University of Sussex

An Evaluation of Identity in Online Social Networking: Distinguishing Fact from Fiction

---

Roya Feizy
Software Systems Group
School of Informatics
University of Sussex

Submitted to the School of Informatics for the degree of Doctor of Philosophy at University of Sussex, Nov 2010

Thesis Committee:

Supervisors . . . . . .

Dr Ian Wakeman
School of Informatics
University of Sussex

Dr Dan Chalmers
School of Informatics
University of Sussex

Thesis Reader . . . . . .

Dr Natalia Beloff
University of Sussex

Thesis Reader . . . . . .

Dr Eamonn O'Neill
University of Bath

Declaration

I hereby declare that this thesis has not been and will not be, submitted in whole or in part to another University for the award of any other degree.

Parts of this thesis have previously appeared in the following publications:

- 'Are Your Friends Who They Say They Are?', ACM Crossroads 16, 2, 19-23, Dec 2009.
- 'Transformation of Online Representation through Time', International Conference on Advances in Social Network Analysis and Mining, Athens, Jul 2009.
- 'Distinguishing Fact and Fiction: Data Mining Online Identities', In Proceedings of 5th International Workshop on Security and Trust Management (STM), France, Sep 2009.

Signature . . . . . .

Roya Feizy

Abstract

Online social networks are understood to replicate the real life connections between people. As the technology matures, more people are joining social networking communities such as MySpace (www.myspace.com) and Facebook (www.facebook.com). These online communities provide the opportunity for individuals to present themselves and maintain social interactions through their profiles. Such traces in profiles can be used as evidence in deciding the level of trust with which to imbue individuals in making access control decisions. However, online profiles have serious implications over the reality of identity disclosure.

There are many reasons why someone may choose not to reveal their true self, which sometimes leads to misidentification or deception. On one hand, the structure of online profiles allows anonymity, which gives users the opportunity to create a persona that may not represent their true identity. On the other hand, we often play multiple identities in different contexts where such behaviour is acceptable. However, realizing the context for each identity representation depends on the individual. As a result, some represented identities will be essentially real, if edited for public view, some will be disguised, and others will be fictitious or humorous.

The millions of social network profiles, and billions of connections between them, make it difficult to formalize an automated approach to differentiate fact from fiction in online self-described identities. How can we be sure with whom we are interacting, and whether these individuals or groups are being truthful with the online identities they present to the rest of the community? What tools and techniques can be used to gather, organize, and explore the available data for informing the level of honesty that should be entrusted to an individual? Can we verify the validity of the identity automatically, based on the available information online?

We aim to evaluate identity representation online and examine how identity can be verified in a less trusted online community. We propose a personality classifier model to identify a user's personality (such as expressive, valid, active, positive, popular, sociable and traceable) using traces of 2.2 million profile features collected from MySpace. We use data mining techniques and social network analysis to extract significant patterns in the data and network structure, and improve the classifier during the cycle of development. We evaluate our classifier model on profiles with known identities such as '*real*' and '*fake*'. Our results indicate that by utilizing people's online, self-reported information, personality, and their network of friends and interactions, we are able to provide evidence for validating the type of identity in a manner that is both accurate and scalable.

Acknowledgments

The PhD years have been very challenging and intense and it's time to conclude this important period in my life. Here I would like to thank those who have influenced me technically and emotionally to reach this moment.

My primary thanks go to my supervisor, Dr Ian Wakeman, who has given me the opportunity to work in his group. I am heartily thankful to him, whose guidance, encouragement and constant support from the initial to the final stage enabled me to develop an understanding of the research subject.

I would like to express my deepest gratitude to my second supervisor, Dr Dan Chalmers, for his invaluable advice throughout this study, whose expertise and understanding added considerably to my knowledge. I am also grateful to the member of the thesis committee, Dr Des Watson, for reading and evaluating reports and for his insightful comments.

I would like to acknowledge the members of the Foundation of Software System Group (FOSS): Dr Anirban Basu, Simon Fleming, Lachhman Dhomeja, Aeshah Alsiyami, James Stanier, Dr Jian Li, Yasir Malkani, Danny Matthews, Renan Krishna and Stephen Naicken. While it has been a diverse and evolving group, it has remained unified with a friendly atmosphere. I especially express my sincere gratitude to Simon Fleming and James Stanier for their kind assistance with thesis proof reading and their valuable comments. I am also indebted to Dr Jon Robinson for his outstanding advice throughout the entire thesis.

This research would not be possible without the generous sponsorship from the Mehr Cultural Foundation. In particular I would like to thank Dr Ghasemi, who has played a large role for partly financing my study. I would also like to extend my thanks to the School of Informatics at the University of Sussex for providing the facilities, resources and the relaxed environment in which I could develop my academic interests.

Further, I wish to thank my family and friends who patiently supported me in many respects during the completion of this research. In particular, my deepest gratitude goes to my beloved family for their moral support during my study at Sussex. Finally, I dedicate this thesis to my soul mate, Pejman Karami, for his extensive encouragement, patience and unconditional love.

Table of Contents

List of Tables

List of Figures

Appendix

Chapter 1

Introduction

"Expose yourself to your deepest fear; after that, fear has no power, and the fear of freedom shrinks and vanishes. You are free."

Jim Morrison

1.1 Introduction

As technology matures more people are leaving a trace of their identity within online social networks. Everywhere we go, both in the real and digital world, we leave a trail of information about who we are, where we've been, who our friends and colleagues are, as well as our purchasing preferences. This information, together with its associated credentials, is a representation of our identity.

