Personality homophily and social-spatial characteristics in online social networks

A thesis submitted in partial fulfilment

of the requirement for the degree of Doctor of Philosophy

Nyala Noë

August 2018

Cardiff University
School of Computer Science & Informatics

Declaration

This work has not been submitted in substance for any other degree or award at this or any other university or place of learning, nor is being submitted concurrently in candidature for any degree or other award.

Signed ..... (candidate)

Date .....

Statement 1

This thesis is being submitted in partial fulfillment of the requirements for the degree of PhD.

Signed ..... (candidate)

Date .....

Statement 2

This thesis is the result of my own independent work/investigation, except where otherwise stated, and the thesis has not been edited by a third party beyond what is permitted by Cardiff University's Policy on the Use of Third Party Editors by Research Degree Students. Other sources are acknowledged by explicit references. The views expressed are my own.

Signed ..... (candidate)

Date .....

Statement 3

I hereby give consent for my thesis, if accepted, to be available online in the University's Open Access repository and for inter-library loan, and for the title and summary to be made available to outside organisations.

Signed ..... (candidate)

Date .....

Statement 4

I hereby give consent for my thesis, if accepted, to be available online in the University's Open Access repository and for inter-library loans **after expiry of a bar on access previously approved by the Academic Standards & Quality Committee**.

Signed ..... (candidate)

Date .....

To my sister, Beryl, and my parents, Ronald and Bettie,
for their continuous support throughout my thesis.

Abstract

Do birds of a feather flock together or do opposites attract? The aim of this thesis is to consider this question in the context of online social networks. Humans, unlike birds, can flock together based on a wide variety of characteristics, such as age, gender, or political affiliation. The tendency of people to assort based on a common trait is referred to as *homophily*. Research into homophilous traits has often overlooked psychological characteristics. In particular, while personality is studied extensively in the context of social media use, it has received little attention in the homophily literature, which is a gap this work endeavours to bridge.

Online social networks have become ubiquitous in our daily lives and understanding their dynamics gives valuable insight into this new form of social interaction. This thesis highlights the importance of personality homophily in shaping online social networks, while also considering the inherent geographic constraints. In offline social networks, geographic proximity allows for frequent face-to-face interactions, which are essential for the formation and maintenance of friendships. Online networks often reflect offline networks, meaning that people still tend to cluster with others who are geographically close. Using datasets from Facebook and Foursquare, we explore the relationship between personality homophily and geographic distance in detail by considering the distance between similar and dissimilar people, and how they differ in their co-location patterns. We find that people assort based on their personality in both social and spatial contexts, although not all aspects of our personality are equally homophilous. Openness to experience and Conscientiousness emerged as the personality

facets with the strongest homophilic tendencies, while Neuroticism appeared to be less homophilous; Agreeableness and Extraversion fall somewhere in the middle. In other words, birds of a feather do seem to flock together, but this depends on the personality facet considered.

Acknowledgements

I would like to thank first and foremost my thesis supervisor, Roger Whitaker, for his incredible support and guidance throughout the past four years. Working together on this project has made this one of the most enjoyable challenges of my life so far. You have been a true mentor to me, and I could not have asked for a better supervisor. I would also like to thank Stuart Allen for his invaluable advice as I started writing my first papers, and eventually, this thesis. Thank you for the many interesting discussions we have had throughout these past years. I would also like to thank Martin Chorley for his support and contributions to a big part of this thesis; as well as being one of the first people to teach me Python. I would also like to thank Thomas Pollet, for advising me to apply to this PhD position, and for supporting me in the publication of my first paper relating to this thesis.

I would also like to thank all the staff at COMSC for their help throughout the PhD process, be it for the research retreat, poster day, or teaching.

I would like to thank all the PhDs at COMSC who have been a part of this adventure with me. There are too many wonderful people to name them all; I will always fondly look back on the many lunch breaks we spent together, the FTS talks we attended, and the games of Werewolf we played during tea at two. Special thanks go to all the members of C/2.13 for making every day both enjoyable and occasionally productive: Roberto Dyke, Peter Sueref, Lauren Hudson, Matthew Nunes, Beryl Noë, and special guest, Kaelon Lloyd.

Special thanks go to Matthew Nunes for being a great friend throughout. Beating this challenge without your help would have been impossible, and I will forever be grateful for the improvements you made to my Windows machine, such as deleting my completely obsolete start menu.

I would also like to give my special thanks to Kaelon Lloyd for his love and support, especially throughout the last hurdles of this PhD. Thank you for all the times you spent reassuring me although you had your own thesis to worry about. I am very grateful to have you by my side, especially this past year.

Finally, I would like to thank my family, without whom I would never have embarked on this journey in the first place. I would like to thank my parents, Ronald and Bettie, and my sister, Beryl, for their continuous support throughout the whole process. Dank je Ronald voor de inspiratie om deze uitdaging aan te pakken. Dank je Bettie voor je liefdevolle aanmoedigingen deze laatste vier jaren (en die daarvoor!). Merci Beryl, qui a toujours été à mes côtés pendant tout ce temps, et qui m'a donné le meilleur soutien, pas seulement ces quatres dernières années, mais toute notre vie.

Contents

List of Figures

List of Tables

List of Acronyms

LBSN Location-Based Social Network

SNS Social Networking Site

BFI Big Five Inventory

IPIP International Personality Item Pool

API Application Programming Interface

MBTI Myers-Briggs Type Indicator

APSI Adolescent Personal Style Inventory

HEXACO PI HEXACO Personality Inventory

NEO-PI Neuroticism, Extraversion, Openness Personality Inventory

EPQ Eysenck Personality Questionnaire

NCS National Character Survey

ICID Inventory of Child Individual Differences

SHL's OPQ SHL Occupational Personality Questionnaire

EPI Eysenck Personality Inventory

TIPI Ten Item Personality Inventory

Chapter 1

Introduction

Online social networks have become ubiquitous in our daily lives. About 2 billion people are registered on one or several social media platforms, such as Facebook, Twitter, LinkedIn, or Reddit. Facebook currently dominates the social media market with over 1.15 billion active daily users as of December 2017 [119].

Interactions between people have become increasingly globalised, as they are no longer constrained by their physical location; but merely, by their access to an internet connection. It is perhaps surprising then, that in fact, even online social networks are subject to the constraints of geographic distance. Geographic distance is one of the biggest constraints to offline social network to date: relationships do not blossom if people cannot meet in the first place to form a tie, or regularly meet up to maintain their friendship. In other words, distance can dramatically change the relationships that flourish and those that fail. While we might have more contacts from across the globe thanks to the Internet, our closest friends, relatives, and partners still tend to be geographically nearest to us. In our increasingly globalised society, it is important to avoid downplaying the lingering impact distance can have on the shape of social networks. Even on anonymous platforms like Twitter or Reddit, we see a geographical clustering of users. People prefer to follow others who are geographically close to them because they prefer to interact with others they can relate to, and who post news and information that are relevant to them or directly affect them. We might be online globally, but we continue to communicate and interact locally.

Social media platforms have become prime fields of study for the observation of such interactions. Advances in online data collection have allowed for studies on unprecedented scales with large and rich datasets. While studies on offline networks rely on self-reported connections to others, studies on online social networks can use the organic ties that are evident from social networking sites.

The most notable data breach in recent history highlights both the importance and sensitivity of data obtained from these social networking sites. Cambridge Analytica used personality data from Facebook users to target them with political advertisment in an attempt to influence their voting behavior in the American presidential elections and the UK Brexit vote of 2016. It is believed that this purposeful targeting of susceptible users has influenced the results of these votes. The extent of this influence is unknown, however, and might be extremely difficult to detect [64]. This major controversy highlights both the reach of social media, its embededness in our everyday lives, and the power of personality data.

Apart from political inclinations, personality can to some extent predict the way we behave, how we perform in our job, and even how happy we are. It is one of the most important characteristics that determines our uniqueness as individuals. The pervasive nature of personality becomes evident as we see how it operates in the background of many life outcomes. As a result, we do not always realise the impact our personality has on our behaviour. Social interactions are an exception to this. Personality clashes and compatibilities are often cited as reasons for relationship failures and successes. So, do birds of a feather flock together, or do opposites attract when it comes to personality?

Homophily, the tendency of people to assort based on a common trait, is a key characteristic of social networks. People assort based on age, gender, religion, socioeconomic background, or political affiliation to only name a few [109]. Personality has received relatively less attention as a potentially homophilic characteristic. Similarity facilitates both tie formation and maintenance with the people we connect to.

Similarity inspires trust, which greatly enhances communication and relationships in general [39]. Trust also greatly facilitates cooperation, a major theme for individuals, organisations, and even countries. Homophily can be seen as a fundamental building block that holds our society together, by fostering relationships between people and eventually leading to increased cooperation.

The purpose of this thesis is to shed light on personality homophily in online social networks. How does personality similarity influence the connectivity between members of an online social network? What is the effect of geographic distance and geographic co-location on personality homophily?

1.1 Objectives and Contribution

The overarching objective of this thesis is to study the interaction between personality homophily and geographic distance in a variety of online social networking contexts. Below, we outline both the objectives and contributions in four distinct, but related areas of this thesis: online social networks, personality homophily, geographic distance, and local & global networks.

Online social networks: Studies on online social networks have exploded in line with their increase in popularity over the past few decades. While some might be sceptical about the ability of social media research to be generalised to offline networks; a promising body of research has shown that offline and online networks do not differ that much from each other in terms of composition, structure, and size [46]. Online social networks are an interesting topic of study in their own right. Online social media data often benefits from large datasets and organic connections. In other words, people do not need to recall who their friends, acquaintances and kin are, the connection can be scraped from their online profile. This prevents accidental omissions and the reporting of uni-directional relationships, where one person might not reciprocate the friendship of another. Facebook requires both parties to accept the connection before they become

online “friends”. The big advantage of using Facebook is its offline-to-online dynamic, meaning that people usually know each other offline before connecting on Facebook [102].

Personality homophily: Homophily is a major theme in social network research, but homophily based on psychological traits has received little attention in the current literature. Most studies concentrate on salient characteristics or traits that tend to be socio-demographic in nature, such as gender, age, or political orientation. The inherent hidden nature of personality and the need to collect data that is not explicitly stated on one’s social media profile or evident through liked pages might have contributed to a lack of investigation on this topic. This thesis aims to fill the gap regarding personality homophily by studying the impact of personality similarity on Facebook friendship connections. We also aim to compare different methods for measuring similarity, in particular by contrasting continuous and categorical measures. Most other studies on personality have used an either or approach, despite both methods having benefits. A categorical approach allows for clearer results and to contrast extremes, while a continuous approach ensures that no information is lost during the data analysis process.

Geographic distance: As emphasized previously, geographic distance plays a major role in the formation and maintenance of social ties. Geographic distance can be considered assortative in itself, in the sense that it connects people based on the common places they frequent. We aim to identify the role geographic distance plays in online connections and how exactly it interacts with personality similarity. Previous work has looked at distance and personality homophily individually, but not together.

Local and global networks: A methodological objective of the thesis is to complement a more global view of the network with local connectivity analyses. While we consider global network characteristics such as network size and transitivity, we also focus on pairwise and triadic connections to capture the effect personality homophily has at the individual level. By combining both approaches, we hope to achieve a comprehensive overview of the dynamics of the social network in regards to both person-

ality homophily and geographic distance. For distance, this translates to considering the absolute geographic distance between people based on their longitude and latitude coordinates; but also to observe what happens at the level of co-location in exactly the same place.

1.2 Thesis Outline

This thesis is structured around three central themes: personality homophily, geographic distance, and online social network structure. Figure 1.1 outlines how the different themes interact with each other and map to chapters. The arrows represent the influence a source theme has on a target theme. The double-sided arrow for personality and homophily indicate that these two themes are intrinsically linked, without a clear source and target between the two. Personality homophily is foremost explored in chapter 3. In chapter 3, the impact of personality on network structure, especially in online social networks, is also addressed. The link between homophily and network structure spans across two chapters, 3 and 4. We introduce the importance of location in online social networks in chapters 4 and 5, in which we look into the effects homophily can have on geographic distance and co-location respectively. Finally, chapter 5 also addresses the interaction between location and network structure by looking into how an increase in co-location can lead to a very different network. As the thesis progresses, the importance of each of these themes and the intrinsic relations between them will become evident.

We will now outline how we endeavour to answer the following research questions at the end of the thesis: How does personality homophily affect the local structure of an online social network? How does personality similarity relate to geographic distance between online friends? Does personality homophily lead to co-location at the same venue?

In the **second chapter**, we outline the vast literature on (1) human sociality and net-

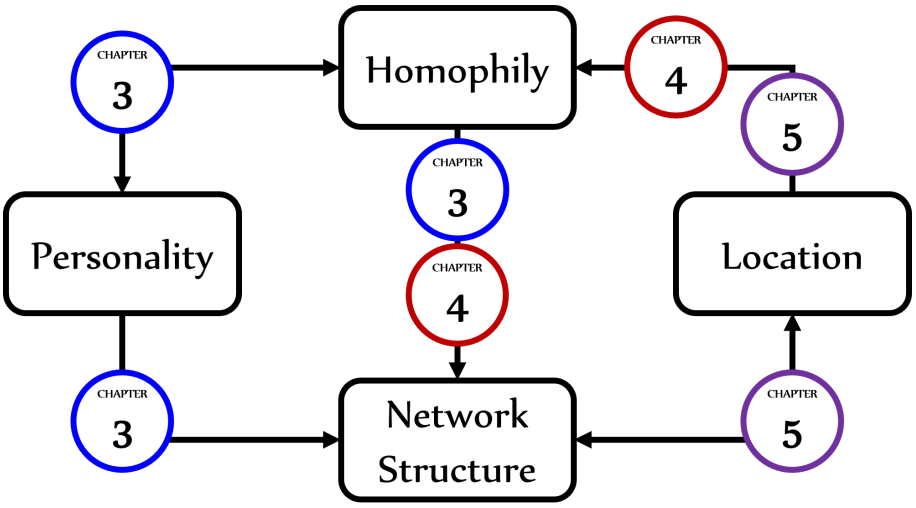


Figure 1.1: Diagram of the core themes of the thesis and their relations.

works, with a focus on online social media; (2) personality, its measurement history, and effects on network structure; (3) geographic distance and how it constraints the expansion and maintenance of our social networks; and finally, (4) personality homophily: our current knowledge on the topic and how we can expand on it.

The **third chapter** will focus on the effects of personality on global social network structure, such as network size and transitivity. We then delve into the homophily aspect of the network. We use different ways of operationalising personality similarity to allow for comparisons with the wider literature, which have used both approaches.

To test hypotheses, we use the MyPersonality dataset [87]. This is an repository of Facebook users personality scores and their connectivity, as well as some basic demographic (age and gender) data. The data is completely anonymised and from voluntary participants willing to share their data for research purposes.

We start by replicating and confirming certain personality relationships with network structure. We first want to replicate the relationship between the different personality facets and network size, as well as transitivity and brokerage. We then investigate whether the online social network derived from a Facebook sample is also subject to personality homophily. In other words, we predict people to be more likely to be

connected if they are similar in personality.

In the **fourth chapter**, we investigate the relationship between geographic distance and personality homophily. The aim of this chapter is to highlight how geographic distance influences personality homophily dynamics in an online social network. We are interested in seeing whether pairs of connected users who are similar in personality and more or less distant to each other.

We use MyPersonality data once again for this chapter. In addition to personality scores and connectivity data, we cross-reference the datasets with geographic location information of the participants.

We make different predictions for each personality facet, based on the qualities and tendencies of each of them. This chapter considers geographic distance at a global level, rather than local co-location between people, which is addressed in the fifth chapter.

In the **fifth chapter**, we take a more local approach to proximity. Instead of focusing on absolute geographic distance, we consider whether people go to the same venues based on their personality. In other words, are people with similar personalities likely to attend the same café or shop at the same store? Unlike in the previous chapters, we consider people's co-location, rather than connection on social media, as an indicator of potential homophily. The aim of this chapter is two-fold. First, it confirms whether personality homophily also occurs at smaller geographic scales; second, it aims to highlight the intrinsic relationship between proximity and personality homophily. Even in the absence of clear-cut relationships, we show that people with the same personality are attracted to the same locations.

A Foursquare dataset with personality scores and checkin history on the location-based social networking application Swarm was used to test hypotheses. The local nature of such Location-Based Social Network (LBSN) applications was more suitable to answer the questions posed in this chapter compared to the previous ones. The checkin history

allowed us to build a network based on co-location, rather than friendship ties. In this model, we let the geographic habits of users determine the connections in the network.

The **sixth chapter** will conclude the thesis. In this chapter, we will summarise and discuss the main findings of the thesis and address its limitations, as well as potential for future work. We will conclude the thesis by aiming to shed as much light as possible on the inner workings of personality homophily in our online social networks, and the effect of geographic distance on the whole process.

1.3 List of Publications

This thesis has lead to the following peer-reviewed publications:

[CiHB16] N. Noë, R.M. Whitaker, M.J. Chorley, Martin, and T.V. Pollet. Birds of a feather locate together? Foursquare checkins and personality homophily. *Computers in Human Behavior*, 58: 343-353, 2016.

[ASONAM16] N. Noë, R.M. Whitaker, and S.M. Allen. Personality Homophily and the Local Network Characteristics of Facebook Personality Homophily and the Local Network Characteristics of Facebook. In: *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 2016.

[CYBER16] N. Noë, R.M. Whitaker, and S.M. Allen. Personality Homophily and Geographic Distance in Facebook. *Cyberpsychology, Behavior, and Social Networking*, 2018.

Chapter 2

Literature Review

The following chapter will discuss the broad literature across the three main themes of this thesis: online social networks, personality homophily, and geographic distance. We will draw from sources in psychology, sociology, and computer science among others, because of the interdisciplinary nature of this project. We keep our focus on how personality and homophily influence tie formation and maintenance, throughout our journey through the current literature on human connections. Why and how do people form friendships with others, and what role do personality and homophily play during this process? To answer this question, we outline the existing literature as follows:

1. Individual differences (Section 2.1)
    (a) individual differences in social connections (Section 2.1.1)
    (b) how we assort into social networks (Section 2.1.2)
2. Social networks (Section 2.2)
    (a) Origins of sociality and how ties are formed and maintained (Section 2.2.1)
    (b) Differences between online and offline networks (Section 2.2.2)
3. Human personality (Section 2.3)
    (a) The history of personality measurement (Section 2.3.1)
    (b) Personality and online behavior (Section 2.3.2)

   (c) Personality and social media use (Section 2.3.3)

   (d) Personality and network structure (Section 2.3.4)

4. Homophily (Section 2.4)

   (a) Existing literature on personality homophily (Section 2.4.2)

   (b) Current state of the art on spatial homophily (Section 2.4.3)

5. Constraints on social relationships imposed by distance (Section 2.5).

6. Current knowledge gap and applications (Section 2.5.3).

We open this chapter with a section on individual differences and their importance in tie formation and maintenance. Throughout this thesis, we will often refer to social relationships as ties, an all-encompassing term for friends, partners, acquaintances, and family members. Individual differences are crucial to our understanding of human individuality. This section will showcase how the different themes of the thesis all relate to each other throughout a person's life and development. We then delve into each theme in more detail, following the structure described above.

2.1 Individual differences

Individual differences are the reason people behave differently in the same situation. Individual differences is a vague term that incorporates all the aspects of a human that make them unique. These aspects can stem from nurture (the environment), or from nature (our genetic predispositions). We may behave differently because we are from different cultures, had a different upbringing, or simply are genetically different [141, 21]. But what about a set of identical twins who grew up in the same home? Unlike fraternal twins, identical twins share the same genetic make-up. Would they not behave and react in the exact same way when confronted with the same situation? And yet, we see differences in behavior between identical twins; they are still each

unique individuals. What distinguishes them are not their physical aspects, which can be so similar that they get frequently confused with one another, but their personality. We structure this review by considering a pair of identical twins and how they navigate their social connections as they grow up.

Personality is a big part of what makes each individual unique. A person's personality can be described using five broad facets according to the Five Factor Model [58]. With the use of personality questionnaires, it is possible to derive scores for each of the facets for a person. Personality modelling and testing is described in more detail in Section 2.3.

To better understand what we mean when we refer to personality, the five facets are defined as follows:

1. **Openness to Experience**: If people are willing to try new things, and are creative; they are considered *open*. People who score low on this facet tend to be more averse to change and more conservative in their approach to life, they are considered *non-open*.

2. **Conscientiousness**: This facet describes how organised and timely a person is. *Conscientious* people have good focus and are less likely to be distracted. People who score low on this facet tend to be disorganised and easily distracted; they are considered *unconscientious*.

3. **Extraversion**: This facet describes how sociable and people-oriented a person is. People who score high on this facet enjoy being around others and seek out opportunities to socialise and connect with others; they are considered *extraverts*. People who score low on this facet tend to prefer or value alone time and do not tend to actively seek out social situations, although they still might enjoy them; they are considered *introverts*.

4. **Agreeableness**: This facet describes how friendly and accommodating a person is. A person scoring high on this facet is considered *agreeable*. A person scoring

low on this facet might be difficult to get along with and unfriendly with the people they interact with; they would be considered *disagreeable*.

5. **Neuroticism**: This facet describes how anxious and prone to worry a person is. A person scoring high on this facet is considered *neurotic*. A person scoring low on this facet tends to be calm and in control of their emotions; they are considered *emotionally stable*.

Personality can be used to predict behavior. It is very difficult for us, or even impossible, to think of two different people with exactly the same personality. There are always small differences in how we behave and react to our environment, and these small differences are due to our own unique personality. At the same time, personality is what makes each individual behave consistently over time and situations [141]. Of course, a person's personality does not exist in a vacuum. Personality is both shaped by our environment and our genes, although the exact precursors for personality are still widely unknown and difficult to pinpoint [21]. It is generally agreed that personality is mostly genetically determined (40%), although our environment can influence how it manifests itself. Shared environment accounts for about 7% of an individual's personality [21]. This means that two individuals who grow up in exactly the same environment, will have about 7% of their personality in common because of it. Bouchard estimated that the rest of the variance found in personality was half due to non-shared environment (26.5%) and the other half due to error of measurement (26.5%). This makes genetic factors the most influential precursor of personality. It could be argued that identical twins, despite growing up in the same home, could be treated differently by their parents. However, it seems that children are treated differently because of their personality, rather than the other way around [21]. We use the example of identical twins to further illustrate the role personality plays throughout our lives.

2.1.1 Connections with others

A critical example of a non-shared environment is an individual's own social network. Despite growing up together, identical twins will eventually make their own friends. Their social networks might still have some overlap as they grow up, especially due to kin ties. Social networks are an essential part of a person's life [68]. Personality has become relevant to social networks, because it has been shown to have an influence on the ties we form with others [8]. Personality contributes to both the quantity and the quality of our social ties. Will our twins have a lot of social connections, but with less emotional depth; or will they have relatively fewer, but closer, friendship ties? The former is often characteristic of the networks of extraverts, while the latter is more likely for introverts. While personality might contribute to our initial network position, this network position might, in turn, reinforce certain personality traits. For example, easily anxious people tend to end up at the periphery of social networks, with relatively fewer connections compared to more adjusted individuals [29, 83]. However, being in such a disadvantageous network position can be stressful in itself, which has potential to reinforce the anxious personality of the individual at the periphery.

behavior in online spaces has garnered increased interest over the past decade. In particular, the way we interact with technology, use it, and how it influences our behavior has received considerable attention. Personality is reflected in the way we use smartphones and social networking applications. These social networking applications have given us the unique capability to keep track of our social network without much effort. Technology allows us to offload cognitive and memory tasks to external devices. This allows us to "extend our mind" and free up valuable cognitive processing [27].

Personality is by far not the only determinant of the position we occupy in our social network, and its composition. Actually, one of the most influential factors in tie formation is simply proximity, even in online social networks [60, 133]. No matter how compatible two people are; if they never have a chance to meet virtually or physically, a social connection cannot form. Proximity can be both temporal and geographic, two

people need to be in the right place at the right time to have an opportunity to connect. As our twins grow up and possibly move away from home, they will encounter different people based on their respective geographic locations and how they decide to spend their time. While proximity is an important prerequisite for meeting new people, it is not always enough. One of our twins might take the bus to work everyday and meet the same people on it. The temporal and geographic proximity conditions are met, yet people rarely make friends on the bus. The main reason is because the environment of a bus is not favourable for new social connections. So, in addition to the proximity condition, people need to engage in a shared activity or goal to be able to connect [90]. A bus does only serve to transport people from A to B and does not give any incentive to people to start communicating, other than boredom. A social club, sport society, a classroom, or a university campus are all places that are already much more favourable than the bus. People can connect based on shared activities, age, or interests respectively [90].

Finally, another important condition for our twins to successfully build new friendships in their respective environments is that the people they encounter need to be receptive to adding a new person to their social circle. First-year undergraduates around the world are often very eager to form new connections. Most of them left their homes for the first time, and have left their high school friends behind, which means they are in need of a new social and support network. As people grow older and settle, they rely more on their existing social networks. People find it more difficult to make new connections as they grow older. This might be on the one hand due to a slight decrease in openness towards change and new people [128], but also because the conditions become less and less favourable for new connections.

2.1.2 Assorting into social networks

Now that all the conditions are met for our twins to make friends, what exactly makes them choose one potential friendship over the other, all else being equal? Will the

twin choose their neighbour on the left or the one on the right? Friendships flourish on commonality. Sharing interests, cultural and social backgrounds all promote a better understanding of the other and helps them relate. Gender and age can also play a role, although this might be more pronounced in children and adolescents. Similarity promotes trust, which in turn facilitates tie formation and maintenance. The tendency of people to connect based on shared or similar characteristics is referred to as *homophily*.

Homophily is an important factor in social network dynamics and tie formation and can be based on a number of socio-demographic characteristics, such as age, gender, religion, political orientation, or socio-economic background [109]. Some of these commonly studied characteristics are partly constrained by external factors. For example, churches and schools are common environments during which social connections are formed, making it likely for people to be of the same religion and age. In short, the environment we live in might constrain the diversity of people we encounter.

Kossinets and colleagues distinguish between induced and choice homophily in this regard [89, 90]. Induced homophily refers to the person selecting or being pushed towards environments that induce homophily without a conscious choice. For example, classes that are separated by age promote people to connect with others of the same age simply because no other option is available. Choice homophily refers to more active choices the person makes in terms of relationships, by selecting people based on similar characteristics within the constrained environments imposed by induced homophily [90]. Online social networks might appear to lift some of the constraints imposed by induced homophily, notably the need to physically share an environment to connect. However, even in online networks, geographic distance remains an influential factor as we will see in more detail in Section 2.5. The Internet has separate online spaces, just like the offline world. People would need to frequent the same virtual space to have a chance to interact and connect. It is also important to consider that having many grounds for commonality does not guarantee a connection to form; individuals need to "click" to become friends. In other words, their personalities need to be compatible.

It is no use having shared interests or backgrounds for two people if they cannot get along because of their inherent disposition.

Personality is an interesting characteristic to consider as a catalyst for social connection. It is not environment-dependent, like religion or socio-economic background, or visible like age or gender. Generally, people need to interact with a person repeatedly to get to know their personality fully; and personality influences the likelihood to interact with others. This aspect of personality is not well-studied. Personality is relatively easy to measure explicitly, but not implicitly, and it often requires the administration of psychological tests. behavior could be used to study personality in the wild, but this might make it difficult to equally capture each of the five facets as they require very different settings to be expressed. Another hurdle to the study of personality in social networks is the difficulty with which information about connections is gathered. The use of online social networks, which display clear and visible connections between users and the use of applications to easily gather personality data makes these type of studies possible. This thesis relies on both methodologies to study personality in online social networks.

2.2 Social networks

Social networking has become a defining feature of the 21st century, partially through the surge in technology and the rise of social media. Its success can be related back to the nature of the human species and its propensity to socialise. We rely on our social network for financial aid and emotional support [159]. If we want to understand why online social networks have become so ubiquitous in our daily lives nowadays, we first need to understand human sociality.

Technology has given us the unique opportunity to keep track of our social groups, maintain connections, and understand their dynamics. This was something that used to only be possible through direct and repeated interactions with our social contacts, but

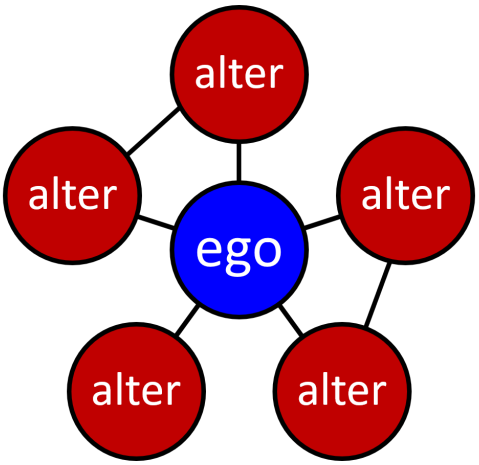


Figure 2.1: Example of an ego-network with one ego (in blue) and five alters (in red). The dots are nodes, while the black lines are edges.

is now much easier via our smartphones, laptops or other technological devices. The following sections will highlight the origins of human social networks, and how online and offline networks differ from each other.

Social network representations are typically composed of nodes, which represent people, and edges, which represent a social connection. In Figure 2.1, the social network of one particular person, the blue dot, is presented. The red dots are the nodes, while the black lines connecting them are the edges. Ego-networks like these are useful to map out the social network of one particular person. The blue dot at the centre of the network is typically referred to as "ego" and is always connected to all the other nodes in the network. The red dots are referred to as "alters". Alters can be acquaintances, friends, family or any other person ego has a social connection with. Alters themselves can be connected to each other as well, but do not have to be.

2.2.1 The origins of sociality and networks

Survival benefits in the face of more powerful predators originally promoted cooperative group-based behavior. Advantages from in-group cooperation created a selection

pressure through which sociality has evolved, resulting in the ability to prosper in large groups with sophisticated social cognition, as exemplified by phenomena such as the Theory of Mind [22] and intentionality [148]. This evolution leaves humans disposed and attracted to the use of digital media and online social networks, in part because they offer an additional channel through which individuals can reason about their social group and potentially maintain their relationships with others. Much reasoning about oneself involves social comparison [51], undertaken to help establish one's overall social positioning, and whether or not to donate precious resources to a third party [160]. This is phylogenetically ancient and embedded in human survival [146], with its suggested origins in evaluating competitors and assessing whether or not to commit resources to challenge a rival in the hierarchy.

The Social Brain Hypothesis posits that evolution of the relatively large human brain is a consequence of the complex social system within which humans live [44]. Cognitive and time resources combine to structure the human's social group as a function of emotional closeness. Our social networks can be organised into layers, with the strongest connections at the centre, and the weakest connections at the periphery. Emotional closeness and interactions decrease across network layers, which together typically cover around 150 relationships, based on the capability of the neo-cortex [43]. Dunbar and Hill have identified that most people have roughly the same organisation for their ego-network in terms of layers: a clique with the 5 closest people to ego, a support group with roughly 15 people with whom ego regularly interacts, followed by the active network, which is typically comprised of around 50 people. These are people ego knows well enough to send Christmas cards to, but would not interact with as regularly as the support group. The last layer is comprised of alters ego knows and can place in a social context, but who are otherwise their weakest ties, which brings up the total number to 150 on average [68].

The main constraints governing the number of relationships are time and opportunity costs, which are exacerbated in resource-constrained environments. Language has

evolved as an efficient way to perform social grooming, with this advantage promoting the expansion of group size from 120 [68, 164]. These meaningful ties have been maintained through physical interaction, with emotional closeness and frequency of contact being strongly related [129]. In particular, face-to-face interactions provide important support to sustain weaker ties. These are of most value for knowledge acquisition and heterogeneity [63], in contrast to emotional or financial resources provided by kin.

Larger social groups generally imply that an individual is putting more effort into maintaining a social group with a larger proportion of weak ties, over which an individual can exercise choice and some control. Based on the convenience of the Internet and ubiquity of digital communication through devices such as smartphones [157], it has been widely speculated that social media and online social networks may help to overcome the cognitive constraints that present a barrier to the increase of average group size beyond 150. This has been motivated by observations on engagement with Facebook for example, where roughly 18% of the global population were daily active users in December 2017 [119]. Despite such technologies enabling individuals to expand their online social networks beyond the restrictions of offline networks, there is a remarkable similarity in structure between the two [46]. However it has been argued that many online social network connections do not represent meaningful relationships, and that cognitive constraints evident in offline networks still persist [45]. Therefore, while offering an additional channel to maintain some weak ties, the existence of a large number of meaningful weak ties in an online setting remains questionable. And while distance might not matter for tie maintenance, it does appear to matter for tie formation [84].

Generally speaking, the rapid evolution of technological capabilities such as the Internet and online social networks are in stark contrast to the speed of biological evolution, which has developed from selection pressures predicated on physical rather than virtual interactions. This means that when the backdrop of physical interactions is diminished, it becomes harder for humans to directly understand and navigate online social

relationships, where there are greater opportunities for adjusting self-presentation and maintaining a degree of self-projection [13]. These issues make observation of human behavior in response to technology important, with social decision making implicit in how we choose to interact through online media. Consequently understanding the interactions between user behavior and online social networks provides significant insight.

2.2.2 Online and offline networks

The relationship between personality and network structure has been extensively studied in offline social networks. A recent review on the subject highlights the main findings for each of the personality facets of the Five Factor Model [137].

In offline networks, extraverts have a preference for social engagement, seek out connections with others, have better social skills, and are good at gaining power and influence. Extraversion is related to larger personal networks. Extraverts' core network tends to have a higher proportion of friends compared to kin, and network expansion is related to tie formation with non-related alters. Interestingly, extraverts might feel like they have bigger networks than they actually do, because they classify people in their network more readily as friends. Extraverts are also more interested in tie formation, rather than tie maintenance [138].

Low neuroticism is useful in workplace networks, because it allows people to maintain their position of power. Neurotic individuals are also motivated to build relationships between non-connected people [83]. Neuroticism seems to be less related to the size of the network, and more to the position within the network. Emotionally stable people are more involved in tie maintenance compared to tie formation.

Agreeableness is related to both tie formation and maintenance. It seems that agreeable people are mostly involved in tie formation because others tend to choose agreeable people as their preferred interaction partners [138]. There is some evidence that agree-

able people tend to occupy central network positions, but this simply might be due to others wanting to connect with them.

Openness to Experience is associated with homophilic tendencies, especially during tie formation, where being open might be useful to make new connections [138].

Conscientiousness does not seem to have an influence on network size, but conscientious individuals tend to occupy central positions in networks related to advice, knowledge, reliability, and performance [85]. The evidence on this effect is mixed however, as some studies have found no evidence of this. Conscientiousness appears to be related to the maintenance of family ties, but not necessarily friendship ties [137].

It is important to consider the difference in homophily between online and offline social networks. Hristova and colleagues [71] contrasted Facebook ties to offline ties for students that lived in two adjacent residences during term time. Homophily in terms of study years and residence (situational homophily) was stronger among online ties, compared to offline ones. On the other hand, political orientation among friends tended to be more diverse online compared to offline. Offline and online health habits were similar, while music preferences were very widely distributed, but did not show a difference in similarity between the offline and online network in general.

The quality of relationships in online social networks can be quantified through different means: frequency of contact, duration of contact, or emotional closeness for example. Research on online social networking sites often focuses on the presence of a connection (friends or followers), but not necessarily on its quality, which can be challenging to establish. Antheunis and colleagues [7] studied the quality of three types of relationships: offline, mixed-mode, and online friendships. Offline referred to friendships that started offline, but could also incorporate online communication. Mixed-mode referred to relationships that started online, but also have an offline component. Online referred to friendships that formed online and only use online communication. The quality of offline friendships has been found to be significantly higher than mixed friendships, which were, in turn, significantly higher in quality than online-

only friendships [7]. The fact that primarily offline relationships were considered of better quality than mixed friendship might indicate that online communication can supplement existing offline relationships, but never truly replace them. However, the type of relationships (offline, online, mixed) did not have an influence on the total duration of the friendship. Also of interest is that the difference in quality between offline and mixed friendships disappeared when compared over time. However, the difference in quality of online-only friendships compared to mixed and offline was maintained over time.

Tie strength has been associated with stronger homophily between two individuals [71]. The more ways two people stay in touch (texting, chatting, face-to-face, etc), the greater their similarity in terms of political affiliation, health, music taste, and situation (residency and cohort) tends to be. The researchers found no clear homophilic tendency for health and political affiliation, mainly due to high homogeneity in the sample, but similarity in music taste and situation was higher as network distance decreased in an offline student network [71].

Despite differences in the quality of relationships, online networks are similar in structure to offline networks [46]. There are cognitive constraints to the number of relationships that a person can maintain at a given time, according to the social brain hypothesis [44]. On average, people can maintain about 150 of such meaningful connections, often referred to as Dunbar's number [44]. It has been argued that social media would lift these cognitive constraints and allow people to stay in contact with more people than ever before. However, actual comparisons show that an online and offline sample had similar means to Dunbar's number with 155.2 and 182.8 alters respectively. This research demonstrates clear parallels between the online and offline worlds. Of interest in particular, is the preservation of a layered structure, which mirrors the one of an offline social network [46]. This layered structure of the network is typically represented as concentric circles around a person, with the core containing the strongest ties, and the subsequent circles containing weaker ties.

2.2.3 Advantages of studying (online) social networks

Clifton and Webster emphasize how social networks have been under-utilised in personality research [30]. Social networks help visualise individual personality in the social context they exist in. This is especially important in the context of personality homophily, as the concept relies on analysing the connections between people based on their personality.

Online social networks are an interesting topic of study in their own right. Online social media research often benefits from large datasets and organic connections. In other words, people do not need to recall who their friends, acquaintances and kin are; these connections can be scraped from their online profile. This prevents accidental omissions and the reporting of uni-directional relationship in which one person might not reciprocate the friendship of another. For example, Facebook requires both parties to accept the connection before they become online "friends". The big advantage of using Facebook is its offline-to-online dynamic, meaning that people usually know each other offline before connecting on Facebook. In a study on the composition of Facebook friends, only 4% were classified as strangers and 2% as online-only. The biggest group was acquaintances with 27%, followed by activity (people met during leisure, work, or study activities) with 24%. Close friendships comprised 21% and maintained (past relationships and friendships) represented 18%. 4% were unaccounted for [102]. How does the distribution of these relationships compare to an offline network? The proportion of close connections decreased significantly with network size ($r = -.33$). This is in line with research on emotional closeness and network layer size [115]. This makes Facebook, and other social networks, a valuable resource for studies on (online) social networks.

2.3 Personality modelling

Personality affects important life outcomes, such as happiness [32, 54] and job satisfaction [81]. Personality can also provide a useful framework for studying behavioral problems in children [47], adolescents [2] or even adults [70].

