Who do we choose to spend our leisure time with? Insights from discrete choice and machine learning models

Shuwei Lin

Institute for Transport Studies (ITS)

University of Leeds

Submitted in accordance with the requirements for the degree of
Master of Philosophy

October 2024

Acknowledgements

I would like to extend my heartfelt gratitude to my supervisors, Chiara and Stephane, for welcoming me into this research programme and providing me with the opportunity to explore my passion for research. I am also deeply grateful to Aravinda for joining the supervisory team and offering his valuable expertise.

Originally, this research was intended as a PhD programme with publication expectations. However, I had to switch to an MPhil as it was not working out for me. I am particularly thankful to my supervisors for their unwavering support during this difficult transition and for their willingness to continue guiding me to complete my degree. Special thanks are also due to Zia and Rachel for their assistance during this period. Zia was always available to offer coaching and help me evaluate my circumstances, while Rachel reminded me to take care of my mental health, providing timely and comforting support.

The COVID-19 pandemic was one of many challenges that made this research journey difficult. I often wonder how things might have been different had I deferred my start date to post-pandemic times. Nonetheless, I am immensely grateful to Chiara for arranging countless online meetings to support my progress.

One of my dreams has been to gain experience as a demonstrator and teaching assistant at a university. I am thankful to Chiara for giving me the opportunity to fulfil this dream through her Transport Data Collection and Analysis module for master's students.

I would like to thank Charisma Choudhury and Ed Manley for serving as my examiners during my transfer process and for their invaluable feedback, which greatly improved my transfer report by enhancing its positioning in the literature and clarifying my ideas. Although some parts of the transfer report were no longer relevant to this thesis, the lessons learned from that process were invaluable.

This research would not have been possible without the data provided by Matthias Kowald. I am grateful not only for the access to the data but also for his patience in answering my questions. I regret that this work did not result in any publications during my MPhil, but I hope to repay his kindness in the future.

I am thankful to the committee of the International Choice Modelling Conference 2022 in Reykjavik, Iceland, for the opportunity to attend and present at my first academic conference. It was exciting to meet scholars I had only encountered through their work, particularly Michael Maness, whose review paper on social influence was one of the starting points in my literature review. I appreciate Michael's invitation for a potential collaboration, although it was unfortunate that it did not materialise.

I am grateful to my colleagues, friends, and fellow research students at this institute. Special thanks to Panos Tsoleridis, David Palma and Thomas Hancock for their assistance with coding in Apollo and discussions, Maximiliano Lizana Maldonado and Zannat for their companionship and research tips, Zihao An for his advice on academic writing and literature review, Alexandra-Elena Vitel for her tips on organising supervisory meetings and progress reviews, and Chen Peng and Zhuoqian Yang for their kindness and support during challenging times. Oguz Tengilimoglu and Pinar Bilgin also deserve thanks for their constant greetings and snacks.

Although Sijin Wu and I never managed to co-author a paper, our discussions on modelling the mobility changes caused by the perceived risk of the pandemic were really enjoyable and enlightening.

Sports have been a cherished part of my life, and I am glad to have found companions at ITS as well. I enjoyed playing football with Thomas Handcock, Zhichao Wang, Victor Cantillo Garcia, and Tangjian Wei; basketball with Aravinda Ramakrishnan Srinivasan and Yueyang Wang; and badminton with Yan Liu, Yaqi Yang, and Yaqi Zhang.

I am deeply thankful to my fellow believers from the Church in Leeds, particularly Nathanael Stone and Rebecca Stone, who have always been just a call or text away, offering support for all my problems, big or small.

My family has been a constant source of support. Despite the health challenges faced by my mother and grandfather over the past few years, we have stood by each other. As the first person in my family to attend higher education, they may not have always understood my struggles, but their unwavering emotional support has been invaluable.

Reflecting on my journey, I realise that I have gained knowledge, skills, and lessons that are transferable to the workplace. I believe the experiences and lessons learned at ITS will be beneficial in my life.

This research, centred on the choices we make about who we spend our time with, has been profoundly shaped by the people who have journeyed with me. I am grateful to all who have been part of this chapter of my life. Thank you!

I acknowledge the funding from the EPSRC, which made this research possible.

Finally, I acknowledge the use of ChatGPT-4, developed by OpenAI, available at https://chat.openai.com/, for assistance in refining the writing style and clarifying the content in some parts of this thesis. ChatGPT helped rephrase and clarify arguments in these parts, ensuring readability without contributing original content or ideas.

Abstract

This thesis investigates the determinants of social contact selection in leisure activities, integrating traditional discrete choice models (DCM) with machine learning (ML) techniques to enhance model specification and explanatory power. The research utilises a dataset collected through snowball sampling in Switzerland, capturing a wide range of respondents' characteristics and the characteristics of their social network members. Employing multinomial logit models, the study reveals how dyadic variables such as age homophily, gender homophily, relationship duration and so on influence leisure activity choices with various social contacts. Furthermore, the incorporation of machine learning techniques, particularly Shapley Additive explanations (SHAP), enriches the model. SHAP highlights predictors that might otherwise be overlooked. It also provides insight into the direction and impact of these predictors, facilitating their interpretation before running a choice model. The findings extend the current understanding of social interaction patterns, advocating for consideration of social networks in data collection and modelling of who people interact with. This thesis also uses machine learning to assist choice modelling, offering an additional tool for analysing social contact preferences in leisure contexts, which implies an additional set (potentially large) of explanatory variables where machine learning models could be useful.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

People's daily activities often involve interactions with others, necessitating several key decisions such as activity type (Calastri, Hess, Daly, & Carrasco, 2017), location of social interaction (van den Berg et al., 2010), travel distance for social interaction (Moore et al., 2013), activity duration (van den Berg et al., 2012a), travel mode for social activity (Sharmeen & Timmermans, 2014), and who to involve in these activities (Habib et al., 2008). The social dimension thus emerges as a fundamental factor influencing choice behaviour.

Social networks are social structures comprising a set of actors (nodes) and the relationships (ties) that connect them (Tindall & Wellman, 2001). These actors can be individuals, groups, organisations, or even nation-states, and the relationships encompass various types of interactions and exchanges, such as flows of resources, information, or support. Social network analysis focuses on understanding how these structures facilitate and constrain opportunities, behaviours, and cognitions. By examining the patterns of relationships, social network analysts aim to explain the effects of these structures on individual and collective outcomes.

These networks significantly influence decision-making processes, impacting choices not only based on individual characteristics and the attributes of alternatives but also through the influence of social network members. Understanding the role of social networks is crucial for comprehensively modelling human behaviour, especially in the context of activity-travel decisions.

In transportation research, the social dimension of decision-making has garnered attention as a core determinant of various choices (Kim et al., 2018; Maness et al., 2015). Various statistical methods have been employed to explore the re-

lationship between social networks and different facets of social activity-travel decisions. These methods generally follow an ego-centric approach (Kim et al., 2018), focusing on the individual's (ego's) network that consists of the ego and their social contacts (alters). This approach considers three types of characteristics: ego-level (decision-maker's characteristics), ego-alter-level (characteristics of alters and their relationships with the ego), and ego-network-level (characteristics of the network itself, such as size). Discrete choice modelling is a common method used in this literature to consider the social dimension or social networks in choice behaviour. Discrete choice models are statistical models used to explain and forecast individuals' choice behaviour. They can be derived from utility theory, where each choice among a set of alternatives is associated with a utility, and a rational decision-maker would choose the alternative that maximises their utility (Train, 2009). The functional form of the utility for each alternative comprises a deterministic part and a random part. The deterministic part includes observable variables that determine the choices, while the random part represents the uncertainty of the modeller (i.e., other factors can explain the choice behaviour of interest but are not observed or available in the data).

One crucial aspect of activity-travel decisions is the "with whom" choice – deciding who to involve in an activity. Research has demonstrated that social context plays a significant role in scheduling activities. For instance, Habib et al. (2008) found that the decision of "with whom" to participate influences the start time and duration of activities. Social activities involving multiple household members tend to start later due to coordination constraints, while those with friends or non-household family members have longer durations compared to those involving only household members. Similarly, Habib and Carrasco (2011) confirmed the importance of the "with whom" variable in understanding the timing and length of social activities, highlighting the interdependencies between social interactions and temporal activity patterns.

Furthermore, social networks significantly affect travel decisions, such as mode choice, travel time, and activity selection. For example, Axhausen (2005) noted that social interactions are a primary driver for travel, particularly leisure travel aimed at meeting friends, relatives, and acquaintances. The spatial distribution of social contacts influences travel patterns, with longer distances leading to increased travel frequency and different mode choices. Additionally, Arentze & Timmermans (2008) developed models predicting activity participation and travel behaviour, considering the presence of companions, which greatly influence both the decision to participate in activities and the choice of travel modes.

Research also underscores the importance of incorporating social factors into

models of travel behaviour to capture the complex dynamics of social interactions and activity participation. For instance, Kim et al. (2018) concluded that the presence and composition of social networks significantly influence travel decisions, such as travel mode choices and activity durations. Similarly, Carrasco & Miller (2009) showed that the characteristics of "with whom" social activities are performed play a crucial role, intertwined with the individual's characteristics. Their multilevel personal networks model demonstrated that individuals are more likely to frequently engage in social activities with close friends and family members, and the spatial distribution of these social contacts affects social activity frequency.

In summary, decision-making processes in the context of activity-travel choices are complex and influenced by a range of factors that extend beyond individual characteristics to include ego-alter-level and network-level characteristics. Understanding these processes, particularly the selection of social contacts in leisure activities provides valuable insights into human behaviour.

While discrete choice techniques remain the key analytical tool for understanding travel behaviour, there is increasing interest in the use of machine learning tools. If one considers the choice variable (e.g., who to interact with) in discrete choice modelling a label associated with observations labels, the question can be thought of as a machine learning problem. Machine learning (ML) is part of artificial intelligence (AI) that employs algorithms and models to learn from data and improve task performance without using explicit instructions. The goal is to provide a computer with vast data to discern patterns, make predictions, or gain insights about the data. There has been increasing interest in synergies between machine learning and choice modelling (Hillel et al., 2021; van Cranenburgh et al., 2022)). Indeed, in the choice modelling community, the integration of machine learning (ML) and choice modelling (CM) has been explored from various perspectives, often highlighting the strengths and limitations of each approach. The following subsections review key studies in this domain, focusing on three main themes: using ML as an alternative to CM, combining the strengths of both ML and CM and employing ML to assist in CM specification.

1.1.1 Machine Learning as an Alternative to Choice Models

One approach in the existing literature is using ML as an alternative to traditional choice models for choice behaviour analysis. Scholars have found that ML models often outperform traditional choice models in terms of predictive accuracy (see

reviews by Hillel et al. (2021) and van Cranenburgh et al. (2022)). These studies emphasise the potential of ML techniques to offer valuable insights into mode choice modelling and improve model specification and estimation time. However, there is still some hesitation in fully embracing ML due to misconceptions about its applications and benefits.

van Cranenburgh & Alwosheel (2019) suggest using an Artificial Neural Network (ANN) as an alternative to the traditional Latent Class discrete choice modelling approach to investigate decision rule heterogeneity. Their ANN-based approach can recognise patterns in travellers' choices that traditional methods might overlook, handling large, complex datasets and uncovering nuanced insights into decision-making processes. Empirical validation against traditional models confirmed the ANN's effectiveness in studying decision rule heterogeneity.

Efforts have also been made to enhance the interpretability of ML models. Alwosheel et al. (2019) introduced the use of prototypical examples from computer vision to help modellers assess learned relationships in an ANN, thereby improving trust in its predictions. This methodology, applied to a Revealed Preference mode choice dataset, allows analysts to evaluate the fundamental relationships learned by the ANN and build trust in well-functioning models.

Similarly, Golshani et al. (2018) estimated the relative importance of each explanatory variable in a neural network and conducted sensitivity analyses to understand their impacts on choices. Using Garson's method to calculate the relative importance of input variables, they improved the interpretability of neural network models and enhanced the understanding of factors driving travel behaviour.

Wang, Wang, & Zhao (2020) proposed a variant of deep neural networks (DNN) that bridges the gap between predictive performance and interpretability. Their study demonstrates that DNNs can provide economic information as complete as classical discrete choice models (DCMs), including various economic indicators. This approach was validated with datasets from Singapore and London, showing that DNNs can learn utility functions and reveal behavioural patterns without prespecification by domain experts. However, challenges such as high sensitivity to hyperparameters, model non-identification, and local irregularity were noted.

Despite these advancements, ML models still face limitations in behavioural and economic analysis due to their less structured approach compared to traditional choice models.

1.1.2 Take the best of both worlds? CM to ML

Given that choice models impose a structured framework on the data, leading to high interpretability, and machine learning models achieve better predictive performance by learning the structure from data, researchers have proposed integrating both approaches to leverage their respective strengths.

One direction of this integration is using choice modelling principles to assist machine learning. For example, Wang, Mo, & Zhao (2020) suggested imposing the Independence of Irrelevant Alternatives (IIA) constraint on a deep neural network to improve its interpretability without significantly compromising predictive accuracy. They developed a novel DNN architecture called alternative-specific utility DNN (ASU-DNN), which uses behavioural knowledge to guide the design of the network. Empirical results demonstrated that ASU-DNN outperforms fully connected DNNs (F-DNN) and other classifiers, achieving higher prediction accuracy and providing more intuitive choice probability functions. Other studies have incorporated matrix factorization methods from recommender systems into random utility models (Athey et al., 2018), thereby enhancing the ability to predict consumer preferences while maintaining a solid theoretical foundation. Athey et al. (2018)) show that matrix factorization can reduce the dimensionality of consumer preference data, making it easier to model complex interactions and price sensitivities. This approach allows for personalized predictions and provides deeper economic insights into consumer behavior by identifying latent characteristics and preferences.

1.1.3 Take the best of both worlds? ML to CM

Another promising direction is using machine learning techniques to assist in the specification of choice models. Wong et al. (2018) employed machine learning to represent latent behavioural variables, serving as an alternative to the Integrated Choice and Latent Variable (ICLV) model when attitudinal indicators are absent. They proposed using restricted Boltzmann machines (RBMs) to model latent behavioural factors by analyzing the relationships between observed choices and explanatory variables. This approach addresses the limitations of ICLV models, such as the reliance on subjective attitudinal data, by inferring latent variables directly from choice data. Sfeir et al. (2021, 2022) proposed using machine learning alternatives for the class membership component of Latent Class Choice Models (LCCM), achieving better prediction accuracy without undermining economic in-

terpretability. In the 2021 study, they introduced a semi-nonparametric LCCM using mixture models, a machine learning approach, to improve the flexibility and accuracy of class membership estimations. The 2022 paper extended this approach by proposing the Gaussian Process Latent Class Choice Model (GP-LCCM), which employs Gaussian Processes to assign individuals to latent classes probabilistically. Tsoleridis et al. (2019) explored integrating clustering techniques, specifically Gaussian Mixture Models (GMMs), within the Latent Class Choice Model (LCCM) framework. Their study found that while traditional LCCMs generally outperformed machine learning methods in prediction accuracy, ML methods provided computational efficiencies, particularly valuable in handling large datasets and complex models.

Both Wong & Farooq (2021) and Sifringer et al. (2020) suggested models where systematic utility is divided into a knowledge-driven part and a data-driven part, enabling the capture of complex patterns while retaining interpretability. Wong & Farooq (2021)'s ResLogit model integrates a deep neural network with a multinomial logit model, capturing non-linear cross-effects and unobserved heterogeneity through residual layers and skip connections. Sifringer et al. (2020)'s approach augments traditional discrete choice models with a neural network component, leading to the Learning Multinomial Logit (L-MNL) and Learning Nested Logit (L-NL) models, which improve predictive performance and parameter estimation accuracy while maintaining interpretability.

1.2 Research Gaps

Research Gap1: Despite the extensive literature on activity-travel behaviour, the determinants of who people interact with remain underexplored. Most studies have focused on how social networks influence the choice of activity or start time, with limited attention to the factors driving the selection of social contacts for various activities, particularly the variables on the ego-alter level. In the literature, incorporating ego-alter level variables has been shown to enhance model performance in various contexts. For example, Sharmeen & Timmermans (2014) demonstrated that including ego-alter variables significantly improved the prediction of interaction frequency between network members, while van den Berg et al. (2012a) found that relationship characteristics (e.g., relationship type and tie strength) enhanced the modelling of activity duration. Kowald et al. (2013) further showed that integrating these variables led to a better understanding of the dis-

tance patterns of social contacts; for example, people tend to keep long-distance relationships with alters of similar age.

However, the existing models often treat social contacts as several aggregated categories, potentially missing out on behavioural nuances. By disaggregating social contact choices for leisure activities into ego-alter level, the proposed social-contact choice models can provide a more detailed understanding of how specific activity types are associated with different social network members. For instance, engaging in sports with a social network member may be influenced more by age and gender homophily, whereas cultural activities with a social network member may be more influenced by education level and relationship duration. This enables a richer representation of the social contact choice in leisure activity, potentially improving model performance and offering deeper insights compared to state-of-the-art models that rely on broad categories of social contacts.

Understanding these determinants is crucial for developing more comprehensive models of human behaviour, which can provide deeper insights into the mechanisms underlying social interactions and activity participation. The primary usefulness of this research lies in enhancing our understanding and knowledge of how individuals make decisions in this specific context, rather than serving an immediate practical application. Engaging in activities with someone from one's social network often involves travel, and the usefulness of this research may depend on other related choices. Thus, while the current model might not have direct standalone applications, it can play an important role in joint modelling approaches, such as those explored in the work of Habib et al. (2008) and Habib and Carrasco (2011). Furthermore, these findings can serve as valuable inputs for simulation models. Furthermore, our insights into the role of distance, for instance, can aid transport planning by highlighting the extent to which individuals engage in activities with their social network members within a specific distance. This understanding can help planners focus on particular transport corridors, such as between two cities, where social interactions are more likely to occur.

Research Gap2: Machine learning is often known as the "black box" model, lacking the interpretations backed by theory. However, it is also known for its performance. Following the stream of "Take the best of both worlds? ML to CM," one identified research opportunity is leveraging machine learning to assist in choice model specification, particularly in the context of social networks. This approach aims to enhance model performance and provide further insights into the factors driving decision-making processes.

1.3 Objectives

This thesis aims to fill this gap by investigating the determinants of social contact selection in leisure activities. Specifically, it seeks to:

1. Identify key factors influencing the "with whom" choice in different social contexts. Develop and validate models that account for these factors, enhancing the explanatory power of activity-travel behaviour models.

2. Integrating machine learning techniques with traditional choice modelling to improve model specification and performance, particularly in the context of social network analysis.

1.4 Thesis Outline

The thesis is structured into a chapter on the data, two main chapters, followed by a conclusion and discussion:

- Chapter 2: This chapter describes the data that is used in both Chapter 3 and Chapter 4

- Chapter 3: This chapter focuses on the selection of social contacts in leisure activities, exploring how individual and contextual factors influence these choices. It utilises data collected through snowball sampling to analyse the patterns and determinants of social contact selection

- Chapter 4: This chapter integrates machine learning techniques with traditional choice modelling to enhance the specification and performance of models predicting social contact selection. It demonstrates the application of Shapley Additive exPlanations (SHAP) in the context of social network analysis.

- Chapter 5: The final chapter summarises the key findings, discusses their implications for theory and practice, and suggests directions for future research.

Chapter 2

Data

This chapter introduces the dataset that underpins the analysis of social contact selection within the context of activity-travel behaviour. This dataset is integral to exploring how individuals make choices regarding their leisure activities and the social contacts they involve. The same dataset will be utilised in both Chapter 3 and Chapter 4.

In Chapter 3, the dataset will be used to investigate the factors influencing the selection of social contacts for various leisure activities. This will provide insights into the social dynamics and patterns of activity participation, focusing on explaining the ”with whom” decision-making process. Chapter 4 will extend this analysis by integrating machine learning techniques with traditional choice modelling to enhance model specification and performance. This approach aims to uncover hidden patterns and improve the interpretability of behavioural models, thereby contributing to a deeper understanding of social network influences on activity-travel behaviour.

The data used in the study was collected in Switzerland by the Institute for Transport Planning and Systems (IVT) of ETH Zurich between January 2009 and March 2011 (Kowald & Axhausen, 2014). It was collected to gain a better understanding of leisure activity engagement. The data was collected through a snowball sample. The rationale of snowball sampling is to conduct the survey with people named by the respondents who are already in the sample. With each iteration, more respondents are recruited, hence the name snowball sampling. An example of the process is demonstrated in Figure 2.1. An ego is a respondent, and social contacts named by the respondents are alters. An initial ego reported four alters. The alters reported by the initial ego in the first iteration (Alter1, Alter2, and Alter4) became egos themselves as they were contacted and successfully recruited into the survey (indicated in blue). Alter3 did not respond (indicated in red). Each new ego in iteration 1 then reported their own alters, leading to the

formation of iteration 2.

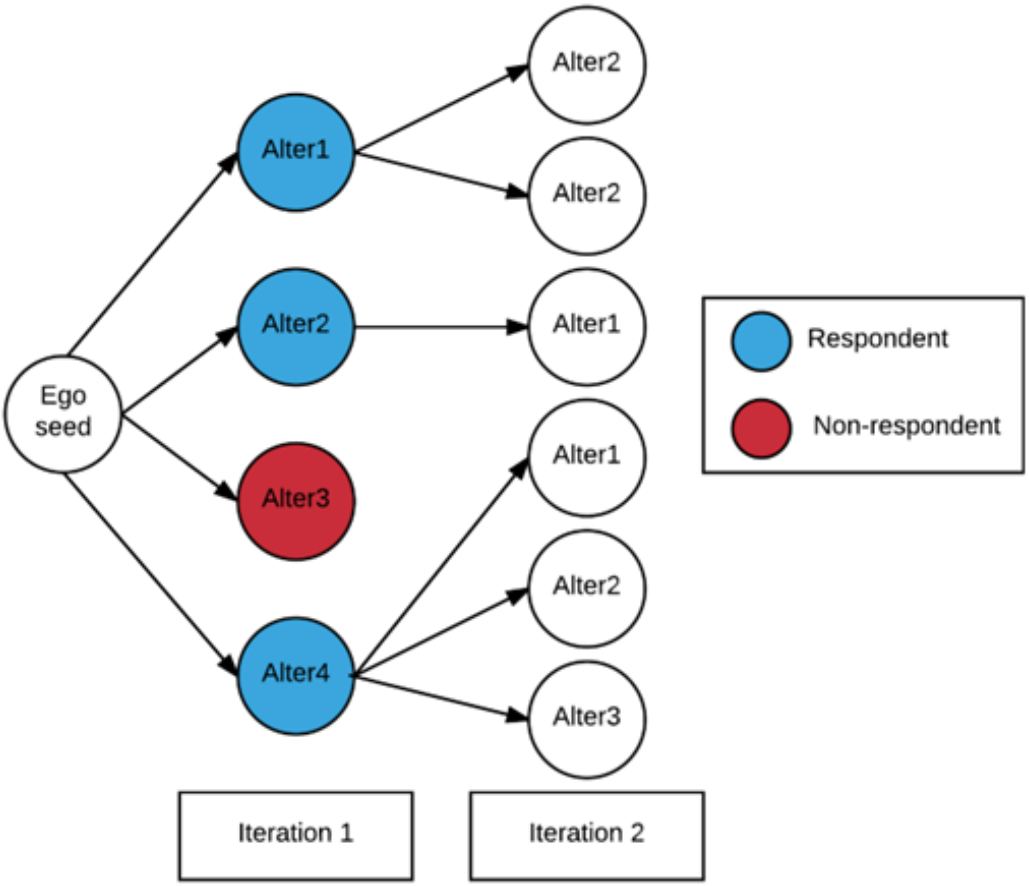

Figure 2.1: Snowball sampling with three iteration levels. Source: Calastri, Hess, Daly, Maness, et al. (2017)

While snowball sampling is effective for studying network structures and understanding the dynamics of social interactions, it comes with several disadvantages. Firstly, it does not meet the criteria for random sampling, leading to selection bias as respondents are selected based on their social connections rather than randomly from the target population. This results in unequally distributed probabilities for individuals to become part of the sample. Secondly, the method can

suffer from homophily bias, where the recruited sample tends to be homogeneous because people are more likely to nominate others who are similar to themselves in behaviour, characteristics and shared preferences. Lastly, there is degree bias, meaning individuals with larger social networks are more likely to be included in the sample multiple times, leading to overrepresenting well-connected individuals. Despite these disadvantages, snowball sampling is particularly suitable for our study because it effectively captures the complex network of social interactions and relationships, which is crucial for examining social contact selection in leisure activities. Moreover, the dataset provides comprehensive insights into the social context and the nature of relationships between respondents, making it invaluable for modelling and understanding activity-travel behaviour. Further details on these disadvantages and the rationale for using this sampling method can be found in Kowald (2013).