Online social networking provides an opportunity for social interaction through digital profiles; this represents an image of the individuals' identity and has a strong link to their personality **[Donath & boyd, 2004]**. Profiles are the collection of demographic information that reveals an image of users and their connections within the community. The published data on profiles may be shared with a network of existing friends, and frequently with strangers. Such profiles may be useful in confirming identities within other contexts, such as in determining the level of trust to place in an individual within a social networking context.

A system that can be used to track a profile and share personal information with others has serious identity implications. An online social identity may be part of a role-playing game or it may be an impersonation, either for play or more nefarious purposes, such as fraud **[Berman & Bruckman, 2001]**. Each of these real or fake identities is associated with profile data and is embedded within a social network. Identity validation has a long history in computer science and translates directly to the pervasive computing context. Trust is now a very hot

topic of research in pervasive computing and it is well known that access control mechanisms will use some form of computational trust **[Kagal *et al.*, 2001]**. One example of this paradigm is the set of social networks embodied in websites, which have become important research areas in both academic and commercial fields. If a person can show proof that he or she is responsible for an online identity through standard public key cryptography, then his/her personal information and relationship with friends can be used to calculate a level of trust.

Online identity has a varying relationship with real world identity, which has been part of the attraction of online interaction for many years. Traditionally our social identity has been interlinked within a physical space and the concept of identity emerges to most people very naturally. At the same time it is very difficult to formalize an approach to identify the reality of an identity presented online in comparison to the real-life identity. One reason is that virtual identities are more adaptable than a real-life identity, and with the opportunity to become anonymous it would be difficult to distinguish the real from the fake identity. Another reason is because we often play multiple identities in different contexts where we are accepted. However, realizing the context for each identity depends on the individual, who assigns the profile according to the preferred context. In addition, where real identity information is disclosed this may raise privacy and security issues. Online profiles have serious identity and privacy implications. As there is a trade-off between privacy and honesty, privacy sometimes encourages dishonesty and honesty can be undermined in online social networking. As a result, some represented identities will be essentially real, if edited for public view. Some will be disguised, but known to those within a group of friends; others will be fictitious or humorous.

In this thesis we explore modes of identity and how people disclose, hide, obfuscate or fabricate their identities within online social networks. A better understanding of these networks and modes of presentation may allow us to determine their value in supporting credentials. We examine how people present themselves through their online space by evaluating their self-described profile and following their network of friends and connections. The focus of our study is the type and amount of published information and the validity of a profile's identity. We aim to identify the reality of profile information and find an accurate measurement to distinguish between real and fake identities online. How can we be sure with whom we are interacting, and whether these connected profiles are being truthful about the online identities presented to the rest of the

community? Can we validate the identity automatically, based solely on the displayed information? Can we use data mining techniques to analyse and evaluate the available data and distinguish the type of identity, such as real or fake? If we wish to use profiles as input to trust decisions, can we measure their authenticity without human intervention, and how stable are such identities over time?

To tackle these questions, we propose a personality classifier, use a machine learning approach to look at traces of people's identities left behind on online social networking sites, and evaluate the validity of those identities. Our research examines personal information on an online social community, and employs MySpace as a case study due to its widespread and diverse population with rich sources of identity information. Similar to other social networking sites, MySpace offers an easy to generate personal web page, along with many features, such as an email service, internal blog, forum, photo sharing, music streaming, and so on. This network allows users to create customised profiles and establish links with other people as friends to commence communication.

We evaluate the quantity and quality of information disclosed on profiles. Therefore, we constructed a spider to crawl a data sample consisting of 2.2 million profiles from MySpace over the course of a two-month observation period. This profile information with unknown identity type was used as a *'validation'* dataset to cluster information more efficiently to identify personality factors and construct our classifier. In addition, we have applied different methods for the collection of personal information where the identity of the user is known and tag them as *'training'* dataset for data mining purposes (see Section **4.3**). We gathered four types of profile: '*real-celebrity'*, '*real-local*', '*fake-celebrity*', and '*fake-invented*'. '*Real-celebrity*' represents the profiles of famous people, mostly celebrities, whose names are listed in the directory of official profiles on MySpace. '*Real-local*' refers to MySpace users at the University of Sussex who responded to our email survey. '*Fake-celebrity*' or impersonator group are users that fabricated the identities of known people, for example celebrities with almost the same identity, such as name and pictures. '*Fake-invented*' are those who replied to our online survey and generated a fake profile (see Section **3.2.2**).

There are two types of identity information: the profile contents "text" and friends "link", which we will focus on to analyse the type of identity. We propose a classifier to identify the characteristics of each profile and predict the type of identity, such as real or fake. We applied different methods and algorithms to

categorize profile attributes together with their friends' connections. Our model classifies collected profile information into a set of characteristics as seven dimensions of personality: '*expressive/anonymous*', *'valid/fantasy*', '*active/inactive*', *'positive/offensive'*, '*popular/isolated'*, '*sociable/unsociable'*, and '*traceable/untraceable*'. These personality factors are identified based on literature review and also through data mining and grouping of information (see Section **3.3.1**). These personality metrics are determined using text mining techniques, such as utilizing several databases to check the validity of information, and the terms and language used against a list of known terms. For instance, checking the existence of a city and country by comparing to a database of known locations collected from online. Additionally, we examined the structure of our sample network, such as centrality **[Russo & Koesten, 2005]** and similarity **[Spertus *et al.*, 2005]** analysis between profiles and their network of friends in order to find a correlation between profile attributes and the type of identity. The link analysis of friendship among groups of friends gives us an opportunity to understand network characteristics, as, according to **[Katona *et al.*, 2009]**, individual characteristics have a significant influence on the community structure and vice-versa. Social network analysis techniques are used to determine the personality factors, such as '*sociability*' and '*popularity*'.