Personality measurement lends itself to deployment through technological means because of the question-based scoring approach. Of interest is therefore the differences in personality models used in the literature. A brief overview of the history of personality measurement is provided, before discussing the different models and tests used nowadays to quantify and qualify personality.

2.3.1 History of Personality Measurement

The Five Factor Model comprises five bipolar facets: Openness to Experience, Conscientiousness, Extraversion, Agreeableness, and Neuroticism. As defined in chapter 1, extraverts are sociable and talkative, whereas introverts are quiet and value alone-time. Agreeable individuals are friendly and easy to get along with, while disagreeable people are not likeable. Highly neurotic individuals are characterized by their persistent negative affect and worry, while emotionally stable individuals experience a healthy balance of emotions instead. Conscientious individuals are focused and driven, while unconscientious people are impulsive and lack self-control. Finally, open individuals are creative, curious and open to new experiences, while non-open individuals are more conservative [41].

Personality represents the fundamental characteristics that describe an individual, and is an influential determinant of behavior, especially in social situations and interactions [58]. Personality is thought to be about 50% [41] genetically determined, with the rest attributed to environmental and cultural influences. The current dominant approaches to characterising personality are lexically based, the origins of which date

back to Galton [55]. Refinement of Galton's efforts have led to a corpus of personality adjectives [58].

Galton first endeavoured to identify the terms that could describe unique human characteristics. A century later, Allport and Odbert, as well as Norman another 30 years later, further refined Galton's efforts to assemble a corpus of personality adjectives [58]. Thurstone was one of the first to identify that this list of $4,500$ personality adjectives could be broken down into five distinct facets. Also from Cattell's original 35 bipolar factors, only 5 facets could be replicated. Further models by Borgatta, Digman & Inouye, and McCrae & Costa were also comprised of 5 facets, each of them based on different variables and with slightly different qualifications for the different facets [58].

The final model of personality that emerges from these five facets, the now widely used Five Factor Model (also referred to as the Big 5), was extensively validated and refined throughout the years [106, 58]. It is noted that variations exist across different human societies in the interpretation of the five facets [65].

The most dominant alternative to the Five Factor Model is the HEXACO model of personality [10]. HEXACO is an abbreviation for the model's six facets, which are: Honesty-Humility, Emotionality, Extraversion, Agreeableness, Conscientiousness, and Openness to experience [95]. An extensive review on this model is provided by Ashton and colleagues [9]. The HEXACO model adds a sixth facet, Honesty-Humility, and replaces Neuroticism with Emotionality, but retains the Openness, Extraversion, Agreeableness and Conscientiousness facets. There is an ongoing debate over which of the two scales should be used for an optimal representation of personality - for example Lee and colleagues [95] argue in favour of HEXACO while Bashiri and colleagues [15] argue against this. However, the Five Factor Model of personality is widely used in the research reviewed in this thesis, with 56 from 109 studies using a scale derived from the Five Factor Model to measure personality. Four papers used the old Eysenck model [67, 5, 69, 6], one paper used the Myers-Briggs Type Indicator (MBTI) [114]

and only one paper used the HEXACO model [115]. Remaining papers focused on discrete psychological factors, such as happiness [20], self-monitoring [110], or the entrepreneurial personality [82] and one dealt with personality pathology [28]. The MBTI model classifies people according to four bipolar facets: Extraversion or Introversion, Sensing or Intuition, Thinking or Feeling, Judging or Perceiving. The Eysenck model contains 3 facets: Psychoticism, Extraversion, and Neuroticism [41].

Table 2.1: keywords used for paper search, divided by theme: personality, social networks, and social media. SNS stands for Social Networking Sites, LBSN stands for Location-Based Social Networks.

personality	**social networks**	**social media**
personality	social networks	social media
Five Factor Model	online social networks	SNS
Big 5	homophily	Twitter
Extraversion	transitivity	Facebook
Neuroticism	clustering	LBSN
Openness to Experience	degree	
Agreeableness		
Conscientiousness		

Figure 2.2 shows the discrepancy in the literature between the Five Factor Model and alternative personality models. 47 papers were selected as a sample for this analysis on the personality models (Figure 2.2), and tests (Figure 2.3) used in the literature.

Google scholar was used to search for papers, using the keywords in Table 2.1. During the search, one keyword from the personality column was entered alongside one keyword from either the social network column, or the social media column. Papers were first evaluated for relevancy based on their abstract. If personality and social networks (or social media) was the main focus of the paper, it was kept; otherwise, it was skipped. Papers also needed to contain at least one measure of personality, and at least one measure pertaining to social networks or social media. This resulted in an initial selection of 109 papers. After this initial phase, papers were reviewed more carefully,

with only the ones explicitly using measures relating to social network connectivity (degree, transitivity, number of friends, strength of connection, and so on) being kept during this second phase. Most papers gathered during the initial phase dealt with social media use, rather than social network structure. However, some of these papers were kept because they addressed some aspect of social networks, like number of friends for example. This process resulted in a selection of 47 papers, which are used to gain a more quantitative insights into the literature over the course of this chapter. The earliest paper is from 2000 and the oldest from 2016. The focus on online networks has skewed the selection to the last few decades. The Five Factor Model and HEXACO were used as search terms because they are the most widely used and validated personality models. As a result, a lot of the papers use the Five Factor Model of personality as shown in Figure 2.2. The Five Factor Model (Big 5) was overrepresented in our selection of papers despite also using more general terms for personality and a competing model (HEXACO) in our search terms, which only emphasizes how widespread its use is.

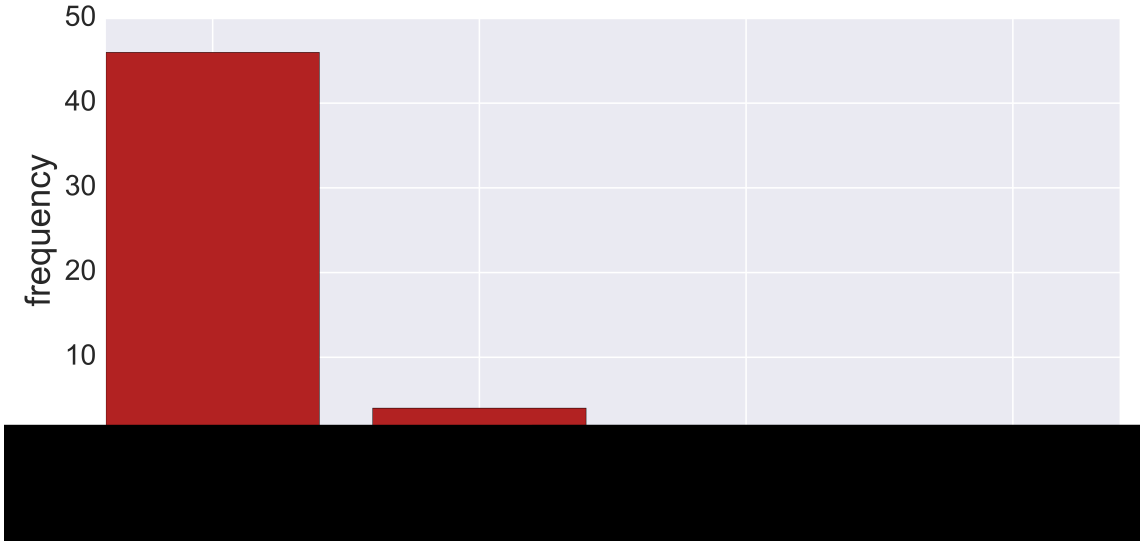


Figure 2.2: Frequency of personality models used in the literature, based on a selection of 47 papers. Acronyms for the personality models are explained in Section 2.3.1.

The papers in Table 2.2 provide a small overview of papers on measuring personality, which gave me a good overview of the two competing models, Big 5 and HEXACO. These papers were selected during a smaller literature review specifically targeting the Big 5 and HEXACO models of personality. The papers were considered important to me, because they were from the people involved in the construction and refinement of the two personality scales, and were frequently used in the introductions of the other papers considered for this chapter, which is also reflected in their citation count. Most papers deal with the Five Factor, but the table also comprises a few papers that highlight the current debate surrounding the use of the HEXACO model instead of the Five Factor Model.

Table 2.2: Key papers on measuring personality and personality scales

papers	Big 5	HEXACO	citation count
McCrae1987	x		6167
Goldberg1990	x		6390
Goldberg1993	x		4669
Digman1990	x		7617
Gosling2003	x		5038
McCrae2005	x		718
Lee2005	x	x	609
Ashton2008		x	336

Several lexical scales to measure the Five Factor Model exist, involving participant response using a 5-point likert scale ranging from strongly disagree to strongly agree. The most extensive scale is NEO-PI-R comprising 240 items [108]. A popular, but shorter version of the scale is the Big Five Inventory (BFI), comprising 44-items [77] which strikes a useful balance between time invested, reliability and validity. Exceptionally short scales with 5-items and 10-items have been developed as well [61]. These still appear to have acceptable reliability, which is useful when time is very constrained. The International Personality Item Pool (IPIP) remains a valuable resource

for researchers, who wish to use the Five Factor Model in their studies [59]. An example scale from the IPIP website, comprised of 50 items, is often used in research, but researchers are free to derive their own scales from the pool.

Figure 2.3 highlights the vast amount of different personality tests used in the literature, based on our selection of 47 papers. According to this figure, the most used personality tests are the 44-item BFI, the 50-item IPIP, and custom-made personality detection algorithms. Both the BFI and IPIP have average length which help their accuracy, while preventing boredom among participants, which can be a problem for the longer tests. The personality detection algorithms are especially prevalent in computer science oriented papers, and vary greatly in the features considered to predict personality. No single established algorithm exists to derive personality from social media features, rather, each study tends to have a bespoke algorithm that works in their particular context. However, the diversity of the personality tests used across the literature can also be a strength. If the same effects can be found with different measures of personality, it only strengthens the validity of these effects. Other models present in Figure 2.3 include the Ten Item Personality Inventory (TIPI), the Adolescent Personal Style Inventory (APSI), the Eysenck Personality Questionnaire (EPQ), the Inventory of Child Individual Differences (ICID), the National Character Survey (NCS), the SHL Occupational Personality Questionnaire (SHL's OPQ), the Big Personality Test, the Eysenck Personality Inventory (EPI), and the HEXACO Personality Inventory (HEXACO PI).

Variation in the level of attention given to different facets in the literature are important to consider. Based on the same selection of 47 papers used previously, we derived the number of significant and non-significant findings for each facet in Figure 2.4. Extraversion and Neuroticism are the two most studied facets in our paper selection, and are often studied together. Unsurprisingly then, most significant effects are found for these two facets, with Extraversion leading in the number of studies finding a significant effect for the facet. Openness to Experience also has a positive ratio in terms of significant effects found compared to non-significant effects, but only barely. On the

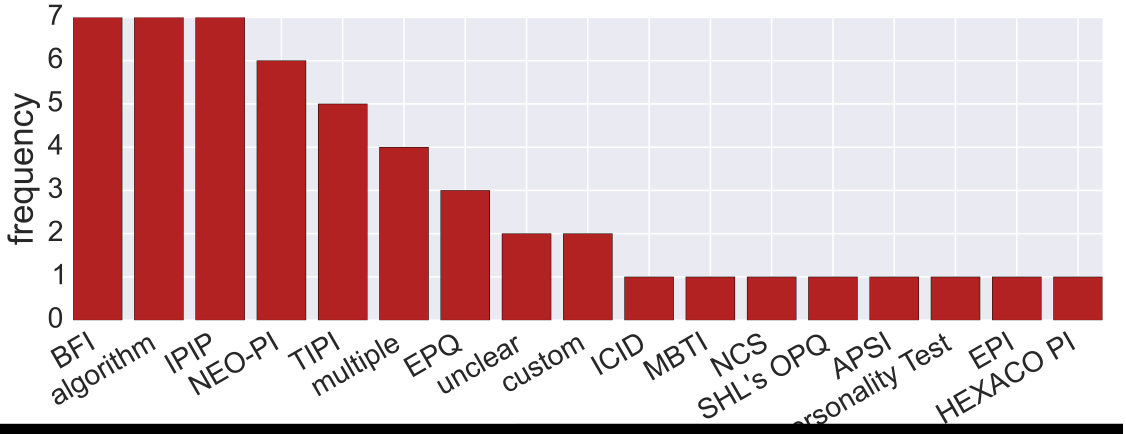


Figure 2.3: Frequency of personality tests used in the literature, based on a selection of 47 papers.

other hand, Conscientiousness and Agreeableness have more non-significant findings than significant ones, which might explain why they are less studied in the literature compared to the other facets.

2.3.2 Personality and online behavior

The Internet has enabled people to expand their social networks beyond offline constraints. Some evidence suggests that people turn to the Internet for social support when it is lacking in their offline networks [120]. By online behavior we understand the navigation habits of users. This relates to the type of sites people frequent, and whether these relate to work, leisure, or social activities. Alongside a more specific focus on social media use, the relationship between personality and Internet use has received much interest in the literature, as outlined below.

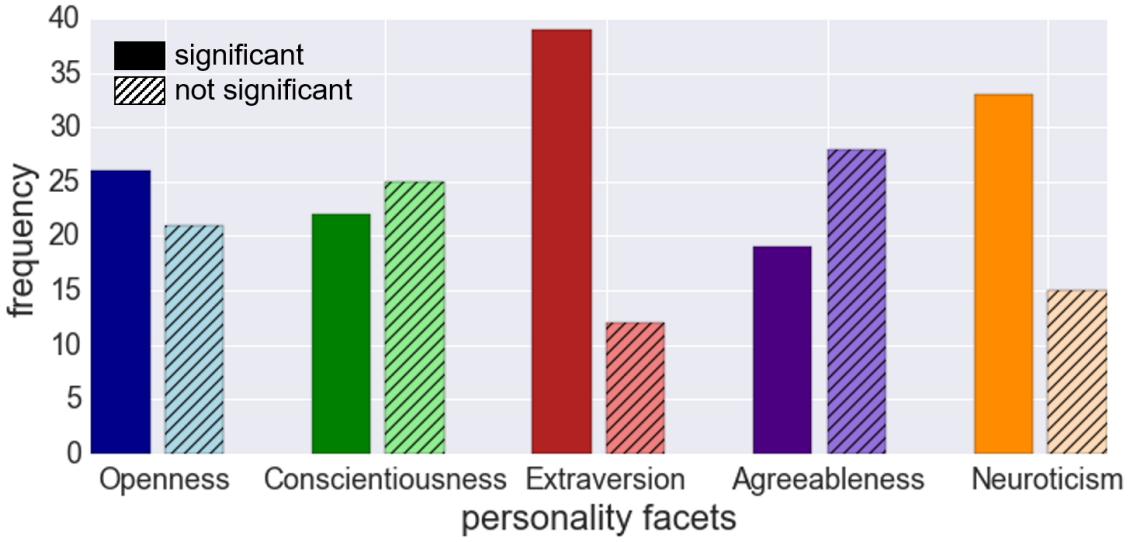


Figure 2.4: Proportion of significant and non-significant findings for each of the facets of the Five Factor Model (Big 5) of personality, based on a selection of 47 papers.

Different Uses of the Internet

Not all personality facets are equally related to all types of Internet use. Landers and colleagues found that Extraversion and Conscientiousness were the only facets to predict overall Internet usage. Even so, Extraversion and Conscientiousness only accounted for 8% of the variance in Internet use. Conscientiousness was found to be negatively correlated with leisure-related Internet use and positively with academic Internet use among adolescents [92]. Extraverts refrain from using the Internet as a substitute for social interactions [6]. They did, however, use the Internet to do research, voice their opinions, and share music [6].

Neuroticism, which has been associated with a lack of perceived social support, also has a negative relationship with Internet use [145], in particular with the leisure facet, which was comprised of online social activities, such as instant messaging and social gaming. Interestingly, Amiel and colleagues also found a negative relationship between instant messaging and Neuroticism. Neurotics were also found to avoid discussion

boards, and showed little interest in participating in them online [6]. Whether the Internet can provide the same quality as face-to-face interactions is still under debate. Internet anonymity promotes self-disclosure, but also deception and misrepresentation. Online interactions have the added disadvantage of lacking the necessary non-verbal cues to convey the correct meaning of a verbal message. Frequent Internet users tend to be introverted if they use the Internet as a source of entertainment [6]. Mitchell and colleagues also highlight that overall Internet use might not be a good predictor for other psychological factors, such as well-being, perceived social support or personality. Rather, heavy uses in specific categories of the Internet, such as entertainment, social networks or information seeking, can be used to predict certain traits or behaviors [114]. In contrast, Extraversion was the only personality facet to have been linked to Internet use at home [69]. None of the other facets showed any significant correlation with any of the Internet uses studied: work, social, home or leisure [69].

Gender differences

Some gender differences have emerged in the relationship between personality and Internet use. Overall, Extraversion is negatively related to loneliness. However, the negative relationship between the Extraversion personality facet and loneliness is mediated by social network size [67]; meaning that introverts are more likely to feel lonely because they have smaller networks than extraverts. Female extraverts especially used the Internet for social communication, while male extraverts used it for leisure activities. Overall, Extraversion was positively associated with a leisurely use of the Internet [67]. Individuals high in Neuroticism, on the other hand, have a greater tendency to use the Internet as a substitute for their social interactions, by expressing a higher desire to belong to groups and avoid loneliness [6]. Neurotics are indeed prone to loneliness, although for a different reason than introverts. Social network size did not mediate the positive relationship between Neuroticism and loneliness. Rather, it seems that neurotic's high negative affect is related to their loneliness. Female neurotics especially

tended to use the Internet for social services [67], suggesting that female neurotics might feel more comfortable communicating with others online. Neurotics also used the Internet to learn about potential threats and to read alternative news [6]. Interestingly, Neuroticism had been found to be negatively associated with Internet use for information services in a previous study [67]. However, this relationship was only significant for males, but not for females.

Online self-disclosure

Most people are more likely to reveal personal information online than in a face-to-face interaction [130], neurotics are especially likely to post accurate personal information on Facebook [5]. Interestingly, a previous study on online behavior and personality found no correlation between any personality facet and information disclosure [135]. However, this study used the 10-item IPIP as a personality measure and relied on self-report for information about online behavior. Neurotics use the Internet to be less lonely [130]. They also appear to feel more comfortable using online communication, especially when their desire to affiliate is high [130]. A perceived lack of social support from other people is characteristic of neurotics, as demonstrated in a communication study [130].

Online services appear to be especially appealing to introverts and neurotics for their anonymity and information control. Online profiles can be rewritten, edited or deleted if not satisfactory. Offline impressions, on the other hand, are hard to change once they have been established. Neurotics and introverts can fulfil their need to belong [17] and other social needs through online interactions, if these are not satisfied through offline ones [5].

2.3.3 Personality and social media use

Most research on Social Networking Site (SNS) have focused on Twitter and Facebook. Twitter is an example of a directed social network in which followers are not necessarily followed back by their followees. Facebook, on the other hand, is composed of undirected social networks, in which people have to accept friend requests before connecting together. Facebook is unlike most social networking sites in that it relies on pre-existing social relationships. Other networks such as Twitter or Tumblr, or more specific SNS such as Pinterest or Deviantart have a more online-to-offline type of relationship formation [130]. An interesting group of social media platforms are LBSN, which have a social component, but focus on location-sharing as their primary activity. Personality has been shown to influence LBSN usage [25, 26] and is of particular interest to this thesis because it incorporates the element of location and geographic distance to the social network, more so than any other social media application. Below, we review the current state of the art on personality and social media use for each personality facet separately.

Conscientiousness

Highly conscientious people are likely to avoid distractions, such as Facebook, and might therefore be less active online [130]. However, other SNS might be more suitable for conscientious users, such as the LBSN Swarm. Swarm is associated with Foursquare, an online website for location recommendations. Swarm's check-in system might be especially favorable for highly conscientious users [26]. This application is relatively less time-consuming compared to Facebook, as the Swarm application only serves to communicate one's location to friends. Interestingly, Conscientiousness was positively associated with influence in a Twitter network, at least when measured with the TIME algorithm [123]. People who score high on the Conscientiousness facet have more friends on Facebook compared to people who score low on the facet [4]. Users also use some Facebook features differently depending on how conscientious

they are. Conscientiousness was negatively correlated with the picture uploading feature [4]. On the other hand, Conscientiousness was positively related to social media use for news and interactions with others [57].

Extraversion

Extraverts were most likely to engage in social activities on Facebook, such as belonging to groups. Interestingly, extraverts have been found to refrain from using the Internet as a substitute for social interactions [6]. Findings from this study also indicate that extraverts do however use the Internet to do research, voice their opinions and share music. Another study found that extraverts do not substitute Facebook for their social life, rather, it complements their offline social activities [130]. This is also in line with a study on online communities and personality. Extraversion was not related to time spent online [135], possibly because introverts are more comfortable with computer-mediated communication. There was, however, a positive correlation between Extraversion and number of friends in online communities [135]. This is also in line with the positive association between high Extraversion and high in-degree in Twitter networks [123]. Extraversion was also positively related to influence on Twitter, albeit only when using the Klout algorithm [123]. Some evidence exists, however, that suggests that extraverts use social media more than introverts [155]. It seems that the nature of the online activity, for example, whether it is social or not, influences extraverts interest in said activity. It has been argued that extraverts have a large circle of friends with relatively weak connections compared to introverts who have fewer friends, but who are more strongly embedded in their social network [139]. In addition, Extraversion is negatively related to loneliness. However, the negative relationship between the Extraversion personality facet and loneliness is mediated by social network size [67]; meaning that introverts are more likely to feel lonely because they have smaller networks than extraverts. Utz found that Extraversion and number of friends influenced perceived popularity of fake social media accounts [150]. Extraver-

ted individuals are perceived as being more popular than introverted individuals, with number of friends further increasing perceived popularity, regardless of Extraversion scores. Interestingly, number of friends plays an important role in social attractiveness for introverted individuals, but not for extraverted ones [150].

An extraverted social media profile was judged as more popular than an introverted one in an experimental setting where participants had to judge the popularity, social attractiveness and communal orientation of bogus social media accounts. Number of friends positively impacted perceived popularity, which was higher when the user profile was manipulated to display a high number of friends (382 friends) compared to a low number (82 friends). Furthermore, the profile was also judged as more popular if these friends appeared to be extraverted, compared to when they appeared to be introverted. When the profile was extraverted, it was judged as having higher communal orientation when the profile's friends were also extraverted, than when they were introverted. There was no effect of Extraversion on communal orientation for the introverted profile condition. For the social attractiveness measure, Extraversion and introversion also had a differential effect. When the profile was introverted, number of friends had a positive effect on perceived social attractiveness. On the other hand, when the profile was extraverted, number of friends did not significantly impact social attractiveness.

Some evidence from Facebook supports these findings [4]. Introverts posted and shared less on Facebook than extraverts, however, when they did, they garnered more interest from their friends than extraverts did, as exemplified by the larger number of likes and comments on their posts [4]. On communication boards, extraverts tended to make twice as many post as introverts did; conscientious users also posted twice as much as unconscientious individuals [12]. Extraverted individual might enjoy communicating and socializing with others, while conscientious users might be more thorough in keeping up with an ongoing discussion, while their more unconscientious counterparts are more likely to get distracted. However, these explanations remain merely speculations, as the content of the posts have not been analysed in this study [12]. Sharing of

personal information, on the other hand, did not seem to be correlated with the Extraversion facet, which appears to be a behavior more closely related to the Neuroticism facet [4].

Neuroticism

In contrast to extraverts, neurotic individuals are more concerned with their online portrayal. This is in line with research that confirmed that extraverts construe their "real-me" through offline social interactions, while introverts and neurotics construe their "real-me" through their online behavior and relationships [5]. In line with this idea, individuals high in Neuroticism are most likely to portray themselves accurately online and to give accurate personal information on their profiles [4]. Monitoring wall posts was an important part of Facebook use for neurotic users [130]. Emotionally stable users, in contrast, preferred the photo sharing aspect of Facebook [130]. Emotionally stable people have a tendency to use social media less overall [57], and neurotic users tend to use social media to share longer posts with more negative sentiments and subjective words. Neurotic Facebook users are also more successful in attracting social support, as evidenced through the large number of comments and likes they get on their posts [139].

Neurotics are also expected to have less friends and be less popular, even online, but evidence is scarce. Low Neuroticism is positively correlated with popularity on Twitter, in terms of in-degree [123], but high Neuroticism did not have any particular effect on either in- or out-degree. Highly neurotic individuals are unpopular online interaction partners. While homophily was found for most personality facets in this communication study, only 4% of neurotics were willing to communicate with fellow neurotics [12]. High Neuroticism seemed to trigger the exact opposite effect of homophily, as most neurotics appeared to actively avoid each other in this communication study. In contrast, emotionally stable users did not seem to have any preference to communicate amongst themselves, they preferred agreeable interaction partners, but were not

avoiding each other either [12].

Openness to Experience

Openness to Experience was positively related to time spent on online communities and number of friends in those communities [135]. Open individuals are also more likely to be sociable through Facebook [130]. It was also found that highly open users tend to use more features of Facebook than others [4]. However, and contrary to expectations, they did not have a deeper knowledge of Facebook functionalities [130, 4]. An explanation for this might be that broader interests also mean less time invested in each one of them, including how Facebook works [130]. Interestingly, and to support this line of reasoning, Openness to Experience was positively correlated with being on Twitter users' reading lists [123]. Reading lists show recommended users to follow on Twitter, suggesting that open users produce interesting and original content. Openness to Experience was positively related to frequency of social media use, but negatively for news and interactions with others [57]. Research from spatial homophily suggests that Openness to Experience might be the strongest predictor of homophilous connections in an LBSN such as Foursquare [79]. Open people were happiest in neighbourhoods with fellow open residents [79]. Openness to Experience was positively correlated to number of sociable and popular venues visited in Foursquare [26].

Agreeableness

Agreeableness did not seem to be related to Facebook use at first, neither in a positive nor in a negative way [130]. However a subsequent review paper found that Agreeableness was positively associated with social media use frequency, but only for news and interactions with other users [57]. This personality trait was also not related to number of friends, contrary to expectation [135]. Highly agreeable people are popular communication partners, especially with open and emotionally stable users [12]. However, few correlations exist between Agreeableness and other social media metrics. For

instance, agreeable users are not more likely to have more friends on Facebook or other online communities than disagreeable users [4, 135, 155]. Agreeableness also did not correlate with time spent online in general, or on social media in particular [4, 135].

2.3.4 Personality and social network structure

Personality affects the way we connect with others, which in turn affects the structure of our social network, as mentioned previously in Section 2.1.2. Social network structure has been used in the diagnosis of personality disorders, highlighting the important relationship between them [28, 29].

The following social network measures are used in the upcoming paragraphs:

1. **network size (or degree)**: the number of connections an individual has in a network
2. **neighbour**: all nodes directly connected to a node A are considered its neighbour
3. **clustering**: measures the degree to which nodes tend to cluster together.
4. **shortest path**: the optimised path between two nodes so that the sum of edges it passes through is minimised.
5. **betweenness centrality**: centrality measure of a node based on the number of shortest path that go through said node.
6. **transitivity**: the extent to which nodes are connected if they share a common neighbour.
7. **Brokerage**: node-based measure that defines the extent to which a node acts as a bridge in the network. Bridges are nodes in a position where they connect otherwise disconnected (groups of) nodes.

Personality and network size

The positive relationship between Extraversion and network size is well-documented [4, 52, 103, 139]. Extraversion was significantly related to degree (network size), in three online [4, 116, 99] and two offline studies [138, 122], and only one study that did not find a significant relationship between Extraversion and network size [130]. Overall, there exists solid evidence for a positive relationship between Extraversion and network size across a range of offline settings: among adolescent pupils [138] and among a more general sample with an average age of 28 years [122], across three layers of their social network: their support group, sympathy group, and the outer layer. Interestingly, a later study replicated the positive relationship between layer size and Extraversion, but only for the support group, not the sympathy group, while the effect on the outer layer was not measured. Additionally, Emotionality and Openness to Experience both had a positive effect on support group size [115]. It is interesting to note that this study [115] used the HEXACO model of personality instead of the Five Factor Model. Also of interest is that the online studies discussed here all used Facebook samples. Facebook is the most prominent example of an online undirected social network based on real life connections, which makes is the social media of choice for network studies.

Extraverted individuals are motivated to engage in social interactions with others, and often are the ones to initiate relationships [138]. It has been suggested that extraverts have indeed more friends than introverts, but introverts are emotionally closer to their friends, while extraverts tend to have less emotionally close social relationships [122]. It is indeed the case that while extraverts have more friends, they are not necessarily emotionally closer to them [129]. It has even been suggested that extraverts consider others as friends more readily, sometimes confusing acquaintanceship for friendship and therefore report a higher number of social connections [7].

Social networks tend to be subject to an Extraversion bias, meaning that extraverts are over-represented, while introverts are under-represented. Interestingly, this bias is es-

pecially strong in the social networks of extraverts and much weaker in the networks of introverts, probably because of a homophily effect. Friends tend to connect with others of similar Extraversion levels [50]. However, Neuroticism has also been associated with number of friends, although this effect is much less consistent in the literature. This arises from a communication study that identified neurotic users as unpopular interaction partners in an online discussion board [12], especially between fellow high scorers on the facet. Network size was also a significant feature to predict Neuroticism in Markovikj et al's personality prediction model [103]. Another study on an online university student network found a small positive relationship between Neuroticism and network size [155]. It is difficult to tell whether Neuroticism is rarely studied in the context of network size or if other studies simply did not report null results.

Transitivity and brokerage

Transitivity is an interesting connectivity feature in social networks, as it is an indication of the embeddedness of a node in its network. In particular, sociable individuals tend to be at the centre of large, loosely connected networks. This would explain why Openness to Experience and Extraversion have previously been negatively associated with transitivity [83, 96, 99]. However, an older study using an online student network did not find a significant relationship between Openness to Experience and transitivity [155]. Differences in networks studied, personality tests used, and methodologies applied could explain these contradictory results. Extraversion, on the other hand, was not directly related to transitivity in the Facebook sample. The significant effect found could be entirely explained through network size [99]. This relationship between transitivity and Extraversion was also found to be significant in the online student network, however, it is unclear if degree was controlled for during the statistical analysis [155].

Brokerage is a measure that quantifies the bridging power of a person in a network. A person with high brokerage can control the flow of information between two otherwise unconnected individuals. This can be a powerful position in an information-based

network especially [48]. However, this broker position can also be stress-inducing, as the cohesion of the network relies on one person. In line with this observation, Neuroticism has been associated with brokerage. In two offline studies, neurotic people have been found to occupy broker positions significantly more often than less neurotic people [83, 82].

Summary of current literature

Personality in general has been shown to be significantly related to social network structure, although results remain sparse in the literature. Extraversion, Openness to Experience, and Neuroticism have been found especially to have significant effects on network structure. Conscientiousness and Agreeableness have received much less attention in general as evidenced in Figure 2.4. This might lead to these two facets to be understudied, but potentially also under-reported in the literature if no significant effects are found. Additionally, some differences emerged between online and offline networks, see Figure 2.5. The evidence for other characteristics, like brokerage and transitivity is much more sparse and inconclusive.

2.4 Homophily

Homophily, the tendency of similar people to connect, is closely related to personality in social networks. As personality drives connections, it becomes logical to consider its influence as an attractive force between people. Homophily is ubiquitous across nature [53], even among mammals [104] and fish [37, 36]. Humans especially tend to befriend others who are similar, a phenomenon referred to as homophily. In social networks, this homophily effect is particularly pronounced. Humans derive similarity with others from a large range of characteristics and interests [109]. Age, gender, religion, ethnicity, political affiliation, education are all examples of potentially homophilous characteristics. Friendships are even subject to genetic homophily: friends tend to be

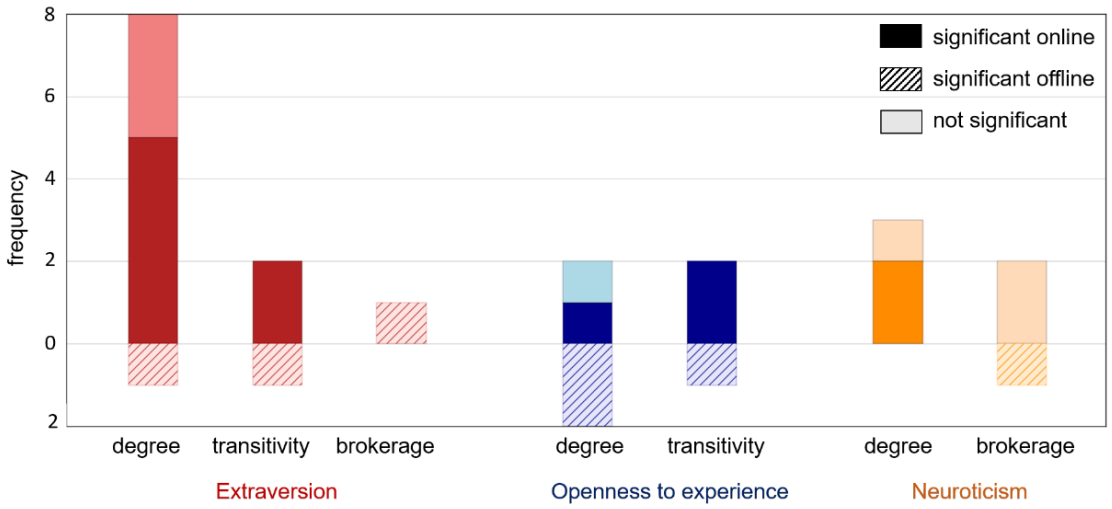


Figure 2.5: Proportion of significant and non-significant findings in the literature for Extraversion, Openness to experience, and Neuroticism, based on a selection of 47 papers.

genetically more similar than non-friends [19]. Similarity induces trust and cooperation, which might explain that people choose friends who are similar to themselves, even down to their genetic make-up

2.4.1 Homophily in social networks

Homophily plays a role in a diverse range of contexts and situations, such as friendship formation, social contagion, the evolution of cooperation; but also more negative aspects, such as social segregation and cultural polarization [53]. While homophily promotes synergy, heterophily focuses on specialisation. Both have their advantages, with homophily promoting cooperation and trust, and heterophily allowing for people to develop specialised skills. In a series of simulation experiments, Fu and colleagues found that homophily tends to evolve even if heterophily is more beneficial [53]. Strong homophily emerges especially when mutation rates are low. Provided pay-off for homophilic interactions are positive, all homophilic individuals, even those weakly preferring heterophily, do better than pure heterophiles. When mobility was modelled into

the experiment, which increased the likelihood of individuals that had interacted before to interact again, homophily still did better compared to heterophily [53]. Homophily emerges under a wide variety of conditions. Fu and colleagues speculate that language could be a strong homophilic driver. The authors also suggest that homophily is beneficial for individuals who have started interacting more and more with unrelated others. Similarity gives the illusion of kin interactions, which are perceived as safer and more trustworthy [53].

Rivera and colleagues [127] propose three main influences in the formation of social ties: assortment, initial network position, and proximity (both in time and space). Why do people prefer to form ties with similar others? People expect similar others to be more likely to accept them, to be trustworthy, and to hold similar beliefs [127]. This reduces maintenance costs, conflicts, and misunderstandings, and ultimately contributes to a long-lasting and healthy relationship. It is no accident that homophily is strongest among intimate pairs and close friendships [127]. However, just because people are similar in one aspect, does not mean they are similar in all aspects. Balance between homophily and compatibility is key. For example, best friends tend to be strongly homophilous on gender, while intimate pairs tend to be compatible instead. However, overall, both types of relationships might be homophilous in some other aspect, like shared interests for example. Social ties tend to be more homophilous than any other form of ties, however. Organizational ties tend to greatly benefit from heterophily, as they enable the collaboration between organisations or companies with different expertise and skills. It is also important to note that every relationship is partly homophilous and partly heterogeneous It therefore makes sense to focus on homophilous and heterophilous characteristics separately, rather than qualifying a relationship as homophilous overall [127].

2.4.2 Personality homophily

Personality homophily is less well studied in the literature than other homophilous characteristics. This thesis aims to fill the current knowledge gap on personality homophily. However, some studies have addressed personality homophily and they are outlined below. As previously discussed, people have a tendency to chose friends who are similar in personality in offline networks [138]. Personality was also found to be homophilous in an online setting [12]. Selfhout and colleagues [138] used an offline social network of adolescent school students, Balmaceda and colleagues [12] derived their data from interactions on an online discussion board. Across these two different contexts, three personality facets emerged as consistently homophilous: Openness to Experience, Extraversion, and Agreeableness. Personality preferences have been explored in communication partner choice [12]. Balmaceda and colleagues found that users with similar personality types were more likely to communicate with one another, especially compared to users with an opposite personality type. For instance, extraverted, agreeable and open users were respectively 80%, 75% and 96% more likely to communicate with a similar user, compared to an opposite user (respectively introverted, disagreeable, non-open). Interestingly, neurotic users were not so likely to communicate with fellow neurotic users (6%). Emotionally stable users' likelihood to communicate with similar others was close to chance levels (58%). Similarly, conscientious users were not as likely to communicate with other conscientious users 42%. Interestingly, extraverted, open, agreeable and emotionally stable users were also more likely to communicate amongst each other [12]. It must be noted that in both samples, the definition of homophily was operationalised differently. Selfhout and colleagues [138] specifically looked at friend selection over a period of time, while Balmaceda and colleagues [12] measured homophily by calculating whether discussions between board members happened between members with similar personality traits or dissimilar traits on each facet.

An interesting phenomenon to consider is the difference between actual and perceived

similarity. In general, people perceive their friends as being more similar to themselves than they actually are. Additionally, perceived similarity has a stronger effect on tie formation online, compared to offline. This is probably due to the fact that people online have less information about their interaction partner compared to face-to-face interactions, especially the absence of non-verbal cues [7]. Furthermore, people expect others to have similar friends [150]. People are perceived as more popular if their friends are mostly perceived as extraverted than if they are perceived as introverted. People were also perceived as socially more attractive when the number of friends matched their perceived personality: extraverts are expected to have a lot of friends, while introverts were expected to have a smaller number of friends [150].

2.4.3 Spatial homophily

As online social networks became popular and could be accessed via smartphones, LBSN networks such as Foursquare and Gowalla started to emerge. Spatial homophily has increased in importance as a result, as LBSN rely on recommendation systems to suggest new locations to customers. Spatial homophily refers to the tendency of similar individuals to assort at common locations [140].