The survey utilized two key components: a name generator and a sociogram. The name generator component asked respondents to list their social network contacts (see Figure 2.2). For each of these contacts, respondents provided key information. The sociogram component then asked respondents to mention activities they do with their social contacts and to list the social contacts that regularly join each activity. The social contacts they listed in the sociogram were limited to those identified in the name generator.

Figure 2.3 is an empirical example of sociogram data. In this example, the ego O named 14 social contacts A-N. Four leisure activity groups exist (group 1 to group 4). Members within the same leisure activity group are fully connected. Alter I join two different leisure activity groups (group 3 and group 4) and thus create an overlap between them. D and alter N do not participate in any leisure activity group (they are known as the isolates).

Please list the people with whom you make plans to spend free time. (Examples: errands, sports, club or organized activities, cultural events, cooking together or going out to eat, taking holidays or excursions together)

No.	Name	No.	Name	No.	Name
Bsp	*Bsp. Doris Musterfrau*	13		27	
		14		28	
1		15		29	
2		16		30	
3		17		31	
4		18		32	
5		19		33	
6		20		34	
7		21		35	
8		22		36	
9		23		37	
10		24		38	
11		25		39	
12		26		40	

If there are other people with whom you discuss important problems, please list them here.

Figure 2.2: The name generators. Source: Kowald (2013)

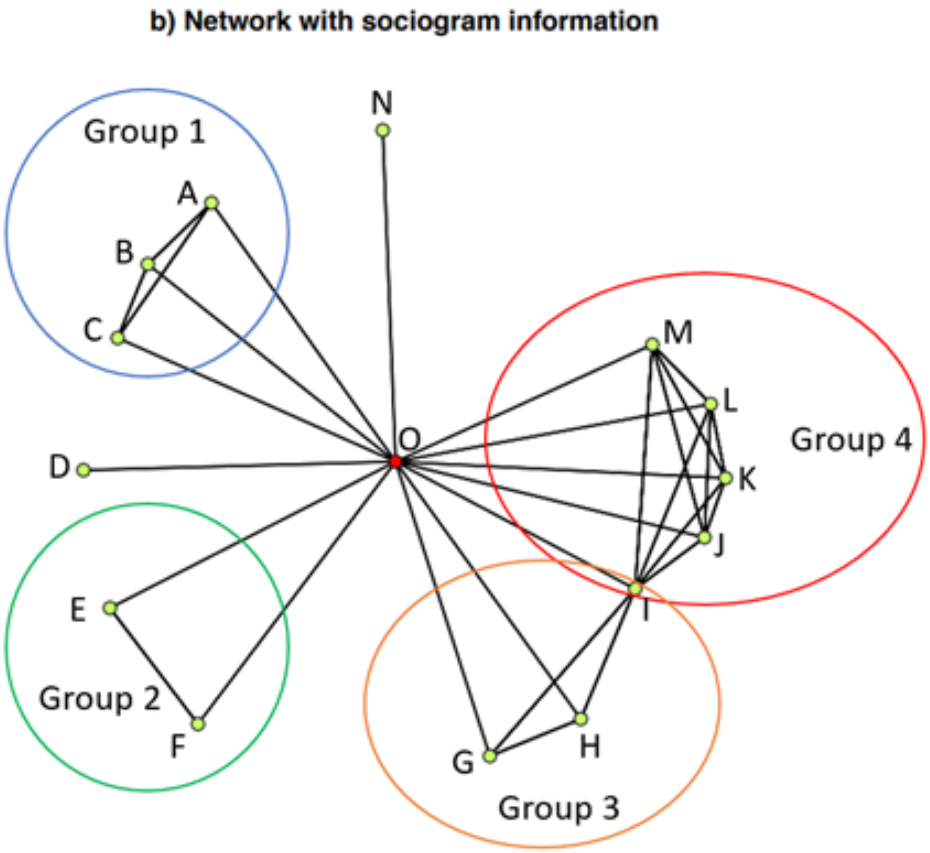

Figure 2.3: An example of sociogram data. The red node in the middle is the ego, and the green dots are the alters. Source: Kowald (2013)

2.1 Data descriptives

Respondents who reported no social contacts or one social contact are excluded from the analysis because there is no choice process in either case. The sample is further reduced, excluding egos who reported 0 activities because there is no other way to know the activities that the egos and alters conduct together. The final sample is made of 639 egos and 14,009 alters. On average, each person reports 22 social contacts. Most people reported more than 5 social contacts, and this is true for over 98% of egos in the sample.

The datasets have many potential explanatory variables (more than 70, excluding identification variables such as iterations in snowball sampling) as displayed in Table 2.1.

Type of variables	**Variable**
Variables on ego level	
	Civil status
	Education level
	Age
	Sex
	Car availability
	Mobility characteristics
	Status of employment
	Household Income
	Number of persons in household
	Driver license
	Birth county
	Citizenship
	Second Citizenship
	Number of main residencies an ego had in course of life
	Number of working people in household
	Number of education places
	Number of years in education
	Internet access
	Mobile phone possession
Variables on ego-alter level	
	Sex

Continued on next page

Table 2.1 – continued from previous page

Type of variables	Variable
	Age
	Core contact
	Ask for help
	Discuss important problems
	Context of meeting
	Variables of locations of 1st meeting
	Variables of home location
	Types of relationship
	Alter's degree centrality
	Alter's betweenness
	Relation duration
	Contact frequencies per year:
(1)	face-to-face
(2)	phone
(3)	e-mail
(4)	SMS
(5)	internet chat
	Distance (KM)
	Similarity of socioeconomics
	Tie strength
	Citizenship
	Language
	Civil status
	Education level
	Preferred language for discussions
Variables on ego-network level	
	Network size (various definitions)
	Network degree (various definitions)
	Degree Centralisation (various definitions)
	Betweenness centralisation (various definitions)
	Network density (various definitions)
	Network composition (proportion)
	Number of components
	Variability of with whom

Continued on next page

Table 2.1 – continued from previous page

Type of variables	Variable
	Homophily of socioeconomics
	Number of cliques (activity groups) reported
	Average number of participants per clique
	Number of strong ties in personal network
	Number of weak ties in personal network
	Varibales of home location

Table 2.1: Independent variables in the data

As demonstrated in Figure. 2.4, the distribution of reported distances is skewed toward the lower end, in fact, over 93% at a distance less or equal to 100 KM. Thus, it is reasonable to censor at 100 KM, as most ego-alter live within this more localised range.

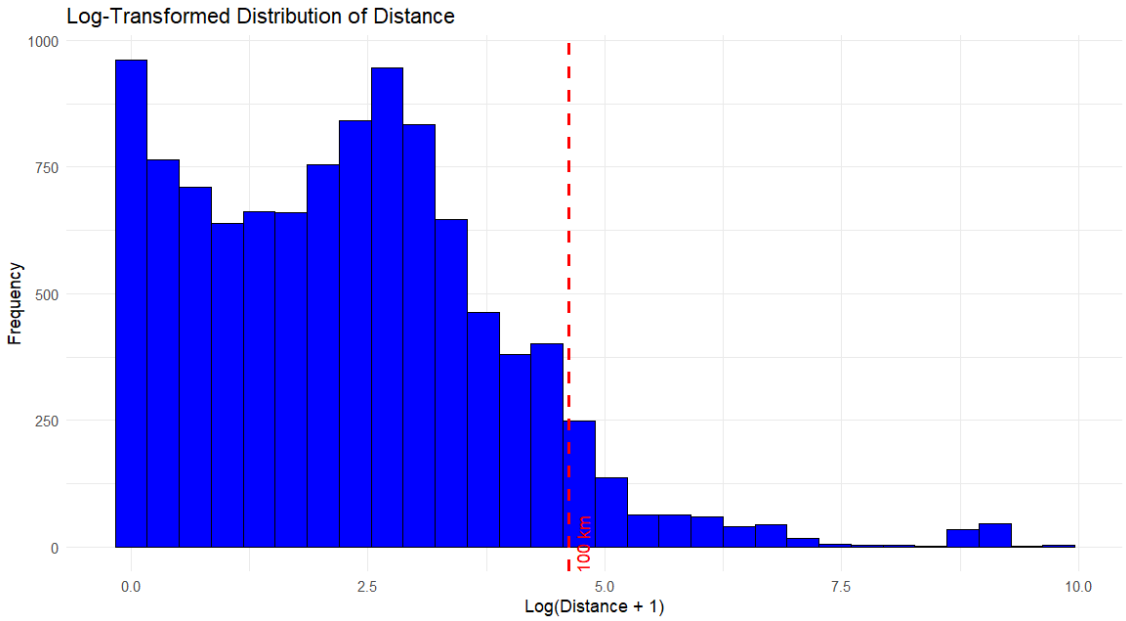

Figure 2.4: Distribution of distance (Log-transformed)

Figure 2.5 presents the log-transformed distribution of relationship duration. Log transformation is applied because we hypothesised that once an ego knows an alter for a long time, the effect of this on the choice diminishes. This means

that the incremental impact of additional time knowing an alter becomes smaller over time. As shown in the figure, most relationships fall within the middle range (i.e., between 2 and 4 on the log scale). This suggests that most relationships have a moderate duration, with fewer relationships being either very short or long.

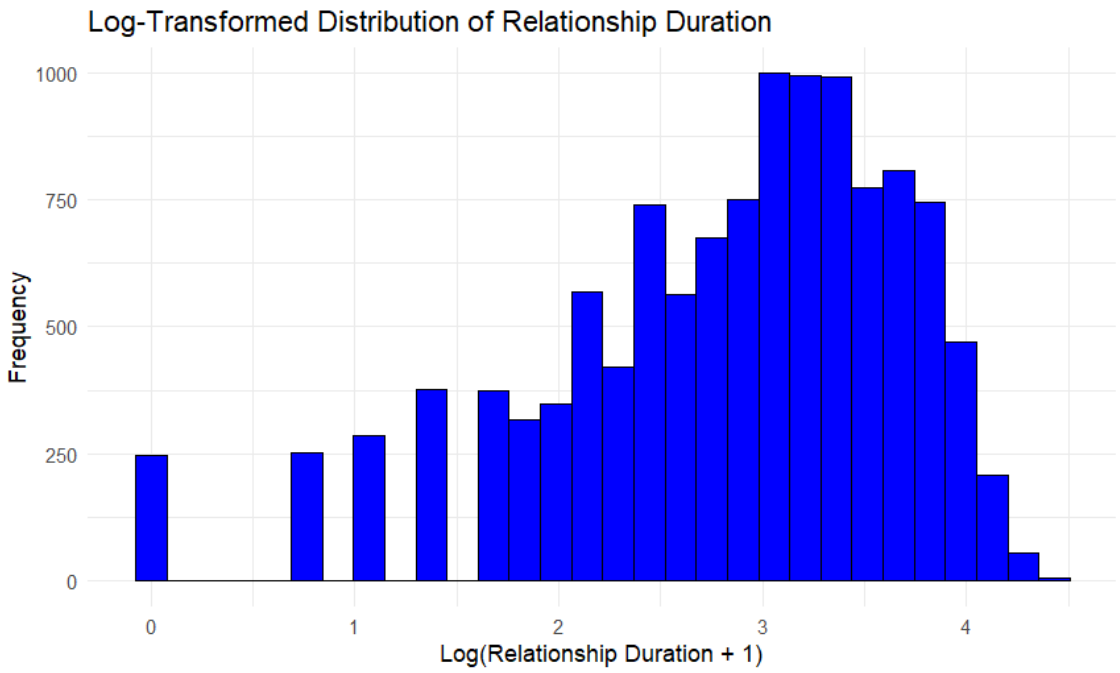

Figure 2.5: Distribution of relationship duration (Log-transformed)

Table 2.2 shows the sample demographics. In our study, age homophily refers to ego-alter pairs with age differences within ten years. A majority of the sample exhibits age homophily, indicating that most ego-alter pairs are close in age. The sample is fairly evenly split between those who are both married and those who are not. There is a higher proportion of ego-alter pairs where both individuals are female. Friends constitute the largest category of ego-alter pairs, followed by acquaintances and relatives.

Table 2.2: Sample demographics

Characteristic	**N**	**%**
Age homophily	9428	68.74%
Not age homophily	3210	23.41%
Both married	6944	50.63%
Not both married	6770	49.37%
Both male	3012	21.96%
Both female	5860	42.73%
Spouse	291	2.12%
Relative 1st degree	1886	13.75%
Other relative	817	5.96%
Married into family	724	5.28%
Friend	5578	40.67%
Acquaintance	3932	28.67%

Table 2.3 details activity participation by age homophily. It can be observed that most activities are conducted with alters who are of similar age, suggesting the influence of age similarity on activity companionship.

Table 2.3: Activity participation by age homophily.

Activity Type	**Age Homophily**	**N (%)**
Excursions		
	Age homophily	141 (79.21%)
	Not age homophily	27 (15.17%)
Sport		
	Age homophily	604 (78.24%)
	Not age homophily	115 (14.90%)
Eating out		
	Age homophily	402 (79.92%)
	Not age homophily	71 (14.12%)
Hobby		
	Age homophily	756 (75.98%)
	Not age homophily	168 (16.88%)

Continue on the next page

Activity Type	Age Homophily	N (%)
Culture		
	Age homophily	104 (64.60%)
	Not age homophily	45 (27.95%)
Visiting		
	Age homophily	33 (58.93%)
	Not age homophily	18 (32.14%)

Table 2.4 provides data on activity participation by marital status homophily. Activities such as culture and sports show varying degrees of participation based on whether both individuals are married.

Table 2.4: Activity participation by marital status homophily

Activity Type	Marital Status	N (%)
Culture		
	Both married	97 (60.25%)
	Not both married	64 (39.75%)
Hobby		
	Both married	531 (53.37%)
	Not both married	464 (46.63%)
Visiting		
	Both married	36 (64.29%)
	Not both married	20 (35.71%)
Excursions		
	Both married	91 (51.12%)
	Not both married	87 (48.88%)
Eating out		
	Both married	258 (51.29%)
	Not both married	245 (48.71%)
Sport		
	Both married	438 (56.74%)
	Not both married	334 (43.26%)

Table 2.5 details activity participation by sex homophily. Activities such as visiting, sports, and hobbies demonstrate varying degrees of participation based on sex homophily.

Table 2.5: Activity participation by sex homophily.

Activity Type	Sex Homophily	N (%)
Visiting		
	Both male	2 (3.57%)
	Both female	37 (66.07%)
Sports		
	Both male	249 (32.25%)
	Both female	291 (37.69%)
Hobby		
	Both male	281 (28.24%)
	Both female	448 (45.03%)
Culture		
	Both male	37 (22.98%)
	Both female	81 (50.31%)
Eating out		
	Both male	104 (20.68%)
	Both female	230 (45.73%)
Excursions		
	Both male	47 (26.40%)
	Both female	70 (39.33%)

Table 2.6 shows activity participation by relationship type. Friends are the most common companions for all activities, indicating their significant role in leisure activities.

Table 2.6: Activity participation by relationship type.

Activity Type	Relationship Type	%
Culture		
	Spouse	2.56%
	Relative 1st degree	5.13%
	Other relative	0%
	Married into family	1.28%
	Friend	55.77%

Continue on the next page

Activity Type	Relationship Type	%
	Acquaintance	35.26%
Sport		
	Spouse	2.65%
	Relative 1st degree	5.16%
	Other relative	2.25%
	Married into family	1.72%
	Friend	55.53%
	Acquaintance	37.70%
Eating out		
	Spouse	2.79%
	Relative 1st degree	6.18%
	Other relative	3.78%
	Married into family	1.59%
	Friend	56.57%
	Acquaintance	29.08%
Hobby		
	Spouse	1.33%
	Relative 1st degree	2.76%
	Other relative	1.13%
	Married into family	0.81%
	Friend	53.02%
	Acquaintance	40.94%
Excursions		
	Spouse	1.16%
	Relative 1st degree	8.72%
	Other relative	4.07%
	Married into family	2.33%
	Friend	52.33%
	Acquaintance	31.40%
Visiting		
	Spouse	5.36%
	Relative 1st degree	37.5%
	Other relative	0%
	Married into family	12.5%
	Friend	32.14%

Continue on the next page

Activity Type	Relationship Type	%
	Acquaintance	12.50%

Chapter 3

Social Contact Selection in Leisure Activities

3.1 Introduction

Several studies have highlighted specific patterns; individuals tend to engage more in joint activities with family members providing emotional support (Lin & Wang, 2014), display a propensity for social activities with neighbours (Carrasco & Miller, 2006), and show a greater frequency of social activities with friends, males, and very close alters (Carrasco & Miller, 2009). Further, it has been shown that people tend to have more social interactions with strong tie contacts (van den Berg et al., 2012b). Collectively, these studies underscore the importance of "with whom" selection in understanding and modelling activity-travel behaviour.

However, the current literature predominantly explores the "with whom" factor as an explanatory component of activity-travel behaviour (choice of activity or start time), with limited focus on identifying determinants of the "with whom" selection process itself. Understanding the determinants of this selection process is crucial because it allows for the development of more accurate and comprehensive models of human behaviour. By improving our models, we can gain deeper insights into the underlying mechanisms that drive social interactions and activity participation. This study aims to enhance the understanding of how individuals choose their social network members for different activities, which can lead to better model fit and more robust behavioural insights. This, in turn, has significant implications for urban planning, transportation policy, and social well-being. Improved "with whom" models can inform data collection and modelling efforts, enabling future studies to incorporate more nuanced social dimensions into their

analyses. With a better "with whom" model, joint models such as the trivariate econometric model used by Habib and Carrasco (2011) and potentially more advanced approaches like the MDCEV (Multiple Discrete-Continuous Extreme Value) model can better account for continuous variables like activity duration alongside social interactions. This advancement will improve the explanatory power of joint models and provide valuable insights for policymakers and researchers aiming to foster more connected and efficient communities.

In addition, individual characteristics significantly affect companionship (i.e., household member and non-household member) choices for leisure activity (Srinivasan & Bhat, 2006). Hasnine et al. (2022) found that factors such as modal accessibility, household size, and age significantly influence decisions about living arrangements (i.e., with parents/family, with a partner, living with a roommate(s), and living alone). This finding is relevant because it highlights how individual characteristics and contextual factors influence decisions about "with whom" to live, which is a specific type of "with whom" selection. Similarly, research has emphasized the role not just of the personal attributes of the individuals but also of the characteristics of their personal networks in shaping "with whom" choices (Habib et al., 2008). For example, a higher proportion of friends in one's network increases the likelihood of participating in social activities with friends. Nonetheless, these studies primarily considered aggregated categories of social contacts such as family, friends, or household members, thereby overlooking important factors like the specific characteristics of individual social network members (alters) and the detailed nature of dyadic relationships. For instance, factors such as the physical distance between the ego and alter, whether both individuals are female or if both are married, can significantly influence "with whom" decisions. Such detailed dyadic variables are often not captured when data and models are based on broader categories, like "with friends" or "with family," which do not account for the nuanced attributes of each individual relationship.

3.1.1 Initial Hypotheses

Based on intuition and guidelines from previous studies in related literature, we propose the following initial hypotheses regarding the potential relevance/significance of certain explanatory variables existing in the dataset across various categories of activities:

- **Age:** Age is shown to influence preferences for socialising with family or friends, especially for older individuals (K. M. N. Habib & Carrasco, 2011).

- **Age Homophily:** Age homophily, where individuals are more likely to engage in activities with others of a similar age, is a well-documented phenomenon in social behaviour (Carrasco & Miller, 2009).

- **Gender:** Gender may play a role in determining the choice of social network members across various activities. Males, for example, may prefer to socialise with friends rather than family, suggesting that gender could influence the choice of social network members for different social activities (K. M. Habib et al., 2008).

- **Gender Homophily:** Gender homophily, the tendency for individuals to interact with others of the same gender, can also play a role in social interactions. van den Berg et al. (2012a) found that gender similarity correlates with the duration of social activities, suggesting that it might also influence the choice of social network members across various social activities.

- **Relationship Type and Household Dynamics:** The nature of the relationship (e.g., friends, family, household members) and the roles individuals occupy within their households/marital status (e.g., household heads, adults with partners) could impact the choice of social network members for social activities (K. M. Habib et al., 2008). Additionally, household size influences social choices, with larger households being less likely to socialise without household members (K. M. N. Habib & Carrasco, 2011).

- **Network Composition:** The composition of an individual's social network (e.g., proportion of friends) may influence the likelihood of participating in social activities with different social network members. A higher proportion of friends in the network increases the probability of engaging in social activities with friends, while a higher number of family members increases the likelihood of participating in family-based activities (K. M. Habib et al., 2008).