In this model, profiles characteristics, together with their friends' connections, can be weighted based on their self-described identity. Using the classified data and evaluating the identity of a clique of friends for each user allows us to construct a rating algorithm for each personality trait. To evaluate our classifier, we first identified which identity elements are more significant in distinguishing the type of identity by extracting main principal components from data. Principal Component Analysis (PCA) uncovers unknown trends in the data that explain the correlations among a set of attributes and simplifies the structure of a set of variables **[Qu *et al.*, 2002]**. Therefore we applied the component analysis techniques to reduce the dimensions in our sample data, which helps to find the significant identity elements and the correlation between each entity. We then found a pattern in information by applying several data mining techniques, such as supervised and unsupervised learning **[Klösgen & Zytkow, 2002]**. Data mining techniques, such as supervised and unsupervised methods, helped us to improve the personality model in a development cycle. Based on the training set (the participants with known identity, such as '*real-celebrity*', '*real-local*', '*fake-celebrity*', and *'fake-invented'*) the accuracy of our personality classifier was evaluated; this shows an average of 82.9% accuracy in predicting the type of identity within our dataset.

In addition, as people often alter their profile information, and their interactions with others on online social networks, we attempt to understand if there is any correlation between the amount of alteration on self-reported profiles and the type of identity. It is challenging to extract profile personalities and formulate the differences between a past and present representation. This transformation of self-presentation can be traced and measured to extract hidden patterns in relation to profile characteristics.

The result of our study shows how examining personality factors can determine the type of identity, such as '*real'* or *'fake*'. We will illustrate what type and amount of information people are willing to disclose. What are the patterns in profile representation and their proportion (according to their age, gender and location)? How does profile information expose a user's characteristics, such as a personality factor? What are the correlations between similarity and centrality in a network and the type of identity? How much of the provided data is valid and identifiable? How does identity transform over a period of time?

To the best of our knowledge this is the first attempt to observe and measure the properties of identity and profile characteristics on online social networking in order to determine the type of identity. The novelty of our research is that we used existing knowledge, algorithm and online information to propose a model to detect and validate the type of identity representation in real time. Our proposed algorithm and classifier can be effectively used to identify the type of identity and, in the future, can be used as an input to trust decisions when combined with an effective recommendation system **[Hsu *et al.*, 2006]**.

This chapter aims to define the problem by highlighting the area that requires more attention and describe our research questions, aims and objectives, contribution, adopted research methods, and research ethics. We will describe our motivations and determination to answer the questions raised from this study. The remainder of this chapter is organised as follows: the next section (Section **1.2**) describes the research problem and how to overcome the dilemma. The primary motivations for this research together with our contributions, research questions, research methods and research ethics are explained in Section **1.3**. Finally the thesis structure (Section **1.4**), including a description of each chapter, will be provided.

1.2 Problem Definition

The rapid evolution of social networking has engendered a new paradigm for collaboration, which offers an opportunity to study human social networking and interaction. In particular, the identity issues confer a variety of research areas for researchers to analyse the main attributes of online social networking. Previous studies on online identity have generally relied on addressing privacy and trust issues, such as **[Acquisti & Gross, 2005]** and **[Dwyer *et al.*, 2007]**. Many researchers also focused on the self-representation of identity in online social networks, such as **[boyd *et al.*, 2004]**, **[Stutzman, 2006]**. A considerable amount of research has also focused on analysing the structure of these social networks to measure their size, shape and available attributes, such as **[Petroczi *et al.*, 2006]** and **[Mislove *et al.*, 2007]**. Although there are some reputation systems based on human ratings (such as eBay), currently there is a lack of any mechanism to measure and enforce a trust model on online social networking. There is little research on evaluating identity to distinguish fact from fiction. In fact, current researchers are unable to explain and measure the accuracy of online profile information. There should be a proper identity evaluation system on social networking sites to identify the type of represented identity to increase the level of trust between individuals.

In our view, studying online identity representation has two main parts: the first refers to validating the identity, whilst the second refers to protecting the identity. Within our investigation we focused on how to validate identity information rather than how to protect identity. Many other concepts involved identity issues, such as privacy, anonymity, multiple identities, predators, and identity fraud, which are out of the scope of this research.

Within this section, we address some major problems that the current generation of social networking services, including friend-aggregators (such as MySpace), are facing today. First, we explain the effect of identity representation and why people misrepresent themselves (Section **1.2.1**). Then we describe the lack of an identity verification system on social networking (Section **1.2.2**). We describe the problems arising from the lack of trust management on online social networking sites (Section **1.2.3**). Finally, we discuss privacy concerns and identity disclosure (Section **1.2.4**).