Two drivers behind homophily are peer influence and social selection. In spatial homophily, peer influence is the result of existing ties that compel people to visit the same locations as their friends. Social selection is an antecedent of friendship ties. People visit the same locations because they are similar, which might lead them to establish friendship ties, based on that similarity [163]. Zhang and colleagues used spatial data from the LBSN Gowalla to investigate the effects of peer influence and social selection on spatial homophily [163]. Globally, peer influence did not have any effect on spatial homophily. However, local peer influence did. Friends who were co-located indeed influenced each other in the venues they visited, especially for sociable venues, such as "food" locations (73% of similarity between Gowalla users explained by local peer influence) and "coffee shop" locations (87% of similarity between Gowalla users ex-

plained by local peer influence). Non-sociable venues, such as "airports" were much less prone to local peer influence (only 8% of similarity between users explained by local peer influence). Social selection showed that some venues attract more diverse crowd, while others attract more specialized crowds. General venues through which a variety of people pass through, such as airports, are characterized by a diverse crowd, as exemplified by their low clustering score and high network size (also referred to as degree) [163]. On the other hand, venues like small coffee shops or restaurants attract more specialized crowds, as exemplified by their high clustering, but low degree [163].

2.5 Geographic distance in social networks

The probability of two people being connected in a friendship-based network is well-known to decrease with distance [97, 91, 11, 14, 40]. Backstrom and colleagues [11] demonstrated that the relationship between distance and connection probability was stronger for medium to large distance, but much weaker at short distances (below 50 miles approximately). Distance also appears to matter less when people are close enough to easily be able to travel to see each other face-to-face, or engage in activities together [7]. Interestingly, frequency of mobile and online communication decreases with distance [93], but the duration of phone calls appears to increase and then level out after 60km [91]. More generally, it is proposed [91] that communication networks appear to have two main levels: short-distance communication has high clustering, but is of short duration, while long distance communication has smaller clustering, but tends to last longer.

2.5.1 Distance as a constraint

Proximity is also an important constraint in social relationships [133]; it is therefore of interest to look at the spatial component of social networks in our study, which have

become more relevant to the wider population through location-based social networks [62, 35, 140]. Proximity, both social and geographic, is essential for friendships to form [90]. In turn, friendships can influence the social venues we frequent [163]. We refer to spatial networks as any network whose nodes are connected based on a geographic or spatial unit. The spatial self has emerged as a new form of identity in online social networks, most of which have location-sharing functionalities [136]. Dedicated Location-based social networks (LBSNs), such as Foursquare or Gowalla are spatial networks with a social component, in which people share their location with friends [136]. Purely spatial networks are interesting in that they tend to be neither assortative, nor disassortative. This makes them stand apart from social networks, that tend to be assortative, and technological networks, which tend to be disassortative [14]. In LBSNs, which have a social component, strong local clustering is observed instead [163]. Spatial constraints modify the graph characteristics of social networks, for example, spatial constraints restrict the appearance of large degrees [14]. Spatial constraints also lead to large fluctuations of betweenness centrality. Hubs are usually central, but get even closer to the centre of gravity of all points in a spatial network [14].

2.5.2 Location Based Social Networks and people's spatial identities

A person's location history is an increasingly important part of their online identity [136]. On a large scale, research has shown that entire regions of the UK could be characterized by the average personality profile of its residents [126]. As a result, Wales could be qualified as "introverted" and "neurotic", while Scotland was found to be "agreeable" and "emotionally stable" [126]. Similarly, neighbourhoods in London have also been identified according to the personalities of its residents [79]. Residents in the centre of London scored, on average, high on openness, but low on Agreeableness. Residents of more peripheral neighbourhoods were low in Neuroticism, but high

in Conscientiousness [79]. Personality homophily has been shown to mitigate the effect of location on life satisfaction [79]. Open individuals were happiest in neighbourhoods with a high prevalence of similar others, while for other personality facets, life satisfaction in a particular location was not affected by the presence of similar others [79].

The mobility patterns of a city's residents have previously been used to derive the overall character of the city in question [35]. People who shared similar interests had a tendency to visit the same network of venues spread across the city, rather than staying confined to their respective neighbourhoods [35]. Furthermore, people are able to derive personal characteristics of others, based on the places they visit [62]. Users of the location-based social networking application Foursquare were able to predict the personalities of patrons based on pictures of the locations they had visited with 69% accuracy [62]. On the other hand, location ambiance and typical activities carried out by visitors proved harder to predict, with only 32% and 33% accuracy respectively [62].

Personality has been shown to influence the use of Foursquare, a Location-Based Social Network (LBSN). Chorley and colleagues found that Conscientiousness was positively related to number of venues visited, while Neuroticism was negatively related to it [25]. Personality appears to be directly linked to our mobility pattern presented through LBSN applications. The question that arises from these findings is whether venues attract people with certain personality traits. This thesis will endeavour to answer this question by analysing the visiting patterns of users with similar personalities.

2.5.3 Applications

Most recommendation systems, whether for movies, clothes, or activities, rely on homophily to predict what a potential consumer might like. This works in two different ways. First, a consumer might like similar products to the ones they have already pur-

chased or consumed. Second, people who are similar, might like the same types of products. Finally, we can also find effects of brand personality on brand loyalty and trust. People might be attracted to brands that match their personality. Brand personality is an emerging field of study, which has developed several models of personality to fit brands [162, 100]. Some brand personality measures have focused on exclusively attributing human characteristics to brands in order to make them more comparable [56]. Others have argued that brand personality and human personality are not comparable, since the former is a concept and does not act like humans do [144]. These studies tend to use measures that involve non-human attributes such as Ruggedness, or much less basic qualities, such as Sophistication [1]. Consumers are likely to like brands with similar personalities, even if the type of product is very different.

It is also of interest to study how humans form and maintain connections online. Social media has become ubiquitous in the developed world. Understanding how online social networks work, especially in contrast to offline networks will contribute to a better understanding of modern humans sociality and social media use. Brand personality is an interesting and exciting field of research, which can benefit many companies by adapting their brand personality to attract their target audience.

Social networks and social media use have been linked to personality pathology [28, 29]. Social network position can help identify potential psychological ailments in patients. Different personalities deal with social media differently, with some more prone to use it than others. Conscientiousness and Extraversion protect against social media use; the former seeing social media as a distraction, and the latter preferring meeting people in person over it. Neuroticism and Introversion can have an interesting relationship with online social media, with the former relying on social media for support and validation, and the latter being able to express their true self more comfortably online than offline [5].

2.6 Motivation and knowledge gap

In this thesis, the aim is to primarily investigate personality homophily in the context of online social networks, namely whether people with similar personalities have a higher tendency to become friends. While investigating personality homophily, we additionally check for distance as a potentially moderating factor. Distance represents one of the main constraints on network structure and tie formation. Furthermore, distance represents a typical example of induced homophily, while personality represents a typical example of choice homophily.

The knowledge gap we aim to fill in this thesis spans across two different topics: personality homophily and geographic distance. Both topics influence the structure of social networks, and in our case, that of online social networks. While personality homophily drives people together and fosters connections between them, geographic distance has the potential to keep people apart and break up connections. By looking at both factors together, we aim to gain a better and more complete understanding of the structure of online social networks. The gaps we aim to fill are as follows:

The role of personality homophily in online social networks: One of the primary aims of the thesis is to complement the knowledge on personality homophily in online social networks. None of the studies carried out so far have used a representative online social network such as Facebook to investigate this phenomenon. Facebook is based on offline networks and mimics its mechanisms, which makes it the social networking site of choice for studies on online networks.

Homophily and geographic distance: One aspect that none of the personality homophily studies mentioned so far [101, 138, 12, 50] take into account are the geographic constraints and their potential influence on homophilous tie formation. We are especially interested in the relationship between distance and personality homophily, and how this influences connections and social network structure. Personality has been shown to affect the way we stay in touch with others at longer distances [72]. This

indicates that distance could have a potential moderating effect on personality homophily. This thesis will address this relationship between personality homophily and distance at a local scale using Foursquare data, but also at a global scale using Facebook data.

Locations as homophilous hubs: The relationship between personality and location has been studied at different levels: countries, cities, and neighbourhoods [126, 79, 35]. However, few have considered the venues we frequent [25, 26]. None have studied venues as a potential attractive force for people with the same personality. People constantly choose one venue over the other when going shopping, or out to eat. We investigate the role personality plays in venue selection, and whether personality homophily drives venue choices.

The aim of this thesis is to fill the knowledge gap outlined above by answering the questions stated at the end of the introduction:

1. What is the relationship between personality homophily and the structure of an online social network? (Chapter 3)
2. What is the relationship between personality homophily and the geographic distance between people of an online social network? (Chapter 4)
3. Are people who are similar more likely to be attracted to the same venues? (Chapter 5)

Chapter 3

Personality Homophily in Facebook

Parts of this chapter have been presented at the peer-reviewed conference ASONAM 2016 and published in its proceedings: N. Noë, R.M. Whitaker, and S.M. Allen. Personality Homophily and the Local Network Characteristics of Facebook Personality Homophily and the Local Network Characteristics of Facebook. In: *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 2016.

3.1 Introduction

In this chapter we explore the relationship between personality and online social network structure with the aim of answering the following research question: *how does personality homophily affect the structure of an online social network?*

While taking into account demographics of age and gender, we assess the extent to which personality homophily is evident in the local network features of Facebook. More specifically, we look at the relationship between different personality facets and common network characteristics, such as degree and transitivity. Furthermore, we explore the effects of personality similarity on the connections between users. We look at different measures of connection to gain a clear understanding of the personality homophily effects at work in an online social network. We first look at the proportion of similar pairs and mixed pairs among connected users, and compare these frequencies

to those of unconnected users. The second approach focuses on strength of connectedness between users, while the third approach focuses on whether a triad of users are all interconnected (transitive) or only partly (intransitive). We use regression models and Welch's t-test to test for statistical significance. These analyses are carried out in a sample network of one of the most widely used social media platforms to date: Facebook.

We consider both a global view of a network of connected Facebook users through their overall degree and transitivity scores, but also their local connections in the form of pairwise relationships and triads.

The data used in this chapter has been obtained through an agreement with the psychometrics group at Cambridge Judge Business School [1]. The MyPersonality data is a collection of datasets obtained from Facebook users with their explicit consent, through a personality questionnaire on Facebook. Users were aware that, if they decided to share their data, it would be used for academic research purposes only. The MyPersonality data has been a valuable academic resource for researchers across the world interested in personality and online social networks. More information on the MyPersonality datasets used in this chapter are provided in Section 3.2.2

In order to answer the research question posed at the beginning, the chapter has been broken down into the following parts:

1. In section 3.1.1, we elaborate on Facebook, as the social media platform on which the dataset used in this chapter is based.

2. In section 3.1.3, we summarise two different sets of hypotheses, one pertaining to network characteristics, and the other to homophily.

3. In section 3.2, we elaborate on the characteristics of our dataset, which is a subset of the MyPersonality data.

[1] https://www.psychometrics.cam.ac.uk/productsservices/mypersonality

4. In section 3.2.3, we define each of the variables obtained or derived from the MyPersonality data, and any equations we used during that process.

5. Planned analyses are outlined in section 3.2.5. Regressions, Welch's t-test, and chi-square tests were used to investigate the hypotheses outlined in section 3.1.3.

6. In section 3.3, we present the results pertaining to each of our hypotheses defined in section 3.1.3. We first present effects relating to age and gender, we then go into more detail regarding degree, transitivity, and finally, homophily.

7. In section 3.4, we assess the support found for each hypothesis, alternative explanations for the findings, and any limitations pertaining to this chapter.

8. In section 3.5, we summarise the main findings of the chapter.

In regards to the findings, we add valuable insight into how individuals are more likely to cluster at the sub-network level, which contributes to the overall structure of the network. Identifying sub-network patterns and how pairwise connections are more likely to form complements research on global network assortment.

3.1.1 Characteristics of Facebook

Facebook is one of the most used social media platforms in existence today and has grown exponentially since it opened to the public in 2005 [151]. The idea of Facebook is to connect with friends using one's real name and identity. This distinguishes it from online discussion forums, like Reddit, which are usually anonymous. Connections are bi-directional, meaning that a friend request must be accepted by the other party before information is shared between the two Facebook profiles. The main directive of Facebook is for a person to share their life experiences with their friends through the numerous options available to them on the platform: photos, videos, posts, groups, events. Facebook is the most used social media platform to date with 1.15 billion active daily users. On average, people spend about 20 minutes on Facebook and 50% of

18-24 year olds go on Facebook first thing in the morning. Social media appears to be an important part of young peoples lives especially [119]. There is a slight gender disparity in Facebook usage: 76% of women and 66% of men are on Facebook [119]. This gender disparity has also lead to the sample presented in this chapter being skewed towards female participants. However, we check for gender effects to account for this imbalance. Facebook has an offline-to-online dynamic, meaning that people usually know each other offline before connecting on Facebook. In a study on the composition of Facebook friends, only 4% were classified as strangers and 2% as online-only. The biggest group was acquaintances with 27%, followed by activity (people met during leisure, work, or study activities) with 24%. Close friendships comprised 21% and maintained (past relationships and friendships) represented 18%. 4% were unaccounted for [102]. For 400 friends, this corresponds to 80 close connections, which would be in the range of one's active network (about 50 connections) [68]. Maintaining one's active network through Facebook makes sense: people even at a distance can be reached and frequent low effort contact, such as wall posts, make it easy to maintain these relationships. Such Facebook posts, especially of important life events, can help keep one's active network involved in one's life, without having to go through the effort of calling or seeing each person individually. While online interactions are not as effective as offline face-to-face ones to maintain ties, they can still help keep someone "present" in other people's mind.

3.1.2 Motivation for Hypotheses

Previous literature extensively covers the influence of an individual's personality on network characteristics, but the influence of personality homophily has not been well addressed. Regarding individual personality, not all effects appear to be robust across replications with different samples, methods and social networking platforms. Several studies have used network characteristics to predict personality [143], or have used Facebook data, such as status updates, for predictive models of personality [103, 49].

Automatic classification systems of personality traits based on network characteristics have also been developed [96]. Other studies have specifically focused on the interplay between personality and network characteristics [85, 83, 99, 52]. For example, the correlation between Extraversion and network size has been repeatedly supported across different networking platforms (i.e., friends on Facebook [4, 52, 103, 139], online communities [135], or followers on Twitter: [123]).

The explosion of online social networks such as Facebook has allowed new ways in which relationships can be maintained, providing insight into fundamental human behaviour [160]. From this perspective, personality has been shown as influential in online settings, especially in how people use social media [31, 130, 4, 112, 26] and the Internet [67, 6, 92] to the extent that it is possible to predict user's personality traits based on their online interactions [88, 86].

We develop hypotheses based on the existing literature, initially considering relationships between individual personality traits and network characteristics, *with hypotheses italicised*.

Network size

Starting with Extraversion and network size, it is prudent to confirm whether or not our dataset reflects the positive correlation evident in previous work [4, 52, 103, 139] (Hypothesis 1b). These results have been obtained in a range of online networks, but each social media platform and dataset is different. Replicating the existing results will give us confidence that the dataset is similar to other social networks studied previously. We extend the existing findings by considering other personality facets, and *hypothesise that there is a basis for Neuroticism to have a negative effect on network size (Hypothesis 1b)*. This arises from a communication study that identified neurotic users as unpopular interaction partners in an online discussion board [12], especially between fellow high scorers on the facet. Network size was also a significant feature to predict Neuroticism in Markovikj et al's personality prediction model [103].

Transitivity

Transitivity measures the embededness of a person in their network. Openness and Extraversion have previously been negatively associated with transitivity [83, 96, 99]. A negative effect of Extraversion on transitivity can possibly be explained through the associated higher degree of nodes while the negative effect of Openness to Experience on transitivity appears to be only significant for men [99]. This study had a smaller sample size ($4,305$ nodes compared to $313,669$ nodes in this dataset) and therefore we reconsider whether Openness to Experience and Extraversion are negatively related to transitivity for our case study (Hypothesis 1d). As in [99], we control for the possible effects of gender and degree, hypothesising that *the effects of Openness to Experience on transitivity are moderated by gender, while effects of Extraversion on transitivity are mediated by degree (Hypothesis 1e).*

Effects of age and gender

Gender plays a significant role in social media use, with females and males using Facebook functions differently depending on their personality [31, 4]. In a study on social media use and identity, extraverts were most likely to construe their offline self as their “real-me", while neurotic users were more likely to construe their online self as their “real-me" [67]. The effect for Extraversion was significant for both males and females, while the effect for neurotic users was driven by female users [67].

Age is another demographic factor that might confound effects of personality. Personality is assumed to remain relatively stable over the lifespan, but some variation occurs. For example, Neuroticism has been found to decrease with age in women, while Agreeableness and Conscientiousness have a tendency to increase [107, 142]. Another study found differences in all Big five facets over the lifespan [128]. We therefore take into account age as a possible confounding factor in our sample, as the MyPersonality sample is quite young, with a mean age of 24. In fact age might also

influence network characteristics, as number of friends on Facebook have been shown to decrease as age increases [52]. Additionally, age and gender have been found to be homophilous in a mobile network study [42]. This same study also highlighted the tendency of young people to expand their social networks, while older people tended to focus on maintaining fewer, but closer ties. Finally, younger people tend to form social connections with both same and opposite gender friends, especially during their reproductive active years. Older people tend to focus on maintaining same-gender ties instead. Gender and age are therefore important control factors that should be taken into account when studying social networks online and offline, considering the gender differences that exist in how social media is used [4], the age-related differences in personality scores [142], and how gender affects fundamental network characteristics, such as transitivity [99]. We therefore hypothesise that differences exist in personality scores for both gender and age (Hypothesis 1a) *and have the potential to moderate personality homophily effects* (Hypothesis 2e). Age and gender indeed both emerge as homophilous factors in previous research [109]. Individuals are more likely to befriend others of the same gender and who are of a similar age.

Homophily

The strongest homophily effects are observed for race and ethnicity, followed by age, religion, education, occupation and gender [109]. Additionally, other factors have also been identified as homophilous in online social networks, such as music taste [71], residential location [71], or simply, overall perceived similarity [7]. Even psychological dispositions, such as happiness or loneliness, have been found to be homophilous in social networks [109, 20].

Notably fewer insights have emerged on personality homophily, although it has been established for behaviour in communication networks. Users high on Extraversion, Agreeableness and Openness prefer to communicate with similar others on an online discussion board [12]. Interestingly, these same personality facets have been found

to be homophilous in offline friendship networks as well [138]. Both studies also found that extraverts were more likely to initiate contact, while agreeable people were the most popular candidates for friendship requests [138] and as online interaction partners [12]. An interesting effect was observed for Neuroticism, which proved to be disassortative in the communication-based network [12].

Based on these observations, *we hypothesise that Facebook users with similar personalities are more likely to be connected than people who have different personalities (Hypothesis 2b)*. Although there are considerable differences in sampling (size and generality), it is reasonable to expect some consistency with [12, 138]. In addition to considering connectedness as a binary variable (connected or not-connected), we consider the strength of the relationship instead. As we have no qualitative information on the actual relationships between Facebook users, we use shared number of friends as a proxy for connectedness. People with many mutual friends tend to be closer than those who have relatively fewer friends in common. Previous work has also highlighted that homophily is stronger the closer two people are. Indeed, homophily has been shown to be strongest among couples, followed by close friends [161].

We also consider an additional measure of connectedness: triangle closure. If similarity in personality indeed predicts connectedness between two nodes, we also expect nodes in closed (transitive) triangles, in which all nodes are directly connected, to be more similar than in open (intransitive) triangles, in which two nodes are only indirectly connected through a broker. We hypothesise that personality similarity is higher in closed triangles than for open triangles for Openness, Extraversion and Agreeableness (Hypothesis 2c). Additionally, we will explore whether similarity in Conscientiousness is related to triangle closure and strength of connectedness. We keep the hypothesis two-tailed because of the lack of prior support for this facet (Hypothesis 2d).

3.1.3 Summary of Hypotheses

1. **Network characteristics and demographics**
   (a) Personality scores differ between genders and across age.
   (b) Extraversion is positively related to network size
   (c) Neuroticism is negatively related to network size.
   (d) Openness to Experience and Extraversion are negatively related to transitivity.
   (e) Effects of Openness to Experience on transitivity are moderated by gender, while effects of Extraversion on transitivity are mediated by network size.
2. **Personality Homophily**
   (a) Users with similar personalities are more likely to be connected than people who have different personalities.
   (b) Similarity in Openness to Experience, Agreeableness, and Extraversion is positively related to strength of connection.
   (c) Personality similarity is higher in closed triangles than in open triangles for Openness, Extraversion and Agreeableness.
   (d) Personality similarity differs for closed and open triangles, and varies with strength of connectedness for Conscientiousness.
   (e) Gender moderates the effect of personality similarity on connectedness.

3.2 Methodology

The data used in this paper was provided by collaboration with the MyPersonality project [87]. The nature of the data does not allow us to make any causal inferences, but does allow us to uncover relationships between personality and social network structure.

3.2.1 Data collection

Permission was granted to use parts of the MyPersonality datasets by the MyPersonality research team [87]. The MyPersonality datasets are based on Facebook and personality information voluntarily disclosed by participants. All data was anonymised by the research team and data was only gathered with explicit consent from participants. Participants could complete a Personality Questionnaire on Facebook. The application showed participants their personality scores on each of the five facets of the Five Factor model at the end of the questionnaire. After receiving their results, they were then asked whether they would be willing to share their profile information and geographic location with the application. Participants did not have to provide this information to get their personality test results. The questionnaire application was available between June 2007 and 2012, and about 40% of participants provided their profile information [87]. Data from friends was not gathered, only the total number of friends were known. Connectivity data was only recorded if both friends had completed the personality questionnaire and accepted to grant access to their profile information. The data as provided by the MyPersonality team is not personally identifiable and cannot be traced back to individuals.

3.2.2 Description of MyPersonality Datasets

Two main datasets were used, the triad dataset and the ego-network dataset. The triad dataset contained information about triangles of friends from Facebook who had filled out a personality questionnaire and provided additional demographic (gender, age) and geographic information. The ego-network dataset contained network metrics such as network size, transitivity, and brokerage of Facebook users, calculated by the MyPersonality research team. The summary of the data provided by MyPersonality is found in Table 3.1.

Table 3.1: Variables provided by MyPersonality from two different datasets: ego-networks and triads.

ego-network dataset	
variable	**description**
transitivity	transitivity score of a user based on their overall network
network size	overall number of Facebook friends of a user
brokerage	brokerage score of a user based on their overall network
triad dataset	
variable	**description**
broker	id of the the broker node of a triad
friend1	id of a Facebook user connected to broker
friend2	id of a second Facebook user connected to broker
transitive	whether friend1 and friend2 are friends
O_b, O_f1, O_f2	Openness score for broker, friend1, friend2 respectively
C_b, O_f1, O_f2	Conscientiousness score for broker, friend1, friend2 respectively
E_b, O_f1, O_f2	Extraversion score for broker, friend1, friend2 respectively
A_b, O_f1, O_f2	Agreeableness score for broker, friend1, friend2 respectively
N_b, O_f1, O_f2	Neuroticism score for broker, friend1, friend2 respectively

Characteristics of ego-network and triad datasets

The ego-network dataset contained information about ego networks; the pre-computed values for node transitivity and network size were of interest for our hypotheses. The intersection of the this dataset with the triad dataset containing personality information resulted in a sample consisting of $9,659$ nodes. The variables considered for each node were the five personality scores from the personality questionnaire (Openness, Conscientiousness, Extraversion, Agreeableness, Neuroticism), their connected alters' personality scores, their network size, and transitivity. The full dataset of $313,699$ cases will be referred to as the triad dataset, while the reduced dataset of $N = 9,659$ will be referred to as the ego-networks dataset. The ego-networks dataset is used to

answer Hypotheses 1d and 1e. The triad dataset of $313,699$ was used to answer hypotheses 1a, and 2b to 2e. A breakdown of the definitions and equations for the different network characteristics used is provided in Section 3.2.3.

Measures of degree and transitivity capture the more global characteristics of a network of Facebook users. However, these values are pre-computed, and do not allow for many additional inferences beyond their relationship with the different personality facets. The personality homophily analyses focus on the local characteristics of the network, which we have more control over through the creation of a network graph, G. This restriction is partly due to the way the data was collected. Rather than focusing on a person and their entire ego-network, the data relied on users being interested in completing a Personality questionnaire. Friendship data between two people could therefore only be assessed if both of them completed the questionnaire. As a result, the average degree in the network derived from the personality data is 6 friends, and global density is very low, see Table 3.2. As the actual network G is very sparse and missing a lot of links that could be there, we prefer using the pre-computed versions of degree and transitivity to get a better picture of their relationship with the different personality facets. We study sub-networks of the overall network in terms of pairs and triads, which represent meaningful connections between people in this network, rather than its overall structure. Figure 3.1 displays a subgraph of G,illustrating how sparsely connected the network is. 1000 edges were randomly selected to be represented in PowerBI, as displaying all of the edges resulted in an unreadable network graph with too many overlapping nodes. As can be seen, the network is composed of many "islands" of users connected to each other but otherwise disconnected from the rest of the graph.

Together, the information from the triad dataset formed a network, denoted graph G, of $313,699$ nodes and $627,503$ edges, see table 3.2 for a breakdown of the characteristics of G.

Table 3.2: Network characteristic values for network G

characteristic	value
number of nodes	$313,699$
number of edges	$627,503$
average clustering	$.14$
graph density	$3.18 * 10^{e-6}$
graph transitivity	$.15$

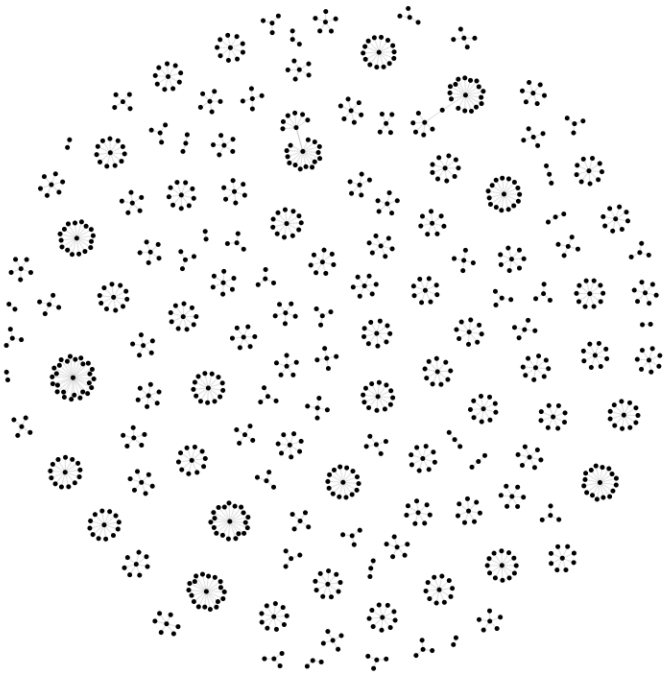


Figure 3.1: Network representation of 1000 randomly selected edges from G

3.2.3 Ego-centric network characteristics

Ego-centric networks offer information about a central actor, ego, and all its connections, referred to as alters. The below definitions and equations apply to every single ego network from the MyPersonality ego-networks dataset [87]. The open-source stat-

istics software R was used to compute most of the network characteristics [147]. The following definitions are applied.

1. **Network size** : total number of nodes in ego's network, including ego.

2. **Transitivity**: this is the fraction of possible triangles that exist involving ego. Transitivity for ego node e is defined as:

$$\frac{2 * e\_triangles}{e\_degree * (e\_degree - 1)} \tag{3.1}$$

   where $e\_triangles$ denotes the number of triangles including e and e_degree is the degree of e.

3. **Strength of connectedness**: The relative number of common neighbours a pair of nodes u,v have, normalised by their combined network size, defined as:

$$\frac{|N(u) \cap N(v)|}{|N(u)| + |N(v)|} \tag{3.2}$$

   where u and v are two connected nodes in the network graph G (see Table 3.2), and N refers to the node's degree in G.

4. **Triangle closure**: Triangle closure is a dichotomous variable, which takes the value of 0 for open triangles and 1 for closed triangles in G. Open triangles are triads in the Facebook dataset that have a broker with two unconnected friends. Closed triangles have brokers with two directly connected friends.

A high transitivity score is an indication of a more connected network. Strength of connectedness and personality variances measure the relative similarity through connectivity, while triangle closure further provides local connectivity information.

3.2.4 Personality measures

Both continuous and categorical measures are used for personality, but were all derived from the scores obtained by Facebook users on the Five Factor model of personality.

Questionnaires used for the MyPersonality dataset are based on the NEO-PI-R developed by Costa and McCrae [33]. The personality questionnaires used ranged from 20 to 100 items and were all derived from the IPIP [59]. Each item was answered on a 5-point Likert scale and responses were averaged across all items to give a single score on each of the five facets for each participant.

Overall personality similarity

We used variance to measure personality similarity in open and closed triads. The equation below is applied to each triad in the dataset, for each facet separately, resulting in five similarity measures for each triad. x represents the facet score for nodes u, v, and w, while $\bar{x}$ represents the mean, and n represents the number of terms in the distribution:

$$\frac{\sum(x)^2}{n-1} - \bar{x}^2 \tag{3.3}$$

Note that a lower variance score represents higher similarity.

Tercile scores

The personality scores for all users were divided into terciles to obtain a categorical personality variable. The values of the cut-offs used for this tercile approach are presented in Table 3.3. Participants scoring equal or below the lower cut-off were considered low scorers. Participants scoring equal or above the upper cut-off were considered high scorers. Participants scoring between the lower and upper cut-off were considered middle scorers. This approach has been successfully applied in previous work [130, 4]. For this study, we consider low and high scorers for each facet, allowing a focus on the facet's extremes. Middle scorers were taken out of the analyses.

We opt to use terciles in our analysis for two main reasons. Firstly, individuals tend to shy away from extreme values in surveys that use midpoints [156]. Creating terciles helps disentangle low, middle and high scorers in view of this natural tendency.

Table 3.3: Terciles cut-offs for personality scores

facet	lower cut-off	upper cut-off
Openness to experience	3.47	4.14
Conscientiousness	2.95	3.74
Extraversion	3.24	3.95
Agreeableness	3.25	3.82
Neuroticism	2.29	2.95

Secondly, terciles allow a clearer demarcation between the stronger and weaker scores, through which the low and high terciles can be used to test hypotheses that concern extreme values (e.g., extraverts and introverts). The use of terciles also increases statistical sensitivity as the effect sizes in these type of studies tend to be small. For similar reasons, this approach has been successfully adopted for the analysis for personality in a number of settings (e.g., [4, 130, 135]).

Based on their personality category (low or high), each pair of nodes is separated into one of three categories for each facet: mutually low scoring pairs (both users scored low on the facet), mutually high scoring pairs (both users scored high on the facet), and mixed pairs (one user scored high, and the other low, on the facet).

Table 3.4: Definition of low and high scorers for each personality facet

facet	low scorer	high scorer
Openness to experience	not open	open
Conscientiousness	unconscientious	conscientious
Extraversion	introverted	extraverted
Agreeableness	disagreeable	agreeable
Neuroticism	emotionally stable	neurotic

3.2.5 Planned analyses

For Hypothesis 1a, multiple linear regression was used for age, and Welch's t-test was used for gender effects. For hypotheses 1b and 1d, linear regression was used. For hypothesis 1e, a mix of linear regression and Welch's t-test was used. Welch's t-test is an alternative version of the Student t-test which is robust against unequal variances and unequal sample sizes [131]. For Hypothesis 2a, we used chi-square to test the difference between expected and observed frequencies of connected same and mixed pairs to non-connected same and mixed pairs. Expected frequencies are generated in $SPSS$ using the chi-square model. These expected frequencies are based on chance levels, but take into account sample size distributions if they differ between the groups that are being compared. In this case, mixed pairs have a smaller N compared to same pairs. This data unbalance is taken into account when expected frequencies are generated. For Hypothesis 2b and 2e, multiple linear regression and ANOVA were used to test the effect of personality similarity on strength of connection, and the moderating effect of gender. For Hypothesis 2c, Welch's t-test was used to compare the differences between open and closed triangles in terms of personality similarity. Open triangles were much more frequent than closed triangles, which prompted the use of Welch's t-test over ANOVA. We adopt a Bonferroni-corrected alpha level of $.001$ to account for the numerous tests carried out in the following analyses. The Bonferroni-corrected alpha level was obtained by dividing the starting alpha level (0.05) by the number of tests carried out (N=50). All statistical analyses are carried out using IBM SPSS 23 [74], while the data restructuring, including the building of network representations were carried out in Python, using the networkx package [152, 66].

3.2.6 Descriptive Statistics

Table 3.5 provides descriptive statistics on the personality scores of the full $N = 313,699$ triad dataset. Missing data accounts for the variation in sample size between

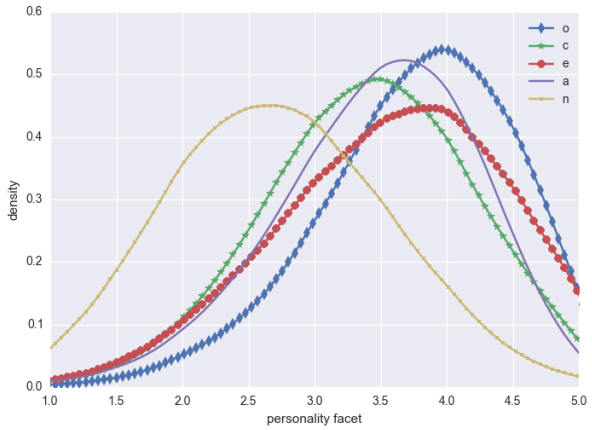


Figure 3.2: Distribution of scores for each of the five personality facets.

the facets. The means and standard deviations for each sample seem typical of an online Internet population [142]. Inter-correlations between facets were small, ranging from 0.01 to -0.36. The highest correlations between facet scores was between Agreeableness and Neuroticism. The distribution of personality scores can be found in Figure 3.2.

Table 3.5: Descriptives for personality scores in the sample

personality facet	**N**	**minimum**	**maximum**	**mean**	**std**
Openness	$313,519$	0.00	5.00	3.76	0.74
Conscientiousness	$313,520$	0.00	5.00	3.40	0.78
Extraversion	$313,521$	0.00	5.00	3.55	0.85
Agreeableness	$313,521$	0.00	5.00	3.47	0.76
Neuroticism	$280,100$	0.00	5.00	2.73	0.83

3.3 Results

We first tested the effects of age and gender on personality to uncover any biases that might arise from these two demographic variables. We then explored the effects of personality on different network characteristics in accordance with our hypotheses. We adopt a conservative alpha-value of $.001$ to account for the large number of tests carried out.

The following standard symbols are used to report test statistics:

1. t stands for t-value, the test statistic from a t-test
2. χ stands for the chi-square test statistic
3. N stands for sample size
4. M stands for mean
5. SD stands for standard deviation
6. SE stands for standard error
7. β stands for standardized regression coefficients
8. B stands for unstandardised regression coefficients
9. p stands for p-value

3.3.1 Age

We explore the relationship between age and personality, with a view to accounting for this as a possible mitigating factor in subsequent analyses. It must be noted, however, that our study is not longitudinal, and therefore it does not measure changes in personality over the lifespan. Instead, the goal is to uncover any differences in personality between younger and older Facebook user cohorts. The triad dataset of $N = 313,699$

cases was used. The Facebook sample was on average quite young, with a mean age of $M = 24.6(7.06)$ and a median age of 22. The youngest person was 18 and the oldest 60. Skewness was $1.84(.006)$, while kurtosis was $3.65(.011)$. This could be because Facebook mainly attracts a young audience. It must also be noted that age information was absent for $N = 130,979$ users, resulting in a reduced sample size of $N = 182,690$, which could further skew the results.

We tested the effect of all five personality facets on age separately first to identify significant effects. Based on the literature [128], we would expect all facets to vary with age at least somewhat. Among these, Conscientiousness ($p < .0001$) and Agreeableness ($p < .0001$) were significant. Surprisingly, Openness ($p = .78$), Extraversion ($p = .24$) and Neuroticism ($p = .67$) did not differ with age, see Table 3.6 for a breakdown of the test statistics. Agreeableness ($B = 0.048$, $SE = 0.022$, $t(2,182590) = 2.15, p = .032$) no longer had a significant effect when controlling for Conscientiousness ($B = 1.40$, $SE = 0.022$, $t(2,182590) = 64.42, p < .0001$), bearing in mind our conservative alpha value of .001; the results for all personality facets can be found in Table 3.6. The model with Agreeableness and Conscientiousness had the highest R^2 value, but was only able to explain 2.5% of variation in age. Age increases by one year with every 1.4 increase in Conscientiousness.

Table 3.6: Relationship between Agreeableness and Conscientiousness scores and age.

dependent variables	**B**	**SE**	**t**	**p-value**
Openness to experience	-0.006	0.022	-0.28	$.78$
Conscientiousness	1.41	0.021	68.76	$< .0001$
Agreeableness	0.51	0.021	23.89	$< .0001$
Neuroticism	-0.009	0.022	-0.42	$.67$
Extraversion	0.023	0.019	1.19	$.24$

As age did not have an effect for most personality facets, it was not further considered as a control variable.

3.3.2 Gender

Gender differences in social media usage and their interaction with personality have been studied in the literature [67, 5], but are not as often considered in studies on personality and social networks, with a few notable exceptions (e.g., [99]). We first explore whether any differences exist between male and female Facebook users, using the triad dataset. The non-parametric Welch's t-test [131] was used to explore differences in personality between male and female users, since homogeneity of variance was violated for Conscientiousness ($p = .006$), Extraversion ($p < .0001$), Agreeableness ($p < .0001$), and Neuroticism ($p < .0001$). Welch's t-test provides a robust test of the equality of means for samples with unequal variances across populations with unequal sample sizes [131]. The results can be found in table 3.7. Openness to Experience, Conscientiousness, Agreeableness, and Neuroticism all significantly differ between males and females. It must be noted, however, that effect sizes are extremely low for these results, varying between $r = .002$ and $r = .003$.

This means that, although a consistent difference between both genders was found, as expected from Hypothesis 1a, these differences are minimal, as seen in their means (Table 3.7).

3.3.3 Network size

To answer Hypothesis 1b on network size, the ego-networks dataset of $N = 9,659$ cases was used. Ego networks ranged from 22 to 1997 alters, with a mean score of 338 $(SD = 299.7)$. Extraversion indeed had a positive effect on network size, with $B = 76.00(3.69), t(9457) = 20.57, p < .0001$. This means that for each point-increase on the Extraversion scale, a user had on average 76 more friends. As predicted by

Table 3.7: Differences in personality between male and female Facebook users.

personality facet	t	p-value	N_F	N_M	mean F (SD)	mean M (SD)
Openness	17.69	$< .0001$	5376	3285	$3.9(0.65)$	$4.0(0.63)$
Conscientiousness	23.67	$< .0001$	5376	3285	$3.5(0.71)$	$3.4(0.72)$
Extraversion	0.70	.40	5376	3285	$3.7(0.81)$	$3.7(0.83)$
Agreeableness	10.77	.001	5376	3285	$3.6(0.68)$	$3.6(0.72)$
Neuroticism	238.09	$< .0001$	5348	3251	$2.8(0.81)$	$2.6(0.81)$

Hypothesis 1b, Neuroticism indeed had a negative effect on network size, with $B = -37.20(3.74), t(9395) = -9.95, p < .0001$. Thus for every point-decrease on the Neuroticism scale, a user loses 37 friends on average. These results confirm previous findings in the literature concerning Extraversion, Neuroticism and network size.

