poorest', right) and nodes with the

[3] The other days and other animations can be seen at `http://www.cs.vu.nl/~efo600/cn2016/`

most connections ('richest', left) evolve over time. More investigation is needed to claim that the preferential attachment is observed here, but the information about the rich and poor nodes suggests that it could be present in our data set.

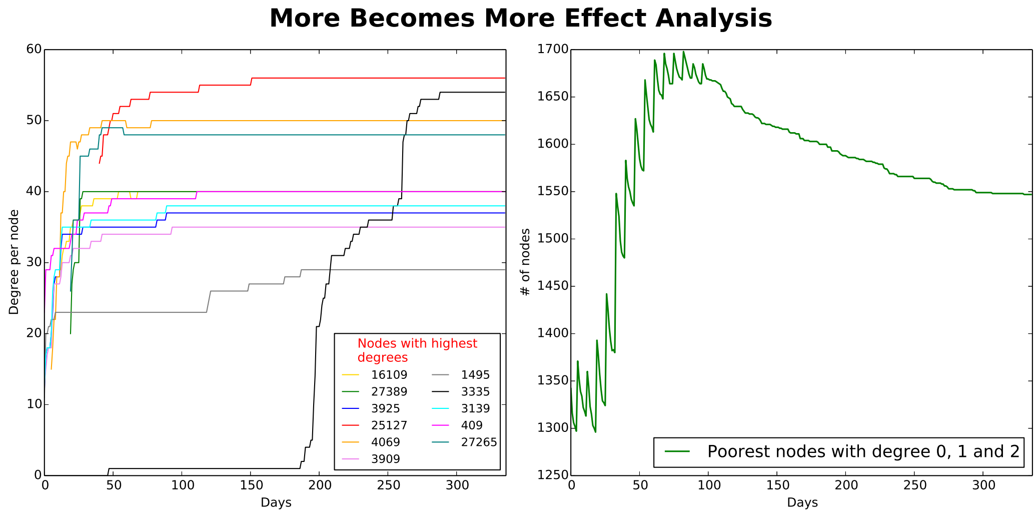


Fig. 6.2: Degree of the richest nodes (left) and number of poor nodes, with $degree \leq 2$ (right)

6.4.2 Largest component and other components

The 'largest component' is the biggest connected component among all components of any graph. Figure 6.3 shows the percentage of the nodes of the graph that are part of the largest component for all time steps in two different scenarios. In the first scenario, all nodes are included in the graph. As can be observed, the average number of nodes in the largest component is 65% after day 296 for the entire graph. The increase in the percentage follows the inclusion of new edges after time 100 (when the number of nodes is stable).

As there are many nodes with degree 0 (isolated nodes), for the second scenario, the nodes with degree 0 were excluded from the graph. In this scenario the percentage of nodes in the largest component goes up to 80%.

Figure 6.4 shows the evolution of the connected components over time. The upper graphic shows the number of components over time. As edges are inserted, many components are joined, explaining the decrease from around 1,200 connected components to almost 600 in the end. The red line shows the number of components bigger than 1, i.e., non isolated nodes. This number goes from 39 on day 1 up to 164 in the last day of the experiment. The number of isolated nodes goes from 1,193 in day 1 down to 492, what explains the high number of components, even after the largest component gathered more than 60% of the nodes of the network.

The correlation between the size of the components and the number of components with a specific size (frequency of occurrence) is shown in the middle part of Figure 6.4 in three graphics, for days 1, 165 and 335. The correlation is significant for all time steps. The three lower graphics show the p value, R squared and standard error for the regression done in all the time steps of the data set. It can be seen that the fit parameter goes from approximately 65% to less than 40% in the end of the experiment. This can be explained by the changes in the largest component, and the joining of previously separated components.

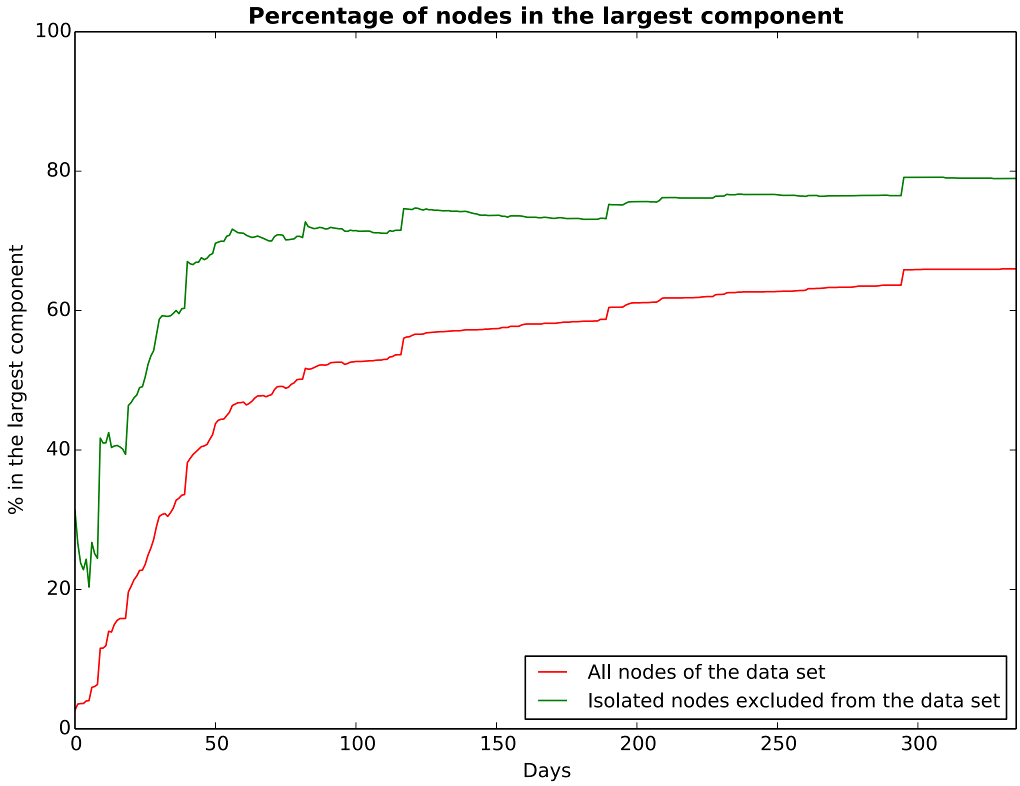


Fig. 6.3: Percentage of the nodes in the largest component. All nodes (lower red line) and nodes with degree larger than 1 (higher green line)

6.4.3 Centrality measurements

As the largest component has most of the nodes and edges, it is also interesting to explore the centrality measurements for this component. The following metrics were analyzed: (1) betweenness centrality, (2) closeness centrality, (3) eigenvector centrality, (4) average shortest path.

The **betweenness centrality** indicates how important a node is for the transfer of information or any kind of spreadable element inside a network. Nodes with higher betweenness have more shortest paths passing through themselves, and therefore can enhance their role in the network. The **closeness centrality** is the proximity of a node to the rest of the network, and it is calculated by the inverse of the sum of the shortest distances between each node and all other nodes in the network. The **eigenvector centrality** is calculated based on the centrality of its neighbors.

The average centrality for all the nodes (betweenness, closeness and eigenvector) is shown in Figure 6.5. The first three graphics on the left show all time steps, while the first three graphics on the right provide a zoomed-in version between day 50 and 336.

The lower graphic in Figure 6.5 shows the average shortest path. The average shortest path for our data set stabilizes around 6.5, a low value as suggested by the theory in [17].

The combined centrality is useful in finding important nodes that combine a good betweenness centrality and closeness centrality. Figure 6.6 shows the combined centrality for all the nodes with degree higher than 1.

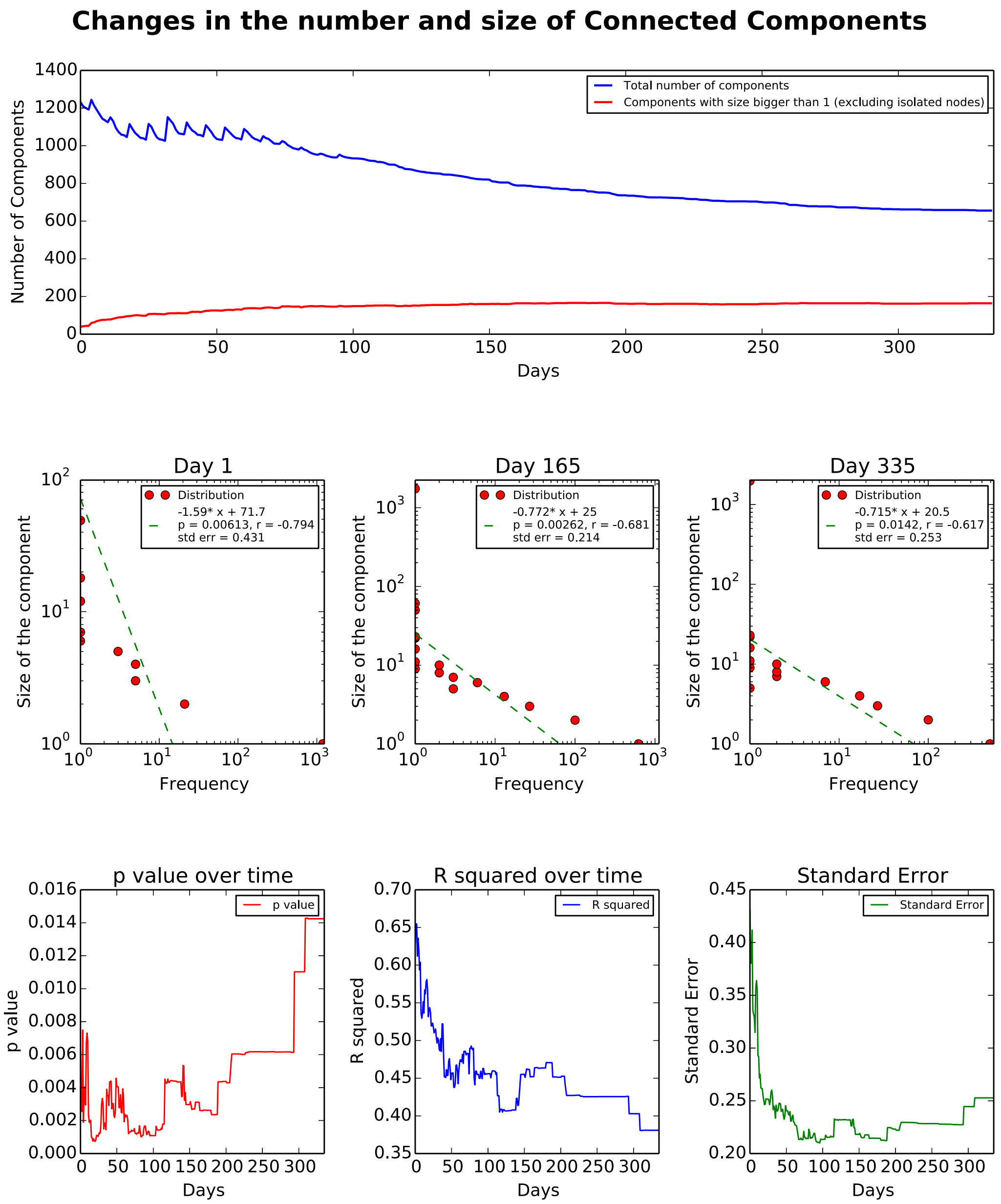


Fig. 6.4: Components analysis. Number of components in the graph over the time (upper), the correlation between the size of the component and the frequency of the size (days 1, 165 and 335) (middle) and the parameters from the linear regression for all time steps (lower)

It is possible to highlight the list of nodes with higher centrality (the most potentially influential nodes in the network). Figure 6.6 shows the most central nodes measures of betweenness, closeness and the combined centrality. As shown in Figure 6.6, nodes 68593 and 3335 are very important for this data set, as they present the highest values for these measurements.

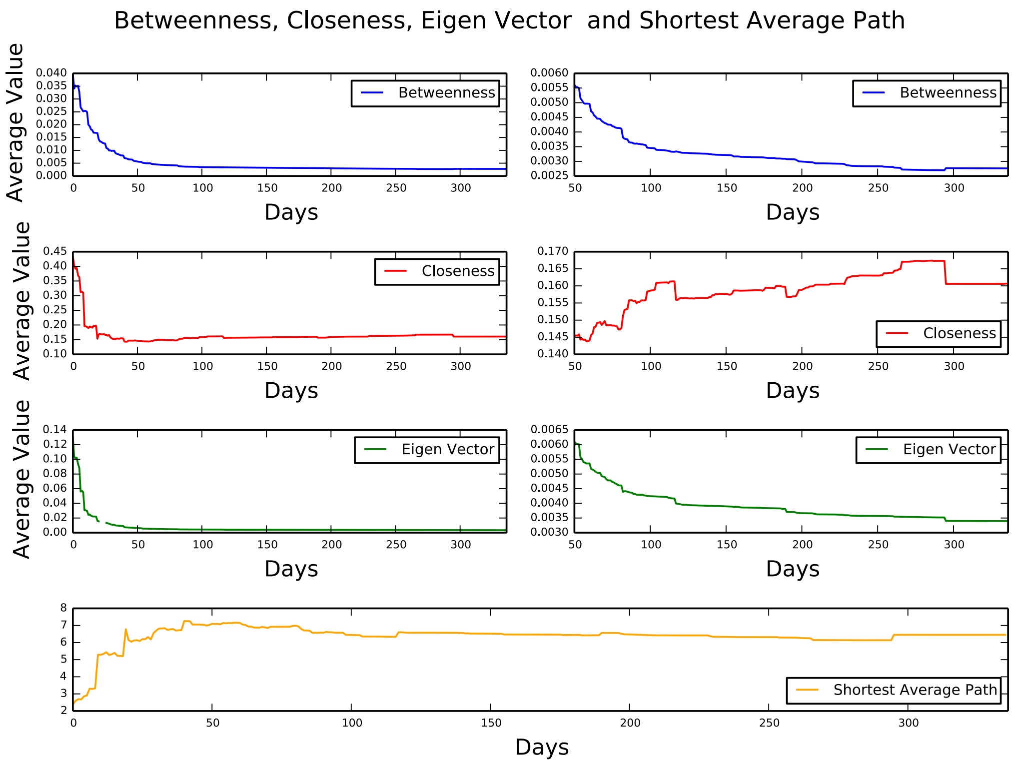


Fig. 6.5: Mean of all centrality measures for all nodes at each time step (six graphics on top) and average shortest path (bottom)

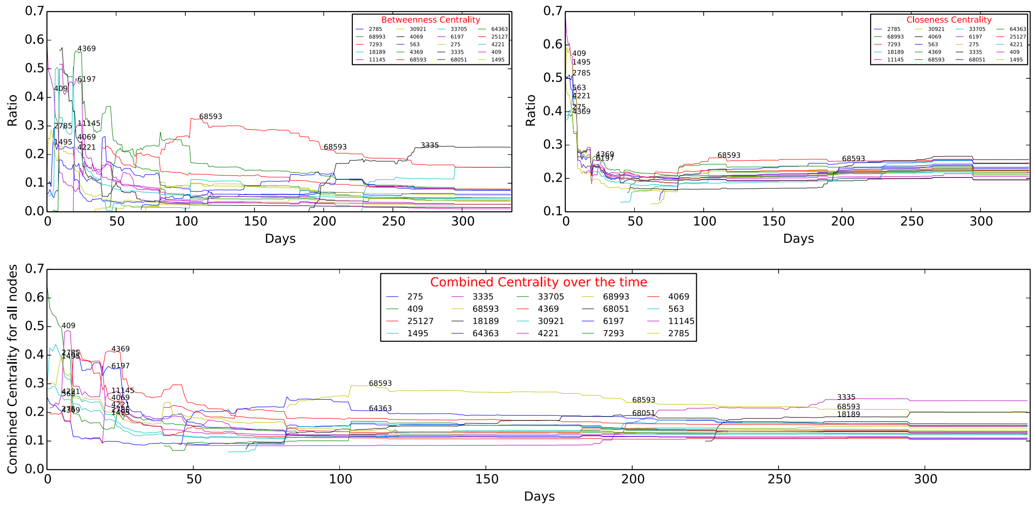


Fig. 6.6: Centrality for most central nodes. Betweenness (upper left) and closeness (upper right) for the 20 nodes with the highest combined centrality. Lower graphic shows the combined centrality measures

6.4.4 Homophily

To investigate homophily according to gender and BMI, the edges were evaluated to determine whether the nodes they connect belong to the same category. The results for the gender analysis follow the equations 6.2 and 6.3. The data set has 51.4% of the nodes of gender male, and 48.6% female. Regarding the BMI of the population, 0.8% are underweight, 33.6% are normal, 34.8% are overweight and 30.8% are obese [18].

Figure 6.7 (left) shows the homophily according to the BMI of the nodes. Two ranges were tested for the nodes: 5.0 and 6.5. For the range of 5.0, the ratio of edges with nodes within the same range is around 50% after day 100, while for range 6.5 this value is increased to around 59%. For both ranges, more than half of the connections are within nodes with close BMI.

Figure 6.7 (right) shows the homophily according to gender. Three calculations were made: (a) edges connecting male-male nodes, (b) edges connecting female-female nodes and (c) edges connecting same gender nodes (male-male plus female-female edges). In this data set, the homophily for women holds for between 50% and 60% of the edges. That means that women connect around half of the time with other women.

For men we observe that more than 60% of the connections are to nodes of the other gender, female. The fact that women have more connections among themselves is know by other studies on gender and relationships [7]. However, the figure also shows that homophily is not present for the male-male connection (i.e., new connections of men are more often with women). When taking both categories together, there is homophily on gender: above 60% of the edges connecting people of the same gender.

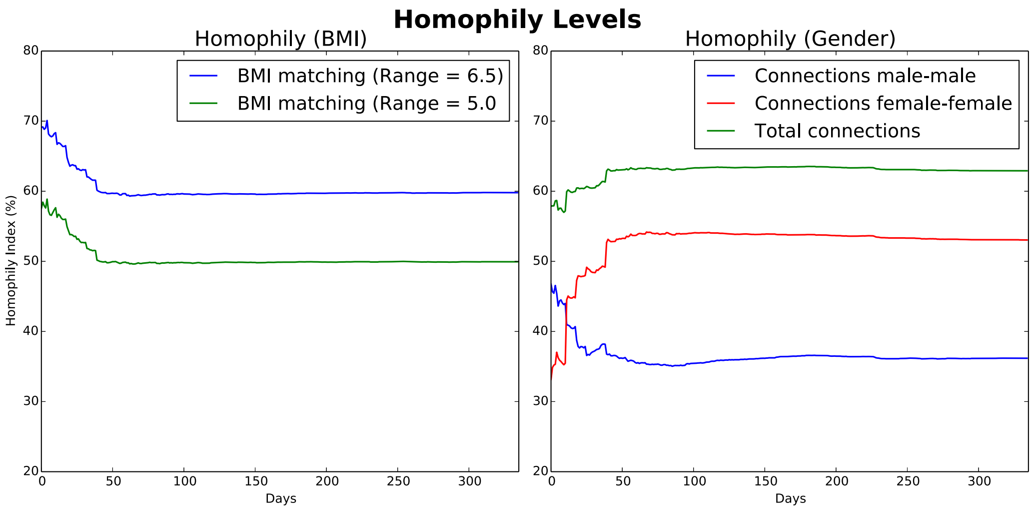
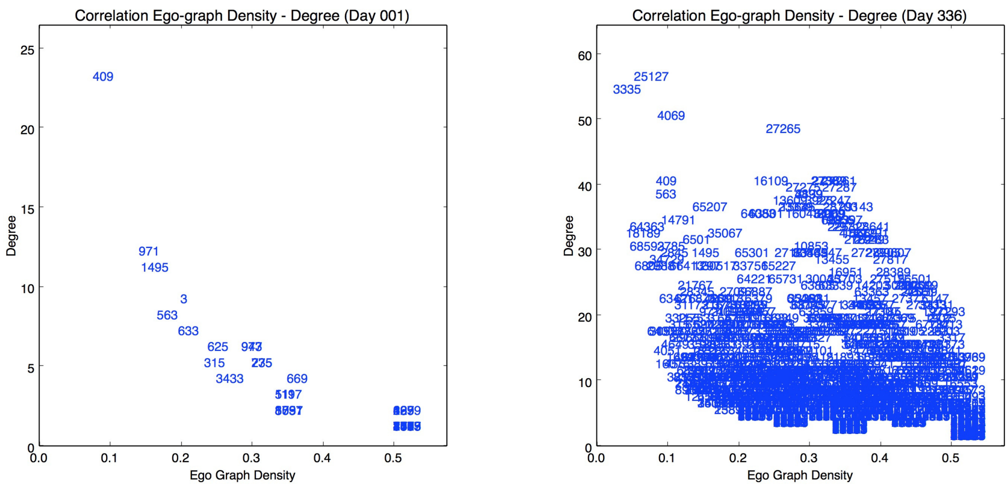


Fig. 6.7: Homophily according to the BMI (left) and gender (right) of the nodes

6.4.5 Identifying influential participants

The dynamic nature of the network is clearly visible from the analyses presented in the previous sections. In previous work we have shown that the more successful participants in the program are, the smaller is the density of their ego-network [10]. This section demonstrates that the set of most influential participants dynamically changes over time. We identify influential participants by comparing properties such as betweenness centrality, closeness centrality, eigenvector centrality, ego-network density and average shortest path.

Figure 6.8 shows the relation between the node degree of each participant and their ego-network density for the first and last day of the experiment. In this graph we're interested in nodes that have a low density yet a growing degree, as they can be bridges on spreading of



(a) First Day

(b) Last Day

Fig. 6.8: The dynamic relation between ego network density and nodes' degrees

emotions, for instance. These are the participants in the top-left quadrant of the graph. Despite the fact that this is just a snapshot, the changes over time provided by the combination of each day's relation can give a better picture of what is happening inside a network.

We plotted graphs for all days of the dynamic network which revealed that the set of nodes that emerges in the top-left quadrant are frequently changing. During the experiment, four leader nodes were in evidence considering the ratio between the degree of the nodes and the ego-network density. Node 409 (from day 1 to 12), node 3069 (from day 13 to 40), node 25127 (from day 41 to 254) and node 3335 (from day 255 to 336).

6.5 Conclusions

In this paper, we have investigated the *dynamic* properties of a longitudinal study of a networked community participating in an online health promotion program. It turned out that studying the dynamics gives additional insights in characteristics of the network. For example, it is shown that the number of components in the network is decreasing while the size of the components is increasing at the same time. The components themselves follow a Power-law distribution at all time steps: there are a few components with many nodes, and a lot of components with only a few nodes. It is also shown that characteristics like betweenness, closeness, eigen vector and average shortest path at the start of the network are very different from the values after 356 days; however it turned out that already after 50 to 100 days most measurements were relatively stable.

The dynamical data set also allowed us to evaluate whether two well-known phenomena of evolving networks are present: homophily and preferential attachment. Our analysis showed that homophily takes place on the aspect BMI and gender; the latter especially for female-female connections. Apart from the possible preferential attachment, more investigation is needed to affirm that it is present in this data set.

Finally, the combination of degree measurements and the density of the ego-network was presented, and we aim to use it to identify people that are potentially influential in their network in further work. Interestingly, the set of people who are influential according to this metric changes during the evolution of the network, even after the moment that the nodes of network have stabilized. This suggest that continuous monitoring the evolution of a network is important to identify such people.

We believe our discoveries and methods can form the basis for automated (health) interventions that exploit the social network for changing behaviours of individuals, and possibly lead us to future discoveries about leadership, spreading of emotions or any other application related to the network's topology and dynamics.

Bibliography

[1] Daron Acemoglu and Asuman Ozdaglar. „Opinion dynamics and learning in social networks“. In: *Dynamic Games and Applications* 1.1 (2011), pp. 3–49 (cit. on p. 92).

[2] Daron Acemoglu, Asuman Ozdaglar, and Ali ParandehGheibi. „Spread of (mis) information in social networks“. In: *Games and Economic Behavior* 70 (2010), pp. 194–227 (cit. on p. 92).

[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 p. 92).

[4] Albert-Laszlo Barabási and Reka Albert. „Emergence of scaling in random networks“. In: *Science* 286.October (1999), pp. 509–512 (cit. on p. 93).

[5] Romy Blankendaal, Sarah Parinussa, and Jan Treur. „A temporal-causal modelling approach to integrated contagion and network change in social networks“. In: *Proceedings of the 22nd European Conference on Artificial Intelligence, ECAI'16*. 2016 (cit. on p. 92).

[6] Nicholas A Christakis and James H Fowler. „The spread of obesity in a large social network over 32 years“. In: *New England journal of medicine* 357.4 (2007), pp. 370–379 (cit. on p. 92).

[7] Steve Duck and Paul H. Wright. „Reexamining gender differences in same-gender friendships: A close look at two kinds of data“. In: *Sex Roles* 28.11-12 (1993), pp. 709–727 (cit. on p. 101).

[8] Nicole B. Ellison, Charles Steinfield, and Cliff Lampe. „The benefits of facebook "friends:" Social capital and college students' use of online social network sites“. In: *Journal of Computer-Mediated Communication* 12.4 (2007), pp. 1143–1168 (cit. on p. 92).

[9] S Eubank, H Guclu, V S Kumar, et al. „Modelling disease outbreaks in realistic urban social networks“. In: *Nature* 429.6988 (2004), pp. 180–184 (cit. on p. 92).

[10] Maartje Groenewegen, Dimo Stoyanov, Dirk Deichmann, and Aart van Halteren. „Connecting with active people matters: the influence of an online community on physical activity behavior“. In: *Social Informatics*. Springer, 2012, pp. 96–109 (cit. on pp. 93, 101).

[11] David Kempe, Jon Kleinberg, and Éva Tardos. „Influential Nodes in a Diffusion Model for Social Networks“. In: *Automata, Languages and Programming* 3580 (2005), pp. 1127–1138 (cit. on p. 92).

[12] Paul F. Lazarsfeld and Robert K. Merton. „Friendship as a Social Process: A Substantive and Methodological analysis“. In: *Freedom and Control in Modern Society* 18 (1954), pp. 18–66 (cit. on p. 95).

[13] Adnan Manzoor, Julienka S. Mollee, Eric F.M. Araújo, Aart T. van Halteren, and Michel C. A. Klein. „Online sharing of physical activity: does it accelerate the impact of a health promotion program?“ In: *Socialcom 2016*. 2016 (cit. on p. 93).

[14] Miller McPherson, Lynn Smith-Lovin, and James M Cook. „Birds of a feather: Homophily in social networks“. In: *Annual review of sociology* 27.1 (2001), pp. 415–444 (cit. on pp. 93, 94).

[15] Alan Mislove, Massimiliano Marcon, Krishna P. Gummadi, Peter Druschel, and Bobby Bhattacharjee. „Measurement and analysis of online social networks“. In: *Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement - IMC '07* (2007), pp. 29–42 (cit. on p. 92).

[16] M E J Newman. „The structure and function of complex networks“. In: *Siam Review* 45.2 (2003), pp. 167–256. arXiv: `0303516 [arXiv:cond-mat]` (cit. on p. 96).

[17] M E J Newman and D J Watts. „Scaling and percolation in the small-world network model“. In: *Physical review. E, Statistical physics, plasmas, fluids, and related interdisciplinary topics* 60.6 Pt B (1999), pp. 7332–7342 (cit. on p. 98).

[18] World Health Organization et al. „Global database on body mass index: an interactive surveillance tool for monitoring nutrition transition“. In: *World Health Organization: Geneva* (2012) (cit. on pp. 95, 101).

[19] John Scott. *Social Network Analysis*. Sage, 2012 (cit. on pp. 92, 94).

[20] Maksim Tsvetovat and Alexander Kouznetsov. *Social Network Analysis for Startups: Finding connections on the social web*. " O'Reilly Media, Inc.", 2011 (cit. on p. 95).

[21] Thomas W Valente. *Network models of the diffusion of innovations*. Vol. 2. Hampton Press (NJ), 1995, pp. 163–164 (cit. on p. 92).

[22] D J Watts and S H Strogatz. „Collective dynamics of 'small-world' networks.“ In: *Nature* 393.6684 (1998), pp. 440–2 (cit. on p. 92).

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)*[1]

7

Explaining changes in physical activity through a computational model of social contagion[1]

"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"[2]

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.

[1]Published as: Mollee J.S., **Araújo E.F.M.**, Manzoor A., van Halteren A.T., Klein M.C.A. (2017) *Explaining Changes in Physical Activity Through a Computational Model of Social Contagion*. In: Gonçalves B., Menezes R., Sinatra R., Zlatic V. (eds) Complex Networks VIII. CompleNet 2017. Springer Proceedings in Complexity. Springer, Cham

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* (ϵ_A), the *openness* of the receiver (δ_B) and the strength of the channel between them (α_{AB}). Together, these factors determine the *connection weight* ω_{AB}. Thus, the impact $\textbf{impact}_{AB}(t)$ of the state of person A on the state of person B is:

$$\textbf{impact}_{AB}(t) = \omega_{AB} q_A \tag{7.1}$$

The aggregated impact $\textbf{aggimpact}_B(t)$ 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 $\textbf{aggimpact}_B(t)$ is calculated as a weighted average of all the impacts of the different connections of a person:

$$\textbf{aggimpact}_B(t) = \sum_{A_i \neq B} w_{AB} q_{A_i}(t) \tag{7.2}$$

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

$$contagion\_effect(t) = \eta_A[\textbf{aggimpact}_B(t) - q_B(t)] \tag{7.3}$$

$$q_B(t + \Delta t) = q_B(t) + contagion\_effect(t)\Delta t \tag{7.4}$$

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$^{\text{th}}$ 2010, all data before that date was disregarded. Community data was available until August 6$^{\text{th}}$ 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$^{\text{th}}$ 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$^{\text{th}}$ 2010 and July 28$^{\text{th}}$ 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.

$$linear\_effect(t) = \frac{0.0005821}{target\_pal(t)} \tag{7.5}$$

$$goal\_achieved(t + \Delta t) = goal\_achieved(t) + linear\_effect(t) \quad (7.6)$$

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 *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.

$$goal\_achieved(t + \Delta t) = goal\_achieved(t) + contagion\_effect(t) + linear\_effect(t) \quad (7.7)$$

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 28th 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.

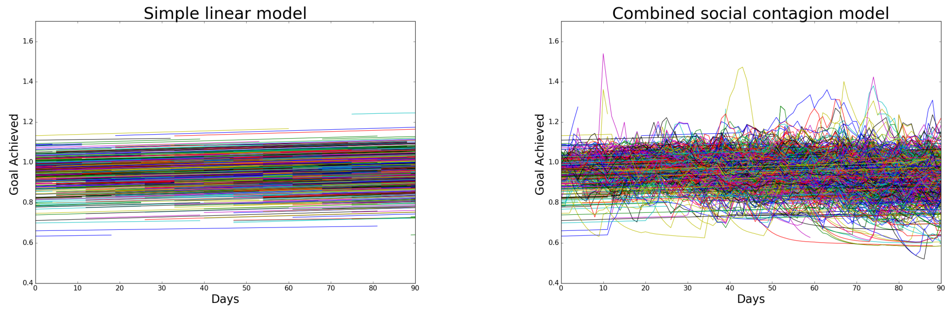


Fig. 7.1: Predictions of the simple linear model (left) and the combined model (right).

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

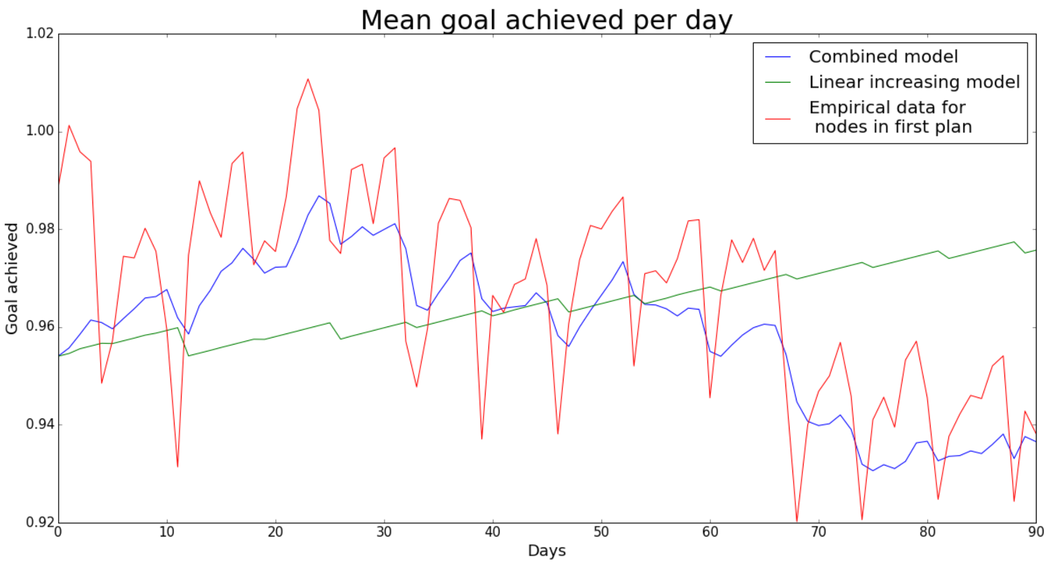


Fig. 7.2: Average predictions of the two models (green: linear, blue: combined), and the empirical data (red).

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

Tab. 7.1: Model evaluations

	Absolute Error		Kendall's correlation test	
	Mean	St. Dev.	Kendall's τ	Kendall's p
Linear Model	0.02212	0.01378	-0.46227	<0.001
Combined Model	0.01321	0.00855	0.53895	<0.001

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.

Bibliography

[1] Eric F. M. Araújo, Michel C. A. Klein, and Aart T. van Halteren. „Social Connection Dynamics in a Health Promotion Network". In: *Complex Networks 2016 - The 5th International Workshop on Complex Networks and their Applications*. 2016 (cit. on p. 112).

[2] Eric F. M. Araújo and Jan Treur. „Analysis and Refinement of a Temporal-Causal Network Model for Absorption of Emotions". In: *Computational Collective Intelligence: 8th International Conference, ICCCI 2016, Halkidiki, Greece, September 28-30, 2016. Proceedings, Part I*. Ed. by Ngoc-Thanh Nguyen, Lazaros Iliadis, Yannis Manolopoulos, and Bogdan Trawiński. Springer, 2016, pp. 27–39 (cit. on p. 111).

[3] Judit Bort-Roig, Nicholas D Gilson, Anna Puig-Ribera, Ruth S Contreras, and Stewart G Trost. „Measuring and influencing physical activity with smartphone technology: a systematic review". In: *Sports medicine* 44.5 (2014), pp. 671–686 (cit. on p. 111).

[4] Tibor Bosse, Rob Duell, Zulfiqar A Memon, Jan Treur, and C Natalie van der Wal. „Agent-based modeling of emotion contagion in groups". In: *Cognitive Computation* 7.1 (2015), pp. 111–136 (cit. on p. 111).

[5] 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 pp. 110, 111).

[6] Vicki S Conn, Adam R Hafdahl, and David R Mehr. „Interventions to increase physical activity among healthy adults: meta-analysis of outcomes". In: *American journal of public health* 101.4 (2011), pp. 751–758 (cit. on p. 110).

[7] Rochelle M Eime, Janet A Young, Jack T Harvey, Melanie J Charity, Warren R Payne, et al. „A systematic review of the psychological and social benefits of participation in sport for children and adolescents: informing development of a conceptual model of health through sport". In: *Int J Behav Nutr Phys Act* 10.98 (2013), p. 1 (cit. on p. 110).

[8] Carol Ewing Garber, Bryan Blissmer, Michael R Deschenes, et al. „American College of Sports Medicine position stand. Quantity and quality of exercise for developing and maintaining cardiorespiratory, musculoskeletal, and neuromotor fitness in apparently healthy adults: guidance for prescribing exercise." In: *Medicine and Science in Sports and Exercise* 43.7 (2011), pp. 1334–1359 (cit. on p. 110).

[9] Philippe J. Giabbanelli, Azadeh Alimadad, Vahid Dabbaghian, and Diane T. Finegood. „Modeling the influence of social networks and environment on energy balance and obesity“. In: *Journal of Computational Science* 3.1–2 (2012), pp. 17–27 (cit. on p. 117).

[10] Maartje Groenewegen, Dimo Stoyanov, Dirk Deichmann, and Aart van Halteren. „Connecting with active people matters: the influence of an online community on physical activity behavior“. In: *Social Informatics*. Springer, 2012, pp. 96–109 (cit. on p. 110).