- **Relationship Duration:** The length of a relationship between individuals has been shown to affect the frequency and mode of social interactions (Calastri, Hess, Daly, Maness, et al., 2017), suggesting that relationship duration might still be a relevant variable in explaining social activity engagement (e.g., engagement in certain types of social activities might decrease with relationship duration but will be compensated by increased communication via other methods)

- **Income:** Income can play a role in the ”with whom” decision. For example, higher income can increase the likelihood of interacting with family members K. M. N. Habib & Carrasco (2011).
- **Physical Distance:** The physical proximity of individuals may be a factor in deciding with whom to engage in social activities. The likelihood of participation might decrease as the distance between individuals increases, making physical distance a potentially significant variable across various activity types (Axhausen, 2005).
- **Education Level:** Education level is strongly linked to individuals' cultural activities, such as museum visits or concert-going (Bourdieu, 2018). Consequently, it might be reasonable to infer that individuals with similar educational backgrounds may engage in cultural activities.
- **Degree Centrality:** Degree centrality, which measures how connected an alter is to other network members, has been shown to positively correlate with social activity frequency (Carrasco & Miller, 2009). Alters with more direct connections are more likely to interact frequently with the ego, suggesting that degree centrality could influence the ”with whom” choice across various activities.
- **Degree Centralisation:** Degree centralisation measures how concentrated connections are in a few central alters. Higher degree centralisation has been linked to a greater propensity for socialising within the network (Carrasco & Miller, 2006), suggesting that it could be a factor in predicting with whom individuals choose to engage in social activities.
- **Network Density:** Network density, defined as the proportion of actual connections to all possible connections in a network, has been shown to correlate with higher interaction frequency between ego and alters (Carrasco & Miller, 2009). A denser network may lead to more frequent social interactions, making this variable relevant to the ”with whom” choice across different activities.

3.2 Method

An ego-alter pair can engage in one of six types of activities. In total, there are 7 distinct alternatives. One of these is the null alternative, indicating the ego is not

engaging in any of the six activities with a given alter.

For each ego i, under the paradigm of utility maximisation, a certain amount of utility is derived from each choice alternative given by:

$$U_{ijn} = \alpha_n + \varepsilon_{ijn} \tag{3.1}$$

where i is the ego; j is the alter;n is one of the alternatives;α_n is an alternative-specific constant; and ε_{ijn} is the disturbance term. The disturbances ε_{ijn} are assumed to be independently and identically distributed (IID) with a Type I Extreme Value distribution (Gumbel distribution).

The utility function in baseline model with just ASC (3.1) was first extended to accommodate the effects of variables on ego level

$$U_{ijn} = \alpha_n + \beta_n X_{in} + \varepsilon_{ijn} \tag{3.2}$$

Where X_{in} the explanatory variables are the ego-level variables; β_n represents the effect of ego-specific characteristics on the probability of choosing alternative n.

Subsequently, the utility function in 3.2 was extended to accommodate the effects of the variables on the ego-network level. By extending Eq. (3.2), the utility of ego i choosing alter j for activity n could then be written

$$U_{ijn} = \alpha_n + \beta_n X_{in} + \gamma_n N_{in} + \varepsilon_{ijn} \tag{3.3}$$

Where N_{in} the explanatory variables are the ego-network level variables; γ_n represents the effect of network characteristics on the probability of choosing alternative n.

The utility function in 3.3 was then extended to accommodate the effects of the dyadic variables (the variables on ego-alter level)

$$U_{ijn} = \alpha_n + \beta_n X_{in} + \gamma_n N_{in} + \zeta_n A_{ijn} + \varepsilon_{ijn} \tag{3.4}$$

where A_{ijn} the ego-alter level variables; ζ_n represents the effect of dyadic variables on the probability of choosing alternative n between ego-alter pairs.

The resulting structure allows us to consider how individual characteristics and preferences and the specific ego-alter relationship affect the utility derived from different activities. Therefore, our model encompasses variables at the ego level and ego-alter level, which helps us examine the impact of these factors on the utility derived from different activities with an alter and, consequently, on the likelihood of different activities being chosen with an alter.

The multinomial logit (MNL) choice model used in this study assumes mutually exclusive alternatives, meaning that each ego-alter pair can only engage in one type of activity at a time. This assumption simplifies the model and is necessary for the proper application of the MNL framework, as overlapping activities would violate the requirement for mutually exclusive alternatives.

To understand the necessity of this simplification, a preliminary analysis of the data revealed that each ego-alter pair does not engage in more than one kind of activity. Table 3.1 and Table 3.2 provide a detailed breakdown of the number of activities and their combinations:

Number of Activity Types	**Frequency**	**%**
0 types of activities	11049	78.82%
1 type of activities	2665	19.02%
2 types of activities	243	1.73%
3 types of activities	46	0.33%
4 types of activities	4	0.03%
5 types of activities	1	0.01%
Other	0	0.00%
Total	**14009**	**100%**

Table 3.1: Frequency of the number of activity types that the ego conducts with a given alter

Table 3.2: Activity Combinations and Frequencies

Activity Combination	**Frequency**
No activities	11049
Hobby	995
Culture	161
Culture, Hobby	10
Visiting	56
Visiting, Hobby	1
Visiting, Culture, Hobby	2
Continued on next page	

Activity Combination	Frequency
Excursions	178
Excursions, Hobby	41
Excursions, Culture	9
Excursions, Visiting	2
Excursions, Visiting, Hobby	1
Eating Out	503
Eating Out, Hobby	52
Eating Out, Culture	4
Eating Out, Culture, Hobby	2
Eating Out, Visiting	1
Eating Out, Excursions	18
Eating Out, Excursions, Hobby	4
Eating Out, Excursions, Visiting	3
Sport	772
Sport, Hobby	41
Sport, Culture	3
Sport, Culture, Hobby	1
Sport, Visiting, Hobby	1
Sport, Excursions	34
Sport, Excursions, Hobby	2
Sport, Excursions, Visiting	3
Sport, Excursions, Visiting, Hobby	2
Sport, Eating Out	28
Sport, Eating Out, Hobby	14
Sport, Eating Out, Culture	4
Sport, Eating Out, Excursions	9
Sport, Eating Out, Excursions, Hobby	1
Sport, Eating Out, Excursions, Culture, Hobby	1
Sport, Eating Out, Excursions, Visiting	1
Total	**14009**

Table 3.2: Frequency of combinations of activities that the ego conducts with a given alter

From Table 3.1, it is evident that most ego-alter pairs (11049 pairs) are not engaged in any activities together, while 2665 pairs are involved in only one type of activity. Only 295 pairs are involved in more than one type of activity, as further

detailed in Table 3.2.

Given the small number of observations involving multiple activities (295 out of a total of 14009 ego-alter pairs), including these combinations as separate alternatives would lead to a sparse data problem. This scarcity could compromise the reliability and stability of the model estimates due to insufficient data for many of the activity combinations.

The decision to simplify the model by removing from the data any scenarios where the ego engages in multiple types of activities with the alters was made to ensure robust estimation and to maintain the tractability of the model. The limited observations for combined activities do not provide a strong statistical foundation for more complex modelling approaches that would incorporate multiple concurrent activities.

An alternative approach could involve creating composite alternatives representing combinations of activities. However, this would exponentially increase the number of alternatives and further complicate the data sparsity issue. Another approach could involve using a nested logit model or a mixed logit model to account for correlations between activities, but these models would also require sufficient data for reliable estimation. For example, although a nested logit model could theoretically account for the correlations between different types of activities, it was not implemented due to data limitations. Specifically, the small number of ego-alter pairs engaging in more than one activity would lead to an insufficient number of observations for reliable estimation of the correlation structure in a nested logit framework. We also explored a multivariate probit (MVP) model to account for possible correlations between activities, but, as we will discuss later in Chapter 5, the data's richness was a limiting factor. The small number of ego-alter pairs engaging in multiple activities compromised the robustness of this model.

In conclusion, the simplification to mutually exclusive alternatives, while an assumption, is justified based on the observed data distribution. The majority of ego-alter pairs engage in zero or one type of activity, supporting the use of the MNL model in its current form. This approach ensures that the model remains statistically robust and computationally feasible, providing reliable insights into the factors influencing activity choices among ego-alter pairs.

The choice probability of alternative n for ego i and alter j is given by

$$P_{ijn} = \frac{e^{f(X_{in}, A_{ijn}; \beta)}}{\sum_{m \in L} e^{f(X_{im}, A_{ijm}; \beta)}} \tag{3.5}$$

where L is the choice set and the denominator is the sum of the exponential of the

systematic utility associated with all alternatives.

In practice, this means that the probability of choosing a particular activity n (such as Sport, Eating out, etc.) by an ego i with an alter j depends on the utility derived from that activity relative to the utility derived from all other available activities in the choice set. All six activities listed in the Data section (section 2) are available to each ego-alter pair. Additionally, there is a null alternative, representing no engagement in any of the six activities. Therefore, the total number of alternatives available in the choice set L for each ego-alter pair is seven, including the null alternative.

This means that for any given ego i and alter j, the model considers these seven alternatives when determining the probability of engaging in a specific activity.

This framework facilitates the understanding of how individual characteristics and specific ego-alter relationships influence the likelihood of different activities being chosen.

3.3 Empirical results

In this section, we present the results of the analysis, focusing on modelling the choice of the type of activities that an ego conducts with a given alter. The activities reported by the respondents have been grouped into several aggregate categories to facilitate a clearer understanding of social behaviour patterns. This categorisation follows the logic of a sociogram, which identifies activities and then lists the social contacts who regularly participate in these activities. This approach allows for a structured analysis of the types of activities and the social dynamics involved. Essentially, we created a choice variable that was previously implicit in the data, enabling a more detailed examination of the participant selection process for different types of activities.

The activities that the egos perform with their alters have been allocated to several aggregate activity categories as follows:

- Sport (e.g., tennis, football, basketball, swimming, skiing)
- Eating out (e.g., dinner, barbeque, coffee, restaurant visits)
- Hobby (e.g., painting, reading, music, gardening, yoga)
- Excursions (e.g., hiking, mountain tours)

- Culture (e.g., concerts, theatre, museums)

- Visiting (e.g., visiting friends and family)

These categories were determined based on the types of activities reported by the respondents. For example, the "Sport" category includes activities such as tennis, football, basketball, swimming, and skiing. The "Eating out" category includes activities such as dinner, barbeque, coffee, and restaurant visits. The "Hobby" category encompasses a wide range of personal interests, such as painting, reading, music, gardening, and yoga. "Excursions" involve outdoor activities such as hiking and mountain tours. The "Culture" category includes arts and culture-related activities, such as concerts, theatre, and museums. The "Visiting" category involves social visits to social network members.

The choice of categories is inspired by the classification of the share of the number of trips according to leisure purposes in the German travel survey (Illenberger, 2012).

The distribution of activity types that egos engage in with their alters in the study data is summarized in Table 3.3. As shown in Table 3.3, hobbies represent the largest category, followed by sports and eating out. This distribution highlights the predominance of personal interests and physical activities in the social interactions of the respondents, while visits and cultural activities are less common.

Activity Type	**Percentage**
Sports	29%
Hobby	37%
Eating Out	19%
Excursion	7%
Visit	2%
Culture	6%

Table 3.3: The table shows the distribution of activity types egos engage in with their alters based on the study data.

3.3.1 Correlation analysis

Correlation analysis was conducted to examine the relationships between the explanatory variables and identify potential multicollinearity issues. A comprehensive analysis has been conducted, considering the nature of each variable.

- Continuous Variables Analysis: Pearson's correlation coefficients calculated for continuous variable pairs, such as "distance" and "relation duration," have shown only weak correlations (e.g., 0.131). These results suggest negligible multicollinearity that would affect parameter estimation.
- Categorical Variables Analysis: Cramer's V statistics were employed for categorical variable pairs. The values ranged from very weak (e.g., 0.022 between "alter's education level" and "ego's sex") to moderate associations (e.g., 0.316 between "ego's sex" and "alter's sex"). While some variables demonstrated moderate associations, none reached a level that would suggest a risk of multicollinearity compromising the model.
- Mixed Variable Types Analysis: For the pairs consisting of one continuous and one categorical variable, a practical assessment approach was taken. The stability of coefficient estimates and robust t-ratios indicated no substantial changes across model specifications when variables were added or removed. For instance, the removal of "relation duration" resulted in minor changes in the estimates for "age homophily" across alternatives such as cultural activities, sports, and dining out, with robust t-ratios suggesting the variable remains significant.
- Specific Findings and Conclusions:
  - The coefficient for "age homophily" (-0.463 to -0.486) and associated robust t-ratios (-1.88 to -2.01) remained stable, affirming that collinearity with "relation duration" is not a concern.
  - A consistent pattern was observed for "age homophily" across various activities, with minimal fluctuation in the estimates and significance levels when "distance" was excluded.
  - For categorical pairs, the moderate association between "ego's sex" and "alter's sex" was notable. However, since our model includes interaction effects between these variables, this association is accounted for and does not imply multicollinearity.

– Further analyses have been conducted to ensure a thorough examination of potential collinearity within our model. They echo the patterns observed in the presented findings, reinforcing our confidence in the model's robustness without delving into repetitive details here.

The analyses suggest that our model is robust with respect to collinearity. The measures of association, significance levels, and stability of estimates across specifications have been carefully considered. Although some variables are moderately associated, the effects are not strong enough to distort the model estimates significantly. The results have been consistent and aligned with theoretical expectations, indicating that the explanatory variables are contributing meaningful and distinct information to the model.

3.3.2 Core Results

The model was estimated using Apollo (Hess & Palma, 2019). The parameter estimates presented in Table 4.5 are evaluated for their statistical significance using robust t-ratios. The model includes a range of ego-level and ego-alter level variables, although some variables, such as income, age, and sex (except for the alternative of playing sport with a given alter), were tested but found to be non-significant and thus not included in the final model. Additionally, the following variables were removed due to endogeneity: frequency of different communication modes, help, discuss important problems, both help and discuss important problems, and contexts of 1st meeting. A number of network features have been suggested from the literature, but we are cautious about including them in our choice model because of the concern of endogeneity (mainly in line with omitted variable bias), including degree centralization, network size, network density, and network degree.

Using the likelihood ratio (LR) statistic to compare Specification 0 (ASC only) with Specification 1 (including variables at the ego level), Specification 2 (adding ego-network level variables), and Specification 3 (further including ego-alter level variables), as shown in Table 3.5, each subsequent model significantly improves in goodness-of-fit relative to its restricted version at a 95% confidence level or higher. Specifically, the LR statistics of 580.7, 134.16, and 627.12 provide strong evidence against the null hypothesis for each comparison, suggesting that the inclusion of ego-level, ego-network, and ego-alter level variables enhances the explanatory power of the original model.

While the likelihood ratio tests show stepwise improvements for each nested model, the AIC values further support the comparison between these specifications (see Table 3.4). Among the tested models, Specification 3 has the lowest AIC value (20123.22), indicating the best balance between model fit and complexity. This result suggests that the inclusion of ego-alter level variables significantly contributes to improving model fit by capturing additional dyadic relationship factors that are not accounted for by ego or ego-network characteristics alone.

Table 3.4: Model comparisons (a)

Model	$\log L$	**AIC**
ASC only (Eq.3.1; Specification 0)	-10678.60	21369.20
ASc plus variables on ego level (Eq. 3.2; Specification 1)	-10388.25	20838.49
ASc plus variables on ego level and ego-network level (Eq. 3.3; Specification 2)	-10321.17	20716.34
ASc plus variables on ego level, ego-network level and ego-alter level (Eq. 3.4; Specification 3)	-10007.61	20123.22

Compared Model	**Restricted Model**	**LR statistic**	**p-value**
Specification 1	Specification 0	580.7	< 0.001
Specification 2	Specification 1	134.16	< 0.001
Specification 3	Specification 2	627.12	< 0.001

Table 3.5: Model comparisons (b)

The results section presents the estimated coefficients for each activity type, where the sign and magnitude of these coefficients provide insights into the relative influence of different factors on the likelihood of engaging in specific activities with alters.

To support the explanations of our findings, we calculated the probabilities of engaging in each type of activity for different levels of explanatory variables (categorical ones). These probabilities were derived using the estimated model, which includes all relevant variables. We filtered the predicted probabilities for each key explanatory variable to focus on specific scenarios/levels of categorical variables (e.g., where age homophily is true). By averaging the probabilities within these levels, we examined the likelihood of engaging in different activities under these specific conditions.

Table 3.7 summarises the average probabilities for activities based on different levels of categorical explanatory variables.

Table 3.6: MNL (base alternative = no joint activity, with its ASC fixed to zero).

	Culture	Excursions	Visiting	Sports	Hobby	Eating Out
ASC	−4.8838 (-5.021)	−4.5926 (-18.183)	−6.9576 (-6.424)	−1.6930 (-3.776)	−1.7649 (-6.928)	−2.5764 (-5.689)
Distance	−0.0170 (-2.271)		0.0409 (2.031)	−0.0380 (-2.176)	−0.0334 (-4.128)	−0.0101 (-2.821)
Distance squared			−0.0005 (-2.431)	0.0003 (1.650)	0.0002 (2.303)	
Relationship duration	−0.1524 (-1.309)				−0.1461 (-2.131)	
Age homophily		0.6328 (2.503)		0.4889 (3.968)		0.2847 (1.362)
Both married				1.2247 (6.650)		
Both male			−1.7604 (-2.660)		0.3489 (2.353)	
Both female			0.6394 (2.098)			
Ego male				−0.7257 (-2.774)		
Household size	0.3931 (3.169)					
Proportion of ego-alter pairs with similar sex	0.0206 (1.774)	0.6328 (2.503)	0.0197 (1.287)	−0.0123 (-1.955)		−0.0117 (-1.658)
Proportion of ego-alter pairs with similar age	−0.0170 (-2.328)			−0.0099 (-1.990)		0.0135 (2.240)
Proportion of ego-alter pairs with similar education				0.0058 (1.859)	0.0073 (2.606)	
Proportion of ego-alter pairs with similar civil status			0.0197 (1.287)	0.0054 (1.501)		
Type of relationship						
Married into family	−1.6376 (-2.212)			−1.4628 (-4.082)	−2.2025 (-5.507)	−1.6340 (-3.430)
Relative 1st degree	−1.2001 (-2.431)			−1.1681 (-4.348)	−1.8902 (-8.169)	−1.2411 (-4.673)
Relative				−1.1217 (-3.786)	−1.8277 (-4.306)	−0.7947 (-2.583)
Spouse					−0.8520 (-2.826)	
Acquaintance						−0.2799 (-1.523)
Missing values coefficients						
Distance	−0.5528 (-1.354)			−0.4330 (-2.308)	−0.6751 (-3.755)	
Relation duration					−0.6013 (-1.989)	

Activity Type	**Similar Age**	**Not Similar Age**	**Both Married**	**Not Both Married**	**Both Female**	**Both Male**
Excursions	0.015	0.009	-	-	-	-
Sports	0.064	0.034	-	-	-	0.083
Eating Out	0.043	0.024	-	-	-	-
Culture	-	-	0.014	0.009	-	-
Visiting	-	-	-	-	0.006	0.001
Hobby	-	-	-	-	0.073	0.093

Table 3.7: Average probabilities for activities based on different explanatory variables.

3.3.3 ASC (Alternative Specific Constant)

The ASCs in this model capture the average effect of all unmeasured variables on the probability of choosing each activity, independent of the measured explanatory variables. These ASCs are crucial for understanding the baseline attractiveness of each alternative, especially within the context of ego-alter relationships. The negative ASCs for all activities indicate a general tendency against engaging in any of the listed activities with alters. The ASC for culture is -3.5113, excursions -4.5963, visiting -5.6770, sports -2.4427, hobby -1.3595, and eating out -2.8036, reflecting varying baseline levels of preference against these activities, with visiting having the strongest negative baseline preference. In other words, all else being equal, one alternative is more preferable than the other depends on the magnitude of the ASCs.

3.3.4 Distance

This continuous variable captures the physical separation (in kilometres) between each ego and their corresponding alter (see Figure 2.4). The distance was censored at 100km, and the variable was then entered into the model in a polynomial form (e.g., $x + x^2$, where x is distance).

The variable "Distance" has negative coefficients for alternatives such as "Eating Out with alters," "Playing Sports with alters," and "Engaging in hobbies with alters," suggesting that an increase in the physical distance between the ego and the alter consistently reduces the likelihood of engaging in these activities together. This indicates that proximity plays a significant role in the propensity to partake in eating, sports, and hobby activities with alters, with greater distances acting as a deterrent. The pattern reverses for "Visiting alters," where the positive coefficient indicates that as the distance increases, the likelihood of engaging in visiting activities actually rises. The positive coefficient for visiting, against the general trend of a negative Alternative-Specific Constant (ASC) and the negative coefficients for other activities, underscores the unique nature of visiting activities. They become more desirable or feasible as the distance increases, contrasting with the reduced likelihood of engaging in other activities under similar distance conditions. This, we recognise, is a counterintuitive result.

The negative coefficient for "Distance squared" in relation to visiting activities suggests a quadratic relationship between distance and the likelihood of engaging in these activities. While the initial positive coefficient for "Distance"

implies that visiting activities become more likely as distance increases, the negative "Distance squared" coefficient indicates that this trend reverses beyond a certain point. Essentially, there is a turning point at which the increasing distance starts to discourage visiting activities, suggesting that extremely long distances eventually become a barrier to visiting. The calculated turning point for visiting activities is approximately 39.3 km. Conversely, the positive coefficients for "Distance squared" with activities like "Playing Sports with alters" and "Engaging in hobbies with alters" suggest a different dynamic (an initial positive influence followed by a negative influence, as demonstrated in Figure 3.1). The calculated turning points for sports and hobby activities are approximately 64 km and 78.5 km, respectively. This could indicate that at moderate distances, the challenges of meeting for these activities are outweighed by other factors, perhaps including the value placed on such activities or the organization of special events that justify travelling the longer distances. This, however, conflicts with the literature. Notably, the model indicates that distance does not significantly affect the utility of excursions with alters, which could be anticipated given the inherent travel component of such activities.

In Figure 3.1, we present simulations that depict the impact of distance on the utility of engaging in various activities with alters, for distances ranging from zero to 100 km. These simulations incorporate the baseline utility of each activity with alters (reflected in the alternative-specific constants) along with the effects of distance and its squared term. The graphical results affirm our interpretations.

These results are in line with the previous studies in terms of the significance of distance in social interactions, for example, in social activity frequency (Carrasco & Miller, 2009) and communication frequency of different communication modes (Calastri, Hess, Daly, Maness, et al., 2017).