3.3.4 Transitivity

Finally, the ego-network dataset was used for our hypotheses on transitivity, 1d and 1e. Transitivity scores ranged from 0.0006 to 0.91, with a mean score of $0.13(SD = 0.16)$. We found a significant negative effect of Openness to Experience on transitivity, $B = -0.011(0.003), t(9458) = -4.19, p < .00001$, as well as for Extraversion, $B = -0.019(0.002), t(9458) = -9.72, p < .0001$. This is in line with Hypothesis 1d. However, we also predicted that network size would mediate the effect of Extraversion on transitivity, see hypothesis 1e. This was not the case in our sample, however. Extraversion remained a significant predictor of transitivity, even when controlling for network size, with $B = -.008(.002), t(9457) = -4.27, p < .0001$.

We further expected gender to mediate the effect of Openness on transitivity (Hypothesis 1e). We once again opt for a Welch's t-test for its robustness. Missing data for gender accounted for the variations in sample size. Transitivity indeed varied between

both genders, with $t(6505) = 12.53, p < .0001$. Men ($M = .14$,$SD = .17$), on average, had networks with higher transitivity than women ($M = .13, SD = .15$). When separated by gender, Openness had a significant negative effect on transitivity for males ($B = -.023(.005), t(5375) = -5.07, p < .0001$), but not for females ($p = .06$). This confirms previous findings that observed a significant relationship between low transitivity and Openness for males only [99] (Hypothesis 1e).

3.3.5 Personality Homophily

In this section, we present the results regarding personality homophily among this sample of Facebook users. In this section, we investigate Hypothesis 2a on the relationship between personality similarity and connectedness in pairs of Facebook users in Section 3.3.5. In Section 3.3.5 on the relationship between personality similarity and strength of connection, we investigate Hypothesis 2b. And finally, in Section 3.3.7 on the difference in personality similarity between open and closed triads of Facebook users, we consider Hypothesis 2c.

Personality similarity in connected and non-connected pairs

We hypothesized that people with similar personalities are more likely to be connected than people who have different personalities (Hypothesis 2a). Pairs of users were separated according to their personality composition (mixed or same) and their connectedness status (connected or non-connected), see Section 3.2.4. Chi-square was then used to test whether proportions of connected similar users are higher than expected, and whether consequently the proportions of mixed user pairs are lower than expected, and vice versa for non-connected pairs. Each chi-square test is carried out on a 2x2 contingency table for each personality facet, with connectedness as rows (connected or non-connected) and type of pair as columns (mixed or same). An example of such a contingency table for Extraversion is shown in Table 3.8. Expected values were gener-

ated by the following formula, where the row represents the pair and the column their observed frequency: $\frac{row\_total * column\_total}{table\_total}$.

This is to ensure that expected values follow the same distribution as observed values. To illustrate this relationship between observed and expected values, we use the example of the Extraversion contingency table (Table 3.8). The sum of each row and column are in bold, while the table total (sum of all cells in the table) is in bolded italics.

Table 3.8: Observed frequencies for the Extraversion facet for connected and non-connected same and mixed pairs.

	same	**mixed**	**total row**
connected	16266	13472	**29734**
not connected	56944	52091	**109035**
total column	**73210**	**65563**	***138773***

To calculate, the expected value for the mixed connected group, we plug in the values from the table into the equation: $\frac{29738 * 65563}{138773} \approx 14049.65$. This is rounded to 14050 in Table 3.9. This process is repeated for all cells in the contingency table for each personality facet, with the summary of all expected and observed values presented in this table.

We expected homophily effects to be especially prevalent for certain facets, such as Openness, Agreeableness, Extraversion, or Conscientiousness, but not necessarily for Neuroticism, which tends to make building relationships with others more difficult. It could be argued that neurotic users bond over the fact that they have trouble making friends in usual social settings (Hypothesis 2a). However, a homophily effect for all five personality facets was found, see Table 3.9. For all facets, connected same pairs were more prevalent than expected, while the frequency of mixed pairs was lower than expected. This effect was reversed for non-connected pairs, further supporting Hypothesis 2a.

Table 3.9: Observed and expected frequencies of similar and mixed pairs for connected and non-connected users, separated by personality facet.

Openness to experience					
connectedness	**pair type**	**observed**	**expected**	χ	$p-value$
connected	mixed	10714	**11333**	80.070	$< .0001$
	same	**13707**	13088		
not connected	mixed	**42817**	42198		
	same	48112	**48731**		
Conscientiousness					
connectedness	**pair type**	**observed**	**expected**	χ	**p-value**
connected	mixed	10834	**11288**	42.069	$< .0001$
	same	**14392**	13938		
not connected	mixed	**41447**	40993		
	same	50167	**50621**		
Extraversion					
connectedness	**pair type**	**observed**	**expected**	χ	**p-value**
connected	mixed	13472	**14050**	57.298	$< .0001$
	same	**16266**	15688		
not connected	mixed	**52091**	51513		
	same	56944	**57522**		
Agreeableness					
connectedness	**pair type**	**observed**	**expected**	χ	**p-value**
connected	mixed	12686	**13080**	29.547	$< .0001$
	same	**14087**	13693		
not connected	mixed	**48124**	47730		
	same	49574	**49968**		
Neuroticism					
connectedness	**pair type**	**observed**	**expected**	χ	**p-value**
connected	mixed	13066	**13635**	58.026	$< .0001$
	same	**15300**	14732		
not connected	mixed	**50803**	50235		
	same	53708	**54277**		

Personality homophily and strength of connection

Finally, we consider the influence of personality homophily on the strength of their connection, as considered in Hypothesis 2b. Personality similarity was calculated for triads in the triad dataset ($N = 313,699$), using Equation 3.3. Triads for which either facet score was missing were not included in the final analysis. We refer to this score as the *difference score*, since a higher value corresponds to a bigger difference between the three nodes in terms of personality.

We base our strength of connectedness measure on the edges in network graph G, created using the triad data from the MyPersonality project [87], see Section 3.2. Strength of connectedness ranged from 0.0 to 0.45, with a mean of 0.06 ($SD = 0.077$) and an $N = 1,048,575$. Table 3.10 details the results of the effect of personality homophily on strength of connectedness.

Table 3.10: Effects of personality similarity on connectedness.

personality facet	β	t	$p - value$
Openness to experience	-0.005	-2.37	.018
Conscientiousness	$-.011$	-11.34	$< .0001$
Agreeableness	$-.008$	-9.47	$< .0001$
Extraversion	$-.010$	-10.23	$< .0001$
Neuroticism	0.006	2.70	.007

Connectedness decreased as difference scores on Agreeableness, Conscientiousness and Extraversion scores increased. This means that, as expected, personality similarity was higher among better connected nodes for Agreeableness and Extraversion (Hypothesis 2b). We also expected this effect for Openness, but this could not be confirmed in our sample. On the other hand, we found an unexpected effect for Conscientiousness.

3.3.6 Interaction with gender

Since we found a significant effect of gender for most personality facets, (Section 3.3.2), we also consider whether gender would moderate the effect of personality similarity on connectedness (Hypothesis 2e). We first checked whether there were gender differences in regard to connectedness. Since connectedness scores are computed pair-wise, we separated our data according to the pairs' gender composition. Pairs were male-male ($N = 200,974$), female-female ($N = 346,762$), or mixed ($N = 466,697$). There was a significant difference in connectedness scores between the different pairs, $t(1,520565) = 476.59$, $p < .0001$. Connectedness was strongest among male-male pairs ($M = 0.062, SD = 0.078$), followed by female-female pairs ($M = 0.057, SD = 0.076$), and finally mixed pairs ($M = 0.056, SD = 0.076$). Since there was no gender difference for Extraversion, we only considered whether there was a difference in the observed negative effect of personality variances in Agreeableness and Conscientiousness on connectedness (Hypothesis 2e). The results can be found in Table 3.11.

Table 3.11: Effects of personality similarity on connectedness by gender pairs.

gender pairs	β	t	$p-value$
Conscientiousness			
male-male	-0.014	-6.37	$< .0001$
female-female	-0.011	-6.76	$< .0001$
male-female	-0.013	-8.83	$< .0001$
Agreeableness			
male-male	-0.019	-8.50	$< .0001$
female-female	-0.002	-1.22	$.221$
male-female	-0.014	-9.85	$< .0001$

Of interest is the observation that for Agreeableness, personality similarity only had a significant effect for male-male and mixed pairs on connectedness, but not for female-female pairs (Table 3.11) with male-male pairs having the stronger negative effect ($\beta = -0.019$) compared to mixed pairs ($\beta = -0.014$). For Conscientiousness, although the strength of the effect varied slightly between pairs, all of them had a negative effect on connectedness (Table 3.11). In accordance with Hypothesis 2e, gender had a moderating effect on the relationship between personality similarity and strength of connectedness for Agreeableness.

3.3.7 Triangle closure

We further hypothesized that personality homophily would be stronger among closed triangles, compared to open triangles (Hypothesis 2c), see Definition 4. Personality similarity was measured between all three nodes in each of the triads. We expect a positive effect of similarity for Openness, Extraversion, and Agreeableness, see hypothesis 2b. We carried out Welch's t-test to uncover any significant differences between open and closed triangles for those facets.

We found a significant effect for Extraversion and Agreeableness, but not for Openness (Table 3.12). Additionally, we also found an effect for Conscientiousness. For all three significant facets, the personality variances were higher among open triangles, as opposed to closed ones (Table 3.12).

As expected from Hypothesis 2c, personality similarity was higher among closed triangles than among open triangles for Openness, Extraversion, and Agreeableness, as well as for Conscientiousness.

Table 3.12: Differences in personality similarity for closed and open triangles.

personality facet	t	$p - value$	M	M
Openness	36.41	$< .0001$	0.34(0.36)	0.33(0.35)
Conscientiousness	76.15	$< .0001$	0.39(0.39)	0.38(0.38)
Extraversion	108.41	$< .0001$	0.44(0.45)	0.42(0.44)
Agreeableness	79.17	$< .0001$	0.38(0.39)	0.37(0.38)
Neuroticism	0.88	.35	0.44(0.41)	0.44(0.41)

3.3.8 Summary of results

1. **Network characteristics and demographics**

   (a) **Personality scores differ between genders and across age.**

   Weak support. Conscientiousness was positively related to age, while agreeableness was negatively related to age. There was no significant effect for the other facets regarding age. Gender effects were found for all facets, except for Exraversion, but actual differences in mean personality scores were extremely low. Men appeared to be slightly more open than women; while women were slightly more conscientious and neurotic than men. Overall, age and gender effects were not significant or very weak in relation with personality.

   (b) **Extraversion is positively related to network size, and Neuroticism is negatively related to network size.**

   Strong support. Extraversion was positively related to network size, while Neuroticism was negatively related to it.

   (c) **Openness to Experience and Extraversion are negatively related to transitivity.**

Strong support. Openness to experience and Extraversion were both negatively related to transitivity.

(d) **Effects of Openness to Experience on transitivity are moderated by gender, while effects of Extraversion on transitivity are mediated by network size.**

Mixed support. Network size did not mediate the effect of Extraversion on transitivity. On the other hand, the effect of openness to experience on transitivity was only significant for men.

2. **Personality Homophily**

(a) **Nodes who are connected are more likely to be similar to each other than nodes who are not.**

Strong support. We found that among connected pairs, same pairs were more prevalent than expected, while mixed pairs were less prevalent. This relationship was reversed among non-connected pairs, were mixed pairs were more prevalent and same pairs less prevalent than expected.

(b) **Similarity in Openness to Experience, Agreeableness, and Extraversion is positively related to strength of connection.**

Mixed support. Similarity in openness to experience and Extraversion scores were indeed positively related to strength of connection among Facebook users. No effect for agreeableness was found however.

(c) **Personality similarity is higher in closed triangles than in open triangles for Openness, Extraversion and Agreeableness.**

Strong support. Similarity in Openness to experience, Extraversion, and Agreeableness was higher for closed triads compared to open triads.

(d) **Personality similarity differs for closed and open triangles, and varies with strength of connectedness for Conscientiousness.**

Strong support. Similarity in Conscientiousness was associated with a

stronger connection among users. Closed triangles were also more similar in Conscientiousness than open triangles.

(e) **Gender moderates the effect of personality similarity on connectedness.**

Mixed support. An interaction with gender was only found for Conscientiousness and Agreeableness. the effect between similarity and strength of connection was strongest among male pairs for conscientiousness and agreeableness, and weakest for female pairs.

3.4 Discussion

The results provide evidence that online network characteristics and personality are intrinsically linked. The personality facets that emerge as related to the shape of our social networks are Extraversion, Neuroticism, Openness to Experience and to a certain extent, Conscientiousness and Agreeableness.

3.4.1 Replication and the moderating role of gender

As expected, network size was positively related to Extraversion, and negatively related to Neuroticism. Extraversion and Openness to Experience were both negatively related to transitivity. We also replicated a gender-dependent effect for Openness to Experience on transitivity [99]: Openness had a significant negative effect on transitivity for males, but not for females. This confirms that Extraversion and Openness are related to networks low in transitivity, in which extraverted and open nodes act as bridges between loosely connected alters. In conjunction with the effects found for network size, this translates to large, loosely connected networks for extraverts.

Of further interest is the moderating role of gender on the relationship between personality similarity and connectedness. In general, connectedness was strongest for

male-only pairs, followed by female-only, and then mixed pairs. For Agreeableness, the effect of personality similarity on connectedness was strongest among male-only pairs, but did not have an effect for female-only pairs. On the other hand, personality similarity continued to have the same positive effect for Conscientiousness, regardless of gender.

3.4.2 Personality homophily results

Furthermore, we studied connectedness in two different ways: first through strength of connection between two nodes, and secondly, through triangle closure. More strongly connected nodes had higher similarity on the Extraversion facet. This effect could also be confirmed for Agreeableness and Conscientiousness, but not for Openness to Experience. Openness to Experience might not have been found to be homophilous, because open individuals, by virtue of their personality, are open to form connections with both similar and dissimilar others. This is highlighted by their transitive networks, which demonstrate a diverse and loosely connected network.

Finally, we also identified that similarity in Openness, Extraversion, Agreeableness, and Conscientiousness was higher in closed triangles compared to open ones. This suggests that personality homophily indeed plays a role in structuring connections in ego-centric networks for certain facets, but not for others. It must be noted, however, that effects in regard to triangle closure and personality homophily were small, with only a few $.01$ differences between closed and open triangles on the personality variance scores. The large MyPersonality dataset of over $300,000$ nodes allowed us to tease out an effect that smaller samples would very likely have missed. Future research will have to determine whether this personality homophily effect is worth investigating further.

3.4.3 Limitations

An alternative explanation for the observed personality homophily effects might be that some personalities are simply more attractive than others: people enjoy the company of friendly and sociable friends. This is especially true for agreeable users, who are the most popular recipients of offline friend requests [138] and online interactions on communication boards [12]. Extraverts on the other hand, are more likely to initiate friendships [138] and reach out to people online [12].

However, we also found personality homophily effects for Conscientiousness, which cannot be explained through the attractive personality explanation. Conscientiousness is usually overlooked as an influential factor in online and offline social networks. Future research in online social networks should nonetheless explore the attractive personality hypothesis, potentially uncovering the most popular personality combinations among online and offline friendship pairs. These results provide further evidence that observed social network features are potentially linked to embedded human characteristics.

Whether personality drives social network structure or vice versa remains unclear, however. A potential explanation could be that personality, being an inherent characteristic, plays a role in determining one's initial network position, which, in turn, is responsible for reinforcing specific personality traits. For example, neurotic people might be in broker positions because of their personality, or become more neurotic because broker positions are stressful [83]. The fact that most facets vary with age supports the idea that personality is malleable [142], but it must also be noted that personality varies only slightly over people's lifespan.

3.5 Conclusion

In conclusion, personality and social network position could be reinforcing each other, and further research needs to explore the possible causal links between social network position and personality. The most important findings are ***highlighted***.

We have replicated some fundamental relationships between Extraversion, Neuroticism and network size, as well as between Extraversion, Openness and transitivity. Gender also played a role: the negative relationship between Openness and transitivity was only significant for males, and the positive relationship between connectedness and Agreeableness was not significant for female-only pairs. Using two different methodologies, triangle closure and strength of connectedness, we have shown that the five personality facets are homophilous to different degrees, and that one facet in particular, Neuroticism, did not appear to be homophilous among Facebook users.

Furthermore, we have uncovered valuable new insights in personality homophily in online social networks. Conscientiousness, Extraversion, and Agreeableness emerge as homophilous facets, both in relation to strength of connection and triangle closure, as predicted. Similarity in Openness to Experience did not appear to be significantly related to strength of connection, but closed triangles of users were more similar in terms of Openness to Experience than open triangles. No homophily effects were found for Neuroticism, which is in line with our expectations. Of interest as well is that users similar in low Neuroticism are also not likely to be connected.

Now that we have established that Facebook users have a tendency to assort based on certain personality facets, we are ready to tackle the question of distance. Distance is a crucial factor in the formation and maintenance of social ties.The next chapter will focus on the interaction between personality and distance on connectedness between people on social media.

Chapter 4

Geographic distance and personality homophily in online social networks

Parts of this chapter have been published as a paper in the peer-reviewed journal *Cyberpsychology, Behavior, and Social Networking*: N. Noë, R.M. Whitaker, and S.M. Allen. Personality Homophily and Geographic Distance in Facebook. *Cyberpsychology, Behavior, and Social Networking*, to appear, 2018

4.1 Introduction

In this chapter, we explore the interaction between personality similarity and geographic distance on online social connections with the aim of answering the following research question: *How does personality homophily and geographic distance affect the formation of links in an online social network?*

We consider the relationships between personality homophily on one of the biggest constraints to human social networks: geographic distance. Our approach is to augment the dataset used in chapter 3 with geographic location data. This allows us to investigate the difference in geographic distance between homophilous pairs, in which both users scored similarly on a particular personality facet, and mixed pairs. We also consider the differences between connected and non-connected pairs of participants from the online social network. While taking into account the geographic distance

between two members of each pair of online users, we also consider if they live in the same country or not, as these might pose additional constraints related to borders, rather than just distance.

In order to answer the research question posed at the beginning, the chapter has been broken down into the following parts:

1. In Section 4.1, we explore the state of the current literature on the effect of geographic distance on social tie formation and maintenance. We summarise the current findings at different levels of granularity for geographic distance, and any expectations for personality effects relating to geographic distance.

2. In section 4.1.6, we summarise the hypotheses for this chapter. We have different hypotheses for each of the five personality facets.

3. In section 4.2, we elaborate on the characteristics of our dataset, which is a similar subset of the MyPersonality data as in chapter 3, but with additional geographic information.

4. Planned analyses are outlined in section 4.2.2. Welch's t-test were mainly used to investigate the hypotheses outlined in section 4.1.6.

5. In section 4.3, we present the results pertaining to each of our hypotheses defined in section 5.2.7. We go through each of the five facets and elaborate on how personality similarity affects geographic distance between pairs of users, also taking into account effects of the country.

6. In section 4.4, we assess the support found for each hypothesis, alternative explanations for the findings, and any limitations pertaining to this chapter.

7. Finally, in section 4.5, we summarise the main findings of the chapter.

The findings highlight that each personality facet has an distinct relationship with geographic distance which depends on the qualities of each of them. However, the ap-

proach in this chapter does not allow us to disentangle what exactly about each facet make them closer or further apart from similar others.

4.1.1 Geographic distance in online social networks

Sociologists have been considering the differences between local and virtual interactions and what they mean for social relationships. It is important to note that the Internet has not replaced existing forms of communication, and mostly serves to reinforce pre-existing relationships between people [158]. In this chapter, we focus on the online interactions at the individual level, with pairs of connected users being the focus of our analysis. This approach is meant to supplement the vast body of research already done at the community level on this subject [158].

People have traditionally been socially and geographically constrained in their choice of friendships, frequently establishing new ties through mutual acquaintances or shared activities [89]. The probability of two people becoming friends is well-known to decrease with geographic distance [97, 91, 11, 14, 40].

Proximity is an important constraint in social relationships [133]; it is therefore of interest to look at the spatial component of social networks in our study, which have become more relevant to the wider population through location-based social networks [62, 35, 140]. Proximity, both social and geographic, is essential for the formation of strong friendship bonds [90]. In turn, friendships influence the social venues we frequent [163].

Through the “global village" metaphor, the Internet is widely seen as transcending these geographical barriers, by providing online representations such as social networks that connect people despite distance [7]. In the absence of relevant non-verbal cues, which humans normally use to decide if they like one another, Internet users might be prone to filling in the blanks left by people’s social media profiles with their own attributions [7]. As a result, they idealize their online interaction partners and per-

ceive them as more similar to themselves than they really are. Interestingly, this effect is also true for offline networks to some extent: third to fifth graders perceived their friends to be more similar in terms of personality than they actually were [101].

Online social networks have also been argued to contribute to feelings of closeness between people through the mere exposure effect. We feel close to far-away others, because we see and read about them everyday on social media [7], something that was not possible before the advent of the Internet. Facebook users have been found to mitigate relationship decay with their more distant ties, by adapting their social network activity to reach these far away connections [153]. Whether this form of tie maintenance is actually effective to prevent decay is not yet well understood. Interestingly, Facebook users who engaged their distant ties via their social media accounts in such a fashion think that it indeed has a positive effect on their relationships [153].

4.1.2 Spatial and social networks

We refer to spatial networks as any network whose nodes are connected based on a geographic or spatial unit. The spatial self has emerged as a new form of identity in online social networks, most of which have location-sharing functionalities [136]. Dedicated LBSN, such as Foursquare or Gowalla are spatial networks with a social component, in which people share their location with friends [136]. Purely spatial networks are interesting in that they tend to be neither assortative, nor disassortative. This makes them stand apart from social networks, that tend to be assortative, and technological networks, which tend to be disassortative [14]. In LBSN, which have a social component, strong local clustering is observed instead [163]. Spatial constraints modify the graph characteristics of social networks, for example, spatial constraints restrict the appearance of large degrees [14]. Spatial constraints also lead to large betweenness centrality fluctuations. Hubs are usually central, but get even closer to the centre of gravity of all points in a spatial network [14].

It seems that the relationship between distance and tie formation might differ depending on the resolution of the distance data. Indeed, [11] demonstrated that the relationship between distance and connection probability was stronger for medium to large distance, but much weaker at short distances (below 50 miles approximately). Interestingly, frequency of mobile and online communication decrease with distance [93], but the duration of phone calls appear to increase and then level out after 60km [91]. More generally, it is proposed [91] that communication networks appear to have two main levels: short-distance communication has high clustering, but is of short duration, while long distance communication has smaller clustering, but tend to last longer.

So, while there is a general trend of friendships decreasing with distance, this relationship does not appear to be linear. Liben-Nowell and colleagues [97] noticed that the probability of friendships indeed decreases with distance, but becomes equal at distances over 1000 km. Liben-Nowell and colleagues ran a geo-routing simulation of the message-forwarding experiment [149] in a LiveJournal network. LiveJournal is an online content creation board on which people can follow each other's updates. The original message-forwarding experiment served to illustrate the small-world effect [113], by demonstrating that a source person can pass on a message to a target person through as little as 6 intermediate friends. The geo-routing experiment aims to connect two cities rather than two people, through the location of LiveJournal users. The simulation reached the target city in 4 steps in 13% of the cases, which is an improvement on the performance of real human subjects [97]. While both studies probably suffered from heavily skewed participant selection (source and targets were participants and cities in the US), it nonetheless serves to illustrate that even on the Internet, our choice of friendships appear to be geographically constrained.

A related observation was made in a mobile communication network: the likelihood of communication decreases with distance but becomes constant after 40 km. It seems that beyond a certain threshold, distance, in fact, does not matter. This could be because people might still be willing to travel to see others who are at a reasonable distance

from themselves, which increases the quality of the relationship. Distance appears to matter less when people are close enough to easily be able to travel to see each other face-to-face, or engage in activities together [7]. At the same time, it is interesting to note that the online network had a higher threshold (1000 km) than the mobile network (40 km). Mobile networks are likely restricted to in-country communication, as, until recently, calling outside of national borders tends to be much more expensive. The difference could also be attributed to the country of study: the LiveJournal network was primarily based in the US which is a much larger country than Belgium, in which the mobile network experiment took place. Regardless of threshold differences, it is important to note that distance matters at relatively smaller distances between people, but this effect tends to stagnate at higher distances.

4.1.3 Geographic distance and personality

Personality has been studied in spatial contexts at different levels of granularity [3, 62, 79, 126]; insights from the literature are outlined below going from countries, to regions, cities, and finally venues, the smallest unit at which individuals can be co-located.

Personality and countries

Personality has been shown to differ between countries [3]. In a major study involving 36 countries, Allik and colleagues uncovered interesting differences in personality distributions across the globe. Similarities in personality are often associated with geographic proximity, such as Germany and Austria, or Canada and the USA; shared ancestry, such as Black South Africa and Zimbabwe; or cultural traditions, such as Hong Kong and Taiwan. The cluster analysis shows that geographic proximity and cultural similarity hold true at higher levels as well. For example Germany, Austria, and German-speaking Swiss are clustered together, while South Koreans join the Hong

Kong and Taiwan group. Ultimately, European and American cultures merge into one cluster, while African and Asian cultures merge into another. Some exceptions occur, such as Turkey being linked to the United States and Canada, while Japan and China are clustered with European and Latin American cultures [3]. People from American and European cultures tend to be high in Openness to experience and Extraversion, and low in Agreeableness. They are outgoing, adventurous, and antagonistic People from Asian and African cultures on the other hand tend to be more introverted, traditional, and compliant. Euro-American cultures tend to reject status hierarchies and score higher in individualism. Neuroticism and Conscientiousness do not appear to have clear geographic distributions. Countries high in neuroticism and low in conscientiousness include Japan, Spain, Russia, and Belgium for example [3]. It is important to consider that language and culture might have an influence on question interpretation, understanding, and the perception of personality compared to others, as the study relied on self-report.

Personality and regions

Regions in the UK can be distinguished based on their differences in personality. For example, Scotland is high in Agreeableness, while Wales scores high on Neuroticism [126]. Several factors could contribute to these regional differences. People with certain personality traits could be attracted to similar types of locations, and therefore end up clustering in the same region. For example, people high in Openness might seek out regions that stimulate their creative and adventurous side, while extraverts will seek out regions with lively cities that allow them to fulfil their social needs. On the other hand, social, economical, or ecological factors characteristic of the region might influence the expression of personality. For example, people might become more anxious and irritable if they are surrounded by neighbours who exhibit similar traits when moving to a region high in Neuroticism [126]. Even ecological factors such as infectious diseases might play a role: Openness to experience and Extraversion are typically low in

regions where infectious diseases are prevalent. People low in Extraversion and Openness to experience limit their social contact, which prevents the spread of illnesses [134]. These explanations might all exist in conjunction and contribute to the observed regional differences to different extents, or even reinforcing one another. As more people with certain traits move to a region that seems to match their personality; the city will become more and more appealing to others with the same trait. At the same time, people already living in the region might exhibit more open or extraverted traits as the people and activities around the change to accommodate the evolving tastes of the region.

Personality and cities

At the city level, differences in neighbourhoods could be observed, with some scoring on average higher on Openness to experience and others higher in Agreeableness [79]. The different personality facets were associated with different neighbourhood characteristics. Openness to experience was associated with higher population density, higher housing prices, higher ethnic and religious diversity, and higher crime rate. This fits well with the characteristics of open individuals who value creativity, diversity, and adventure. Low agreeableness was also associated with high density, high crime rates, and high housing prices. Jokela and colleagues suggest that the low Agreeableness scores might be a result of urban alienation [79]. The study also investigated life satisfaction across several neighbourhoods in London, but found interesting personality confounds. Life satisfaction in general has a strong positive relationship with Extraversion and Emotional stability, and is not moderated or mediated in any way by neighbourhood characteristics. However, it appeared that open individuals were happiest when living in neighbourhoods that exhibited characteristics associated with openness: heterogeneous population and high density. Open individuals were happiest in neighbourhoods with others who scored high on Openness to experience. This relationship was only found for Openness to experience, and not for any other facets

[79].

Personality and venues

At the finest level of granularity lie the venues that individuals visit, where the perceived personality of patrons could be used to infer the ambiance of a venue [62]. Previous work has also shown that personality facets, such as Conscientiousness, Openness, and Neuroticism, are related to Foursquare usage [25, 26].

Conscientiousness was positively associated with the number of venues visited. This might be explained by the fact that conscientious people visit a larger number of venues compared to others, or that they are more consistent in their check-in patterns.

Neuroticism was negatively associated with the numbers of venues visited. This might be explained by the fact that neurotic users curate their check-ins more as they might be more concerned about what others think of their mobility patterns. On the other hand, neurotic users might on average have less activities because of their smaller social networks, and therefore less opportunities to visit venues and check-in at them. Finally, neurotic users might in general be less inclined to leave their house, further contributing to the lack of venues to check-in to.

Conclusion

The Internet and online social media provide an alternative platform for individuals to find common ground through shared interests, or similar attitudes and beliefs. Despite the opportunity that the Internet provides to create relationships without meeting, the importance of physical proximity in tie formation is also present in online social networks [133]. It is therefore important to consider the influence these spatial influences can have on the formation, shape, and evolution of social networks.

4.1.4 Motivation for Hypotheses

We hypothesize that people with certain personality traits are more affected by distance than others, although there is a lack of studies in the literature tackling the effect of personality similarity on distance between people who are connected online. Expectations for each individual facet and the motivation for its associated hypothesis are outlined in the sections below.

4.1.5 Distance and connectedness

We first make the general prediction than connected users are, on average, in closer geographic proximity than non-connected users, see Hypothesis 1, as friendships are more likely to blossom as distance decreases, even in online settings [7, 133, 46].

Openness to Experience

Open people might be more likely to be connected to people further away, as they are more likely to initiate contacts with a range of different people. Additionally, open individuals are more likely to move from their home state to another state in the US [78]. The willingness to travel ties in with their adventurous nature. We are predicting to observe the highest distance between two open people who are connected, especially compared to two conservative people. Both open individuals are likely to travel more and might end up in very different locations, while still staying in touch with the ties they made with others during their travels, especially if they are like-minded open people. In contrast, we would expect more conservative people to be located geographically closer to each other, as they tend to stay in an environment they are familiar with and are much less likely to want to experience different cities or countries (Hypothesis 2).

Conscientiousness

Conscientious people are organized and mindful of others [58], which might make it more likely for them to stay in contact, even if they are further away and face-to-face contact is not possible. We therefore expect the connection between fellow conscientious people to prevail even at longer distances, compared to pairs of friends who are low in Conscientiousness (Hypothesis 3). People scoring low in conscientiousness are disorganised and might lose contact with other who have moved away as they do not keep up regularly with their friends. Regular communication is crucial to maintain relationships at a distance, as face-to-face contact is not possible. Family ties are more resilient to such negligence, but friendships tend to decay if they are not maintained. However, Conscientiousness was not associated with between or within state migration, so conscientious people might not necessarily live further away from their friends compared to non-conscientious people. However, we might see that non-conscientious people lose Facebook connections at longer distances compared to conscientious users, which will be reflected in their average distance to friends.

Extraversion

We hypothesize that distance matters for Extraversion: people who are extraverts rely on physical, rather than online, activities to form friendships and maintain their social bonds [130]. Such activities require relative geographic proximity and therefore there is a basis for distance among extraverts to be lower compared to introverts (Hypothesis 4). Especially among pairs of extraverts, we expect distance to be much closer compared to pairs of introverts. Interestingly, Extraversion was associated with increased migration, but within US states and not between them [78]. This might be related to the fact that extraverts like to move but still remain close to their social contacts, and be able to easily travel back home.

Agreeableness

Agreeable people are popular friendship and communication partners, whether offline or online [12, 138] People might be motivated to stay in contact with others who are friendly. People who are high in Agreeableness might also be more likely to maintain contact with others. We therefore expect agreeable people to be connected, despite the distance (Hypothesis 5), especially if both members of the pair score high on this facet. On the other hand, people who scored high on Agreeableness were more likely to remain in their home state than to migrate [78]. They tend to be community-oriented and might prefer to build up a strong social network in their home state, rather than to branch out. Despite this effect, we still predict that agreeable people are connected at greater distances, because of their popularity as friendship partners.

Neuroticism

It is harder to make predictions for Neuroticism. Previous research suggests that neurotic people tend to have smaller groups of friends, and might use online interactions to substitute offline ones [73]. This might indicate that distances between mutually neurotic people could be larger than between emotionally stable pairs, as they might find more connections online. On the other hand, neurotic people might also need emotional support that is readily available, making functional long distance relationships less likely (Hypothesis 6). Neuroticism was not associated with migration, which might indicate that Neurotic people are not likely to move far away from their initial contacts [78]. Emotionally stable individuals might also be able to hand long distance relationships better, so we expect neurotic users to be connected at smaller distances compared to emotionally stable ones.

4.1.6 Summary of Hypotheses

1. Users who are Facebook friends are in closer geographic proximity than users who are not connected.

2. Mutually open pairs tend to be further apart, while mutually not-open pairs tend to live in close proximity to one another, compared to mixed pairs.

3. Mutually conscientious pairs maintain friendships at greater distances, compared to mixed and mutually unconscientious pairs.

4. Mutually extraverted pairs tend to be in closer proximity to one another, compared to mutually introverted pairs and mixed pairs.

5. Mutually agreeable pairs are connected at greater distances, compared to disagreeable or mixed pairs.

6. Mutually neurotic pairs tend to maintain friendships at shorter distances, compared to emotionally stable or mixed pairs.

4.2 Methods

From the MyPersonality triads dataset [87], we use a subset of variables, which contains personality, geographic, and demographic information of 300,669 Facebook users (Table 4.1).

4.2.1 Characteristics of the MyPersonality dataset

As explained in chapter 3 Section 3.2.2, the MyPersonality data is derived from Facebook users who answered a Personality questionnaire and provided access to their Facebook information and location. We use the same triad dataset discussed in chapter 3, which contains personality and demographic information of $300,669$ Facebook users

Table 4.1: Description of variables used in this chapter

variable	refers to	type	unit
personality tercile score	user	categorical	low middle high
personality similarity score	pair of users	categorical	same low same high mixed
connectedness	pair of users	categorical	connected not connected
country	pair of users	categorical	same different
distance	pair of users	continuous	kilometres

(Table 4.1). From this triad dataset, we derive pairs of connected users and their personality. Another dataset contained geographic distance between the members of each triad, which was cross-referenced with the triad dataset to obtain the final sample of $217,604$ pairs, as some had missing geographic location. This information was collected with the explicit consent of the user, which they could refuse to give.

Geographic distance

This measure refers to the distance (in km) between two users in a pair, inferred from their latitude and longitude at the moment of completing the MyPersonality questionnaire. Figure 4.1a shows all available geographic locations of Facebook users from the entire MyPersonality repository, which comprises over 1.7M users. Figure 4.1 shows the geographic distribution of Facebook users who have completed the personality questionnaire. Facebook users in this subset are overwhelmingly American (99.94%) compared to the general Facebook population presented in Figure 4.1a, where Amer-

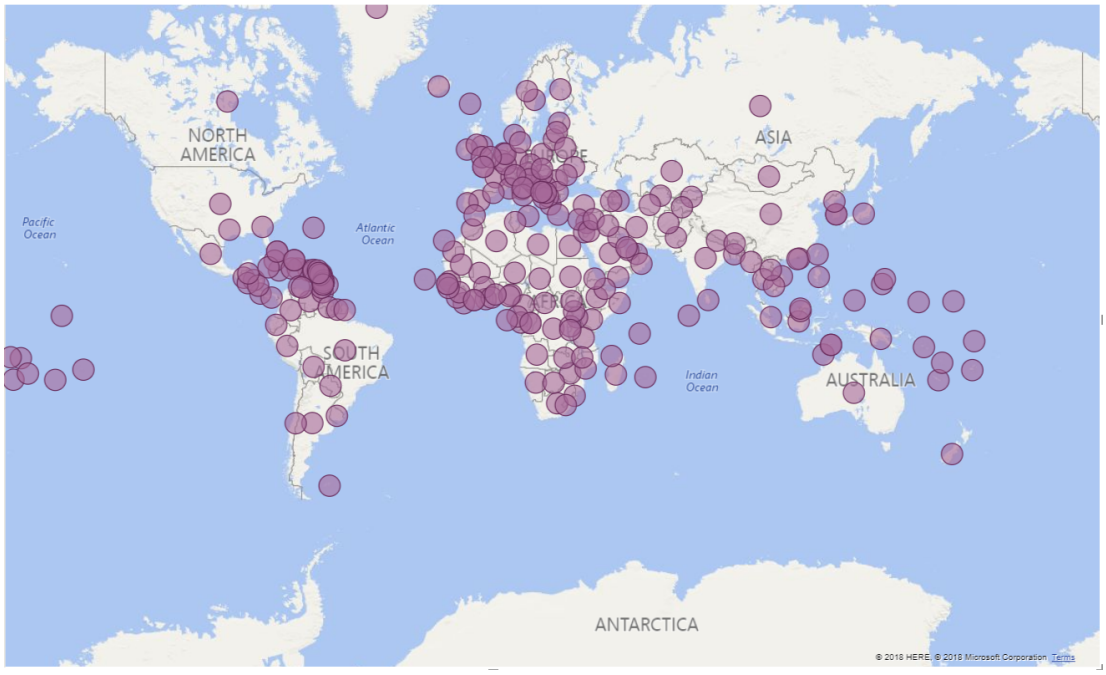
icans only make up about 55% of the users, followed by the United Kingdom (12.5%), Canada (5%), and Australia (4%). The subset of users who completed the Personality questionnaire come from 51 different countries with the United Kingdom, Germany, Canada, and France ranking after the United States in terms of number of users. The fact that the Personality questionnaire was presented in English and most Facebook users are American could explain this bias in the dataset.