[11] Adnan Manzoor, Julienka S. Mollee, Eric F.M. Araújo, Aart T. van Halteren, and Michel C. A. Klein. „Online sharing of physical activity: does it accelerate the impact of a health promotion program?“ In: *Socialcom 2016*. 2016 (cit. on pp. 110, 113).

[12] M D Mifflin, S T St Jeor, L A Hill, et al. „A new predictive equation for resting energy expenditure in healthy individuals“. In: *The American Journal of Clinical Nutrition* 51.2 (1990), pp. 241–7. eprint: `http://ajcn.nutrition.org/content/51/2/241.full.pdf+html` (cit. on p. 113).

[13] Russell R Pate, Michael Pratt, Steven N Blair, et al. „Physical activity and public health: a recommendation from the Centers for Disease Control and Prevention and the American College of Sports Medicine“. In: *Jama* 273.5 (1995), pp. 402–407 (cit. on p. 110).

[14] Mitesh S Patel, David A Asch, and Kevin G Volpp. „Wearable devices as facilitators, not drivers, of health behavior change“. In: *Jama* 313.5 (2015), pp. 459–460 (cit. on p. 110).

[15] James F Sallis and Neville Owen. *Physical activity and behavioral medicine*. Vol. 3. SAGE publications, 1998 (cit. on p. 111).

[16] Cosma Rohilla Shalizi and Andrew C Thomas. „Homophily and contagion are generically confounded in observational social network studies“. In: *Sociological methods & research* 40.2 (2011), pp. 211–239 (cit. on p. 111).

[17] Jerry Ed Suls and Thomas Ashby Ed Wills. *Social comparison: Contemporary theory and research.* Lawrence Erlbaum Associates, Inc, 1991 (cit. on p. 110).

[18] Rena R Wing and Robert W Jeffery. „Benefits of recruiting participants with friends and increasing social support for weight loss and maintenance.“ In: *Journal of consulting and clinical psychology* 67.1 (1999), p. 132 (cit. on p. 110).

[19] World Health Organization. „Global recommendations on physical activity for health“. In: (2010) (cit. on p. 110).

[20] Rick S Zimmerman and Catherine Connor. „Health promotion in context: the effects of significant others on health behavior change“. In: *Health Education & Behavior* 16.1 (1989), pp. 57–75 (cit. on p. 110).

8 Using Simulations for Exploring Interventions in Social Networks: Modeling Physical Activity Behaviour in Dutch School Classes[1]

> *"Some day you will be old enough to start reading fairy tales again."*
>
> — **C.S. Lewis**
> "The Lion, the Witch and the Wardrobe"[2]

Abstract

The reduction of childhood obesity through the promotion of a healthy lifestyle is one of the most important public health challenges at the moment. It is known that the unhealthy habits of children can cause unavoidable side effects in their early stage of life, including both physical and mental consequences. This work considers that the physical activity level of children is a behaviour that can be spread throughout the social relations of children in their daily life at school. Therefore, the aim of this work is to define what the best strategy is to find 'targets' (i.e., influential children that can initiate behavioural change) for physical activity (PA) interventions that would affect the average PA of a population of Dutch school classes. We tuned a model based on the influence of the children's peers in their social network, based on the data set from the *MyMovez* project – Phase I. Five intervention strategies were implemented, and their efficacy was compared. Once the targets were chosen, an increase of 17% was applied to their initial PA. Then, the diffusion model was run to verify the improvement on the PA of the whole network after one year. We discuss implications of the simulation results on which strategies may be used to make informed choices about the setup of social network interventions and future model improvements. Our results show that targeting more vulnerable children (i.e. in a worse environment) and applying a network optimization algorithm are the best solutions for this data set indicating that future interventions should aim for these two strategies.

[1]Published as **Araújo, E.**, Simoski, B., Woudenberg, T., Bevelander, K., Smit, C., Buijs, L., Klein, M. and Buijzen, M. *Using Simulations for Exploring Interventions in Social Networks – Modeling Physical Activity Behaviour in Dutch School Classes.* In Proceedings of 8th International Conference on Simulation and Modeling Methodologies, Technologies and Applications (SIMULTECH 2018), pages 414-425. ISBN: 978-989-758-323-0 «**nominated to best student paper award**»

8.1 Introduction

One of the most important public health challenges is the prevention and reduction of childhood obesity. It has worldwide priority because the prevalence of childhood overweight and obesity is still rising. Childhood obesity has persisting effects on adult adiposity and can lead to diseases such as diabetes and cardiovascular diseases [31]. Evidence is accumulating which shows that the social environment is an important factor underlying the development of inappropriate weight gain due to its powerful impact on energy-balance related behaviours [7]. Youth are especially susceptible to environmental influences and are surrounded by influential individuals (i.e., role models such as family and peers) who support and/or undermine their health behaviours. For example, studies have shown that individual peers as well as peer groups shape a youth's consumption behaviour and physical activity [8, 9, 20]. Numerous studies have shown that the health of individuals is connected to each other and that social networks influence people's well-being [13, 24].

Social network interventions aim to use influential individuals to correct unhealthy behaviours within social networks by letting them promote specific health behaviours [28]. It is suggested that when influential individuals stimulate and spread the targeted behaviour successfully, the behaviour will turn into a group norm supporting long-term behaviour change. The term network interventions "describes the process of using social network data to accelerate behaviour change or improve organizational performance" [26]. Social network interventions have been successful in reducing behaviours such as smoking and unsafe sex [4, 15, 27]. To date, there is great necessity to target the physical activity of youth because they are even less active than previous generations and the majority of adolescents do not meet daily guidelines of being active for at least 60 minutes [12]. Hence, social network interventions are now not only dedicated to reducing childhood obesity by targeting water and sugar-sweetened drinks but also increasing physical activity [12, 14, 23, 30].

An important step in the design of such interventions is the selection of influential individuals. This is usually called "influence maximization" [6, 16], which is the task of selecting a small subset of nodes (seed nodes) in a social network that could maximize the spread of influence. Many algorithms have been suggested and developed, e.g. [5, 17, 19]. However, most of these papers focus on the efficiency of the algorithm for selecting the influential nodes. The aim of this paper is to explore how a diffusion model can be used to compare the effect on the spread of behaviour of: (1) different ways to build the network from questionnaires, (2) different strategies for selecting influential nodes, and (3) different percentages of the people that are targeted. In order to do this, we created an agent-based model that is supported by real data collected from Dutch primary and secondary school children [2]. The diffusion model is based on the one used by Beheshti et al. [1] and Giabbanelli et al. [11]. We tuned the model parameters on actual data on physical activity collected among Dutch school children.

In this paper, we first discuss the literature on using agent-based models for predicting the effect of contagion in social networks. Then, in Section 8.3, we describe the data and model that we used and the ways in which we can generate network graphs from the questionnaires. In Section 8.4, we report on the simulations that we have performed to tune the model and to compare the different strategies and networks. Finally, we discuss the consequences of our findings in Section 8.5.

8.2 Background

This section starts by presenting an overview of previous research studies that implemented agent-based models to predict intervention effects in social networks. Most of them consider the social or peer influence on the agent's particular health-related behaviour. The basis of our model is an agent-based model (ABM) of network diffusion of obesity behaviour, that looks at both environmental and social influences on physical activity and energy intake in a network. We continue by looking in detail at the work of Beheshti et al. [1] and Giabbanelli et al. [11], as our model builds on the model introduced in these papers.

There are numerous factors to be considered when selecting an appropriate network intervention, such as the type of network data, environmental context (e.g. geographic distance), network structure (highly centralized network versus decentralized network), prevalence of behaviour or agent's personal characteristics. In the following paragraphs, we review previous research on network-based interventions for reducing obesity and/or increasing physical activity in networks. Most of them compare different targeting methods for testing their effectiveness in diffusion of health behaviour.

The effects of targeting the most connected individuals as opposed to random individuals was investigated by El-Sayed et al. [21], with the goal of reducing population obesity in a social network. They looked at two different interventions, the first one preventing obesity, and the latter treating obesity, both by targeting 10% of the population. They concluded that targeting the most connected individuals may not be effective in reducing obesity in the network.

The selection of intervention strategy should depend on the purpose (the goal) of the particular intervention, as suggested by Zhang et al. [33]. They have studied network interventions for increasing children's physical activity on a real-life social network, composed of 81 children living in low socioeconomic status neighborhoods, out of which 41% was labeled as overweighted or obese. They used three different network intervention strategies and concluded that targeting opinion leaders is better for increasing the physical activity levels in the network as a whole, while targeting intervention in the most sedentary children is best to increase their own physical activity levels.

Zhang et al. [32] have studied the effect of both the social networks dynamics, and peer influence on overweight in adolescents in a real-life social network. They proposed an ABM for simulating the environment, and conducted several experiments on modifying the network's dynamics or changing strength of peer influence, to get a more refined model. They showed that peer influence can significantly affect those who are overweight. Bigger peer influence lowers the prevalence of being overweight, especially in low-obesity networks. On the other hand, in high-obesity networks, inducing stronger peer influence can have an unwanted reversed effect of further increasing the population's weight.

A "bottom-up" agent-based approach was introduced by Trogdon and Allaire [25], modeling the food consumption and friend selection at individual levels, for weight loss interventions. They have shown that the underlying social network can influence the effect of population-level interventions. Looking at the network structure, they have concluded that aggregate effects of population-level interventions are bigger in clustered networks, compared to scale-free networks. In addition, targeting particular agents of the network can be important for social network interventions. Selecting the most popular obese agents for the weight loss intervention, resulted in greater weight loss in the population than selecting a random assortment of obese targets.

Looking at the related work, we can conclude that literature gives contradictory outputs, possibly as a result of the complexity of social networks and the numerous factors that can influence network interventions as explained above. Choosing leaders as targets for intervention is shown as effective in [32], [25], as opposed to [21]. Evaluating new methods of targeting obesity interventions is needed in order to create both cost-effective and time-effective social networks interventions. Following this idea, Beheshti et al. [1] have developed an ABM, an adaptation of the model proposed by Giabbanelli et al. [11], that simulates the results of five targeting approaches, and integrates three key factors that influence the diffusion of intervention effects in a social network. These factors are: personal characteristics of agents, social network ties (social influence) and environmental influence. The authors propose two network interventions, the first one with the aim of reducing energy intake and the latter for increasing physical activity, both targeting 10% of the population. The individual traits of the agents, like BMI, sex, energy intake, environment, etc. were attributed based on the NLSY79 data set. The same data set was used to validate the model, and compare the simulated weight changes trends of the model, with the historical weight trends of the NLSY79 dataset. Comparing the effectiveness of the proposed targeting strategies, they concluded that targeting based on network information, outperforms more traditional targeting approaches like selecting high-risk agents or vulnerable categories (e.g., obese or low-income agents, respectively). Their most efficient targeting method is based on influence maximization and is explained in details in Section 8.3. Beheshti et al. [1] simulations are based on an artificial network built following a power law degree distribution and homophily properties. In this work, a real social network is used, which is derived from data collected through surveys, as explained in Section 8.3.

8.3 Methods

This section presents the methods used for this research. First we describe the data collected and how the characteristics of the population of children were used for the design of the ABM. Next we explain the model in detail, as well as the process of tuning the parameters to better fit the model to the empirical data. Then, we explain the strategies tested for selecting the targets for the interventions.

8.3.1 The data

The data have been collected in the *MyMovez* project – Phase I [2]. This is a large-scale cross-sequential cohort study among school children (N=953; 8-12 and 12-15 years-old) from 21 primary and secondary schools in the Netherlands. The *MyMovez* project – Phase I consists of five data collection waves over 3 years, starting in 2016: February/March 2016 (Wave 1), April/May 2016 (Wave 2), June/July 2016 (Wave 3), February/March 2017 (Wave 4) and February/March 2018 (Wave 5). In this paper, we used data from the 4 first waves, as the data collected for wave 5 is still being processed. The collected data contains information about the children's social network, media consumption, psychological determinants of behaviour, physical environment, eating behaviour, socialization characteristics and physical activities. The children were surveyed in many aspects through the *MyMovez* application on a research smartphone provided by the project in order to collect their impressions about their classmates and their own routine and habits. Participants' weight and height are measured individually by a trained researcher following standard procedures (without shoes but fully clothed) in Wave 2 and 4. The BMI is calculated as $\frac{weight(kg)}{height(m)^2}$. Data on physical activity is collected using a wearable device (bracelet) that tracks the steps of the participants for 5 days in a row (week and weekend days). For more detailed information, see the *MyMovez* project [2]. For the current study, we selected school classes with more than 80% of participation in the experiment, resulting in 26 classes out of 196. 455 participants were removed from the data set for not taking part in the selected classes. The total number of children after the cleaning of the data (removal of participants with missing data) is 451. The data was processed using Python 3 and the NetworkX library combined with Pandas data frames.

8.3.2 Model implementation

The network-oriented model used for this work is based on diffusion dynamics of behaviour throughout a social network. That means we assume that behaviour change regarding obesity aspects (physical activities and energy intake) are spread throughout one's relationships. The model is based on the work of Giabbanelli et al. [11] and some of the adaptations of Beheshti et al. [1] were also taken into account, as explained below. Two main factors are considered as determinants for the agents' behaviour change: the influence via the social network and the environmental influence.

Social network influence is the influence from peers, i.e. those people who are connected to the agents. The *environmental influence* is based on the social-economical conditions of each child. Many factors are important to assess the lifestyle of children and quantify it. In this paper, we focus on the spread of physical activity (PA) as a measure of healthy behaviour. Therefore the interventions applied will increase the PA of the selected participants.

The PA changes for the simulations are calculated in 3 steps, according to the method presented by Giabbanelli et al. [11]:

1. the influence on the individual by their friends;

2. the combination of the friends influence with the influence from the environment; and
3. a threshold used to decide if an individual's PA will be changed or not.

Beheshti et al. [1] and Giabbanelli et al. [11] treated the connections as binary variables, where a connection has a weight of 1 in case a relationship exists and 0 otherwise. For this work we measured the strength of the connections as float numbers between 0 and 1, as is going to be explained in Section 8.3.3. For this reason, we adjusted the formulas for step (1) regarding the weights of the edges as being part of the calculation of the peers' influences. For step (1), equation 8.1 show the friends influences based on the weight of the connections and the difference between the states of PA.

$$inf_{PA_i}(t) = \frac{\sum_j (PA_j(t-1) - PA_i(t-1)) \times w_{(j,i)}}{\sum_j w_{(j,i)}} \tag{8.1}$$

A positive inf_{PA_i} for node i means that the overall influence from i's friends is positive towards the PA of agent i. In these circumstances, a good environment will further increase PA, while a bad environment would do the opposite. For the simulation, the environment is beneficial when $0 < env < 1$ and harmful when $1 < env < 2$. The environment calculation is explained in detail in Section 8.3.4. Equations 8.2 and 8.3 show how step (2) is calculated, combining the influence of the peers with the influence of the environment.

$$inf_{PA_i(t),env} = env \times inf_{PA_i(t)},\ if\ inf_{PA_i} < 0 \tag{8.2}$$

$$inf_{PA_i(t),env} = \frac{inf_{PA_i(t)}}{env},\ if\ inf_{PA_i} \geq 0 \tag{8.3}$$

The last part of the influence spread is to compare the amount of influence with the given threshold. Beheshti et al. [1] defined the values for low and high thresholds for EI and PA as 0.002 and 0.2. When testing these values, many problems with convergence and steepness were raised in our simulations. For that reason, we went back to the original model, by Giabbanelli et al. [11] and kept only one threshold, applying a simulated annealing algorithm to fine tune it, as explained in Section 8.3.5. The threshold is used to define the minimum amount of impact that is going to cause the behaviour change to take effect. Equations 8.4 and 8.5 show the final value for PA in the next time step t, where $factor = 1 + I_{PA}$, in case $inf_{PA_i(t),env} > 0$, and $factor = 1 - I_{PA}$ otherwise.

$$PA_i(t) = PA_i(t-1),\ if\ |inf_{PA_i(t),env}| < T_{PA} \tag{8.4}$$

$$PA_i(t) = PA_i(t-1) \times factor,\ if\ |inf_{PA_i(t),env}| \geq T_{PA} \tag{8.5}$$

8.3.3 Building the network

The simulations in [1] are based on an artificial network, which is built following a power law degree distribution and homophily properties. In this work, a real social network is used, which is derived from data collected through surveys. Twelve questions about the relationships and impressions of other classmates were asked. In our experiment, we compared three different subsets of questions to build the network:

1. **(Friendship)** One question regarding friendship: "who are your friends?";

2. **(General)** 6 general questions, including the question about friendship regarding respect, advice, leadership and who they would like to resemble; and

3. **(All)** The questions from 1 and 2 (above) plus questions regarding physical activities and food intake behaviours, 12 in total. The extra questions are related to peers that influence you to eat healthier, exercise and practice sports.

The edges of our network are bidirectional, and they account for the amount of influence that the origin of the edge has on the destination. Every question generates a nomination from node i to j. This nomination is interpreted as the influence that node j has on i. The more nominations a node i gives to j, the stronger the influence of node j is over i, and therefore the value for the edge $w_{j,i}$ is higher. Each question receives a different weight of 0 or 1. For each subset of questions, a different configuration of the weights for the questions q_n is given. The total weight for the edge from node j to node i is given by equation 8.6.

$$w_{j,i} = \frac{\sum_{n=1}^{k}(q_n \; nomination_{i,j})}{\sum_{n=1}^{k} q_n} \tag{8.6}$$

The use of different subsets of questions is a way of mapping different networks based on levels of influence that can affect a social network. We assume that the question about friendship explains well who the people that children prefer to spend their time together with are, but it can be biased towards other children in the class who they want to be like, but are not as close as they want to be. For instance, popular children in the classroom might influence others, even if they don't spend much time together, or are considered friends. Table 8.1 shows the density of the networks for each of the classes selected. As can be seen, the networks generated by the single friendship question alone are the least dense, while the networks generated by using all questions present more connections that raise other levels of influence between the children.

To verify if there are significant differences between the three generated networks, a Hamming distance was applied to the edges in the graphs. Table 8.2 shows the Hamming distance for each of the classes and the three possible comparisons between the generated graphs. As is shown, the distance is close to zero to almost all the classes when comparing the graph for all questions (3) and the graph of general

Tab. 8.1: Density of the networks generated by the different subsets of questions.

Class	Friendship	General	All
1	0.45	0.67	0.75
2	0.29	0.65	0.71
3	0.34	0.66	0.80
4	0.69	0.85	0.86
5	0.48	0.65	0.72
6	0.58	0.80	0.84
7	0.30	0.46	0.51
8	0.64	0.88	0.95
9	0.60	0.78	0.87
10	0.70	0.79	0.80
11	0.53	0.83	0.87
12	0.53	0.67	0.76
13	0.63	0.81	0.89
14	0.42	0.68	0.73
15	0.50	0.76	0.84
16	0.57	0.78	0.81
17	0.72	0.89	0.89
18	0.46	0.69	0.76
19	0.64	0.91	0.96
20	0.32	0.62	0.73
21	0.40	0.64	0.77
22	0.39	0.59	0.68
23	0.62	0.89	0.91
24	0.50	0.77	0.85
25	0.43	0.67	0.70
26	0.25	0.58	0.72

Tab. 8.2: Hamming Distance between the three graphs generated by the subsets of questions

Classes	All x General	General x Frienship	All x Friendship
1	0.08	0.22	0.30
2	0.06	0.36	0.41
3	0.14	0.32	0.46
4	0.02	0.15	0.17
5	0.06	0.17	0.24
6	0.04	0.22	0.26
7	0.05	0.15	0.20
8	0.07	0.24	0.31
9	0.09	0.18	0.27
10	0.00	0.10	0.10
11	0.03	0.30	0.33
12	0.09	0.14	0.23
13	0.08	0.18	0.26
14	0.05	0.26	0.30
15	0.07	0.26	0.34
16	0.04	0.21	0.25
17	0.00	0.17	0.17
18	0.07	0.23	0.30
19	0.05	0.27	0.32
20	0.11	0.29	0.40
21	0.14	0.24	0.37
22	0.09	0.20	0.29
23	0.02	0.27	0.29
24	0.08	0.27	0.35
25	0.03	0.24	0.27
26	0.14	0.33	0.47

questions (2), while the friendship graph (1) presents a bigger distance compared to the other two. That means that the network generated by all questions and the network generated by the general questions present almost the same edges in their graphs, while the graph generated with the friendship question alone has many different edges from the other networks.

8.3.4 Agents characteristics

The agent-based model for the behaviour spread explained in Section 8.3.2 requires some information about the agents. More specifically, it is necessary to know the PA-level of each agent, as well as the influence of the environment on each of them, calculated using socio-economic status (SES). The BMI is also important to know for the interventions that target high risk children. Here we explain how these characteristics of the children were extracted from the empirical data set.

Environment

One of the most influential factors for a healthy lifestyle is a person's living environment. Family wealth can be a good predictor for a child's healthy living environment, with better opportunities for healthy eating and physical activity facilities. The Family Affluence Scale (FAS) is a simple metric created to avoid the difficulties that youth have in reporting family income or other measures for wealth [3]. Boyce et al. [3] argue that the FAS measures are related to food intake habits and to physical activity, meaning that this questionnaire can also be a good predictor for factors related to health aspects of a youth's lives.

In *MyMovez* project the participants were asked the following questions:

1. Does your family own a car, van or truck? (No [0]; Yes, one [1]; Yes, two or more [2]);

2. Do you have your own bedroom for yourself? (No [0]; Yes [1]);

3. During the past 12 months, how many times did you travel away on holiday with your family? (Not at all [0]; Once [1]; Twice [2]; More than two [3]); and

4. How many computers does your family own? (None [0]; One [1]; Two [2]; More than two [3]).

The model used for the simulation consider that more obesogenic environments have a scale factor between 1 and 2, while healthier environments present a scale from 0 to 1 [1]. An environment factor of 1 means neutral. We normalized the values so all four questions have the same weight in the overall calculation.

BMI

The Body Mass Index (BMI) is the metric used to define children with higher risk. That means that the higher the BMI the higher the risk of a child to become an obese adult. We compared the BMI of the children from the same class and sorted them to define the best targets for high risk interventions.

In [11] the overall change in BMI is used as the outcome measure of the effect of interventions. Their assumption is that the population is made of adults, and their height is fixed. But for children the same assumptions do not stand, as children have a much more dynamic and complex process of growing [22], which is followed by their BMI. For that reason, in our experiments we do not use the BMI as the outcome measure, but the PA level instead. The BMI is only used for the selection of the targets to apply the interventions.

PA

The participants in the experiment were asked to wear a Fitibit Flex bracelet for 5 days. The device measured their steps in continuous time with minute to minute precision. The PA used for the simulations was based on the number of steps. In [11] the PA of the simulated agents were drawn as a normal distribution with a mean value of 1.53, the level of sedentary individual according to Food and United Nations [10]. To normalize the values for PA in our data set we took the mean PA (in steps) and converted it to 1.53, in order to keep the same scale presented in the previous work.

The initial PA for the simulations is calculated as the average of the 3 first waves, and the final PA is given by wave 4. Waves 1, 2 and 3 are closer in time to each other, and also closer to the initial date of the experiment, while wave 4 is 1 year further.

8.3.5 Parameter tuning

Two parameter tuning algorithms are used to fine tune the thresholds and the speed factors of the model.

First we applied a grid search in the bi-dimensional space with the threshold (T_{PA}) and the factor (I_{PA}) of change. These variables are explained in Section 8.3.2 in detail. The grid search guided us to a subspace where the best results were obtained. Then we applied simulated annealing to fine tune our optimization search. The simulated annealing algorithm is an optimization combinatorial method for problem solving. It is inspired by condensed matter physics, where annealing denotes a process in which a solid in a heat bath is heated up to a maximum temperature so all particles are liquid, and then cooled down slowly so the solid particles are reorganized [29].

We started our simulations with a temperature of 1.0, and the cooling factor was 0.9, until the temperature was less than 0.01. For each temperature we explored 20 neighbors. The parameters found by the simulated annealing are used for the spread of behaviour model.

8.3.6 Strategies for selecting targets for intervention

The aim of this work is to explore the effect of different strategies to find targets for PA interventions. Several selection strategies were implemented, and the effect of applying an (imaginary) intervention to the selected agents on the overall PA is simulated.

The effect of the imaginary intervention to an individual is modeled as an increase of 17% of their initial PA. This value is taken from [1] and is chosen based on other research that apply different sorts of strategies to increase the amount of PA of people. The initial states of the remaining nodes is based on the empirical data. Then, the diffusion model is run to verify the improvement on the PA of the whole network after one year.

The strategies to select the targets are:

1. Higher risk children (BMI);
2. More vulnerable children (Environment);
3. Most central children in the network (degree centrality);
4. Optimized selection based on the impact of the children in the whole network;
5. Random selection of the targets.

Strategies (1) and (2) are based on the children's characteristics. Strategy (1) uses the BMI as indicator for the risk (the higher the BMI, the higher the risk). For the BMI we used the data about the children's height and weight in the first wave, as shown in Section 8.3.4. Strategy (2) targets vulnerable children. Vulnerability is measured based on the child's environment. The worse the environment (i.e. lower socio-economic status), the higher the vulnerability. For environment we used the FAS measures, as explained in Section 8.3.4. Strategies (3) and (4) are based on network characteristics. Strategy (3) is based on degree centrality of the nodes given by the Python toolbox NetworkX 2.1. The degree centrality for a node v is the fraction of nodes it is connected to. These values are normalized by dividing them by the maximum degree possible in the graph.

Strategy (4) selects the k nodes that propagate (or influence) the other nodes in the network the most. This is a simple algorithm in which the diffusion model algorithm is run for each of the nodes in the network after applying the intervention to each of these nodes separately. After running for all the nodes, the selected agent is the one that causes the biggest increase of PA in the whole network. Then after the first node is selected, the same is done to the other nodes that were not selected, together with the first node in the subset of targets. This strategy is based on the "influence maximization" algorithms used for viral marketing and advertising [5, 18]. The difference here is that instead of searching for nodes being "activated" as an objective function, we quantify the impact spread throughout the network as our optimization goal. Beheshti et al. [1] used the same strategy, but the goals were related to the decrease of the number of obese people in the network as the goal. For our optimization algorithm we are interested in the overall increase in PA for all the participants.

Lastly, strategy (5) selects the targets by random. This strategy is useful to verify if the other strategies provide better results than just selecting targets without any criteria.

As a third variable (in addition to the different methods to build the network and the different selection strategies), we compared the effect of the amount of different percentages of people to which the (imaginary) intervention was applied. We compared three different fractions: 10%, 15% and 20% of the nodes in a class. Table 8.3 shows the ratios between boys and girls per class, as well as the amount of targets to select in each fraction selection.

Tab. 8.3: Ratio between boys and girls per class and amount of targets for each fraction selection

Class	Total of kids	Boys	Girls	10%	15%	20%
1	18	9	9	2	3	4
2	20	13	7	2	3	4
3	20	14	6	2	3	4
4	12	6	6	1	2	2
5	19	6	13	2	3	4
6	20	10	10	2	3	4
7	25	12	13	2	4	5
8	28	15	13	3	4	6
9	14	7	7	1	2	3
10	16	8	8	2	2	3
11	20	10	10	2	3	4
12	18	7	11	2	3	4
13	17	12	5	2	3	3
14	14	10	4	1	2	3
15	11	9	2	1	2	2
16	17	7	10	2	3	3
17	11	7	4	1	2	2
18	14	9	5	1	2	3
19	9	6	3	1	1	2
20	21	10	11	2	3	4
21	11	8	3	1	2	2
22	20	8	12	2	3	4
23	18	8	10	2	3	4
24	19	9	10	2	3	4
25	20	7	13	2	3	4
26	19	6	13	2	3	4
Total	451	233	218	45	70	91

8.4 Results

The experiments aim to explore the effect of different networks, different target selection strategies and different fractions of targets on the overall PA of the social network. We first present the results of the model tuning on the empirical data, and then we present the results of the different simulation scenarios.

8.4.1 Tuning of threshold and change factor

Section 8.3.2 explained the model and the two parameters that should be tuned in order to fit the empirical data to the simulations: threshold T_{PA} and the factor of change I_{PA}. T_{PA} is the amount of influence a child needs to receive from their peers combined with the environment to be affected and change their behaviour. The factor of change I_{PA} is the amount of influence that is going to be propagated to the receiving node within the network. We used a grid search followed by a simulated annealing algorithm to minimize the difference between the empirical data and the simulated data and find the best values for T_{PA} and I_{PA}. The tuning was performed in two steps: first, a grid search algorithm was applied to identify the search space, and then a simulated annealing algorithm was used to find the optimal values.

Grid Search

The grid search was performed with interval steps of 0.05. The values for T_{PA} and I_{PA} were tuned for each of the networks generated with: (1) all questions, (2) general questions and (3) friendship questions. As can be seen in Figure 8.1, the space where $I_{PA} < 0.35$ showed the smaller errors. The grid search for the other graphs follow almost the same pattern as the one presented in Figure 8.1.

Simulated annealing

After the space was defined, the simulated annealing was run. As a result of the grid search, the value of I_{PA} was restricted to a maximum of 0.4. To calculate the error (i.e. the difference between the empirical data and the simulation) we compared the simulation outcomes with the data in waves 1 and 4. Waves 2 and 3 were ignored because they are too close to Wave 1. For the simulated annealing, the initial temperature was 1.0, with an alpha (cooling factor) of 0.9, and a number of neighbors explored was 20. Increasing the number of neighbors or slowing down the speed of the cooling process didn't significantly improve the results obtained. The initial parameters for the simulated annealing were $threshold = 0.2$ and $I_{PA} = 0.05$. Figure 8.2 shows the space of search on the simulated annealing algorithm for the graph created from the general questions.

The best results for threshold and I_{PA} for each of the three networks are presented in table 8.4.

Fig. 8.1: Grid search for the general questions.

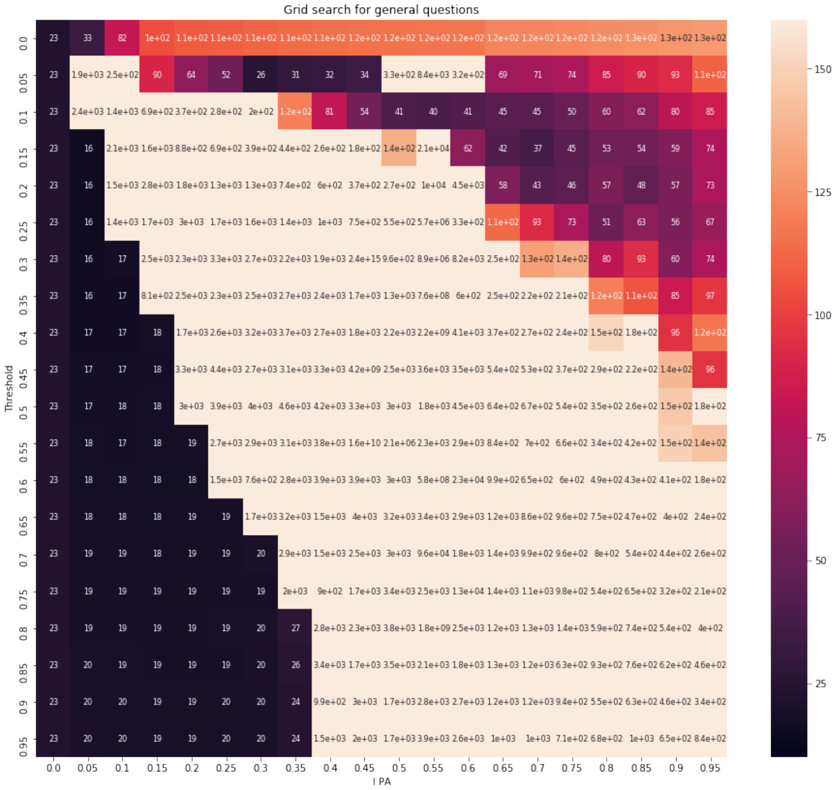
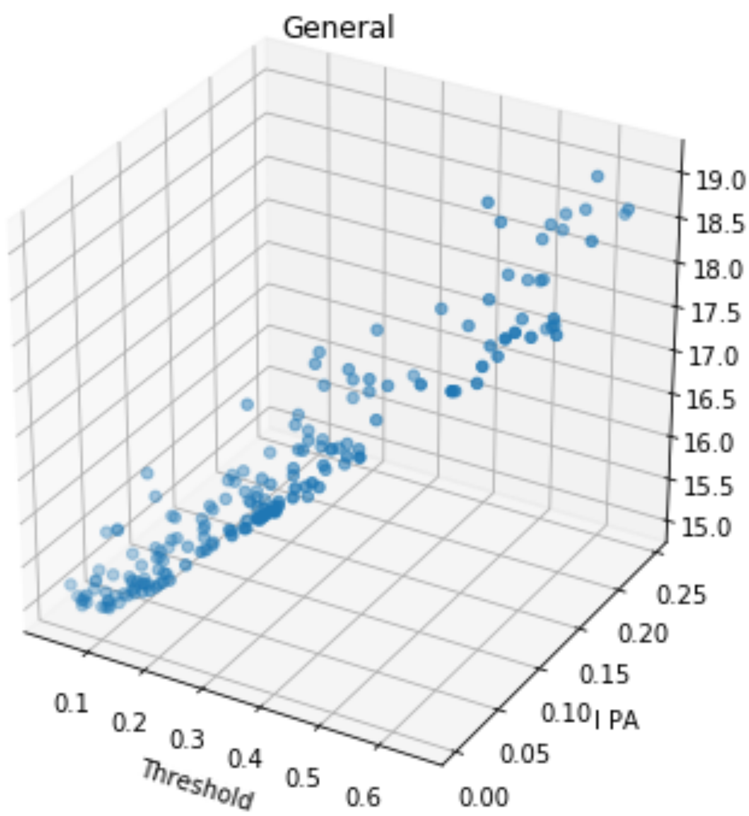


Fig. 8.2: Simulated annealing space of search explored for the general questions.



Tab. 8.4: Best threshold and factor of change for the three networks generated with (1) all questions, (2) general questions and (3) friendship question

	All (1)	**General (2)**	**Friendship (3)**
Threshold	0.0942	0.0588	0.0426
I_{PA}	0.0055	0.0057	0.0041

Tab. 8.5: Differences between initial and end mean PA for the network generated with all questions and 20% of the nodes selected for intervention.

	Diff (day 0)	**Diff (day 364)**	**Overall diff**
High Risk	0.0491	0.0469	-0.0022
Vulnerability	0.0515	0.0871	0.0356
Random	0.0516	0.05599	0.0044
Centrality	0.0548	0.0623	0.0075
Optimized	0.0508	0.0985	0.0476

8.4.2 Exploring different strategies

After fine tuning the model parameters, the diffusion model was run combined with the strategies to select targets for intervention in the network. For each class, three different percentages of the children were chosen as targets. To select the targets, we compared the use of 5 different strategies: (1) Random selection; (2) High-risk selection; (3) Vulnerability selection; (4) Degree centrality selection; and (5) Optimization selection.