1.2.1 Self-representation of Identity

In real-world identity representation, physical appearance and body language may reveal some information about people and their personality **[Suler, 2002]**. In contrast, in online environments it is more difficult to decide and trust the represented identity. This is because we are mainly dealing with text, images and possibly a trace of interaction between people in online profiles. On one hand, people reveal a variety of information on their profile, and with unknown viewers (such as close friends, parents, employer, and strangers) sometimes they are uncertain of how to perform their identity **[Goffman, 1959]**. Therefore, people might hide their identity, make a fantasy character or impersonate other known people. In addition, the structure of profiles allows anonymity and the use of pseudonyms; this gives people the opportunity to create a persona that may not represent the true self. According to **[Ford & Strauss, 2008]**, anonymity enables people to adopt different online personas and often appears to undermine accountability that motivates people towards misrepresentation. On the other hand, some people may expose their true identity, which may encourage the fear of losing privacy in some contexts.

The first question to ask is why people misrepresent themselves online? A more revealing question is why they do not misrepresent? There are some sociological and psychological answers to these questions, which are outside of the scope of our research. However, according to **[Donath & boyd, 2004]** the key aspect to online identity is that people want others to be able to discover more about them. Sometimes they represent themselves in the way they would like other people to see them. According to **[boyd, 2002]** there are many reasons why people may not reveal their true self, such as fear of isolation, rejection and losing privacy. She also discusses that true identity is sometimes risky to expose (e.g. for employment reasons).

1.2.2 Deficiency of Identity Verification

As people engage in cyber interaction, there is a need for methods to link the real world identity with a virtual identity. One of the primary problems we face today in online social networking is the verification of identity disclosure. Online social networking sites provide the ability to easily and quickly sign up and generate a profile with little authentication process. According to **[Fairhurst, 2003]**, one attempt to ensure the verification of identity is to introduce

biometrics to literally translate physical identifiers into digital terms. Biometric systems authenticate users through physical identifiers, such as a fingerprint, iris scan, facial scan, or signature dynamics. However, to the best of our knowledge, biometric authentications are not implemented on online social networking systems yet.

Although MySpace attempts to protect a user's personal information, the system is currently rather unsophisticated and lacks any verification infrastructure. There is no central management and it is easy to join the site without any accurate authentication, which makes it a target for a range of attacks. For instance, online predators take advantage of the simple access to published personal information on profiles to trace their target (see Section **2.2.4** for more details on MySpace issues). Therefore, a technology is required in the social networking environment that allows identity content to be verified for a trust decision. There should be well-defined mechanisms, frameworks and standards that detect and validate identity, ideally based on the user's roles, current location, and personal preferences. However, checking the accuracy of represented identity and detecting online deception is not an impossible task although it requires substantial knowledge. Within this research, we mainly focus on how to validate a user's identity by examining profile information using machine-learning techniques.

1.2.3 Lack of Trust Management

In face-to-face social interactions, physical factors, such as facial expression and body language, may help us to make a trust decision **[Eckel & Wilson, 2003]**. In online social networking we can only build trust based mainly on a profile's content, such as self-described text, photos and friend's connections. Since profile information can be misrepresented for privacy or personal reasons, this may create new threats and security issues. There is currently a lack of mechanisms and solutions to ensure trust when people collaborate and share personal information on online social networking sites. A trust level between participants is required in order to support collaborative activities, while protecting sensitive information used in the collaboration.

According to **[Dwyer *et al.*, 2007]**, compared to other social networking sites, MySpace is a less trusted community. For instance, in comparison to Facebook MySpace faces more problems with spammers and predators. Facebook friends are usually real-world friends, whilst on MySpace people often connect to anyone

to gain a greater number of friends. Facebook also has more privacy options to protect profile information; this is in contrast to MySpace.

Understanding the characteristics of profiles in online social networking may lead us to proposing an identity model, which validates profile's identity for further trust decisions. The automatic verification of identity combined with a recommendation system will make the social networking environment a safer place for further communication and interaction. Such a system will support users when making friends and collaborating through social networking.

1.2.4 Privacy Implications

Providing the true identity would not be an efficient solution to build a trusted community, as many people are concerned about their loss of privacy online. In particular, online social networking can raise significant problems, such as identity fraud, predators and spammers: privacy protection becomes an increasingly important concern to online users. Although significant numbers of online social network users have indicated awareness about their privacy, some may still underestimate privacy protection and are at risk of over-disclosing their personal information **[Squicciarini *et al.*, 2009]**. Social collaborative identity systems should include a privacy protection solution, ideally based on a mixture of technical, social and legal mechanisms **[Wisse & Jansen, 2006]**.

Within this research, we do not consider the general privacy threat in online social communities. We rather focus on the technological aspects of how to determine the type of identity to build a more trusted community for further collaboration and socializing.

1.3 Aims and Objectives

In the world of online social networking it is challenging to identify profile characteristics and detect those who construct a fake persona. In many cases virtual lies are indistinguishable from truths, and only the person who created a persona would know whether he/she has been honest or not. It is difficult to pinpoint if a profile's information is true or false. Our primary aim is to analyse an online community to detect profile characteristics in order to distinguish between real and fake personas.

Our main objective is to analyse different types of identity to explore how people disclose, hide or fabricate their identities within their social networks. A better understanding of these networks and modes of representation may allow us to determine their value in supporting credentials. This mechanism would detect identity attributes and predict how likely it is for a profile to represent a true identity. Using machine learning techniques and evaluating the profile's contents along with the friend's connections, will give us a prospect of constructing a classifier model to distinguish different types of identities.

The end system should be able to protect the user against deception by verifying identity and making the user aware of the identity of a potential friend. For instance, when a user receives a friend request, the system should check the potential friend's profile and determine the type of profile. This type of system would act as a firewall and support users in identifying trustworthy friends and block non-trusted friends. The system also should assist identity fraud detectors to estimate the validity of an identity holder. So, online social networking would be a more trusted environment for collaboration and sharing personal information.