Personality tercile scores

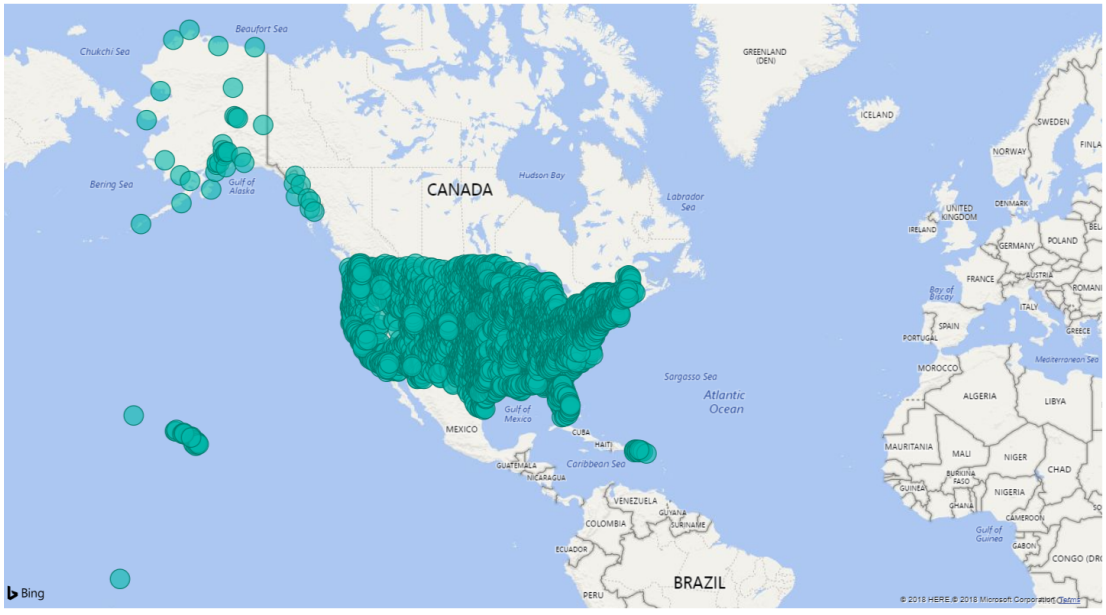
The personality scores for all users were divided into terciles to obtain a categorical personality variable, as described in Section 3.2.4 in chapter 3. The same cut-offs were used as those presented in Table 3.3. For each facet, we consider low and high scorers, allowing a focus on the facet's extremes and clear-cut comparisons, such as between extraverts and introverts, for example. For each pair of connected users on Facebook, we categorize their relative scores into one of three categories. For *same low pairs* of users, both users are low scorers; for *mixed pairs* of users, one user is a low scorer, while the other is a high scorer; for *same high pairs* of users, both users are high scorers.

Sample size

Distance and personality data was available for $289,557$ pairs of users from the MyPersonality dataset. $71,953$ pairs of users had missing data for the distance measure, bringing the total number of valid pairs to $217,604$. For Openness to experience, this results in $86,354$ remaining pairs. For Conscientiousness, this results in $87,514$ remaining pairs; $105,033$ pairs remaining for Extraversion; $93,607$ for Agreeableness; and $98,823$ pairs for Neuroticism. The varying numbers between the facets are a result of missing data for the personality scores. All sample sizes for the different facets are within at least two standard deviations from the mean sample size, making them



(a) Geographic location of Facebook users (each dot represents a general location of the users, such as regions or countries



(b) Geographic location of Facebook users who have completed the personality questionnaire (each dot represents a specific location, latitude and longitude coordinates recorded by the phone of the user when completing the MyPersonality questionnaire)

Figure 4.1: Geographic spread of general Facebook users and Facebook users who completed the personality questionnaire.

comparable. Each pair of connected users on Facebook was assigned to one of three categories.

4.2.2 Planned analyses

We first analysed the average distance to friends of Facebook users, depending on their own personality tercile score, but independently of the personality of their friends. This is to uncover any inherent tendencies of people of certain personality dispositions. Same pairs were then compared to mixed pairs, with the expectation of finding a significant difference in mean geographic distance, based on Hypotheses 2 to 6. Welch's t-test was used for all analyses. This is an alternative to the Student t-test, which is robust against unequal sample sizes and unequal variances [131].

Welch's t-test is used to evaluate the effect of personality pairs on geographic distance. We compare same pairs to mixed pairs, with the expectation of finding a difference in geographic distance, based on Hypotheses 2, 3, 4, 5, and 6, as outlined previously (Section 4.1.6). We also compare same low scoring pairs and same high scoring pairs, as distance might differ substantially between pairs of extraverts and pairs of introverts, for example. Bootstrapping (1,000 samples) is applied using IBM SPSS Version 23, as geographic distance was right-skewed. This method uses case resampling with replacement [74]. The results presented in this chapter include the bootstrapped parameter estimates, but with original N values for the sample size. All statistical analyses are carried out using IBM SPSS 23 [74], while the data restructuring was carried out in Python [152]

To consider connectedness, we separate our sample into connected and non-connected pairs. We expect to find differences in distance among connected users, but not among non-connected users. According to Hypothesis 1, we expect distance to be significantly lower among connected users, compared to non-connected users, which we evaluate using Welch's t-test. Welch's t-test was used because connected and non-

connected pairs differ significantly in sample size. For our hypotheses concerning personality and geographic distance, we evaluate each hypotheses among connected and non-connected users. We expect to find effects of personality among connected users, but not among non-connected ones.

We control for the effect of country in this relationship, as users who live in different countries are more likely to have higher geographic distances between them, which might somewhat skew the results. We also consider the fact that regional differences in personality might exist [126], which could manifest itself at the country level. Therefore, in addition to separating the sample according to connected and non-connected users, we also separate it according to whether the pairs live in the same country or not. We might find that for some facets, people who are similar in personality live closer together, regardless if they are connected or not.

In light of the numerous tests conducted, we adopt a Bonferroni-corrected alpha-level of $.0014$. The False Discovery Rate (FDR) was also checked with the Benjamini-Hochberg procedure. Both methods help guard against Type I errors. The Bonferroni-corrections modify the alpha level by dividing the usual alpha level ($.05$) by the number of tests carried out (N = 35). The Benjamini-Hochberg procedure is also applied to the p-values, as an alternative to the Bonferroni-corrections, using an online calculator [124].

Additional standard statistical symbols used in this chapter are outlined below:

1. W stands for Welch's t-test statistic
2. $Cohen'sd$ is a measure of effect size

4.3 Results

We expected pairs of connected Facebook users to be closer in terms of geographical distance than users who were not connected to each other (see Hypothesis 1). Indeed,

users who are connected were in significant closer geographic proximity (M $= 422$km, $SD = 893$), than users who were not friends on Facebook (M $= 638km$, $SD = 1,029$), with $t(1, 2161090) = 37254, p < .0001$.

Using Welch's t-test, we found significant differences between low scorers and high scorers in terms of distance to friends for all facets (Table 4.2). W stands for the Welch's test statistic and Cohen's d was used as a measure of effect size. Open people had friends living further away on average, compared to not-open individuals. Conscientious people, on average, tended to live further apart from their friends compared to unconscientious Facebook users. Agreeable users also tended to live further away from their friends compared to disagreeable users. Extraverted users, on the other hand, tended to live closer to their friends, compared to introverted users. Similarly, neurotic users had friends closer to them, on average, as compared to emotionally stable Facebook users. Analyses were also carried out comparing middle scorers to low or high scorers. Middle scorers consistently scored in between low and high scorers in terms of geographic distance to their friends. These findings were left out, as they did not contribute any more insights than already presented in Table 4.3. Subsequent analyses carried out in this chapter focus on low and high scorers for this reason as well.

4.3.1 Openness to Experience

For the Openness to Experience facet (Hypothesis 2), there was a significant effect for connected pairs, who lived in the same country, $t(2, 10411.74) = 11.254$, $p < .0001$. Results from the Welch's t-tests are presented in Table 4.3.

Pairs of same low scorers in openness to experience ($M = 417$km, ($SD = 876.3$)) were significantly closer geographically to each other than people in mixed pairs ($M = 484$km, ($SD = 958.8$)). Homogeneous high scorers ($M = 502$km, ($SD = 991.5$)) were significantly further apart than same low pairs, but not compared to mixed pairs ($p = .027$). This might be explained by the fact that conservative people do not tend to

Table 4.2: Mean distance (in km) between users and their friends for each facet, separated by high and low scorers.

Facet	Score	M	SD	N	W	p-value	Cohen's d
Open	Low	561.38	807.81	49185	305.04	<.0001	0.1
	High	649.15	878.48	64392			
Consc	Low	557.16	807.57	46480	390.03	<.0001	0.1
	High	657.34	882.06	66508			
Extra	Low	622.93	859.15	52702	18.55	<.0001	0.0
	high	601.87	834.72	70378			
Agree	Low	602.7	842.61	56005	14.33	<.0001	0.0
	High	621.55	859.45	61628			
Neuro	Low	607.97	838.56	57565	0.93	.33	N/A
	High	603.31	847.27	64543			

move around, and stay in close proximity to similar others, while open people tend to travel further away, and therefore are spread out around the country.

There was a significant effect of personality on distance for people who lived in the same country, even though they were not connected, $t(2, 30858.23) = 55.509$, $p < .0001$. Homogeneous low scoring pairs ($M = 608$km, $SD = 987.5$) were significantly closer together than mixed pairs ($M = 665$km, $SD = 1055.8$). In turn, mixed pairs were significantly closer than same high scoring pairs ($M = 730$km, $SD = 1116.9$), see Table 4.3.

Finally, there was a marginally significant effect for pairs of users who were connected, but who did not live in the same country. Pairs of same high scorers ($M = 53km$, $SD = 84.8$) appeared to be geographically closer together than mixed pairs ($M = 520km$, $SD = 797.7$). However, this result was not significant ($t(1, 24.898) = 8.425, p = .008$) when taking into account our Bonferroni-corrected alpha level.

Table 4.3: Welch's t-test results of the comparison between same low, same high, and mixed pairs for Openness to experience, separated by country.

Openness to experience					
country	**connectedness**	**pair type**	**df**	**W**	**p-value**
same	connected	same low / mixed	12385	21.249	$< .0001$
		same high / same low	11038	21.249	$< .0001$
	not connected	same high / mixed	55454	47.938	$< .0001$
		same low /mixed	42003	25.632	$< .0001$
		same high / same low	35447	104.966	$< .0001$

4.3.2 Conscientiousness

For Conscientiousness (Hypothesis 3), there was a significant effect of personality on distance among connected pairs, who lived in the same country, $t(2, 8567.102) = 40.065$, $p < .0001$. Results from the Welch's t-tests are presented in Table 4.4.

Pairs of same low scorers (M = 399km, $SD = 924.3$) and mixed pairs (M = 440km, $SD = 912.1$) did not differ significantly in terms of geographic distance ($t(1, 5263.25) = 4.462, p = .035$), when adopting a Bonferroni-corrected alpha-level of $.0014$. However, with a FDR-corrected p-value of $.048$ using the Benjamini-Hochberg procedure, this difference is statistically significant. Same high scoring pairs (M = 548km, $SD = 1000.8$) were significantly further apart than mixed pairs, with $t(1, 17280.6) = 55.51$, $p < .0001$. Homogeneous low and same high pairs also differed significantly, as shown in Table 4.4.

For non-connected pairs in the same country, their relative conscientiousness scores had a significant effect on the average geographic distance between them, $t(2, 27051.61) = 116.546$, $p < .0001$. Homogeneous high scoring pairs (M = 760km, $SD = 1107.2$) tended to be geographically further apart than mixed pairs (M = 653km, $SD = 1043.2$). Homogeneous low scorers (M = 591km, $SD = 1033.2$), were in turn geographically closer than mixed pairs, see Table 4.4.

For pairs of non-connected users in different countries, there was a significant effect of Conscientiousness on distance, $t(2, 127.729) = 7.329$, $p = .001$. Pairs of same low scorers lived significantly closer together (M=489km, ($SD = 714.2$) than pairs of same high scorers (M=1453km, $SD = 2026.1$). There was no significant difference between mixed pairs and same low or high pairs, with $t(1, 123.134) = 4.111, p = 0.045$ and $t(1, 121.187) = 5.012, p = .027$ respectively.

It appears that country, rather than connectedness, has a stronger effect on the distance between conscientious Facebook users.

Table 4.4: Welch's t-test results of the comparison between same low, same high, and mixed pairs for Conscientiousness, separated by country. *not significant with Bonferroni-approach, but FDR-corrected p-value of $.048$ **is significant.**

Conscientiousness					
country	**connectedness**	**pair type**	**N**	**W**	**p-value**
same	connected	same high / mixed	17379	55.51	$< .0001$
		same low/ mixed	11650	4.461	.035*
		same high / same low	11805	56.328	$< .0001$
	not connected	same high / mixed	57273	139.794	$< .0001$
		same low /mixed	39192	26.451	$< .0001$
		same high / same low	37202	183.309	$< .0001$
different	not connected	same low / same high	114	13.695	$< .0001$

4.3.3 Extraversion

For the Extraversion facet (Hypothesis 4), results were mostly in line with our predictions. Results from the Welch's t-tests are presented in Table 4.5.

Among connected pairs, same low scoring pairs lived the furthest apart ($M = 545$km, $SD = 998.6$), compared to mixed pairs ($M = 465$km, $SD = 927.4$). Same high scoring pairs lived closest together ($M = 420$km, $SD = 897.6$), compared to mixed

pairs, and same low scoring pairs, $t(2, 11488) = 24.879, p < .0001$. This provides further support for our hypothesis: extraverts like to physically meet up with others to maintain their relationship, which might explain their closer proximity.

People from the same country who were not connected tended to have different average distances between them based on their personality scores ($t(2, 35815.02) = 33.178, p < .0001$). Indeed, pairs of same low scorers (M = 728km, $SD = 1087.6$) tended to be geographically further apart, compared to mixed pairs (M = 687km, $SD = 1077.2$). In turn, same high scoring pairs were geographically closer (M = 641km, $SD = 1041.1$) than mixed pairs. Extraverts might tend to choose cities or neighbourhoods that are lively, which translates to a closer proximity to each other as they choose similar regions to live in.

When comparing pairs of non-connected users, from different countries, Extraversion did not have an effect on distance ($t(2, 109.624) = 0.361, p = .70$). There was also no effect for pairs of users who were connected, but lived in different countries ($t(2, 28.27) = 0.093, p = .91$). For extraverts specifically, we expected them to be geographically closer together than mixed pairs or introverted pairs (see Hypothesis 4).

Table 4.5: Welch's t-test results of the comparison between same low, same high, and mixed pairs for Extraversion, separated by country.

		Extraversion			
country	**connectedness**	**pair type**	**N**	**W**	**p-value**
same	connected	same low / mixed	15263	20.89	$< .0001$
		same high / mixed	19723	11.987	.001
		same high / same low	13210	49.151	$< .0001$
	not connected	same high / mixed	67438	30.778	$< .0001$
		same low /mixed	51542	14.042	$< .0001$
		same high / same low	42348	59.189	$< .0001$

4.3.4 Agreeableness

We found no support for Hypothesis 5 in regards to country and geographic distance. When looking at pairs of users, there was no significant effect of personality on distance for non-connected users, regardless of whether they lived in the same country ($t(2, 54.358) = 0.358, p = .701$) or not ($t(2, 36650.64) = 1.701, p = .183$). Surprisingly, also for people who were connected, there was no significant effect of personality on distance, again whether they were in the same country ($t(2, 11213.26) = 0.786, p = .456$) or not ($t(2, 12.778) = 0.257, p = .777$). It seems that the Agreeableness facet does not play a role in the relationship between geographic distance and personality homophily.

4.3.5 Neuroticism

As with Agreeableness, we found few significant effects of neuroticism on distance for pairs of users (Hypothesis 6). Significant results for Neuroticism are presented in Table 4.6.

For connected pairs, there was no significant effect, whether pairs of users lived in a different country ($t(2, 23.668) = 0.606, p = .554$) or in the same country ($t(2, 11097.52) = 2.552, p = .078$).

For non-connected users, there was no significant effect for people in different countries ($t(2, 111.505)p = .0.021$), but there was one for people within the same country, $t(2, 37276.92) = 7.585, p = .001$. Homogeneous pairs of low scoring users ($M = 704$km, $SD = 1073.9$) lived further apart than mixed pairs ($M = 672$km, $SD = 1063$), as shown in Table 4.6. Homogeneous high scoring pairs ($M = 661$km, ($SD = 1060.7$)) did not differ significantly from mixed pairs in terms of geographic distance ($t(1, 53575.7) = 1.36, p = .244$). However, they were significantly closer together than low scoring pairs. It could be that, like extraverts, neurotic people are attracted to specific places or regions, which explains this effect.

Table 4.6: Welch's t-test results of the comparison between same low, same high, and mixed pairs for Neuroticism, separated by country.

Neuroticism					
country	**connectedness**	**pair type**	**N**	**W**	**p-value**
same	not connected	same low /mixed	50901	9.665	.002
		same high / same low	39246	14.703	$< .0001$

4.3.6 Summary of results

Table 4.7 gives an overview for the most important results in this chapter and their effect sizes. This table only includes connected pairs and does not separate samples based on country. As is often the case with personality, effect sizes are small for all relationships summarised in Table 4.7. Implications for all findings and their effect sizes are discussed in Section 4.4.

Table 4.7: Welch's t-test results for Openness to Experience, Conscientiousness, and Extraversion.

pair type	N1/N2	W	p-value	Cohen's d
Openness to Experience				
Same low / mixed	3770 / 8641	14.34	<.0001	0.1
Same low / same high	3770 / 7282	21.20	<.0001	0.1
Conscientiousness				
Same high / mixed	8789/3046	58.09	<.0001	0.1
Same high / same low	8789/8623	57.65	<.0001	0.2
Same low / mixed	8623/3046	4.461	.035	0.0
Extraversion				
Same low / mixed	4380/10901	21.03	<.0001	0.0
Same high / mixed	8861/10901	11.83	.001	0.0
Same high / same low	8861/4380	49.13	<.0001	0.1

1. **Users who are Facebook friends are in closer geographic proximity than users who are not connected.**

   Strong support. Facebook users who were connected were indeed significantly closer to each other than Facebook users who were not connected.

2. **Mutually open pairs tend to be further apart, while mutually not-open pairs tend to live in close proximity to one another, compared to mixed pairs.**

   Mixed support. Open Facebook users tended to be further apart from their friends compared to non-open individuals. Also mutually open pairs tended to be further apart than mutually non-open pairs, but there were not significant differences in geographic distance between mutually open pairs and mixed pairs. On the other hand, mutually non-open pairs were significantly closer together than mixed pairs.

3. **Mutually conscientious pairs maintain friendships at greater distances, compared to mixed and mutually unconscientious pairs.**

   Strong support. Conscientious users were geographically further apart from their friends than unconscientious users were. Mutually conscientious pairs were significantly further apart than mixed pairs. Mutually unconscientious pairs, on the other hand, were geographically closer together than mixed pairs.

4. **Mutually extraverted pairs tend to be in closer proximity to one another, compared to mutually introverted pairs and mixed pairs.**

   Strong support. Extraverted users were geographically closer to their friends, compared to introverted users. Mutually extraverted pairs were significantly closer together than mixed pairs. On the other hand, mutually introverted pairs were geographically further apart than mixed pairs.

5. **Mutually agreeable pairs are connected at greater distances, compared to disagreeable or mixed pairs.**

No support. Agreeable users were geographically further apart from their friends than disagreeable users. There were no significant differences in geographic distance between mutually agreeable, mutually disagreeable and mixed pairs.

6. **Mutually neurotic pairs tend to maintain friendships at shorter distances, compared to emotionally stable or mixed pairs.**

   Weak support . Neurotic and emotionally stable users were not different in terms of their geographic distance to their friends. A significant effect was only found among non-connected pairs in the same country, where mutually neurotic pairs were geographically closer compared to mutually emotionally stable pairs. Mutually emotionally stable pairs were also geographically further apart compared to mixed pairs.

4.4 Discussion

The aim of this chapter was to study the interaction of geographic distance and personality on the connection between Facebook users.

In accordance with our hypotheses, we found that geographic distance between a pair of users differed depending on their personality. Notably, we find that people who scored high on the openness to experience, conscientiousness, and agreeableness were geographically further apart from their friends compared to people who scored low on these same facets. On the other hand, high scorers on the Extraversion and Neuroticism facet were found to be geographically closer to their Facebook friends, compared to low scorers on those facets. It is important to note, however, that most of these relationships have small effect sizes. However, we have used conservative methods to control for false positive rates through Bonferroni-corrections. We therefore believe that despite the small effect sizes, the uncovered relationship between personality and geographic distance is worth investigating further.

To further investigate this effect, we compared connected pairs of users, based on their personality composition (same or mixed). We confirmed that conscientious and open pairs of users were indeed further apart than their low-scoring counterparts, as well as pairs with mixed personality scores on these facets. We also found that extraverted pairs of users were closer together geographically, compared to introverted pairs and mixed pairs. We did not, however, find any effects for the Agreeableness or Neuroticism facet.

4.4.1 Openness to experience

We find that Facebook friends who both score high on Openness to experience are significantly further apart than Facebook friends with mixed Openness scores, and Facebook friends who both score low on Openness to experience. Openness to experience is positively related to living in heterogeneous neighbourhoods [126], where befriending people from different cultures and countries is more likely. It could be that mutually open Facebook friend pairs first lived in the same neighbourhood and then one or both moved away. Indeed, Openness to experience is also associated with increased migration between US states [78]. When we consider that the average age of our sample is in the early 20s, it is likely to include many students, who have many opportunities to travel abroad and form new social connections. The more open ones are likely to take a gap year to travel or to go on exchange abroad, during which they might have initialised friendships that are then maintained through Facebook despite the distance once the exchange or travel period is over. In addition, high Openness to experience is associated with a tendency to stay in touch through online means, and to build strong online social networks [72]. It is therefore likely that Facebook is an important tool for open individuals to stay in touch with distant others, particularly those they feel similar to and who might have the same tendencies to stay in touch through online means.

4.4.2 Conscientiousness

We find that Facebook friends who both score high on Conscientiousness are more likely to live further apart than pairs of friends who score low on Conscientiousness. Conscientiousness was not associated with migration in a US sample [78], so the explanation for this pattern is probably different than for open people. This finding might be confounded by the intrinsic disorganised nature of unconscientious people: they might be less consistent in their communication patterns with others who are further away, and therefore lose connections that cannot be maintained through face-to-face contact more easily than conscientious Facebook users. Interestingly, conscientious users are less likely to stay in touch with others through online means [72]. This appears to contradict our speculation on conscientious users staying in touch *online*. A Facebook friendship for conscientious users might be indicative of a stronger offline relationship, as conscientious users do not tend to use online social media as much. They might therefore be staying in touch with distant others through different means, which are not captured through online interactions such as phone calls or texting. This facet in particular deserves further investigation, as it consistently emerges as homophilous, although the existing literature rarely mentions it as a possible candidate.

4.4.3 Extraversion

Pairs of extraverted users were significantly closer geographically than pairs of introverted users. Extraverts rely on direct contact with their social contacts to stay in touch with them [23]. Social media is only used as a means to report on activities carried out with friends, and generally not as the primary tool to stay in contact with them [5]. This is further confirmed by the finding that Extraversion is not related to building strong online relationships [72]. This strongly suggests that extraverts prefer to physically meet up with their friends and geographic proximity is crucial for the planning of joint social activities. For introverted pairs, online-based relationships might be more likely,

which could explain the increase in distance between them compared to extraverted pairs. Introverts construe their “real-me" as how they present themselves online, so they might be more comfortable to communicate via more indirect means, and might not require frequent face-to-face activities.

4.4.4 Agreeableness

We find no difference in geographic distance between agreeable pairs of Facebook friends, disagreeable pairs, or mixed pairs. Agreeable people tend to stay in their home state and contribute to social cohesion in their neighbourhoods [78]. Agreeableness was also positively associated with the strength of online relationships, which can help maintain relationships at long distances. However, it has also been found that agreeable people are the most popular interaction partners online [12]. This might indicate that people are motivated to stay in touch with them regardless of distance, which could explain the lack of effect overall.

4.4.5 Neuroticism

We find that pairs of neurotic users are geographically closer together compared to pairs of emotionally stable Facebook friends. We suggested that this might be due to the fact that neurotic users are restricted in the social connections they form and therefore stay in close proximity to their existing contacts so they can rely on them for support. Interestingly, while neurotic users prefer indirect ways of communicating such as text messages [23], Neuroticism is negatively associated with forming strong online connections [72]. This might mean that neurotic users are not good at maintaining relationships at longer distances, which mostly rely on online communication. Neuroticism was also not associated with migration [78], making it less likely for neurotic users to move around and therefore increase the distance between them and their social contacts. Travelling or moving to another state, country or even within the same region

can be a stressful experience, which might be exacerbated if a person is naturally prone to anxiety and worry. Neurotic people might therefore avoid such experiences, which contributes to their more local connections compared to emotionally stable Facebook users.

4.4.6 Limitations

Further work will have to focus on the reasons behind these effects. The static nature of the current dataset does not allow to answer this question, but a more dynamic network approach, which follows people and their connections as they move or settle somewhere, would be able to give further insight. A recent paper found that personality influences the way we keep in touch with others, which shows the importance of such an approach [72]. It is also important to consider the quality of the relationships. Offline only rrelationships have been shown to be the best in terms of quality, followed by mixed-mode relationships, and online-only ones [7]. However, it is also important to note that online and offline networks mirror each other in terms of structure, making readily-available social media data an acceptable proxy for recreating people's social networks [46]. Finally, with the exponential increase in the use of social media over the last decade, it is valuable to consider online networks as a worthy subject of study in their own right.

A limitation of the dataset is that the location of participants of the personality questionnaire do not necessarily match their actual location of residence. The distance metric took into account their location at the moment they completed the questionnaire, while the country variable was derived from a different dataset and referred to their country of residence (as inputted by them on Facebook). In addition, locations reported on Facebook might not always be accurate. People might forget or neglect to update their location when moving to a new city or country. How many of these cases are present in the dataset is unknown. This means that people's location in this analysis might not match up to their usual location, which could skew the computed distances

between pairs and skew the interpretation of the results.

Future work will need to address how such relationships evolve with distance in a longitudinal context, which also takes into account other variables, such as frequency and quality of contact, which have been found to be essential for the maintenance of social ties. Personality homophily and how it relates to network structure and social connections remains an understudied phenomenon in the current literature. To the best of our knowledge, this study is one of the first to link personality homophily and geographic distance together.

4.5 Conclusion

Personality homophily had a distinct relationship with geographic distance, depending on the personality facet studied. The most important findings are ***highlighted***

Pairs of extraverted friends tended to live closer together than pairs of introverted or mixed friends. On the other hand, pairs of open friends tended to live further apart when in the same country. Conscientious users were also on average further apart from each other, compared to unconscientious ones. We found no significant effects for Agreeableness, which might mean that this facet is not affected by geographic distance as much as the other ones. The finding for neuroticism was puzzling, as only effects for non-connected users were found. This might mean that geographic location might play a role in attracting people with similar personalities to the same location.

The findings from this chapter provide evidence to support that notion that personality mediates the sustenance of online relationships over geographical distance. This is considered at low resolution, with distance measured in kilometres, and participants spread potentially between countries. In an attempt to further explore the possibility that personalities are attracted to the same location, we investigate the effect of personality on co-location in the next chapter. More specifically, we explore how personality

similarity is related to the number of common venues users of a location-based social media application, frequent.

Chapter 5

Personality homophily in a location-based social network

Parts of this chapter have been published in the peer-reviewed journal Computer in Human Behavior in 2016: N. Noë, R.M. Whitaker, M.J. Chorley, Martin, and T.V. Pollet. Birds of a feather locate together? Foursquare checkins and personality homophily. *Computers in Human Behavior*, 58: 343-353, 2016.

5.1 Introduction

In this chapter, we explore the relationship between personality and physical co-location with the aim of answering the following research question: *Are people co-located at the same venue more likely to be similar?*

We move away from the broad scope of geographic distance into a much more granular view of distance: the different venues we frequent everyday. In this chapter, we do not only redefine our definition of distance, we also adjust what we understand under "connection". Instead of viewing connection as a clear friendship request on Facebook or a follow on Twitter, we define connection in this Chapter as co-location at a particular venue.

LBSN change the way we view distance. Instead of an absolute measure of distance which was considered in the previous chapter, LBSN consider distance in more relative

terms, with co-location at the same venue being more meaningful than the actual distance between them. LBSN rely on checkins at venues by users, the choice of a user to signal to their friends that they are at a particular locations by “checking-in". The venue represents any location people can co-exist in: a shop, a train station, a restaurant. Following the locations people visit through these checkins can give us valuable insights into their preferences, despite not providing any evidence of their mobility throughout their day or a month. The attraction to a certain venue creates commonality for all the people who decided to visit a particular venue. It is on the basis of this commonality that we can derive a meaningful network of LBSN users who are connected through the common choices of venues they have made. This commonality is solely based on geography, and not on time. Time is not considered, as the ground for commonality is the venue, rather than a physical interaction. Many considerations go into the choice of a shop or restaurant in terms of convenience, price, and enjoyability. The one we consider in this chapter is our personality. People with similar personalities might be attracted to similar venues, although few studies have ventured to explore this particular relationship. For example, we would expect introverted people to prefer quieter places, while extraverts would prefer more social places. In recording people’s checkins we capture which venues they want to be associated with, rather than the venues they actually go to. This creates an interesting bias, as the venues that fit a person’s self-identity and personality best are more likely to be captured.

This way, LBSN provide a new view of distance, by not considering absolute geographic metrics such as latitudes and longitudes, but by giving a meaningful representation of what it means to be at the same location. They distinguish themselves from other forms of social media in that regard, as they rely on the physical world to create a virtual footprint of visited locations in the online world. They enable us to study human co-location with an unprecedented level of fidelity, as the venue ensures that each recorded co-location is indeed meaningful and not merely two people passing each other on a busy road. The characteristics of LBSN are further expanded on in Section 5.1.1.

The dataset used in this chapter is derived from a study carried out by Chorley and colleagues [25, 26]. To explore user personality and location-based activity, Chorley and colleagues collected data through the Foursquare Personality Experiment [25], which was designed to allow users of the Foursquare [1] location-based social network to participate in anonymous collection of their checkins and personality profile in return for a visualisation of their own personality relative to others at locations where common checkins are made. This novel approach naturally incentivises participation and has allowed viral participant recruitment "in-the-wild" to be accomplished, resulting in data from 174 anonymous participants who have collectively checked in 487,398 times at 119,746 venues. Chorley and colleagues [25] gave valuable insight through correlational analyses of personality facets and check-in behaviours, but we extend this work with additional considerations. First, we build a geographic network of Foursquare users based on their common checkins to map out the connections people make by visiting the same locations. We then consider the effects of personality homophily on the geographic connections between users, and how venues can attract users with the same personality.

This way, the same dataset enables us to link preferences for the same venues to similarities in personality. As in the previous chapter, we introduce a measure of strength for connectedness by distinguishing between pairs of users who frequent several of the same venues, as opposed to those who only have a few in common. While we consider the distance dimension, we do not take into account the time dimension when measuring co-location. Based on the strength of connection, we build several network graphs with the users as nodes and the number of venues each connected user pair has in common as the edge. The occurrence of expected connections is compared to that of the observed ones in each graph, using chi-square tests.

In order to answer the research question, the chapter has been broken down into the following parts:

[1] Foursquare have reorganised their business model since the study and checkins are now made through a dedicated application called Swarm: `http://www.swarmapp.com`

1. In section 5.1.1, we elaborate on the significance of LBSN applications and their use for research.

2. In section 5.1.2, we focus on the motivation behind the use of LBSN applications.

3. In Section 5.1.3, we elaborate on the idea of spatial homophily and how personality relates to it.

4. In Section 5.2, we provide the motivation behind each of our hypotheses, and summarise them.

5. In Section 5.3, we present the methodology and dataset used in this chapter to answer the different hypotheses.

6. In Section 5.4, we present the results pertaining to each of the hypotheses; we first look at the characteristics of the user network obtained from the dataset in Section 5.4.1, then at their personality scores in Section 5.4.2, and finally we assess the co-occurrence of personality in the network in Section 5.4.3, and summarise the findings for each hypothesis in 5.4.3.

7. In section 5.5, we discuss the findings of this chapter, their implication, and any limitations.

8. In Section 5.6, we summarise the main findings of the chapter.

We find that there were no homophily effects when considering users' overall personality profile. In other words, number of common locations between users was not significantly related to their similarity on their overall personality profile. More interesting findings emerge when we look at each facet separately. We find that Openness to Experience and Conscientiousness have strong homophilous effect in regards to venue co-location. Next, we find partial support for Agreeableness and Neuroticism being homophilous, but no evidence for Extraversion. Conscientiousness and Extraversion both present surprises: we expected strong effects for Extraversion but found none, while

we expected no effects for Conscientiousness and found strong support. This mismatch might be explained by the fact that our hypotheses are mainly informed from findings regarding standard social networks, rather than spatial ones. Spatial networks might have different rules of association than social ones, which makes them an interesting avenue for new research regarding homophilous tendencies.

5.1.1 Characteristics of Location-Based Social Networks

With the advent of the Internet and the popularity of social networking, it has become possible to understand this concept through the electronic ties that individuals choose to make with each other, leading to a wide range of insights from large electronic data sources. Despite these recent advances, relatively little is known about the manifestation of homophily in a physical context, thus the *extent to which similar people have a preference for visiting the same places* is an important question to ask. Unfortunately, a significant barrier to answering this question has been convenient data collection on a large scale, which until recently has been challenging to accomplish without access to dedicated location tracking equipment. However, the recent advent of smartphones and LBSN allows new progress to be made.

Checkins give particular insight into the venues that an individual chooses to record as important, interesting or relevant. However in some SNS such as Facebook and Google+, the checkin functionality has been introduced as a secondary function, built on top of other online social networking functionality. Foursquare is different in this regard, originating with checkins as its primary function, and with limited secondary content provision. These factors, combined with a rich Application Programming Interface (API) on which third party applications can be developed, have led to Foursquare being a popular basis for academic insight to a range of human behaviours. Primarily these have concerned physical activity, such as relating to patterns made by users (e.g., [118]) and with a high degree of location data aggregation. This has led to insights into the effect of social relationships and routine on spatial behavior

for example [24].

5.1.2 User Motivation

A LBSN users' checkin behavior may be motivated by several factors, such as establishing a social connection with friends, discovering new places to visit, keeping track of already visited places, fighting boredom and gamification [98]. LBSN allow users to select certain locations as a means of self-presentation, referred to as the *spatial self* [136]. This is frequently consistent with other forms of online self-presentation and can involve venue avoidance to counter associations with perceived negative places [98]. Users have been found to control the volume of checkins in different ways, avoiding spamming their social networks with too many checkins and giving thought to self-presentation [136]. Different levels of consistency (i.e., venue selection) have been reported. Some users consistently check in to any place they visit, while others select their checked in locations more carefully, based on how interesting or deserving they deem the place to be [98]. Audience management is a further aspect of user behavior in LBSNs, with users sharing different checkins with different groups of friends and acquaintances. In some cases, interesting checkins, meaning checkins at unusual or new venues, were reserved for Twitter and Facebook, while more general checkins were shared with friends [34].

These factors mean that the checkin is a potentially noisy signal with varying purposes between individuals. To some degree, checkins represent a unique footprint which is characteristic of the individual user, and are worthy of investigation as a means to understand human behavior. However, limited existing studies have addressed the role of checkins in relation to individual differences such as personality. [154] have considered the personality characteristics that correlate with individuals sharing checkins in Facebook, and in [25], the personality traits of individual users have been correlated with observed checkins.

5.1.3 The Emergence of Spatial Homophily

Given that personality is a potential predictor for behaviour and attitudes in a range of situations [58], it is possible that personality-based homophily may support the attraction of like individuals for a wide range of scenarios [140]. One conceivable scenario where personality may have a homophilic effect relates to the type of location that individuals choose to visit. So-called *spatial homophily* has only recently been considered [163, 121], and captures the attraction of individuals, who are in some sense similar, to common locations.

Taking the volume, diversity and broad categorisation of venues visited as variables, the first examination of human mobility behaviour at street level, in relation to human personality [25] identified a number of interesting correlations. In particular, conscientiousness positively correlated with the number of venues visited, openness positively correlated with checkins at both sociable and popular venues, and neuroticism negatively correlated with the number of sociable venues visited.

Recent work [136] has proposed that people may use the places that they visit to build an online representation of themselves. Hence, potentially the characteristics of people can be derived from the locations that they choose to affiliate with through checkins. [62] demonstrated that impressions of a place and its visitors could systematically be derived from the Foursquare user profiles of its visitors. Participants were able to accurately predict the personality of a typical visitor of a specific location, based on the Foursquare profiles of actual visitors ($ICC = .69$). Ambiance ($ICC = .32$) and typical activities of visitors ($ICC = .33$) of a specific place had far lower agreement. On a larger scale, [35] demonstrated that a city's character could be derived from the mobility patterns of its residents. Similar people tended to visit a network of venues within a neighbourhood or region of a city that form a comprehensive whole, rather than individual locations [35].

Personality has also been related to spatial location and to spatial homophily. For ex-

ample, different neighbourhoods in London have different personality profiles [79]. Here it was identified that the centre of London has a higher prevalence of high Openness to experience and low Agreeableness, while neighbourhoods further away from the city centre are low in Neuroticism and high in Conscientiousness. [79] also showed that personality mitigated the effect of neighbourhood on life satisfaction. More specifically, open individuals were the happiest in neighbourhoods with a high number of fellow open people. This suggests that personality-homophily can have important implications for life satisfaction in specific London neighbourhoods [79]. Personality not only characterizes specific neighbourhoods, but evidence has been presented that it may characterize entire countries, as explored in chapter 4.

The places considered through spatial homophily need not be restricted by one's residential neighbourhood or region, however. For example, [80] identified clusters of individuals, such as gym enthusiasts or art enthusiasts, who had similar interests in venues consistent with their Foursquare checkins. Interestingly, the venues visited by individuals within the same cluster were spread throughout the city, rather than being confined to a particular neighbourhood. Specific types of locations, rather than general geographic areas, can therefore be places where people with similar personality traits assort. This contributes to the motivation for our investigation.

5.2 Motivation for Hypotheses

Our focus concerns observing signals of homophily through common LBSN checkins and similarity of personality. The extent of the effect of individual differences in personality on the similarity of locations visited remains unknown. Developing further understanding of this issue is our objective, while acknowledging that checkin activity represents only a subset of human physical behavior and a conscious but noisy signal, with different motivations for its use (see Section 5.1.2).

Based on previous findings (e.g., [163]) it is possible some venues may play a greater

role in facilitating spatial homophily than others, such as leisure venues (e.g. sports centre) and sociable avenues, (e.g. night-life spots), as compared to venues people only pass through as a necessity and with little option for choice or self-expression (e.g., transport hubs). Furthermore, each checkin may serve as a signal to social network followers concerning personal affiliations with places that they feel are important.