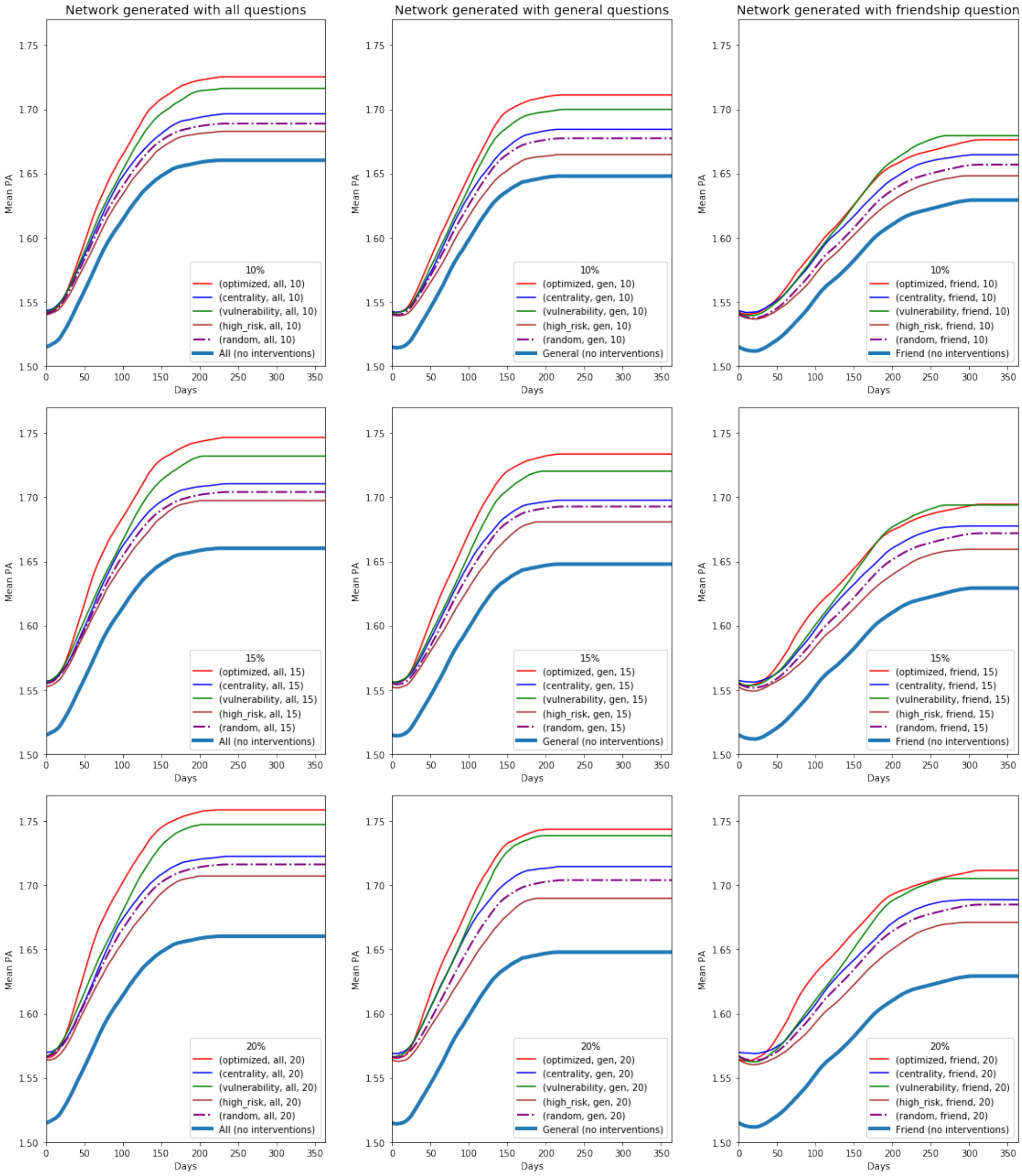
Figure 8.3 shows the comparisons between the random selection of 10, 15 and 20% of the agents for the three networks generated with all questions, general questions and friendship question.

We used the random selection (1) of the children to receive interventions on their PA as a baseline. We run the random selection for 100 samples and used the mean to evaluate the average impact of this method. Increasing the number of tests didn't cause any difference to the results.

The selection of the high-risk agents (2) is based on the BMI of the children in the classes. The children with a higher BMI have a higher risk of obesity later on. As shown in Figure 8.3, the "high risk" is the worse intervention causing the smallest influence on increasing PA. Table 8.5 shows the differences between the initial (and final) mean PA of the whole network with no interventions and the initial (and final) mean PA for the networks with the interventions applied. The overall difference is how much the difference of the mean PAs increased (or decreased) from the beginning to the end of the simulation (start-to-end difference). The overall difference for the high risk is the only negative one, meaning that after one year of simulation the mean PA of this intervention is closer to the simulations without interventions.

The vulnerability (3) was used to select the children with the least socio-economic situation as targets for the interventions. For most of the scenarios, this strategy

Fig. 8.3: Simulation of average PA for all the strategies for the different percentages of targets in the three rows. The left column shows the results for the network generated with all questions. The center column is for the network generated with the general questions set. The right column shows the results for the network generated using the friendship question alone.



Tab. 8.6: Differences from the mean PA for each intervention to the simulations without any intervention. The results are referred to the network generated from all the questions combined.

	Beginning (day 0)			End (day 364)			Difference			Difference (%)		
	10%	15%	20%	10%	15%	20%	10%	15%	20%	10%	15%	20%
High risk	0.0250	0.0374	0.0491	0.0225	0.0371	0.0469	-0.0024	-0.0003	-0.0022	0.00	-88.16	664.07
Vulnerability	0.0267	0.0408	0.0515	0.0560	0.0718	0.0871	0.0294	0.0310	0.0356	0.00	5.69	14.71
Random	0.0256	0.0398	0.0516	0.0286	0.0439	0.0560	0.0030	0.0041	0.0044	0.00	37.31	8.45
Centrality	0.0280	0.0418	0.0548	0.0362	0.0503	0.0623	0.0082	0.0085	0.0075	0.00	3.81	-12.28
Optimized	0.0268	0.0401	0.0508	0.0650	0.0863	0.0985	0.0382	0.0462	0.0476	0.00	20.92	3.03

was one of the two best interventions, performing almost as well as the optimized solution especially for the network created with the single friendship question. Strategy (3) shows an overall difference of 0.0356, the second highest (see Table 8.5).

Strategies based on network degree centrality (4) were also tested and compared with the others. This intervention presents better results than the random selection (1), but are not as good as the strategies (3) and (5). Some positive improvement in the distance to the expected changes in the network without interventions is also perceived (Table 8.5).

The highest overall PA increase are in the interventions based on the optimization algorithm. This strategy is beaten by strategy (3) in one of the nine scenarios, though it shows the best results for all the other remaining scenarios. Table 8.6 presents the values for the differences between the strategies applied and the simulation with no interventions. The column Beginning (day 0) is the initial point and the difference between the mean PA for each strategy and no intervention. The same is valid for the column End (day 364), that represents the distance (or difference) on the last day of simulation. The column Difference is calculated as the value for the day 364 minus the day 0 for each of the percentages. The Difference (%) column represents how much increase (or decrease) of PA is caused by increasing the amount of intervention targets from 10 to 15% of the agents, and from 15 to 20% of the agents. The results from table 8.6 are important in order to make the right decision of selecting the appropriate nodes for real-life intervention. Therefore, increasing the fraction for the "high risk" strategy is not beneficial, as there is a degradation of the difference between mean PA for this strategy and for the simulation without any interventions.

Choosing targets based on centrality shows that selecting 15% of agents is better than selecting 20%. For random and vulnerability selections the improvements caused when increasing the percentage of targets shows that 20% is a good choice, while for the optimized strategy there is a very small percentage increase from 15 to 20% (3.03%), which would require further investigation to decide if selecting 20% of the nodes instead of 15% is a good decision.

8.5 Discussion

The simulation results allow us to discuss potential strategies that may be used to make informed choices about ways to improve simulation models as well as the setup of social network interventions.

The strategy of targeting children with high risk (i.e., higher BMI) turned out to be the least successful compared to the other strategies in this specific data set and model. It should be noted that our sample had a small variance in BMI. The majority of the BMIs are healthy. Therefore, future studies should test whether a more heterogeneous sample would generate different findings. In addition, the model can be further improved. For example, if BMI would be used as the dependent variable, it would show how the overall prevalence of obesity in social networks is reduced by the intervention. For this, a more complex model is required which accounts for the dynamic change in the BMI categorization in adolescents. The inclusion of the information regarding the energy intake of the children would also enrich the model by providing another independent variable which directly affects the obesity of the group. The difficulties of including this variable are related to the data collection, as it would be required to have a method for assessing the children's food intake habits. Targeting children who reported to live in a less wealthy home environment appeared to be one of the best solutions and very close to—or even better than—the optimized strategy in some scenarios. This insight can also be helpful for the selection of targets in the absence of the network structure, as a personal characteristic provides good improvement of the overall PA of the whole group.

The change in the networks based on different subsets of question did not reflect drastically different results for the scenarios. That is a good indicator that the weights of the connections present some stability when you compare the different graphs, and it does not affect the spread of behaviour process. This suggests that previous peer-driven intervention studies which used subsets of 5 nomination questions to identify the peer leaders have chosen a sufficient number of questions [4, 23]. For future research and data collection, these results indicate that a small subset of nomination questions can be asked to the participants, without losing quality on the description of the influential ties and overburdening the participants with questions.

With respect to the fraction of people that are targeted, our results show that it differs per strategy whether it is beneficial to increase the target group. This is a finding with important consequences, as the costs of an intervention usually increases linearly with the increase of the size of the target group.

8.6 Conclusion

This paper shows how an agent-based model can be used to explore the effect of different scenarios on the diffusion of physical activity in children. It is the first study in which simulations are based on real social networks and the model has been tuned on actual measurements of physical activity in the network. Study findings indicate that social network interventions aimed at increasing physical activity should take the socio-economic status of children into account. In addition, a small subset of peer nomination question is sufficient to map the children's social network.

In this study, we have simulated the effect of targeting specific children and applying an intervention that increases in 17% their PA. We have compared 5 strategies for the selection, in 3 differently generated networks and with 3 different percentages of targets selected. We compared these 45 different scenarios with the expected results when no intervention is applied. Although this is a comprehensive strategy, the model used in the simulations is a straightforward diffusion model. As a future work, we would like to test other, more realistic, contagion models on the same data set. For example, an option is to use a differential equation model that includes personality traits to account for other factors that play a role in the process of social contagion. Other potential ways to improve the results would be to combine an optimized strategy with the vulnerability trait of the agents to verify if it can show better results than of two strategies alone.

Acknowledgements The research leading to these results has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP7/2007-2013) / ERC grant agreement number [617253].

Appendix

This Appendix contains the questions from the survey regarding the relationships between the children in the MyMovez data collection. The questions were divided in 3 subgroups to measure the strength of the connections between the participants: (1) All questions; (2) GENERAL questions; (3) GENERAL/Friendship question alone.

1. (GENERAL/Friendship) Who are your friends?
2. (GENERAL/Advice) Who do you ask for advice?
3. (GENERAL/Leader) Who are the leaders or take the position of leader in a group?
4. (GENERAL/Respect) Who do you respect?
5. (GENERAL/Social Facilitation) With whom do you spend time?
6. (GENERAL/Want to be) Whom do others want to look like? (+)
7. (DIETERY/Com Network) With whom do you talk about what they eat and drink?
8. (DIETERY/Modeling) Who eats or drinks products that you would like to eat or drink?
9. (DIETERY/Impression management) Of whom do you think it is important that they see you as someone who eats and drink healthly?
10. (PA/Impression Management) Of whom do you think it is important that they see you as someone who is physically active?
11. (PA/Com Network) Who do you talk about physical activity and sports?
12. (PA/Modeling) Who are exercising in a way you also want to exercise? (MEDIA/Com Network) With whom do you talk about what do you see on TV or internet?

Bibliography

[1] Rahmatollah Beheshti, Mehdi Jalalpour, and Thomas A. Glass. „Comparing methods of targeting obesity interventions in populations: An agent-based simulation". In: *SSM - Population Health* 3 (2017), pp. 211–218 (cit. on pp. 122–127, 130–132).

[2] Kirsten E Bevelander, Crystal R Smit, Thabo J van Woudenberg, et al. „Youth's social network structures and peer influences: study protocol MyMovez project–Phase I". In: *BMC public health* 18.1 (2018), p. 504 (cit. on pp. 122, 125).

[3] William Boyce, Torbjorn Torsheim, Candace Currie, and Alessio Zambon. „The Family Affluence Scale as a Measure of National Wealth: Validation of an Adolescent Self-Report Measure". In: *Social Indicators Research* 78.3 (Sept. 2006), pp. 473–487 (cit. on p. 130).

[4] Rona Campbell, F Starkey, J Holliday, et al. „An informal school-based peer-led intervention for smoking prevention in adolescence (ASSIST): a cluster randomised trial". In: *The Lancet* 371.9624 (2008), pp. 1595–1602 (cit. on pp. 122, 139).

[5] Wei Chen, Chi Wang, and Yajun Wang. „Scalable influence maximization for prevalent viral marketing in large-scale social networks". In: *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM. 2010, pp. 1029–1038 (cit. on pp. 122, 132).

[6] Wei Chen, Yajun Wang, and Siyu Yang. „Efficient influence maximization in social networks". In: *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '09*. New York, New York, USA: ACM Press, 2009, p. 199. arXiv: 1204.4491 (cit. on p. 122).

[7] Nicholas A Christakis and James H Fowler. „The spread of obesity in a large social network over 32 years". In: *New England journal of medicine* 357.4 (2007), pp. 370–379 (cit. on p. 122).

[8] Kayla De La Haye, Garry Robins, Philip Mohr, and Carlene Wilson. „How physical activity shapes, and is shaped by, adolescent friendships". In: *Social science & medicine* 73.5 (2011), pp. 719–728 (cit. on p. 122).

[9] Kayla De la Haye, Garry Robins, Philip Mohr, and Carlene Wilson. „Obesity-related behaviors in adolescent friendship networks". In: *Social Networks* 32.3 (2010), pp. 161–167 (cit. on p. 122).

[10] Food and Agriculture Organization of the United Nations. *Human energy requirements: Report of a Joint FAO/WHO/UNU Expert Consultation*. 2004 (cit. on p. 131).

[11] Philippe J. Giabbanelli, Azadeh Alimadad, Vahid Dabbaghian, and Diane T. Finegood. „Modeling the influence of social networks and environment on energy balance and obesity“. In: *Journal of Computational Science* 3.1 (2012), pp. 17–27 (cit. on pp. 122–126, 130, 131).

[12] Pedro C Hallal, Lars Bo Andersen, Fiona C Bull, et al. „Global physical activity levels: surveillance progress, pitfalls, and prospects“. In: *The lancet* 380.9838 (2012), pp. 247–257 (cit. on p. 122).

[13] Ross A Hammond. „Social influence and obesity“. In: *Current Opinion in Endocrinology, Diabetes and Obesity* 17.5 (2010), pp. 467–471 (cit. on p. 122).

[14] Ruth F Hunter, Kayla de la Haye, Jennifer Badham, et al. „Social network interventions for health behaviour change: a systematic review“. In: *The Lancet* 390 (2017), S47 (cit. on p. 122).

[15] Jeffrey A Kelly, Janet S St Lawrence, Yolanda E Diaz, et al. „HIV risk behavior reduction following intervention with key opinion leaders of population: an experimental analysis.“ In: *American journal of public health* 81.2 (1991), pp. 168–171 (cit. on p. 122).

[16] David Kempe, Jon Kleinberg, and Éva Tardos. „Maximizing the Spread of Influence Through a Social Network“. In: *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD ’03. Washington, D.C.: ACM, 2003, pp. 137–146 (cit. on p. 122).

[17] Bo Liu, Gao Cong, Yifeng Zeng, Dong Xu, and Yeow Meng Chee. „Influence spreading path and its application to the time constrained social influence maximization problem and beyond“. In: *IEEE Transactions on Knowledge and Data Engineering* 26.8 (2014), pp. 1904–1917 (cit. on p. 122).

[18] Flaviano Morone and Hernán A Makse. „Influence maximization in complex networks through optimal percolation“. In: *Nature* 524.7563 (2015), p. 65 (cit. on p. 132).

[19] Huy Nguyen and Rong Zheng. „On Budgeted Influence Maximization in Social Networks“. In: *IEEE Journal on Selected Areas in Communications* 31.6 (June 2013), pp. 1084–1094. eprint: `arXiv:1204.4491v2` (cit. on p. 122).

[20] Sarah-Jeanne Salvy, Kayla De La Haye, Julie C Bowker, and Roel CJ Hermans. „Influence of peers and friends on children’s and adolescents’ eating and activity behaviors“. In: *Physiology & behavior* 106.3 (2012), pp. 369–378 (cit. on p. 122).

[21] Abdulrahman M El-Sayed, Lars Seemann, Peter Scarborough, and Sandro Galea. „Are network-based interventions a useful antiobesity strategy? An application of simulation models for causal inference in epidemiology“. In: *American journal of epidemiology* 178.2 (2013), pp. 287–295 (cit. on pp. 123, 124).

[22] Yvonne Schönbeck, Henk Talma, Paula van Dommelen, et al. „Increase in prevalence of overweight in Dutch children and adolescents: a comparison of nationwide growth studies in 1980, 1997 and 2009“. In: *PloS one* 6.11 (2011), e27608 (cit. on p. 130).

[23] Crystal R Smit, Rebecca NH de Leeuw, Kirsten E Bevelander, William J Burk, and Moniek Buijzen. „A social network-based intervention stimulating peer influence on children’s self-reported water consumption: A randomized control trial“. In: *Appetite* 103 (2016), pp. 294–301 (cit. on pp. 122, 139).

[24] Kirsten P Smith and Nicholas A Christakis. „Social networks and health“. In: *Annu. Rev. Sociol* 34 (2008), pp. 405–429 (cit. on p. 122).

[25] Justin G Trogdon and Benjamin T Allaire. „The effect of friend selection on social influences in obesity“. In: *Economics & Human Biology* 15 (2014), pp. 153–164 (cit. on p. 124).

[26] Thomas W Valente. „Network interventions“. In: *Science* 337.6090 (2012), pp. 49–53 (cit. on p. 122).

[27] Thomas W Valente, Beth R Hoffman, Annamara Ritt-Olson, Kara Lichtman, and C Anderson Johnson. „Effects of a social-network method for group assignment strategies on peer-led tobacco prevention programs in schools“. In: *American journal of public health* 93.11 (2003), pp. 1837–1843 (cit. on p. 122).

[28] Thomas W Valente and Patchareeya Pumpuang. „Identifying opinion leaders to promote behavior change“. In: *Health Education & Behavior* 34.6 (2007), pp. 881–896 (cit. on p. 122).

[29] Peter JM Van Laarhoven and Emile HL Aarts. „Simulated annealing“. In: *Simulated annealing: Theory and applications*. Springer, 1987, pp. 7–15 (cit. on p. 131).

[30] Thabo J Van Woudenberg, Kirsten E Bevelander, William J Burk, et al. „A randomized controlled trial testing a social network intervention to promote physical activity among adolescents“. In: *BMC public health* 18.1 (2018), p. 542 (cit. on p. 122).

[31] UNICEF WHO, C Mathers, et al. „Global strategy for women's, children's and adolescents' health (2016-2030)“. In: *Organization* 2016.9 (2017) (cit. on p. 122).

[32] J Zhang, L Tong, PJ Lamberson, et al. „Leveraging social influence to address overweight and obesity using agent-based models: the role of adolescent social networks“. In: *Social science & medicine* 125 (2015), pp. 203–213 (cit. on pp. 123, 124).

[33] Jun Zhang, David A Shoham, Eric Tesdahl, and Sabina B Gesell. „Network interventions on physical activity in an afterschool program: An agent-based social network study“. In: *American journal of public health* 105.S2 (2015), S236–S243 (cit. on p. 123).

Part IV

Use of Contagion Models for Perceptions and Emotions

"And may I ask, oh, Lucy Daughter of Eve," said Mr. Tumnus, "how you have come into Narnia?"

"Narnia? What's that?" said Lucy.

"This is the land of Narnia," said the Faun, "where we are now; all that lies between the lamp-post and the great castle of Cair Paravel on the eastern sea. And you– you have come from the wild woods of the west?"

"I– I got in through the wardrobe in the spare room," said Lucy.

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

A Temporal-Causal Model for Spread of Messages in Disasters[1]

9

"I have received no assurance that anything we can do will eradicate suffering. I think the best results are obtained by people who work quietly away at limited objectives, such as the abolition of the slave trade, or prison reform, or factory acts, or tuberculosis, not by those who think they can achieve universal justice, or health, or peace. I think the art of life consists in tackling each immediate evil as well as we can."

— **C.S. Lewis**
"The weight of glory"[2]

Abstract

In this paper we describe a temporal-causal model for the spread of messages in disaster situations based on emotion contagion and awareness works. An evaluation of the model has been done by simulation experiments and mathematical analysis. Parameter tuning was done based on two scenarios, including a credible message and a dubious message. The results are useful for the prediction of reactions during disasters, and can be extended to other applications that involve panic and supportive systems to assist people.

[1]Published as: **Fernandes de Mello Araújo E.**, Franke A., Hosain R.W. (2017) *A Temporal-Causal Model for Spread of Messages in Disasters*. In: Nguyen N., Papadopoulos G., Jędrzejowicz P., Trawiński B., Vossen G. (eds) Computational Collective Intelligence. ICCCI 2017. Lecture Notes in Computer Science, vol 10449. Springer, Cham 2017

9.1 Introduction

The possibility of disaster and tragedy are a constant in everyone's lives. Apart from the problems caused by humans themselves (due to political decisions, errors, etc.), natural disasters also frequently occur in many places on Earth.

In 1953, a flood was caused by an extremely heavy storm in the Netherlands, England, Scotland and Belgium, called the North Sea Flood [4]. In the aftermath of this disastrous flood, a protective construction in the form of the Delta Works was created in the Netherlands in order to protect the country against similar natural disasters. Despite the protection of this new construction, a new flood is still a possibility. This was shown in a Dutch TV series called 'Als de dijken breken', or directly translated, 'If the dikes break'. This series raises the question of whether people are prepared for such a natural disaster [6]. What makes the scenario of 1953 different from nowadays is the current use of technology for communication. During the Twin Towers attack on the 9/11 in 2001 in New York, survivors made phone calls during the evacuation, most of which were not directed to emergency personnel, but to relatives, friends and family [1].

The spread of information in emergency situations can bring panic or can calm people down, like for relatives and friends that would remain in a stressful state in case of a lack of information about someone involved in a tragedy. In order to understand this scenario, we propose a temporal-causal model that considers how the act of sending a message could influence people's behaviour through social contagion. Some of the questions that guided us are related to how the information (message) is received. How people react regarding to the context, the sentimental and the emotional charge of the message? Are they unable to perform any action, or intentionally not taking action when the message is not taken seriously? Does the source and the type of communication define if the message is serious or not?

In section 9.2 we discuss the background theory. Section 9.3 discusses the temporal-causal network model in detail, with both a conceptual and a numerical representation. Section 9.4 is dedicated to the parameter tuning and datasets. In section 9.5 are the simulation results in multiple scenarios, and the mathematical analysis. Lastly, section 9.6 will be the discussion of the paper.

9.2 Social Contagion and Behaviour in Disaster Situations

Modeling disaster situations is a huge challenge. Especially, because it is impossible to simulate realistic scenarios for this sort of event. Because of this, our model considers similar situations and combines different works that explain parts of the cognition of humans and presents some solid ground to build upon.

Blake et al. [1] studied the reactions of survivors of the World Trade Center attack on 9/11, in 2001. Over 20% of the survivors that participated in his research had made phone calls during the evacuation, and 75% of these calls were directed to relatives, friends and family, and not emergency personnel. The survivors wanted to inform their relatives about the situation, their whereabouts and warn them about what was going on. They also used the calls, text messages and emails (on Blackberry devices) to gather more information on the situation from outside during the evacuation process. In our model we consider the emergence of social media as a trend, and possibly an easy way to communicate with people during a disaster.

Paton [5] developed a model of disaster preparedness using the knowledge about the social cognitive preparation system. This model describes how people prepare for disaster situations that might occur in the future and how different factors can affect that process. The focus on disaster situations in the future is different from our approach as we want to look into the spontaneous occurrence of a disaster and how people respond here. Paton shows that there are clear indications that anxiety or fear can play a motivating or demotivating effect in preparing for a disaster.

Bosse et al. [2] propose a temporal-causal model for emotion contagion based on interaction between individuals. It shows how specific traits of people define how they affect each other. For our model we assume that the messages, used to communicate, carry some subjective content. This can be seen as the sentiment of the message. Furthermore, it carries other cues to which the cognitive system will pay attention on the attempt to unfold and understand, for instance, how serious the message is.

Thilakarathne and Treur [8] present a computational agent-based model to simulate emotional awareness and how this may affect the execution of an action. Thilakarathne and Treur [7] also introduce a neurologically inspired agent model which makes distinctions between prior and retrospective awareness of actions. Those concepts are used in our model, as a way of tracking awareness in our agents who will receive messages. Our model follows the Network-Oriented modeling approach, proposed by Treur [9].

9.3 Agent-based model

This section presents the designed temporal-causal model. The numerical representation for the connections in the network is also shown. This model represents a scenario where the person has a perception of the environment, and starts receiving messages from another person about a possible disaster happening. The internal states are based on the emotions and potential actions of the person. Figure 10.1 shows the conceptual representation of the temporal-causal model.

Table 9.1 describes the meaning of each state in the model. We based our model on the previous models by Thilakarathne and Treur [7] and Thilakarathne and Treur [8], but without the element of ownership. In addition, the model has been extended in the field of emotion and sentiment, both positive and negative. Despite the fact that we realize culture could be of some influence, we decided to not include this in the model.

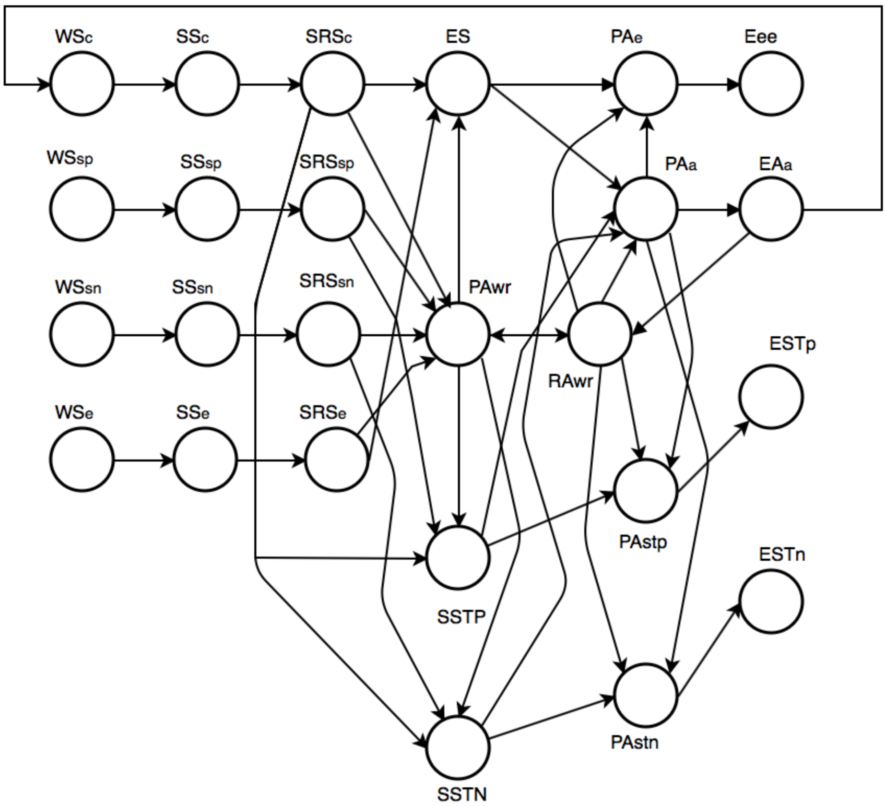


Fig. 9.1: Temporal causal network model

The scenario for this model can be understood as someone who receives a message claiming that something bad is happening, and the person starts to investigate the environment to see if there is any cue that matches the message. It can be an alarm message about a storm approaching. In this case, the perceptions about the environment could come from taking a look outside of a window.

The model has four different world state inputs. The critical awareness of hazards [5] are represented as the world state *context* (the environmental context at the moment) WS_c. In order to include the anxiety factor, we added emotion (scariness, shock, excitement and shame, based on Ekman's basic emotions [3]), positive sentiment and negative sentiment, respectively as world states WS_e, WS_{sp} and WS_{sn}.

These external inputs are then sensed and lead to the sensor states SS_c, SS_{sp}, SS_{sn} and SS_e. Subsequently they proceed to the sensory representation states, SRS_c, SRS_{sp}, SRS_{sn} and SRS_e, which indicate how intense the stimuli is perceived by the person.

The emotional state ES is the current emotional state of the person, influenced by the *context* and *emotion* stimuli. The prior-awareness $PAwr$ is the awareness state of a person before they have executed any action. The $PAwr$ can then be suppressed by the retrospective-awareness, $RAwr$, when the person actually has executed an action which might have changed their awareness state. The two states for the sentiments are the positive, $SSTP$, and the negative state, $SSTN$. These two states

Tab. 9.1: External and internal states of the model

Notation	Description
WS_c	World state context *c*
WS_{sp}	World state sentiment positive *sp*
WS_{sn}	World state sentiment negative *sn*
WS_e	World state emotion *e*
SS_c	Sensor state context *c*
SS_{sp}	Sensor state sentiment positive *sp*
SS_{sn}	Sensor state sentiment negative *sn*
SS_e	Sensor state emotion *e*
SRS_c	Sensory representation of context *c*
SRS_{sp}	Sensory representation of sentiment positive *sp*
SRS_{sn}	Sensory representation of sentiment negative *sn*
SRS_e	Sensory representation of emotion *e*
ES	Emotion state
$PAwr$	Prior-awareness state
$RAwr$	Retrospective-awareness state
$SSTP$	Sentiment state positive
$SSTN$	Sentiment state negative
PA_e	Preparation for action emotion *e*
PA_a	Preparation for action *a*
PA_{stp}	Preparation for action sentiment state positive *stp*
PA_{stn}	Preparation for action sentiment state positive *stn*
EE_e	Expressed emotion *e*
EA_a	Execution of action *a*
EST_p	Expressed sentiment state positive *p*
EST_n	Expressed sentiment state negative *n*

are defined as the current sentiment state of the person, whether the stimuli had a positive or negative sentimental charge.

The model has four similar preparation states. The state preparation of a person to express an emotion, PA_e, leads to the external state expressed emotion EE_e. The state preparation for action PA_a prepares a person to execute an action, which leads to the external state execution of action, EA_a. Lastly, both states preparation for action sentiment state positive stp, PA_{stp} and preparation for action sentiment state negative stn, PA_{stn}. These two preparation states lead to the expressed sentiment state positive EST_p and expressed sentiment state negative EST_n.

Appendix A (`www.cs.vu.nl/~efo600/iccci17/appendixA.pdf`) shows all connections between the states within the model and where each of the combined functions were used. The temporal-causal behaviour of the model is based on the methods proposed by Treur [9]:

1. At each time point t each state Y in the model has a real number value in the interval $[0, 1]$, denoted by $Y(t)$.

2. At each time point t each state X connected to state Y has an *impact* on Y defined as $\textbf{impact}_{X,Y}(t) = \omega_{X,Y} X(t)$ where $\omega_{X,Y}$ is the weight of the connection from X to Y.

3. The *aggregated impact* of multiple states X_i on Y at t is determined by a *combination function* $\mathbf{c}_\gamma$**(..)** where X_i are the states connected to state Y.

$$
\begin{aligned}
\textbf{aggimpact}_\gamma(t) &= c_\gamma(\textbf{impact}_{X_1,Y}\textbf{(t), ..., impact}_{X_k,Y}\textbf{(t)}) \\
&= c_\gamma(\omega_{X_1,Y} X_1(t), ..., \omega_{X_k,Y} X_k(t))
\end{aligned}
\tag{9.1}
$$

4. The effect of $\textbf{aggimpact}_\gamma\textbf{(t)}$ on Y is exerted over time gradually, depending on *speed factor* η_γ

$$
\mathbf{d}Y(t)/\mathbf{d}t = \eta_\gamma[\textbf{aggimpact}_\gamma(t)\text{- } Y(t)]
\tag{9.2}
$$

5. Thus the following difference and differential equation for Y are obtained:

$$
\mathbf{d}Y(t)/\mathbf{d}t = \eta_\gamma[\mathbf{c}_\gamma(\omega_{X_1,Y} X_1(t), ..., \omega_{X_k,Y} X_k(t)) - Y(t)]
\tag{9.3}
$$

The two combination functions used in our model were the identity and the advanced logistic (alogistic) functions. The identity function is $\mathbf{c}_\gamma(V) = \mathbf{id}(V) = V$, while equation 9.4 shows the advanced logistic function. The results of the simulations are shown in Section 9.5.

$$
\begin{aligned}
\mathbf{c}_\gamma(V_1, ...V_k) &= \textbf{alogistic}_{\sigma,\tau}(V_1, ...V_k) \\
&= \left(\frac{1}{1+e^{\sigma(V_1+...+V_k-\tau)}} - \frac{1}{1+e^{\sigma\tau}}\right)(1+e^{-\sigma\tau})
\end{aligned}
\tag{9.4}
$$

9.4 Data generation and parameter tuning

This section describes which datasets were used to tune the parameters of the model. It is, furthermore, described how the parameters were tuned.

9.4.1 Experimental datasets

Finding experimental data on the cognitive reactions during the spread of messages in disaster situations is difficult. The data concerning the messages is mostly missing or protected by messenger services. The information about cognitive states of people in the context of message receiving is difficult to obtain, due to limitations on the extraction of the data. Therefore, we created two different experimental datasets based on our understanding of the problem and on the literature. Both datasets contain information about a person receiving a message about a disastrous situation that might occur.

The first experimental dataset defines the course of all 25 states for a person receiving a message through the telephone with a tensed emotion and negative sentiment. This person is easily influenced by the sentiment and emotion of the message and believes that the message is true. In figure 9.2 the course of the 25 states in the first dataset is shown. Bosse et al. [2] state that the emotion of a person is influenced by the impact of the emotion of another person and the person's own belief. We assume that the impact of the emotion is big, because this person received a telephone call and this person is easily influenced. It is considered that the person will react spreading the message.

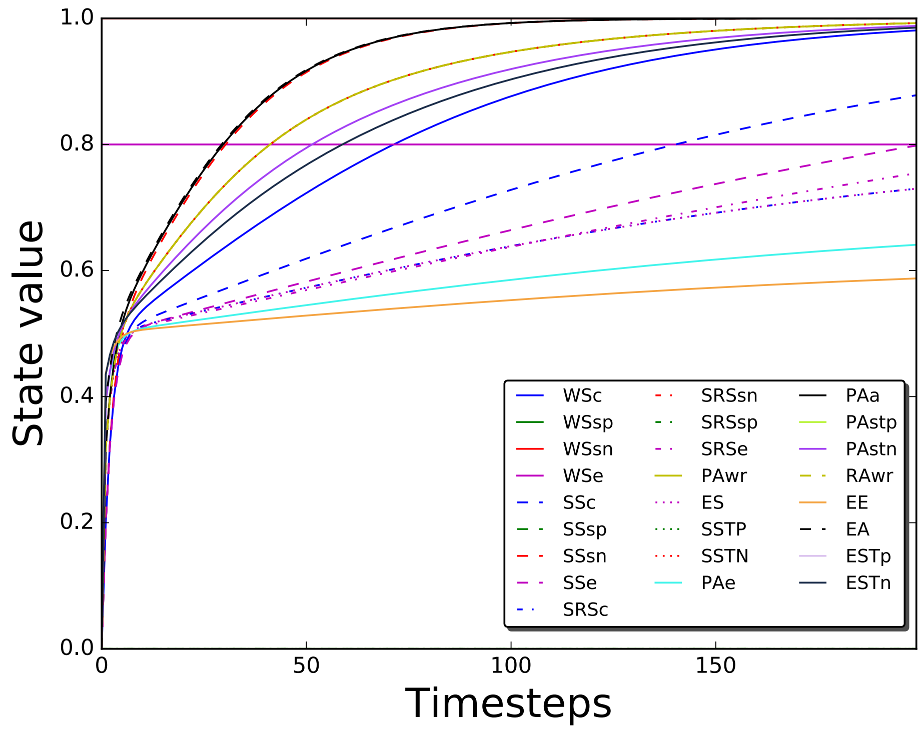


Fig. 9.2: Experimental dataset: Easily influenced person

The second experimental dataset defines the course of all 25 states for a person that receives a message through a messenger service. The observed emotion of the sender is happy and positive. This person is influenced by the emotion and sentiment of the message, however, believes that the message is not accurate enough to spread. In figure 9.3 the course of the second dataset is shown. It can be seen that this person's emotion and sentiment also approach the observed emotion and sentiment. The reason for this is that the person believes that the sender is happy and has a positive sentiment and is, therefore, influenced by the happy and positive sender. Because the value of the expected action is below 0.5 and this person believes that the message is not accurate enough, it is decided that this person will not revert to the action of spreading the message.