To the best of our knowledge this study is the first attempt to implement such a system; it examines the content of a profile and detects false or true identities using both machine learning techniques and social network analysis.

1.3.1 Research Questions

The previous section highlighted important problems that affect current identity systems in social networking environments. The main question we aim to answer according to our research problem is to explore whether it is possible to generate a model that automatically detects and validates the type of identity:

> *By examining identity information from online profiles, can we determine the type of a user's identity? What methods can be used to distinguish between real and fake profiles based on a user's self-presentation online?*

In addition, we are aiming to answer the following questions:

What are the main factors in deciding real or fake persona? How do personality factors (such as expressiveness, validity, traceability, activity, positivity, popularity and sociability) determine the type of identity? How efficiently can we build a personality classifier, and model a rating algorithm to value each personality? How do people present themselves in online social communities?

What type and amount of information do they disclose about themselves? How does the MySpace structure impact on the construction of identity and can this be improved by building a trust model?

1.3.2 Contributions

The key contributions to our research are to evaluate profile attributes and learn the representation patterns of each identity trait by employing a set of known techniques, such as social network analysis and data mining to examine different types of identity. The main steps of our research are as follows:

- Accumulate profile information and their friend's connection from the MySpace network;
- Build a classifier to cluster the identity traits into more categorized characteristics in a development cycle using data mining techniques;
- Determine the validity and traceability of disclosed information;
- Examine the use of language to detect offensive and positive attributes;
- Measure the similarity between friends' attributes;
- Determine the centrality attributes (such as in-degree, out-degree) within our sample network;
- Reduce the data dimension using principal component analysis;
- Evaluate the classifier using several data mining techniques for an accurate prediction;
- Measure the evolution of identity over time;
- Improve the classifier based on findings in a development cycle.

We hypothesize that identity contents and network structure of online social communities are significant entities to distinguish real from fake. Based on our findings contributed from several techniques used, we propose and test the following hypothesis:

- The existing methods, such as data mining, social network analysis and principal component analysis can be applied to determine the type of online profiles;
- Proposed personality factors, such as '*expressive/anonymous*', *'valid/fantasy*', '*active/inactive*', *'positive/offensive'*, *'popular/isolated'*, '*sociable/unsociable'*, and '*traceable/untraceable*', can determine the type of identity with 82.9% accuracy;

- The centrality attributes, such as in-degree and out-degree in the network, have a simple relationship with the type of identity, as isolated nodes are highly associated with *'fake'* profiles;
- We discover through similarity analysis that there is a correlation between friends' similarity and the type of identity. For instance *'real'* profiles are more similar to their friends than *'fake'* profiles;
- There is a strong connection between identity disclosure and the number of friends. For instance, the higher number of friends associated with the higher level of *'expressive'* characteristics;
- There is a relationship between privacy and the type of identity. For instance, the less sensitive data, such as 'zodiac' (birth sign) are disclosed more compared with 'age' and 'location';
- There is a correlation between the type of identity representation and the amount of transformation in self-described profiles over time. For instance, *'real'* profiles are more transformed over time than *'fake'* profiles.

1.3.3 Research Methods

In order to meet our research goal and answer our research questions, we have developed a mixture of qualitative and quantitative research methods. The following are a selection of employed research methods:

- **Theoretical:** We studied the background literature by reviewing related papers and discussion on different approaches to examine online social networking and identity representation.
- **Case Study:** We captured user's information from a large number of MySpace profiles and observed profile information on this social network as a case study.
- **Survey:** We collected information from local MySpace users through an email survey to find out their self-rating according to their level of honesty; we also collected data about how they rate their friends' level of honesty. In addition, we distributed an online form and accumulated a number of generated fake identities from participants.

- **Data Analysis:** We examined profiles and network information using text mining and network analysis to identify the key personality specifications and their frequency.
- **Algorithmic:** We applied several predictive algorithms to implement a personality classifier, measure the similarity attributes, and examine the transformation analysis.
- **Evaluation:** We evaluated the accuracy of our classifier by applying different methods, such as data mining, principal component and social network analysis, looking at the efficiency of each method and measuring any false-positive and false-negative predictions. This method is the most significant approach by continually modifying and evaluating the accuracy of our personality classifier in a life cycle of development.

1.3.4 Research Ethics

As the nature of this research contains sensitive identity information, we confirm that we are using the collected data for research purposes only. We assume that the profile information that is on public display can be used for our research purpose, such as data analysis and profile observation. Also by creating an identity code for each profile, we made each identity anonymous within this study (see Section **3.3.4.2**). Therefore, we are not disclosing any personal information during this study or in the future.

Also, as the depth of online identity disclosure has raised many concerns for privacy, with respect to participants' privacy the information obtained from research participants remains confidential. In addition, during this thesis we express acknowledgment and cite any references used within this thesis. We also confirm the honesty in reporting results and the originality of our research.

1.4 Structure of the Thesis

This thesis is organised as follows:

In order to understand the research problems, objectives and contributions, this thesis began with an introduction and an overview of the research problem: this includes identity representation, validation of identity, trust management and privacy concerns (Section **1.2**). We explained the objectives and contributions

that we are focusing on within this study in Section **1.3**. The main research questions are explained in Section **1.3.1**. The contribution and research methods used within this study are described in Sections **1.3.2** and **1.3.3**. The ethical implications of our research are also explained in Section **1.3.4**.