3.3.5 Relationship duration

This variable indicates for how many years the ego and each alter have known each other. A log transformation is applied to the values because the model performs better with a log transformation than without in terms of log-likelihood, AIC and BIC. We have also considered the interaction terms between relationship duration and family members to explore the potential differences in how long-standing family relationships impact social contact choices compared to other types of relationships. However, the inclusion of such interaction terms did not lead to a significant improvement in the model's statistical performance, so we decided to

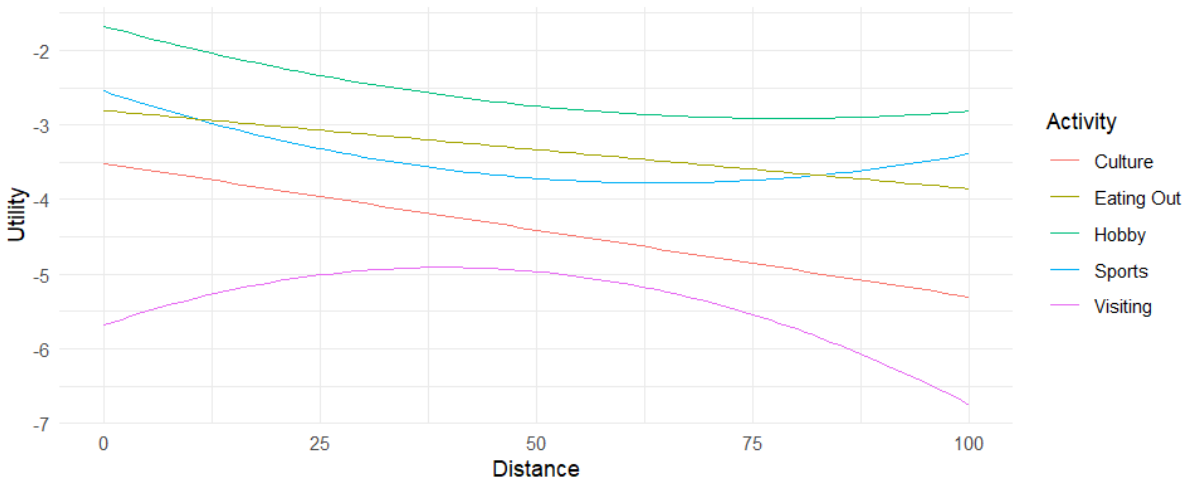

Figure 3.1: Effect of distance on utility.

retain the simpler specification without the interaction term.

The significant and negative coefficients for "relationship duration (number of years)" in the context of "cultural activities with alters" and "engaging in hobby activities with alters" suggest that the longer the ego has known the alter, the less likely they are to engage in these activities together, as reflected in Figure. 3.2, where we represent the effect of relationship duration on utility given the estimated parameters, analogously to the relationship with distance in Fig. 3.1, though engaging in hobby activities with alters is preferred over cultural activity with alters for longer-term relationships. This trend is particularly significant in light of the negative ASC, which generally indicates a tendency towards not choosing any activity over non-engagement. The findings imply that, as the duration of the relationship between the ego and the alter increases, the probability of choosing to participate in cultural and hobby activities diminishes. This could reflect a shifting dynamic in long-standing relationships, where the preference for engaging in cultural and hobby activities together wanes over time. Consequently, the relative undesirability of these types of activities becomes more pronounced as the relationship between the ego and the alter grows older. This is also in line with the general finding in the literature that face-to-face time interaction, in general, reduces as the ego knows their alter longer (van den Berg et al., 2012b; Kowald, 2013; Frei & Ohnmacht, 2016; Calastri, Hess, Daly, Maness, et al., 2017) indeed

engaging in various leisure activities with the alters is a kind of face-to-face communication mode. Similar to the interpretation by Calastri, Hess, Daly, Maness, et al. (2017) and her colleagues for the negative coefficient for face-to-face communication frequency, it is likely that the longer an ego knows an alter, the more physically separated they are making engaging in leisure activities with alters logistically challenging.

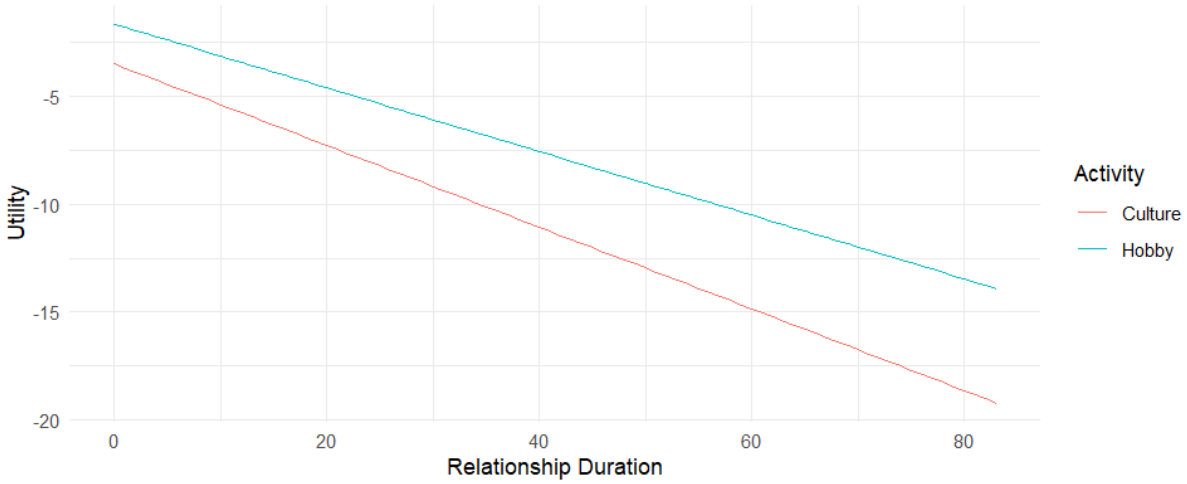

Figure 3.2: Effect of relationship duration on utility.

3.3.6 Age homophily

This study categorises pairs as either "age homophily" or "non-age homophily", as shown in Table 2.3, with the latter serving as the base category. The decision to model age homophily as a dummy variable was influenced by related literature, where this is one of the approaches that has been adopted. However, another approach that includes age difference as a continuous variable could be employed instead, which could provide better model performance and interpretability (e.g., this will help address potential boundary issues that arise with a fixed threshold, where small differences (e.g., 11 years) might still represent age similarity but fall outside the defined boundary). However, the performance of the two models is very similar, with only a minor difference in log-likelihood and AIC. This suggests that neither model is significantly outperforming the other. The former approach is chosen in this work for easier comparison with some of the literature on activity-travel behaviour.

The model reveals insights about the role of age homophily in the selection of activities with alters. Age homophily has positive coefficients for alternatives like "Eating Out with alters", "Playing Sport with alters", and "Going on an Excursion with alters", indicating that age similarity between the ego and alter increases the relative appeal of these activities, even though the general inclination might still

be against engaging in any activity with any alter. The magnitude of these coefficients suggests a hierarchy of preferences, with excursions being the most preferred, followed by eating out and sports, within age-similar dyads. Conversely, the absence of a significant coefficient for activities like "Visiting", "Culture", or "Hobby" suggests that age similarity does not significantly influence the choice of these activities with alters. This differential impact of age homophily across activities provides insights into social preferences.

The probabilities table (Table 3.7) further supports these findings. It shows a higher probability of choosing activities like "Excursions" (0.015 vs 0.009), "Sports" (0.064 vs 0.034), and "Eating Out" (0.043 vs 0.024) for age-homophily pairs compared to non-age-similar pairs. This empirical evidence aligns with the model's coefficients, underscoring the significance of age homophily in social activity choices with alters.

In light of the detailed age homophily data presented in Table 3.8, our findings gain additional depth and specificity. Notably, the top three counts of ego-alter pairs within age homophilous pairs offer interesting insights. The number of ego-alter pairs in the 25-39 age group aligns with the observed trend of age homophily being a significant factor in the preference for "Playing Sport with Alters". This is consistent with the general understanding that younger individuals are more inclined towards physical activities, thus reflecting a higher propensity for engaging in sports within this age group. The prevalence of the 40-59 age group ego-alter pairs may explain the significant coefficient of age homophily for the preference of "Engaging in Excursions with Alters". This demographic, often characterized by greater financial stability and a desire for leisure activities, may prefer excursions as a means of social engagement, reinforcing the influence of age homophily in this choice. The prominence of the 60-79 age group pairs within ego-alter pairs with similar age supports the finding of a significant parameter of age homophily for "Eating Out with Alters". This preference could be attributed to retired individuals having more leisure time and a tendency to socialize in more relaxed and accessible settings, such as dining out.

3.3.7 Both married

There are two dummy variables that report the marital status of ego and alter. Respondents and their social contacts could be married or not married. We included an interaction term between these two variables, to consider ego and alters who are both married (i.e, consider the marital status of both members of the ego-alter

Table 3.8: Age Homophily by Age Group

Ego Age Group	**Alter Age Group**	**N**	**%**
40-59	40-59	4780	38.34
60-79	60-79	1340	10.75
25-39	25-39	1072	8.60
40-59	25-39	470	3.77
40-59	60-79	510	4.09
60-79	40-59	376	3.02
25-39	40-59	356	2.86
18-24	18-24	262	2.10
25-39	18-24	88	0.71
18-24	25-39	77	0.62
60-79	80+	49	0.39
80+	60-79	27	0.22
80+	80+	19	0.15

pair, but it does not imply that the ego and alter are married to each other, see Table 2.4). In this case, only the "cultural activities with alters" coefficient is significant and has a positive sign. The positive coefficient for the variable "both married" in the context of the "Engaging in cultural activities with alters" alternative signifies that when both the ego and the alter are married, there is an increased likelihood of participating in cultural activities together compared to dyads where this is not the case. This suggests that a shared married status enhances the attractiveness of cultural engagements amongst pairs. Notably, this tendency is evident despite a negative ASC, which indicates a general trend towards not engaging in activities. Thus, the condition of both ego and alter being married appears to counteract the overall predisposition against activity engagement, at least in the realm of cultural activities (as it is not significant in other alternatives).

The probabilities table (Table 3.7) further supports these findings. It shows a higher probability of choosing cultural activities for pairs who are both married compared to non-married pairs. This empirical evidence aligns with the model's coefficient, underscoring the significance of pairs being both married in social activity choices with alters.

3.3.8 Sex homophily

There are two dummy variables that report the gender of the ego and alter (male or female). We included two interaction terms between these two variables to consider ego and alters who are both male and both female (i.e., consider the gender of both members of the ego-alter pair, see Table 2.5).

The positive coefficients for "Ego and alter both male" in relation to alternatives like "Playing Sport with alters" and "Engaging in a hobby with alters" indicate that when both the ego and the alter are male, there is an increased likelihood of engaging in these activities relative to all other gender pairings (except for the case of visiting alters, this is relative to ego and alters of different gender). Despite the negative ASC, which implies a general disinclination towards activity engagement, male-male dyads show a particular propensity for sports and hobbies. A potential explanation could be that when a male ego considers playing sports, an alter being male is naturally more appealing simply due to the level of physicality and competition a male aler could offer as opposed to a female alter. Similarly, a male-male pair might be more likely to have similar interests, e.g., shooting, as opposed to other pairings. Conversely, the negative coefficient for "Visiting with alters" suggests that male-male pairs are less likely to choose vis-

iting activities compared to other gender combinations. This denotes a specific aversion towards visiting activities within male-male pairings, further accentuating the general non-engagement trend as captured by the negative ASC.

The positive coefficient associated with the "Ego and alter both female" for the "Visiting alters" alternative suggests that when the ego and alter are both female, there is a higher likelihood of choosing to engage in visiting activities compared to pairs of mixed genders. This increased propensity is observed even in the context of an overall lesser inclination to engage in any activities. Therefore, the data imply a specific preference for visiting activities among female-female dyads, which stands out against the general trend of non-engagement.

The probabilities table (Table 3.7) further supports these findings. It shows a higher probability of choosing visiting activities for pairs who are both female compared to pairs both male, a higher probability of choosing sports and hobby activities for pairs who are both male compared to pairs both female. This empirical evidence aligns with the model's coefficients, underscoring the significance of sex homophily in social activity choices with alters.

3.3.9 Type of relationship

The "Type of Relationship" variable captures the essence of the ego-alter connection, distinguishing among various relational ties as depicted in Table **??**. Friendship serves as the reference category, offering a baseline against which other relationships are measured.

The model outputs reveal that non-friendship ties—spanning familial to acquaintance levels—are consistently associated with lower propensities to engage in shared activities like "Eating Out with alters", "Playing Sports with alters", "Engaging in cultural activities with alters" and "Engaging in hobbies with alters" (further accentuating the general non-engagement trend as captured by the negative ASC). This trend suggests that activities typically associated with voluntary social leisure are less preferred when the alter is not a friend. The negative coefficients for familial and acquaintance relationships across these activities indicate a nuanced preference structure. For example, the hierarchy of least preferred activities for marital family ties—starting with cultural activities and extending to hobby activities—may reflect the ingrained social rituals or perceived appropriateness of engaging in certain activities within specific relational contexts.

The pronounced aversion across all relationship types, apart from friendship, to participating in the listed activities may reflect the social dynamics where friend-

ship is the primary gateway to leisurely social interaction. The differential magnitude of the coefficients across relationship types could be hinting at varying degrees of relational distance and the social obligations that come with them. For instance, familial relationships may entail a sense of obligation that potentially dampens the appeal of more casual or recreational activities that are otherwise freely chosen among friends.

The distinct patterns observed underscore the intricate interplay between social ties and activity choices, suggesting that the nature of the relationship significantly sways the decision to engage in certain activities with alters. It also points to the potential role of cultural and social expectations in shaping these patterns, especially in the Swiss context where private leisure time is highly valued and potentially reserved for certain types of relationships.

Our findings are in line with the existing studies in that they often would find positive coefficients in reference to friends (Carrasco & Miller, 2009; van den Berg et al., 2012b), which is also our case here.

3.3.10 Missing values analysis

In our study, missing values in the explanatory variables were modelled as a separate category of explanatory variables. This approach allows us to handle the non-random nature of missing data and incorporate it into the model without losing significant portions of observations. This methodology is similar to the approach taken by Calastri, Hess, Daly, Maness, et al. (2017) in their study of social network interactions, where they treated missing values as a separate category to understand underlying patterns and avoid bias from data elimination.

In our model, missing values, particularly for distance and relationship duration, reveal interesting patterns. For instances where the ego or the alter's home location is not reported, significant negative coefficients are noted in activities like engaging in culture, sports, and hobbies with alters. This pattern suggests that the absence of location data, likely due to unknown addresses, is associated with a lower likelihood of participation in these activities with alters. The tendency for missing values predominantly at the alter-level supports the hypothesis that unrecognized distances serve as a barrier to engagement. This aligns with the broader understanding that physical proximity is a key facilitator of social interactions.

The missing values in the context of relationship duration present a similarly insightful narrative. The observed coefficients mirror those where the duration is known but with higher magnitudes in the case of missing data. This could

be indicative of long-standing relationships where the ego might struggle to recall the exact duration of the relationship, a situation perhaps more common in well-established social ties. Such ties, characterized by familiarity and longevity, could naturally influence the utility derived from engaging in hobbies with alters, justifying the parallel effect on utility in both known and unknown cases.

For the type of relationship, significant coefficients in the missing values are evident for hobbies and eating out with alters. This pattern could suggest a nuanced effect where the nature of the relationship, when unspecified, influences the preference for certain activities with the alters. It could be that in cases where the relationship type is unclear or less defined, there is a tendency to lean towards less close activities like playing sports than eating out.

3.4 Conclusion

This study provides a detailed examination of the determinants influencing the selection of social network members for various leisure activities at the ego-alter level. Our findings indicate that age homophily, physical distance, marital status, and the type of relationship significantly influence the likelihood of participating in specific activities with alters.

Our analysis revealed several key insights:

- **Age Homophily**: Individuals are more likely to engage in activities such as sports, eating out, and excursions with alters who are close in age. This suggests that age similarity plays a significant role in social activity choices.

- **Physical Distance**: The likelihood of engaging in leisure activities decreases as the physical distance between ego and alter increases, except for visiting activities, which initially increase with distance but decrease beyond a certain point. This highlights the importance of proximity in facilitating social interactions.

- **Marital Status**: Pairs where both the ego and alter are married are more likely to participate in cultural activities together.

- **Type of Relationship**: Friends are the most common companions for leisure activities, highlighting the importance of friendship ties over familial or acquaintance relationships in these contexts.

Snowball sampling, while effective for studying network structures, inherently includes biases such as homophily and degree bias. Homophily bias occurs because respondents tend to recruit others who are similar to themselves, leading to a more homogeneous sample. Degree bias arises because individuals with larger social networks are more likely to be included multiple times, which can overrepresent well-connected individuals.

To address these biases, several measures were implemented:

1. **Initial Iteration Representativeness**: The initial sample was carefully selected to ensure it was representative of the target population. This included using a diverse range of starting points for the snowball chains to capture a wide variety of social network structures.

2. **Balancing Homophily Bias**: The study employed two consecutive subsamples to balance homophily bias. This approach ensured that the sample included individuals from various social backgrounds and network sizes.

3. **Data Quality and Imputation**: The study used homophily-based imputation strategies to handle missing data and reduce biases. This method leveraged the observed homophily patterns to accurately impute missing values, thereby improving the overall data quality.

Despite these efforts, it is acknowledged that snowball sampling does not aim for full representativeness. Instead, it focuses on understanding network structures. The degree bias and homophily are part of the tool's inherent characteristics, and while they can be mitigated, they cannot be entirely eliminated. However, the study's exploratory nature and the measures taken ensure that the findings provide valuable insights into the dynamics of leisure activities within social networks. More details on these methodologies can be found in Kowald (2013).

Future research could address these limitations by incorporating data from diverse geographic and cultural settings. Previous research has successfully examined distance patterns in personal networks across multiple countries, including Canada, Switzerland, the Netherlands, and Chile (Kowald et al., 2013), demonstrating that such comparative studies can yield valuable insights. Exploring more complex models, such as nested logit or mixed logit models, could provide a deeper understanding of the interplay between multiple concurrent activities.

Improved "with whom" models can enhance the explanatory power of joint models, such as the trivariate econometric model and the MDCEV (Multiple Discrete-Continuous Extreme Value) model, by better accounting for continuous variables like activity duration alongside social interactions.

Chapter 4

Machine learning assisted choice model specification in a context of social contact selection in leisure activities

4.1 Introduction

Traditional choice model specification involves a base model, which is then systematically refined by adding and combining variables (e.g., the interaction term between two variables). This process is driven by statistical significance, intuition, insights from prior research and the reasonableness of the findings. However, it is generally impractical and often practically impossible to try all specifications, even with only a few explanatory variables, let alone the scenarios where the data consists of a large set of explanatory variables.

Machine learning methods, on the other hand, excel at identifying hidden correlations within large datasets (Hillel et al., 2021; van Cranenburgh et al., 2022; Wang et al., 2021). Previous studies have attempted to help the choice modellers in specification searching by framing this as an optimisation problem solved through iterative search algorithms (Ortelli et al., 2021).

Thus, some authors have suggested using feature importance ranking of the machine learning model (Hillel et al., 2019) to help the modellers in selecting the explanatory variables (even before looking into how this variable should be specified in the model), which gives an additional aspect of considerations in developing the choice model alongside other aspects aforementioned.

Some other authors take another step back and suggested using feature importance ranking of the machine learning model (Hillel et al., 2019) to help the

modellers in selecting the explanatory variables (even before looking into how this variable should be specified in the model), which gives an additional aspect of considerations in developing the choice model alongside other aspects aforementioned. Feature importance ranking in machine learning is a technique used to identify and rank the relative importance of various features (or variables) in a dataset with respect to the predictive power they hold in a model. This ranking is based on the principle that not all features contribute equally to a prediction model's accuracy. Some features might strongly influence the outcome (e.g., distance might be a strong predictor in an activity-travel-related model), while others may have little to no impact. The process of feature importance ranking involves using statistical techniques or algorithms to quantify the extent to which each feature contributes to the model's performance.

Our study introduces the assisted specification approach to the domain of social network analysis, a context distinct from its traditional applications in transportation, such as mode choice studies (Ali et al., 2023). Unlike transport contexts, where choices are often explained by factors such as time, cost, convenience and etc., social network analysis look into the complexities of social relationships and their multifaceted impacts on behavior. This domain's inherent complexity presents unique challenges and opportunities for methodological innovation. The traditional choice modelling approach may not fully capture the subtleties of these social influences. By applying feature importance ranking within this novel context, our research aims to uncover hidden patterns and correlations that are uniquely present in social networks but might be overlooked by the conventional approach. Our investigation responds to calls for further testing and case studies of the assisted specification approach (Hillel et al., 2019) by demonstrating its applicability in social network analysis, aiming to contribute to broader methodological exposure, highlighting the value of integrating machine learning techniques with traditional choice modelling to enhance the interpretability of behavioural models.

In contrast to Hillel et al. (2019), our study will employ SHapley Additive exPlanations (SHAP) for feature importance ranking (Lundberg & Lee, 2017). This technique offers an alternative theoretical perspective based on game theory, as opposed to information theory. We would attempt to see if what we perform here with machine learning can be a valuable complement to choice modelling. For example, improving the variable selection process in choice modelling and unearthing hidden correlations could provide new insights into activity-travel behaviour. Ultimately, this chapter would serve as further evidence for bridging choice modelling and machine learning.

4.2 Method

In supervised learning—a foundational method of machine learning described comprehensively in Hastie et al. (2009)—models are trained using labelled data to predict outcomes based on input features. In this research, the objective is to classify each ego-alter pair into one of several activity categories, showcasing a typical application of supervised learning where the model learns to map inputs (e.g., dyadic variables) to outputs (activity types). Our method incorporates a two-step approach. Initially, we implement and compare various machine learning models, such as tree-based models and neural networks, to identify the model that performs best. Once the best-performing model is selected, we apply SHAP to this model to obtain detailed insights from the model, such as the contribution of each feature toward the prediction of the outcome.

In our context of imbalanced datasets, where one class significantly outweighs the others (i.e., one label/alternative is chosen around 80% of the time), relying solely on overall accuracy as a performance metric can be misleading. This is because a model that always predicts the majority class will achieve high accuracy but will not necessarily be effective in identifying instances of the minority class. Therefore, a combination of the metrics will be considered to gain a comprehensive understanding of a model's performance.

The key metrics used to evaluate our models are:

1. Accuracy: The ratio of correctly predicted instances to the total instances. While high accuracy might seem desirable, in the context of imbalanced data, it can be misleading because it does not account for the model's ability to predict minority class instances.

$$\text{Accuracy} = \frac{\text{True Positives (TP)} + \text{True Negatives (TN)}}{\text{Total Number of Samples}}$$

2. Recall (Sensitivity or True Positive Rate): The ratio of correctly predicted positive observations to all observations in the actual class.

$$\text{Sensitivity} = \frac{\text{True Positives (TP)}}{\text{False Negatives (FN)} + \text{True Positives (TP)}}$$

3. Specificity (True Negative Rate): The ratio of correctly predicted negative observations to all observations in the actual negative class.

$$\text{Specificity} = \frac{\text{True Negatives (TN)}}{\text{False Positives (FP)} + \text{True Negatives (TN)}}$$

4. Balanced Accuracy: The average of recall and specificity, providing a more balanced measure of performance across both classes.

$$\text{Balanced Accuracy} = \frac{\text{Sensitivity + Specificity}}{2}$$

Label Encoding (Integer Encoding) is convenient but inappropriate when dealing with nominal variables, as it imposes an ordinal relationship where none exists. We employed one-hot encoding for nominal variables in our dataset to address this. For instance, the variable "Ego-Alter relation: Is alter a kin-contact?" with levels such as close relative, non-relative, other relative, and missing data was converted into four binary variables (dummy variables), each indicating the presence or absence of a specific category.