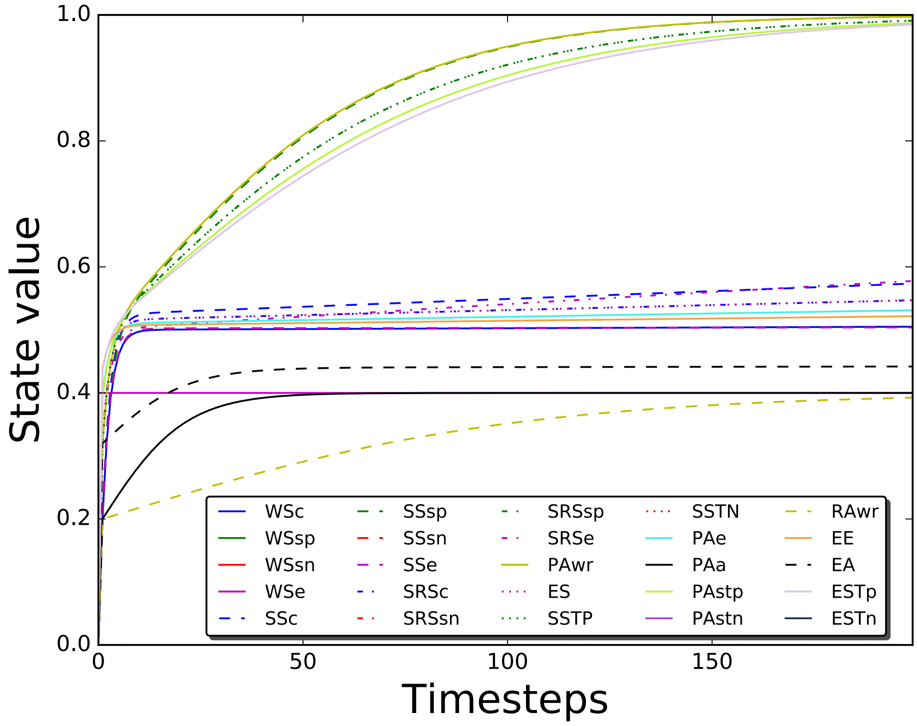


Fig. 9.3: Experimental dataset: Difficult to influence person

9.4.2 Parameter tuning methods

To tune the model, optimal values for the speed factor values η, the steepness σ and the threshold τ in the advanced logistic formulas should be found. For reasons of simplicity, we have decided to use a speed factor value of 0.4 for each state, because we assume that the changes of the states within a person to deal with receiving a disastrous message does not occur quickly, however, not too slowly either.

The steepness and threshold of the advanced logistic function is more difficult to manually define. In this model there are 14 states that use the advanced logistic function. Thus, in total there are 28 steepness and threshold values that need to be tuned. The domain of the steepness and threshold values is assumed to be $[-\inf, \inf]$.

This is another reason why manually tuning the parameters is difficult. We chose the mean squared error (MSE) as the objective function that has to be minimized.

We used random search to tune the parameters. To make parameter tuning with this method more tractable, we tuned 2 parameters instead of 28, and we searched within a uniformly distributed domain of $[-e^{-5}, e^{5}]$. This method provided errors of approximately 0.10 on the first dataset and of approximately 0.09 on the second dataset. When using the corresponding steepness and threshold values (see Table 9.2) in the model, this gave us an acceptable simulation.

Tab. 9.2: The resulting MSE, steepness and threshold after tuning.

	MSE	σ	τ
Dataset 1	0.103	5.859	-5.945
Dataset 2	0.094	0.847	-5.565

Tab. 9.3: Initial values of the input states of scenarios 1 and 2

Input states	Initial values scenario 1	Initial values scenario 2
WS_c	0.0	0.0
WS_{sp}	0.0	1.0
WS_{sn}	1.0	0.0
WS_e	0.8	0.4

Then, we performed random search with the modeling choice of tuning 28 parameters. This gave us an error of approximately 0.08, which is lower than tuning 2 parameters. However, when using the corresponding steepness and threshold values in the model, we got an abnormal simulation. We tried to decrease the domain to $[-e^{-2}, e^2]$. However, decreasing the domain did not make a difference.

9.5 Simulations and Results

In this section the simulation results are given. Three scenarios are simulated. For all simulations the steepness and threshold values from table 9.2 are used for each of the 14 states that use the advanced logistic function. We used a step size of $\Delta t = 0.1$ for all simulations, and the speed factor value η is 0.4.

The value of all connection weights are 1.0, except for the connection weights of $(SRS_c, SSTP)$, $(SRS_c, SSTN)$, $(PAwr, SSTP)$, $(PAwr, SSTN)$, (PA_a, PA_{stp}), (PA_a, PA_{stn}). The values of these weights are 0.01, because $\omega_{SRS_{sp} \to SSTP}$ and $\omega_{SRS_{sn} \to SSTN}$ should weigh the heaviest to calculate the state values of SSTN and SSTP.

9.5.1 Scenario 1: Receiving a tensed and negative phone call

In this first scenario, a person receives a phone call from another person informing then that a disastrous situation might occur. The person observes that the sentiment of the message is negative, and that the emotion of the message is tensed. This person is easily influenced by the sentiment and emotion of the message, also because he/she received a telephone call, which is assumed to be a credible source. This person, therefore, believes more easily that the message is true.

The initial values of the input states can be found in table 9.3. We have defined a tensed emotion to be 0.8 and a happy emotion to be 0.4. This is based on the valence and arousal model of Valenza et al. [10]. In figure 9.4 all states are depicted for this person. In figure 9.5 only the input and output states are depicted.

The output states EST_n, EA_a, EE_e and input state WS_c go up to around 0.9 due to the message's effect at the person. Since an action (i.e. spreading a message about a disastrous situation) increases awareness about the danger, the WS_c value increases.

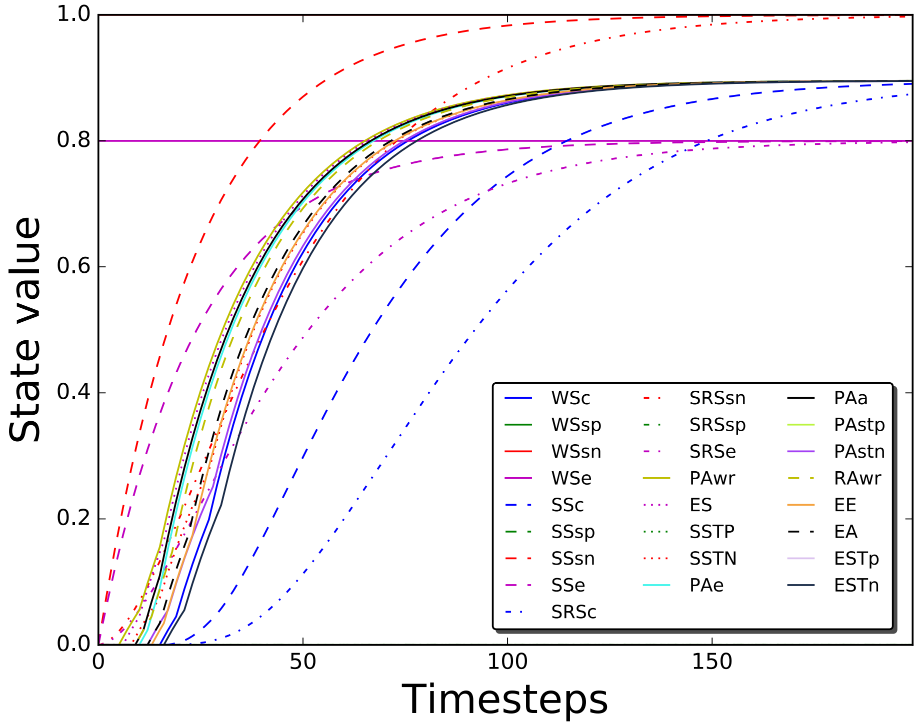


Fig. 9.4: Scenario 1: all states

9.5.2 Scenario 2: Receiving a happy and positive text message

In this scenario a person receives a message through a messenger service (textual) about a disastrous situation that might occur. This person, however, observes that the message has a positive sentiment and a happy emotion. This person is influenced by the message, however, does not believe that it is true enough to spread the message as it is more difficult to determine the sentiment and emotion of a textual message.

The initial values of the input states can be found in table 9.3. In figure 9.6 the simulation for this person is shown with all states. In figure 9.7 only the input and output states are shown for this person.

It can be seen that the EST_n is 0 over time, as the observed sentiment (WS_{sn}, WS_{sp}) of the message was positive. However, the EST_p state does not approximate to 1 throughout the simulation. This was expected since the connection weights that are important to define the EST_p are all 1.0. As expected, the E_e is higher than the WS_e, because this person observed a happy emotion and is influenced by it. The E_a is around 0.45. It is assumed that this person will not take action to spread the message, because this value is below 0.5.

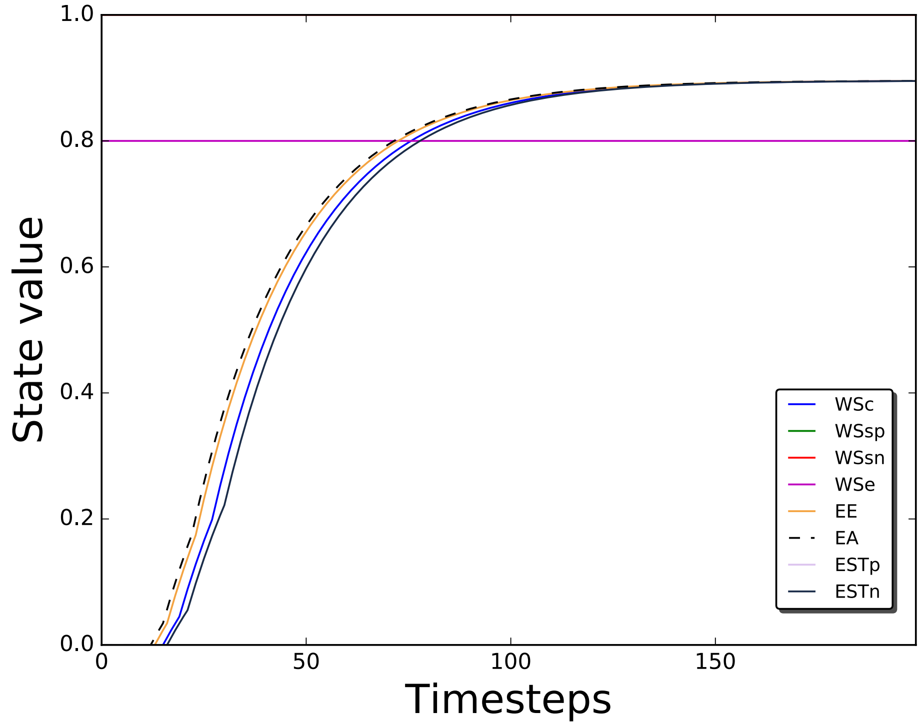


Fig. 9.5: Scenario 1: input and output states

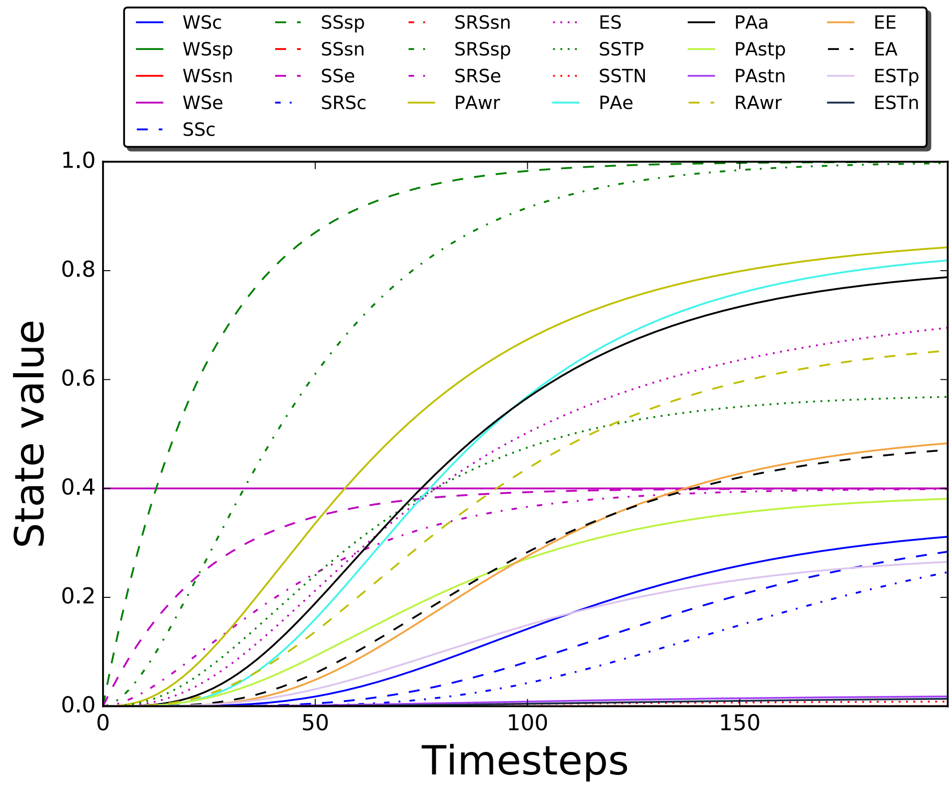


Fig. 9.6: Scenario 2: all states

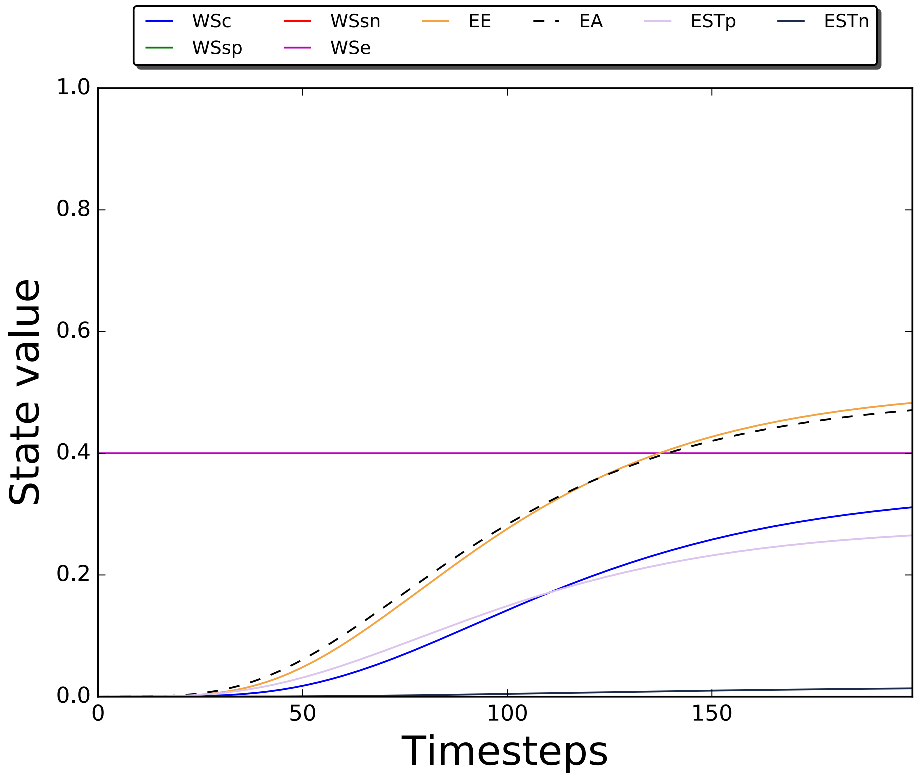


Fig. 9.7: Scenario 2: input and output states

9.5.3 Scenario 3: Outputs as inputs

In this scenario the person from scenario 1 receives a telephone call informing that a disaster might occur and takes action by sending a textual message to the person from scenario 2.

Tab. 9.4: Initial values of the input states of the two persons in scenario 3

Input states	Initial values person 1	Initial values person 2
WS_c	0.0	person1 E_a
WS_{sp}	0.0	person1 EST_{sp}
WS_{sn}	1.0	person1 EST_{sn}
WS_e	0.8	person1 E_e

The initial values can be found in table 9.4. In figure 9.8 the outputs of the first person and the outputs of the second person are shown. The outputs of the first person are used as inputs for the second person.

It can be seen that person 1 has a high EST_{sn}, and that person 2 also gets an EST_{sn} because of person 1. However, the EST_{sn} of person 2 is much lower than that of person 1. It is also shown that person 1 has a tensed emotion (i.e. $E_e = 0.8$), however, person 2 has a rather happy emotion (i.e. $E_e = 0.4$). It can, thus, be seen that person 2 is not influenced easily by the emotional state of person 1 as expected. The E_a of person 2 approximates a value of 0.4. Therefore, it is assumed that person 2 does not spread the message further.

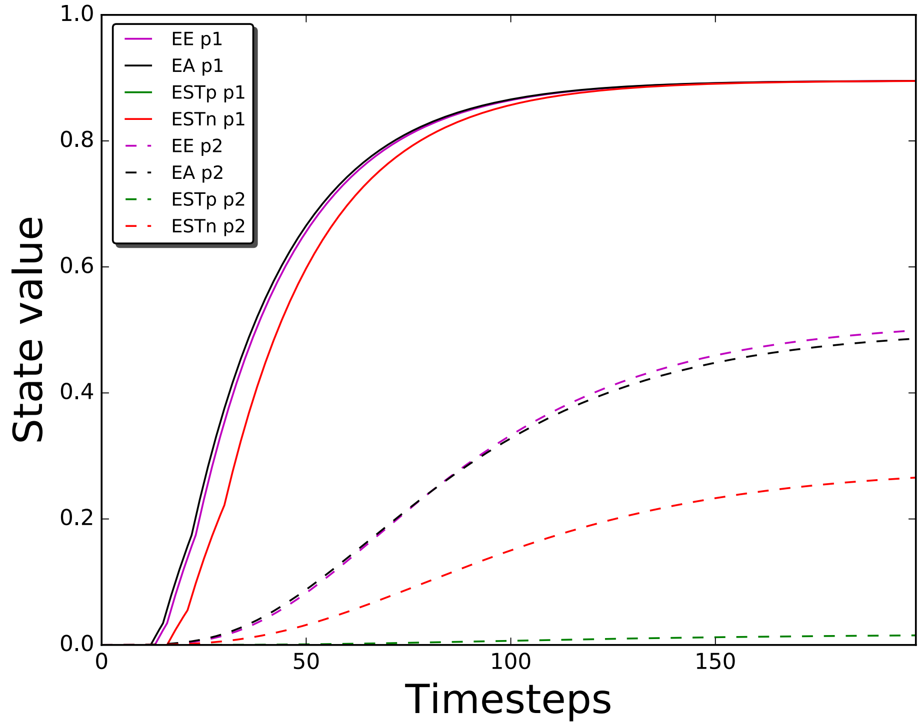


Fig. 9.8: Scenario 3: Outputs of person 1 as inputs for person 2

9.5.4 Mathematical analysis of the model

To determine when the model is in equilibrium, we check when the states reach their stationary points. For instance, a state Y has a stationary point if $\vec{d}Y(t)/\vec{d}(t) = 0$. The model is in equilibrium if every state has a stationary point at certain time t. Taking into account the difference and differential equations used in the model, the stationary point equation can be written as:

$$Y(t) = \vec{c}_Y(\omega_{X_1,Y}X_1(t), ..., \omega_{X_k,Y}X_k(t)) \quad (9.5)$$

As an example we are going to determine the stationary points for the states SS_c, SS_{sn}, and SS_e. The verification method is the substitution of values in the stationary point equations. To determine the stationary points, the person from the simulation in scenario 1 (figure 9.4) is used, however, with a longer simulation time and with a $\Delta t = 0.05$ (see figure 9.9).

The model was run until $Y(t + \Delta t) = Y$ holds. A stationary point for state SS_c was found at time point 531, with state value 0.8952. For state SS_e the stationary point is at time step 366 (state value of 0.7995). A stationary point for state SS_{sn} was found at time point 377 with state value 0.9995.

The connection states of SS_c, SS_e and SS_{sn} are respectively WS_c, WS_e and WS_{sn}. The state values of these connection states at the time points 531, 366 and 377

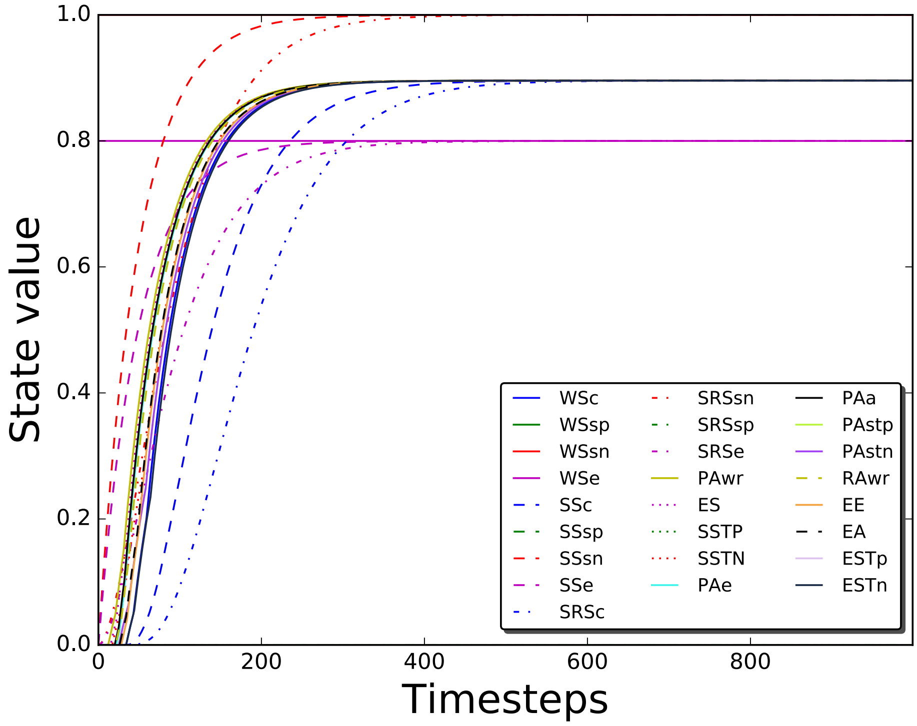


Fig. 9.9: Simulation for mathematical analysis

are, respectively, 0.8957, 0.8 and 1.0. The connection weights are are all 1.0. When substituting these values in equation 9.5 we can see that the equation holds with an accuracy of $< 10^{-2}$:

We found the stationary points for all states in the model. When taking into account that every state has to be stationary at time point t for the model to be in equilibrium, we can observe that the model is in equilibrium at time point 531 for the proposed set up.

9.6 Discussions and future works

In this paper a computational model is presented in order to model people's behaviour on spreading messages in disaster situations. The model was designed as a temporal causal network model, following the approach of Treur [9], moreover, inspired and based on findings from previous research, as discussed in the background information, section 2.

The proposed model can be a base for any type of disaster situation and can easily be extended in future research. Validation of this model is very difficult, if even possible, due to the lack of empirical data, because often observations during disaster situations are missing. Therefore, experimental data was created based on literature and experience to perform parameter tuning.

Within the scope of this paper we decided to incorporate different types of communication methods through the simulated scenarios. It could be interesting however to take those communication methods to the next level as well in order to learn why particular methods are more credible than others, or how messages spread more easily through some channels than others. Personality traits are somewhat incorporated in the simulation scenarios as well, however, for future research more traits should be explored. The same goes for culture and other possible influences.

Bibliography

[1] S Blake, E Galea, H Westeng, and AJP Dixon. „An analysis of human behaviour during the WTC disaster of 11 September 2001 based on published survivor accounts“. In: *Proceedings of the 3rd international symposium on human behaviour in fire*. Vol. 1. 2004, p. 3 (cit. on p. 148).

[2] Tibor Bosse, Mark Hoogendoorn, Michel CA Klein, et al. „Modelling collective decision making in groups and crowds: Integrating social contagion and interacting emotions, beliefs and intentions“. In: *Autonomous Agents and Multi-Agent Systems* 27.1 (2013), pp. 52–84 (cit. on pp. 149, 153).

[3] Paul Ekman. „Basic Emotions“. In: *Handbook of Cognition and Emtion* (1999) (cit. on p. 150).

[4] Alexander Hall. „The North Sea Flood of 1953“. In: *Arcadia* 5 (2013) (cit. on p. 148).

[5] Douglas Paton. „Disaster preparedness: a social-cognitive perspective“. In: *Disaster Prevention and Management: An International Journal* 12.3 (2003), pp. 210–216 (cit. on pp. 149, 150).

[6] Wilfred Takken. „Wat doen we als de Randstad onder water staat?“ In: *NRC handelsblad* (2016) (cit. on p. 148).

[7] Dilhan J. Thilakarathne and Jan Treur. „Modelling Prior and Retrospective Awareness of Actions“. In: (2013), pp. 62–73 (cit. on p. 149).

[8] Dilhan J. Thilakarathne and Jan Treur. „Modelling the Dynamics of Emotional Awareness“. In: ECAI'14 (2014), pp. 885–890 (cit. on p. 149).

[9] Jan Treur. „Network-Oriented Modeling and Its Conceptual Foundations“. In: *Network-Oriented Modeling: Addressing Complexity of Cognitive, Affective and Social Interactions*. Cham: Springer International Publishing, 2016, pp. 3–33 (cit. on pp. 149, 151, 160).

[10] Gaetano Valenza, Antonio Lanata, and Enzo Pasquale Scilingo. „The role of nonlinear dynamics in affective valence and arousal recognition“. In: *IEEE transactions on affective computing* 3.2 (2012), pp. 237–249 (cit. on p. 155).

10 A Computational Cognitive Model for Political Positioning and Reactions in Web Media[1]

> *"What you see and what you hear depends a great deal on where you are standing. It also depends on what sort of person you are."*
>
> — **C.S. Lewis**
> "The Magician's Nephew"[2]

Abstract

This paper presents a computational cognitive model about political positioning inspired on recent insights from neuroscience and psychology. We describe a model that takes into consideration the individual structures of the brain and the environmental influences that may interfere on how a political positioning is shaped in a social media environment. Results also consider the social contagion effect. We use the model to provide simulations of how the political positioning of artificial agents change under different circumstances.

[1]Published as: **E. F. de Mello Araújo** and M. Klein, *A computational cognitive model for political positioning and reactions in web media*, 2017 IEEE 16th International Conference on Cognitive Informatics & Cognitive Computing (ICCI*CC), Oxford, 2017, pp. 414-422.

10.1 Introduction

Politics and social media are currently strongly related and the relation became very important in many fields, including political sciences, sociology, psychology and more recently neuroscience and computer science. Neuroscience is interested in understanding how to correlate brain structures with political positioning. Computer science provides tools and insight for doing polls, making predictions based on web media crawling and data mining, and using web media for political campaigns.

Despite the fact that the majority of the population is not directly involved in politics, the political positioning of the population and the empowerment given by social media, with its usage as a campaign platform by the candidates in recent elections around the globe illustrate that more attention should be paid to this subject.

Not only is it important to understand how media influences the political positioning, it is also relevant to understand possible side effects of political campaigns that use web media to influence voters and to advertise their ideas and proposals for society.

Over the years, many researchers have tried to allocate people in political boxes, besides all the various questionnaires and tools created to evaluate people's political mindsets. Neuroscience has also brought up new discoveries that challenge the traditional concepts and theories developed until a decade ago.

We believe that it is of great interest for science to see the consequences of these discoveries on a large scale, using computational simulations to predict or mimic reality and reveal whose outcomes would be expected for processes of reasoning around politics.

In this paper we present a model that describes the political positioning of an individual based on recent discoveries in psychology and neuroscience. Our cognitive model considers the prior structure of the brain and the stable personality traits and the possible influences that might change (or reinforce) his political positioning while interacting with social media (i.e. reading messages in a social web platform). This model can be applied in many applications, like polls simulations, support systems for political platforms and for agent-based modeling.

In the remainder of this paper, we first discuss the background on the (neuro-) psychological aspects of political behaviour. Second, we discuss the literature on the effect of mood and persuasion on the interpretation of messages. Then, we describe our model and show simulations of scenario's with the model. Finally, we discuss the implication of our findings and the possible applications.

10.2 Psychological aspects of Political Behaviour

Political attitudes and voting behaviour are well explored topics for a long time, with many works in psychology, social sciences and (more recently) neuroscience [4, 5, 7, 14]. Many studies at different moments in history have different perspectives about the differences between conservatives and liberals (or progressists), from the 1940s to our current society. The main goal of most of the research is related to find personality traits that correlate to right-left (or conservative-liberal) voting choices or attitudes.

10.2.1 Political Psychology

Caprara et al. [4] defines personality as *a set of dynamic, self-regulatory systems that emerge and operate over the life course in the service of personal adaptations*. These internal systems will define how affective, cognitive, and motivational processes run, taking into account the individual and collective goals of the person. To preserve the sense of identity, there should be coherence and continuity in behavioural patterns across these settings [3].

Caprara et al. [4] explain how traits and values, two personality aspects, define political orientation towards the right-left (or liberal-conservative) spectrum, concluding that values have a stronger matching with the political orientation than the traits. Traits are enduring dispositions, whereas values are enduring goals. Traits reveal what people are like, while values are connected to what people consider important, and relate to standards for judging behaviour, events and people.

The values play an important role on the self-regulation mechanism of people. People may change their behaviour in order to align their actions with their values and reduce discrepancies. On the other hand, traits may also affect values, especially those that converge to the trait itself, justifying behaviour and contributing to the self-regulatory mechanism of the individual [4].

Carney et al. [5] study shows that two traits from the Big Five, openness to new experiences and conscientiousness, capture many of the ways people define their political orientation. Conscientiousness is related to how careful or vigilant a person is. It is related to efficiency and organization, self-discipline, planned behaviour (instead of spontaneous). It is connected also to behaviours like being neat and systematic. Openness to new experiences, in the other hand, involves six dimensions, including active imagination (fantasy), aesthetic sensitivity, attentiveness to inner feelings, preference for variety, and intellectual curiosity, and can be treated as a combination of these characteristics. People with this trait tend to be non-conventional and contemporary about their actions and appearance.

Carney et al. [5] assume that two core aspects define the position right-left: (a) acceptance *versus* rejection of inequality and (b) preference for social change *versus* preservation of the societal status quo.

Carney et al. [5] presents three different experiments to show how personality relates to the political orientation or people. The first experiment is an intake questionnaire that tries to relate the Big Five with the political positioning, and the other two experiments investigate non-verbal clues as the behaviour during an interview and the way people organize their spaces at work and at home. He correlates gestures, halting speech, smiling behaviour, tidiness, decoration and accessories of the ambient where people leave (among many other aspects) to their position toward politics.

In all experiments there is a potential support to other observations over the past decades [1, 9, 10, 13, 19, 27, 31, 35] where personality differences covary with political orientation.

Caprara et al. [4] used the FFM (Five Factor Model), a shorter version of the Big Five factors for personality to track the traits, and the Portrait Values Questionnaire (PVQ) [25] for the ten value constructs (power, achievement, hedonism, stimulation, self-direction, universalism, benevolence, tradition, conformity and security). This theory has been tested in more than 200 samples and 67 countries, and the results can vary according to each country's reality. Political voting follows the values sustained by the persons. Parties that threaten someone's values usually tends to lose their vote.

From the ten values studied by Caprara et al. [4], four predicted well votes for center-left or center-right coalitions in Italy: (1) universalism; (2) security; (3) tradition; and (4) conformity. In general, voters who give high priority to self-direction (against tradition) and universalism but low priority to power and security tend to vote for parties that stand for individual freedom and social programs to help eliminate poverty, and not for parties that emphasize nationalism and the need for law and order.

The friendly and open aspects of personality are more related to center-left positions. People that are more energetic and dominant have higher tendency to assimilate the programs from center-right that includes individual entrepreneurship and business freedom. That means that self-confidence and assertiveness are traits of a candidate that is more attractive to center-right people.

In this way, there is a consensus that politicians that emphasize a commitment to social justice (universalism) or family values (tradition and security) appeal to basic values that shape people's attitudes regarding ideological issues, and therefore are good predictors of people's voting behaviour.

To bring all these works to an unified notation for political psychology, Carney et al. [5] proposes a framework for interpreting and integrating the body of theories and findings. Using the Big Five questionnaire and correlating the aspects of this range of personality traits, it well accepted that conservatives present a higher score in *Conscientiousness* (e.g. definite, persistent, self-controlled, moralistic, etc.), while liberals score higher in *Openness to experience* (including tolerance, expressiveness, open-minded, etc.). Other traits from the Big Five (Extraversion, Agreeableness and Neuroticism) don't present strong evidences to match political positioning or attitudes.

Most recent theories have an agreement that fit an uncertainty-threat model of political orientation, as stated by Jost et al. [13]:

> We regard political conservatism as an ideological belief system that is significantly (but not completely) related to motivational concerns having to do with the psychological management of uncertainty and fear. Specifically, the avoidance of uncertainty (and the striving for certainty) may be particularly tied to one core dimension of conservative thought,

> resistance to change... Similarly, concerns with fear and threat may be linked to the second core dimension of conservatism, endorsement of inequality... Although resistance to change and support for inequality are conceptually distinguishable, we have argued that they are psychologically interrelated, in part because motives pertaining to uncertainty and threat are interrelated.

10.2.2 Political Neuroscience

The political psychology is much more developed than the neuroscience of politics. Nevertheless, the emergence of neuroscience brought refreshment for the studies in politics towards behaviour and attitudes, helping to reinforce some prior discoveries from psychology and shining more light to the understanding of the phenomenons, mainly regarding racial prejudice, partisan bias and motivated political cognition, nature of left-right differences in political orientations and the dimensional structure of political attitudes [14].

The contribution from neuroscience is still questionable for many scientists, mainly because most part of the works are interested in finding correlations between behaviour or attitudes and areas that are activated in the brain. In these cases, the most announced warning prevails: correlation does not imply causation. As the brain is a very complex structure where many processes happen simultaneously, scientists are warned to be careful about their statements [12, 29].

Even thou, many studies have shown that two parts of the brain are important on defining traits of people [2, 12, 24, 29, 30]. The anterior cingulate cortex (ACC) is the part of the brain where error detection, conflict monitoring, and evaluating or weighting different competing choices happen [16]. This part of the brain filters the possible variables in a decision making process giving different levels of importance to each of them. A bad functioning of the ACC leads to schizophrenia, for instance. It is also known that strong and overwhelming emotions can disturb your capacity of evaluating and making decisions. This is basically seen when you try to make an exam right after receiving some bad news about a relative that just passed away.

For the politics, the ACC tells us a lot about the capacity of putting emotions aside and making decisions in complex environments. The development of the ACC helps people to solve complex problems, and to be more resilient minds towards difficult situations. Those people are able to adapt better to changes, and are related to liberal positioning in the literature [4].

The amygdala is a part of the brain related to empathy and emotions. It holds the emotion, learning and memory formation processes, and there is evidence that amygdala's volume is positively correlated with dispositional fearfulness [23]. A hyperactive amygdala also causes illness, like Borderline Personality Disorder. A small amygdala (lack of emotions) is also connected to psychopathy. Kanai et al. [16] observed that conservatism (measured in terms of ideological self-placement) was positively associated with larger right-amygdala volume. Oxley et al. [20] findings show that conservatism is associated with greater physiological startle response to threatening stimuli.

Other aspect studied is the system justification. This theory holds that the need to keep the status quo is a psychological trait of someone who wants to attain certainty and order [15]. As the amygdala is associated with reactions towards threats and uncertainty [33], it was hypothesized that system-justification trends could be associated with greater amygdala volume, as could be observed in a sample with 48 U.S. college students.

10.3 The Effect of Mood and Persuasion on Interpretation of Messages

Besides the traits and anatomy of the brain aspects, it is also important to understand how people are influenced and what effect social contagion has on people's opinions, sentiments and behaviours.

Social and behavioural contagion is the propensity of someone to copy or mimic other person's behaviour, opinion or even sentiment. This can be caused by exposure to media coverage about some specific and famous actor, or even by just having a relationship with someone [6, 28, 34]. Social psychologists study the factors behind this phenomenon, and many works [17, 32, 36] have acknowledge the affective effect on people exchanging experiences or opinions throughout different sorts of models.

It is also known that the mood affects the way influences are perceived, in other words, how persuasion is affected by the mood of a person. Schwarz et al. [26] present a vast number of works where participants' mood is elated in order to verify if people's engagement in extensive and detailed processing of arguments is affected or not. The arguments used can be videos with campaigns, texts or even some conversation in order to convince a person to donate money or sign a petition.

Based on many theoretical works about cognitive responses that leads to persuasion and potential attitude change [11, 22], the main conclusion of Schwarz et al. [26] is that the process of interpreting persuasive messages can be explained in two parts. First, people may pay more attention to the content and the implications of the arguments within it, or in a content-oriented approach called "systematic processing". Second, receivers do engage with a depicted evaluation of the message, but rely on the cues that are present in the message, like the messenger, key words, etc. They call this approach "heuristic processing". The former is a central route while the latter is a peripheral route to evaluate the content of a message and process it.

These results can help us to understand the role of mood in the process of interacting with arguments, which often happens in social media and political debates.

- individuals in an elated mood are less likely to engage in extensive processing of the presented arguments than people in a mildly depressed or unaltered mood;

- providing time to processing of the message allows the person to overcome the effects of mood [18];
- individuals in a bad or good mood are both influential, but people in a bad mood are less likely to be convinced by poor arguments, or low quality attempts to persuasion;
- recipients are not able to differentiate good and bad messages or arguments when they were in good mood by the time of encoding the message, but were able to differentiate the quality of the messages in a neutral mood;
- attitude change via the central route is more stable than attitude change through peripheral route [21].