Given this context, we consider the implications of personality facets on spatial homophily in the following sections. As the literature on spatial homophily and location-based social networks is limited, we additionally consider the usage of online social networks and user personality.

5.2.1 Openness to Experience

Recent research from spatial homophily [79] suggests that openness to experience might be the strongest predictor of homophilous connections in an LSBN such as Foursquare. Openness to experience was also positively correlated with visiting sociable and popular venues [25]. In terms of online social networks, open people tend to enjoy a diverse network of friends [155] and are frequent users [155, 130, 135]. The motivation for use of online social networks by highly open users is most likely tied to their novelty [4]. Therefore, one could infer that in a LBSN setting, open users might seek popular venues, because such locations appeal to them through their novelty and originality. Sociable venues might be attractive because open people tend to enjoy socializing with and meeting new people. Additionally, by virtue of their curiosity, open people might have a tendency to assort at common venues that are new and interesting to them. However, this could lead to widespread dispersion of checkins, reducing scope for spatial overlap and common checkins, thus resulting in lower spatial homophily. In terms of low openness scoring, such individuals may have a tendency to congregate at a more limited range of familiar places, affecting likelihood of common checkins being detected.

5.2.2 Extraversion

In terms of LBSN, Extraversion has not been found to correlate with any particular checkin behaviors [25], but their high sociability characteristics might make them likely to assort at sociable venues nonetheless [139]. When using Facebook, extraverts post and share updates about their social life through photos and events more often than introverts; and have, unsurprisingly, a bigger network of friends in online communities [4, 135, 123]. Therefore LBSN might be especially suited to extraverts who like to readily share the events and offline activities they take part in through online means [4]. However we could equally find that extraverts are attracted by a diverse range of venues, and therefore do not display the predicted homophilous behavior. Furthermore, in terms of online behavior, extraverts have been found to refrain from using the Internet as a substitute for social interactions [6]. This means that extraverts could use LBSN consistent with meeting friends or partaking in social activities. For online social networks it has also been argued that extraverts, although enjoying a vast number of friends and being less prone to loneliness, tend to have less well connected neighbours, while introverts are embedded in strongly connected networks, albeit with fewer neighbours [139, 67]. Introverts post and share less on social media, however, when they do, they gain more likes and comments than their extraverted counterparts [5], providing support for the idea that introverts are embedded in small, but tight-knit social networks. Homophily has been shown to be stronger in smaller communities [94], we could therefore find introverts to be more homophilous than extroverts, including in a location-based social network such as Foursquare.

5.2.3 Conscientiousness

For online activity, Conscientiousness was found to be negatively correlated with leisure-related Internet use and positively with academic Internet use among adolescents [92]. It has been argued that conscientious users tend to stay focused on their tasks, which

makes them less likely to engage in distracting behaviours, such as going on Facebook [130]. Conscientious users have more friends on Facebook than unconscientious users, but also use some Facebook features less [4]. Conscientiousness has been linked to the use of LBSN through Foursquare [25], being positively correlated with the number of venues visited. The nature of the Foursquare application might be especially suitable for conscientious users: they consistently remember to checkin at the venues they visit, unlike their more disorganized counterparts. There is no indication that being a consistent LBSN user increases their likelihood to checkin to the same venues, however. Previous social network and communication studies have not identified conscientiousness as playing a role in homophilous processes of other social networks [12, 4, 130]. Therefore, in terms of spatial homophily the basis for specific expectations for the conscientiousness facet to be assortative are limited. However, a conscientious user's consistent checkin behavior might increase the likelihood of detecting homophilic effects.

5.2.4 Agreeableness

Overall, agreeableness appears assortative in a communication setting, but does not seem to be specifically correlated to online behaviour or social networking site use [4, 130, 135]. It was also uncorrelated with venue checkins in Foursquare [25]. Other than a friendly atmosphere, it is difficult to speculate on what aspects of a venue attract agreeable individuals. Agreeableness is a personality facet that is most related to social interactions between acquainted individuals, which might be difficult to capture from LBSN data when the relations between users are not known. Communication between users, the only aspect that agreeable individuals have proven homophilous on [12], cannot be assessed. We therefore expect that agreeable LBSN users would not necessarily increase likelihood of attraction to similar venues.

5.2.5 Neuroticism

Neuroticism, which has been associated with a lack of perceived social support, also has a negative relationship with Internet use [145], in particular with leisure usage such as instant messaging and social gaming [6]. Neurotic people have been found to avoid discussion boards, showing little interest in participating in them online [6]. Unsurprisingly, neurotics are avoided as online interaction partners on discussion boards, even by other neurotics [12]. Whether these avoidance patterns are reflected in their spatial behavior is unclear. Emotionally stable users preferred to communicate with agreeable users, but not with each other [12]. It seems that neurotic people tend to have difficulties forming and maintaining social relationships online and offline [155]. However, neurotic individuals are speculated to be more comfortable in some online settings, as they are more likely to construe their online persona as their 'real-me' [5], which they create in LBSN by regulating their checkins [136]. This 'altered' version of their profile might therefore be an inaccurate reflection of their 'true', offline personality. Despite this, neuroticism was found to be negatively correlated with number of checkins to sociable venues [25]. We have seen little evidence for personality homophily for Neuroticism in previous chapters. No relation to spatial homophily was identified in [79] and therefore we expect to detect no spatial homophily effect for neuroticism, but one might expect highly neurotic users to be disassortative.

5.2.6 Overall personality profile

Analysing each of the five personality traits separately gives us valuable insight into homophily processes. However, given the spatial context in which homophily is being investigated, we further consider the overall personality profile. Concerning LBSNs, [62] found that participants were able to accurately predict the personality of typical visitors of a venue, solely based on images from Foursquare. Additionally, previous studies on ties in social networks found that similarity in three of the five facets (Extraversion,

agreeableness, openness to experience) promotes tie formation [138]. It remains unclear from this study whether tie formation is especially strong among people who score similarly on all three facets at once. According to [109], the stronger the connection between two people, the higher their similarity. In line with this assertion, the homophily effect appears to be especially strong among spouses and close friends [109].

However, in the present study, connection between people represents the extent of commonality (i.e., number of checkins) at a location in a LBSN, rather than a direct human relationship. To the extent of our knowledge, this is the first time connection strength has been assessed in this way. But based on previous work on close ties and personality [109, 138] and predictions based on Foursquare activity [62], there is some basis to hypothesise that increased commonality at which checkins are made positively influences overall personality similarity.

5.2.7 Summary of Hypotheses

Based on Sections 5.2.1-5.2.6 we summarise the hypotheses as follows:

1. Open users have a greater tendency to be co-located at the same venues.
2. Spatial homophily and conscientiousness are not correlated.
3. Extraverted users have a greater tendency to be co-located at the same venues.
4. Spatial homophily and agreeableness are not correlated.
5. Neurotic users have a lesser tendency to be co-located at the same venues.
6. Greater similarity in overall personality profile implies a greater tendency to to be co-located at the same venues.

5.3 Methodology

Data for this chapter was collected through the Foursquare Personality Experiment [25], which has already given important insights into the correlation between different personality facets and Foursquare use. The dataset was used to create a network of users based on their common checkins to inform the hypotheses outlined in the previous section.

5.3.1 Data collection

The data for the study was collected from an open web-based participatory study [25] that was created to examine checkin behaviour and personality of volunteer users of the Foursquare LBSN. Based on substantial software engineering, this was open to all Foursquare users, and referred to as the 'Foursquare Personality Experiment', which allowed an individual's checkin history to be assessed while undertaking a questionnaire-based assessment of the user's personality. The 'Foursquare Personality Experiment' was launched on the 19th of November 2012 [25]. It is recognised that the higher the number of items in the personality questionnaire, the more accurate the personality assessment [61] and the most recent and standard version of the Big 5 personality questionnaire, the NEO-PI-3, is comprised of 240 items [108]. However to maximise completion rates, the 44-item BFI was used [18], with answers represented on the Likert scale 1 to 5. Volunteers were incentivised by providing feedback on their personality as compared to others who checked in at the same venues.

The data collection involved participation from 218 Foursquare users. Personality data was given by 183 of these users. Of these 9 users did not have any checkins, leaving a total of 174 users for analysis. In terms of internal consistency, within the BFI questionnaire the Extraversion facet was comprised of 8 items ($\alpha = .87$), the agreeableness facet of 9 items ($\alpha = .81$), conscientiousness of 9 items ($\alpha = .82$), neuroticism of 8 items ($\alpha = .83$) and openness to experience of 10 items ($\alpha = .83$). Checkin vari-

ables from this data set were assessed in detail [25], addressing correlations concerning number of checkins, number of distinct venues visited, number of checkins at sociable venues, number of sociable venues visited and the average popularity of venues visited. Figure 5.1 shows the geographic location of the Foursquare users in the dataset. There is a good geographic spread of users across North America, Europe, and Asia, but relatively few users from Africa, South America, and Australia.

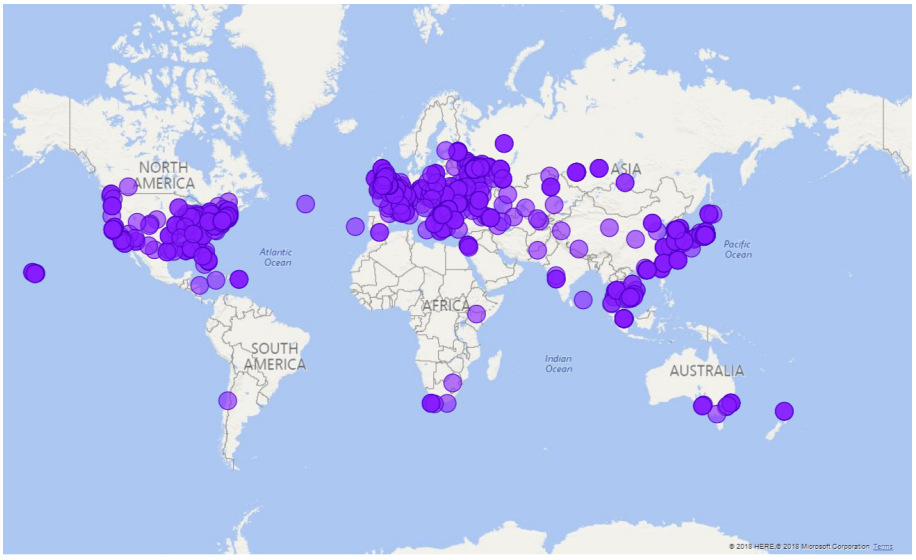


Figure 5.1: Geographic location of Foursquare users who have completed the personality questionnaire (each dot represents the latitude and longitude of the venues users who completed the questionnaire have checked into).

5.3.2 Definitions

To model spatial homophily we use a graph-based representation, defined as follows.

Definition 1. *Define a graph $G = (V, E)$ where node $v \in V$ represents a unique LBSN user, and edge $\{u, v\} \in E$ represent the common checkin of u and v at 1 or more locations. For an edge $e \in E$, let the weight of e, denoted e_w, indicate the number of common venues at which u and v have checked-in.*

Definition 2. *Let $d_G(v, w)$ be the number of edges in G adjacent to v with weight at least w. Then $G_w = (V_w, E_w)$ is the graph with $V_w = \{v \in G : d_G(v, w) > 0\}$ and where $e \in E_w$ if and only if e has edge weight of at least w.*

Graph G allows commonality between individuals, based on checkins, to be assessed. To model the relative ranking of an individual's personality score we label the nodes as described in Definition 3. Network edges can be prone to measurement errors. Repeated measurement has been suggested as a way to combat such errors [117]. If an edge is repeatedly observed to exist at different measurement times, it is likely to be a real edge. In this case, measurements are repeated over locations, rather than time, as network edges represent a common check-in. Each participant has a record of check-ins to specific venues, which have a unique id, and are recognised by the Swarm application as such. If two participants have checked in at the same venue, then they are considered to have a common check-in. For each pair of participants, the number of such common check-ins can be summed up. To this effect, G_w allows for the creation of graphs whose edges we can be more confident about. A unique observation of a co-location might be prone to measurement error, while repeated co-location at different venues gives greater confidence that two users are indeed co-located.

Definition 3. *For graph $G = (V, E)$ each node $v \in V$ is labelled with a five-dimensional vector $(v_1, \ldots, v_5)$. v_i indicates the facet value for the ith personality facet, which collectively represent openness, conscientiousness, Extraversion, agreeableness and neuroticism.*

The facet value v_i represents the tercile (first, second or third in ascending rank) in which v's personality score is categorised, relative to all nodes within V for the ith facet. We opt to use terciles in our analysis for the same reasons presented in the chapters 3 and 4: Terciles allow a clearer distinction between strong and weak scores, through which the first and third terciles can be used to test hypotheses that concern extreme values (e.g., extraverts and introverts). For similar reasons, this approach has been successfully adopted for the analysis for personality in a number of settings

(e.g., [4, 130, 135]). The cut-offs for the tercile categories for this dataset are presented in Table 5.1. Participants scoring equal or below the lower cut-off are considered low scorers, while participants scoring equal or above the upper cut-off are considered high scorers. Participants scoring in between the lower and upper cut-off are considered middle scorers.

Table 5.1: Terciles cut-offs for personality scores

tercile group	lower cut-off	upper cut-off
Openness to experience	3.60	4.20
Conscientiousness	3.11	3.55
Extraversion	2.63	3.5
Agreeableness	3.11	3.89
Neuroticism	2.50	3.25

Assessment of tercile use

Finally we check that representing personality facets by tercile, as commonly adopted in other work (e.g., [130, 4]), retains strong correlation with raw average personality scores from the completed questionnaires. A strong correlation is expected if terciles are indeed a good representation of the raw scores. Let u_i denote the ith personality facet for node u. For a pair of users u, v such that $u, v \in G$, we define the sum of absolute difference between personality profiles as $SAD_{u,v} = \sum_{i=1}^{5} |u_i - v_i|$. When facet values represent terciles (i.e., 1, 2 or 3), this metric is denoted by $SAD^T_{u,v}$. When facet values represent raw personality scores (i.e., a Likert scale rating in the range 1-5), the metric is denoted by $SAD^R_{u,v}$. For all $u, v \in G$, the correlation between $SAD^T_{u,v}$ and $SAD^R_{u,v}$ is significant and strong for all personality facets (openness: $r = .88, p = .0001$; conscientiousness: $r = .89, p = .0001$; Extraversion: $r = .91, p = .0001$; agreeableness: $r = .92, p = .0001$; neuroticism: $r = .90, p = .0001$). This provides confidence that terciles are representative of the raw personality scores.

5.3.3 Creation of the spatial network: G

From the checkin data and personality data collected in Section 5.3.1 a graph G is constructed consistent with Definitions 1 and 3. G has $|V| = 173$ and $|E| = 5373$, representing an edge density of approximately 35%. Edge weights reach a maximum of 319, with a mean of 2.92 and standard deviation of 10.85. In total 8075 unique venues are represented from 347 Foursquare venue categories. A graphical representation of the network graph G is presented in Figure 5.2, and was built using the open source graph visualisation software, Gephi [16].

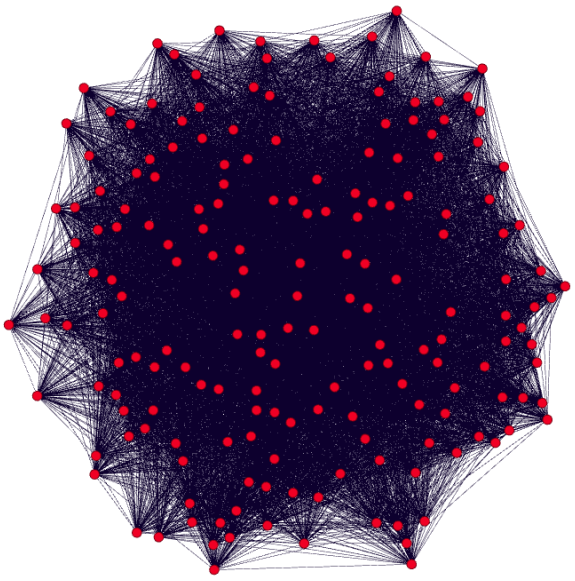


Figure 5.2: Representation of the network graph, G, with nodes representing Foursquare users and edges their co-location at one venue at least.

5.3.4 Creation of subgraphs G_w: G_1, G_2, G_6

Three subgraphs of G, G_1, G_2, and G_6 were generated according to Definition 2. $w = 1$, $w = 2$, and $w = 6$ represent meaningful cut-offs for edge weight, when

these are distributed according to terciles (Table 5.2). The edge weights were originally divided according to terciles, but the high frequencies of 1's skewed the results towards one category. The cut-offs generated by the tercile categorization were manually adjusted until the low, middle, and top group had the most similar sample sizes between them. As mentioned in Section 5.2.6, homophily effects might increase as connections between nodes grow stronger. Consequently, G_1 represents the graph with the weakest connections (1 common check-in to create an edge), G_2 represents a subgraph with moderate connections (at least 2 common check-ins to create an edge) and G_6 represents a subgraph with strong connections (at least 6 common check-ins to create an edge). We present the results of the analyses in subsequent sections for G_1, G_2 and G_6.

Table 5.2: Terciles cut-offs for edge weights, e_w

tercile group	lower cut-off	upper cut-off	N
low	1	1	1592
middle	2	5	1871
top	6	319	1910

Characteristics of G_w

G_1 has $|V_1| = 173$ and $|E_1| = 5373$, representing an edge density of approximately 36%. Mean node degree in G_1 is 62.12 (std $= 35.65$), with a range from 1 to 137. G_2 has $|V_2| = 170$ and $|E_2| = 3781$, which represents an edge density of approximately 26%. Mean node degree in G_2 is 44.48 (std $= 30.27$) with a range from 1 to 122. Finally, G_6 has $|V_6| = 164$ and $|E_6| = 1910$, with an edge density slightly above 14%. In G_6, mean node degree was 23.29 (std $= 20.34$) with a range from 1 to 85.

5.3.5 Planned analyses

We first investigate the general structure and characteristics of the three subgraphs G_w to understand how taking into account strength of connection changes the global structure of the network in Section 5.4.1. Personality scores of all Foursquare users are compared to those of a general Internet population using t-tests in Section 5.4.2. Personality scores between the different subgraphs G_w are also compared using t-tests with the expectation to find no significant differences. Finally, correlations between personality scores are also assessed in order to highlight any natural tendencies of two facets being more likely to co-occur.

To test for significance relating to the structure of G_w, we benchmark G_w against a set of random graphs R_w^*, where each graph $R_w \in R_w^*$ has the same dimensions as G_w (i.e., same number of nodes and edges). Therefore each node v in R_w corresponds to a node v in G_w, and the corresponding five dimensional facet value vector for v (Definition 3) is fixed for each v in R_w. Thus the personality profile associated with nodes in R_w remains fixed with edges randomised. We use $|R_w^*| = 1000$ and $\bar{R}$ indicates the hypothetical average graph in R^*. A similar approach is commonly used in social network analysis (e.g., [163], [37]).

For every statistic for $\bar{R}$, R_w is generated a 1000 times, and the average results of the statistic are reported. For example, co-occurrences of similar and dissimilar pairs of users on each personality facet are assessed using chi-square tests. Observed frequencies are taken from each subgraph G_w and compared to expected frequencies from its corresponding random subgraph $\bar{R}$. These expected frequencies are generated as an average of the frequencies of co-occurrences observed in 1000 different generations of R_w.

All statistical analyses are carried out using IBM SPSS 23 [74], while the data restructuring, including the building of network representations were carried out in Python, using the networkx package [152, 66]

5.4 Results

We first present the findings of the comparative analyses between the different subgraphs G_w in terms of their network characteristics in Section 5.4.1. In Section 5.4.2, we present these comparative analyses in terms of personality, and also in comparison to a general internet population. In section 5.4.3, we assess the support for each of the hypotheses defined in Section 5.2.

5.4.1 Network characteristics of each subgraph G_w

We compare the network characteristics of each subgraph to get a better understanding of their comparative structure, also in relation to their equivalent random graphs created, R_w.

Degree distribution

Degree differed significantly between $G1$, $G2$, and $G6$. Median node degree in G_2 (median=44) is significantly lower than in G_1 (median=62), $U = 10,451, Z = -4.63, p = .0001$. Median node degree in G_6 (median=16.5) is, in turn, significantly lower than in G_2, $U = 8,122, Z = -6.60, p = .0001$.

Degree for G_1 is not normally distributed ($W(173) = 0.97, p = .002$). A skewness value of ($S = 0.009$) indicates that the distribution is close to being symmetrical around the mean, suggesting that the right skew of the distribution is limited. Kurtosis values of $K = -0.93$ suggest a platykurtotic distribution, which is qualified by less extreme values at either tails and a flattening of the values around the mean, when compared to a normal distribution [38]. Degree for G_2 and G_6 follow a similar distribution with kurtosis values of $K = -0.72$ and $K = 0.19$ respectively. Skewness values were $S = 0.38$ for G_2 and $S = 0.96$ for G_6 (Figure 5.3).

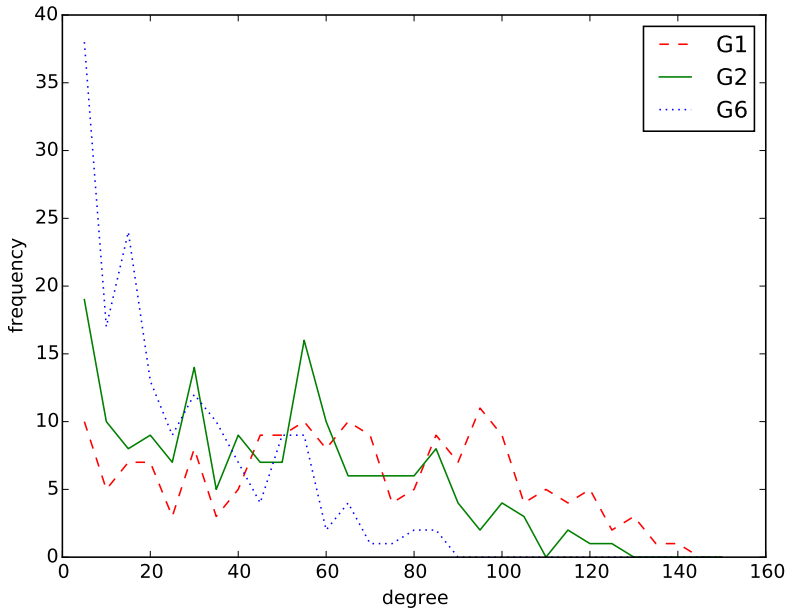


Figure 5.3: Degree distribution for G_1, G_2, and G_6

Clustering coefficients

Clustering statistics are presented in Table 5.3, and were calculated using the networkx package in Python [66], which uses the node clustering formula as defined by Saramaki et al [132]. Significantly higher clustering is seen in G_1 (mean $= .73$, std $= 0.18$) as compared to R_1 (mean $= .36$, std $= 0.00036$), with $U = 870, Z = -15.21, p = .0001$. Clustering was also higher for G_2 compared to R_2 ($U = 1566, Z = -14.46, p = .0001$) and for G_6 compared to R_6 ($U = 3828, Z = -12.06, p = .0001$). This suggests that checkins indeed have a tendency to cluster at particular locations and are not randomly distributed.

5.4.2 Personality characteristics of Foursquare users

The Foursquare users in the sample considered scored around the midpoint of 3 on the Likert scale for most personality facets as shown in Table 5.4, with the highest score

Table 5.3: Descriptives for node clustering in graphs G_w and $\bar{R}_w$

dependent variable	minimum	maximum	mean	std
observed (G_1)	0	1	.73	0.18
random ($\bar{R}_1$)	.36	.36	.36	0.00035
observed (G_2)	0	1	.70	0.22
random ($\bar{R}_2$)	.25	.25	.25	0.00045
observed (G_6)	0	1	.61	0.29
random ($\bar{R}_6$)	.13	.13	.13	0.0007

for openness to experience and the lowest score for neuroticism.

Table 5.4: Descriptives for personality scores

personality facet	**N**	**minimum**	**maximum**	**mean**	**std**
openness to experience	174	1.20	5.00	3.87	0.61
conscientiousness	174	2.00	5.00	3.43	0.65
Extraversion	174	1.13	5.00	3.15	0.84
agreeableness	174	1.89	5.00	3.56	0.64
neuroticism	174	1.00	4.50	2.91	0.73

Comparative personality scores for G_w

Personality scores from Foursquare users of G_1, G_2 and G_6 were similar to the users considered in G (Table 5.5).

Mean personality scores remained consistent across all subgraphs G_1, G_2 and G_6 even though each subgraph had fewer nodes than the parent graph, G, see figure 5.4. This gives confidence that despite reductions in sample size, subgraphs G_1, G_2 and G_6 are comparable in terms of personality.

Table 5.5: Descriptives for personality scores in G_6

personality facet	N	minimum	maximum	mean	std
openness to experience	164	1.20	4.90	3.89	0.60
conscientiousness	164	2.11	5.00	3.45	0.65
Extraversion	164	1.13	5.00	3.15	0.84
agreeableness	164	1.89	5.00	3.57	0.64
neuroticism	164	1.00	4.38	2.89	0.72

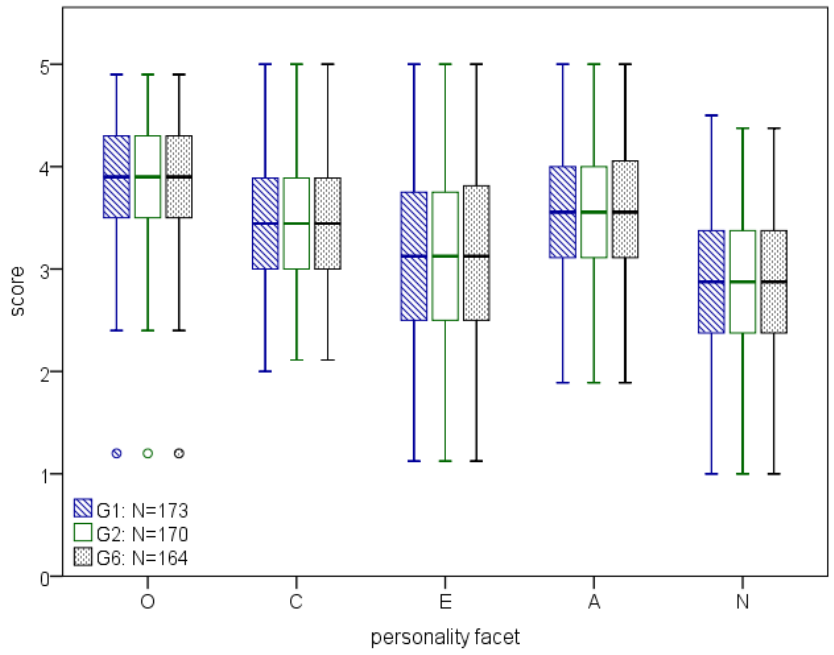


Figure 5.4: Scores of each personality facet for G_1, G_2, and G_6. O: Openness to experience, C: Conscientiousness, E: Extraversion, A: Agreeableness, N: Neuroticism.

Comparison with general internet population

We compare the aggregate personality profile for all Foursquare users (Table 5.4) with results obtained for the general Internet population [142], assuming a sample aged 30 years old. The mean and standard deviation rather than raw scores were available for each facet, and the comparison sample was larger ($N = 3007$). These

mean scores are available in [142]. Using a two-sample t-test, the two samples were compared for each facet separately. Foursquare users in our sample scored similarly on openness to experience ($t(3180), p = .12$) and marginally lower on extraversion ($t(3180) = 1.86, p = .06$). However, Foursquare users in our sample scored significantly lower on the conscientiousness facet (mean $= 3.43$, std $= 0.65$) compared to the general Internet population (mean $= 3.63$, std $= 0.72$), $t(3180) = 3.09, p = .002$. The Foursquare users in our sample also scored significantly lower on the agreeableness facet (mean $= 3.56$, std $= 0.64$) compared to the general Internet population (mean $= 3.67$, std $= 3.69$), $t(3180) = 2.25, p = .02$. Finally, Foursquare users scored significantly lower on neuroticism as well (mean $= 2.91$, std $= 0.73$), compared to the general internet population (mean $= 3.22$, std $= 0.84$), $t(3180) = 4.78, p < .0001$. However, it must be noted that effect sizes for these differences were small (conscientiousness: $d = .11$; agreeableness: $d = .08$; neuroticism: $d = .17$). In conclusion, our Foursquare sample exhibited some small, albeit significant, differences with a general internet population in terms of personality traits. Generalizability of our subsequent findings to other populations, especially non-internet ones, might therefore be limited.

Table 5.6: Pearson correlations across all personality facets of graph G, *significant at $p < .05$, **significant at $p < .001$.

	openness	conscientiousness	extraversion	agreeableness
conscientiousness	.03			
Extraversion	.29**	.26**		
agreeableness	.18*	.17*	.14	
neuroticism	-.11	-.18*	-.24*	-.32**

In Table 5.6 we present the correlation between facets for all Foursquare users. Ideally absolute correlations should be no more than around $r = |.30|$ for facets to be tested without confounding each other. All inter-facet correlations are within or around this threshold with the greatest being neuroticism and agreeableness ($r = -.32$) which is overall weak and deemed acceptable for independent analysis.

5.4.3 Assessing personality co-occurence

By considering the co-occurrence of similar personality facets at connected nodes in G_w, we are able to assess personality homophily in the context of common checkin locations. Significance is determined by comparison of G_w against $\bar{R_w}$. We firstly assess each facet in isolation, using tercile values. Only personality scores attaining the first and third terciles are considered in our analysis. This avoids ambiguity of mid-scale personality characteristics and focuses on the polar opposite strengths. Thus for graph G_w and personality facet i, all node pairs u, v where $\{u, v\} \in E_w$ and either $u_i = v_i = 1$, $u_i = v_i = 3$ or $u_i = 1$ and $v_i = 3$ are considered. The frequency of the same low facet value connections (both users scored in the 1st tercile), the same high facet value connections (both users scored in the 3rd tercile) and dissimilar facet value connections (one user scored in the 1st tercile and the other in the 3rd tercile) are assessed by comparison with $\bar{R_w}$. The results of the chi-square test on the observed frequencies (from graph G_w) and expected frequencies (from graph $\bar{R_w}$) of each combination, and for each personality facet separately, are presented in Table 5.7. This approach allows us to directly address hypotheses **H1**-**H5**. FDR-corrections were carried out using an online calculator [124], with the alpha level maintained at $.05$. The raw p-values are given in Table 5.7, but FDR-corrected p-values which are above the alpha-level are marked by an asterisk and discussed in-text.

For an individual facet, it is feasible for multiple co-occurrence relationships to be simultaneously significant. For example, given the fixed number of users in tercile 3, a significantly higher number of high facet value connections (i.e., both users in tercile 3) necessitates potentially fewer connections from such nodes to those in tercile 1, which may result in significantly lower dissimilar facet value connections (one user scored in the 1st tercile and the other in the 3rd tercile). Given these dependencies our primary focus concerns low to low or high to high facet interactions. All test statistics for the results discussed in the following paragraphs are presented in Table 5.7.

Table 5.7: Parameter Estimates for the effect of pairwise association type on observed and expected frequencies per personality facet for G_1, G_2, and G_6. Emotion. stable is an abbreviation for emotionally stable. *FDR-corrected p-values above alpha level.

	$\mathbf{G_1}$				$\mathbf{G_2}$				$\mathbf{G_6}$			
combination	NG_1	NR_1	χ^2	$p-value$	NG_2	NR_2	χ^2	$p-value$	NG_6	NR_6	χ^2	$p-value$
openness to experience												
not open- not open	530	617	13.86	**.0001**	360	406	5.84	**.016**	200	205	0.14	.71
not open-open	1336	1384	2.24	.13	962	959	0.013	.91	507	478	2.35	.13
open-open	866	751	20.47	**.0001**	626	547	13.34	**.0001**	313	270	7.98	**.005**
conscientiousness												
unconscientious-unconscientious	325	517	78.90	**.0001**	217	349	55.00	**.0001**	98	168	31.98	**.0001**
unconscientious-conscientious	1168	1306	19.26	**.0001**	781	916	26.26	**.0001**	381	449	13.46	**.0001**
conscientious-conscientious	981	797	49.88	**.0001**	723	583	39.75	**.0001**	408	287	60.04	**.0001**
Extraversion												
introverted-introverted	617	555	7.72	**.005**	417	391	1.93	.17	211	204	0.27	.60
introverted-extraverted	1332	1273	3.58	.058	940	884	4.63	.031*	468	464	0.05	.83
extraverted-extraverted	680	706	1.10	.29	489	482	0.12	.73	247	253	0.16	.69
agreeableness												
disagreeable-disagreeable	417	407	0.27	.61	302	261	6.92	**.009**	186	141	15.51	**.0001**
disagreeable-agreeable	1135	1057	7.17	**.007**	809	723	12.65	**.0001**	449	380	15.64	**.0001**
agreeable-agreeable	718	660	5.81	**.02**	505	482	1.26	.26	262	244	1.52	.22
neuroticism												
emotion. stable-emotion. stable	519	479	3.67	.055	370	349	1.39	.24	190	175	1.42	.23
emotion. stable-neurotic	1080	1106	0.77	.38	745	795	3.98	.046*	332	393	11.92	**.001**
neurotic-neurotic	520	617	17.23	**.0001**	375	435	9.35	**.002**	198	213	1.19	.28

Openness to Experience

We expected high facet value connections for openness occurring significantly more often than otherwise expected by chance. This was supported by the data for G_1 ($p = .0001$), G_2 ($p = .0001$) and G_6 ($p = .005$). On the other hand, observed frequency of low facet value connections were significantly below expectations for G_1 ($p = .0001$) and G_2 ($p = .016$), but not for G_6 ($p = .71$).

Conscientiousness

We expected high facet value connections for conscientiousness being not significantly different from chance. Contrary to expectations, conscientious users follow a similar pattern of homophily as open individuals. Observed frequency of high facet value connections was significantly above expectations for G_1, G_2, and G_6 (all $p = .0001$), while low facet value connections were significantly below expectations for G_1, G_2, and G_6 (all $p = .0001$). The observed frequency of dissimilar facet value connections is significantly below expectations for conscientiousness for G_1, G_2, and G_6 (all $p = .0001$).

Extraversion

We expected high facet value connections for Extraversion being significantly above expectation. Evidence does not support this hypothesis and interestingly it is further observed that the low facet value connections for Extraversion are significantly above expectation for G_1 ($p = .0001$), but not for G_2 ($p = .17$) and G_6 ($p = .60$). Dissimilar facet value connections were not significantly above expectations for G_2 either, when taking into account the FDR-corrected p-value ($p = .031$, FDR-corrected $p - value = .055$).

Agreeableness

We expected high facet value connections for agreeableness being insignificant as compared to expectation. This is indeed the case for G_2 ($p = .26$) and G_6 ($p = .22$), but evidence suggests that high facet value connections for agreeableness are significantly above expectation for G_1 ($p = .006$). Surprisingly, low facet value connections are significantly above expectations for G_2 ($p = .009$) and G_6 ($p = .0001$); this is, however, not the case for G_1 ($p = .61$). Dissimilar facet value connections are significantly above expectations for G_1 ($p = .007$), G_2 ($p = .0001$), and G_6 ($p = .0001$).

Neuroticism

We expected high facet value connections for neuroticism being significantly below expectation. This is supported by the data for G_1 ($p = .0001$) and G_2 ($p = .002$), but not for G_6 ($p = .28$). Dissimilar facet value connections are, on the other hand, significantly above expectation for G_6 ($p = .001$). Dissimilar facet value connections do not differ significantly from expectations for G_2 when taking into account the FDR-corrected p-value ($p = .046$, FDR-corrected $p - value = .08$)

Overall personality profile

We expected that overall personality profiles correlate with a greater tendency to checkin at common venues. This can be assessed using the *SAD* measure as a similarity metric, applying the raw personality scores as defined in Section 5.4.2. Contrary to our hypothesis, SAD scores were similar between graph G_1 (mean $= 3.95$, std $= 1.72$) and graph $\bar{R}_1$ (mean $= 3.97$, std $= 1.71$), $F(1, 10744) = 0.62$, $p = .43$. Similarly, there was no significant difference in SAD scores for G_2 ($p = .84$) and G_6 ($p = .77$) as compared to $\bar{R}_2$ and $\bar{R}_6$.

Summary of results

1. **Open users have a greater tendency to checkin at common venues.**

   Strong support. Open users were significantly co-located in G_1, G_2, and G_6. On the other hand, observed frequency of low facet value connections were significantly below expectations for G_1 and G_2, but not for G_6.

2. **Spatial homophily and conscientiousness are not correlated.**

   No support. Observed frequency of high facet value connections was significantly above expectations for G_1, G_2, and G_6, while low facet value connections were significantly below expectations for all graphs. The observed frequency of dissimilar facet value connections is significantly below expectations for G_1, G_2, and G_6.

3. **Extraverted users have a greater tendency to checkin at common venues.**

   No support. Interestingly it is further observed that the low facet value connections for Extraversion are significantly above expectation for G_1, but not for G_2 and G_6.

4. **Spatial homophily and agreeableness are not correlated.**

   Mixed support. This is indeed the case for G_2 and G_6, but evidence suggests that high facet value connections for agreeableness are significantly above expectation for G_1. Surprisingly, low facet value connections are significantly above expectations for G_2 and G_6; this is, however, not the case for G_1. Dissimilar facet value connections are significantly above expectations for G_1, G_2, and G_6.

5. **Neurotic users have a lesser tendency to checkin at common venues.**

   Mixed support. This is supported by the data for G_1 and G_2, but not for G_6. Dissimilar facet value connections are, on the other hand, significantly above expectation for G_6.

6. **Greater similarity in overall personality profile implies a greater tendency to checkin at common venues.**

   No support. SAD scores were similar between graph G_1 and graph $\bar{R}_1$. Similarly, there was no significant difference in SAD scores for G_2 and G_6 as compared to $\bar{R}_2$ and $\bar{R}_6$.

5.5 Discussion

Previous work on personality homophily has focused on the direct attraction between people with similar personality profiles, such as through evidence of particular relationships (e.g., friendships) or interactions between people (e.g., communication). In contrast, this chapter addresses personality homophily in the spatial dimension, with connections being defined through commonality of location, as indicated by checkins. Each individual effectively filters whether a visit to a location is recorded by a checkin, and the personality traits themselves could affect the emphasis an individual places on this action [25]. These issues are consistent with the new role that LBSNs play in augmenting human behavior, which has to date received relatively little attention, and results should be interpreted in this context. We note that as compared with other scenarios in which homophily has been addressed, assortative individuals in spatial homophily may be strangers, with limited or implicit awareness of the other individuals with which they assort. Existing literature has very limited coverage of this scenario, meaning that the characteristics of common locations are the indirect attractors driving personality homophily, rather than the characteristics of other LBSN users.