4.2.1 ML models

4.2.1.1 Data

Given the imbalanced nature of our dataset, where some classes are significantly underrepresented, traditional data splitting and performance metrics could lead to misleading conclusions. To mitigate this, we adopted a stratified sampling approach to maintain the proportion of each class in both training and testing datasets (85% and 15% respectively). Stratified sampling was chosen to ensure that the class distribution in the training and test sets is consistent with the overall dataset. This approach prevents scenarios where random sampling might result in some subsets having very few or no instances of the minority class, which could lead to biased model training and evaluation. By preserving the class ratios, stratified sampling ensures that the model is exposed to both classes in a balanced way during training and testing. Furthermore, balanced accuracy is used as one of the main evaluation metrics, stratified sampling is a natural complement because it allows for a fair comparison of performance across classes. Balanced accuracy considers both the majority and minority classes equally, and stratified sampling ensures that the minority class is represented in all subsets, thus avoiding misleading results/evaluations caused by highly skewed data distributions. This strategy ensures that our models are trained and validated on a representative sample of the original data, enhancing the reliability of our predictive performance. While other techniques like oversampling or undersampling could also address class imbalance, they were not chosen in this study because the focus was on preserving the

original structure and relationships in the data rather than altering them (oversampling duplicates examples from minority classes, which can lead to overfitting. Undersampling, on the other hand, reduces the number of samples in the majority class, which can cause loss of potentially valuable information).

The decision to adopt an 85-15 split for the training and test datasets was made to ensure that a substantial portion of the data was used for model training, while still maintaining an adequate sample for performance evaluation. Allocating 85% of the data for training aids in capturing sufficient information, which is essential given the complexity of the model and the range of explanatory variables under consideration. The remaining 15% provides a reliable basis for evaluating the model's generalisability and mitigating risks associated with overfitting.

To examine whether the model's performance is sensitive to the specific choice of split ratio, additional experiments were conducted using different proportions (e.g., 70-30, 80-20, and 90-10). The results indicated that the model's performance across different metrics remained relatively stable, with only minor differences observed. For example, the balanced accuracy values for the random forest binary classifier predicting "eating out" with alters were:

- 70-30 split: 0.6977
- 80-20 split: 0.7044
- 85-15 split: 0.6987
- 90-10 split: 0.7261

Although the 90-10 split yielded a slightly higher balanced accuracy, it resulted in a reduced test set size, thereby limiting its reliability for performance evaluation. Conversely, the 70-30 split provided a larger test set but a smaller training set, which could impact the model's capacity to learn from underrepresented classes.

The 85-15 split was selected as it offers a balanced compromise, ensuring that the training set is sufficiently large to capture complex patterns, while the test set is robust enough to evaluate model performance reliably. This choice ensures that class proportions are preserved through the use of stratified sampling, thereby enhancing the reliability of the evaluation.

Moreover, it is worth noting that there is no universally optimal rule for choosing a split ratio. As highlighted by Joseph (2022), the optimal training/testing ratio often varies depending on the specific data characteristics and the model

complexity, and commonly used splits such as 80-20 or 70-30 are largely based on heuristic considerations. Given this, the observed stability of our results across multiple splits indicates that the model's performance is not unduly sensitive to the specific choice of split ratio, and the use of the 85-15 split is justified for the data in this study.

During the training of our machine learning models, we used balanced accuracy as the evaluation metric within a cross-validation framework. This involved splitting the data into multiple folds, training the model on some folds, and validating it on the remaining folds. The balanced accuracy metric was used to assess the model's performance on the validation sets, guiding the selection of optimal model parameters. By averaging the balanced accuracy across all folds, we ensured that our model's performance was robust and not biased towards any particular class. The cross-entropy loss was used to train the models by optimising probabilistic predictions. However, since some of the activity choices are imbalanced, balanced accuracy was used as the evaluation metric during the cross-validation (CV) process to select the best model. This combination allowed us to effectively train the models while ensuring balanced performance across different classes during evaluation.

A brief summary of the ML models used in this chapter is provided. More details can be found in Hastie et al. (2009).

4.2.1.2 Decision Trees

A decision tree is a popular machine learning model used for both classification and regression tasks. It operates by repeatedly splitting the data into distinct sub-groups based on the attributes of the data. Each split aims to maximise the homogeneity of the resultant sub-groups regarding the outcome variable. The quality of these splits is determined using impurity measures such as Gini impurity and entropy: - **Gini Impurity:** $Gini(t) = 1 - \sum_{i=1}^{k} p_i^2$ - **Entropy:** $Entropy(t) = -\sum_{i=1}^{k} p_i \log_2(p_i)$ where p_i represents the proportion of class i samples at node t. We used Gini Impurity.

The decision tree begins with the root node, which contains the entire dataset. This node represents the initial state before any splits have been made.

As we move down the tree, the dataset is split at several internal nodes. Each node represents a decision point that bifurcates the data based on a specific condition or attribute. For example: An internal node may pose a question such as "Is the number of edges greater than 116?" Depending on the answer, the dataset is divided into two paths: one for data points that satisfy the condition (right branch)

and another for those that do not (left branch). The branches culminate in leaf nodes, where no further splitting occurs. Each leaf node provides a prediction or decision based on the path followed from the root. Typically, a leaf node will specify:

- The predicted class (e.g., plays sports together: yes or no)
- A measure of confidence or error rate associated with the prediction
- The proportion of the dataset that reaches this leaf

The decision-making process in a decision tree is analogous to following a flowchart where each decision leads down a path to a final outcome. Starting from the root, one evaluates the attributes of the case in question and follows the path dictated by the conditions at each node until reaching a leaf. The prediction at the leaf node is then used as the outcome for that case.

In developing the decision tree model for predicting whether individuals are likely to engage in sports together (and so on for the other types of activity), it was crucial to adjust the model's complexity to optimise both its accuracy and ability to generalise. This optimisation process was conducted using the `caret` package in `R`, which supports efficient parameter tuning via cross-validation and allows for adjustments to the tree complexity. The model was trained using a 5-fold cross-validation approach. This technique involves splitting the dataset into five equal parts, each part being used once as a validation set while the remaining data serve as training sets. Such a strategy ensures comprehensive model evaluation and helps in avoiding overfitting.

The complexity parameter (`cp`) in the decision trees serves as a control for the minimum required improvement in model fit at each node. By experimenting with different `cp` values, the model can be adjusted to balance between being overly complex and overly simplified:

- Parameter Range: Multiple `cp` values were tested to identify the best `cp` value that allows the model to maintain an appropriate level of complexity without sacrificing predictive accuracy. The criterion for selecting the best `cp` value was based on achieving the highest balanced accuracy across the cross-validated datasets.
- Evaluation Metric: The primary metric for model evaluation was balanced accuracy, which provides a robust measure for the effectiveness of the model in classifying the two outcomes.

With the optimal complexity parameter identified, the decision tree model is finely tuned to predict sports participation (and other types of activities) among ego-alter pairs.

4.2.1.3 Random Forests

Random Forests build on the foundation of decision trees, enhancing the model's predictive power and robustness by incorporating ensemble learning principles. This method involves deploying multiple decision trees, each trained on a distinct bootstrap sample of the data. A bootstrap sample is a random subset drawn with replacement, meaning some observations may appear more than once, while others may not appear at all. This ensures variability among the trees, which is crucial for reducing overfitting.

Each tree in a Random Forest is constructed by selecting a random subset of features at each decision split—this approach is termed 'feature bagging'. By not always using all features at each split, it decreases the correlation between individual trees, significantly enhancing the model's ability to generalise and further reducing the variance of the model without significantly increasing the bias.

The predictive performance of Random Forests primarily hinges on reducing variance while retaining low bias. Each tree aims to minimise either the Gini Impurity or Entropy at each split (we used the former).

Random Forests employ bagging (bootstrap aggregating) to stabilise the variance. In classification, it uses a majority voting system among the N trees. Each tree votes for a class and the class with the majority of votes becomes the model's prediction.

In our study, each Random Forest model was trained using a 5-fold cross-validation technique, facilitated by the `caret` package, which ensured a comprehensive validation approach. This package also handled the tuning of key parameters such as the number of trees (`ntree`) and the number of features at each split (`mtry`). These parameters are critical for optimising the Random Forest model to achieve the best trade-off between bias and variance.

The model's effectiveness was evaluated using the balanced accuracy metric.

Random forests typically achieve higher accuracy and better generalisation capabilities through these ensemble techniques than individual decision trees, particularly on complex datasets with interactions among attributes. The ensemble nature of Random Forests helps manage larger datasets and effectively handles unbalanced data by balancing error rates across different classes.

4.2.1.4 Boosting

Building further on the principles of ensemble learning, Boosting represents a different methodology aimed at creating a strong predictive model by combining multiple weak models, typically decision trees. Unlike Random Forests, which build trees independently and combine them through averaging or majority voting, Boosting involves adding trees sequentially to correct the errors made by previous trees.

XGBoost (eXtreme Gradient Boosting) stands as one of the most efficient and powerful implementations of the gradient boosting framework. It enhances the boosting technique with both speed and performance optimisations, making it highly effective for a wide range of data science applications and competitions.

XGBoost improves on the traditional boosting methods by using a more regularised model formalisation to control over-fitting, which gives it better performance. At its core, XGBoost utilises the gradient boosting algorithm, where new models are created that predict the residuals or errors of prior models and then added together to make the final prediction. This is often done through a technique called gradient descent to minimise a loss function. Unlike other boosting techniques that grow trees greedily, XGBoost uses a depth-first approach where the growth of a tree is halted early if it doesn't lead to a minimum reduction in the loss function. This results in a more optimal and generalised tree. The loss function used here is logistic loss. Additionally, XGBoost automatically handles missing values and supports sparse data input, making it versatile for various datasets.

In the application to our dataset for predicting sports participation among ego-alter pairs, XGBoost was configured to optimise the binary logistic objective, which is essential for our binary classification problem. Using a robust cross-validation framework, the model parameters were finely tuned, focusing on key parameters like the number of gradient-boosted trees (`n rounds`), tree complexity (`max depth`), and the learning rate (η).

The XGBoost model was trained using the same 5-fold cross-validation as previous models to align with our methodological rigour. This ensured a comprehensive assessment and validation against overfitting, maintaining consistency with our analytical approach.

The effectiveness of the XGBoost model was evaluated using the balanced accuracy metric, allowing for a direct comparison with both the decision tree and Random Forest models. Given the sequential improvement nature of Boosting, XGBoost was expected to show an increment in the balanced accuracy metric,

demonstrating its superior ability to accurately classify positive and negative outcomes.

4.2.1.5 Support Vector Machine

Support Vector Machines (SVM) introduce yet another powerful approach, primarily distinguished by its basis in statistical learning theory. Unlike decision trees that partition the data space into segments using lines or curves, SVMs construct a hyperplane or set of hyperplanes in a high-dimensional space, which can be used for classification, regression, or other tasks. The effectiveness of SVMs lies in their ability to find the maximum margin separator between classes, making them exceptionally good at recognising subtle patterns in complex datasets.

The core idea behind SVM is to find the hyperplane that best divides a dataset into two classes, where the margin of separation between the classes is maximised. This is particularly powerful for linearly separable data but extends to non-linear boundaries using the kernel trick. By applying different kernel functions, such as polynomial, radial basis function (RBF), or sigmoid, SVMs can perform non-linear classification, subtly capturing complex relationships between features without necessitating transformations on the raw data itself.

Incorporating SVMs into our study, we employed the `svmRadial` method, which utilises an RBF kernel, known for its flexibility and suitability for various data distributions. The training was conducted within the `caret` framework, allowing us to leverage its robust cross-validation capabilities to fine-tune the model. This involved adjusting parameters like the cost of constraints violation (C) and the γ parameter, which defines the reach of a single training example. Cost (C) determines the trade-off between achieving a low error on the training data and minimising the model complexity for better generalisation. A higher C attempts to correctly classify all training examples by giving the model freedom to select more samples as support vectors. γ defines how far the influence of a single training example reaches, with low values meaning 'far' and high values meaning 'close'. The right gamma parameters can affect the smoothness of the decision boundary.

The performance of the SVM model was rigorously evaluated using the balanced accuracy metric, ensuring consistency with the evaluation criteria applied to our previous models. This metric was particularly pertinent to our SVM analysis due to its sensitivity in distinguishing between the classifier's ability to maintain high true positive rates while minimising false positives.

Using a 5-fold cross-validation approach, similar to that used for both the Ran-

dom Forest and XGBoost models, ensured that the SVM's evaluation was comprehensive and robust, effectively demonstrating its capacity to generalise across unseen data.

4.2.1.6 Neural Networks

Neural Networks offer a sophisticated computational approach that mimics the human brain's operations to recognise patterns and solve complex problems. Unlike the previously discussed models that emphasise individual learners or hyperplanes, Neural Networks consist of layers of interconnected nodes or 'neurons' that work collaboratively to make decisions. This model excels in environments where relationships between inputs and outputs are non-linear and intricate.

Neural Networks are structured in layers: an input layer that receives the data, one or more hidden layers that compute the inputs, and an output layer that produces the final decision. Each neuron in these layers is connected with adjustable weights, which are tuned during the training process to minimise the prediction error. The power of Neural Networks lies in their ability to learn these weights through backpropagation, effectively adjusting them based on the gradient of the error with respect to the network's configuration.

For our study on predicting activity participation among ego-alter pairs, we utilised a standard feedforward Neural Network with multiple hidden layers, trained using the backpropagation algorithm. The architecture included:

- Input Layer: Matching the number of features in the dataset.
- Hidden Layers: Two hidden layers with 3 and 5 neurons respectively.
- Output Layer: A single neuron output layer to classify the binary outcome of activity participation.

The decision to use two hidden layers with 3 and 5 neurons, respectively, was driven by the necessity to construct a model that balances complexity with performance and generalisation capabilities. The initially larger networks led to practical issues such as excessive model weights and potential overfitting. The network size is then systematically reduced, which indicated that a simpler model was more appropriate for the dataset at hand. This approach is consistent with best practices in machine learning, where model complexity should not exceed what is supported by the data.

Neural Networks compute the output from each neuron in the hidden layers and the output layer using an activation function. For our binary classification

task, the sigmoid function is used at the output layer to squash the output between 0 and 1, representing the probability of class 1:

$$\sigma(z) = \frac{1}{1+e^{-z}}$$

During training, we use the binary cross-entropy loss function, which measures the performance of our classification model. The loss function is defined as:

$$L(y,\hat{y}) = -\left[y\log(\hat{y}) + (1-y)\log(1-\hat{y})\right]$$

Backpropagation is used to update the weights in the network. The weights are adjusted according to the gradient of the loss function with respect to each weight:

$$w_{ij}^{(new)} = w_{ij}^{(old)} - \eta \frac{\partial L}{\partial w_{ij}}$$

This ensures that the model learns to minimize the loss function, improving its accuracy over time. The learning rate η controls the step size during the weight updating phase.

In addition, feature scaling through standardisation was applied to ensure that all input variables have a mean of zero and a standard deviation of one, making them comparable in scale. The training process involved tuning the network with a grid specifying the size and decay parameters. The network was optimised for the balanced accuracy metric, with parameters controlled for the maximum number of iterations and the maximum number of weights allowed.

The Neural Network's performance was assessed using the balanced accuracy metric, as with the other models, to maintain evaluation consistency across all analytical methods. Given the Neural Network's capacity for modelling complex relationships and its flexibility in layer and neuron configurations, it often excels in capturing subtle nuances in data that other models might miss.

During our initial experiments with neural networks, particularly when exploring configurations suitable for our dataset, we encountered a challenge related to the model's complexity and the data. As the complexity of the neural network architecture increases, so does the number of weights that need to be optimised. This can lead to a scenario known as the 'curse of dimensionality', where the parameter space becomes so large that the available data are sparse, making effective training and generalisation difficult. Given this challenge, we decided against pursuing deeper neural network architectures.

4.2.2 SHAP

SHAP (SHapley Additive exPlanations) is a machine learning explainability framework that offers insights into the contribution of each feature to a model's prediction at the individual data point level. Derived from cooperative game theory, SHAP values quantify the impact of each feature by considering all possible combinations of features in a model.

SHAP is rooted in the Shapley value concept from cooperative game theory, which allocates payouts to players based on their contribution to the total payout. In machine learning, "players" are the features used in a model, and the "payout" is the prediction output. The Shapley value provides a fair distribution of the "payout" among the features, considering the contribution of each feature to the prediction in all possible combinations of features.

SHAP assigns each feature an importance value for a particular prediction by defining an explanation model as a linear function of binary variables. The general form of the explanation model g is:

$$g(z') = \phi_0 + \sum_{i=1}^{M} \phi_i z'_i$$

Where z' is a binary vector representing the presence (1) or absence (0) of each feature, ϕ_i are the feature attributions, and M is the number of features. This framework ensures that the sum of the feature contributions (including the baseline ϕ_0) approximates the model output $f(x)$.

To calculate SHAP values, we use the concept of marginal contributions, which involves computing the change in the expected model prediction when conditioning on each feature. The SHAP value for a feature i is given by:

$$\phi_i(f,x) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} [f(S \cup \{i\}) - f(S)]$$

Where:

- N is the set of all features,
- S is a subset of features excluding i,
- $|S|$ is the size of the subset S,
- $f(S)$ is the model prediction with the features in subset S,

- $f(S \cup \{i\})$ is the model prediction with the features in subset S plus feature i.

Using SHAP for feature importance provides three key outputs:

- Global Feature Importance: This shows the overall importance of each feature across all predictions, akin to what information gain would provide. It will show us, say, the type of relationship is generally the most important feature in classifying ego-alter pairs into different types of activities. However, it does not indicate how different values of the type of relationship affect the classification into each category, nor does it give insights into individual ego-alter pair classifications, which is why we suggested SHAP as an alternative

- Feature Influence: SHAP values illustrate how each feature value impacts the prediction for a specific data point. For instance, an alter being a friend might increase the SHAP value for playing sport (positive impact) but decrease it for visiting (negative impact).

- Local Interpretation: Consider two ego-alter pairs with the following characteristics:

  - **Ego-alter pair A:**
    - Alter homophily
    - Alter is a friend
    - Lives 50 KM away from the ego
  - For ego-alter pair A, SHAP might reveal that the alter being the same age contributes to it being classified as playing sport, while the alter being a friend also contributes but to a lesser extent.
  - **Ego-alter pair B:**
    - Not age homophily
    - Alter is a relative
    - Lives 10 KM away from the ego
  - For ego-alter pair B, the SHAP values might indicate that the ego-alter pairs of different ages/generations strongly influence its hobby classification, and the short distance reinforces this classification.

The analysis would help give insights in terms of expectation of statistical significance (global feature importance), signs of the parameters (feature influence) and behavioural interpretation (local interpretation, e.g., preview of whether it is reasonable or intuitive or in line with literature even before running a choice model).

Eventually, we would want to see what choice specification we ended up with. Is it the same as the one we obtained in the previous chapter? If it is, this machine learning-assisted choice model specification further confirms/validates our previous model. If not, does it offer additional behavioural insight we previously overlooked?

By integrating these insights into our choice model, we aim to uncover hidden patterns and correlations that are unique to social networks but might be overlooked by conventional approaches. This methodology allows us to enhance the variable selection process, validate our previous models, and provide new behavioural insights.

In conclusion, SHAP values offer a rigorous and theoretically grounded approach to interpreting machine learning models. By quantifying feature contributions in a way that ensures fairness and consistency, SHAP values enhance our understanding of model predictions and support more informed decision-making in choice model specification.

4.3 Results and discussion

4.3.1 Comparison with ML models

The data is an imbalanced data (see Table 4.1)

Category	None	Hobby	Culture	Visiting	Excursions	Eating Out	Sport
Percentage	80.57%	7.26%	1.17%	0.41%	1.3%	3.67%	5.63%

Table 4.1: Percentage of ego-alter pairs for different types of activities

Since accuracy is not the best measure of the performance of a machine learning model for an imbalanced data, we should explore other measures that take into account of this nature of the data and therefore gives an less biased view on the

performance. The performance of a machine learning model can be summarised in a confusion matrix :

TN	FP
FN	TP

Table 4.2: Confusion Matrix

Where TP denotes the number of true positives etc.. So, the top row consists of negative predictions and the left-hand column of actual negative observations. Below is the decision tree for the target variable playing sport with alters, introduced here for illustration of the metrics introduced earlier)

	Reference	
Prediction	0	1
0	1920	114
1	7	16

Table 4.3: Confusion Matrix for decision tree

Confusion Matrix is interpreted as the following:

- True Negatives (TN): 1920 - The model correctly predicted the majority class (0) most of the time.
- False Negatives (FN): 114 - These are instances where the model incorrectly predicted the negative class (0) when the actual class was positive (1).
- True Positives (TP): 16 - The model correctly predicted the minority class (1), which is typically harder in imbalanced datasets.
- False Positives (FP): 7 - Instances where the model incorrectly predicted the positive class.

One common measure is **Accuracy**:

$$\text{Accuracy} = \frac{\text{TP + TN}}{\text{Total Number of Samples}} = \frac{1920+16}{2057} \approx 94.12\%$$

While this might seem high, it's not very informative due to the imbalanced nature of our data.

Another measure is **Sensitivity**:

$$\text{Sensitivity} = \frac{\text{TP}}{\text{FN + TP}} = \frac{16}{114+16} \approx 12.31\%$$

This is relatively low, reflecting the model's limited effectiveness in correctly identifying actual positive cases (class 1).

Specificity:

$$\text{Specificity} = \frac{\text{TN}}{\text{FP + TN}} = \frac{1920}{7+1920} \approx 99.64\%$$

This is very high, indicating the model's effectiveness at identifying negative cases (class 0) as expected.

Balanced Accuracy:

$$\text{Balanced Accuracy} = \frac{\text{Sensitivity + Specificity}}{2} = \frac{0.1231+0.9964}{2} \approx 55.97\%$$

Since this takes into account both sensitivity and specificity, it provides a more realistic picture of model performance, showing that it is not performing well overall.

The holdout validation results for all models are shown in the table below. The GBDT model performs best, achieving the highest balanced accuracy.

Model	**Sports**	**Hobby**	**Visiting**	**Excursions**	**Eating Out**	**Culture**
Decision Tree	0.5361	0.5301	0.9370	0.6197	0.5375	0.5275
Random Forest	0.7270	0.6949	0.8328	0.6399	0.6912	0.7056
GBDT	0.7472	0.7143	0.8333	0.6598	0.6987	0.7629
SVM	0.7047	0.7124	0.7998	0.5677	0.6144	0.6384
NN	0.5970	0.6083	0.7725	0.6633	0.6817	0.5868

Table 4.4: Balanced Accuracy of Different Models

4.3.2 SHAP

One-vs-All (OvA), also known as one-versus-the-rest, is a heuristic strategy employed to adapt binary classification algorithms for multi-class classification tasks (Murphy, 2012). This method entails decomposing the multi-class dataset into several binary classification challenges. For each distinct binary problem, a binary classifier is trained, where the class in focus is labelled as positive, and all other classes are labelled as negative. For example, the choice of an ego to play sports with an alter is positive (or 1) when an ego chooses to do sport with an alter otherwise negative (or 0), this negative means not only no engagement but also other alternatives other than sport. Subsequently, predictions are made based on the model that shows the highest confidence level. In other words, there will be 6 binary classifiers corresponding to each alternative/binary problem, the one with highest probability will be used for the prediction say of a specific ego-alter pair.