It is also important to highlight the bidirectional process of emotions / mood affecting the cognitive system of message interpretation and vice versa. According to Frijda [8], "events that satisfy the individual's goals, or promise to do so, yield positive emotions; events that harm or threaten the individual's concerns lead to negative emotions". This relation between emotions and cognitive states is implemented in our model explained forward, when messages that diverge from the person's opinion will affect his or her mood state, while the mood state also affects the interpretation of the message and ultimately the feeling for the reception of the influence by the message.

10.4 Agent-based model for political positioning on web media

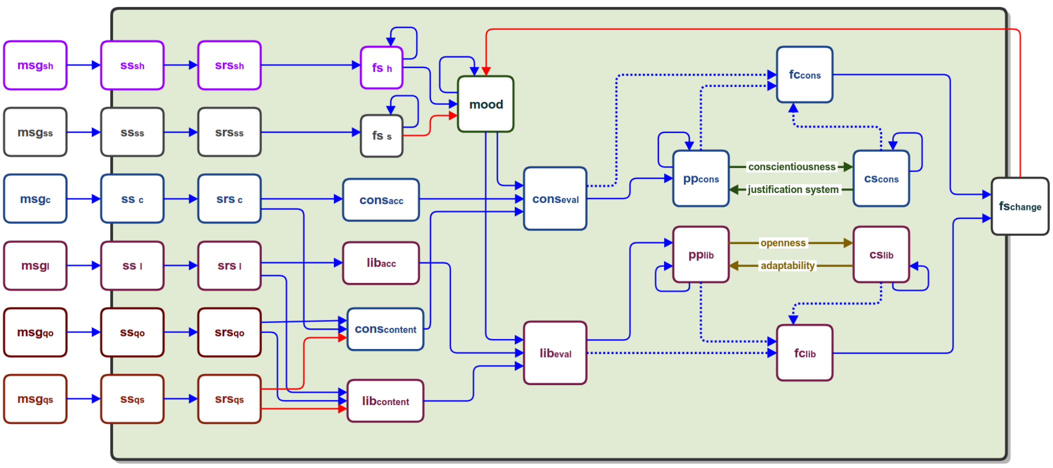


Fig. 10.1: Computational model for political changes and reactions towards web media messages.

In this section we present our model based on the literature explored in the previous sections, and shown in Figure 10.1.

We use a temporal-causal modeling approach based on time series where the future states are dependent on the past states [32]. Our model uses cognitive states to describe the process of spread of information. Technically, the model is built upon mathematical equations of type 10.1, where η_Y is considered a speed factor and Δt is the time step size for each interaction over state Y.

$$Y(t + \Delta t) = Y(t) + \eta_Y [c_Y(\omega_{X_1,Y} X_1(t), \ldots, \omega_{X_k,Y} X_k(t)) - Y(t)]\Delta t \tag{10.1}$$

The combination function $c_Y(...)$ is calculated as the result of the impact of all the neighbors of the state Y that have some connection towards it. It is used a different cumulative function combined with specific calculations for each state.

The agent is influenced through web media messages that are received and interpreted. Each message contains 6 different characteristics related which are evaluated by the cognitive model: (a) positive sentiment (msg_{sh}) and negative sentiment (msg_{ss}), both ranging from 0 to 1; (b) conservative political position (msg_c) and liberal political position (msg_l), both ranging from 0 to 1; and (c) quality, which refers to the objectiveness (msg_{qo}) and subjectiveness (msg_{qs}) of the message, also ranging from 0 to 1. We consider that objective messages present better quality on its content.

The agent can react by supporting or rejecting the message as a result of how comfortable the changes and towards which direction the political change is dragged because of the message. This reactions are shown in the fs_{change} state of the model, and range from 0 to 1, where 0 means that the agent agrees with the message, while 1 represents complete disagreement with the message. This state also elates the mood of the agent, as sustained by previous theory about persuasion and mood [26].

There is a limited time of exposition to each message, which is counted in time steps in this model. In our simulations the agent is exposed for 20 time steps per message. A sequence of messages is given in order so the changes can be observed.

As explained in the previous section, the mood influences the way the message is interpreted. People in a good mood tends to be less analytic and more guided by the cues. The states $cons_{acc}$ and $cons_{content}$, as well as lib_{acc} and $lib_{content}$ are the part of the process of persuasion that will receive influence from the $mood$ state and define the inner interpretation of the message by the individual.

States $cons_{content}$ and $lib_{content}$ refer to the central flow (or route) of evaluation of the content of the message, while lib_{acc} and $cons_{acc}$ are related to the peripheral system. The $mood$ is the tuner for the level of criticism of the agent, which is responsible to define if the agent will follow the two phases (neutral to bad mood), or relax on the phase of content processing. The judging phase bases its evaluation on cues of the message, like the messenger and the keywords of the text, while the processing phase performs a more detailed analysis of the content, considering how strong is the argument through the quality of the message evaluation (srs_{qo} and srs_{qs}).

The $mood$ state is influenced by the sentiment of the message received, and also by fs_{change}, which is the result of the differences between the agent's political opinion and the interpreted message. A loop is created in order to keep the previous mood influencing the next time step for state maintenance. The same holds for fs_h and fs_s.

For the changes in the political position of the agent, the theories presented before based on the self-regulatory system [4] are represented in the relation between the political position states (pp_{cons} and pp_{lib}) and the coherence system states (cs_{cons} and cs_{lib}). These relations are responsible for the traits of the agent and its original tendency towards a specific political position. That means that if the messages affect the agent towards a very far political position from its original position, the agent will have a mechanism that drags its opinion back and try to keep the agent coherent and comfortable with itself through the coherence system. The political positioning states represent the liberalness and conservativeness of the agent.

The reactions of the agent are measured according to the differences between the message and the agent's own positioning. For that, fc_{cons} and fc_{lib} will give a value between 0 and 1 showing how far the message is from the conservative and liberal positions of the agent, and fs_{change} will catch the two information and responds with the overall feeling for change for the agent receiving the message.

Table 10.1 shows the explanation for each node in the model.

The connections and the cumulative functions are shown in Appendix A[3]. Most part of the connections use the identity (id) function or advanced logistic ($alogistic$), as will be explained later.

As shown in section 2, some traits are correlated to political positioning on the scale between conservative or liberal. In our model, we used four traits to define how the coherence systems (cs_{cons} and cs_{lib}) and political opinion states (pp_{cons} and pp_{lib}) affect each other: openness, conscientiousness [5], justification system and adaptability [15]. Those traits are responsible to define how open the agent is for changes, or how rigid it is to change its mind. So for a conservative agent, we assume that the justification system will influence it more, while for liberals, the openness is a big factor that will open it for changes.

The other concepts presented before are integrated to the four we chose to include in this model, due to similarities in the concepts and in the correlation between neuro-scientific and psychological discoveries. This is the case, for instance, for the concepts of conformity [4], preference for social change [5] and sensitivity for threatening [15].

Figure 10.1 shows how these four traits represent four connections between political positioning and coherence systems for the model. Those values vary according to the characteristics of the agent implemented, and as will be seen later on the simulations, and are constant over time.

[3] `http://www.cs.vu.nl/~efo600/iccicc17/AppendixA.pdf`

State name	Information about the node	Range
msg_{sh} and msg_{ss}	Positive and negative sentiment of the message	$[0,1]$
msg_c and msg_l	Conservative and liberal political positioning of the message	$[0,1]$
msg_{qo} and msg_{qs}	Objectiveness (high quality) and subjectiveness (low quality) of the message	$[0,1]$
ss_{sh} and ss_{ss}	Sensor state for positive and negative sentiment of the message	$[0,1]$
ss_c and ss_l	Sensor state for conservative and liberal political position of the message	$[0,1]$
ss_{qo} and ss_{qs}	Sensor state for quality/subjectivity of the message	$[0,1]$
srs_{sh} and srs_{ss}	Sensory representation for positive and negative sentiments of the message	$[0,1]$
srs_c and srs_l	Sensory representation for conservative and liberal political position of the message	$[0,1]$
srs_{qo} and srs_{qs}	Sensory representation for quality/subjectivity of the message	$[0,1]$
$mood$	Emotion state for the agent	$[0,1]$
$cons_{acc}$ and lib_{acc}	Message judging phase of the argument interpretation system	$[0,1]$
$cons_{content}$ and $lib_{content}$	Message processing of content phase of the argument interpretation system	$[0,1]$
$cons_{eval}$ and lib_{eval}	Processing of the information from the two phases of message interpretation	[0,1]
fc_{cons} and fc_{lib}	Feeling sensor of the conservative and liberal political opinion change	$[0,1]$
fs_{change}	Feeling sensor of the overall political opinion change	$[0,1]$
pp_{cons} and pp_{lib}	Conservative and liberal political position of the agent	$[0,1]$
cs_{cons} and cs_{lib}	Trait from coherence system for agent's conservative and liberal political positions	$[0,1]$

Tab. 10.1: States in the model

The edges in red have a negative effect over the affected state, while the blue connections have a positive effect. The combination function for states using the advanced logistic function are calculate as shown in equation 10.2.

$V_1, ..., V_k$ are all the products of the states of the neighbors by the weight of the link between them and the current state.

$$\begin{aligned} c(V_1, ..., V_k) = \textbf{alogistic}_{\sigma}, \tau(V_1, ..., V_k) = \\ = [1/(1+e^{-\sigma(V_1+...+V_k-\tau)}) - (1/(1+e^{\sigma\tau}))](1+e^{-\sigma\tau}) \end{aligned} \tag{10.2}$$

For $id(.)$ functions that have a negative weight, we calculate the next state as equation 10.3, where $A(t)$ is the previous state status for time t and $\omega_{a,Y}$ is the strength of the connection between states A and Y.

$$Y(t+\Delta t) = Y(t) + \eta_Y[(1 - \textbf{abs}(\omega_{A,Y})A(t)) - Y(t)]\Delta t \tag{10.3}$$

The $mood$ state is the main actor on the simulation of the peripheral and central routes of interpretation of the message. If the value for mood is higher (better mood), the $eval$ states will value more the $content$ (central route) and less the acc (peripheral route). This is simulated for $cons_{eval}$ and lib_{eval} as shown in equation 10.4.

$$\begin{aligned} Y(t+\Delta t) &= Y(t) + \eta_Y[c_Y(...) - Y(t)]\Delta t \\ c_Y &= \textbf{alogistic}(mood(t)(Y_{content}(t)) + (1 - mood(t))(Y_{acc}(t))] \end{aligned} \tag{10.4}$$

The values for the political position states for the next time step (pp_{cons} and pp_{lib}) are calculated based on the coherence system (cs_{cons}, for the conservative position, and cs_{lib} for the liberal position) and on the evaluation of the message received ($cons_{eval}$ for the conservativeness of the message evaluated, and lib_{eval} for the liberalness of the message processed). The political position states and coherence system states have a self connection for persistence of the status.

A very important characteristic of our model is that political position and the coherence system can have a stronger or weaker relation based on the traits of the agent, defined by the factors justification system, conscientiousness, adaptability and openness, all explored in Section 2.

Aiming to identify messages that are strange to the agent, the feeling about the change state (fs_{change}) calculates the difference between the actual political position, the coherence system and the perceived political position of the message in order to identify whenever the message is in conflict with agent's mindset. The fs_{change} state is the result of two other states that calculate the differences for liberal and conservative positions, fc_{lib} and fc_{cons}, respectively.

Equation 10.5 shows the formula for the combination function of fc_{lib}, and can be replicated to fc_cons just by changing the states connected to fc_{lib} for the others connected to fc_{cons}.

$$\begin{aligned} c_{fc_{lib}} = \textbf{abs}[\omega_{lib_{eval},fc_{lib}} lib_{eval}(t) - \omega_{pp_{lib},fc_{lib}} pp_{lib}(t)] + \\ + \textbf{abs}[\omega_{cs_{lib},fc_{lib}} cs_{lib}(t) - \omega_{lib_{eval},fc_{lib}} lib_{eval}(t)] \end{aligned} \tag{10.5}$$

All states fs_{change}, fc_{cons} and fc_{lib} use advanced logistic function to calculate the next state value.

All the connections and their respective combination functions are shown in Appendix A[3].

10.5 Simulations

This section presents a number of simulations with the proposed model, showing the political position changes over time for agents with different traits.

Specifically, we built four different agents and verified if they follow the expected pattern observed in other works in literature presented before concerning social contagion [32]. The values for the traits of the agents are shown in table 10.2.

Traits / Values	**Agent 1**	**Agent 2**	**Agent 3**	**Agent 4**
Openness	0.9	0.1	0.5	0.5
Adaptability	0.9	0.1	0.5	0.5
Conscientiousness	0.1	0.9	0.5	0.5
System justification	0.1	0.9	0.5	0.5

Tab. 10.2: Traits of the simulated agents

Agent 1 is representative for a very liberal agent, with high openness and adaptability, while agent 2 is very conservative, with high values for conscientiousness and system justification. Agent 3 is moderate, and therefore does not have high values for any of the traits. Agent 4 is also balanced in its opinion, but presents higher values for conservativeness and liberalness. It can be understood as an agent with strong convictions that match both sides (right and left) or even a rigorous willing to find the best from both positions.

Each agent was initiated with states for political position, coherence system and mood, accordingly to table 10.3.

States / Values	**Agent 1**	**Agent 2**	**Agent 3**	**Agent 4**
pp_{cons}	0.10	0.75	0.10	0.60
pp_{lib}	0.80	0.05	0.10	0.50
cs_{cons}	0.20	0.80	0.15	0.55
cs_{lib}	0.70	0.10	0.20	0.52
$mood$	0.50	0.50	0.50	0.50

Tab. 10.3: Initial states for the simulated agents

As can be observed, agents 1 and 2 are opposites in their positions, while agent 3 is moderate and presents less extremist opinions. Agent 4 has similar characteristics as agent 3 but with stronger levels of political positioning. All 4 agents have an average mood of 0.5 to start with.

For each state that is calculated with an advanced logistic function a steepness (σ) and a threshold (τ) was chosen in order to get the expected behaviour from the agent. Also different speed factors were chosen for $mood$, pp_{cons}, pp_{lib}, cs_{cons} and cs_{lib} due to their very specific characteristics. The coherence system is a very stable part of the cognition, and therefore harder to change, while the mood is much changeable than the political position of the agent. For our simulations, we choose the speed for the coherence system to be 10^{-5}, the speed for the political position changes to be 5×10^{-4} and the speed for the mood changes as 10^{-3}.

To test the changes of the states in the different agents, we simulated that they received 9 different kinds of message in sequence. So each agent received a stream with 60 messages containing the same characteristics regarding political position, quality of the message and sentiment. The agent was exposed to each message for 20 time steps. The sorts of messages can be seen in table 10.4.

Message / Values	**Cons/Lib**	**Happy/Sad**	**Obj/Subj**
A	1.0/0.0	1.0/0.0	1.0/0.0
B	1.0/0.0	0.5/0.5	1.0/0.0
C	1.0/0.0	0.0/1.0	1.0/0.0
D	0.5/0.0	1.0/0.0	1.0/0.0
E	0.5/0.0	0.5/0.5	1.0/0.0
F	0.5/0.0	0.0/1.0	1.0/0.0
G	0.0/1.0	1.0/0.0	1.0/0.0
H	0.0/1.0	0.5/0.5	1.0/0.0
I	0.0/1.0	0.0/1.0	1.0/0.0

Tab. 10.4: Messages used for the simulations

Figure 10.2 shows the changes on political position and coherence system for a liberal agent which gets conservative messages in sequence. The steepness of the political positions are bigger than the coherence system, as expected. It is also possible to observe that the liberalism opinion decreases due to the messages and also because of the coherence system of the agent.

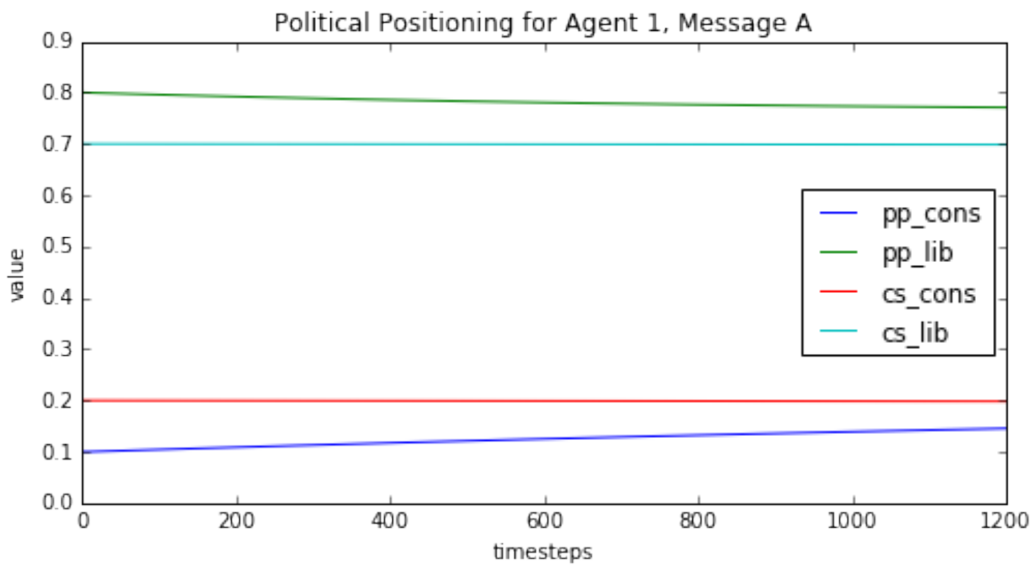


Fig. 10.2: Simulation of conservative message to liberal agent.

Mood is attached to the feeling of change, and it can be observed for the same liberal agent the mood being affected when it receives messages with neutral sentiment (messages type B) but that create conflicts with agent's opinion. Figure 10.3 shows the feeling for change and the reception of the sentiment of a neutral message.

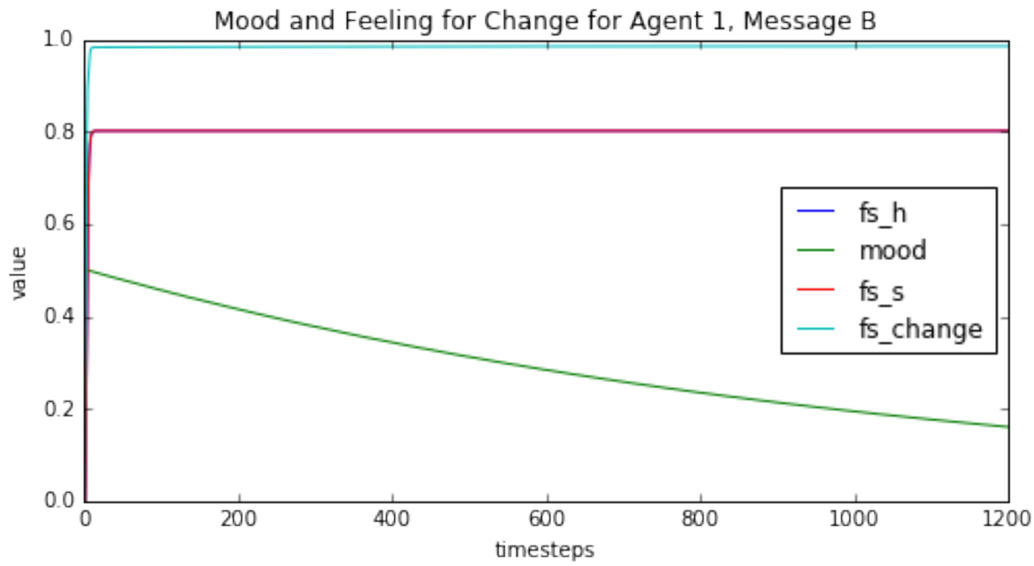


Fig. 10.3: Simulation of conservative message with neutral sentiment to liberal agent.

To understand how the mood affect the interpretation of the message, figure 10.4 shows the evaluation of the message for the same liberal agent while receiving messages of type B (conservative, neutral sentiment and objective). The evaluation for conservativeness or liberalness floats as the mood goes down. The values get closer to the real value as the mood gets worst, which follows the expected theory that a person in bad mood would be more critic to arguments.

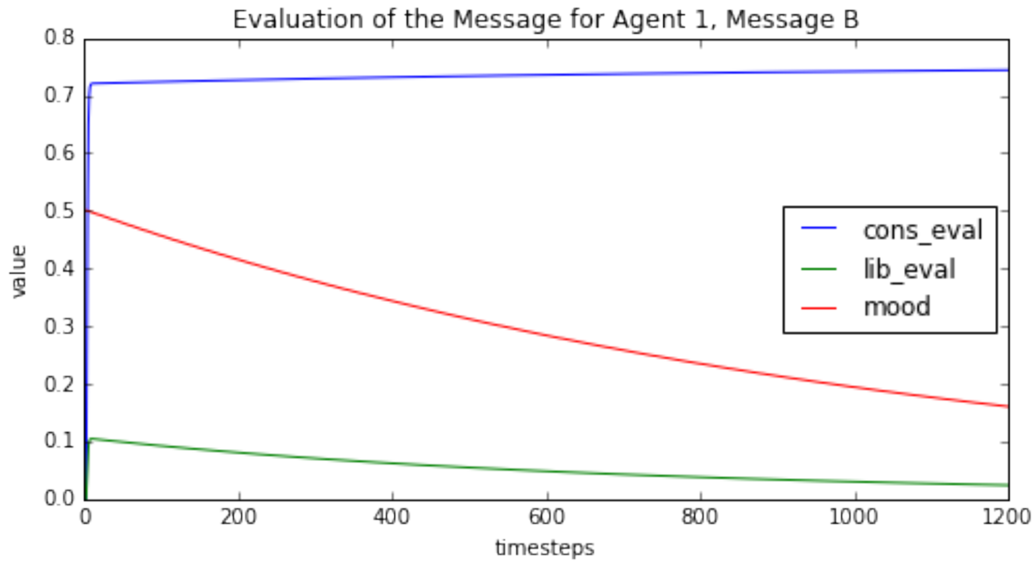


Fig. 10.4: Interpretation of the message for agent 1, message B.

For agent 3, a moderate agent, the scenario of changes is higher. Figure 10.5 shows a moderate agent receiving conservative messages with bad sentiment. As the criticism of the agent becomes higher, the changes are also affected. For this agent, the sequence of messages influence it more than its own coherence system for liberalism.

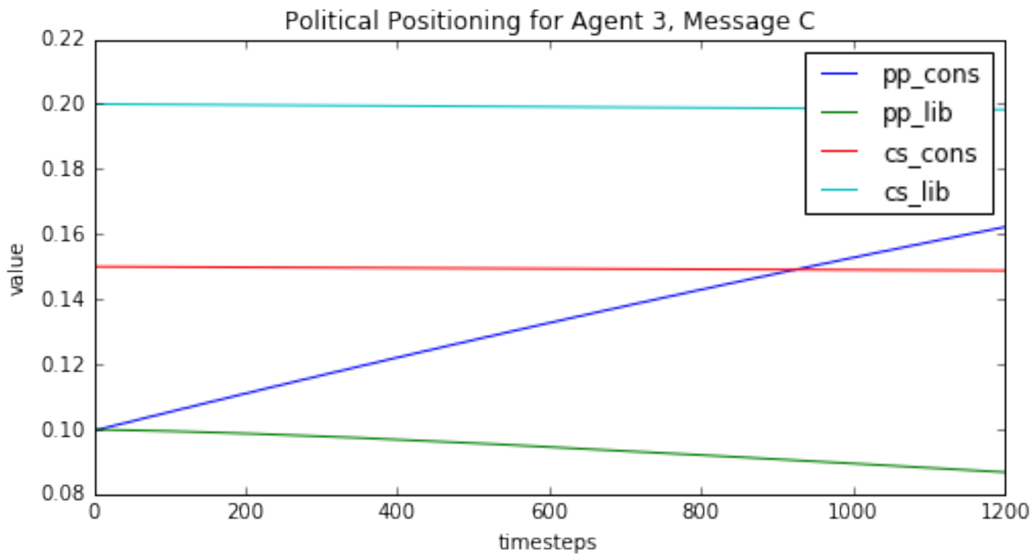


Fig. 10.5: Changes on political position for agent 3.

For agent 3 it is also possible to observe the feeling of change decreasing as the agent adjust its own convictions to a more conservative position. As the feeling of change decreases, the mood is also affected and turns flatter instead of following the curve down. Figure 10.6 shows the results for these states and agent 3.

Figure 10.7 shows agent 4 getting changed in its positions when confronted with a message that goes against its predominant opinion. As conservativeness and liberalness are close to each other, this transition can happens. This can be a good simulation for people who tends to be more balanced on their analysis of political issues.

All the simulations regarding all the messages and agents can be seen in Appendix B (`http://www.cs.vu.nl/~efo600/iccicc17/AppendixB.pdf`).

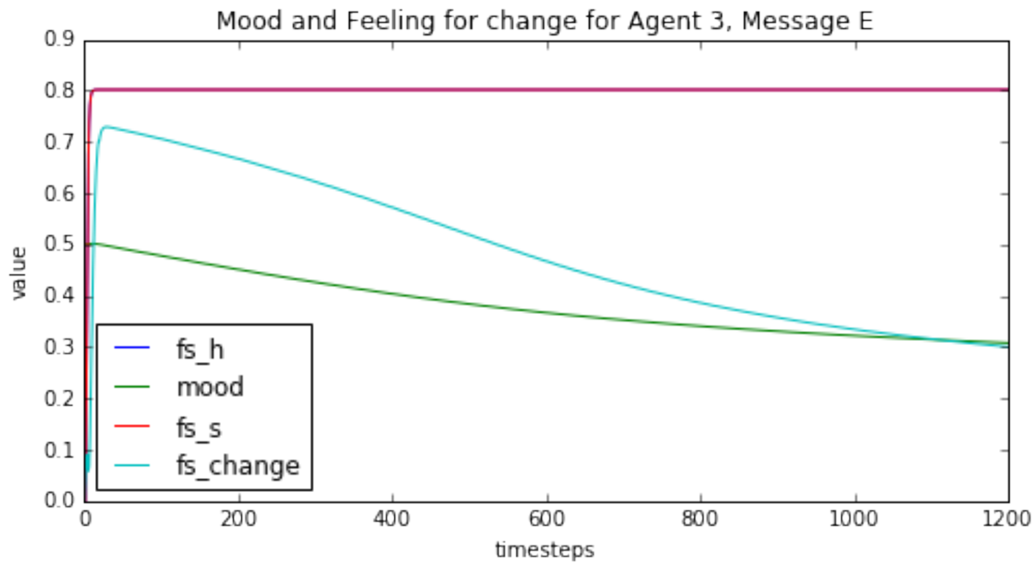


Fig. 10.6: Changes on political position for agent 3.

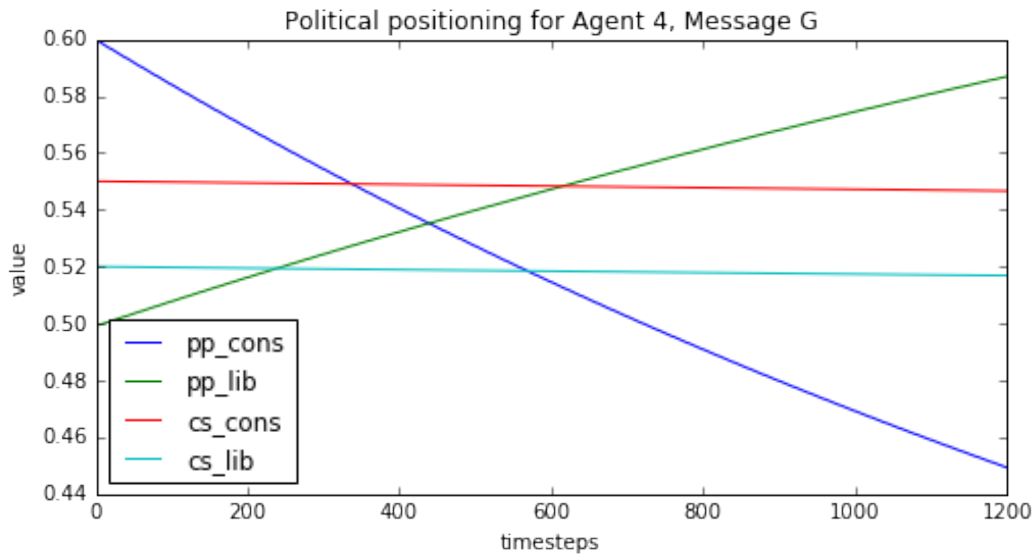


Fig. 10.7: Changes on political position for agent 3.

10.6 Discussion and future work

Based on recent literature about the way in which the political positioning of people is influenced by (neuro-)psychological factors as well as social factors, we have presented a model that describes how the political positioning of individuals is influenced by information that we receive via (social) media. Given the large role that social media plays in modern societies in general and in political campaigns specifically, this model can contribute to a better understanding of the political dynamics in the modern world.

The current model is only a first attempt to formally describe the process. A clear limitation is the fact that the model is ignorant to other societal factors that have influence on people's political positioning. However, the aim of this paper was not to provide a complete model of political positioning, but to focus on the role of receiving and interpreting information via social media networks. Extensions that include other factors can be made in the future. Another limitation is the fact that the model has not been validated in practice. Given the complex nature of the topic, this is inherently difficult, if even possible.

The model can be used in several application scenario's. First, one could use the model to simulate and possibly evaluate the effect of potential social media campaigns. Secondly, it can be used to further study (neuro-)psychological and social theories about influence and contagion of political opinions. Further empirical research is required to verify the way in which the model describe those processes. Thirdly, the model could form a basis for virtual social media characters that should behave in a realistic and human-like fashion. Our current research is focusing on the implementation of such an agent and the evaluation of its behaviour in a context with human agents.

Bibliography

[1] Theodor W Adorno, Else Frenkel-Brunswik, Daniel J Levinson, and R Nevitt Sanford. *The authoritarian personality.* Harpers, 1950 (cit. on p. 168).

[2] John T Cacioppo, Gary G Berntson, Tyler S Lorig, et al. „Just because you're imaging the brain doesn't mean you can stop using your head: a primer and set of first principles." In: *Journal of personality and social psychology* 85.4 (2003), p. 650 (cit. on p. 169).

[3] Gian Vittorio Caprara and Daniel Cervone. *Personality: Determinants, dynamics, and potentials*. Cambridge University Press, 2000 (cit. on p. 167).

[4] Gian Vittorio Caprara, Shalom Schwartz, Cristina Capanna, Michele Vecchione, and Claudio Barbaranelli. „Personality and Politics: Values, Traits, and Political Choice". In: *Political Psychology* 27.1 (Feb. 2006), pp. 1–28 (cit. on pp. 166–169, 173).

[5] Dana R Carney, John T Jost, Samuel D Gosling, and Jeff Potter. „The Secret Lives of Liberals and Conservatives: Personality Profiles, Interaction Styles, and the Things They Leave Behind". In: *Political Psychology* 29.6 (2008) (cit. on pp. 166–168, 173).

[6] 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. 170).

[7] John Duckitt and Chris G. Sibley. „Personality, Ideological Attitudes, and Group Identity as Predictors of Political Behavior in Majority and Minority Ethnic Groups". In: *Political Psychology* 37.1 (Feb. 2016), pp. 109–124 (cit. on p. 166).

[8] Nico H Frijda. „The laws of emotion." In: *American psychologist* 43.5 (1988), p. 349 (cit. on p. 171).

[9] Erich Fromm. *Man for himself: An inquiry into the psychology of ethics*. Vol. 102. Routledge, 2013 (cit. on p. 168).

[10] Erich Fromm. *The heart of man*. Lantern Books, 2011 (cit. on p. 168).

[11] Anthony G Greenwald. „Cognitive learning, cognitive response to persuasion, and attitude change". In: *Psychological foundations of attitudes* (1968), pp. 147–170 (cit. on p. 170).

[12] John T Jost and David M Amodio. „Political ideology as motivated social cognition: Behavioral and neuroscientific evidence". In: *Motivation and Emotion* 36.1 (2012), pp. 55–64 (cit. on p. 169).

[13] John T Jost, Jack Glaser, Arie W Kruglanski, and Frank J Sulloway. „Political conservatism as motivated social cognition." In: *Psychological Bulletin* 129.3 (2003), pp. 339–375 (cit. on p. 168).

[14] John T Jost, H Hannah Nam, David M Amodio, and Jay J. Van Bavel. „Political neuroscience: The beginning of a beautiful friendship". In: *Political Psychology* 35.SUPPL.1 (2014), pp. 3–42 (cit. on pp. 166, 169).

[15] John T Jost, Brian A Nosek, and Samuel D Gosling. „Ideology: Its resurgence in social, personality, and political psychology". In: *Perspectives on Psychological Science* 3.2 (2008), pp. 126–136 (cit. on pp. 170, 173).

[16] Ryota Kanai, Tom Feilden, Colin Firth, and Geraint Rees. „Political orientations are correlated with brain structure in young adults". In: *Current biology* 21.8 (2011), pp. 677–680 (cit. on p. 169).

[17] Bibb Latane et al. „The psychology of social impact". In: *American psychologist* 36.4 (1981), pp. 343–356 (cit. on p. 170).

[18] Diane M Mackie and Leila T Worth. „Processing deficits and the mediation of positive affect in persuasion." In: *Journal of personality and social psychology* 57.1 (1989), p. 27 (cit. on p. 171).

[19] Robert R McCrae. „Social consequences of experiential openness." In: *Psychological bulletin* 120.3 (1996), p. 323 (cit. on p. 168).

[20] Douglas R Oxley, Kevin B Smith, John R Alford, et al. „Political attitudes vary with physiological traits". In: *science* 321.5896 (2008), pp. 1667–1670 (cit. on p. 169).

[21] Richard E Petty and John T Cacioppo. „The elaboration likelihood model of persuasion". In: *Communication and persuasion*. Springer, 1986, pp. 1–24 (cit. on p. 171).

[22] Richard E Petty, Thomas M Ostrom, and Timothy C Brock. *Cognitive responses in persuasive communications: A text in attitude change*. 1981 (cit. on p. 170).

[23] Ellen AA van der Plas, Aaron D Boes, John A Wemmie, Daniel Tranel, and Peg Nopoulos. „Amygdala volume correlates positively with fearfulness in normal healthy girls". In: *Social cognitive and affective neuroscience* (2010), nsq009 (cit. on p. 169).

[24] Russell A Poldrack. „Can cognitive processes be inferred from neuroimaging data?" In: *Trends in cognitive sciences* 10.2 (2006), pp. 59–63 (cit. on p. 169).

[25] Shalom H. Schwartz. „Universals in the Content and Structure of Values: Theoretical Advances and Empirical Tests in 20 Countries". In: ed. by Mark P. Zanna. Vol. 25. Advances in Experimental Social Psychology. Academic Press, 1992, pp. 1–65 (cit. on p. 168).

[26] Norbert Schwarz, Herbert Bless, and Gerd Bohner. „Mood and persuasion: Affective states influence the processing of persuasive communications". In: *Advances in experimental social psychology* 24 (1991), pp. 161–199 (cit. on pp. 170, 172).

[27] Jim Sidanius. „Intolerance of ambiguity and socio-politico ideology: A multidimensional analysis". In: *European Journal of Social Psychology* 8.2 (1978), pp. 215–235 (cit. on p. 168).

[28] Geoffrey M Stephenson and Geoffrey T Fielding. „An experimental study of the contagion of leaving behavior in small gatherings“. In: *The Journal of Social Psychology* 84.1 (1971), pp. 81–91 (cit. on p. 170).

[29] Alexander G Theodoridis and Amy J Nelson. „Of BOLD claims and excessive fears: A call for caution and patience regarding political neuroscience“. In: *Political Psychology* 33.1 (2012), pp. 27–43 (cit. on p. 169).

[30] Cibu Thomas and Chris I Baker. „Teaching an adult brain new tricks: a critical review of evidence for training-dependent structural plasticity in humans“. In: *NeuroImage* 73 (2013), pp. 225–236 (cit. on p. 169).

[31] Silvan Tomkins. „Left and right: A basic dimension of ideology and personality.“ In: (1963) (cit. on p. 168).

[32] J. Treur. *Network-Oriented Modeling: Addressing Complexity of Cognitive, Affective and Social Interactions*. Understanding Complex Systems. Springer International Publishing, 2016 (cit. on pp. 170, 172, 176).

[33] Paul J Whalen. „The uncertainty of it all“. In: *Trends in cognitive sciences* 11.12 (2007), pp. 499–500 (cit. on p. 170).

[34] Ladd Wheeler. „Toward a theory of behavioral contagion.“ In: *Psychological Review* 73.2 (1966), p. 179 (cit. on p. 170).