In Chapter **2** we intend to continue the literature study on research papers and discuss the related works. A clear understanding of the research background and theory of identity and online representation is described in Section **2.1**. The research background includes the theoretical approaches to the notion of identity, digital self-representation, social and multiple identities, and an introduction to MySpace social networking. In Section **2.2**, we look at online identity issues (such as privacy, anonymity, and trust implication), which opens up a wide direction for many researchers. The related works are included in Section **2.3**, which describes current and past approaches to overcome online identity issues. Related works, such as identity management systems, evaluation of online communities, social network analysis (such as friendship and similarity analysis), data mining, deception detection, and recommendation systems are explained in this section. Finally, the summary of the literature review is included in Section **2.4**.

Chapter **3** describes the overall research approaches, including data accumulation and modelling the classifier. Section **3.2** describes the data collection methods, such as crawling and survey study along with brief descriptions of data. Section **3.3** explains the procedure of modelling our classifier. Section **3.3.1** describes why we selected seven personality factors, and Section **3.3.2** defines each personality factor. In Sections **3.3.3** and **3.3.4** we describe the text mining and network analysis methods used to extract and rate each personality factor. Section **3.4** also describes the analysis of evolutionary features of identity representation by examining self-described identity profiles of the same person over different periods of time. We conclude this chapter with a discussion and summary of our research approach in Section **3.5**.

Empirical techniques are described in Chapter **4**, which evaluates our proposed classifier in order to determine each type of identity. Section **4.2** explains the procedure for principal component analysis, such as component and rotation analysis (Section **4.2.1**), measuring the correlation between each personality and the influence on predicting the identity type (Section **4.2.2**). Section **4.3** describes the data mining techniques, such as data pre-processing, supervised and unsupervised learning. We take advantage of existing machine learning techniques to identify patterns in our data. By analysing different algorithms, we

are able to explain the prediction accuracy through a confusion matrix in Section **4.3.4**.

Chapter **5** presents the results that correspond to our findings from previous analysis. Section **5.1** demonstrates some statistical results and explains the properties of our dataset. We present the statistical relationship between each entity, including the initial analysis, further personality factors and extracted patterns in our selected data. Exploratory results are illustrated in Section **5.2**, including the results from our social network analysis, principal component analysis and data mining approaches. Section **5.3** also explains the evolutionary results, such as the transformation of a profile's feature and the evolutionary features of the social network. This chapter concludes with a discussion and a summary of our findings in Section **5.4**.

Finally, this thesis concludes with a discussion on our findings and conclusions in Chapter **6**. Section **6.1** includes a discussion of our results, including the efficiency of our classifier model, and how more advanced and sophisticated classifiers could be implemented in the future. We conclude our thesis with a conclusion in Section **6.2**, including an overview of a future system in Section **6.2.1**, and our research limitations in Section **6.2.2**. The opportunities for further research, together with a summary of the thesis, are also explained in Sections **6.2.3** and **6.2.4**.

Chapter 2

Literature Review

"If you establish an identity, you build a monster-and that's right, you've got to live with it. Of course, you can enjoy it, too."

George Shearing

Online identity and social networking are active research areas with a large input from computer science, sociology and psychology, evident from the large collection of recent and past papers. Researchers from a range of diverse fields are currently working on the issues of identity representation on online social networking, for example **[Donath & boyd, 2004]**, **[Mislove *et al.*, 2007]**, **[Stutzman, 2006]**. Studying the relevant research papers gives us a greater understanding of our current research problem and the approaches to be taken.

This literature chapter is organised as follows: Section **2.1** presents a clear understanding of the research background including the definition of identity, digital self-representation, social and multiple identities, and an overview of online social networking focusing on the MySpace community. Section **2.2** looks at online identity issues such as privacy, anonymity and trust implication. MySpace identity issues such as fake identities are also highlighted in Section **2.2.4**. The related works are included in Section **2.3**, which describes the variety of existing approaches used to overcome online identity issues. These works include identity management systems, social network analysis (such as friendship and similarity analysis), data mining, deception detection, and recommendation systems. The summary of the literature review in Section **2.4** concludes this chapter.

2.1 Research Background

It was intended to conduct a research background studying related papers within the following fields. First the identity was defined in Section **2.1.1**,

describing what identity really means in the different areas and how context is related to identity. The digital representation was described in Section **2.1.2** including the description of social and multiple identities (Section **2.1.3**). The importance of online social networking sites is also highlighted and, in particular, the MySpace community in Section **2.1.4**.

2.1.1 The Notion of Identity

A discussion of digital identity within social networking sites cannot be complete without setting a definition of identity. The main question that naturally emerges when investigating identity is certainly about the notion of identity itself. The subject of identity is interesting and emerges very naturally within people. If someone is asked what defines an identity, a number of attributes by which one can be distinguished from others are likely to be listed such as name, date of birth, place of birth, nationality and so on. In practice a combination of techniques, such as physical identification (for example fingerprints, DNA and iris recognition), legal documents (birth certificate and passport), and social identification (such as club membership) can be applied to distinguish one identity from another.

In the physical world the body provides a convincing definition of identity. According to **[Donath *et al.*, 1999]** "*the norm is: one body, one identity*". Though the self may be complex and changeable over time and circumstance, the body provides more stability to identity representation. **[Pato, 2003]** defines the identity of an individual as "*the set of information known about a person*", which is issued by a relevant authority to determine someone's identity. In more common terms, it could be said that identity is the means that distinguishes one from another **[Baier *et al.*, 2003]**. In mathematics, identity is a relation where each element is similar to itself. For instance, 'x' is identical to exactly 'x' and equal in its value. From this conception **[Gutierrez & Feigenbaum, 2006]** define the notion of human identity as a comparison method as follows: "*two humans may not be identical and they must not share a single identity*".