Hence, instead of presenting a single feature importance ranking for a multi-class classifier—which could potentially lead to confusion—it is rational to present individual feature importance rankings for each of the six binary classifiers (excluding null class or alternative). This method simplifies the interpretation of feature importance rankings. We have further verified that it gives no different results in terms of the feature importance ranking with either approach practically with our dataset. That is when comparing the feature importance ranking of the multi-class classifier with the binary classifiers, feature rankings agree.

4.3.2.1 Feature importance rankings

The results in Figure 4.1 to 4.6 can be understood in the following way:

- Variable Names/Feature: The y-axis lists the variable names, these represent the different features used in the model.

- Importance Scores: The x-axis shows the importance scores derived from SHAP values. Higher scores indicate that the feature has a greater impact on model predictions.

- Ranking of Features: The length of each bar represents the magnitude of the feature's importance. The features are ranked from top to bottom, with the top feature ('Centralization in network') having the highest importance in Figure 4.1.

In SHAP analysis, the importance of a feature is measured by the average magnitude of the SHAP value across all samples. A feature with a high SHAP value would typically have a stronger impact on the model's prediction, either pushing the prediction higher or lower.

From Figure 4.1, we can see the following: "Centralization in-network", "share of alters with same-sex as ego", and "No. of people in household", these features are the most important, indicating they have the most significant influence on the decision of engaging in cultural activities with alter. The features "context of meeting: club" also seem to have a substantial impact on predictions, though less than the top three features. This implies that the context of meetings plays an important role in the model's decision-making. The features like "network density", "network degree" etc. suggest that network-related variables are also influential in the model, although to a lesser extent than the top features. Practitioners could look at such a plot to determine which features they might want to examine further to understand behaviour.

Furthermore, by comparing the importance of features across different classes, we can deduce which factors are universally important and which are specific to certain activities. For example, the type of relationship (i.e., non-realtive) is important across many classes; it suggests that the fundamental relationship between ego and alter is a key determinant in the decision to engage in any activity with alters.

As a result of the SHAP investigation:

- Several network features have been suggested across the feature importance rankings of classifiers, including degree centralization, network size, number of educational places an ego had in the course of life, network density, network degree, number of edges in the network, mean participant of activity groups, components in the network, number of weak ties.

- Figure 4.1 (culture) suggests Share of alters with same-sex, Share of alters with the same education, and No. of people in the household

- Figure 4.2 (eating out) suggests a different way of classifying the type of relationships for "better" model performance, i.e., instead of the type of relationship incorporated into the choice model kin classification (in Chapter 2 we employed the following classification: acquaintance/friend/married into family/relative/relative 1. degree/spouse; But feature importance seems to suggest a different classification: close relative/non-relative/other relative).

The implication here is that a decision should be made on which classification of relationship should be used considering both the performance and behavioural interpretations.

- Figure 4.3 (Excursions) suggests the Share of alters with the same civil status.
- Figure 4.4 (Hobby) suggests Alter not a relative and Share of alters with the same education
- Figure 4.5 (Sport) suggest Alter not a relative
- Figure 4.6 (Visiting) suggest Share of alters with the same education and No. of the main residency an ego has.

A key insight is that such analysis helps modeller in deciding which definition of a variable to use when there are multiple definitions possible.

We then applied what we had learned from these results to inform our new choice model specification, using the model developed in Chapter 3 as the baseline model (i.e., Eq. 3.4; Specification 3).

The network features, or variables on the ego-network level that we have identified, have not been included in the choice model in the previous chapter. However, we are cautious about including them in our choice model because of the concern of endogeneity (see 3.3.2). The model remains the same.

From Figure 4.1 (culture): Share of alters with same-sex and No. of people in the household, both have significant and positive coefficients, which would have been included in the model anyway. The share of alters with the same education is not significant. No updates on the model.

From Figure 4.2 (eating out): Alter not a relative has a significant and positive coefficient, 0.9599***, as opposed to relatives and close relatives. Missing values have a significant but negative coefficient, -2.2797***. The model is updated by removing married into family, 1st degree relative, relative and acquaintance, replaced by non-relative and missing value coefficient for the type of relationship for the utility function of this alternative.

From Figure 4.3 (Excursions): Share of alters with the same civil status is tested, but no significant result is found. No update on the model. No further updates on the model following the change in the previous step.

From Figure 4.4 (Hobby): Alter not a relative and missing values both have significant and positive coefficients, 1.7843*** and 1.1077***, respectively, as

opposed to relatives and close relatives. Share of alters with the same education has a significant and positive coefficient of 0.0085*. In the former case, the model is updated by removing married into family, 1st degree relative, relative and spouse, replaced by non-relative and missing value coefficient for the type of relationship for the utility function of this alternative; in the latter case, this is what we would have found anyway.

From Figure 4.5 (Sport): Alter not a relative has a significant and positive coefficient, 1.0951***, as opposed to relatives and close relatives. The model is updated by removing married into family, 1st degree relative and relative, replaced by non-relative for the utility function of this alternative.

From Figure 4.6 (Visiting): Share of alters with the same education and No. of the main residency an ego have was tested, but neither was significant. No further updates on the model following the change in the previous step.

The results from Chapter 3 are displayed in Table 4.5 and the results of the ML-assisted choice model specification are summarised in Table 4.6.

Using the LR statistic to compare the model from Chapter 3 with the ML-assisted model (Model 2), as displayed in Table 4.7, Model 2 is significantly different from Model 1 at well beyond the 95% confidence level. This result confirms that the ML-assisted choice model significantly increases the goodness of fit of the original model. The very low p-value indicates strong statistical evidence against the null hypothesis that the earlier model is sufficient, thereby supporting the more complex model's validity in capturing essential variations in the data that are not accounted for by the previous model.

Practitioners could look at such a plot (i.e, Figure 4.1 to Figure 4.6) to determine which features they might want to examine further to understand behaviour.

Furthermore, by comparing the importance of features across different classes, we can deduce which factors are universally important and which are specific to certain activities. For example, if 'alter relation' is important across many classes, it suggests that the fundamental relationship between ego and alter is a key determinant in the decision to engage in any activity with alters.

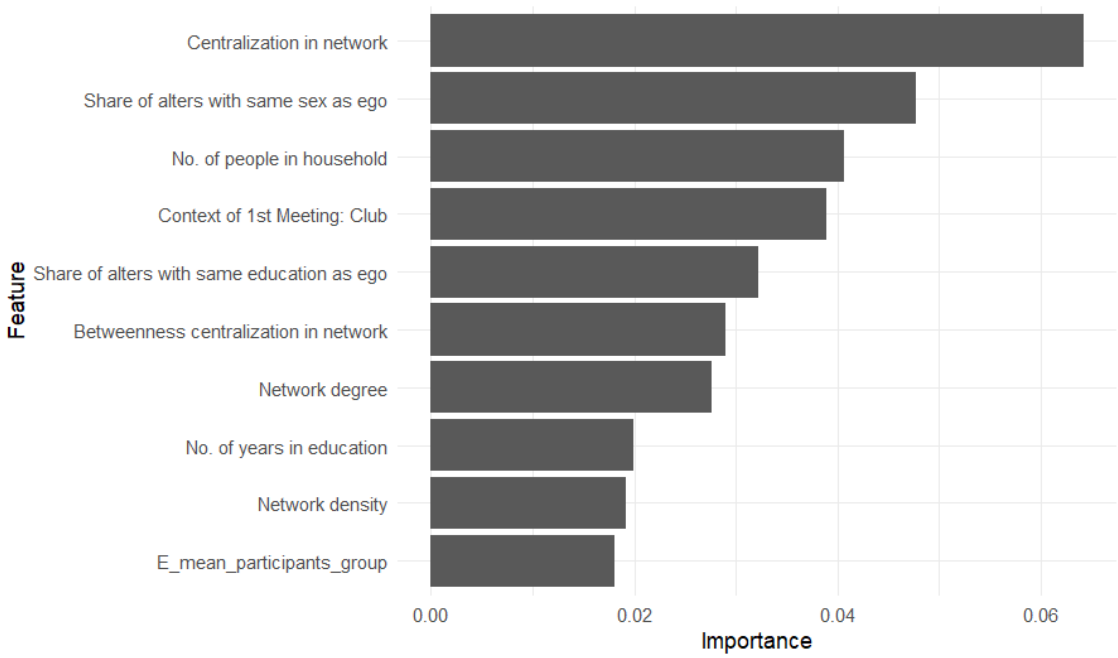

Figure 4.1: feature importance ranking of the binary classifier for the target variable ego engaging in cultural activities with alters

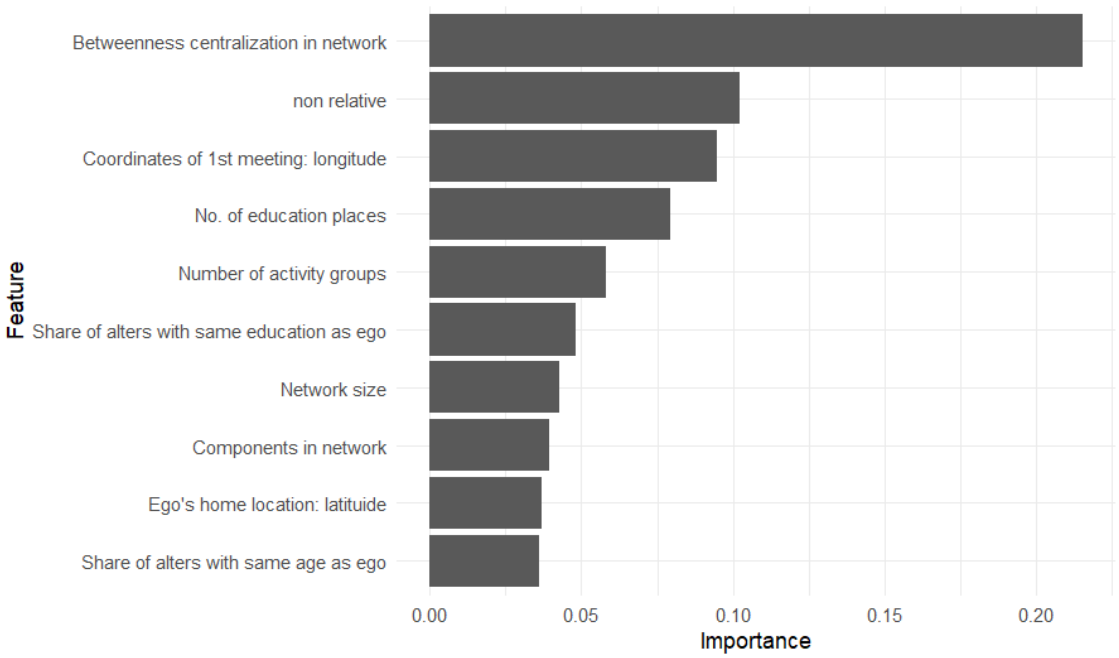

Figure 4.2: feature importance ranking of the binary classifier for the target variable ego eating out with alters

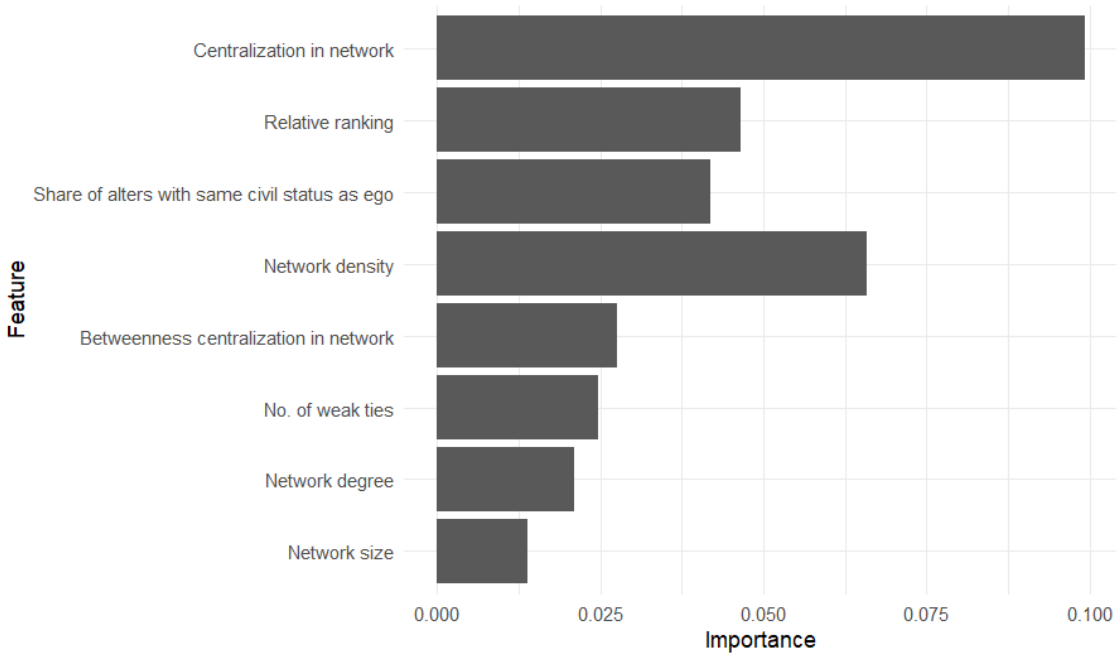

Figure 4.3: feature importance ranking of the binary classifier for the target variable ego going on excursions with alters

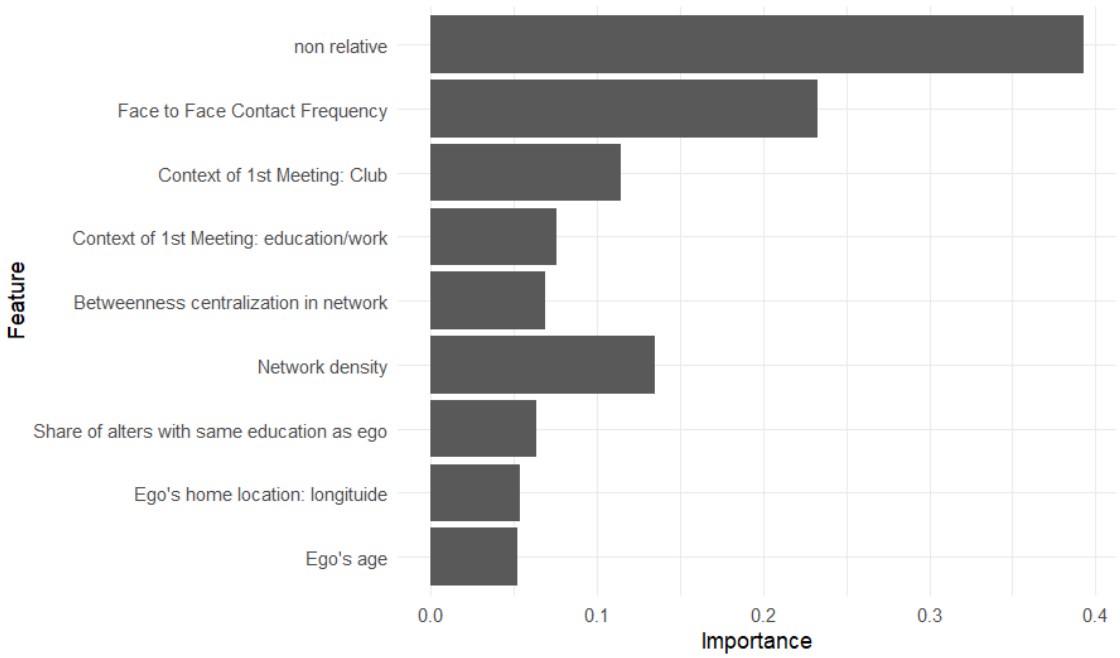

Figure 4.4: feature importance ranking of the binary classifier for the target variable ego engaging in hobbies with alters

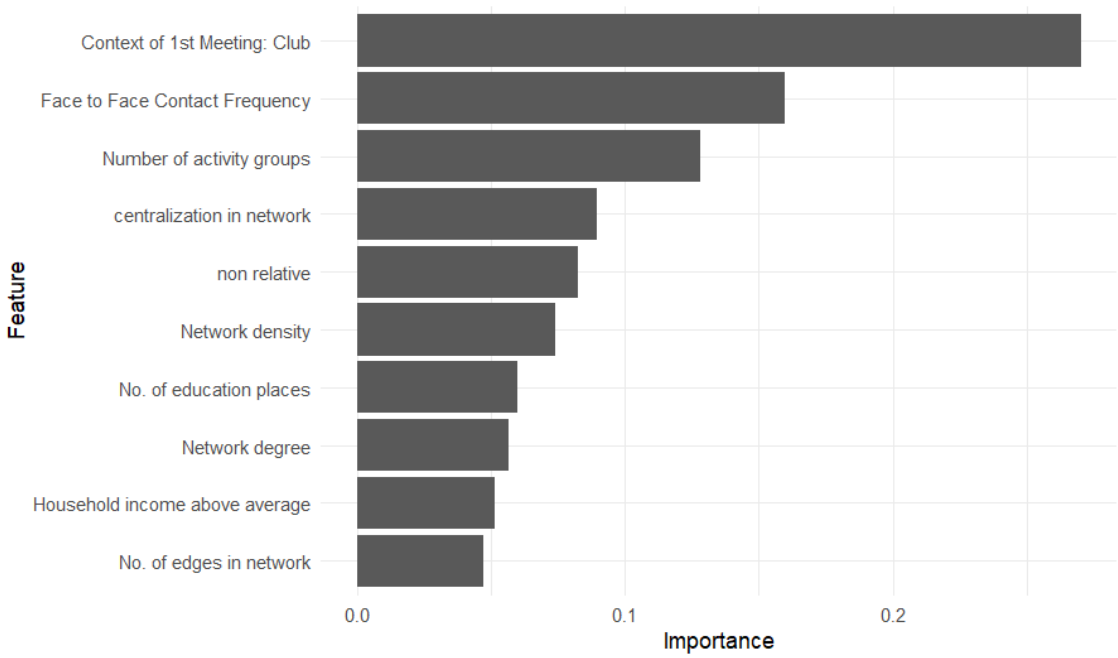

Figure 4.5: feature importance ranking of the binary classifier for the target variable ego playing sports with alters

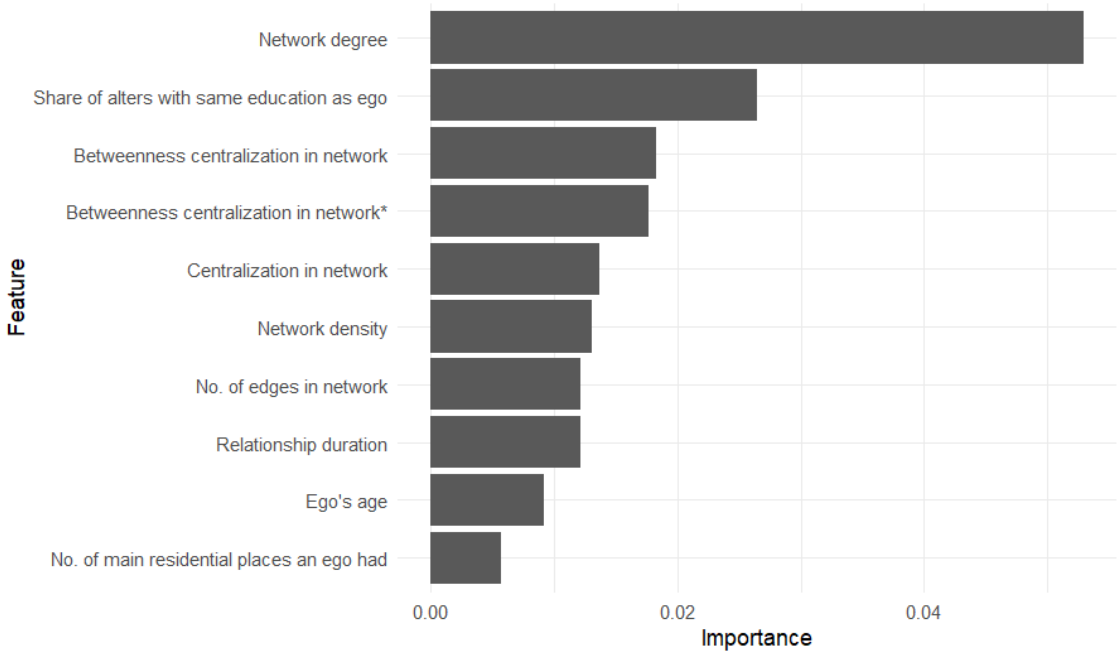

Figure 4.6: feature importance ranking of the binary classifier for the target variable visiting interactions between ego-alter pairs

Table 4.5: MNL Model results from chapter 2 (base alternative = no joint activity, with its ASC fixed to zero).

	Culture	Excursions	Visiting	Sports	Hobby	Eating Out
ASC	−4.8838 (-5.021)	−4.5926 (-18.183)	−6.9576 (-6.424)	−1.6930 (-3.776)	−1.7649 (-6.928)	−2.5764 (-5.689)
Distance	−0.0170 (-2.271)		0.0409 (2.031)	−0.0380 (-2.176)	−0.0334 (-4.128)	−0.0101 (-2.821)
Distance squared			−0.0005 (-2.431)	0.0003 (1.650)	0.0002 (2.303)	
Relationship duration	−0.1524 (-1.309)				−0.1461 (-2.131)	
Age homophily		0.6328 (2.503)		0.4889 (3.968)		0.2847 (1.362)
Both married				1.2247 (6.650)		
Both male			−1.7604 (-2.660)		0.3489 (2.353)	
Both female			0.6394 (2.098)			
Ego male				−0.7257 (-2.774)		
Household size	0.3931 (3.169)					
Proportion of ego-alter pairs with similar sex	0.0206 (1.774)	0.6328 (2.503)	0.0197 (1.287)	−0.0123 (-1.955)		−0.0117 (-1.658)
Proportion of ego-alter pairs with similar age	−0.0170 (-2.328)			−0.0099 (-1.990)		0.0135 (2.240)
Proportion of ego-alter pairs with similar education				0.0058 (1.859)	0.0073 (2.606)	
Proportion of ego-alter pairs with similar civil status			0.0197 (1.287)	0.0054 (1.501)		
Type of relationship						
Married into family	−1.6376 (-2.212)			−1.4628 (-4.082)	−2.2025 (-5.507)	−1.6340 (-3.430)
Relative 1st degree	−1.2001 (-2.431)			−1.1681 (-4.348)	−1.8902 (-8.169)	−1.2411 (-4.673)
Relative				−1.1217 (-3.786)	−1.8277 (-4.306)	−0.7947 (-2.583)
Spouse					−0.8520 (-2.826)	
Acquaintance						−0.2799 (-1.523)
Missing values coefficients						
Distance	−0.5528 (-1.354)			−0.4330 (-2.308)	−0.6751 (-3.755)	
Relation duration					−0.6013 (-1.989)	

Table 4.6: Updated Model results with new estimates and significance levels. Variables in bold indicate changes from the baseline model.