5.5.1 Personality facets

Overall, the hypotheses were not fully supported, which is in part reflective of the basis on which they were formulated, being informed by the dominant literature concerning

online social networks rather than homophily in the context of location-based social networks. When considering all personality facets simultaneously (**H5.6**), personality profile similarity did not correlate with common checkins.

Of the individual personality facets considered, only the hypothesis on openness was strongly supported (**H5.1**). This complements previous findings [25], where Openness to experience was found to be correlated with checkins to popular and sociable venues. Combining these observations, it is feasible that popular and sociable venues could be an underlying feature attracting open people to common locations. This is also consistent with the observation that people *low on Openness* tend to be conservative in their choices and this may manifest itself with preference for checkins at familiar locations, instead of exposure to new locations that reflect additional diversity. As a result, individuals with low Openness scores might co-locate with similar others less often, due to reduced opportunities to do so, with this reflected in checkin behavior.

Partial support was found concerning agreeableness (**H5.4**), which exhibited more homophilous tendencies than first anticipated. These unexpected results are of interest given that across the existing literature, of all the personality facets explored, findings concerning Agreeableness have generally featured the least. However this facet may have more significance for spatial homophily because disagreeableness is consistent with the inclination to be critical of others [58, 111]. This may manifest itself in specific and stringent standards for the locations they visit. As a result, disagreeable people are more inclined to visit common locations from a much smaller subset of venue types, in contrast to their agreeable counterparts.

Partial support was found for neuroticism as well (**H5.5**). By virtue of their personality, individuals high in Neuroticism are much more likely to use electronic media to present themselves favourably online [130], although they also tend to provide accurate personal information [5, 130]. Furthermore, neurotic individuals might be less inclined to visit locations in the first place, resulting in fewer opportunities to gain common checkins with others. This makes spatial homophily effects less likely to exist for

neurotic personalities, which is in line with our findings. It is interesting to note, however, that the spatial behavior of neurotics offline mirrors the communication behavior of neurotics online, in the sense that they seem to be less likely to be co-located and communicate, respectively, with one another.

No support was found concerning Extraversion (**H5.3**). This indicates that extraverts might not be commonly attracted to specific characteristics of a location, or may not be consistent in displaying checkins based on the location's characteristics. From existing literature, extraverts are known to use social media as a means to portray their social activities but it does not replace their social interactions [4], nor do they construe their online self-representation as part of their identity [5]. Consequently it is possible that these features of Extraversion are dominant in spatial homophily. Introverts, however, may pursue checkins at locations with common characteristics, which are aligned with the facet (e.g., quietness). However, it must be noted that this homophilic effect disappears with increased commonality of checkins (i.e., $w = 2, 6$) and so we discount this for further consideration.

The conscientiousness facet proved to be assortative, which was not anticipated (**H5.2**) and is of particular interest. Conscientious users follow a similar pattern of homophily as open individuals. These results extend the observation in [25] that Conscientiousness and number of checkins in Foursquare correlate, indicating that venue selection has an important role to play for this personality facet. It is possible that Conscientiousness in conducting checkins may well lead to increases in volume which in turn increase the likelihood of common checkins. However, certain characteristics of locations might be especially attractive to conscientious people, such as a well-organized, distraction-free environments, which increases the likelihood of visiting locations that have these characteristics in common, and instigating a checkin.

5.5.2 Strength of connection

We also hypothesised that connection strength could have an effect on overall personality similarity, taking into account all factors simultaneously. However, there was no significant difference between either weakly, moderately or strongly connected users, suggesting that the existence of a connection, rather than its strength, had an effect on personality similarity in the specific context of a location-based social network. In other words, even if users had only visited one common venue, they were already more likely to be similar in terms of personality, compared to users who had never been to the same venue. However, there was no difference in overall personality differences between co-located users and users who had never been to the same venues. This was assessed using the sum of absolute differences (SAD) applied to the raw score on the five factor personality profiles.

5.5.3 Future research

Results for all hypotheses, including those that are unsupported in the current study, present interesting avenues for future research. While we identified which personality facets might play a role in spatial homophily, we can only speculate on the ways these facets contribute to the observed homophily effect. For example, open individuals could be attracted to venues because they are popular or new, while introverts are attracted to quiet places. Open individuals might also value different characteristics than introverts. Atmosphere might be an important characteristic for them, while introverts value the location of the venue more, for example. Future research will have to determine whether personality indeed predicts a preference for distinct characteristics of visited venues, and identify what these characteristics precisely are.

These findings reflects the indirect nature of spatial homophily where the attraction between participants is a function of location and checkins. Personality seems to influence spatial and non-spatial homophily differently. Both for social (e.g., friend-

ship [138] or communication [12]) and spatial contexts, openness to experience appears to have a positive impact on homophily. Similarly, neuroticism appears to negatively affect homophily in both spatial and social contexts [12]. However, while Extraversion is homophilous in social contexts [138, 12], it does not appear to have any particular effect on spatial homophily. On the other hand, conscientiousness appears to play a role in spatial homophily, but not in social homophily. Finally, Agreeableness, which appears to be homophilous among friends [138] but not among online communication partners [12], did not have a significant influence on spatial homophily, as predicted. However, an interesting trend emerged with disagreeable people, who seemed to assort at common locations, while nothing in the literature seems to indicate that disagreeable people associate in social settings [12, 138]. Future research can address current shortcomings in the explanations given for the observed spatial homophily effect and what characteristics in the studied venues drive this effect.

5.5.4 Limitations

A limitation of this particular analysis is that it is much harder to capture similarity as the number of dimensions increases, and the five personality facets are only weakly correlated, making it less likely to find an effect based on their aggregated scores. A further potential issue of using SAD to measure overall personality is the loss of information. Measuring personality scores of users results in loss of information as they are the average of the aggregate scores from the 44 questionnaire items.

It could of course be argued that these findings, in particularly for disagreeableness, may occur as a consequence of the underlying LBSN database which could skew the availability of pre-existing checkin opportunities around particular locations. We feel this is unlikely given the extent of coverage of Foursquare in the developed world, and the user-generated phenomenon of venue creation leads to multiple checkins sometimes representing the same location, which diminishes the detection of spatial homophily. A further consideration is that users refrain from making checkins, resulting in

a loss of information and skewed results. The nodes in the graph-based representation of spatial homophily might therefore appear more clustered than they actually are, and clustering in G_1 is indeed high with a mean of $.72$. However, this is in line with the small world effect often found in networks with a limited amount of nodes [113]. There is also a notable absence of hubs in our graph-based representation of spatial homophily, which the small world effect also predicts [113]. Degree decreases significantly as commonality increases, while clustering stays relatively constant. A possible explanation is that increased commonality reduces the number of connected individuals in the homophily network, but does not drastically alter the interconnectedness of those same individuals.

It is important to understand the constraints that are inherent in the study, as compared to lab-based experimentation. The open participatory nature of this survey means that conventional controls are relaxed with a view to obtaining data that cannot be conveniently accessed by any other means. Selection by this mechanism is a necessary compromise that allows us to gain new insights, but these need to be interpreted with caution. To comply with data collection through the third party API, we have been unable to request personal details when collecting user data. However, the broad characteristics of Foursquare usage is consistent with early adopters of technology, who are motivated by new forms of knowledge sharing (e.g., [76]). As discussed in [25], this means that robust generalisation cannot be made to a wider population, but new insights are provided within a restricted context. As such, it is noted that Foursquare users are not necessarily representative of the general population. In Section 5.4.2 personality results from the collected data are compared with those of a general Internet population [142]. Results show that subject to particular assumptions, Foursquare users in this study were significantly lower on their Conscientiousness, Agreeableness and Neuroticism, but with a small effect size.

5.6 Conclusion

Valuable insights have been gained into the co-location patterns of people with similar personalities. The most important findings are ***highlighted***.

Our findings further consolidate the importance of individual differences in homophilic processes of social networks. We have extended the findings from chapter 4, which looked at geographic distance from a very global view, to a much more granular view: the individual venues we frequent every day. This chapter has demonstrated that homophily also acts at "street-level", with people with similar personalities assorting at common locations.

Considering the results overall, Openness and Conscientiousness persist as the most dominant personality traits that are present in spatial homophily, which is consistent with the role that LBSNs fulfil. Assortativity effects for disagreeableness are an interesting insight into possibly more selective individual decision-making that may result from critical assessment of locations. Homophilous effects were also found for Agreeableness and Neuroticism, but not for Extraversion. It is important to consider these findings in the context of LBSNs, which differ from other social media platforms in terms of their location rather than social-driven purpose.

From this investigation, we consider that there is a basis for spatial homophily as a consequence of personality, and through checkins, LBSNs provide a new form of data for its assessment. Unanticipated results for disagreeableness are of particular interest and signal possible effects concerning decision-making and location. This indicates that different venue types and distinctive characteristics may be attractors for people with particular selective tendencies. Brand associations and the local extent of alternative choice could well be influential factors in driving personality-based spatial homophily. This represents a significant new research direction that requires considerable development and to further uncover human behaviour crossing physical, psychological and technological boundaries. The results serve to reaffirm the value and power of new

forms of data obtained from mobile and social technology. In particular, the nature of spatial homophily differs considerably from homophily that captures direct attraction.

Chapter 6

Discussion and Conclusion

Through recent technologies that support online interactions, we are now able to understand the role of personality in online social networks like never before. This gives insight into how these networks are structured and how differences at the individual level drive their organisation, with particular reference to similarity and attraction, and spatial considerations. We explored how spatial elements of social networks, such as geographic distance or co-location, shape the social connections within them.

In chapter 3, we explored the relationship between people's similarity in personality and their likelihood to be connected in an online social network through direct friendship links on Facebook. In Chapter 4, geographic distance was explored at a high level in Facebook, which is primarily a social network, but which has started adopting the use of location-sharing as one of its many features. This investigation was taken a step further in Chapter 5, where the focus was on Foursquare, a LBSN, whose primary function is to share locations with friends. In this chapter, we explore the relationship between spatial and personality components through co-location at the venues people frequent. We have shown that distance, at both the global (Chapter 4) and local (Chapter 5) level play an important role for personality homophily.

In this chapter, the main findings from the thesis (Section 6.1) are summarised and their implications discussed. In section 6.2, we review: the differences that have been found between the five personality facets (Section 6.2.1); potential alternatives to the personality measures used in this thesis (Section 6.2.2); the relative advantages and dis-

advantages of using continuous and categorical measures of personality in this study (Section 6.2.3); These represent important factors that have emerged through this research. As with other personality studies, small effect sizes are typically seen for most findings, and we discuss their context and implications (Section 6.2.4). The generalizability of the results based on the samples used in this thesis is also considered (Section 6.2.5). We highlight the applications and impact of the research carried out for this thesis in Section 6.3. Finally, in Section 6.4 we draw our overall conclusions, which extend to application of the current findings and their possible impact on society.

6.1 Summary of Results

The key finding of this thesis is that users of online social media platforms have a tendency to assort according to their personality. Overall, the strongest homophily effects found in this thesis were for Openness to Experience and Conscientiousness, followed by Extraversion and Agreebleness, and finally, Neuroticism. The same personality facets that are homophilous in a social context (Facebook), were also found to be homophilous in a spatial context (Foursquare).

6.1.1 Personality homophily in Facebook (Chapter 3)

This chapter showed that **people primarily derive meaningful commonality from Openness to Experience, Agreeableness and Conscientiousness**. There was also some evidence for Agreeableness being a homophilous trait, but only in analyses that compared triads, rather than pairs. There was no effect for Neuroticism, suggesting that people do not value similarity on this trait in their friendships. For Extraversion specifically, we find no relation to strength of connection (number of mutual friends) between users (Chapter 3). This could be related to the tendency of extraverts to have more numerous, but also more superficial relationships with people. Strength of connection might not be as valued by extraverts; as a result, two users do not necessarily

need to be strongly connected for a homophily effect for Extraversion to be evident. Also of interest is the finding that overall similarity in personality, taking into account all five facets together, did not yield any significant results. It appears that users form friendships based on similarity in personality, but only for specific facets: Openness to Experience, Conscientiousness, and Agreeableness.

6.1.2 Personality homophily and geographic distance (Chapter 4)

We next explored the relationship between personality homophily and geographic distance in the same Facebook sample. Results suggest that **maintaining long distance relationships requires more effort** (pairs of mutually open or mutually conscientious users lived further apart than pairs of mutually not-open or mutually un-conscientious respectively). On the other hand, **people inclined towards face-to-face interactions are more likely to live in close proximity to their friends** (pairs of extraverted users lived closer together than pairs of introverted users). There were no clear effects for Neuroticism or Agreeableness. Differences in personalities might account for these effects. Not-open people might be more likely to stay in familiar places, while open individuals might travel more and be more inclined to move away from home. Extraverts like being able to make plans with friends, while introverts do not need the same level of face-to-face interaction to maintain friendships. Conscientious people might be more organised in keeping in touch with distant others.

6.1.3 Personality Homophily and local co-location (Chapter 5)

At a more local level of geographic distance (co-location at the same venue), we find that **the same facets identified in the Facebook sample are also homophilous in the Foursquare sample**. Open, conscientious, introverted, and agreeable users were most likely to be found at venues with patrons that shared their personality traits. On the other hand, neurotic and not-open users almost seemed to avoid venues with sim-

ilar others, indicating they might be disassortative. Also of interest is the fact that strength of connection influenced these results. For example, the homophily effect for agreeableness disappeared when only taking into account pairs that were strongly co-located. Interestingly, disagreeable pairs became homophilous when they had a lot of venues in common. Together, the findings seem to be in line with the literature as homophily effects are expected for Openness to Experience, Extraversion, and Agreeableness. Conscientiousness is correlated with Foursquare use, so effects for this facet tie in nicely with the more LBSN-oriented literature [26].

6.2 Discussion and Future Work

In the following sections, we discuss the limitations and methodological choices made in this thesis and provide avenues for future research that can build upon the foundation laid out in this thesis.

6.2.1 Consistency with the literature

This thesis found a small number of differences in the importance of facets compared to previous work. Firstly, the findings for Conscientiousness are surprising, as this facet did not previously emerge as homophilous. Social media platforms can be very specialised, which might highlight different personality facets. For example, Facebook is very focused on social interactions with friends and families, while Foursquare is very location-focused. Personality facets that are homophilous in these contexts might be slightly different from the ones in offline networks [138] or discussion boards [12]. On the other hand, the finding that Openness to Experience is consistently homophilous is in line with previous work [138, 12]. Extraversion and Agreeableness were also identified as facets likely to be homophilous alongside Openness to Experience. However, results for these two facets were not as consistent in this thesis as for Openness

to Experience. Future work on different contexts for personality homophily could give valuable insights into how different facets might be expressed differently in varying contexts.

Finally, very little evidence of homophily was found for Neuroticism, which is in line with the literature. It seems that Neuroticism even had disassortative effects among Foursquare users: people who were neurotic were less likely to frequent the same locations as similar others (Chapter 5). Neurotic individuals have trouble building social relationships in general because of their anxiety and proneness to worry. This might explain the lack of homophily on this facet, as two neurotic users might be reinforcing each other and break their relationship more easily, resulting in no evidence of a connection in Facebook or Foursquare. This is consistent with previous literature who found that neurotic users were disassortative in an online communication network [12].

High and Low Scorers

Another consideration that needs to be addressed in this regard is that of high scorers and low scorers. It might not be enough to consider whether two users are similar on a specific facet, but also how their actual score might influence personality homophily dynamics. For example, a pair of introverted users might act very differently from a pair of extraverted userst.

In Chapter 3, we find that high scorers tend to be homophilous while low scorers are not for some facets, but the opposite for other facets. For geographic distance (Chapter 4), we also find differences depending on whether people are low or high scorers. In fact, we find more significant results between similar low-scoring pairs and similar high-scoring pairs, than between similar and mixed pairs. In Chapter 5, mixed pairs were as prevalent as expected by chance levels. Only mixed pairs on conscientiousness were significantly less prevalent than expected. This weakens the conclusion for personality homophily across all facets, except Conscientiousness in Chapter 5.

In other words, while we find some personality homophily effects, especially in the Facebook network, it might interesting to focus on the differences uncovered between low-scoring users (not-open, introverted, disagreeable, neurotic, unconscientious) and high-scoring users (open, extraverted, agreeable, emotionally stable, conscientious) and their dynamics with each other.

Overall Personality Profile

Overall, it seems that individual facets drive homophily, rather than similarity on the overall personality profile. Of course, each user has a full personality profile, which could interfere with looking at each facet separately. Two users could have the same score on Extraversion, but be opposites on all other facets. Would these two users be more likely to be connected than two users who score the same on each facet? This does not appear the case, both in the Facebook dataset (Chapter 3) and the Foursquare dataset (Chapter 5). However, both chapters had an all-or-nothing approach, with either considering each facet separately or all of them together. It would have been preferable to compare pairs on two, three, or four facets. However, this would have yielded a significant amount of comparisons to carry out because of the number of possible combinations. In total, 100 comparisons would have been made, if all possible combinations were to be taken into account (45 combinations for two facets, 45 for three facets, 10 for four facets). An alternative, of course, would be to compare pairs based on the number of common facet pairs, instead of looking at each combination separately. This would require a much lighter analysis. Future work could investigate whether having more than one facet in common contributes to stronger, or more likely, connections between people.

6.2.2 Personality measures used

The analyses carried out in this thesis are based on self-reported survey-based personality measures from the Five Factor model of personality. With personality surveys, a balance must be struck between boredom, which is a problem for longer scales, and accuracy, which is a problem for shorter scales . The 44-item BFI seems to be a good length to guarantee accuracy and engagement from the person completing the survey.

Using a mix of measures for personality might be detrimental to the personality comparisons being carried out, as different surveys might yield different results. However, some diversity in personality measurement might be a positive thing. All IPIP based scales should yield very similar results and are constructed so that they reach similar conclusions regardless of the actual items used [59]. Diversity might also give us confidence in the interpretation of the results. Personality homophily has been found in both the mixed scale sample (Facebook) and the BFI sample (Foursquare), showing that regardless of the measurement used, homophily was found. This consistency shows that homophily is pervasive across different contexts and ways of measuring personality.

A criticism of survey-based personality questionnaires is that people's attitudes do not always translate to their behavior. We acknowledge that behavioral measures of personality are lacking in this thesis. However, some behavioral measures of personality are actually captured and confounded by other measures, such as network size or mobility. For example, the size of a person's network could be used as a behavioral expression of how extraverted a person is. In Chapter 3, we replicate the well-known correlation between Extraversion and number of friends. However, as number of friends was a measure of interest, we could not have used it as a proxy for Extraversion in this case. Different behavioral measures would have needed to be combined to build a good model of each personality facet, which was not the focus on the current thesis, but would be extremely valuable as future work.

Getting accurate personality measures constitutes a non-trivial task. It would have been valuable to supplement self-reported measures with measures from close friends or family. People might, even unconsciously, portray themselves as closer to their ideal selves than they actually are. Friends and family might also be biased, however. People tend to inject their own personality when rating others, a phenomenon called the self-based heuristic. This is especially prevalent for traits that are difficult to rate [125]. On the other hand, more independent observers might not have enough knowledge of the person to make an accurate judgement. It must also be noted that anonymity is an important aspect of social media-based studies, and asking friends for personality-ratings would have complicated the collection of data from Facebook and from the Foursquare application.

From a more philosophical perspective, we are the only ones who can truly know ourselves. We are the only ones who actually know what we think and how we behave at any moment in our lives. There are always parts of our personality which are hidden from external observers, either because they are not relevant, or because they are voluntarily concealed. With this consideration in mind, self-reported personality might be the closest we can get to one's true personality, despite the inherent biases. Survey-based personality measures are still widely used in the personality literature. We think they work as an acceptable proxy for a person's personality, although we acknowledge that other measures could have been used to supplement the self-reported ones.

6.2.3 Terciles and continuous scores

Throughout the thesis, we have used a mix of categorical and continuous measures to qualify the personality of users. The tercile approach was usually preferred, and consisted of dividing people into low, middle, and high scorers on each facet. Typically, only low and high scorers were then considered for the analyses, as this allowed us to compare meaningful personality *types* such as introverts and extraverts. The literature often focuses on extremes [130, 4, 12] making comparisons with existing findings

easier. Research has also found stronger effect sizes when considering low and high scorers rather than continuous scores [8], which might especially valuable for studies in which small effects are to be expected. The downside of categorization is that it can lead to loss of information which can lead to wrongful conclusions. On the other hand, categorization can allow us to make more clear-cut comparisons between scores which are particularly interesting because they fall on each extreme of the scale (low and high scorers). Additionally, categorization can be beneficial in statistical analyses whose effect sizes are known to be small.

Continuous measures have also been used in this thesis, to supplement the categorical ones, such as the variance in personality within a triad in Chapter 3 or SAD scores (sum of absolute differences) in Chapter 5. Continuous and categorical results often lead to similar conclusions. In Chapter 3, results with continuous scores (variance among triangles) actually lead to significant results for more facets than with categorical measures (comparing the prevalence of same pairs compared to mixed pairs). This shows the importance to consider both continuous and categorical measures for personality-based research.

6.2.4 Effect Sizes

Most effect sizes reported throughout the chapters are small. This is consistent with personality studies in which small effect sizes are common [4, 130]. This does not imply that the results are meaningless. It simply means that the personality homophily effects in the samples used are not strong. However, they are of high value because they are relatively consistent. Throughout the thesis, we find support for personality homophily for all five facets, except Neuroticism. This is across a range of very different samples (Facebook users and Foursquare users) and contexts. We consider global geographic distance, as well as local co-location, and in each context, we find significant effects regarding personality. Studies with large sample sizes do not need to rely on strong effect sizes to find meaningful effects in the data [105]. Consistency, rather than

strength of effect is what makes these findings interesting and worthwhile investigating further.

The problem with small effect sizes is that they often need large sample sizes to be detected [105]. This is not a problem in our samples, as the data collection on online social media allowed us to reach an acceptable number of participants. This might also be the reason why personality homophily is a a relatively recently studied phenomenon, as large-scale personality collection was not possible before the advent of the internet and mobile computing. The need for a large sample size might hinder replication studies, especially offline ones, for whom large samples are difficult to obtain. Nonetheless, with the advent of Big Data and the possibility to collect large amounts of data, obtaining large samples will likely be a problem of the past.

Alpha-level and p-value corrections

The problem of multiple comparisons and the increased chance of Type I errors has to be addressed in this thesis. In other words, the more statistical tests carried out, the more likely it is to find spurious significant effects. To correct for this bias, we have used Bonferroni corrections in Chapters 3, 4, and 5. In Chapter 4, we additionally considered FDR-corrected p-values with the Benjamini-Hochberg procedure. While Bonferroni corrections are applied before the tests are being carried out and change the alpha-level, FDR p-value corrections are carried out afterwards and modify the p-values instead. The problem with Bonferroni corrections is that they might be too conservative and not reject null hypotheses when they should. FDR corrections are more lenient, on the other hand. It must be noted that at the number of tests carried out in each Chapter, both approaches would lead to similar results, as differences only start to emerge at an excessive number of comparisons (100 tests and more). For example, in Chapter 4, FDR-corrected p-values made one result significant (see Table 4.7), which would have been considered not significant with the Bonferroni-approach. Whether a conservative approach or a more lenient approach is preferred is debat-

able. More lenient corrections might be especially interesting in exploratory research, while a more conservative approach might be preferable for one-tailed hypotheses. It must also be considered that a more conservative approach might be preferable to make conclusions stronger despite expected small effect sizes. Most hypotheses in this thesis are one-tailed, so we think that Bonferroni corrections were appropriate. The Bonferroni-approach also helps set an alpha-level from the start whose interpretation is unequivocal. With the Benjamini-Hochberg procedure, interpretations can change after all results are in, as the correction relies on the distribution of all p-values.

In short, while FDR corrections are very valuable, we think that a more conservative approach using Bonferroni corrections was more appropriate in this case. We encourage future work to look into personality homophily in different contexts (online, offline, different platforms) and through different methods of observation (survey, observation, experimental), and analyses (continuous or categorical measures, different tests and corrections) to see which effects survive and remain consistent across different contexts, methodologies, and analyses.

6.2.5 Generalizability of results

It is important to consider that the samples were collected from a very specific slice of the population, and are online-based rather than from a more traditional offline approach.

The advantages of online-based social networks is that they allow for the collection of organic and non-invasive connections between people, with their explicit consent. Collection of offline network connections typically require participants to write down manually all the people they can think of as being close to them. Following Dunbar's network model, researchers might only collect the most inner layers of the network to facilitate data collection. This approach often neglects the outer layers, which can be very important in certain aspects of life [63]. People can also forget who they are

connected to on Facebook, but the digital trace of that connection enables researchers to rely on that rather than people's memories. An alternative offline method which also circumvents memory problems are observational measures where people are being seen interacting with each other. The edges of such a network can be weighted with the frequency and length of observed interactions to qualify the edge in terms of connection strength. However, such an approach is often carried out in very specific contexts, such as play-interactions between school children of the same class. These settings are self-contained in both people (the pupils in the class) and location (the playground), making such observations possible. Observational work using adults in real-life situations would be close to impossible. Observational networks also do not allow for us to qualify the type of relationships between people. Two classmates might interact frequently, but their interactions might be antagonistic, rather than friendly. A connection will be present in the observational network, which would be absent in the survey-based one or a Facebook-based one.

The downside of online connections is that they do not come with a measure of emotional closeness, which is often the case for survey-based offline network studies. This means that in the Facebook network, a connection between two people in a couple will be seen as the same as two people who once went to high school together but have not interacted since graduation. Gaps in the network will also be frequent, as people who are not on Facebook or Foursquare cannot be captured. This is one of the main reasons Chapter 3 and 4 with the Facebook sample have focused on triadic and pairwise interactions, rather than on the reconstruction of a complete social network which was bound to have many gaps. For the Foursquare network, a reconstruction was possible, as connections were based on co-location rather than electronic friendship ties.

Network data is bound to be error-prone, by capturing links that do not really exist or missing ones that do. A recent paper by Newman discusses the implications of noisy network data and proposes ways of combatting errors that arise from data collection [117]. Especially non-connections are interesting in this respect. The way the Face-

book sample was collected, it was not possible to tell whether a non-existent connection was due to two people never being in touch in the first place, or because two people wilfully unfriended each other. Both scenarios have very different consequences for the interpretation of the results presented in this thesis. The absence of a connection, as opposed to an unfriending action, also does not clarify whether two people never had the chance to meet (absence of induced homophily), or whether two people met and decided that they would not befriend each other on facebook, because of personal differences. One proposition is to collect data at different points in time and check which links in the network remain consistent. We have somewhat used this approach in Chapter 5 by constructing different graphs with an increasing number of co-locations necessary to make an edge (1 for $G1$, 2 for $G2$, and 6 for $G6$). Errors might be especially prevalent in $G1$, as the application might have recorded an erroneous co-location. However, with 2 co-locations registered, errors are already less likely. With 6 locations in common, it is actually very likely that two people were co-located at some point, as an erroneous co-location for 6 different venues is starting to get extremely unlikely.

More care could have been given to the Facebook network in this respect, however, we have focused on pairwise connections and triads for this reason. Facebook friendships are unlikely to be erroneous as they take two people (one requesting, the other accepting) to be made. Of course, the question of emotional closeness between these two people arises. We tried to model this with the number of shared friends between people, as people are likely to meet their friends' friends and become friends in turn [75]. However, this does not fully capture people's real closeness to each other and future work would have to focus on the emotional closeness aspect of Facebook friendships in regards to personality homophily.

6.3 Research Impact

6.3.1 Novelty

This thesis builds on previous work carried out on personality homophily [137, 12]. Work that focuses on personality homophily remains sparse in the literature, with most insights regarding personality in online spaces gained from work surrounding the impact of personality on social media use. This thesis aims to fill this gap by having conducted an extensive investigation of personality homophily in both a social context through Facebook, and a spatial one, through Foursquare.

Distance and how much it truly matters for online interactions has been studied before [133]. The geographic idiosyncrasies associated with personality have also been the subject of in-depth studies at different levels of granularity [3, 62, 79, 126]. However, this thesis is the first to consider spatial elements of connections built on personality homophily specifically. We do this by taking into consideration geographic distance between users connected on the basis of personality homophily. We also consider co-location and how it might drive personality homophily, which has not been studied before.

In this thesis we have focused on pairwise connections and triads as meaningful components of the network. The study of recurring patterns in networks is an exciting and ever-evolving field of study, which places emphasis on the local, rather than the global characteristics of a network.

6.3.2 Applications

The importance of homophily is ubiquitous across recommendation systems. Over the course of the thesis, we have shown that homophily is not equally strong for all five facets. It is therefore important to consider facets separately rather as a whole when using personality for a recommendation system.

In Chapter 5 we have shown that people with similar personality traits are likely to frequent the same venues. Recommendation systems for location-based entertainment such as museums, restaurants, and so on could benefit from matching potential patrons to venues based on their personality. Future work would need to investigate what aspects of venue selection are driven by personality to make these recommendations systems truly effective.

We speculated that disagreeable people tend to frequent the same venues because they have high expectations regarding service and quality. Introverted people might be attracted to quiet locations which suit their personality. It is possible that no clear effect was found for extraverts because they feel at ease in a range of different venues. Open individuals might be attracted to diverse or unique locations. However, these remain speculations as future work will have to determine whether this is indeed the case. Some work has been carried out on the relationship between personality of patrons and the venues they visited, but these only dealt with the perception of patrons' personality matching the venues' ambiance, rather than the actual personality of the patrons [62]. Based on the findings in Chapter 5, a more in-depth investigation of venue types and their links to personality could be an interesting avenue for new research.

Beyond recommendation systems, we have gained insights into relationship formations and the importance of homophily in online social networks. The understanding that not all personality facets are equal in their impact is crucial. This is especially true for the Conscientiousness facet. Conscientiousness is traditionally understudied as an influential factor in social media, because it is well known that users high in conscientiousness avoid the distractions that come from social media platforms [130]. Of particular interest is the behaviour of conscientious people who still partake in social media, but might have a more healthy relationship with it compared to a neurotic user for example.

We also highlight the value of distinguishing people according to their personality scores into extraverts and introverts, or open and not-open. People tend to identify as

either an introvert or an extravert, but rarely know their exact personality score. This facilitates interpretation as people are given understandable personality qualities rather than an abstract survey score.

6.3.3 Impact on Society

Personality is an ubiquitous part of everyone's lives, as are social connections. Understanding how they influence one another can help us understand when relationships fall apart, or on the contrary, when they are successful. Personality disorders can have devastating effects on those who suffer from them and the people around them [28]. People who suffer from personality disorders often end up on the fringes of their social networks, with little support from others. This becomes a vicious cycle where the disadvantageous network position enhances the expression of the personality disorder. While personality is mostly static, improving one's social network position can have a beneficial effect on the way we express our personality and its consequences. Homophily is the glue that holds social networks together, and associating with others we are compatible with can have beneficial consequences for our well-being and social network.

6.4 Final Conclusion

This thesis provides evidence, within the assumptions and limitations previously stated, that there is a small but significant relationship between personality and the structure of our online interactions, not only socially, but also spatially. We have shown that personality homophily is evident in a Facebook network, with geographic distance playing a role for some of the facets. We have also shown that at a more local geographic level, clear homophily effects emerge for some of the personality facets, but not all. Taken together, these findings demonstrate that our online social networks are highly structured and reflect inherent characteristics of human nature.

GNU Free Documentation License

Version 1.2, November 2002

0. Preamble

The purpose of this License is to make a manual, textbook, or other functional and useful document *free* in the sense of freedom: to assure everyone the effective freedom to copy and redistribute it, with or without modifying it, either commercially or non-commercially. Secondarily, this License preserves for the author and publisher a way to get credit for their work, while not being considered responsible for modifications made by others.

This License is a kind of "copyleft", which means that derivative works of the document must themselves be free in the same sense. It complements the GNU General Public License, which is a copyleft license designed for free software.

We have designed this License in order to use it for manuals for free software, because free software needs free documentation: a free program should come with manuals providing the same freedoms that the software does. But this License is not limited to software manuals; it can be used for any textual work, regardless of subject matter or whether it is published as a printed book. We recommend this License principally for works whose purpose is instruction or reference.

1. Applicability and Definitions

This License applies to any manual or other work, in any medium, that contains a notice placed by the copyright holder saying it can be distributed under the terms of

this License. Such a notice grants a world-wide, royalty-free license, unlimited in duration, to use that work under the conditions stated herein. The "Document", below, refers to any such manual or work. Any member of the public is a licensee, and is addressed as "you". You accept the license if you copy, modify or distribute the work in a way requiring permission under copyright law.

A "Modified Version" of the Document means any work containing the Document or a portion of it, either copied verbatim, or with modifications and/or translated into another language.

A "Secondary Section" is a named appendix or a front-matter section of the Document that deals exclusively with the relationship of the publishers or authors of the Document to the Document's overall subject (or to related matters) and contains nothing that could fall directly within that overall subject. (Thus, if the Document is in part a textbook of mathematics, a Secondary Section may not explain any mathematics.) The relationship could be a matter of historical connection with the subject or with related matters, or of legal, commercial, philosophical, ethical or political position regarding them.

The "Invariant Sections" are certain Secondary Sections whose titles are designated, as being those of Invariant Sections, in the notice that says that the Document is released under this License. If a section does not fit the above definition of Secondary then it is not allowed to be designated as Invariant. The Document may contain zero Invariant Sections. If the Document does not identify any Invariant Sections then there are none.

The "Cover Texts" are certain short passages of text that are listed, as Front-Cover Texts or Back-Cover Texts, in the notice that says that the Document is released under this License. A Front-Cover Text may be at most 5 words, and a Back-Cover Text may be at most 25 words.

A "Transparent" copy of the Document means a machine-readable copy, represented in a format whose specification is available to the general public, that is suitable for revising the document straightforwardly with generic text editors or (for images composed of pixels) generic paint programs or (for drawings) some widely available drawing editor, and that is suitable for input to text formatters or for automatic translation to a variety of formats suitable for input to text formatters. A copy made in an otherwise Transparent file format whose markup, or absence of markup, has been arranged to thwart or discourage subsequent modification by readers is not Transparent. An image format is not Transparent if used for any substantial amount of text. A copy that is not "Transparent" is called "Opaque".

Examples of suitable formats for Transparent copies include plain ASCII without markup, Texinfo input format, LaTeX input format, SGML or XML using a publicly available DTD, and standard-conforming simple HTML, PostScript or PDF designed for human modification. Examples of transparent image formats include PNG, XCF and JPG. Opaque formats include proprietary formats that can be read and edited only by proprietary word processors, SGML or XML for which the DTD and/or processing tools

are not generally available, and the machine-generated HTML, PostScript or PDF produced by some word processors for output purposes only.

The "Title Page" means, for a printed book, the title page itself, plus such following pages as are needed to hold, legibly, the material this License requires to appear in the title page. For works in formats which do not have any title page as such, "Title Page" means the text near the most prominent appearance of the work's title, preceding the beginning of the body of the text.

A section "Entitled XYZ" means a named subunit of the Document whose title either is precisely XYZ or contains XYZ in parentheses following text that translates XYZ in another language. (Here XYZ stands for a specific section name mentioned below, such as "Acknowledgements", "Dedications", "Endorsements", or "History".) To "Preserve the Title" of such a section when you modify the Document means that it remains a section "Entitled XYZ" according to this definition.

The Document may include Warranty Disclaimers next to the notice which states that this License applies to the Document. These Warranty Disclaimers are considered to be included by reference in this License, but only as regards disclaiming warranties: any other implication that these Warranty Disclaimers may have is void and has no effect on the meaning of this License.

2. Verbatim Copying

You may copy and distribute the Document in any medium, either commercially or noncommercially, provided that this License, the copyright notices, and the license notice saying this License applies to the Document are reproduced in all copies, and that you add no other conditions whatsoever to those of this License. You may not use technical measures to obstruct or control the reading or further copying of the copies you make or distribute. However, you may accept compensation in exchange for copies. If you distribute a large enough number of copies you must also follow the conditions in section 3.

You may also lend copies, under the same conditions stated above, and you may publicly display copies.

3. Copying in Quantity

If you publish printed copies (or copies in media that commonly have printed covers) of the Document, numbering more than 100, and the Document's license notice requires Cover Texts, you must enclose the copies in covers that carry, clearly and legibly, all these Cover Texts: Front-Cover Texts on the front cover, and Back-Cover Texts on the

back cover. Both covers must also clearly and legibly identify you as the publisher of these copies. The front cover must present the full title with all words of the title equally prominent and visible. You may add other material on the covers in addition. Copying with changes limited to the covers, as long as they preserve the title of the Document and satisfy these conditions, can be treated as verbatim copying in other respects.

If the required texts for either cover are too voluminous to fit legibly, you should put the first ones listed (as many as fit reasonably) on the actual cover, and continue the rest onto adjacent pages.

If you publish or distribute Opaque copies of the Document numbering more than 100, you must either include a machine-readable Transparent copy along with each Opaque copy, or state in or with each Opaque copy a computer-network location from which the general network-using public has access to download using public-standard network protocols a complete Transparent copy of the Document, free of added material. If you use the latter option, you must take reasonably prudent steps, when you begin distribution of Opaque copies in quantity, to ensure that this Transparent copy will remain thus accessible at the stated location until at least one year after the last time you distribute an Opaque copy (directly or through your agents or retailers) of that edition to the public.

It is requested, but not required, that you contact the authors of the Document well before redistributing any large number of copies, to give them a chance to provide you with an updated version of the Document.

4. Modifications

you may copy and distribute a Modified Version of the Document under the conditions of sections 2 and 3 above, provided that you release the Modified Version under precisely this License, with the Modified Version filling the role of the Document, thus licensing distribution and modification of the Modified Version to whoever possesses a copy of it. In addition, you must do these things in the Modified Version:

A. Use in the Title Page (and on the covers, if any) a title distinct from that of the Document, and from those of previous versions (which should, if there were any, be listed in the History section of the Document). You may use the same title as a previous version if the original publisher of that version gives permission.

B. List on the Title Page, as authors, one or more persons or entities responsible for authorship of the modifications in the Modified Version, together with at least five of the principal authors of the Document (all of its principal authors, if it has fewer than five), unless they release you from this requirement.

C. State on the Title page the name of the publisher of the Modified Version, as the publisher.

D. Preserve all the copyright notices of the Document.

E. Add an appropriate copyright notice for your modifications adjacent to the other copyright notices.

F. Include, immediately after the copyright notices, a license notice giving the public permission to use the Modified Version under the terms of this License, in the form shown in the Addendum below.