People naturally create different identities within different contexts, for example as family members, citizens or patients. Since, nowadays people change roles, activities and duties more frequently due to modern social communities, there are different ways to express identity based on numerous contexts. Typically, when individuals are performing a certain role, they are selective in terms of what information they reveal. **[Schilit *et al.*, 1994]** identified important context

aspects as being where you are, who you are and what resources are near. The authors define context based on surroundings, which provide additional sources of information such as where, who and what. **[Jøsang & Pope, 2005]** also describe an identity as "*a representation of an entity in a specific application domain*", where the user may have different identities within different contexts. **[boyd, 2002]** stresses that people negotiate their appearance based on different facets of their personality that are associated with different roles or contexts. Therefore, based on different situations, people often present a particular facet of their identity. According to **[Windley, 2005]**, there are many ways to think about what identity is, such as "*how we define ourselves and how others see us*".

There are many complex philosophical implications surrounding the subject of identity. In our view there is no particular definition for the word identity as it can be used differently within different contexts. We define an identity as a core element about who we are as individual and social beings with a set of characterising attributes. It is about how we present ourselves (verbally, on paper or electronically) and how others see us (such as our reputation and honesty).

2.1.2 Online self-representation

The concept of digital representation is a broad area for researchers. It is a complex issue ranging from philosophical to practical concepts on cyberspace. A digital identity is the representation of physical identity that can be distributed within a network for representation and interaction with other people. In his book *Digital Identity*, Phillip J Windley defines digital identity as "*data, which uniquely describes a persona or things and contains information about the subject's relationships to other entities*" **[Windley, 2005]**.

Researchers such as **[Schau & Gilly, 2003]** employ the theories of self-presentation, possessions, and computer-mediated environments. This research was conducted to understand how online self-presentation relates to the physical performance of identity, and the motivation of people who create websites that change based on context. The purpose of this theory is based on Goffman's theories of identity and social performance, which explains how self-representation reflects external social observations **[Goffman, 1959]**. Goffman's thesis indicates that self-presentation is a component of identity that projects a desired impression.

A digital profile has become a general mechanism for presenting one's identity online. It provides an opportunity to present people as who they really are, or as who they would like to be. Our identity profile comprises a subset of personal information to describe what we look like, how we behave, where we live, how others can get in touch with us as well as our personal and professional circumstances. According to **[Donath & boyd, 2004]**, the creation of a social network website represents an individual's public persona and their network of connections to others. **[boyd & Heer, 2006]** emphasise how profiles have become a common mechanism for presenting one's identity online. In particular, **[boyd, 2004]** examines social relations and the motivation of using such a social community by taking a sociological study amongst Friendster users. Other research also focuses on how people present their identity when faced with an unknown audience **[boyd *et al.*, 2004]**.

According to **[Suler, 2002]** there are five factors to examine how people manage their identity representation online: these include the level of awareness, fantasy or reality, and positive or negative attributes. This paper investigates the question of what is one's true identity and how online communities can be trusted. The author also suggests that in real world identification people are wearing masks and do not always reveal what they think and feel, and it becomes more difficult to make a trust decision.

2.1.3 Social and Multiple Identities

Who we are? This is a simple question, but it does not have a simple answer because of the way we represent our identity within society. People are more connected than ever before due to the rise of online social networking that shifts identity from the individual to the social realm. Therefore, another part of our identity is given by the social connection to a particular community, also called social identity. According to **[Grayson, 2002]**, "*identity does not inhere within us; it is a social construct granted by others"*.

In the world of communication and interaction, the knowledge of whom we communicate with is essential to decide which information to make available. Goffman perceives social interaction as an interactive performance, where actors present themselves based on the reaction they receive from others. This idea can distinguish between the terms "*expressions given and expressions given off"* **[Goffman, 1956]**. According to **[boyd, 2002]**, "*social interaction is a negotiation of identities between people in a given environment*". The author considers

identity as two parts of personal and social representation. She indicates that people often realize who they are in association with other people around them. **[Donath, 2007]** also defines identity as who we are while social identity is about what type of person we are.

There is no limit to the number of online identities that an individual can create. For example, a person may have two identities associated with being both a student and an employee at the same university. Furthermore, sometimes people do not just identify themselves as unique individuals, but as part of a group or community. While this group identity seems to be just another identity tag, it is becoming part of our identity **[Baier *et al.*, 2003]**. The fact is that a person may associate with a different identity representation, yet this multiple identity does not necessarily mean that any of them are false. According to **[boyd, 2002]**, people maintain multiple accounts that represent different facets of their internal identity in association with particular contexts.

According to **[Cameron, 2005]** due to the numerous contexts in which identity is presented online, a single identity is not sufficient and a logical method of using multiple identity systems is required. It is increasingly difficult and time consuming to keep track of all of our names and authentication mechanisms in the networks we use, especially since we often identify ourselves differently according to the type of communication in which we wish to participate. According to the 'Laws of Identity' one reason there is no single, centralized system is "*because the characteristics that would make any system ideal in one context will disqualify it in another*" **[Cameron, 2005]**. On the other hand according to **[Damiani *et al.*, 2003]**, "*maintaining multiple identities as separate and independent named sets of attributes or credentials obviously poses huge management problems*". It would be more efficient if people could create and manage a core digital identity by going to only one website and by only having to make changes once such as OpenID **[Recordon, 2006]**. However, such a system may cause more serious problems; for instance, if one central identity is stored on an authenticating server to verify a person, what would happen if the server was hacked? How would that impact on all the sites that use the same identity?