[35] Glenn D Wilson, James Ausman, and Thomas R Mathews. „Conservatism and art preferences.“ In: *Journal of Personality and Social Psychology* 25.2 (1973), p. 286 (cit. on p. 168).

[36] Fei Xiong and Yun Liu. „Opinion formation on social media: an empirical approach“. In: *Chaos: An Interdisciplinary Journal of Nonlinear Science* 24.1 (2014), p. 013130 (cit. on p. 170).

11 Detecting Dutch political tweets perceptions: A classifier based on voting system using supervised learning[1]

> *"The humans are often puzzled to understand the range of his opinions —- why he is one day almost a Communist and the next not far from some kind of theocratic Fascism —- one day a scholastic, and the next prepared to deny human reason altogether —- one day immersed in politics, and, the day after, declaring that all states of this world are equally 'under judgment.' We, of course, see the connecting link, which is Hatred."*
>
> **— C.S. Lewis**
> "The Screwtape Letters"[2]

The task of classifying political tweets has been shown to be very difficult, with controversial results in many works and with non-replicable methods. Most of the works with this goal use rule-based methods to identify political tweets. We propose here two methods, being one rule-based approach, which has an accuracy of 62%, and a supervised learning approach, which went up to 97% of accuracy in the task of distinguishing political and non-political tweets in a corpus of 2.881 Dutch tweets. Here we show that for a data base of Dutch tweets, we can outperform the rule-based method by combining many different supervised learning methods.

[1]Published as: **Fernandes de Mello Araújo E.** and Ebbelaar D. (2018). *Detecting Dutch Political Tweets: A Classifier based on Voting System using Supervised Learning*. In Proceedings of the 10th International Conference on Agents and Artificial Intelligence - Volume 2: ICAART, pages 462-469.

11.1 Introduction

Social media platforms became an excellent source of information for researchers due to its richness in data. Social scientists can derive many studies from behaviour in social media, from preferences regarding brands, political orientation of mass media to voting behaviour [1, 7, 19]. Most of these works rely in text mining techniques to interpret the big amount of data which would mostly not be processed manually in a feasible time.

One of the most used social media platforms for text mining is Twitter, a microblogging service where users post and interact with messages called "tweets", restricted to 140 characters. As of June 2017, about 500 million tweets are posted to Twitter every day[3]. Twitter presents an API for collecting data, and many works have been using it including for political analysis [7, 15, 17].

Natural language processing (NLP) is a computer science method for processing and understanding natural language, mostly vocal or textual. Despite the fact that NLP has been used for decades, processing tweets has brought new challenges to this field. The limited amount of information (limited number of characters in each message) induces the users of the Twitter platform to ignore punctuation, shorten longer words and creates abbreviations for common used expressions, as FYI (for your information). Most recently the increasing use of emojis (ideograms and smileys used in the message) also raised new features that NLP algorithms have to deal with.

Much works are currently combining NLP techniques for twitter messages in order to assess information about public political opinions, aiming mostly to predict the results of elections or referendums. One of the steps to develop a NLP system to classify tweets is the filtering of tweets that concern politics from other topics of discussion. Rule-based techniques like the use of keywords to identify political tweets resulted in 38% of the tweets being falsely classified in our collected Dutch corpus. This is far from ideal and therefore other classification methods should be evaluated.

We have used in this work a supervised learning approach for classification of the tweets in political or non-political, a machine learning technique where a function is generated from labeled training data. An algorithm analyzes a dataset and generates an inferred function, which can be used to classify unseen instances. We aim to examine whether classifying political content from Twitter using a supervised learning approach outperforms a rule-based method, leading to more accurate analyses of political content. To construct the classifier, a corpus of 2.881 Dutch tweets was first collected over a time period of two months. The corpus was manually tagged using a web application built for this project. Tweets were then pre-processed and extra features were extracted from metadata to optimize for classification. Various machine learning algorithms were trained using the tagged dataset and accuracies were compared to find the right models. Eventually, the five best performing models were combined to make a classifier that uses a voting system.

[3]https://www.omnicoreagency.com/twitter-statistics/

The structure of this paper is as follows. Section 11.2 discusses related work. In Section 11.3 the method for collecting, tagging and pre-processing data is explained, followed by the process of building the classifier and an explanation of the models in Section 11.4. The results are shown in Section 11.5 followed by future research and discussion in Section 11.6.

11.2 Related work

One of the earliest studies to use Twitter for political analysis aims to predict the German federal elections using data from Twitter, concluding that the number of messages mentioning a party reflects the election result [21]. They collected all tweets that contained the names of the six parties represented in the German parliament or selected prominent politicians related to these parties. With a rule-based method to identify tweets as being politically relevant, they stated that the number of messages mentioning a party reflects the election result.

A similar method of counting Twitter messages mentioning political party names was applied to predict the 2011 Dutch Senate election [18]. The results were contradictory with [21], concluding that counting the tweets that mention political parties is not sufficient to obtain good election predictions.

He et al. [8] analyzed tweet messages leading to the UK General Election 2010 to see whether they reflect the actual political scenario. They have used a rule-based method. A model was proposed incorporating side information from tweets, i.e. emoticons and hashtags, that can indicate polarities. Their search criteria included the mention of political parties, candidates, use of hashtags and certain words. Tweets were then categorized as in relevance to different parties if they contain keywords or hashtags. Their results show that activities on Twitter cannot be used to predict the popularity of election parties.

A study by Conover et al. [5] investigated how social media shapes the networked public sphere and facilitates communication between communities with different political orientations. Two networks of political communication on Twitter were examined leading up to the 2010 U.S. congressional midterm elections. A political communication was identified as any tweet containing at least one politically relevant hashtag. To identify an appropriate set of political hashtags, a tag co-occurrence discovery procedure was performed. They began by seeding the sample with the two most popular political hashtags. For each seed, they identified the set of hashtags with which it co-occurred in at least one tweet and ranked the results. They stated that when the tweets in which both seed and hashtag occur make up a large portion of the tweets in which either occurs, the two are deemed to be related. Using a similarity threshold they identified a set of unique hashtags. This method is more advanced than the previously discussed methods but lacks recall of political content. Hong et al. [9] showed that only 11% of all tweets contain one or more hashtags. While this study was conducted on Twitter data in general and not just political content, one can still assume that far from all political relevant tweets contain a hashtag.

Several studies have also used Twitter data to predict the political orientation of users. Some with great success where accuracies are reported over 90% [4, 12]. However, Cohen and Ruths [3] discovered that reported accuracies have been systemically overoptimistic due to the way in which validation datasets have been collected, reporting accuracy levels nearly 30% higher than can be expected in populations of general Twitter users meaning that tweet classifiers cannot be used to classify users outside the narrow range of political orientation on which they were trained.

Maynard and Funk [14] used NLP advanced techniques to classify tweets and their political orientation without much success. They conclude that machine learning systems in annotated corpus of tweets could improve their method.

As showed, most of the works aim to categorize tweets regarding their political positioning without removing those which follow their rule-based method but do not have political content. We consider that filtering the tweets with a very good accuracy tool is a way of improving the results presented by previous works. If tweets can be classified as political or not political before they pass through other processes, better results can be obtained.

11.3 Collecting and processing the tweets

This project consists of data collection, data cleaning, tagging of the messages and finally the processing and analysis of the results.

11.3.1 Collecting data

To collect the tweets we have used the Twitter Streaming API[4]. The API pushes data in real-time, and provides a search mechanism that can be based on keywords, user-names, language or locations. The tweets that match the criteria are pushed directly to the destination defined in your code. The public stream can push approximately 1% of all the Twitter data[5]. The full stream of data can be accessed using the Twitter Firehose but is fairly costly. For this work, a sample of the data was sufficient enough to train a classifier and therefore the Streaming API was used.

For the collection of the corpus, the abbreviations of the Dutch political parties and the names of their leaders were used as the set of keywords shown in Table 11.1. Hashtags are not included because a hashtag will only match the given hashtag and not the keyword without the hashtag. For example '#Twitter' will only match tweets containing '#Twitter' whereas using just 'Twitter' will match 'Twitter' and '#Twitter'. Therefore adding hashtags for parties or names would be redundant. The first and last names of the politicians were searched separately because it was noticed that people rarely address Dutch politicians by their full name in tweets. Besides the keywords, a language filter was used to only push Dutch tweets. Data was streamed in intervals over a time period of two months to make sure the results were not

[4]https://dev.twitter.com/streaming/overview

[5]https://brightplanet.com/2013/06/twitter-firehose-vs-twitter-api-whats-the-difference-and-why-should-you-care/

Tab. 11.1: Keywords used to filter the tweets collected

Party	Leader
VVD	Mark, Rutte
PVV	Geert, Wilders
CDA	Sybrand, Haersma, Buma
D66	Alexander, Pechtold
GL	Jesse, Klaver
PvdA	Lodewijk, Asscher
SP	Emile, Roemer
CU	Gert-Jan, Segers
PvdD	Marianne, Thieme
50plus	Henk, Krol
SGP	Kees, Staaij
DENK	Tunahan, Kuzu
FvD	Thierry, Baudet

influenced by major events. After removing duplicates, this resulted in a total of 2.881 tweets.

11.3.2 Cleaning the data

The Twitter Streaming API returns the collected data in a JSON format. We cleaned up the data by extracting the relevant features as username, text, expanded_url, extended_text, retweeted_status and reply_status. The utility of each feature will be explained in this section.

The collected corpus contained duplicate tweets. In order to automatically remove duplicates from the dataset, URLs had to be temporarily removed because Twitter creates unique URLs for every tweet using their t.co service which shortens URLs. After the removal of duplicates, the URLs were placed back.

Because Twitter shortens the URLs, potential information gets lost, so the shortened URL was replaced by features extracted from the expanded_url feature which contains the original URL. This was done by splitting up the URL using the Python *urlpase* package. Special characters and Dutch stop words were removed using the NLTK stop word corpus[6]. An additional set of frequent URL words was also removed containing words such as 'www', 'html' and 'com'. This way only relevant words would remain. An example of the feature extraction from an URL is shown in Table 11.2.

Tab. 11.2: Feature extraction example

URL	https://t.co/C7fwW3eE5p
Features extracted	fd, economie, politiek, asscher, sluit, deal, soepeler, ontslagrecht

[6] http://www.nltk.org/book/ch02.html

To further extract as much information as possible, retweets (tweets that are shared by another user) had to be replaced with the original text because sometimes the text of a retweet is truncated. Tweets can also contain an extended_text feature. When this was the case, the text was replaced with the extended_text feature. This method ensures that the full text is displayed. Replies lack context and therefore make accurate tagging hard or impossible. In order to include replies, additional steps should be taken to link replies to tweets. However, for this project, replies were removed from the dataset. Finally, the clean dataset was exported to a CSV file and passed on to the tagging system.

11.3.3 Tagging the tweets

In order to use supervised learning, the tweets had to be manually tagged first. This was done using a web application that was built for this project. The goal was to create a tagging system that can also be used for future projects. Another option would be using the Amazon Mechanical Turk website for tagging, but since the dataset is relatively small and domain specific (Dutch politics) the self-built application was a better option. The interface can be seen in Figure 11.1. The app shows one tweet at a time and a tweet could be tagged as either political or non-political by clicking the green or the red button. Tags were saved in a database which could be downloaded as a CSV file to transfer back to the program. A distribution of the tagged tweets is shown in Table 11.3.

Tab. 11.3: Collected tweets

Total	2.881
Political	1.823 (62,0%)
Non-political	1.058 (38,0%)

While the set of keywords only contained politically relevant words, 38% of the tweets are tagged as non-political. Most of this noise comes from tweets where people mention the first name of a political leader but refer to someone else. There are also cases where political leaders are mentioned, but not in a political way. For example, the Dutch prime minister went skydiving during the collection of data. Therefore it contains some tweets commenting on the jump, mentioning the Prime Minister, but has nothing to do with politics.

11.3.4 Rule-based method

To extract political tweets using a rule-based method, tweets were classified as politically relevant if they contained at least one of the keywords from Table 11.1. Most of the works shown in Section 11.2 use the same approach.

In this case, the Twitter Streaming API basically acts as the classifier by only pushing tweets that contain at least one of the keywords provided in the search. To calculate the accuracy we only have to verify which tweets contain the keywords but are not related to political topics of discussion.

Fig. 11.1: Tagging application

Fig. 11.2: Bag-of-words feature representation

11.4 Structure of the classifier

In order to build a classifier, the tweets first had to be converted to a mathematical feature representation. This was done using the bag-of-words model [10, 11]. In this model, the text is represented as the bag (multiset) of its words. The bag-of-words model is often used in methods of text classification where the frequency of occurrence of each word is used as a feature for training a classifier. An example of such a feature representation is shown in Figure 11.2. To achieve this, the *Countvectorizer* module was used [16].

Before the bag-of-words could be created, Dutch stop words and special characters were removed and text was converted to lowercase. This was done to ensure that only relevant words would remain and names would have the same form, independent of uppercase use. With the removal of special characters, emoticons were also removed. While emoticons can contain sentimental information, they were never a deciding factor to classify a tweet in this dataset. The characters # and @ (frequently used Twitter characters) were also removed in this process but the words following the characters remained. This way mentions, replies and hashtags referring to parties and leaders have the same form.

While analyzing the word frequencies of the total corpus, it was noticed that some specific politically irrelevant words occurred frequently. These were mostly words related to events. Since the data was collected in intervals over a relatively short time period, these words were removed to ensure the classifier would not overfit on these irrelevant words. Stemming [13] and the use of tf-idf [6] did not improve results. The 1.000 most frequent words were used for the bag of words.

To run the machine learning process, each tweet was converted using the Countvectorizer. The information used is username, text and the features extracted from the URL when present in the tweets JSON output. The set of feature representation of the tweets was then split up into a training (80%) and testing (20%) set. This way an estimation of the classifier's performance can be made. The training data was finally passed on to a series of eight machine learning models from the Scikit-learn Python module:

- Logistic regression
- Linear discriminant analysis
- K-nearest neighbors
- Classification and regression trees
- Random forest
- Gaussian naive bayes
- Support vector machines
- Neural network

The **Logistic Regression (LR)** is a linear machine algorithm based on the statistical logistic function, also known as the sigmoid function, as shown in figure 11.1.

$$1/(1+e^{-value}) \tag{11.1}$$

The function takes on an S-shaped curve and can take any real-valued number and map it between 0 and 1. LR is used for two-class (binary) classification problems. The algorithm makes predictions by linearly combining input values using weights or coefficient values. LR performs well on numerical data with lots of features and is often used for a first look at the dataset because it is computationally fast. Besides that, the model is not so prone to overfitting.

Overfitting can occur when a model is very complex, such as having too many parameters relative to the amount of data. A model that has been overfit will overreact to minor fluctuations in the training data and therefore will have a poor predictive performance [2].

Linear Discriminant Analysis (LDA) is another linear machine learning algorithm used for multi-class classification problems that can also be used for binary classification. LDA uses the statistical properties of each class calculated from the data. It takes the mean and the variance of a single input variable for each class and uses the LDA equation to make predictions.

While training the LDA model on this dataset a warning occurred stating that the variables are collinear. This means that the predictors are correlated. This is not optimal for LDA because it involves computing a matrix inversion, which is not accurate if the determinant is close to zero. Therefore we expect this model to not perform well on our dataset.

K-Nearest Neighbors (KNN) is a non-linear algorithm that uses the entire dataset for representation, with no learning required. Predictions are made using the K most similar instances (neighbors) in the training set. To calculate which instances are most similar (closest), the Euclidean distance measure is often used, which takes the square root of the sum of the squared differences between a new point and an existing point across all input attributes. KNN can be used for both regression and classification problems but can perform poorly on high dimensional datasets.

Classification and Regression Trees (CART) is a non-linear decision tree algorithm. As the name indicates, the CART variant can be used for classification and regression problems. The CART model is represented as a binary tree. Each root node represents a single input variable and a split point on that variable. The last nodes of the tree, called the leaf nodes, contain an output variable which is used to make predictions. CART is computationally fast and robust to noise and missing values. The model is also easy to interpret visually when the trees only contain several levels.

The **Random Forest (RF)** algorithm is another form of a decision tree that constructs multiple decision trees during training. To classify a new input, each of the trees in the forest makes a classification. The algorithm then chooses the classification that occurs the most. Regular decision trees are prone to overfitting to their training set, RF corrects for this. However, the RF is harder to visually interpret than a regular decision tree.

The **Gaussian Naive Bayes (NB)** is also a non-linear algorithm used for binary and multi-class classification. The probability of a hypothesis is calculated using Bayes Theorem given prior knowledge of the dataset. It makes predictions based on the probabilities of each class in the training dataset and the conditional probabilities of each input value given each class value. NB is computationally fast and simple to implement but relies on independence assumption and will not perform well if this assumption is not met.

Support Vector Machines (SVM) split up data in a two-dimensional space using a hyperplane. A hyperplane is chosen to best separate the data by their classes. The hyperplane is established by learning from the training data. Predictions are made using this line by feeding a new value to the line equation. The algorithm then calculates whether the value is above or below the line to classify the input. SVM can model complex, nonlinear relationships, are robust to noise and good at text classification [20]. This model is therefore expected to perform well on this dataset.

Neural Network (NN) algorithms are inspired by the structure and functionality of the brain. Calculations are made using an interconnected group of neurons, that pass on information once a certain threshold is met. NNs are used to model relationships between data, to find patterns in data and can also be used for classification. NNs are extremely powerful and can model very complex relationships without the need to understand the underlying data. NNs are good at classifying images, video and even human-intelligence type tasks like driving.

To get a baseline performance estimation, the models were trained using the default settings. The algorithms were evaluated using cross-validation. Cross-validation is a method where the training set is split up into K-folds. The algorithm is then trained on K-1 folds and tests its accuracy on the remaining fold that was not used for training. This process is repeated K times where every time another fold is used for testing. After training and testing on all the possible folds, the mean accuracy is calculated. So cross-validation combines the average prediction error to derive a more accurate estimate of the performance of the model. For this project, 10-folds were used and the random seed was reset before each test to make sure that the evaluation of each algorithm was done using exactly the same data splits to ensure that the results are directly comparable.

11.5 Results

This section presents the results for the two methods used to classify the tweets: a rule-based and a supervised learning methods.

11.5.1 Rule-based method

As explained in the Section 11.3, the accuracy of the rule-based method is measured by comparing the tweet corpus collected by the API to the results obtained by manually tagging the tweets. From the total 2.881 tweets, only 1.823 tweets were actually politically relevant, resulting in an accuracy of 62%.

11.5.2 Supervised learning

As explained in the previous section, we have run eight cross-validation models to find a good fit for our data set. The mean accuracy from the cross-validation per model was calculated and resulted in the scores shown in Figure 11.3 and Table 11.4

As can be observed in table 11.4, LDA, KNN and NB are outperformed by the other models by more than 10%. Therefore these models were excluded from the final classifier. The five remaining models were then trained on the whole training set and used to make predictions on the test set. This process was repeated 10 times

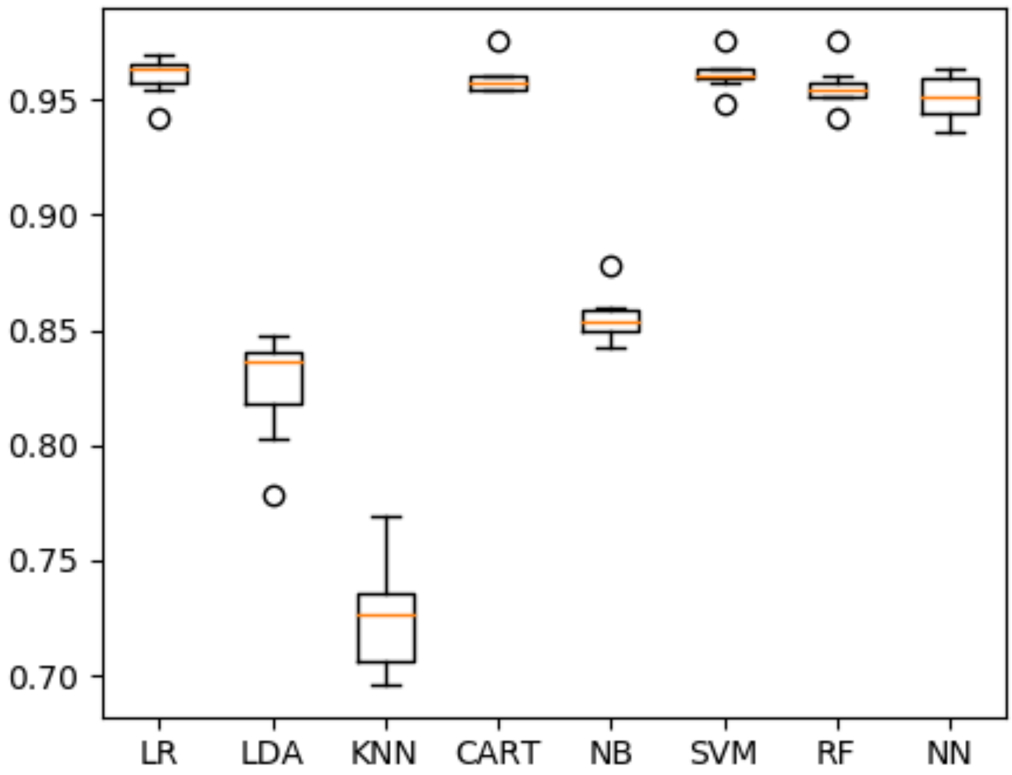


Fig. 11.3: Training set accuracies

with different training/test splits resulting in the average accuracies shown in Table 11.5.

The accuracies are very similar, therefore the models were combined to check whether it would improve performance. This was done by using a voting system. Since there are five models, a vote will always have a majority. If three models classify a tweet as 'political' and two as 'non-political', the final prediction will be 'political' and vice versa. With the combination of models, the accuracy on the test set went up by roughly 1% depending on the training/test split, resulting in an average accuracy of 97%.

Accuracy can be misleading though. A model with a lower accuracy can sometimes have a greater predictive power. This can occur when there is a class imbalance which is the case for this dataset. The classification report in Figure 11.4 provides a breakdown of the classes by precision, recall and f1-score where 'N' and 'Y' correspond to non-political and political tweets respectively. The classification report shows that the classifier slightly underperforms (93%) in classifying non-political tweets as non-political but overall performs well and therefore the accuracy measure is not misleading.

Tab. 11.4: Cross validation results

Model	Accuracy
LR	0.96
LDA	0.83
KNN	0.73
CART	0.96
NB	0.86
SVM	0.96
RF	0.96
NN	0.95

Tab. 11.5: Test set results

Model	Accuracy
LR	0.96
CART	0.95
SVM	0.96
RF	0.96
NN	0.95

11.6 Discussion

The classification of tweets for the prediction of political elections and people's opinions in social media became very controversial, leading to completely different results when using rule-based methods for this purpose. We trust that there is a potential improvement in those results by separating tweets that are related to political topics before classifying them as supportive to certain parties of political positions.

This work presents a method based on more than one machine learning algorithm to define the content of messages shared in Twitter concerning the topic of discussion as political or non-political. In our method, the five best performing machine learning models were combined to create a voting system that can distinguish never before seen political from non-political Dutch tweets with an accuracy of 97%. The usage of this method can be extended to studies related to spread of political opinion on social media, political interpretation of social media content, and can also be applied to other problems related to classification of text content.

The results show that using a supervised learning approach to identify political tweets instead of a rule-based method could result in more representative datasets which could then lead to more accurate analyses of political content from Twitter. The method described in this paper could help to solve the contradictory results from previous studies discussed in here.

While the results of this study are sound, further research should be done to investigate how the classifier transfers to other, but similar corpora. Cohen and Ruths [3] showed that tweet classifiers cannot be used to classify users outside the narrow

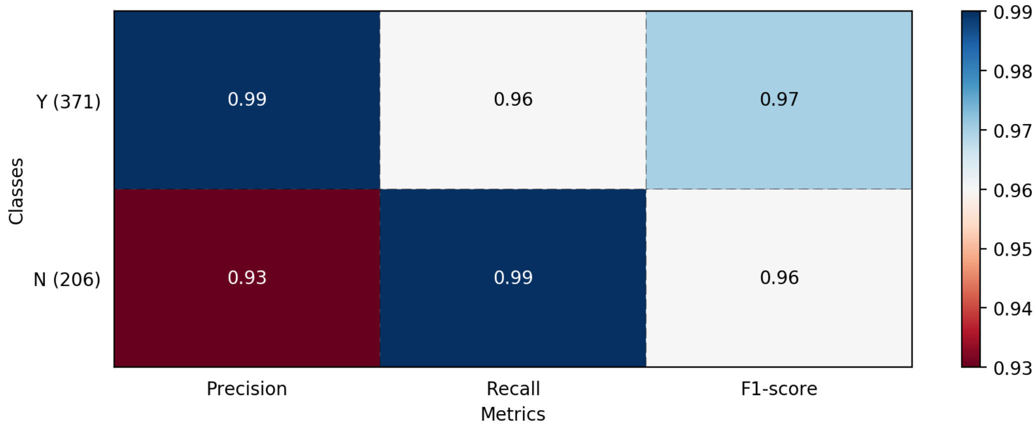


Fig. 11.4: Classification report for political (Y) and non-political (N) tweets

range of political orientation on which they were trained. However, their study was done on the classification of the political orientation of users and not political tweets in general.

Our classifier was trained and tested on a small dataset collected over a short period of time (2.881 tweets in a two months time span). The political agenda changes over time and thus also the political subjects which people tweet about. A classifier should be held up to date by adding new training data and increasing the sample size.

The set of keywords used for the collection of political tweets is also limited. The set included the abbreviations of the Dutch political parties and their leaders but there are others ways to address politics. For example by using the words 'Senate' or 'Prime Minister'. Thus the set of keywords could be extended according to the desired application.

The method used in this research also lacks a technique to process replies. A solution to this could be to link the reply to the original tweet, and separate both texts. This can be very useful when studying the effect of the spread of messages in social networks.

Finally, the machine learning models could be tweaked further to optimize the results. In this process, called hyperparameter optimization, the model settings are adjusted accordingly to the dataset. Future work is going to be carried in improving the parameters of the models. We also aim to use the classifier in other works related to social network analysis of political positions and social contagion of political opinions in networks.

Bibliography

[1] Sitaram Asur and Bernardo A Huberman. „Predicting the future with social media“. In: *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on*. Vol. 1. IEEE. 2010, pp. 492–499 (cit. on p. 188).

[2] Michael A Babyak. „What you see may not be what you get: a brief, nontechnical introduction to overfitting in regression-type models“. In: *Psychosomatic medicine* 66.3 (2004), pp. 411–421 (cit. on p. 194).

[3] Raviv Cohen and Derek Ruths. „Classifying political orientation on Twitter: It's not easy!“ In: *ICWSM*. 2013 (cit. on pp. 190, 198).

[4] Michael D Conover, Bruno Gonçalves, Jacob Ratkiewicz, Alessandro Flammini, and Filippo Menczer. „Predicting the political alignment of twitter users“. In: *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on*. IEEE. 2011, pp. 192–199 (cit. on p. 190).

[5] Michael Conover, Jacob Ratkiewicz, Matthew R Francisco, et al. „Political Polarization on Twitter.“ In: *ICWSM* 133 (2011), pp. 89–96 (cit. on p. 189).

[6] Bruce Croft, Donald Metzler, and Trevor Strohman. *Search Engines: Information Retrieval in Practice*. 1st. USA: Addison-Wesley Publishing Company, 2009 (cit. on p. 194).

[7] Jennifer Golbeck and Derek Hansen. „A method for computing political preference among Twitter followers“. In: *Social Networks* 36 (2014), pp. 177–184 (cit. on p. 188).

[8] Yulan He, Hassan Saif, Zhongyu Wei, and Kam-Fai Wong. „Quantising opinions for political tweets analysis“. In: *LREC 2012, Eighth International Conference on Language Resources and Evaluation*. 2012 (cit. on p. 189).

[9] Lichan Hong, Gregorio Convertino, and Ed H Chi. „Language Matters In Twitter: A Large Scale Study.“ In: *ICWSM*. 2011 (cit. on p. 189).

[10] Thorsten Joachims. „Text categorization with support vector machines: Learning with many relevant features“. In: *Machine learning: ECML-98* (1998), pp. 137–142 (cit. on p. 193).

[11] Bing Liu. „Sentiment analysis and opinion mining“. In: *Synthesis lectures on human language technologies* 5.1 (2012), pp. 1–167 (cit. on p. 193).

[12] Wendy Liu and Derek Ruths. „What's in a Name? Using First Names as Features for Gender Inference in Twitter.“ In: *AAAI spring symposium: Analyzing microtext*. Vol. 13. 2013, p. 01 (cit. on p. 190).

[13] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. New York, NY, USA: Cambridge University Press, 2008 (cit. on p. 194).

[14] Diana Maynard and Adam Funk. „Automatic detection of political opinions in tweets“. In: *Extended Semantic Web Conference*. Springer. 2011, pp. 88–99 (cit. on p. 190).

[15] Saif M. Mohammad, Xiaodan Zhu, Svetlana Kiritchenko, and Joel Martin. „Sentiment, emotion, purpose, and style in electoral tweets“. In: *Information Processing and Management* 51.4 (2015) (cit. on p. 188).

[16] F. Pedregosa, G. Varoquaux, A. Gramfort, et al. „Scikit-learn: Machine Learning in Python“. In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830 (cit. on p. 193).

[17] Ashwin Rajadesingan and Huan Liu. „Identifying users with opposing opinions in Twitter debates“. In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 8393 LNCS (2014), pp. 153–160. eprint: 1402.7143 (cit. on p. 188).

[18] Erik Tjong Kim Sang and Johan Bos. „Predicting the 2011 dutch senate election results with twitter“. In: *Proceedings of the workshop on semantic analysis in social media*. Association for Computational Linguistics. 2012, pp. 53–60 (cit. on p. 189).

[19] Karolina Sylwester and Matthew Purver. „Twitter language use reflects psychological differences between democrats and republicans“. In: *PloS one* 10.9 (2015), e0137422 (cit. on p. 188).

[20] Simon Tong and Daphne Koller. „Support vector machine active learning with applications to text classification“. In: *Journal of machine learning research* 2.Nov (2001), pp. 45–66 (cit. on p. 196).

[21] Andranik Tumasjan, Timm Oliver Sprenger, Philipp G Sandner, and Isabell M Welpe. „Predicting elections with twitter: What 140 characters reveal about political sentiment.“ In: *ICWSM* 10.1 (2010), pp. 178–185 (cit. on p. 189).

Part V

Discussion and Evaluation

It was the Unicorn who summed up what everyone was feeling. He stamped his right fore-hoof on the ground and neighed, and then cried:

“I have come home at last! This is my real country! I belong here. This is the land I have been looking for all my life, though I never knew it till now. The reason why we loved the old Narnia is that it sometimes looked a little like this. Bree-hee-hee! Come further up, come further in!”

C.S. Lewis, *The Last Battle (1956)*[1]

12 Discussions and Conclusion

The use cases discussed in this thesis show the relevance of studying human connections, as technology now enables humans to be connected as never before. This fact leads us to consider what the consequences of such an integrated world are, where it is possible to know what almost anyone is thinking, doing, eating, watching, listening to, etc. The sharing of all this personal information and how it shapes other people's behaviours, perceptions and emotions was the main motivation to explore models of social contagion provided throughout this thesis. Our main research question is:

How can we create and validate computational models that explain social influence and contagion in social networks?

To answer this question, we explored many different scenarios in which the spread of perceptions, emotions and physical activity behaviour is happening inside of a network structure. In this section, we summarize the results of this work and discuss what can be done going forward. Section 12.1 contains the discussion of the work presented in the previous chapters. Section 12.3 presents a debate surrounding the limitations of this work and the nature of the methods used. Section 12.4 presents potential future work to follow up on with the results provided by this research. Lastly, Section 12.5 highlights the contributions of the author of this thesis in each of the chapters presented previously.

12.1 Discussion of the contributions of this thesis

The research presented here contains several contributions to the state of the art on the computational modeling of behaviours, perceptions and emotions spread in social networks. The results are relevant as a methodological source for other studies that aim to model how people can change based on their relations and interactions. We aim to provide all the methodological details necessary to replicate and extend the models presented in this thesis. Our first contribution is an improved model for social contagion based on differential equations systems, detailed in Chapter 2. Throughout the thesis we explore many ways of applying this model to different contexts. The major part of this thesis is dedicated to the spread of physical activity behaviour. We used data sets from various groups of people randing from children to adults and from a small data set of 25 participants to a much larger data set of 5,000 people. Here we sought to explore and shed light on the dynamics of the changes in the physical activity levels of the participants. These explorations can be seen in Chapters 4, 5, 7 and 8.

Besides the analysis of the data sets, other problems were tackled to explain the social contagion effect based on the data collected, and to understand the effect of applying interventions to improve the physical activity levels of people. To find good candidates for interventions in a social network, we explored the dynamics of the social network being built over time in a health promotion program, as seen in Chapter 6. We also presented different methods to find intervention targets in a network of children in Dutch schools based on either their centrality in the network, or on their personal characteristics. This is seen in Chapter 8.

In Chapter 3 we proposed a method to find the traits of a group of young adults in a physical activity behaviour experiment. We try to understand how the relations in this group of people affect their behaviours over time. For that, it is important to find the traits of the participants that contribute to the social contagion effect, i.e. openness and expressiveness. The method provides an alternative to the limitations of self-reports which are a very biased source of data about people, as well as a pure machine learning approach that can be very impersonal and unconnected to the knowledge provided by real life information.

The model presented in the first part of the thesis can also be extended to different applications beyond healthy lifestyle and physical exercise. This is shown in Chapter 9 where we used the same method applied to a context of the spread of messages in disaster situations. A cognitive model was built based on a temporal-causal approach to account for the responses of people to receiving warning messages regarding potential disasters as they happen. Another model using this very similar approach was built in Chapter 10 in order to understand and predict people's reactions when interacting on web media with political posts. This work is based on a multidisciplinary investigation that embodies neuroscience and social psychology to validate the models.

Lastly in Chapter 11, this thesis presented a classification mechanism using machine learning techniques for political messages from Twitter (or Tweets). This classification is relevant, as the models proposed required that the context and the inputs for the individual are quantified. In this work, we proposed a mixed method for the classification of the Tweets where many methods are applied and a "jury" defines if the content is related to politics or not. The high accuracy ($\sim 97\%$) of this method is remarkable and provides a consistent way of quantifying the political content spread in the Twitter platform.

Other attempts to simulate social contagion can be found in related literature. [18] presents a review in different models of people involved in a dynamic of opinions exchange. For the models presented in the work of Hegselmann, Krause, et al., the sum of all the connections between the agents ($a_{ij} \geq 0$) is 1 due to the mathematical framework of the models. The models presented in this thesis do not have the same structure. Even though it would be possible to change the scale by some normalization of the original weights for the edges, the representation wouldn't be realistic towards the original purpose of the work. Hegselmann, Krause, et al.'s models also permit that loops exist. That means that an agent i can disregard everyone's opinions, meaning that $a_{ii} = 1$ and $a_{ij} = 0$ for $j \neq i$. The result of [18] is a stochastic matrix for the connections, with all the values of the rows summing up to 1. In our work, we use deterministic equations to address the changes in the state of the agents and in the edges' weights. [18] reviewed many works with stochastic models [10, 21, 13, 14] and analyses a time-variant for situations "where agents put in the course of time more and more weight on their own opinion and less weight on the opinion of others". The only resemblance with some of the works listed by Hegselmann, Krause, et al. is developed by [20], where bounded confidence is used to define to whom the agents are going to consider the opinions. Some similarities with this approach can be seen in Chapter 8's model based on the works of [2] and [15] , where the threshold is used to define to what extent the behaviour of the agent is affected and prone to change. The same strategy can be found in [9, 23], all of them related to opinion adjustments, and not to behaviour, as proposed in part of this thesis. The temporal-causal model with differential equations presented in Chapter 2 and used throughout this thesis includes the time scale as a factor of the model together with a speed factor, not presented in any of the works explored in [18]. The nonlinear equations used are also deterministic, to which specific mathematical analysis are required. This work explores new and more advanced ways of addressing other problems with some similarities with the group opinion formation, but respecting the complexity of phenomena not considered as the target applications of the analysis performed by [18].

Flache et al. [12] also review three different classes of models for social influence, proposing new frontiers for future research. The models of this thesis are more related to the models of *assimilative social influence*, where individuals always influence each other towards reducing opinion differences. Chapter 8 presents a new approach by including a threshold to the influence received that would be closer to the class of models with *similarity biased influence*.