	Culture	Excursions	Visiting	Sports	Hobby	Eating Out
ASC	−4.8938 (-5.056)	−4.5902 (-18.173)	−6.9583 (-6.434)	−2.4027 (-4.663)	−3.5230 (-10.934)	−3.6630 (-7.802)
Distance	−0.0170 (-2.272)		0.0408 (2.029)	−0.0406 (-2.302)	−0.0347 (-4.280)	−0.0104 (-2.857)
Distance squared			−0.0005 (-2.430)	0.0003 (1.766)	0.0002 (2.436)	
Relationship duration	−0.1533 (-1.319)				−0.1486 (-2.163)	
Age homophily		0.6292 (2.489)		0.4419 (3.748)		0.0143 (2.350)
Both married				1.1638 (6.359)		
Both male			−1.7584 (-2.658)		0.3348 (2.268)	
Both female			0.6448 (2.114)			
Ego male				−0.7262 (-2.736)		
Household size	0.3928 (3.161)					
Proportion of ego-alter pairs with similar sex	0.0206 (1.785)			−0.0144 (-2.188)		−0.0119 (-1.684)
Proportion of ego-alter pairs with similar age	−0.0170 (-2.318)			−0.0109 (-2.156)		0.0143 (2.350)
Proportion of ego-alter pairs with similar education				0.0051 (1.666)	0.0072 (2.575)	
Proportion of ego-alter pairs with similar civil status			0.0196 (1.288)	0.0044 (1.248)		
Non-relative				1.0951 (5.312)	1.7843 (10.396)	0.9599 (5.233)
Married into family	−1.6083 (-2.173)					
Relative 1st degree	−1.1875 (-2.406)					
Missing values coefficients						
Distance	−0.5540 (-1.355)			−0.4071 (-2.196)	−0.6823 (-3.808)	
Relation duration					−0.6073 (-2.009)	
Type of relationship					1.1077 (2.810)	−2.2797 (-2.209)

Table 4.7: Comparison of Nested Models Using Likelihood Ratio Test

ML assisted choice model	Model from Chapter 3	LR Statistic	p-value
Model 2	Model 1	29.72	2.369×10^{-4}

4.3.2.2 Feature influence

Feature Importance Order (Activity type sport): The features are listed on the y-axis in descending order of importance. Context of 1st meeting: club is the most important feature, while face to face contact frequency is the least important among the top features displayed. Next to each feature name is the mean SHAP value. This value indicates the average impact of the feature on the model's output. A higher absolute value means the feature has a greater effect on the model's output. The x-axis shows the SHAP values, representing the change in log-odds of the output due to each feature. Points to the right of zero increase the log-odds of the dependent variable, while points to the left decrease the log-odds.The colour represents the feature's value for each observation. Purple indicates higher values, and yellow indicates lower values. This allows us to see if higher or lower values of a feature are associated with increasing or decreasing the prediction. Each dot represents an individual observation (an ego-alter pair in our dataset). The spread of the dots indicates the variability of this feature's impact across different observations.

In terms of how this is useful for a choice modeller, an example would be (the fifth row in Figure 4.11, which is non-relative): the lower feature value is associated with the yellow colour on the plot. Because these dots are on the left-hand side of zero, they are contributing negatively to the model output (which is true as reflected in Table 4.6, had we included not a non-relative, in the model instead of non-relative, we would expect a negative sign). A positive SHAP value means that this feature increases the log odds of the dependent variable (playing sport with an alter in our example). This implies that an ego-alter pair characterised by a non-relative relationship is more likely to play sports with an alter. Therefore, you would expect a positive impact of the alter being a non-relative on the likelihood of choosing to play sport with alter. Similar observations can be made in Figure 4.10 and Figure 4.8. We indeed found the positive signs of the coefficients associated with this variable in the choice model in Table 4.6.

A lack of clear colour separation could indicate that the relationship between the feature and the outcome is not linear or is influenced by interactions with other features. For instance, in Figure 4.11, ”face-to-face contact frequency” may have a different impact on the likelihood of engaging in sports with an alter depending on other features.

This is useful to a choice modeller because the choice modeller would expect the sign of the parameter associated with the variables included in the model, at least from a performance perspective. Any disagreement in the model's results in the comparison of feature influence serves as a signal that may prompt the modeller to check if something is wrong in the modelling process.

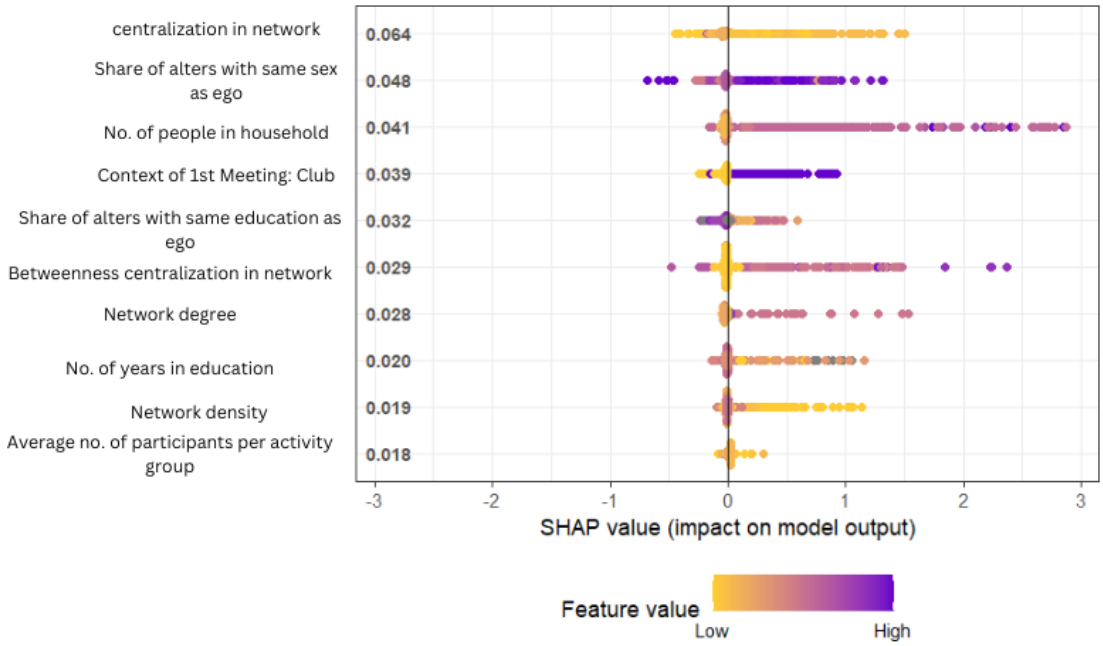

Figure 4.7: feature influence of the binary classifier for the target variable ego engaging in cultural activities with alters

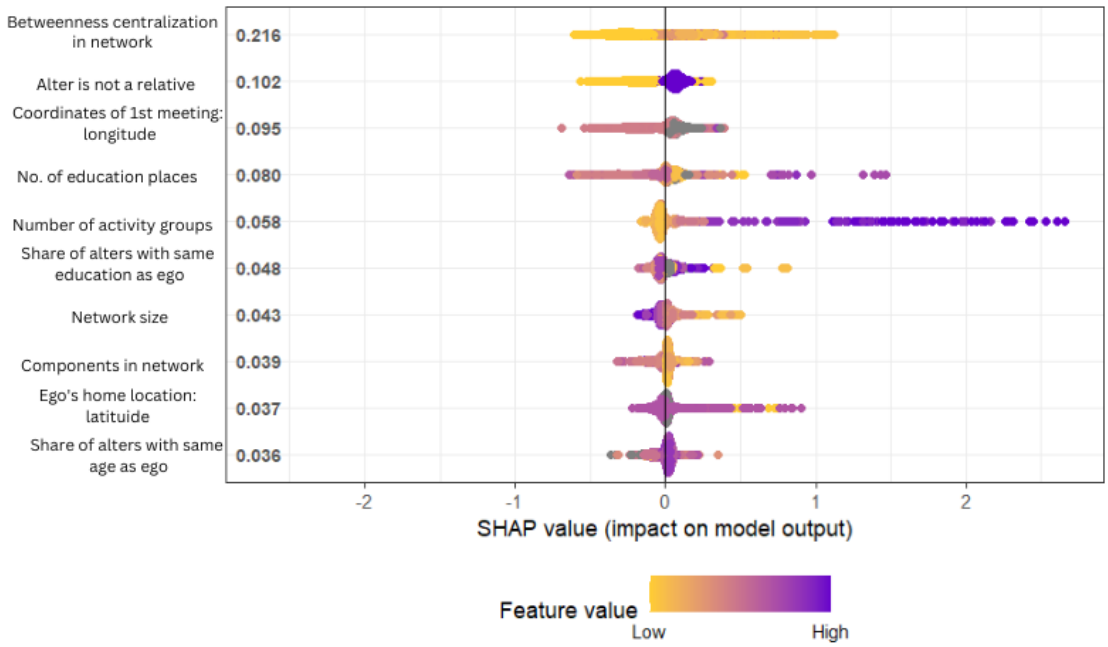

Figure 4.8: feature influence of the binary classifier for the target variable ego eating out with alters

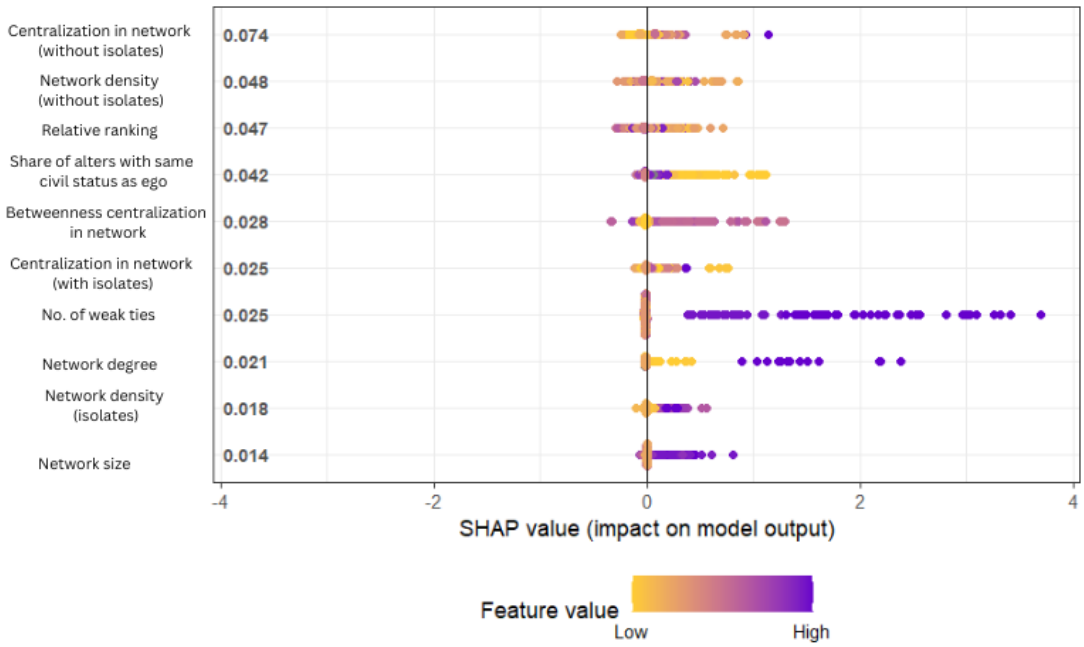

Figure 4.9: feature influence of the binary classifier for the target variable ego going on excursions with alters

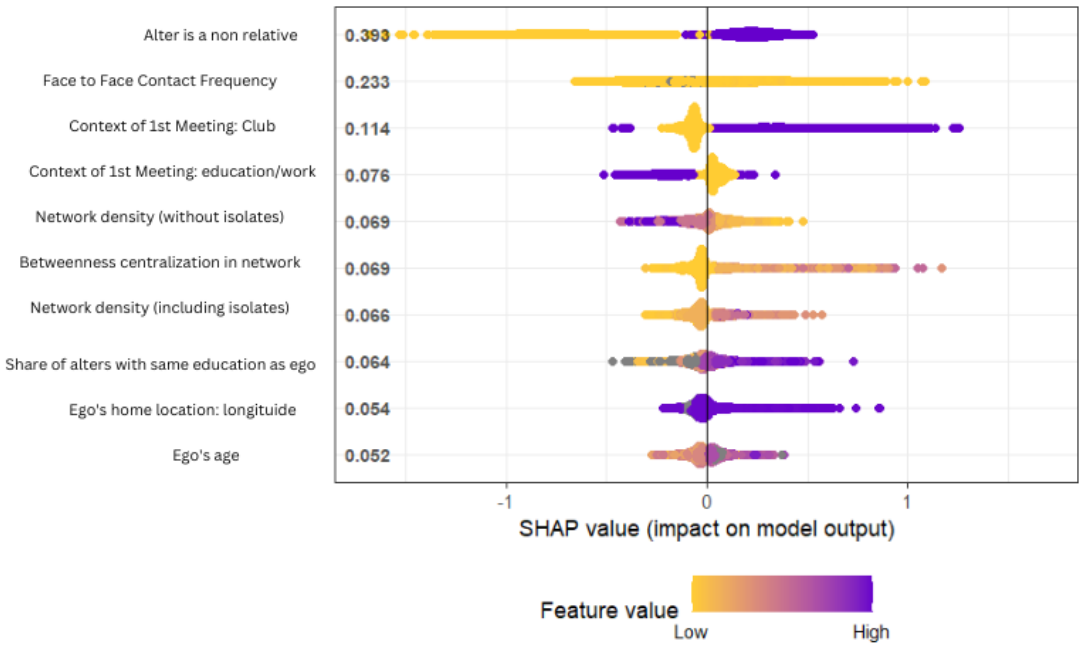

Figure 4.10: feature influence of the binary classifier for the target variable ego engaging in hobbies with alters

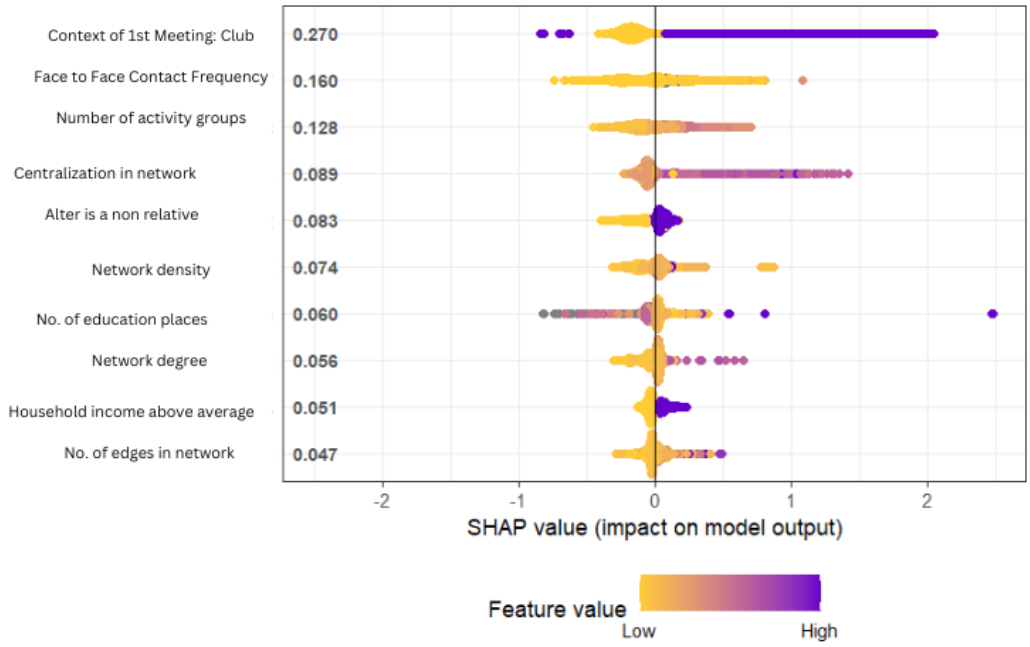

Figure 4.11: feature influence of the binary classifier for the target variable ego playing sports with alters

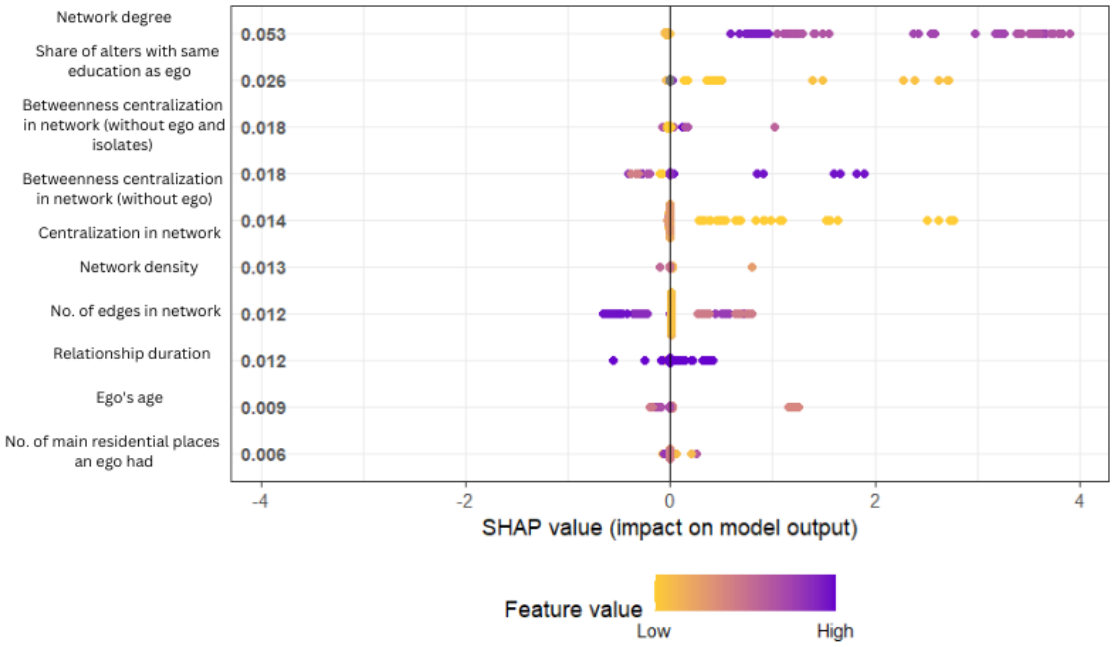

Figure 4.12: feature influence of the binary classifier for the target variable visiting interactions between ego-alter pairs

4.3.2.3 Local interpretation

Force plots are a type of visualization used to show the contribution of each feature to the model's prediction for a specific instance. Each feature's effect is displayed as a force that either increases or decreases the model's output. In a force plot, base Value $E[f(x)]$ is the starting point for the SHAP value calculation. Output Value $f(x)$ is the difference between the base value, and this value is explained by the SHAP values of the features for this instance. Features that push the prediction higher (to the right of the base value) have positive SHAP values and are shown in yellow. These features contribute to an increase in the prediction from the base value. Features that push the prediction lower (to the left of the base value) have negative SHAP values and are shown in purple. These features contribute to a decrease in the prediction from the base value. The length of each colored bar represents the magnitude of that feature's contribution. The longer the bar, the greater the impact of that feature on the model's prediction.

Figure 4.13 and Figue 4.14 are SHAP values to explain the prediction of

playing sports with two different ego alter pairs. In the first pair, probability-decreasing effects such as the context of 1st meeting not in a club are offset by increasing effects such as alter is a non-relative. In the second pair, being a non-relative and face-to-face contact frequency increases the probability of the pair engaging in sports.

As demonstrated here, local interpretation gives the choice modeller an opportunity to examine the behaviour of the individuals in the data, thereby verifying any assumption or hypothesis they might hold before the formal investigation. We recognise that choice modelling is ultimately a result on an aggregate level, but such check as a pre-modelling can still benefit the modeller, for example, if the decision rule such as utility maximisation is really a reasonable assumption to hold by doing so (are an alternative rule demonstrated by some individuals and therefore more consideration should be invested before going down the path of utility maximisation decision rule).

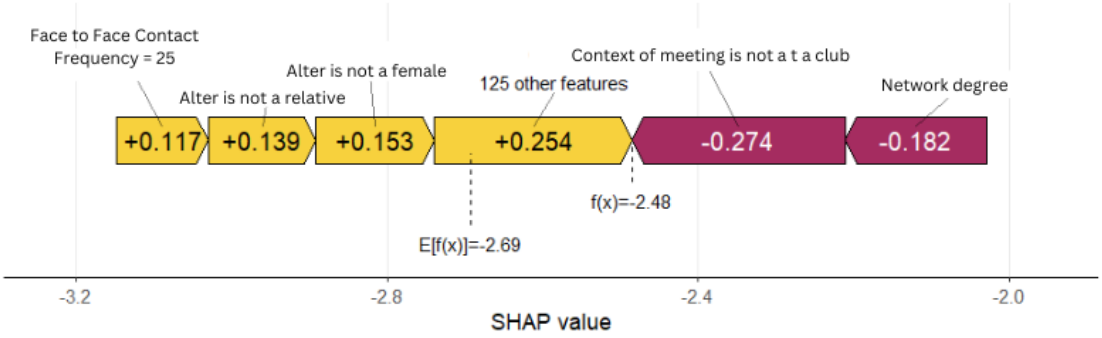

Figure 4.13: force plot of row id 1 of the binary classifier for the target variable ego playing sports with alters

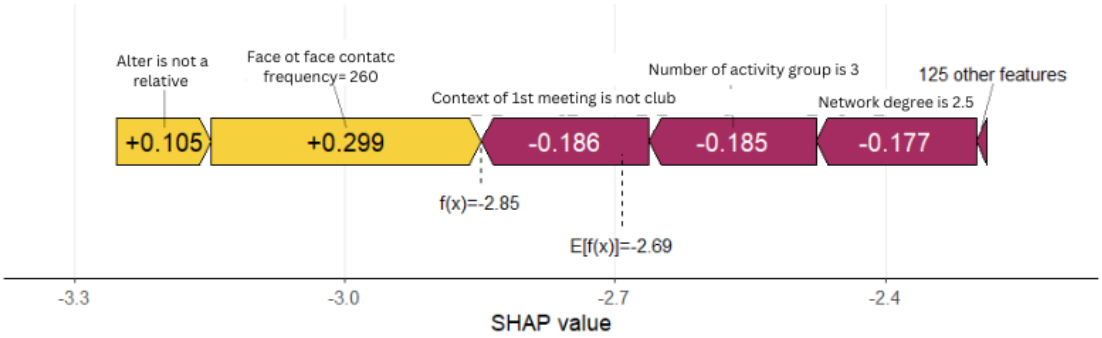

Figure 4.14: force plot of row id 22 of the binary classifier for the target variable ego playing sports with alters

4.4 Conclusion

This chapter has explored the integration of machine learning (ML) techniques, specifically SHapley Additive exPlanations (SHAP), into choice model specification within the context of social network analysis. Our study underscores the potential of combining ML and traditional choice models to enhance both the predictive accuracy and interpretability of behavioural models.