G. Preserve in that license notice the full lists of Invariant Sections and required Cover Texts given in the Document's license notice.

H. Include an unaltered copy of this License.

I. Preserve the section Entitled "History", Preserve its Title, and add to it an item stating at least the title, year, new authors, and publisher of the Modified Version as given on the Title Page. If there is no section Entitled "History" in the Document, create one stating the title, year, authors, and publisher of the Document as given on its Title Page, then add an item describing the Modified Version as stated in the previous sentence.

J. Preserve the network location, if any, given in the Document for public access to a Transparent copy of the Document, and likewise the network locations given in the Document for previous versions it was based on. These may be placed in the "History" section. You may omit a network location for a work that was published at least four years before the Document itself, or if the original publisher of the version it refers to gives permission.

K. For any section Entitled "Acknowledgements" or "Dedications", Preserve the Title of the section, and preserve in the section all the substance and tone of each of the contributor acknowledgements and/or dedications given therein.

L. Preserve all the Invariant Sections of the Document, unaltered in their text and in their titles. Section numbers or the equivalent are not considered part of the section titles.

M. Delete any section Entitled "Endorsements". Such a section may not be included in the Modified Version.

N. Do not retitle any existing section to be Entitled "Endorsements" or to conflict in title with any Invariant Section.

O. Preserve any Warranty Disclaimers.

If the Modified Version includes new front-matter sections or appendices that qualify as Secondary Sections and contain no material copied from the Document, you may at your option designate some or all of these sections as invariant. To do this, add their titles to the list of Invariant Sections in the Modified Version's license notice. These titles must be distinct from any other section titles.

You may add a section Entitled "Endorsements", provided it contains nothing but endorsements of your Modified Version by various parties — for example, statements of peer review or that the text has been approved by an organization as the authoritative definition of a standard.

You may add a passage of up to five words as a Front-Cover Text, and a passage of up to 25 words as a Back-Cover Text, to the end of the list of Cover Texts in the Modified Version. Only one passage of Front-Cover Text and one of Back-Cover Text may be added by (or through arrangements made by) any one entity. If the Document already includes a cover text for the same cover, previously added by you or by arrangement made by the same entity you are acting on behalf of, you may not add another; but you may replace the old one, on explicit permission from the previous publisher that added the old one.

The author(s) and publisher(s) of the Document do not by this License give permission to use their names for publicity for or to assert or imply endorsement of any Modified Version.

5. Combining Documents

You may combine the Document with other documents released under this License, under the terms defined in section 4 above for modified versions, provided that you include in the combination all of the Invariant Sections of all of the original documents, unmodified, and list them all as Invariant Sections of your combined work in its license notice, and that you preserve all their Warranty Disclaimers.

The combined work need only contain one copy of this License, and multiple identical Invariant Sections may be replaced with a single copy. If there are multiple Invariant Sections with the same name but different contents, make the title of each such section unique by adding at the end of it, in parentheses, the name of the original author or publisher of that section if known, or else a unique number. Make the same adjustment to the section titles in the list of Invariant Sections in the license notice of the combined work.

In the combination, you must combine any sections Entitled "History" in the various original documents, forming one section Entitled "History"; likewise combine any sections Entitled "Acknowledgements", and any sections Entitled "Dedications". You must delete all sections Entitled "Endorsements."

6. Collections of Documents

You may make a collection consisting of the Document and other documents released under this License, and replace the individual copies of this License in the various documents with a single copy that is included in the collection, provided that you follow the rules of this License for verbatim copying of each of the documents in all other respects.

You may extract a single document from such a collection, and distribute it individually under this License, provided you insert a copy of this License into the extracted document, and follow this License in all other respects regarding verbatim copying of that document.

7. Aggregation with Independent Works

A compilation of the Document or its derivatives with other separate and independent documents or works, in or on a volume of a storage or distribution medium, is called an "aggregate" if the copyright resulting from the compilation is not used to limit the legal rights of the compilation's users beyond what the individual works permit. When the Document is included in an aggregate, this License does not apply to the other works in the aggregate which are not themselves derivative works of the Document.

If the Cover Text requirement of section 3 is applicable to these copies of the Document, then if the Document is less than one half of the entire aggregate, the Document's Cover Texts may be placed on covers that bracket the Document within the aggregate, or the electronic equivalent of covers if the Document is in electronic form. Otherwise they must appear on printed covers that bracket the whole aggregate.

8. Translation

Translation is considered a kind of modification, so you may distribute translations of the Document under the terms of section 4. Replacing Invariant Sections with translations requires special permission from their copyright holders, but you may include translations of some or all Invariant Sections in addition to the original versions of these Invariant Sections. You may include a translation of this License, and all the license notices in the Document, and any Warranty Disclaimers, provided that you also include the original English version of this License and the original versions of those notices and disclaimers. In case of a disagreement between the translation and the original version of this License or a notice or disclaimer, the original version will prevail.

If a section in the Document is Entitled “Acknowledgements”, “Dedications”, or “History”, the requirement (section 4) to Preserve its Title (section 1) will typically require changing the actual title.

9. Termination

You may not copy, modify, sublicense, or distribute the Document except as expressly provided for under this License. Any other attempt to copy, modify, sublicense or distribute the Document is void, and will automatically terminate your rights under this License. However, parties who have received copies, or rights, from you under this License will not have their licenses terminated so long as such parties remain in full compliance.

10. Future Revisions of this License

The Free Software Foundation may publish new, revised versions of the GNU Free Documentation License from time to time. Such new versions will be similar in spirit to the present version, but may differ in detail to address new problems or concerns. See `http://www.gnu.org/copyleft/`.

Each version of the License is given a distinguishing version number. If the Document specifies that a particular numbered version of this License “or any later version” applies to it, you have the option of following the terms and conditions either of that specified version or of any later version that has been published (not as a draft) by the Free Software Foundation. If the Document does not specify a version number of this License, you may choose any version ever published (not as a draft) by the Free Software Foundation.

ADDENDUM: How to use this License for your documents

To use this License in a document you have written, include a copy of the License in the document and put the following copyright and license notices just after the title page:

> Copyright © `YEAR YOUR NAME`. Permission is granted to copy, distribute and/or modify this document under the terms of the GNU Free Documentation License, Version 1.2 or any later version published by the Free

> Software Foundation; with no Invariant Sections, no Front-Cover Texts, and no Back-Cover Texts. A copy of the license is included in the section entitled "GNU Free Documentation License".

If you have Invariant Sections, Front-Cover Texts and Back-Cover Texts, replace the "with... Texts." line with this:

> with the Invariant Sections being `LIST THEIR TITLES`, with the Front-Cover Texts being `LIST`, and with the Back-Cover Texts being `LIST`.

If you have Invariant Sections without Cover Texts, or some other combination of the three, merge those two alternatives to suit the situation.

If your document contains nontrivial examples of program code, we recommend releasing these examples in parallel under your choice of free software license, such as the GNU General Public License, to permit their use in free software.

Bibliography

[1] Jennifer L. Aaker. Dimensions of brand personality. *Journal of Marketing Research*, pages 347–356, 1997.

[2] Joyce Akse, William W. Hale, Rutger C.M.E. Engels, Quinten A.W. Raaijmakers, and Wim H.J. Meeus. Personality, perceived parental rejection and problem behavior in adolescence. *Social Psychiatry and Psychiatric Epidemiology*, 39(12):980–988, 2004.

[3] Jüri Allik and Robert R. McCrae. Toward a geography of personality traits: Patterns of profiles across 36 cultures. *Journal of Cross-Cultural Psychology*, 35(1):13–28, 2004.

[4] Yair Amichai-Hamburger and Gideon Vinitzky. Social network use and personality. *Computers in Human Behavior*, 26(6):1289–1295, 2010.

[5] Yair Amichai-Hamburger, Galit Wainapel, and Shaul Fox. On the internet no one knows i'm an introvert: extroversion, neuroticism, and internet interaction. *Cyberpsychology & Behavior*, 5(2):125–128, 2002.

[6] Tel Amiel and Stephanie Lee Sargent. Individual differences in internet usage motives. *Computers in Human Behavior*, 20(6):711–726, 2004.

[7] Marjolijn L. Antheunis, Patti M. Valkenburg, and Jochen Peter. The quality of online , offline , and mixed mode friendships among users of a social networking site. *Cyberpsychology: Journal of Psychosocial Research on Cyberspace*, 6(3):1–13, 2012.

[8] Jens B. Asendorpf and Susanne Wilpers. Personality effects on social relationships. *Journal of personality and Social Psychology*, 74(6):1531, 1998.

[9] Michael C. Ashton and Kibeom Lee. The prediction of honesty-humility-related criteria by the {HEXACO} and five-factor models of personality. *Journal of Research in Personality*, 42(5):1216 – 1228, 2008.

[10] Michael C. Ashton and Kibeom Lee. The hexaco–60: A short measure of the major dimensions of personality. *Journal of Personality Assessment*, 91(4):340–345, 2009.

[11] Lars Backstrom, Eric Sun, and Cameron Marlow. Find me if you can: improving geographical prediction with social and spatial proximity. In *Proceedings of the 19th International Conference on the World Wide Web*, pages 61–70. ACM, 2010.

[12] Jose Maria Balmaceda, Silvia Schiaffino, and Daniela Godoy. How do personality traits affect communication among users in online social networks? *Online Information Review*, 38:136–153, 2013.

[13] John A. Bargh, Katelyn Y.A. McKenna, and Grainne M. Fitzsimons. Can you see the real me? activation and expression of the "true self" on the internet. *Journal of Social Issues*, 58(1):33–48, 2002.

[14] Marc Barthélemy. Spatial networks. *Physics Reports*, 499(1-3):1–101, 2011.

[15] Hassan Bashiri, Usha Barahmand, Z. Saeed Akabri, G. Hossein Ghamari, and Asghar Vusugi. A study of the psychometric properties and the standardization of hexaco personality inventory. *Procedia-Social and Behavioral Sciences*, 30:1173–1176, 2011.

[16] Mathieu Bastian, Sebastien Heymann, and Mathieu Jacomy. Gephi: An open source software for exploring and manipulating networks. 2009.

[17] Roy F. Baumeister and Mark R. Leary. The need to belong: desire for interpersonal attachments as a fundamental human motivation. *Psychological Bulletin*, 117(3):497, 1995.

[18] Verónica Benet-Martínez and Oliver P. John. Los cinco grandes across cultures and ethnic groups: Multitrait-multimethod analyses of the big five in spanish and english. *Journal of Personality and Social Psychology*, 75(3):729, 1998.

[19] Jason D. Boardman, Benjamin W. Domingue, and Jason M. Fletcher. How social and genetic factors predict friendship networks. *Proceedings of the National Academy of Sciences*, 109(43):17377–17381, 2012.

[20] Johan Bollen, Bruno Goncalves, Guangchen Ruan, and Huina Mao. Happiness is assortative in online social networks. *Artificial Life*, 17:237–251, 2011.

[21] Thomas J. Bouchard et al. Genes, environment, and personality. *Science*, pages 1700–1700, 1994.

[22] Inge Bretherton and Marjorie Beeghly. Talking about internal states: The acquisition of an explicit theory of mind. *Developmental Psychology*, 18(6):906, 1982.

[23] Sarah Butt and James G. Phillips. Personality and self reported mobile phone use. *Computers in Human Behavior*, 24(2):346–360, 2008.

[24] Eunjoon Cho, Seth A. Myers, and Jure Leskovec. Friendship and mobility: user movement in location-based social networks. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1082–1090. ACM, 2011.

[25] Martin J. Chorley, Gualtiero B Colombo, Stuart M. Allen, and Roger M. Whitaker. Visiting patterns and personality of foursquare users. In *Cloud and Green Computing (CGC), 2013 Third International Conference on*, pages 271–276. IEEE, 2013.

[26] Martin J. Chorley, Roger M. Whitaker, and Stuart M. Allen. Personality and location-based social networks. *Computers in Human Behavior*, 46:45–56, 2015.

[27] Andy Clark and David Chalmers. The extended mind. *Analysis*, 58(1):7–19, 1998.

[28] Allan Clifton, Eric Turkheimer, and Thomas F. Oltmanns. Improving assessment of personality disorder traits through social network analysis. *Journal of Personality*, 75(October):1007–1032, 2007.

[29] Allan Clifton, Eric Turkheimer, and Thomas F. Oltmanns. Personality disorder in social networks: Network position as a marker of interpersonal dysfunction. *Social Networks*, 31(1):26–32, 2009.

[30] Allan Clifton and Gregory D. Webster. An introduction to social network analysis for personality and social psychologists. *Social Psychological and Personality Science*, 8(4):442–453, 2017.

[31] Teresa Correa, Amber Willard Hinsley, and Homero Gil de Zúñiga. Who interacts on the web?: The intersection of users' personality and social media use. *Computers in Human Behavior*, 26(2):247–253, 2009.

[32] Paul T. Costa and Robert R. McCrae. Influence of extraversion and neuroticism on subjective well-being: happy and unhappy people. *Journal of Personality and Social Psychology*, 38(4):668, 1980.

[33] Paul T. Costa and Robert R. McCrae. *The NEO personality inventory: Manual, form S and form R*. Psychological Assessment Resources, 1985.

[34] Henriette Cramer, Mattias Rost, and Lars Erik Holmquist. Performing a check-in: emerging practices, norms and'conflicts' in location-sharing using foursquare. In *Proceedings of the 13th International Conference on Human Computer Interaction with Mobile Devices and Services*, pages 57–66. ACM, 2011.

[35] Justin Cranshaw, Raz Schwartz, Jason I. Hong, and Norman M. Sadeh. The livehoods project: Utilizing social media to understand the dynamics of a city. In *Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media*, pages 58–65. AAAI, 2012.

[36] Darren P. Croft, Jens Krause, Safi K. Darden, Indar W. Ramnarine, Jolyon J. Faria, and Richard James. Behavioural trait assortment in a social network: patterns and implications. *Behavioral Ecology and Sociobiology*, 63(10):1495–1503, 2009.

[37] D.P. Croft, R. James, A.J.W. Ward, M.S. Botham, D. Mawdsley, and J. Krause. Assortative interactions and social networks in fish. *Oecologia*, 143(2):211–219, 2005.

[38] Christine Dancey and John Reidy. *Statistics without maths for psychology*. Pearson, 6th edition, 2014.

[39] Lisa M. DeBruine. Facial resemblance enhances trust. *Proceedings of the Royal Society of London B: Biological Sciences*, 269(1498):1307–1312, 2002.

[40] Pierre Deville, Chaoming Song, Nathan Eagle, Vincent D. Blondel, Albert-László Barabási, and Dashun Wang. Scaling identity connects human mobility and social interactions. *Proceedings of the National Academy of Sciences*, 113(26):7047–7052, 2016.

[41] John M. Digman. Personality structure: Emergence of the five-factor model. *Annual Review of Psychology*, 41:417–440, 1990.

[42] Yuxiao Dong, Yang Yang, Jie Tang, Yang Yang, and Nitesh V. Chawla. Inferring user demographics and social strategies in mobile social networks. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 15–24. ACM, 2014.

[43] Robin I.M. Dunbar. Neocortex size as a constraint on group size in primates. *Journal of Human Evolution*, 22(6):469–493, 1992.

[44] Robin I.M. Dunbar. The social brain hypothesis. *Evolutionary Anthropology*, 9(10):178–190, 1998.

[45] Robin I.M. Dunbar. Do online social media cut through the constraints that limit the size of offline social networks ? *Royal Society Open Science*, 3(150292), 2016.

[46] Robin I.M. Dunbar, Valerio Arnaboldi, Marco Conti, and Andrea Passarella. The structure of online social networks mirrors those in the offline world. *Social Networks*, 43:39–47, 2015.

[47] David J. Ehrler, J. Gary Evans, and Ron L. McGhee. Extending big-five theory into childhood: A preliminary investigation into the relationship between big-five personality traits and behavior problems in children. *Psychology in the Schools*, 36(6):451–458, 1999.

[48] Martin G. Everett and Thomas W. Valente. Bridging, brokerage and betweenness. *Social Networks*, 44:202–208, 2016.

[49] Golnoosh Farnadi, Susana Zoghbi, Marie-Francine Moens, and Martine De Cock. Recognising personality traits using Facebook status updates. In *Proceedings of the Workshop on Computational Personality Recognition (WCPR13) at the 7th international AAAI Conference on Weblogs and Social Media (ICWSM13)*. AAAI, 2013.

[50] Daniel C. Feiler and Adam M. Kleinbaum. Popularity, similarity, and the network extraversion bias. *Psychological Science*, 26(5):593–603, 2015.

[51] Leon Festinger. A theory of social comparison processes. *Human Relations*, 7(2):117–140, 1954.

[52] Adrien Friggeri, Renaud Lambiotte, Michal Kosinski, and Eric Fleury. Psychological aspects of social communities. In *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom)*, pages 195–202. IEEE, 2012.

[53] Feng Fu, Martin A. Nowak, Nicholas A. Christakis, and James H. Fowler. The evolution of homophily. *Scientific Reports*, 2:845, 2012.

[54] Adrian Furnham and Chris R. Brewin. Personality and happiness. *Personality and Individual Differences*, 11(10):1093–1096, 1990.

[55] Francis Galton and S. Okamoto. *Measurement of character*. Chapman and Hall, 1884.

[56] Maggie Geuens, Bert Weijters, and Kristof De Wulf. A new measure of brand personality. *International Journal of Research in Marketing*, 26(2):97–107, 2009.

[57] Homero Gil de Zuniga, Trevor Diehl, Brigitte Huber, and James Liu. Personality Traits and Social Media Use in 20 Countries: How Personality Relates to Frequency of Social Media Use, Social Media News Use, and Social Media Use for Social Interaction. *Cyberpsychology, Behavior and Social Networking*, 20(9):540–552, 2017.

[58] Lewis R. Goldberg. An alternative description of personality : The big-five factor structure. *Journal of Personality and Social Psychology*, 59(6):1216–1229, 1990.

[59] Lewis R. Goldberg, John A. Johnson, Herbert W. Eber, Robert Hogan, Michael C. Ashton, C. Robert Cloninger, and Harrison G. Gough. The international personality item pool and the future of public-domain personality measures. *Journal of Research in Personality*, 40(1):84–96, 2006.

[60] Jacob Goldenberg and Moshe Levy. Distance is not dead: Social interaction and geographical distance in the internet era. *arXiv preprint arXiv:0906.3202*, 2009.

[61] Samuel D. Gosling, Peter J. Rentfrow, and William B. Swann. A very brief measure of the big-five personality domains. *Journal of Research in Personality*, 37(6):504–528, 2003.

[62] Lindsay T. Graham and Samuel D. Gosling. Can the ambiance of a place be determined by the user profiles of the people who visit it? In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, pages 145–152. AAAI, 2011.

[63] Mark S Granovetter. The strength of weak ties. *Social Networks*, pages 347–367, 1977.

[64] The Guardian. The cambridge analytica files. `https://www.theguardian.com/news/series/cambridge-analytica-files`, Accessed: 10-06-2018.

[65] Michael Gurven, Christopher Von Rueden, Maxim Massenkoff, Hillard Kaplan, and Marino Lero Vie. How universal is the big five? testing the five-factor model of personality variation among forager–farmers in the bolivian amazon. *Journal of Personality and Social Psychology*, 104(2):354, 2013.

[66] Aric Hagberg, Pieter Swart, and Daniel S. Chult. Exploring network structure, dynamics, and function using networkx. Technical report, Los Alamos National Lab.(LANL), Los Alamos, NM (United States), 2008.

[67] Yair Amichai Hamburger and Elisheva Ben-Artzi. Relationship between extraversion and neuroticism and the different uses of the internet. *Computers in Human Behavior*, 16(4):441–449, 2000.

[68] Russell A. Hill and Robin I.M. Dunbar. Social network size in humans. *Human Nature*, 14(1):53–72, 2003.

[69] Peter Hills and Michael Argyle. Uses of the internet and their relationships with individual differences in personality. *Computers in Human Behavior*, 19(1):59–70, 2003.

[70] Rick H. Hoyle. Personality processes and problem behavior. *Journal of Personality*, 68(6):953–966, 2000.

[71] Desislava Hristova, Mirco Musolesi, and Cecilia Mascolo. Keep your friends close and your Facebook friends closer : A multiplex network approach to the analysis of offline and online social ties. *International Conference on Web and Social Media*, pages 206–215, 2014.

[72] Han Chung Huang, T. C.E. Cheng, Wei Fan Huang, and Ching I. Teng. Who are likely to build strong online social networks? The perspectives of relational cohesion theory and personality theory. *Computers in Human Behavior*, 82:111–123, 2018.

[73] David John Hughes, Moss Rowe, Mark Batey, and Andrew Lee. A tale of two sites: Twitter vs. Facebook and the personality predictors of social media usage. *Computers in Human Behavior*, 28(2):561–569, 2012.

[74] SPSS Inc. Ibm spss statistics for windows, version 23.0. Armonk, NY: IBM Corp.

[75] Matthew O. Jackson and Brian W. Rogers. Meeting strangers and friends of friends: How random are social networks? *American Economic Review*, 97(3):890–915, 2007.

[76] Tanja Jadin, Timo Gnambs, and Bernad Batinic. Personality traits and knowledge sharing in online communities. *Computers in Human Behavior*, 29(1):210–216, 2013.

[77] Oliver P. John, Eileen M. Donahue, and Robert L. Kentle. The big five inventory—versions 4a and 54, 1991.

[78] Markus Jokela. Personality predicts migration within and between us states. *Journal of Research in Personality*, 43(1):79–83, 2009.

[79] Markus Jokela, Wiebke Bleidorn, Michael E. Lamb, Samuel D. Gosling, and Peter J. Rentfrow. Geographically varying associations between personality and life satisfaction in the london metropolitan area. *Proceedings of the National Academy of Sciences*, pages 1–6, 2015.

[80] Kenneth Joseph, Chun How Tan, and Kathleen M. Carley. Beyond local, categories and friends: clustering foursquare users with latent topics. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, pages 919–926. ACM, 2012.

[81] Timothy A. Judge, Daniel Heller, and Michael K. Mount. Five-factor model of personality and job satisfaction: A meta-analysis. *Journal of Applied Psychology*, 87(3):530–541, 2002.

[82] Yuval Kalish. Bridging in social networks: Who are the people in structural holes and why are they there? *Asian Journal of Social Psychology*, 11(1):53–66, 2008.

[83] Yuval Kalish and Garry Robins. Psychological predispositions and network structure: The relationship between individual predispositions, structural holes and network closure. *Social Networks*, 28(1):56–84, 2006.

[84] Andreas Kaltenbrunner, Salvatore Scellato, Yana Volkovich, David Laniado, Dave Currie, Erik J. Jutemar, and Cecilia Mascolo. Far from the eyes, close on the web: Impact of geographic distance on online social interactions. *Workshop on Online Social Networks (WOSN)*, pages 19–24, 2012.

[85] Katherine J. Klein, Beng-chong Lim, Jessica L. Saltz, and David M. Mayer. How do they get there ? an examination of the antecedents of centrality in team networks. *Academy of Management Journal*, 47(6):952–963, 2004.

[86] Michal Kosinski, Yoram Bachrach, Pushmeet Kohli, David Stillwell, and Thore Graepel. Manifestations of user personality in website choice and behaviour on online social networks. *Machine Learning*, 95(3):357–380, 2014.

[87] Michal Kosinski, Sandra C. Matz, Samuel Gosling, Vesselin Popov, and David Stillwell. Facebook as a social science research tool: Opportunities, challenges, ethical considerations and practical guidelines. *American Psychologist*, 2015.

[88] Michal Kosinski, David Stillwell, and Thore Graepel. Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences*, 110(15):5802–5805, 2013.

[89] Gueorgi Kossinets and Duncan J. Watts. Empirical analysis of an evolving social network. *Science*, 311(5757):88–90, 2006.

[90] Gueorgi Kossinets and Duncan J. Watts. Origins of homophily in an evolving social network. *American Journal of Sociology*, 115(2):405–450, 2009.

[91] Renaud Lambiotte, Vincent D. Blondel, Cristobald De Kerchove, Etienne Huens, Christophe Prieur, Zbigniew Smoreda, and Paul Van Dooren. Geographical dispersal of mobile communication networks. *Physica A: Statistical Mechanics and its Applications*, 387(21):5317–5325, 2008.

[92] Richard N. Landers and John W. Lounsbury. An investigation of big five and narrow personality traits in relation to internet usage. *Computers in Human Behavior*, 22(2):283–293, 2006.

[93] Jonas Larsen, Kay W. Axhausen, and John Urry. Geographies of social networks: meetings, travel and communications. *Mobilities*, 1(2):261–283, 2006.

[94] Jacques Launay and Robin I.M. Dunbar. Does implied community size predict likeability of a similar stranger? *Evolution and Human Behavior*, 36(1):32–37, 2015.

[95] Kibeom Lee, Babatunde Ogunfowora, and Michael C. Ashton. Personality traits beyond the big five: Are they within the hexaco space? *Journal of Personality*, 73(5):1437–1463, 2005.

[96] Bruno Lepri, Jacopo Staiano, Erez Shmueli, Fabio Pianesi, and Alex Pentland. The role of personality in shaping social networks and mediating behavioral change. *User Modeling and User-Adapted Interaction*, pages 1–33, 2016.

[97] David Liben-Nowell, Jasmine Novak, Ravi Kumar, Prabhakar Raghavan, and Andrew Tomkins. Geographic routing in social networks. *Proceedings of the National Academy of Sciences of the United States of America*, 102(33):11623–11628, 2005.

[98] Janne Lindqvist, Justin Cranshaw, Jason Wiese, Jason Hong, and John Zimmerman. I'm the mayor of my house: examining why people use foursquare-a social-driven location sharing application. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2409–2418. ACM, 2011.

[99] Jan-Erik Lnnqvist, Juha V.A. Itkonen, Markku Verkasalo, and Panu Poutvaara. The five-factor model of personality and degree and transitivity of Facebook social networks. *Journal of Research in Personality*, 50:98–101, 2014.

[100] Didier Louis and Cindy Lombart. Impact of brand personality on three major relational consequences (trust, attachment, and commitment to the brand). *Journal of Product & Brand Management*, 19(2):114–130, 2010.

[101] Mary E Lutz. Personality homophily among third through fifth grade students: relations with social identity and conformity. *Master Thesis*, 2009.

[102] Adriana M. Manago, Tamara Taylor, and Patricia M. Greenfield. Me and my 400 friends: The anatomy of college students' Facebook networks, their communication patterns, and well-being. *Developmental Psychology*, 48(2):369–380, 2012.

[103] Dejan Markovikj, Sonja Gievska, Michal Kosinski, and David Stillwell. Mining Facebook data for predictive personality modeling. In *Proceedings of the 7th international AAAI conference on Weblogs and Social Media (ICWSM 2013), Boston, MA, USA*, pages 23–26, 2013.

[104] Jorg J.M. Massen and Sonja E. Koski. Chimps of a feather sit together: chimpanzee friendships are based on homophily in personality. *Evolution and Human Behavior*, 35(1):1–8, 2014.

[105] Sandra C. Matz, Joe J. Gladstone, and David Stillwell. In a world of big data, small effects can still matter: a reply to boyce, daly, hounkpatin, and wood (2017). *Psychological Science*, 28(4):547–550, 2017.

[106] Robert R. McCrae and Paul T. Costa. Validation of the five-factor model of personality across instruments and observers. *Journal of Personality and Social Psychology*, 52(1):81–90, 1987.

[107] Robert R. McCrae and Paul T. Costa Jr. A five-factor theory of personality. *Handbook of Personality: Theory and research*, 2:139–153, 1999.

[108] Robert R. McCrae, Paul T. Costa Jr, and Thomas A. Martin. The NEO-PI-3 : A more readable revised NEO personality inventory. *Journal of Personality Assessment*, 84(3):37–41, 2005.

[109] Miller McPherson, Lynn Smith-Lovin, and James M. Cook. Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27:415–444, 2001.

[110] Ajay Mehra, Martin Kilduff, and Daniel J. Brass. The social networks of high and low self-monitors: Implications for workplace performance. *Administrative Science Quarterly*, 46(1):121–146, 2001.

[111] Brian P. Meier and Michael D. Robinson. Does quick to blame mean quick to anger? the role of agreeableness in dissociating blame and anger. *Personality and Social Psychology Bulletin*, 30(7):856–867, 2004.

[112] Minas Michikyan, Kaveri Subrahmanyam, and Jessica Dennis. Can you tell who i am? neuroticism, extraversion, and online self-presentation among young adults. *Computers in Human Behavior*, 33:179–183, 2014.

[113] S. Milgram. The small world problem. *Psychology Today*, 1(1):61–67, 1967.

[114] M. E. Mitchell, J. R. Lebow, R. Uribe, H. Grathouse, and W. Shoger. Internet use, happiness, social support and introversion: A more fine grained analysis of person variables and internet activity. *Computers in Human Behavior*, 27(5):1857–1861, 2011.

[115] Catherine Molho, Sam G.B. Roberts, Reinout E. de Vries, and Thomas V. Pollet. The six dimensions of personality (hexaco) and their associations with network layer size and emotional closeness to network members. *Personality and Individual Differences*, 99:144–148, 2016.

[116] Kelly Moore and James C. McElroy. The influence of personality on Facebook usage, wall postings, and regret. *Computers in Human Behavior*, 28(1):267–274, 2012.

[117] M. E. J. Newman. Network reconstruction and error estimation with noisy network data. *ArXiv e-prints*, March 2018.

[118] Anastasios Noulas, Salvatore Scellato, Cecilia Mascolo, and Massimiliano Pontil. An empirical study of geographic user activity patterns in foursquare. *ICwSM*, 11:70–573, 2011.

[119] Dan Noyes. The top 20 valuable Facebook statistics – updated march 2018, 2018.

[120] Zizi Papacharissi and Alan M. Rubin. Predictors of internet use. *Journal of Broadcasting & Electronic Media*, 44(2):175–196, 2000.

[121] Konstantinos Pelechrinis and Prashant Krishnamurthy. Socio-spatial affiliation networks. *Computer Communications*, 2015.

[122] Thomas V. Pollet, Sam G. B. Roberts, and Robin I. M. Dunbar. Extraverts have larger social network layers. *Journal of Individual Differences*, 32(3):161–169, 2011.

[123] Daniele Quercia, Michal Kosinski, David Stillwell, and Jon Crowcroft. Our Twitter profiles, our selves: Predicting personality with Twitter. In *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on*, pages 180–185. IEEE, 2011.

[124] Joaquim Radua and Anton Albajes-Eizagirre. Fdr online calculator. `https://www.sdmproject.com/utilities/?show=FDR`, Accessed: 29-10-2018.

[125] Rebecca E. Ready, Lee Anna Clark, David Watson, and Kelley Westerhouse. Self-and peer-reported personality: Agreement, trait ratability, and the "self-based heuristic". *Journal of Research in Personality*, 34(2):208–224, 2000.

[126] Peter J. Rentfrow, Markus Jokela, and Michael E. Lamb. Regional personality differences in great britain. *PloS One*, 10(3):1–20, 2015.

[127] Mark T. Rivera, Sara B. Soderstrom, and Brian Uzzi. Dynamics of dyads in social networks: Assortative, relational, and proximity mechanisms. *Annual Review of Sociology*, 36:91–115, 2010.

[128] Brent W. Roberts, Kate E. Walton, and Wolfgang Viechtbauer. Patterns of mean-level change in personality traits across the life course: a meta-analysis of longitudinal studies. *Psychological Bulletin*, 132(1):1, 2006.

[129] Sam G.B. Roberts and Robin I.M. Dunbar. The costs of family and friends: an 18-month longitudinal study of relationship maintenance and decay. *Evolution and Human Behavior*, 32(3):186–197, 2011.

[130] Craig Ross, Emily S. Orr, Mia Sisic, Jaime M. Arseneault, Mary G. Simmering, and R. Robert Orr. Personality and motivations associated with Facebook use. *Computers in Human Behavior*, 25(2):578–586, 2009.

[131] Graeme D. Ruxton. The unequal variance t-test is an underused alternative to student's t-test and the mann–whitney u test. *Behavioral Ecology*, 17(4):688–690, 2006.

[132] Jari Saramäki, Mikko Kivelä, Jukka-Pekka Onnela, Kimmo Kaski, and Janos Kertesz. Generalizations of the clustering coefficient to weighted complex networks. *Physical Review E*, 75(2):027105, 2007.

[133] Salvatore Scellato and Cecilia Mascolo. Distance matters: Geo-social metrics for online social networks. *Proceedings of the 3rd conference on Online social networks*, pages 1–8, 2010.

[134] Mark Schaller and Damian R. Murray. Pathogens, personality, and culture: disease prevalence predicts worldwide variability in sociosexuality, extraversion, and openness to experience. *Journal of Personality and Social Psychology*, 95(1):212, 2008.

[135] Johann Schrammel, Christina Koffel, and Manfred Tscheligi. Personality traits , usage patterns and information disclosure in online communities. In *Proceedings of the 23rd British HCI Group Annual Conference on People and Computers: Celebrating People and Technology*, pages 169–174. HCI, 2009.

[136] Raz Schwartz and Germaine R. Halegoua. The spatial self: Location-based identity performance on social media. *New Media & Society*, pages 1–18, 2014.

[137] Mary Selden and Adam S. Goodie. Review of the effects of Five Factor Model personality traits on network structures and perceptions of structure. *Social Networks*, 2017.

[138] Maarten Selfhout, William Burk, Susan Branje, Jaap Denissen, Marcel van Aken, and Wim Meeus. Emerging late adolescent friendship networks and big five personality traits: A social network approach. *Journal of Personality*, 78(2):509–538, 2010.

[139] Jianqiang Shen, Oliver Brdiczka, and Juan Liu. A study of Facebook behavior : What does it tell about your neuroticism and extraversion ? *Computers in Human Behavior*, 45:32–38, 2015.

[140] Ryne A. Sherman, Christopher S. Nave, and David C. Funder. Properties of persons and situations related to overall and distinctive personality-behavior congruence. *Journal of Research in Personality*, 46(1):87–101, 2012.

[141] Mark Snyder. The influence of individuals on situations: Implications for understanding the links between personality and social behavior. *Journal of Personality*, 51(3):497–516, 1983.

[142] Sanjay Srivastava, Oliver P. John, Samuel D. Gosling, and Jeff Potter. Development of personality in early and middle adulthood: set like plaster or persistent change? *Journal of Personality and Social Psychology*, 84(5):1041, 2003.

[143] Jacopo Staiano, Bruno Lepri, Nadav Aharony, Fabio Pianesi, Nicu Sebe, and Alex Pentland. Friends don't lie: inferring personality traits from social network structure. In *Proceedings of the 2012 ACM conference on ubiquitous computing*, pages 321–330. ACM, 2012.

[144] Yongjun Sung and Jooyoung Kim. Effects of brand personality on brand trust and brand affect. *Psychology & Marketing*, 27(7):639–661, 2010.

[145] Rhonda J. Swickert, James B. Hittner, Jamie L. Harris, and Jennifer A. Herring. Relationships among internet use, personality, and social support. *Computers in Human Behavior*, 18(4):437–451, 2002.

[146] Shelley E. Taylor and Marci Lobel. Social comparison activity under threat: Downward evaluation and upward contacts. *Psychological Review*, 96(4):569, 1989.

[147] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2015.

[148] Michael Tomasello, Malinda Carpenter, Josep Call, Tanya Behne, and Henrike Moll. In search of the uniquely human. *Behavioral and Brain Sciences*, 28(5):721–727, 2005.

[149] Jeffrey Travers and Stanley Milgram. An experimental study of the small world problem. In *Social Networks*, pages 179–197. Elsevier, 1977.

[150] Sonja Utz. Show me your friends and i will tell you what type of person you are: How one's profile, number of friends, and type of friends influence impression

formation on social network sites. *Journal of Computer-Mediated Communication*, 15(2):314–335, 2010.

[151] José Van Dijck. *The culture of connectivity: A critical history of social media*. Oxford University Press, 2013.

[152] G. van Rossum. Python tutorial. Technical Report CS-R9526, Centrum voor Wiskunde en Informatica (CWI), Amsterdam, May 1995.

[153] Jessica Vitak. Facebook makes the heart grow fonder: relationship maintenance strategies among geographically dispersed and communication-restricted connections. In *Proceedings of the 17th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, pages 842–853. ACM, 2014.

[154] Dashun Wang, Dino Pedreschi, Chaoming Song, Fosca Giannotti, and Albert-Laszlo Barabasi. Human mobility, social ties, and link prediction. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1100–1108. ACM, 2011.

[155] Stefan Wehrli. Personality on social network sites: An application of the five factor model. *ETH Zurich Sociology*, 2009.

[156] Bert Weijters, Elke Cabooter, and Niels Schillewaert. The effect of rating scale format on response styles: The number of response categories and response category labels. *International Journal of Research in Marketing*, 27(3):236–247, 2010.

[157] Barry Wellman, Anabel Quan Haase, James Witte, and Keith Hampton. Does the internet increase, decrease, or supplement social capital? social networks, participation, and community commitment. *American Behavioral Scientist*, 45(3):436–455, 2001.

[158] Barry Wellman, Anabel Quan-Haase, Jeffrey Boase, Wenhong Chen, Keith Hampton, Isabel Díaz, and Kakuko Miyata. The social affordances of the internet for networked individualism. *Journal of Computer-Mediated Communication*, 8(3):JCMC834, 2003.

[159] Barry Wellman and Scot Wortley. Different strokes from different folks: Community ties and social support. *American Journal of Sociology*, 96(3):558–588, 1990.

[160] Roger M. Whitaker, Gualtiero B. Colombo, Stuart M. Allen, and Robin I.M. Dunbar. A dominant social comparison heuristic unites alternative mechanisms for the evolution of indirect reciprocity. *Scientific Reports*, 6:31459, 2016.

[161] Wu Youyou, David Stillwell, H. Andrew Schwartz, and Michal Kosinski. Birds of a feather do flock together: behavior-based personality-assessment method reveals personality similarity among couples and friends. *Psychological Science*, 28(3):276–284, 2017.

[162] Joachim Zentes, Dirk Morschett, and Hanna Schramm-Klein. Brand personality of retailers–an analysis of its applicability and its effect on store loyalty. *The International Review of Retail, Distribution and Consumer Research*, 18(2):167–184, 2008.

[163] Ke Zhang and Konstantinos Pelechrinis. Understanding spatial homophily: the case of peer influence and social selection. In *Proceedings of the 23rd international conference on World wide web*, pages 271–282. ACM, 2014.

[164] W.X. Zhou, Didier Sornette, Russell A. Hill, and Robin I.M. Dunbar. Discrete hierarchical organization of social group sizes. *Proceedings of the Royal Society of London B: Biological Sciences*, 272(1561):439–444, 2005.