[12] claim that "more empirical work is needed testing and underpinning micro-level assumptions about social influence as well as macro-level predictions". This thesis presents many different ways of validation of a contagion model through empirical data in an integration of theory and reality, an empirically-based computational approach for social contagion model. As highlighted above, the model presented in Chapter 2 is distinguished from other classic models studied in previous decades, and proper mathematical and simulation analysis are provided to understand the characteristics of this temporal-causal model. [12] present three patterns for models of social influence with different convergence characteristics for the opinions of the agents, namely consensus formation, clustering and bi-polarization. The model presented in Chapter 2 could generate any of the three classes, depending on the context and on the dynamics of the network structure.

The model for political positioning presented in Chapter 10 and the model for disasters reactions presented in Chapter 9 are not validated in real data, but based on literature. As Flache et al. [12] state, referring to other attempts to model social influence, "in many contributions authors derive the theoretical assumptions they make both from fundamental psychological theories about social influence and from empirical evidence, thus 'calibrating' models in a broad sense". The political positioning model is an attempt to summarize the knowledge on political psychology and political neuroscience in order to fill some of the gaps on the understanding of political positioning shaping. This is a new approach for ABMs that could help to answer Robert Axelrod's question "if people tend to become more alike in their beliefs, attitudes, and behavior when they interact, why do not all such differences eventually disappear?" [1] by bringing the complexities of the human brain into consideration when evaluating the mechanisms of perceptions (or beliefs) formation.

Centola [8] data set has some similarities with the data presented in Chapters 5, 6 and 7. Even though Centola [8] gathers data of people in a health promotion network, there are strong differences between what was done by Centola [8] and this work. Firstly, [8] is interested in comparing two different network topologies in order to verify which one spreads the health-related behaviour farther. The participants of the experiments were designed to be part of one topology or another (a clustered lattice network or a random network). The topologies were fixed before the participant joined the network, and the number of neighbours was the same across conditions. Participants couldn't see the identity of their peers. Differently, this thesis does not take into account the network topology as a factor in the behaviour spread, even though the topology can be considered relevant to fully understand the social contagion phenomenon. The data set used in this thesis contains two populations and permitted the participants to connect with others without restrictions. Therefore, the resulted network contains all the connections desired by the people in it without limitations. For this reason, Chapter 6 explores the formation of the network in order to verify what type of network is created when it is built from the beginning.

Most of the effort presented in this work is towards generating the network of people in real life rather than defining how the network should be set in order to better perform the spread of behaviour. Another consistent difference is in the fact that the health-related behaviour adopted by the participants in [8] is the adoption of a certain program and not an objective measure of their physical activity as done in this thesis.

Centola [7] presents a similar set up for the network system to the one explained above in [8], but aims to verify if connecting people with similar characteristics (creating homophilic ties) would increase or decrease the chances of obese and non-obese people to adopt a diet agenda after being exposed to other peers adopting it. The network structure was fixed, and the individuals were distributed in two groups: a homophilious group where the characteristics of the individuals (BMI, age, sex, etc.) were taken into account to populate the network, and, a random group where the individual attributes were not considered to place the nodes in network. [7] showed that homophily significantly increased adoption both among the obese and non-obese members of the community. The results presented by [7] are really relevant to strengthen the concept of homophily in social relations and spread of behaviour. This thesis presented some homophily analysis in a network built for a health promotion program in Chapter 6, but as stated above, the network wasn't fixed. Besides, the aim of the work wasn't to compare how far the diffusion of one specific program adoption goes. The participants of the data set presented in Chapter 6 had the option to opt in or out the network system alone. The results found in this thesis account for the quantified spread of behaviour using the PAL of the individuals over time. Therefore, instead of a binary condition (adopt or not adopt a diet agenda, for example), this thesis presented a concrete work on studying the quantification of the behaviour spread, a new exploration in the vast literature investigated in this field.

12.2 Research questions

The aim of this thesis is to understand *how we can create and validate computational models that explain social influence and contagion in social networks*. To achieve this goal, 3 subquestions were raised to guide our investigation.

1. How can we design and use temporal-causal models based on networks to better understand and describe social contagion taking into account personal characteristics?

2. How can we predict changes in behaviour using the relationships and data related to physical activity and how can we measure social contagion in a social network regarding people's PAL?

3. What are potential applications for modeling behaviour in social networks and how can we apply the knowledge of temporal-causal network modeling using different contexts and methodologies?

The following subsections aim to explain how the research questions connect with the work presented throughout this thesis.

12.2.1 Research question 1

How can we design and use temporal-causal models based on networks to better understand and describe social contagion taking into account personal characteristics?

We have shown that temporal-causal models are very suitable for modeling social contagion in networks. The spread of any sort of behaviour, perception or emotion should have a temporal dimension to account for the dynamics of the change. That means, it takes time for someone to adapt to new perceptions, or to be affected by an external agent's emotions. The causal effects of the interactions between the nodes is also relevant, as the contagion is only possible between agents connected to each other and able to expose themselves reciprocally. Our proposed model is in line with other social contagion studies that show that "the empirical research has tended to confirm that the hypothesis that human behaviour clusters in both space and time even in the absence of coercion and rationale" [22].

For a model to be useful, it is essential that it provides a correct representation of the process in reality. Finding a mathematical representation for social contagion is a challenge, as there is no guarantee that the structure chosen to explain a certain event is the most suitable one. To provide an accurate representation of reality, effort is required to validate it. This can be done by showing that empirical data can be explained by the proposed model. Differential equations are used in many other investigations with the intent to provide time-dependent dynamics of some phenomenon [11, 19]. In Chapter 2 we presented our model to account for the social contagion of behaviour. For this we have adapted a previous model developed by [4] and provided evidence that the model was stable enough to explain contagion of behaviours, perceptions and emotions on a bigger scale than the previous model.

This was a good improvement, mainly because it removed the instability of the changes in the states of the individuals in the simulation of bigger groups of nodes. The new model provides more realistic results as the changes in the states are not abrupt or out of the limits defined for the scenarios.

We explored many methods to model the network connections of the individuals in our data sets. To create a realistic model that accounts for the changes caused by the relationships of people, it is also relevant to consider how the connections between people are determined. The relationships of a person are the way through which social contagion happens. It is through the interaction with friends, family and other people that behaviour, perceptions and emotions are changed. We proposed a few methods to describe and quantify the relationship of groups of people in different scenarios in order to fill the model with a reliable estimate of how strong the ties are between the individuals in different data sets. In Chapter 4, we applied some questionnaires to a group of young adults to define what the network structure was during the experiment. The task was to define how strong the connections are, and who influences whom in an oriented network. That is, two peers can have different levels of relationship, and therefore levels of influence, when looking at how the first affects the second, and vice-versa. In Chapter 5 however, the edges of the network structure are not oriented. The social network structure provided by the health promotion program did not permit that we assess the level of relationship between the participants in detail, providing only the friendship requests and the acceptance of them. For this data set the edges were dynamic. That is, they were generated over time, as the program was being run. Because of the dynamics of the connections, we proposed a method where a network is generated per each day of the experiment, so the changes in the network structure can be reflected in the spread of contagion and therefore in the model proposed. To better understand how the structure of the data set used in Chapter 5 is changed, we investigated the social network characteristics over time in Chapter 6. In Chapter 6 we presented the dynamics of the connections in this network, the number of edges increasing over time and the changes in the position of nodes, as those are factors that can affect the spread of behaviour and potential predictions using the contagion model. All the studies on the network connections and structure dynamic are a contribution to the building of reliable and consistent models. The network structure is extremely relevant for the validity of the results of the simulations using the models proposed in this thesis. The design of realistic network edges will certainly provide good understanding and more accurate predictions for the state of the network.

In this thesis we have shown a few methods to gather information about personal traits and how to quantify them. Besides knowing how the network structure of the people is, it is also relevant to know two specific personality traits of the individuals: openness and expressiveness. These traits are the indicators of the potential change in a person (openness) and of the level of extroversion and articulateness (expressiveness), and are important to determine the amount of influence received and sent to the neighbors of an individual. Therefore, finding these two traits is also important to make the proposed contagion model more accurate. In Chapter 3 we proposed a new method to combine the use of a Big Five questionnaire applied to a group of young adults with some optimized algorithms to better define the personal characteristics in the context of our contagion model for spread of behaviour. These methods try to improve the results obtained and shown in Chapter 4, where the self-reported questionnaires were used alone to define the openness and expressiveness of the participants.

We have shown that we can use dynamic models to analyze which network questions are more relevant to ask in order to obtain a realistic representation of the social connections of individuals. In Chapter 8 we used self-reports from children in the schools from The Netherlands to build the network structure. The children were asked in many ways how they perceive their classmates. 16 questions in total were used to understand who the popular children are, who dresses well, who is a good person to talk about food, who the friends of the participants are, etc. From these nomination questions we came out with three different oriented networks based on different sets of questions. We were interested in learning if raising less questions to the children would provide us similar networks. This was an attempt to diminish an overload on the participants by asking less questions and keep the confidence about their connections built from the data. Also in Chapter 8 we used a different mathematical model to account for the change in the behaviour of the children. This mathematical model turned out to be very sensitive to the threshold parameters, which resulted in necessary adaptations to keep the results realistic. The original model for this work is based on the research described in [2, 16].

Finally, this thesis presented a body of work on how to model a social network for social contagion using different mathematical methods. Some of the problems encountered when modeling the spread of behaviour are studied in Chapters 2 , 4, 5, 6, 7 and 8.

12.2.2 Research question 2

How can we predict changes in behaviour using the relationships and data related to physical activity and how can we measure social contagion in a social network regarding people's Physical Activity Level (PAL)?

In the previous research question we have shown that this thesis provides methods and ways to model a contagion model that accounts for the spread of behaviour, perceptions and emotions. The more realistic a model becomes, the more it becomes a reliable measurement for predictions and interventions.

We have shown in Chapter 4 that our social contagion model can predict around 80% of the cases if the PAL of an individual is going to increase or decrease based on the information provided for the network structure and the daily PAL collected for a period of one month. The high accuracy for the contagion model in this scenario implies that the social spread of behaviour is real and can be modeled using the model proposed. The results of this chapter also provided the information on the amount of physical activity behaviour spread in a social network of young adults.

Chapter 7 shows that the model presented in Chapter 2 performs better at describing the pattern seen in a data set of approximately 2,400 people than a simple linear model. We have shown in this chapter that some of the dynamics of the PALs in the network can be explained by social contagion processes. First we provided a simple linear model that has been derived by a regression analysis. Then we compared with the social contagion model mixed with a steady increase in the PAL due to the community effect (see Chapter 5) and showed that the model outperforms the simple linear model. As far as we know, this is 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.

To predict changes in the behaviour of a group of children in Dutch schools, we presented a study on the many scenarios where interventions are applied to the participants. We have shown that selecting individuals based on their socio-economic situation to increase their PALs is comparable to using social network optimization algorithms in the data set collected by the MyMovez Project [3]. The results are shown in Chapter 8, which presents the exploration on the spread of behaviour in a group of schools in the Netherlands. The children were tracked using Fitbit devices and answered questions regarding their social relationships with their classmates. In this work we generated the initial network and state for each of the participants and applied 5 different interventions in a percentage of the class increasing the PAL of the individuals that were selected. This research demonstrates how we can use data to measure the social contagion and provide predictions according to the model used.

12.2.3 Research question 3

What are potential applications for modeling behaviour in social networks and how can we apply the knowledge of temporal-causal network modeling using different contexts and methodologies?

Our hypothesis is that the use of temporal-causal network models can be extended to many other contexts and applications that involve people and the spread of behaviours, perceptions and emotions through social ties. As the scope of the model presented is not restricted to the physical activity, we dedicated a part of the thesis to show how to apply the same methods to the spread of messages in disaster situations (Chapter 9), and for cognitive modeling of political positioning changes while interacting with web media posts (Chapter 10).

Chapter 9 explores the spread of emotions in situations of disaster. For this, an Agent-Based Model was introduced accounting for the expressed emotion and potential actions of a person when they receive a message warning about a potential disaster scenario. Many simulations were performed to show that the model simulates what it is expected to happen when a tense and negative phone call or a happy and positive text message are received by the agent.

Chapter 10 presents a cognitive model which is also based on differential equations and temporal-causal relations, to explain how a political perception (or positioning) of a person is affected by exposure to web media tweets. This paper is based on research results from neuroscience and social psychology, and also considers that the social contagion of opinions plays an important role in the shaping of political positioning. Chapter 11 is a complementary work of Chapter 10, as it provides a machine learning classifier to categorize the tweets that would be used as inputs for our cognitive model.

Thus, we have shown that the model can be extended to other contexts where the social contagion is observed. Many other applications can be derived from the methods used in this thesis, incorporating the social contagion whenever it is relevant to explain the phenomenon studied. Section 12.4 of this current chapter will give good examples of potential future work following up on the methods and results obtained here.

12.3 Limitations of this work

The work presented in this thesis is relevant for the understanding of the social contagion processes and the modeling of social networks to explain and predict the spread of behaviours, perceptions and emotions. The task of creating realistic models is a very complex task that requires knowledge from the other fields. We dedicate this section to bring up the limitations of this work that require special attention.

When evaluating the spread of PAL, we were faced with many problems to define a fair and solid method to understand the results. To validate a contagion model is a difficult task that requires a good data set with sufficient participants, reliable information about personal characteristics and a complete longitudinal trace of the behaviour studied (in this case, the physical activity behaviour). The work presented in this thesis tried to address all these problems, but still faces the problems that the data sets don't provide all the necessary information. Most of the work here could be improved by collecting data for a longer time period. Unfortunately, to reach all the requirements of a perfect model would require more time and resources than was available.

In Chapter 4 we collected data over 30 days of a small group of people (25 young adults). One could claim that more data should be necessary to validate the results, or that a longer period of time should be taken to provide a better longitudinal perspective in the changes. Although these remarks are relevant and true, collecting behavioural data is not an easy task. For Chapter 4, the short period of time was relevant to reduce the seasonal effect on the data, but didn't avoid the lack of other potential explanations for the changes in the PAL of the participants, such as weather, injuries, holidays or other motives for people to exercise less (or more) than usual.

Although Chapter 4 presents a small data set, more explorations were done in Chapters 3, 5 and 7 using a data set of originally 5,000 participants over a time period of more than 300 days. In this data set, we had to acknowledge the fact that a high number of participants had dropped out or didn't record the data properly, causing missing data and sparseness of the data points. That means, even with a lot of data available, we struggled to build a reliable analysis. The conclusions were based on a very strict cleaning process to filter out any dirt from the data set (i.e. participants who dropped out, individuals that started the plan in other time windows, etc.), but it certainly limited our understanding of the whole population's behaviour. Therefore, we acknowledge that the data management and filtering were bottlenecks that could have an influence on our analysis and outcomes. Nevertheless, we worked with the aim to provide reliable explanations for all the decisions made and for all the steps taken to derive the final results.

The approach used to define personal characteristics in our models can be considered too generic. It would require additional work on how to define the personal traits (e.g. openness and expressiveness) to create a more reliable metric for these characteristics. In Chapter 4 for instance, we adapted the Big-5 questionnaire [17] to account for the openness and expressiveness of the participants. In Chapter 7 we set all the expressiveness and opennesses to 0.5 due to the lack of information about the personal traits. We believe that better results would be obtained if this data was provided through a very specific, well tested and validated process. Due to the difficulties on getting data related to the personal traits of individuals, we proposed a new method in Chapter 3. This method needs further investigation and more study using a different data set.

The applicability of our models should also be understood from the perspective of the context. Modeling behaviour change in groups is highly dependent on the context, i.e. the time of the year when data was collected, the level of details provided by the participants, the commitment of the participants to the data collection, the socio-economic characteristics of the population, etc. Therefore, it is important to be aware that our models and our validation studies are relevant in the context of the research we have done. Results cannot be generalized to any population, so for other populations it is important to verify if the characteristics of the individuals within the population, network and tools are similar to the ones used in our work.

Some of the explorations presented in this thesis do not have any data, due to the nature of the models. This is the case as seen in Chapters 9 and 10. Cognitive models are based on brain states that are not easily accessible by any existing tool so far. In these cases, we created data based on our assumptions and the expected functioning of the states modeled. It is important to acknowledge that as the state of the art in these fields become more advanced and provide better details about the phenomena studied, changes in the models may be needed to incorporate the new discoveries.

12.4 Future Work

This thesis presents an exploration on the modeling of social contagion in social networks within different contexts. This section describes potential future investigations based on the work presented in the previous chapters.

In Chapter 2, we presented a new proposal for the contagion model. In this proposal, the speed factor calculation is adapted to avoid unrealistic changes in the states of the nodes. Further investigations could explore how the new speed factor alternatives affect the results of previous research, and how they can be combined with the model for emotion contagion spirals from [5]. It would also be an interesting work to verify how this model can be validated in other contexts such as the perception and emotions' contagion.

The new method to define traits of people in a network for sharing their physical activity achievements presented in Chapter 3 could be improved in several ways. As the data set is quite small, we would have to test the method for finding the personality traits of people using optimization and machine learning algorithms within a bigger data set. This method can also be very useful for other applications, such as for the personal traits of users of web media, or for finding relevant traits of people in programs for behaviour changes (i.e. drug addiction, leadership training, etc.). Using the method proposed could help to divulge other results and therefore improve the understanding of the advantages and limitations of this method. Further applications using the method proposed could be also developed. For instance, an application for defining potentially depressed people in a network based on self reports and on machine learning techniques could be helpful identifying nodes that can be a support for people who are struggling with depression.

In Chapter 4 we collected data of the PAL of a group of young adults over a period of 30 days in order to understand the dynamics of the social contagion of behaviour in their network of relations. A longer and bigger set up would be required to provide stronger findings that would support the spread of the behaviour in the network. Other improvements could include considering the changes in the connections over time to account for the real changes in the relationships, and to try to apply interventions in some of the nodes to propagate a positive behaviour faster. The results presented are good indicators that the social contagion can explain changes in a group of people. Therefore, it can be expected that applications seeking to find good candidates for interventions in order to improve the lifestyle and health of a population can take advantage of the results obtained by this research. The social contagion model can be used to predict the increase/decrease of the PAL and provide inputs for a potential change in the structure of the network, or some interventions in the people's states.

Even though the data set shown in Chapters 5 and 7 provides a much bigger sample with more people and a longer experiment, many challenges are still to be tackled in future investigations. It is important to study the effect of other factors on the physical activity level, such as the community size and structure. That way, research can further uncover phenomena that are at the basis of the beneficial effects of online social networks in health promotion programs. For this, the results from Chapter 6 can be a good starting point. The study of the dynamics of the network creation provides insights to how the ties are formed and how the position of the nodes in the network are changed. The combination of degree measurements for the nodes and the density of the ego-network can be used to identify people who are potentially influential in their network in further work. The results show that continuously monitoring the evolution of a network is important to identify such people. Future work can use the outcomes of the network analysis to form the basis for automated (health) interventions that exploit the social network for changing behaviours of individuals. This could lead us to future discoveries about leadership, spread of emotions or any other application related to a network's topology and dynamics.

Besides modeling social contagion and the spread of behaviour, applying interventions in a group of people is also a task that requires deep investigation. In Chapter 8 we use an adapted model to select children from Dutch schools with the aim to improve the PAL of the whole group. Future work has to be carried out in order to verify if our proposed contagion model is a better predictor for the behavioural change in the data set than the currently used diffusion model. Other strategies can also be drawn to define better ways to select the intervention nodes in the network. The data set used in this work can be studied further and additional findings could better explain the dynamics of the network, or improve the mapping of the relationships questions to a network. The use of the method proposed in Chapter 3 could possibly also be used in this data set to find the personality traits of the children.

The results provided by the research in Chapter 8 can be used to define a good set of questions in order to generate the network of a classroom of children and teenagers from the same school. It also provides knowledge of what happens when some intervention is applied in the network. Therefore, more research can still be done to verify if the interventions caused the changes that were expected by the model proposed, and if so, how can these results be repeated in other groups (i.e. different social-economic situation contexts).

Additional work can also be carried out by finding other contexts to apply the temporal-causal modeling method. Chapters 9 and 10 proposed two cases for the use of contagion principles, and from these models other real applications can also be created. The model for the spread of messages in disaster situations could be extended to simulate a population in a situation of disaster or panic. The political cognitive model can also be extended such that it is suited for a network to investigate the spread of political opinions in social networks. Further validation using empirical data would also be useful to provide stronger evidence that the models are correct. The work presented in Chapter 11 is a starting point for the creation of real world applications using the model created in Chapter 10.

Understanding and modeling social contagion still has many unanswered questions and potential applications. The results obtained by this research are useful for future applications for health lifestyle promotion, understanding perceptions and modeling the spread of emotions involving new technologies that aim to use the personal traits and the connections of people to encourage them to be more active, eat healthier or both. These technologies could be extended to applications that account for loneliness, depression, drug addiction, or any other behaviour that can be considered shaped within a social context.

12.5 Contributions and chapters overview

This section presents the contributions of the author in each of the research works presented throughout the thesis and the chapters overview.

Chapter 2 has been published as: ***Fernandes de Mello Araújo E.***, Treur J. (2016) Analysis and Refinement of a Temporal-Causal Network Model for Absorption of Emotions. In: Nguyen NT., Iliadis L., Manolopoulos Y., Trawi´nski B. (eds) Computational Collective Intelligence. ICCCI 2016. Lecture Notes in Computer Science, vol 9875. Springer, Cham.

My contribution to this chapter includes the proposal of a new measurement for the speed factor in the mathematical equation that accounts for the aggregated impact of the other nodes. I also coded the simulated scenarios, plotted the graphics, discussed the results and wrote part of the manuscript in partnership with the other author. This paper presents an improvement for the emotion contagion model previously created by Bosse et al. [6] that is used to account for the spread of behaviour in many other chapters of this thesis.

Chapter 3 has been 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).

My contribution to this chapter includes the proposal of the new method to define personality traits, the coding of the analysis and algorithms, the plotting of the graphics, the discussion of the results and writing of the manuscript. In this chapter we present a method that combines intake questionnaires with some optimization algorithm in order to better tune the traits of openness and expressiveness of a group of young adults in an experiment of sharing behaviour.

Chapter 4 has been published as: ***Eric F. M. Araújo***, Anita V. T. T. Tran, Julia S. Mollee, Michel C. A. 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.

My contribution to this chapter includes the proposal of the experiment, the supervision of the data collection, the configuration of devices to track the PAL of the participants, the coding and analysis of the results, the discussion of the results and the writing of the manuscript. This work aimed to use a social contagion model based on temporal-causal relations to predict the changes in the PAL of 25 young adults from the same course. The results showed that in more than 80% of the cases, the contagion model is a good predictor for the increase or decrease of the PAL based on the social relations and personality traits of the individuals in the data set.

Chapter 5 is an extension of the paper published as: A. Manzoor, J. S. Mollee, ***E. F. M. Araújo***, A. T. V. Halteren and M. C. A. Klein. Online Sharing of Physical Activity: Does It Accelerate the Impact of a Health Promotion Program?, 2016 IEEE International Conferences on Big Data and Cloud Computing (BDCloud), Social Computing and Networking (SocialCom), Sustainable Computing and Communications (SustainCom) (BDCloud-SocialCom-SustainCom), Atlanta, GA, 2016, pp. 201-208.

My contribution to this chapter includes the data analysis, involving the understanding of the data set, the management of the data, the building of the networks and the filtering of nodes, the coding of the tasks to obtain the results, the discussion of the results and the writing of the manuscript. This chapter presents a statistical analysis of a data set of a health promotion program where the participants could opt to participate in a community for sharing their PALs over time with their friends (connections). We were interested in understanding what the effect of being part of a community is and what the effect of being connected to other participants in the same experiment is.

Chapter 6 has been published as: ***de Mello Araújo, E. F.***, Klein, M., van Halteren, A. (2016, November). Social Connection Dynamics in a Health Promotion Network. In International Workshop on Complex Networks and their Applications (pp. 773-784). Springer, Cham.

My contribution to this chapter includes the discussion of the research questions, data manipulation, coding of the analysis, discussion of the results and writing the manuscript. This work is a social network analysis of the dynamics of the edges in a data set of people connected in a health promotion program, where they could share their daily PAL with their connections. We look for leaders in the network and the dynamics of a network generation.

Chapter 7 has been published as: Mollee, J. S., ***Araújo, E. F.***, Manzoor, A., van Halteren, A. T., Klein, M. C. (2017, March). Explaining Changes in Physical Activity Through a Computational Model of Social Contagion. In Workshop on Complex Networks CompleNet (pp. 213-223). Springer, Cham.

My contribution to this chapter includes the discussion of the research methods, the coding of the simulations, the generation of the graphics and results, the discussion of the findings and the writing of the manuscript. This work tries to explain the changes in the PAL of a group of people in a health promotion program by using the social contagion model combined with the effect found in being a part of the community. The results showed that the model outperforms a simple linear model.

Chapter 8 has been published as: **Araújo, E.**, Simoski, B., Woudenberg, T., Bevelander, K., Smit, C., Buijs, L., Klein, M. and Buijzen, M. *Using Simulations for Exploring Interventions in Social Networks - Modeling Physical Activity Behaviour in Dutch School Classes.* In Proceedings of 8th International Conference on Simulation and Modeling Methodologies, Technologies and Applications (SIMULTECH 2018), pages 414-425. ISBN: 978-989-758-323-0.

My contribution to this chapter includes the discussion of the research questions, the definition of the methods, the coding of the simulations, the generation of the results, the discussion of the findings and the writing of the manuscript. This work aims to find the best intervention methods to increase the overall PAL of a group of children in a Dutch school class. This paper was nominated to the best student paper award of the conference.

Chapter 9 has been published as: ***Fernandes de Mello Araújo E.***, Franke A., Hosain R.W. (2017) A Temporal-Causal Model for Spread of Messages in Disasters. In: Nguyen N., Papadopoulos G., Jedrzejowicz P., Trawinski B., Vossen G. (eds) Computational Collective Intelligence. ICCCI 2017. Lecture Notes in Computer Science, vol 10449. Springer, Cham 2017.

My contribution to this chapter includes the discussion about the model, as well as the formulation of the model, the discussion of the results and the writing and review of the manuscript. This work presents a cognitive model for the reactions caused by messages and calls alerting a disaster scenario. The model is based on neuroscience and psychology findings and is modeled to predict people's behaviour within these contexts.

Chapter 10 has been published as: ***E. F. de Mello Araújo*** and M. Klein, "A computational cognitive model for political positioning and reactions in web media," 2017 IEEE 16th International Conference on Cognitive Informatics & Cognitive Computing (ICCI*CC), Oxford, 2017, pp. 414-422.

My contribution to this chapter includes the literature review and the elaboration of the model, the coding and simulation of the model, the discussion of the findings and the writing of the manuscript. This work presents a cognitive model that represents the political positioning change of a person when interacting with Tweets from web media. The work is based on neuroscience and social psychology findings, and is a great contribution to the understanding of how we can map changes in people's positioning.

Chapter 11 has been published as: ***Fernandes de Mello Araújo E.*** and Ebbelaar D. (2018). *Detecting Dutch Political Tweets: A Classifier based on Voting System using Supervised Learning*. In Proceedings of the 10th International Conference on Agents and Artificial Intelligence - Volume 2: ICAART, pages 462-469.

My contribution to this chapter includes the conception of the method to be used, the supervision of the student that collected the data and performed the machine learning techniques, the discussion of findings and the review and writing of the manuscript. This work provides a very accurate classifier for the task of distinguishing political and non-political tweets.

The following article was published but not included in the thesis: Mollee J.S., ***Araújo E.F.M.***, Klein M.C.A. (2017) Exploring Parameter Tuning for Analysis and Optimization of a Computational Model. In: Benferhat S., Tabia K., Ali M. (eds) Advances in Artificial Intelligence: From Theory to Practice. IEA/AIE 2017. Lecture Notes in Computer Science, vol 10351. Springer, Cham.

Bibliography

[1] Robert Axelrod. „The dissemination of culture: A model with local convergence and global polarization". In: *Journal of conflict resolution* 41.2 (1997), pp. 203–226 (cit. on p. 207).

[2] Rahmatollah Beheshti, Mehdi Jalalpour, and Thomas A. Glass. „Comparing methods of targeting obesity interventions in populations: An agent-based simulation". In: *SSM - Population Health* 3 (2017), pp. 211–218 (cit. on pp. 207, 211).

[3] Kirsten E Bevelander, Crystal R Smit, Thabo J van Woudenberg, et al. „Youth's social network structures and peer influences: study protocol MyMovez project–Phase I". In: *BMC public health* 18.1 (2018), p. 504 (cit. on p. 212).

[4] Tibor Bosse, Rob Duell, Zulfiqar A Memon, Jan Treur, and C Natalie van der Wal. „Agent-based modeling of emotion contagion in groups". In: *Cognitive Computation* 7.1 (2015), pp. 111–136 (cit. on p. 209).

[5] 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), pp. 48–67 (cit. on p. 214).

[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. 217).

[7] Damon Centola. „An experimental study of homophily in the adoption of health behavior". In: *Science* 334.6060 (2011), pp. 1269–1272 (cit. on p. 208).

[8] Damon Centola. „The spread of behavior in an online social network experiment". In: *science* 329.5996 (2010), pp. 1194–1197 (cit. on p. 208).

[9] Guillaume Deffuant, David Neau, Frederic Amblard, and Gérard Weisbuch. „Mixing beliefs among interacting agents". In: *Advances in Complex Systems* 3.01n04 (2000), pp. 87–98 (cit. on p. 207).

[10] Morris H DeGroot. „Reaching a consensus". In: *Journal of the American Statistical Association* 69.345 (1974), pp. 118–121 (cit. on p. 206).

[11] Keisuke Ejima, Kazuyuki Aihara, and Hiroshi Nishiura. „Modeling the obesity epidemic: social contagion and its implications for control". In: *Theoretical Biology and Medical Modelling* 10.1 (2013), p. 17 (cit. on p. 209).

[12] Andreas Flache, Michael Mäs, Thomas Feliciani, et al. „Models of Social Influence: Towards the Next Frontiers." In: *Journal of Artificial Societies & Social Simulation* 20.4 (2017) (cit. on p. 207).

[13] Noah E Friedkin and Eugene C Johnsen. „Social influence and opinions". In: *Journal of Mathematical Sociology* 15.3-4 (1990), pp. 193–206 (cit. on p. 206).

[14] NE Friendkin and EC Johnsen. „Social influence networks and opinion change". In: *Adv Group Proc* 16 (1999), pp. 1–29 (cit. on p. 206).

[15] Philippe J. Giabbanelli, Azadeh Alimadad, Vahid Dabbaghian, and Diane T. Finegood. „Modeling the influence of social networks and environment on energy balance and obesity". In: *Journal of Computational Science* 3.1 (2012), pp. 17–27 (cit. on p. 207).

[16] Philippe J. Giabbanelli, Azadeh Alimadad, Vahid Dabbaghian, and Diane T. Finegood. „Modeling the influence of social networks and environment on energy balance and obesity". In: *Journal of Computational Science* 3.1–2 (2012), pp. 17–27 (cit. on p. 211).

[17] 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. 214).

[18] Rainer Hegselmann, Ulrich Krause, et al. „Opinion dynamics and bounded confidence models, analysis, and simulation". In: *Journal of artificial societies and social simulation* 5.3 (2002) (cit. on pp. 206, 207).

[19] Alison L Hill, David G Rand, Martin A Nowak, and Nicholas A Christakis. „Infectious disease modeling of social contagion in networks". In: *PLOS computational biology* 6.11 (2010), e1000968 (cit. on p. 209).

[20] Ulrich Krause. „A discrete nonlinear and non-autonomous model of consensus formation". In: *Communications in difference equations* 2000 (2000), pp. 227–236 (cit. on p. 207).

[21] Keith Lehrer. „Social consensus and rational agnoiology". In: *Synthese* 31.1 (1975), pp. 141–160 (cit. on p. 206).

[22] 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. 209).

[23] G. Weisbuch, G. Deffuant, F. Amblard, and J.-P. Nadal. „Interacting Agents and Continuous Opinions Dynamics". In: *Heterogenous Agents, Interactions and Economic Performance*. Ed. by Robin Cowan and Nicolas Jonard. Berlin, Heidelberg: Springer Berlin Heidelberg, 2003, pp. 225–242 (cit. on p. 207).

List of Figures

2.1 Full channel connections for 3 to 18 members. 29
2.2 Three types of relations for the speed factor depending on number of nodes: left: η_A increasing linearly as the number of members increases middle: η_A increases up to a limit, due to a weighted speed factor right: η_A increases according to an advanced logistic function 30
2.3 Comparison for scenario 1(a) between the 3 speed factors 35
2.4 Scenario 3 and the different speed factors used to calculate the emotion levels 35

3.1 Methodology for tuning the questionnaire questions weights (left). Methodology for tuning the traits for each participant (right). 45
3.2 Mean and standard deviation for the 100 best solutions using the grid search. 47
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). 48
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). 48

4.1 Network structure of the experiment. 60
4.2 Physical Activity Level achievement of participant EXP01 during 30 days. 61
4.3 Graphs with trendlines of real data and simulation results of the model. 65

5.1 Number of nodes in each of the components 75
5.2 The network components used for the community analysis in scenario 2. 77
5.3 Physical activity levels (PALs) of connected vs. not connected participants during 84 days of the intervention, including linear trendlines. . 79
5.4 Physical activity levels (PALs) of community vs. non-community participants during 84 days of the intervention, including linear trendlines. . 80

6.1 Degree distribution in days 1, 100 and 336 (top) and p value for slope, R squared and standard error (bottom) 96
6.2 Degree of the richest nodes (left) and number of poor nodes, with $degree \leq 2$ (right) 97
6.3 Percentage of the nodes in the largest component. All nodes (lower red line) and nodes with degree larger than 1 (higher green line) 98
6.4 Components analysis. Number of components in the graph over the time (upper), the correlation between the size of the component and the frequency of the size (days 1, 165 and 335) (middle) and the parameters from the linear regression for all time steps (lower) 99

6.5 Mean of all centrality measures for all nodes at each time step (six graphics on top) and average shortest path (bottom) 100
6.6 Centrality for most central nodes. Betweenness (upper left) and closeness (upper right) for the 20 nodes with the highest combined centrality. Lower graphic shows the combined centrality measures 100
6.7 Homophily according to the BMI (left) and gender (right) of the nodes 101
6.8 The dynamic relation between ego network density and nodes' degrees 102

7.1 Predictions of the simple linear model (left) and the combined model (right). 115
7.2 Average predictions of the two models (green: linear, blue: combined), and the empirical data (red). 116

8.1 Grid search for the general questions. 135
8.2 Simulated annealing space of search explored for the general questions. 135
8.3 Simulation of average PA for all the strategies for the different percentages of targets in the three rows. The left column shows the results for the network generated with all questions. The center column is for the network generated with the general questions set. The right column shows the results for the network generated using the friendship question alone. 137

9.1 Temporal causal network model 150
9.2 Experimental dataset: Easily influenced person 153
9.3 Experimental dataset: Difficult to influence person 154
9.4 Scenario 1: all states 156
9.5 Scenario 1: input and output states 157
9.6 Scenario 2: all states 157
9.7 Scenario 2: input and output states 158
9.8 Scenario 3: Outputs of person 1 as inputs for person 2 159
9.9 Simulation for mathematical analysis 160

10.1 Computational model for political changes and reactions towards web media messages. 171
10.2 Simulation of conservative message to liberal agent. 177
10.3 Simulation of conservative message with neutral sentiment to liberal agent. 178
10.4 Interpretation of the message for agent 1, message B. 178
10.5 Changes on political position for agent 3. 179
10.6 Changes on political position for agent 3. 180
10.7 Changes on political position for agent 3. 180

11.1 Tagging application 193
11.2 Bag-of-words feature representation 193
11.3 Training set accuracies 197
11.4 Classification report for political (Y) and non-political (N) tweets . . . 199

List of Tables

4.1 Parameters for personal and social characteristics. 55
4.2 Statements measuring expressiveness and openness. (Reversed statements are marked by *.) 59
4.3 Structural characteristics of the network. 61
4.4 Comparison of line tendency between contagion model and real data[a] and comparison of line tendency between contagion model and real data after removing trendlines with a slope less than 0.03 [b]. Correct predictions[c] are shown in the last row. 62
4.5 Mean Squared Errors of prediction lines 63

5.1 Meta-Properties of Intervention and Control Groups of Scenario 1 . . . 74
5.2 Meta-Properties of Selected Components (Intervention Group) of Scenario 2 76
5.3 Characteristics of Participants in Selected Components (Intervention Group) of Scenario 2 78
5.4 Detailed Characteristics of Component E 78
5.5 Analysis using multiple linear regression for scenario 1 80
5.6 Analysis using multiple linear regression for scenario 2 81
5.7 Physical activity level data in long format for community and non-community individuals for scenario 2 81
5.8 Analysis of scenario 1 using generalized linear regression 82
5.9 Analysis using linear mixed effects modeling for scenario 1 82
5.10 Comparison of standard linear regression model with random intercept model for scenario 1 83
5.11 Analysis of scenario 2 using generalized linear regression 83
5.12 Analysis using linear mixed effects modeling for scenario 2 84
5.13 Comparison of standard linear regression model with random intercept model 84

7.1 Model evaluations 116

8.1 Density of the networks generated by the different subsets of questions. 128
8.2 Hamming Distance between the three graphs generated by the subsets of questions 129
8.3 Ratio between boys and girls per class and amount of targets for each fraction selection 133
8.4 Best threshold and factor of change for the three networks generated with (1) all questions, (2) general questions and (3) friendship question 136
8.5 Differences between initial and end mean PA for the network generated with all questions and 20% of the nodes selected for intervention. . . . 136

8.6 Differences from the mean PA for each intervention to the simulations without any intervention. The results are referred to the network generated from all the questions combined. 137

9.1 External and internal states of the model 151
9.2 The resulting MSE, steepness and threshold after tuning. 155
9.3 Initial values of the input states of scenarios 1 and 2 155
9.4 Initial values of the input states of the two persons in scenario 3 158

10.1 States in the model 174
10.2 Traits of the simulated agents 176
10.3 Initial states for the simulated agents 176
10.4 Messages used for the simulations 177

11.1 Keywords used to filter the tweets collected 191
11.2 Feature extraction example 191
11.3 Collected tweets 192
11.4 Cross validation results 198
11.5 Test set results 198

SIKS Dissertation Series

2011-01 Botond Cseke (RUN), Variational Algorithms for Bayesian Inference in Latent Gaussian Models

2011-02 Nick Tinnemeier (UU), Organizing Agent Organizations. Syntax and Operational Semantics of an Organization-Oriented Programming Language

2011-03 Jan Martijn van der Werf (TUE), Compositional Design and Verification of Component-Based Information Systems

2011-04 Hado van Hasselt (UU), Insights in Reinforcement Learning; Formal analysis and empirical evaluation of temporal-difference

2011-05 Bas van der Raadt (VU), Enterprise Architecture Coming of Age - Increasing the Performance of an Emerging Discipline.