The key findings of this research are:

1. **Enhanced Variable Selection**: By employing SHAP for feature importance ranking, we identified significant variables that influence the likelihood of individuals engaging in various leisure activities with their social contacts (e.g., number of people in the household in the alternative of engaging in cultural activities with alters) This method provided a more nuanced understanding of variable contributions compared to traditional approaches.

2. **Improved Model Performance**: The ML-assisted choice model demonstrated superior performance compared to traditional models. The integration of SHAP values allowed for more accurate identification of relevant features, which improved the overall goodness-of-fit of the model, as evidenced by the likelihood ratio test.

3. **Behavioral Insights**: The application of SHAP provided further behavioural insights into social contact selection in leisure activities. It highlighted the importance of variables such as centralization in the network, the type of relationship between ego and alter, and demographic similarities (e.g., age and sex homophily). These insights are crucial for understanding the underlying dynamics of social interactions.

4. **Feature Influence**: SHAP values provided insights into the influence of individual features on the model's predictions. This technique allowed us to see how different values of each feature affect the classification of specific instances, offering a clear view of the positive or negative impact of each feature.

5. **Local Interpretation** SHAP also facilitated local interpretation, enabling the examination of model predictions at an individual level. This was particularly useful for understanding how specific characteristics of ego-alter pairs influenced the likelihood of engaging in different activities, providing detailed behavioural insights that complement the aggregate-level findings.

6. **Methodological Contribution**: This study contributes to the broader methodological discourse by demonstrating the applicability and benefits of integrating ML techniques, like SHAP, with traditional choice modelling. It is a small step toward bridging the gap between choice modelling and machine learning.

In summary, integrating SHAP into choice modelling represents a modelling approach for enhancing model specification and gaining deeper behavioural insights. While this approach may not be universally necessary, its potential benefits make it valuable to the choice modeller's toolkit. Future research could further explore the applicability of this method in other contexts and extend its use to capture more complex interactions within behavioural models.

This chapter contributes to the advancement of methodological practices by demonstrating the utility of ML-assisted choice model specification in social network analysis, as well as offering a pathway for future research to build upon

these findings. The study demonstrates how machine learning techniques, such as feature importance, feature influence, and local interpretation, can support choice modellers in the pre-modelling stage. These techniques are not intended to replace discrete choice modelling, which has a strong theoretical foundation rooted in utility maximisation, statistical theory, and interpretability. Machine learning models, while primarily built for predictive purposes, lack the level of interpretability and policy relevance that are often important to choice modellers.

In this thesis, we do not advocate for the use of machine learning as a substitute for discrete choice models. Instead, we propose that machine learning serves as a complementary tool in the pre-modelling stage. Specifically, insights obtained from machine learning can guide the modeller in identifying potentially relevant variables or specifications, thereby leading to better model formulations. Additionally, machine learning models can provide reassurance regarding the predictive accuracy of a choice model, offering a form of external validation.

Thus, rather than being the end product, machine learning can be a useful tool in the toolbox of a choice modeller. It can aid in the model specification process and ensure that the resulting discrete choice model is both statistically robust and theoretically grounded.

Chapter 5

Discussion and conclusions

This thesis explores the determinants of social contact selection in leisure activities through the integration of traditional discrete choice models (DCM) and machine learning (ML) techniques. Utilising a snowball-sampled dataset from Switzerland, the study employed multinomial logit models (MNL) to analyse how various dyadic variables influence leisure activity choices with different social contacts. Additionally, machine learning models were used, and Shapley Additive explanations (SHAP) were employed to interpret the best-performed machine learning, and the insights obtained were used to enhance the choice model specification and, thus, the explanatory power. In the following sections of this chapter, we will go over the key gaps and demonstrate how they have been addressed.

5.1 Significance and Contributions

The research presented in this thesis offers contributions to the field of choice modelling, particularly in understanding the social dimension of activity-travel behaviour. By addressing the determinants of social contact selection, this research provides a nuanced understanding of how social factors influence individuals' activity choices. Specifically, it highlights the roles of dyadic variables in shaping social interactions. This enhanced understanding adds depth to existing studies by uncovering the factors that govern social contact choices, which are often under-explored in the literature.

Furthermore, the integration of machine learning techniques with traditional modelling approaches enhances the methodological toolkit available to choice modellers. This integration allows for the identification and analysis of complex patterns and relationships within datasets, which traditional methods might miss. For instance, the use of machine learning in variable selection offers a more effi-

cient and potentially more accurate approach to choice model specification. This methodological approach not only improves model performance but also offers further insights into the factors driving decision-making processes. By contributing to bridging the gap between machine learning and choice modelling, this research contributes to a growing body of literature that seeks to leverage advanced analytical techniques to enhance our understanding of human behaviour, and here in this thesis, particularly in social networks and activity-travel contexts.

5.2 Addressing Identified Gaps

5.2.1 Social Contact Selection in Leisure Activities

- Gap Identified: Previous research often oversimplified the choice of who people interact with by not accounting for detailed dyadic characteristics.

- Contribution: Chapter 2 presented an in-depth analysis of the determinants influencing social contact selection in leisure activities, employing a multinomial logit (MNL) model to understand how various factors affect the likelihood of engaging in different activities within ego-alter pairs. The choice of the MNL model was guided by literature and theoretical considerations that emphasise the need to capture the exclusive nature of activity choices within ego-alter pairs.

Key findings include:

- **Age Homophily**: Individuals are more likely to engage in activities such as sports, eating out, and excursions with alters who are close in age (within 10 years of age difference). This highlights the significant role of age similarity in social activity choices.

- **Physical Distance of home locations between ego-alter pairs**: The likelihood of engaging in leisure activities generally decreases as the physical distance between ego and alter increases, with the exception of visiting activities, which initially increase with distance but decrease beyond a certain point (39.3 KM for visiting, 64 KM for sports and 78.5 KM for hobby). This underscores the importance of proximity in facilitating social interactions.

- **Marital Status**: Pairs, where both the ego and alter are married, show a higher likelihood of participating in cultural activities together.
- **Type of Relationship**: Friends are the most common companions for leisure activities, indicating the primacy of friendship ties over familial or acquaintance relationships in these contexts.
- **Relationship Duration**: Longer relationship durations are associated with a decreased likelihood of engaging in cultural and hobby activities together, reflecting a changing dynamic where these activities become less common over time.
- **Gender Homophily**: Female-female pairs have a higher likelihood of visiting activities, whereas male-male pairs are more inclined towards sports and hobbies. This reflects the influence of gender on activity preferences.

5.2.2 Machine Learning-Assisted Choice Model Specification

Chapter 4 introduced the use of machine learning (ML) techniques to assist in choice model specification. Key findings include:

- **Enhanced Variable Selection**: Employing SHapley Additive exPlanations (SHAP) for feature importance ranking improved the identification of significant variables influencing the likelihood of individuals engaging in various leisure activities with their social contacts.
- **Improved Model Performance**: The ML-assisted choice model demonstrated superior performance compared to traditional models. The integration of SHAP values allowed for a more accurate identification of relevant features, enhancing the overall goodness-of-fit.
- **Behavioral Insights**: The application of SHAP provided deeper behavioural insights, highlighting the importance of variables such as network centralisation, relationship type, and etc.
- **Feature Influence**: SHAP values provided insights into the influence of individual features on the model's predictions. This technique allowed for a detailed understanding of how different values of each feature affected the classification of specific instances.

- **Local Interpretation**: SHAP facilitated local interpretation, enabling the examination of model predictions at an individual level. This was particularly useful for understanding how specific characteristics of ego-alter pairs influenced the likelihood of engaging in different activities, providing detailed behavioural insights that complement the aggregate-level findings.

5.3 Discussion

Chapter 3 advances the understanding of social contact selection in leisure activities by investigating the "with whom" choice in various leisure contexts. The study emphasizes the importance of dyadic characteristics in influencing the selection of companions for leisure activities. This focus provides a nuanced view of how personal and relational factors shape social interactions.

Chapter 4 demonstrates that integrating traditional choice modelling with machine learning techniques offers a powerful approach to understanding social contact selection in leisure activities. The enhanced variable selection process improves model performance and provides richer behavioural insights. The research contributes to the broader methodological discourse by demonstrating the applicability of integrating ML techniques with traditional choice modelling. The use of SHAP for feature importance ranking (primarily) alongside feature influence and local interpretation is a small step towards bridging the gap between predictive performance and interpretability.

A key assumption in modelling the choice process in this research is excluding ego alter pairs with multiple activity engagement, which made the investigation of multiple joint activities impossible. Further insights into behavioural complementarities and substitutions were also therefore unable to be explored using a multivariate probit (MVP) model (Greene, 2003)

It is important to note that the data used in this study was not specifically collected for the purpose of with whom choice modelling. Despite these limitations, progress was made in understanding the factors influencing social contact selection in leisure activities. The insights gained point towards promising directions for future research and data collection efforts, which could enable a more thorough exploration of these complex social behaviours.

5.4 Outlook and limitations

During the course of this research, several aspects of choice modelling using the available data were explored but did not yield the desired results. Two specific approaches encountered limitations due to data insufficiencies and model complexities.

One area of exploration was the implementation of a hybrid choice model to account for latent variables (Walker, 2001). The intention was to model a latent variable that could capture underlying preferences that some individuals might have for certain social contacts (alters) in leisure activities, even when these alters presented identical characteristics (e.g., same gender, age, similar distance from ego). The hypothesis was that underlying preferences (such as YuanFen or Ci Chang in Chinese culture) play a significant role in these preferences.

Despite building a hybrid choice model using relative ranking as an indicator, the indicator proved to be insufficiently motivated. The lack of rich data—meaning the heterogeneity in the data was not extensive enough—prevented a comprehensive exploration of this hypothesis. The relative ranking indicator did not adequately capture the nuanced underlying preferences due to the insufficient variability in the data.

As illustrated in the work of Calastri et al. (2020), ranking can serve as an effective indicator of relationship strength in social network analysis, providing a potential avenue for capturing underlying preferences more effectively in future studies. However, further refinement and more targeted data collection are necessary to support this approach.

We also estimated an MVP model; while the MVP model yielded interesting preliminary results, the data's richness was again a limiting factor. Specifically, the number of observations for ego-alter pairs engaging in multiple joint activities was very limited. This scarcity compromised the robustness of the model's estimations and limited the ability to derive conclusive insights about the complementarities and substitutions in activities. The limited data on multiple joint engagements constrained the model's potential to reveal comprehensive behavioural patterns.

These limitations underscore the importance of targeted data collection for future research. For hybrid choice models to be effective, future data collection efforts should aim to capture more detailed indicators that accurately reflect latent preferences. Additionally, this opens another opportunity for research, extending from Chapter 4, towards bridging the gap between machine learning and

choice modelling (in the research stream of machine learning in assisting choice modelling). Specifically, we could take this as a problem where indicators are absent and employ restricted Boltzmann machines (RBMs) to represent latent behavioural variables (Wong et al., 2018).

The challenges faced, and the results obtained highlight the need for rich, detailed data in advancing our understanding of social interaction dynamics and their influence on activity-travel behaviour. Future research that incorporates these recommendations could yield models with greater explanatory power, providing deeper insights for academic research.

In addition, our study utilised offline social network data, a resource increasingly difficult to access compared to online social network data (which itself is increasingly difficult to access due to legal and privacy issues). This raises the question of whether methodologies and insights from the analysis of online social networks, where many machine learning models have been successfully applied (Luceri et al., 2019; Qiu et al., 2018; Tang et al., 2013; Cuzzocrea et al., 2020), can be adapted to offline social networks. A natural example would be a graphical neural network applied to social networks. For choice modellers, it is worth investigating which tools and methods from online social network analysis could benefit the study of offline social networks. Moreover, understanding whether insights obtained from online social network data can be translated to offline contexts is an interesting research avenue. Conversely, choice modelling could provide structured approaches that could enhance the interpretation of machine learning models in online social network studies. Can the synergies between machine learning and choice modelling be explored in this very context of social network data in online and offline forms?

5.5 Future research

One natural future research direction is to model the choice process explored in this study while relaxing some of the assumptions imposed due to data limitations. By doing so, future research could provide insights into the complementary and substitution patterns of multiple activities that the ego conducts with their alters, thereby contributing further to the understanding of activity-travel behaviour in this context.

Additionally, applying the established methodology to more recent datasets would be beneficial, especially those after rare events such as the COVID-19 pan-

demic, which occurred after the data used in this study was collected. This would enable an investigation of how such unprecedented events may have altered social interaction patterns and activity participation with social network members.

As mentioned in Chapter 3, future research could also explore applying machine learning in the way demonstrated here but with larger and richer datasets—preferably datasets containing a wide range of explanatory variables, or even big data or passively generated data. This approach would allow researchers to better showcase the potential benefits of using machine learning techniques to enhance discrete choice models in a more convincing manner.

References

Ali, A., Kalatian, A., & Choudhury, C. F. (2023). Comparing and contrasting choice model and machine learning techniques in the context of vehicle ownership decisions. *Transportation Research Part A: Policy and Practice*, *173*, 103727.

Alwosheel, A., Van Cranenburgh, S., & Chorus, C. G. (2019). 'computer says no'is not enough: Using prototypical examples to diagnose artificial neural networks for discrete choice analysis. *Journal of choice modelling*, *33*, 100186.

Arentze, T., & Timmermans, H. (2008). Social networks, social interactions, and activity-travel behavior: a framework for microsimulation. *Environment and Planning B: Planning and Design*, *35*(6), 1012–1027.

Athey, S., et al. (2018). The impact of machine learning on economics. *The economics of artificial intelligence: An agenda*, 507–547.

Axhausen, K. W. (2005). Social networks and travel: Some hypotheses. *Social dimensions of sustainable transport: transatlantic perspectives*, 90–108.

Bourdieu, P. (2018). Distinction a social critique of the judgement of taste. In *Inequality* (pp. 287–318). Routledge.

Calastri, C., Hess, S., Daly, A., & Carrasco, J. A. (2017). Does the social context help with understanding and predicting the choice of activity type and duration? an application of the multiple discrete-continuous nested extreme value model to activity diary data. *Transportation Research Part A: Policy and Practice*, *104*, 1–20.

Calastri, C., Hess, S., Daly, A., Maness, M., Kowald, M., & Axhausen, K. (2017). Modelling contact mode and frequency of interactions with social network members using the multiple discrete–continuous extreme value model. *Transportation Research Part C: Emerging Technologies*, *76*, 16–34.

Calastri, C., Hess, S., Palma, D., & dit Sourd, R. C. (2020). Capturing relationship strength: A choice model for leisure time, frequency of interaction and ranking in name generators. *Travel Behaviour and Society*, *20*, 290–299.

Carrasco, J. A., & Miller, E. J. (2006). Exploring the propensity to perform social activities: a social network approach. *Transportation*, *33*(5), 463–480.

Carrasco, J.-A., & Miller, E. J. (2009). The social dimension in action: A multilevel, personal networks model of social activity frequency between individuals. *Transportation Research Part A: Policy and Practice*, *43*(1), 90–104.

Cuzzocrea, A., Leung, C. K., Deng, D., Mai, J. J., Jiang, F., & Fadda, E. (2020). A combined deep-learning and transfer-learning approach for supporting social influence prediction. *Procedia Computer Science*, *177*, 170–177.

Frei, A., & Ohnmacht, T. (2016). Egocentric networks in zurich: Quantitative survey development, data collection and analysis. *Social Networks and Travel Behaviour, 1st Edition. Ashgate Publishing, Ltd*, 51–98.

Golshani, N., Shabanpour, R., Mahmoudifard, S. M., Derrible, S., & Mohammadian, A. (2018). Modeling travel mode and timing decisions: Comparison of artificial neural networks and copula-based joint model. *Travel Behaviour and Society*, *10*, 21–32.

Greene, W. H. (2003). *Econometric analysis*. Pearson Education India.

Habib, K. M., Carrasco, J. A., & Miller, E. J. (2008). Social context of activity scheduling: Discrete–continuous model of relationship between "with whom"

and episode start time and duration. *Transportation Research Record*, *2076*(1), 81–87.

Habib, K. M. N., & Carrasco, J.-A. (2011). Investigating the role of social networks in start time and duration of activities: Trivariate simultaneous econometric model. *Transportation Research Record*, *2230*(1), 1–8.

Hasnine, M. S., Chung, B., & Nurul Habib, K. (2022). How far to live and with whom? role of modal accessibility on living arrangement and distance. *Transportmetrica A: Transport Science*, 1–23.

Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction* (Vol. 2). Springer.

Hess, S., & Palma, D. (2019). Apollo: A flexible, powerful and customisable freeware package for choice model estimation and application. *Journal of choice modelling*, *32*, 100170.

Hillel, T., Bierlaire, M., Elshafie, M., & Jin, Y. (2019). Weak teachers: Assisted specification of discrete choice models using ensemble learning. In *8th symposium of the european association for research in transportation, budapest.*

Hillel, T., Bierlaire, M., Elshafie, M. Z., & Jin, Y. (2021). A systematic review of machine learning classification methodologies for modelling passenger mode choice. *Journal of choice modelling*, *38*, 100221.

Illenberger, J. (2012). Social networks and cooperative travel behaviour.

Joseph, V. R. (2022). Optimal ratio for data splitting. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, *15*(4), 531–538.

Kim, J., Rasouli, S., & Timmermans, H. J. (2018). Social networks, social influence and activity-travel behaviour: a review of models and empirical evidence. *Transport Reviews*, *38*(4), 499–523.

Kowald, M. (2013). *Focussing on leisure travel: The link between spatial mobility, leisure acquaintances and social interactions* (Unpublished doctoral dissertation). ETH Zurich.

Kowald, M., & Axhausen, K. W. (2014). Surveying data on connected personal networks. *Travel Behaviour and Society*, *1*(2), 57–68.

Kowald, M., van den Berg, P., Frei, A., Carrasco, J.-A., Arentze, T., Axhausen, K., ... Wellman, B. (2013). Distance patterns of personal networks in four countries: a comparative study. *Journal of Transport Geography*, *31*, 236–248.

Lin, T., & Wang, D. (2014). Social networks and joint/solo activity–travel behavior. *Transportation Research Part A: Policy and Practice*, *68*, 18–31.

Luceri, L., Braun, T., & Giordano, S. (2019). Analyzing and inferring human real-life behavior through online social networks with social influence deep learning. *Applied network science*, *4*(1), 1–25.

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, *30*.

Maness, M., Cirillo, C., & Dugundji, E. R. (2015). Generalized behavioral framework for choice models of social influence: Behavioral and data concerns in travel behavior. *Journal of transport geography*, *46*, 137–150.

Moore, J., Carrasco, J.-A., & Tudela, A. (2013). Exploring the links between personal networks, time use, and the spatial distribution of social contacts. *Transportation*, *40*(4), 773–788.

Murphy, K. P. (2012). *Machine learning: a probabilistic perspective*. MIT press.

Ortelli, N., Hillel, T., Pereira, F. C., de Lapparent, M., & Bierlaire, M. (2021). Assisted specification of discrete choice models. *Journal of choice modelling*, *39*, 100285.

Qiu, J., Tang, J., Ma, H., Dong, Y., Wang, K., & Tang, J. (2018). Deepinf: Social influence prediction with deep learning. In *Proceedings of the 24th acm sigkdd international conference on knowledge discovery & data mining* (pp. 2110–2119).

Sfeir, G., Abou-Zeid, M., Rodrigues, F., Pereira, F. C., & Kaysi, I. (2021). Latent class choice model with a flexible class membership component: A mixture model approach. *Journal of choice modelling*, *41*, 100320.

Sfeir, G., Rodrigues, F., & Abou-Zeid, M. (2022). Gaussian process latent class choice models. *Transportation Research Part C: Emerging Technologies*, *136*, 103552.

Sharmeen, F., & Timmermans, H. (2014). Walking down the habitual lane: analyzing path dependence effects of mode choice for social trips. *Journal of Transport Geography*, *39*, 222–227.

Sifringer, B., Lurkin, V., & Alahi, A. (2020). Enhancing discrete choice models with representation learning. *Transportation Research Part B: Methodological*, *140*, 236–261.

Srinivasan, S., & Bhat, C. R. (2006). Companionship for leisure activities. *Innovations in Travel Demand Modeling*, *2*, 129–136.

Tang, J., Wu, S., & Sun, J. (2013). Confluence: Conformity influence in large social networks. In *Proceedings of the 19th acm sigkdd international conference on knowledge discovery and data mining* (pp. 347–355).

Tindall, D. B., & Wellman, B. (2001). Canada as social structure: Social network analysis and canadian sociology. *Canadian Journal of Sociology/Cahiers canadiens de sociologie*, 265–308.

Train, K. E. (2009). *Discrete choice methods with simulation*. Cambridge university press.

Tsoleridis, P., Choudhury, C. F., & Hess, S. (2019). Capturing heterogeneity in mode choice decisions: Comparing and combining machine learning and discrete choice models. *Institute for Transport Studies, University of Leeds. Working paper*.

van Cranenburgh, S., & Alwosheel, A. (2019). An artificial neural network based approach to investigate travellers' decision rules. *Transportation Research Part C: Emerging Technologies*, *98*, 152–166.

van Cranenburgh, S., Wang, S., Vij, A., Pereira, F., & Walker, J. (2022). Choice modelling in the age of machine learning-discussion paper. *Journal of Choice Modelling*, *42*, 100340.

van den Berg, P., Arentze, T., & Timmermans, H. (2010). Location-type choice for face-to-face social activities and its effect on travel behavior. *Environment and Planning B: Planning and Design*, *37*(6), 1057–1075.

van den Berg, P., Arentze, T., & Timmermans, H. (2012a). A latent class accelerated hazard model of social activity duration. *Transportation Research Part A: Policy and Practice*, *46*(1), 12–21.

van den Berg, P., Arentze, T., & Timmermans, H. (2012b). A multilevel path analysis of contact frequency between social network members. *Journal of geographical systems*, *14*(2), 125–141.

Walker, J. L. (2001). *Extended discrete choice models: integrated framework, flexible error structures, and latent variables* (Unpublished doctoral dissertation). Massachusetts Institute of Technology.

Wang, S., Mo, B., Hess, S., & Zhao, J. (2021). Comparing hundreds of machine learning classifiers and discrete choice models in predicting travel behavior: an empirical benchmark. *arXiv preprint arXiv:2102.01130*.

Wang, S., Mo, B., & Zhao, J. (2020). Deep neural networks for choice analysis: Architecture design with alternative-specific utility functions. *Transportation Research Part C: Emerging Technologies*, *112*, 234–251.

Wang, S., Wang, Q., & Zhao, J. (2020). Deep neural networks for choice analysis: Extracting complete economic information for interpretation. *Transportation Research Part C: Emerging Technologies*, *118*, 102701.

Wong, M., & Farooq, B. (2021). Reslogit: A residual neural network logit model for data-driven choice modelling. *Transportation Research Part C: Emerging Technologies*, *126*, 103050.

Wong, M., Farooq, B., & Bilodeau, G.-A. (2018). Discriminative conditional restricted boltzmann machine for discrete choice and latent variable modelling. *Journal of choice modelling*, *29*, 152–168.