2011-06 Yiwen Wang (TUE), Semantically-Enhanced Recommendations in Cultural Heritage

2011-07 Yujia Cao (UT), Multimodal Information Presentation for High Load Human Computer Interaction

2011-08 Nieske Vergunst (UU), BDI-based Generation of Robust Task-Oriented Dialogues

2011-09 Tim de Jong (OU), Contextualised Mobile Media for Learning

2011-10 Bart Bogaert (UvT), Cloud Content Contention

2011-11 Dhaval Vyas (UT), Designing for Awareness: An Experience-focused HCI Perspective

2011-12 Carmen Bratosin (TUE), Grid Architecture for Distributed Process Mining

2011-13 Xiaoyu Mao (UvT), Airport under Control. Multiagent Scheduling for Airport Ground Handling

2011-14 Milan Lovric (EUR), Behavioral Finance and Agent-Based Artificial Markets

2011-15 Marijn Koolen (UvA), The Meaning of Structure: the Value of Link Evidence for Information Retrieval

2011-16 Maarten Schadd (UM), Selective Search in Games of Different Complexity

2011-17 Jiyin He (UVA), Exploring Topic Structure: Coherence, Diversity and Relatedness

2011-18 Mark Ponsen (UM), Strategic Decision-Making in complex games

2011-19 Ellen Rusman (OU), The Mind's Eye on Personal Profiles

2011-20 Qing Gu (VU), Guiding service-oriented software engineering - A view-based approach

2011-21 Linda Terlouw (TUD), Modularization and Specification of Service-Oriented Systems

2011-22 Junte Zhang (UVA), System Evaluation of Archival Description and Access

2011-23 Wouter Weerkamp (UVA), Finding People and their Utterances in Social Media

2011-24 Herwin van Welbergen (UT), Behavior Generation for Interpersonal Coordination with Virtual Humans On Specifying, Scheduling and Realizing Multimodal Virtual Human Behavior

2011-25 Syed Waqar ul Qounain Jaffry (VU), Analysis and Validation of Models for Trust Dynamics

2011-26 Matthijs Aart Pontier (VU), Virtual Agents for Human Communication - Emotion Regulation and Involvement-Distance Trade-Offs in Embodied Conversational Agents and Robots

2011-27 Aniel Bhulai (VU), Dynamic website optimization through autonomous management of design patterns

2011-28 Rianne Kaptein (UVA), Effective Focused Retrieval by Exploiting Query Context and Document Structure

2011-29 Faisal Kamiran (TUE), Discrimination-aware Classification

2011-30 Egon van den Broek (UT), Affective Signal Processing (ASP): Unraveling the mystery of emotions

2011-31 Ludo Waltman (EUR), Computational and Game-Theoretic Approaches for Modeling Bounded Rationality

2011-32 Nees-Jan van Eck (EUR), Methodological Advances in Bibliometric Mapping of Science

2011-33 Tom van der Weide (UU), Arguing to Motivate Decisions

2011-34 Paolo Turrini (UU), Strategic Reasoning in Interdependence: Logical and Game-theoretical Investigations

2011-35 Maaike Harbers (UU), Explaining Agent Behavior in Virtual Training

2011-36 Erik van der Spek (UU), Experiments in serious game design: a cognitive approach

2011-37 Adriana Burlutiu (RUN), Machine Learning for Pairwise Data, Applications for Preference Learning and Supervised Network Inference

2011-38 Nyree Lemmens (UM), Bee-inspired Distributed Optimization

2011-39 Joost Westra (UU), Organizing Adaptation using Agents in Serious Games

2011-40 Viktor Clerc (VU), Architectural Knowledge Management in Global Software Development

2011-41 Luan Ibraimi (UT), Cryptographically Enforced Distributed Data Access Control

2011-42 Michal Sindlar (UU), Explaining Behavior through Mental State Attribution

2011-43 Henk van der Schuur (UU), Process Improvement through Software Operation Knowledge

2011-44 Boris Reuderink (UT), Robust Brain-Computer Interfaces

2011-45 Herman Stehouwer (UvT), Statistical Language Models for Alternative Sequence Selection
2011-46 Beibei Hu (TUD), Towards Contextualized Information Delivery: A Rule-based Architecture for the Domain of Mobile Police Work
2011-47 Azizi Bin Ab Aziz (VU), Exploring Computational Models for Intelligent Support of Persons with Depression
2011-48 Mark Ter Maat (UT), Response Selection and Turn-taking for a Sensitive Artificial Listening Agent
2011-49 Andreea Niculescu (UT), Conversational interfaces for task-oriented spoken dialogues: design aspects influencing interaction quality
2012-01 Terry Kakeeto (UvT), Relationship Marketing for SMEs in Uganda
2012-02 Muhammad Umair (VU), Adaptivity, emotion, and Rationality in Human and Ambient Agent Models
2012-03 Adam Vanya (VU), Supporting Architecture Evolution by Mining Software Repositories
2012-04 Jurriaan Souer (UU), Development of Content Management System-based Web Applications
2012-05 Marijn Plomp (UU), Maturing Interorganisational Information Systems
2012-06 Wolfgang Reinhardt (OU), Awareness Support for Knowledge Workers in Research Networks
2012-07 Rianne van Lambalgen (VU), When the Going Gets Tough: Exploring Agent-based Models of Human Performance under Demanding Conditions
2012-08 Gerben de Vries (UVA), Kernel Methods for Vessel Trajectories
2012-09 Ricardo Neisse (UT), Trust and Privacy Management Support for Context-Aware Service Platforms
2012-10 David Smits (TUE), Towards a Generic Distributed Adaptive Hypermedia Environment
2012-11 J.C.B. Rantham Prabhakara (TUE), Process Mining in the Large: Preprocessing, Discovery, and Diagnostics
2012-12 Kees van der Sluijs (TUE), Model Driven Design and Data Integration in Semantic Web Information Systems
2012-13 Suleman Shahid (UvT), Fun and Face: Exploring non-verbal expressions of emotion during playful interactions
2012-14 Evgeny Knutov (TUE), Generic Adaptation Framework for Unifying Adaptive Web-based Systems
2012-15 Natalie van der Wal (VU), Social Agents. Agent-Based Modelling of Integrated Internal and Social Dynamics of Cognitive and Affective Processes.
2012-16 Fiemke Both (VU), Helping people by understanding them - Ambient Agents supporting task execution and depression treatment
2012-17 Amal Elgammal (UvT), Towards a Comprehensive Framework for Business Process Compliance
2012-18 Eltjo Poort (VU), Improving Solution Architecting Practices
2012-19 Helen Schonenberg (TUE), What's Next? Operational Support for Business Process Execution
2012-20 Ali Bahramisharif (RUN), Covert Visual Spatial Attention, a Robust Paradigm for Brain-Computer Interfacing
2012-21 Roberto Cornacchia (TUD), Querying Sparse Matrices for Information Retrieval
2012-22 Thijs Vis (UvT), Intelligence, politie en veiligheidsdienst: verenigbare grootheden?
2012-23 Christian Muehl (UT), Toward Affective Brain-Computer Interfaces: Exploring the Neurophysiology of Affect during Human Media Interaction
2012-24 Laurens van der Werff (UT), Evaluation of Noisy Transcripts for Spoken Document Retrieval
2012-25 Silja Eckartz (UT), Managing the Business Case Development in Inter-Organizational IT Projects: A Methodology and its Application
2012-26 Emile de Maat (UVA), Making Sense of Legal Text
2012-27 Hayrettin Gurkok (UT), Mind the Sheep! User Experience Evaluation & Brain-Computer Interface Games
2012-28 Nancy Pascall (UvT), Engendering Technology Empowering Women
2012-29 Almer Tigelaar (UT), Peer-to-Peer Information Retrieval
2012-30 Alina Pommeranz (TUD), Designing Human-Centered Systems for Reflective Decision Making
2012-31 Emily Bagarukayo (RUN), A Learning by Construction Approach for Higher Order Cognitive Skills Improvement, Building Capacity and Infrastructure
2012-32 Wietske Visser (TUD), Qualitative multi-criteria preference representation and reasoning
2012-33 Rory Sie (OUN), Coalitions in Cooperation Networks (COCOON)
2012-34 Pavol Jancura (RUN), Evolutionary analysis in PPI networks and applications
2012-35 Evert Haasdijk (VU), Never Too Old To Learn – On-line Evolution of Controllers in Swarm- and Modular Robotics
2012-36 Denis Ssebugwawo (RUN), Analysis and Evaluation of Collaborative Modeling Processes
2012-37 Agnes Nakakawa (RUN), A Collaboration Process for Enterprise Architecture Creation

2012-38 Selmar Smit (VU), Parameter Tuning and Scientific Testing in Evolutionary Algorithms
2012-39 Hassan Fatemi (UT), Risk-aware design of value and coordination networks
2012-40 Agus Gunawan (UvT), Information Access for SMEs in Indonesia
2012-41 Sebastian Kelle (OU), Game Design Patterns for Learning
2012-42 Dominique Verpoorten (OU), Reflection Amplifiers in self-regulated Learning
2012-43 Withdrawn
2012-44 Anna Tordai (VU), On Combining Alignment Techniques
2012-45 Benedikt Kratz (UvT), A Model and Language for Business-aware Transactions
2012-46 Simon Carter (UVA), Exploration and Exploitation of Multilingual Data for Statistical Machine Translation
2012-47 Manos Tsagkias (UVA), Mining Social Media: Tracking Content and Predicting Behavior
2012-48 Jorn Bakker (TUE), Handling Abrupt Changes in Evolving Time-series Data
2012-49 Michael Kaisers (UM), Learning against Learning - Evolutionary dynamics of reinforcement learning algorithms in strategic interactions
2012-50 Steven van Kervel (TUD), Ontologogy driven Enterprise Information Systems Engineering
2012-51 Jeroen de Jong (TUD), Heuristics in Dynamic Sceduling; a practical framework with a case study in elevator dispatching
2013-01 Viorel Milea (EUR), News Analytics for Financial Decision Support
2013-02 Erietta Liarou (CWI), MonetDB/DataCell: Leveraging the Column-store Database Technology for Efficient and Scalable Stream Processing
2013-03 Szymon Klarman (VU), Reasoning with Contexts in Description Logics
2013-04 Chetan Yadati (TUD), Coordinating autonomous planning and scheduling
2013-05 Dulce Pumareja (UT), Groupware Requirements Evolutions Patterns
2013-06 Romulo Goncalves (CWI), The Data Cyclotron: Juggling Data and Queries for a Data Warehouse Audience
2013-07 Giel van Lankveld (UvT), Quantifying Individual Player Differences
2013-08 Robbert-Jan Merk (VU), Making enemies: cognitive modeling for opponent agents in fighter pilot simulators
2013-09 Fabio Gori (RUN), Metagenomic Data Analysis: Computational Methods and Applications
2013-10 Jeewanie Jayasinghe Arachchige (UvT), A Unified Modeling Framework for Service Design.
2013-11 Evangelos Pournaras (TUD), Multi-level Reconfigurable Self-organization in Overlay Services
2013-12 Marian Razavian (VU), Knowledge-driven Migration to Services
2013-13 Mohammad Safiri (UT), Service Tailoring: User-centric creation of integrated IT-based homecare services to support independent living of elderly
2013-14 Jafar Tanha (UVA), Ensemble Approaches to Semi-Supervised Learning Learning
2013-15 Daniel Hennes (UM), Multiagent Learning - Dynamic Games and Applications
2013-16 Eric Kok (UU), Exploring the practical benefits of argumentation in multi-agent deliberation
2013-17 Koen Kok (VU), The PowerMatcher: Smart Coordination for the Smart Electricity Grid
2013-18 Jeroen Janssens (UvT), Outlier Selection and One-Class Classification
2013-19 Renze Steenhuizen (TUD), Coordinated Multi-Agent Planning and Scheduling
2013-20 Katja Hofmann (UvA), Fast and Reliable Online Learning to Rank for Information Retrieval
2013-21 Sander Wubben (UvT), Text-to-text generation by monolingual machine translation
2013-22 Tom Claassen (RUN), Causal Discovery and Logic
2013-23 Patricio de Alencar Silva (UvT), Value Activity Monitoring
2013-24 Haitham Bou Ammar (UM), Automated Transfer in Reinforcement Learning
2013-25 Agnieszka Anna Latoszek-Berendsen (UM), Intention-based Decision Support. A new way of representing and implementing clinical guidelines in a Decision Support System
2013-26 Alireza Zarghami (UT), Architectural Support for Dynamic Homecare Service Provisioning
2013-27 Mohammad Huq (UT), Inference-based Framework Managing Data Provenance
2013-28 Frans van der Sluis (UT), When Complexity becomes Interesting: An Inquiry into the Information eXperience
2013-29 Iwan de Kok (UT), Listening Heads
2013-30 Joyce Nakatumba (TUE), Resource-Aware Business Process Management: Analysis and Support
2013-31 Dinh Khoa Nguyen (UvT), Blueprint Model and Language for Engineering Cloud Applications
2013-32 Kamakshi Rajagopal (OUN), Networking For Learning; The role of Networking in a Lifelong Learner's Professional Development
2013-33 Qi Gao (TUD), User Modeling and Personalization in the Microblogging Sphere
2013-34 Kien Tjin-Kam-Jet (UT), Distributed Deep Web Search
2013-35 Abdallah El Ali (UvA), Minimal Mobile Human Computer Interaction
2013-36 Than Lam Hoang (TUe), Pattern Mining in Data Streams

2013-37 Dirk Börner (OUN), Ambient Learning Displays
2013-38 Eelco den Heijer (VU), Autonomous Evolutionary Art
2013-39 Joop de Jong (TUD), A Method for Enterprise Ontology based Design of Enterprise Information Systems
2013-40 Pim Nijssen (UM), Monte-Carlo Tree Search for Multi-Player Games
2013-41 Jochem Liem (UVA), Supporting the Conceptual Modelling of Dynamic Systems: A Knowledge Engineering Perspective on Qualitative Reasoning
2013-42 Léon Planken (TUD), Algorithms for Simple Temporal Reasoning
2013-43 Marc Bron (UVA), Exploration and Contextualization through Interaction and Concepts
2014-01 Nicola Barile (UU), Studies in Learning Monotone Models from Data
2014-02 Fiona Tuliyano (RUN), Combining System Dynamics with a Domain Modeling Method
2014-03 Sergio Raul Duarte Torres (UT), Information Retrieval for Children: Search Behavior and Solutions
2014-04 Hanna Jochmann-Mannak (UT), Websites for children: search strategies and interface design - Three studies on children's search performance and evaluation
2014-05 Jurriaan van Reijsen (UU), Knowledge Perspectives on Advancing Dynamic Capability
2014-06 Damian Tamburri (VU), Supporting Networked Software Development
2014-07 Arya Adriansyah (TUE), Aligning Observed and Modeled Behavior
2014-08 Samur Araujo (TUD), Data Integration over Distributed and Heterogeneous Data Endpoints
2014-09 Philip Jackson (UvT), Toward Human-Level Artificial Intelligence: Representation and Computation of Meaning in Natural Language
2014-10 Ivan Salvador Razo Zapata (VU), Service Value Networks
2014-11 Janneke van der Zwaan (TUD), An Empathic Virtual Buddy for Social Support
2014-12 Willem van Willigen (VU), Look Ma, No Hands: Aspects of Autonomous Vehicle Control
2014-13 Arlette van Wissen (VU), Agent-Based Support for Behavior Change: Models and Applications in Health and Safety Domains
2014-14 Yangyang Shi (TUD), Language Models With Meta-information
2014-15 Natalya Mogles (VU), Agent-Based Analysis and Support of Human Functioning in Complex Socio-Technical Systems: Applications in Safety and Healthcare
2014-16 Krystyna Milian (VU), Supporting trial recruitment and design by automatically interpreting eligibility criteria
2014-17 Kathrin Dentler (VU), Computing healthcare quality indicators automatically: Secondary Use of Patient Data and Semantic Interoperability
2014-18 Mattijs Ghijsen (UVA), Methods and Models for the Design and Study of Dynamic Agent Organizations
2014-19 Vinicius Ramos (TUE), Adaptive Hypermedia Courses: Qualitative and Quantitative Evaluation and Tool Support
2014-20 Mena Habib (UT), Named Entity Extraction and Disambiguation for Informal Text: The Missing Link
2014-21 Kassidy Clark (TUD), Negotiation and Monitoring in Open Environments
2014-22 Marieke Peeters (UU), Personalized Educational Games - Developing agent-supported scenario-based training
2014-23 Eleftherios Sidirourgos (UvA/CWI), Space Efficient Indexes for the Big Data Era
2014-24 Davide Ceolin (VU), Trusting Semi-structured Web Data
2014-25 Martijn Lappenschaar (RUN), New network models for the analysis of disease interaction
2014-26 Tim Baarslag (TUD), What to Bid and When to Stop
2014-27 Rui Jorge Almeida (EUR), Conditional Density Models Integrating Fuzzy and Probabilistic Representations of Uncertainty
2014-28 Anna Chmielowiec (VU), Decentralized k-Clique Matching
2014-29 Jaap Kabbedijk (UU), Variability in Multi-Tenant Enterprise Software
2014-30 Peter de Cock (UvT), Anticipating Criminal Behaviour
2014-31 Leo van Moergestel (UU), Agent Technology in Agile Multiparallel Manufacturing and Product Support
2014-32 Naser Ayat (UvA), On Entity Resolution in Probabilistic Data
2014-33 Tesfa Tegegne (RUN), Service Discovery in eHealth
2014-34 Christina Manteli (VU), The Effect of Governance in Global Software Development: Analyzing Transactive Memory Systems.
2014-35 Joost van Ooijen (UU), Cognitive Agents in Virtual Worlds: A Middleware Design Approach
2014-36 Joos Buijs (TUE), Flexible Evolutionary Algorithms for Mining Structured Process Models
2014-37 Maral Dadvar (UT), Experts and Machines United Against Cyberbullying
2014-38 Danny Plass-Oude Bos (UT), Making brain-computer interfaces better: improving usability through post-processing.

2014-39 Jasmina Maric (UvT), Web Communities, Immigration, and Social Capital
2014-40 Walter Omona (RUN), A Framework for Knowledge Management Using ICT in Higher Education
2014-41 Frederic Hogenboom (EUR), Automated Detection of Financial Events in News Text
2014-42 Carsten Eijckhof (CWI/TUD), Contextual Multidimensional Relevance Models
2014-43 Kevin Vlaanderen (UU), Supporting Process Improvement using Method Increments
2014-44 Paulien Meesters (UvT), Intelligent Blauw. Met als ondertitel: Intelligence-gestuurde politiezorg in gebiedsgebonden eenheden.
2014-45 Birgit Schmitz (OUN), Mobile Games for Learning: A Pattern-Based Approach
2014-46 Ke Tao (TUD), Social Web Data Analytics: Relevance, Redundancy, Diversity
2014-47 Shangsong Liang (UVA), Fusion and Diversification in Information Retrieval
2015-01 Niels Netten (UvA), Machine Learning for Relevance of Information in Crisis Response
2015-02 Faiza Bukhsh (UvT), Smart auditing: Innovative Compliance Checking in Customs Controls
2015-03 Twan van Laarhoven (RUN), Machine learning for network data
2015-04 Howard Spoelstra (OUN), Collaborations in Open Learning Environments
2015-05 Christoph Bösch (UT), Cryptographically Enforced Search Pattern Hiding
2015-06 Farideh Heidari (TUD), Business Process Quality Computation - Computing Non-Functional Requirements to Improve Business Processes
2015-07 Maria-Hendrike Peetz (UvA), Time-Aware Online Reputation Analysis
2015-08 Jie Jiang (TUD), Organizational Compliance: An agent-based model for designing and evaluating organizational interactions
2015-09 Randy Klaassen (UT), HCI Perspectives on Behavior Change Support Systems
2015-10 Henry Hermans (OUN), OpenU: design of an integrated system to support lifelong learning
2015-11 Yongming Luo (TUE), Designing algorithms for big graph datasets: A study of computing bisimulation and joins
2015-12 Julie M. Birkholz (VU), Modi Operandi of Social Network Dynamics: The Effect of Context on Scientific Collaboration Networks
2015-13 Giuseppe Procaccianti (VU), Energy-Efficient Software
2015-14 Bart van Straalen (UT), A cognitive approach to modeling bad news conversations
2015-15 Klaas Andries de Graaf (VU), Ontology-based Software Architecture Documentation
2015-16 Changyun Wei (UT), Cognitive Coordination for Cooperative Multi-Robot Teamwork
2015-17 André van Cleeff (UT), Physical and Digital Security Mechanisms: Properties, Combinations and Trade-offs
2015-18 Holger Pirk (CWI), Waste Not, Want Not! - Managing Relational Data in Asymmetric Memories
2015-19 Bernardo Tabuenca (OUN), Ubiquitous Technology for Lifelong Learners
2015-20 Lois Vanhée (UU), Using Culture and Values to Support Flexible Coordination
2015-21 Sibren Fetter (OUN), Using Peer-Support to Expand and Stabilize Online Learning
2015-22 Zhemin Zhu (UT), Co-occurrence Rate Networks
2015-23 Luit Gazendam (VU), Cataloguer Support in Cultural Heritage
2015-24 Richard Berendsen (UVA), Finding People, Papers, and Posts: Vertical Search Algorithms and Evaluation
2015-25 Steven Woudenberg (UU), Bayesian Tools for Early Disease Detection
2015-26 Alexander Hogenboom (EUR), Sentiment Analysis of Text Guided by Semantics and Structure
2015-27 Sándor Héman (CWI), Updating compressed colomn stores
2015-28 Janet Bagorogoza (TiU), Knowledge Management and High Performance; The Uganda Financial Institutions Model for HPO
2015-29 Hendrik Baier (UM), Monte-Carlo Tree Search Enhancements for One-Player and Two-Player Domains
2015-30 Kiavash Bahreini (OU), Real-time Multimodal Emotion Recognition in E-Learning
2015-31 Yakup Koç (TUD), On the robustness of Power Grids
2015-32 Jerome Gard (UL), Corporate Venture Management in SMEs
2015-33 Frederik Schadd (TUD), Ontology Mapping with Auxiliary Resources
2015-34 Victor de Graaf (UT), Gesocial Recommender Systems
2015-35 Jungxao Xu (TUD), Affective Body Language of Humanoid Robots: Perception and Effects in Human Robot Interaction
2016-01 Syed Saiden Abbas (RUN), Recognition of Shapes by Humans and Machines
2016-02 Michiel Christiaan Meulendijk (UU), Optimizing medication reviews through decision support: prescribing a better pill to swallow
2016-03 Maya Sappelli (RUN), Knowledge Work in Context: User Centered Knowledge Worker Support
2016-04 Laurens Rietveld (VU), Publishing and Consuming Linked Data
2016-05 Evgeny Sherkhonov (UVA), Expanded Acyclic Queries: Containment and an Application in Explaining Missing Answers

2016-06 Michel Wilson (TUD), Robust scheduling in an uncertain environment
2016-07 Jeroen de Man (VU), Measuring and modeling negative emotions for virtual training
2016-08 Matje van de Camp (TiU), A Link to the Past: Constructing Historical Social Networks from Unstructured Data
2016-09 Archana Nottamkandath (VU), Trusting Crowdsourced Information on Cultural Artefacts
2016-10 George Karafotias (VUA), Parameter Control for Evolutionary Algorithms
2016-11 Anne Schuth (UVA), Search Engines that Learn from Their Users
2016-12 Max Knobbout (UU), Logics for Modelling and Verifying Normative Multi-Agent Systems
2016-13 Nana Baah Gyan (VU), The Web, Speech Technologies and Rural Development in West Africa - An ICT4D Approach
2016-14 Ravi Khadka (UU), Revisiting Legacy Software System Modernization
2016-15 Steffen Michels (RUN), Hybrid Probabilistic Logics - Theoretical Aspects, Algorithms and Experiments
2016-16 Guangliang Li (UVA), Socially Intelligent Autonomous Agents that Learn from Human Reward
2016-17 Berend Weel (VU), Towards Embodied Evolution of Robot Organisms
2016-18 Albert Meroño Peñuela (VU), Refining Statistical Data on the Web
2016-19 Julia Efremova (Tu/e), Mining Social Structures from Genealogical Data
2016-20 Daan Odijk (UVA), Context & Semantics in News & Web Search
2016-21 Alejandro Moreno Célleri (UT), From Traditional to Interactive Playspaces: Automatic Analysis of Player Behavior in the Interactive Tag Playground
2016-22 Grace Lewis (VU), Software Architecture Strategies for Cyber-Foraging Systems
2016-23 Fei Cai (UVA), Query Auto Completion in Information Retrieval
2016-24 Brend Wanders (UT), Repurposing and Probabilistic Integration of Data; An Iterative and data model independent approach
2016-25 Julia Kiseleva (TU/e), Using Contextual Information to Understand Searching and Browsing Behavior
2016-26 Dilhan Thilakarathne (VU), In or Out of Control: Exploring Computational Models to Study the Role of Human Awareness and Control in Behavioural Choices, with Applications in Aviation and Energy Management Domains
2016-27 Wen Li (TUD), Understanding Geo-spatial Information on Social Media
2016-28 Mingxin Zhang (TUD), Large-scale Agent-based Social Simulation - A study on epidemic prediction and control
2016-29 Nicolas Höning (TUD), Peak reduction in decentralised electricity systems - Markets and prices for flexible planning
2016-30 Ruud Mattheij (UvT), The Eyes Have It
2016-31 Mohammad Khelghati (UT), Deep web content monitoring
2016-32 Eelco Vriezekolk (UT), Assessing Telecommunication Service Availability Risks for Crisis Organisations
2016-33 Peter Bloem (UVA), Single Sample Statistics, exercises in learning from just one example
2016-34 Dennis Schunselaar (TUE), Configurable Process Trees: Elicitation, Analysis, and Enactment
2016-35 Zhaochun Ren (UVA), Monitoring Social Media: Summarization, Classification and Recommendation
2016-36 Daphne Karreman (UT), Beyond R2D2: The design of nonverbal interaction behavior optimized for robot-specific morphologies
2016-37 Giovanni Sileno (UvA), Aligning Law and Action - a conceptual and computational inquiry
2016-38 Andrea Minuto (UT), Materials that Matter - Smart Materials meet Art & Interaction Design
2016-39 Merijn Bruijnes (UT), Believable Suspect Agents; Response and Interpersonal Style Selection for an Artificial Suspect
2016-40 Christian Detweiler (TUD), Accounting for Values in Design
2016-41 Thomas King (TUD), Governing Governance: A Formal Framework for Analysing Institutional Design and Enactment Governance
2016-42 Spyros Martzoukos (UVA), Combinatorial and Compositional Aspects of Bilingual Aligned Corpora
2016-43 Saskia Koldijk (RUN), Context-Aware Support for Stress Self-Management: From Theory to Practice
2016-44 Thibault Sellam (UVA), Automatic Assistants for Database Exploration
2016-45 Bram van de Laar (UT), Experiencing Brain-Computer Interface Control
2016-46 Jorge Gallego Perez (UT), Robots to Make you Happy
2016-47 Christina Weber (UL), Real-time foresight - Preparedness for dynamic innovation networks
2016-48 Tanja Buttler (TUD), Collecting Lessons Learned
2016-49 Gleb Polevoy (TUD), Participation and Interaction in Projects. A Game-Theoretic Analysis
2016-50 Yan Wang (UVT), The Bridge of Dreams: Towards a Method for Operational Performance Alignment in IT-enabled Service Supply Chains

2017-01 Jan-Jaap Oerlemans (UL), Investigating Cybercrime

2017-02 Sjoerd Timmer (UU), Designing and Understanding Forensic Bayesian Networks using Argumentation

2017-03 Daniël Harold Telgen (UU), Grid Manufacturing; A Cyber-Physical Approach with Autonomous Products and Reconfigurable Manufacturing Machines

2017-04 Mrunal Gawade (CWI), Multi-core Parallelism in a Column-store

2017-05 Mahdieh Shadi (UVA), Collaboration Behavior

2017-06 Damir Vandic (EUR), Intelligent Information Systems for Web Product Search

2017-07 Roel Bertens (UU), Insight in Information: from Abstract to Anomaly

2017-08 Rob Konijn (VU) , Detecting Interesting Differences:Data Mining in Health Insurance Data using Outlier Detection and Subgroup Discovery

2017-09 Dong Nguyen (UT), Text as Social and Cultural Data: A Computational Perspective on Variation in Text

2017-10 Robby van Delden (UT), (Steering) Interactive Play Behavior

2017-11 Florian Kunneman (RUN), Modelling patterns of time and emotion in Twitter #anticipointment

2017-12 Sander Leemans (TUE), Robust Process Mining with Guarantees

2017-13 Gijs Huisman (UT), Social Touch Technology - Extending the reach of social touch through haptic technology

2017-14 Shoshannah Tekofsky (UvT), You Are Who You Play You Are: Modelling Player Traits from Video Game Behavior

2017-15 Peter Berck (RUN), Memory-Based Text Correction

2017-16 Aleksandr Chuklin (UVA), Understanding and Modeling Users of Modern Search Engines

2017-17 Daniel Dimov (UL), Crowdsourced Online Dispute Resolution

2017-18 Ridho Reinanda (UVA), Entity Associations for Search

2017-19 Jeroen Vuurens (UT), Proximity of Terms, Texts and Semantic Vectors in Information Retrieval

2017-20 Mohammadbashir Sedighi (TUD), Fostering Engagement in Knowledge Sharing: The Role of Perceived Benefits, Costs and Visibility

2017-21 Jeroen Linssen (UT), Meta Matters in Interactive Storytelling and Serious Gaming (A Play on Worlds)

2017-22 Sara Magliacane (VU), Logics for causal inference under uncertainty

2017-23 David Graus (UVA), Entities of Interest — Discovery in Digital Traces

2017-24 Chang Wang (TUD), Use of Affordances for Efficient Robot Learning

2017-25 Veruska Zamborlini (VU), Knowledge Representation for Clinical Guidelines, with applications to Multimorbidity Analysis and Literature Search

2017-26 Merel Jung (UT), Socially intelligent robots that understand and respond to human touch

2017-27 Michiel Joosse (UT), Investigating Positioning and Gaze Behaviors of Social Robots: People's Preferences, Perceptions and Behaviors

2017-28 John Klein (VU), Architecture Practices for Complex Contexts

2017-29 Adel Alhuraibi (UvT), From IT-BusinessStrategic Alignment to Performance: A Moderated Mediation Model of Social Innovation, and Enterprise Governance of IT"

2017-30 Wilma Latuny (UvT), The Power of Facial Expressions

2017-31 Ben Ruijl (UL), Advances in computational methods for QFT calculations

2017-32 Thaer Samar (RUN), Access to and Retrievability of Content in Web Archives

2017-33 Brigit van Loggem (OU), Towards a Design Rationale for Software Documentation: A Model of Computer-Mediated Activity

2017-34 Maren Scheffel (OU), The Evaluation Framework for Learning Analytics

2017-35 Martine de Vos (VU), Interpreting natural science spreadsheets

2017-36 Yuanhao Guo (UL), Shape Analysis for Phenotype Characterisation from High-throughput Imaging

2017-37 Alejandro Montes Garcia (TUE), WiBAF: A Within Browser Adaptation Framework that Enables Control over Privacy

2017-38 Alex Kayal (TUD), Normative Social Applications

2017-39 Sara Ahmadi (RUN), Exploiting properties of the human auditory system and compressive sensing methods to increase noise robustness in ASR

2017-40 Altaf Hussain Abro (VUA), Steer your Mind: Computational Exploration of Human Control in Relation to Emotions, Desires and Social Support For applications in human-aware support systems

2017-41 Adnan Manzoor (VUA), Minding a Healthy Lifestyle: An Exploration of Mental Processes and a Smart Environment to Provide Support for a Healthy Lifestyle

2017-42 Elena Sokolova (RUN), Causal discovery from mixed and missing data with applications on ADHD datasets

2017-43 Maaike de Boer (RUN), Semantic Mapping in Video Retrieval

2017-44 Garm Lucassen (UU), Understanding User Stories - Computational Linguistics in Agile Requirements Engineering

2017-45	Bas Testerink (UU), Decentralized Runtime Norm Enforcement
2017-46	Jan Schneider (OU), Sensor-based Learning Support
2017-47	Jie Yang (TUD), Crowd Knowledge Creation Acceleration
2017-48	Angel Suarez (OU), Collaborative inquiry-based learning
2018-01	Han van der Aa (VUA), Comparing and Aligning Process Representations
2018-02	Felix Mannhardt (TUE), Multi-perspective Process Mining
2018-03	Steven Bosems (UT), Causal Models For Well-Being: Knowledge Modeling, Model-Driven Development of Context-Aware Applications, and Behavior Prediction
2018-04	Jordan Janeiro (TUD), Flexible Coordination Support for Diagnosis Teams in Data-Centric Engineering Tasks
2018-05	Hugo Huurdeman (UVA), Supporting the Complex Dynamics of the Information Seeking Process
2018-06	Dan Ionita (UT), Model-Driven Information Security Risk Assessment of Socio-Technical Systems
2018-07	Jieting Luo (UU), A formal account of opportunism in multi-agent systems
2018-08	Rick Smetsers (RUN), Advances in Model Learning for Software Systems
2018-09	Xu Xie (TUD), Data Assimilation in Discrete Event Simulations
2018-10	Julienka Mollee (VUA), Moving forward: supporting physical activity behavior change through intelligent technology
2018-11	Mahdi Sargolzaei (UVA), Enabling Framework for Service-oriented Collaborative Networks
2018-12	Xixi Lu (TUE), Using behavioral context in process mining
2018-13	Seyed Amin Tabatabaei (VUA), Computing a Sustainable Future
2018-14	Bart Joosten (UVT), Detecting Social Signals with Spatiotemporal Gabor Filters
2018-15	Naser Davarzani (UM), Biomarker discovery in heart failure
2018-16	Jaebok Kim (UT), Automatic recognition of engagement and emotion in a group of children
2018-17	Jianpeng Zhang (TUE), On Graph Sample Clustering
2018-18	Henriette Nakad (UL), De Notaris en Private Rechtspraak
2018-19	Minh Duc Pham (VUA), Emergent relational schemas for RDF
2018-20	Manxia Liu (RUN), Time and Bayesian Networks
2018-21	Aad Slootmaker (OUN), EMERGO: a generic platform for authoring and playing scenario-based serious games
2018-22	Eric Fernandes de Mello Araújo (VUA), Contagious: Modeling the Spread of Behaviours, Perceptions and Emotions in Social Networks