2.1.4 Online Social Networking (MySpace)

Social networking sites are online spaces that offer individuals an opportunity to present themselves and maintain social interaction through their profiles. Today, people with no knowledge of web designing can quickly create a

free webpage portraying an online identity and representing themselves to the rest of the community. The earliest online networking website was sixdegrees.com, which was established in 1997 and closed after four years, but has been followed by many other social networking sites **[boyd & Ellison, 2007]**. Some popular examples of social networking communities are forums and event listing sites (e.g. orkut.com), work related contexts (e.g. linkedin.com), relationship and dating services (e.g. match.com), college communities (e.g. facebook.com that originated as a way of connecting college students), networks of friends (e.g. friendster.com), and music and other interests (e.g. myspace.com).

MySpace is one of the largest social networking sites, offering a customized personal profile, posting of images and comments, and searching profiles to find friends who share common interests. The name 'MySpace' implies that it is a personal presenting space and the central elements are personal information within profiles. This website was established by Tom Anderson (the current president and an alumnus of the University of California) and a group of programmers in July 2003. Later, in 2005, Rupert Murdoch's News Corporation bought MySpace's parent company (Intermix Media) for $580 million. Since 2006 various countries and language specific versions of the site have been released **[MySpace Wikipedia]**. MySpace relies on advertising as the main profit making stream, utilising a variety of advertising opportunities such as banner ads, album promotions, services and products, sponsorships, streaming media and event invitation **[Trendmaker, 2006]**.

Currently this online community has universal appeal; however it is especially attractive to teenagers and those in their early twenties who are interested in music, culture and are often seeking to establish their sense of self within a like-minded community. MySpace has various community and group features, for example, blogging, an e-mail service, instant messaging (IM), online dating, and media sharing within profiles. In addition, the growing number of MySpace members and new features led mobile phone providers to release a series of mobile phones that employ MySpace mobile services through a hand-held device and grant community mobile services anywhere. The population of this social site is increasing rapidly with a current population of 256 million users (as of December 2009). MySpace, initially intended for musicians, gradually became a space for friends to articulate their network and meet others through their profiles. We selected the MySpace network as a case study for our research due to its variety of population.

2.2 Online Identity Concerns

Online social networking provides an overview of a user's identity through a digital profile that represents an image of the individual's identity. If we were able to believe in the validity of this information, we could use the profile information to make decisions about whether to trust someone in a variety of contexts. Since people may have concerns about loss of privacy, identity theft, phishing, etc., we cannot be sure about the reliability of the information provided.

While in the physical world many governments perceive national identity cards as a solution for many problems, the issues associated with the use of digital identities in the context of online social networking are not yet solved. Currently there is no standard model for authentication and authorization of identity management online. Therefore, for people who want to be able to trust the other end of the connection, there is a lack of any structure to verify the friends/opponents in social networking. Furthermore, an individual is likely to have many different versions of their identity that are difficult to manage effectively and accurately. Within this section we aim to highlight some online identity issues such as privacy (Section **2.2.1**), anonymity (Section **2.2.2**), trust and honesty (Section **2.2.3**). In Section **2.2.4** some MySpace concerns such as fake identities and predators that this community faces today in are discussed.

2.2.1 Identity Disclosure and Privacy

Online identity fraud is a rapidly growing crime due to the poor privacy practices on the Internet **[Coates *et al.*, 2000]**. Privacy has become an increasingly important concern to online users, especially in contexts where personal information is disclosed for interactions and communication. According to **[boyd *et al.*, 2002]**, privacy is a dynamic process between the desire of being alone and the desire of interacting with others. They state three common definitions of privacy as:

- The right to be left alone;
- Control of personal information;
- Encrypted data and communications;

A system such as online social networking that shares personal information with others may have serious privacy implications. For instance, there are increasing

concerns about third parties who have access to the information contained in a profile, and misuse or sell people's identities **[Zarandioon *et al.*, 2009]**. There are no technological controls on how an Internet application provider uses someone's personal information and people have little control over their identity distribution.

Currently, there are some systems that identify the user-level personal information disclosure within an effective privacy management framework. For instance, **[Lederer *et al.*, 2003]** use techniques to authorize users in order to customise their privacy preferences by manual configuration. The authors believe that the people whose privacy is in question should make decisions about privacy. Although privacy systems cannot be an absolute solution to protect identity, they have to be flexible enough for an individual to manage privacy in a range of social contexts.

2.2.2 Anonymity and Pseudonymity

According to **[Jøsang & Pope, 2005]**, "*pseudonym is an identifier where only the party that assigned the pseudonym knows the real world identity behind it*". Anonymity is used extensively in digital environments, even for serious or essential online activities **[Pfitzman & Hansen, 2008]**. In some situations, such as online gaming, creating a pseudonym persona is more acceptable, however for other purposes, such as exchanging goods and services, finding friends and seeking employees, it is not acceptable. The right to be anonymous encourages individuals' freedom of expressing themselves. On one hand, anonymity may provide people with the opportunity to extend their reputations based on the quality of their ideas rather than their identity itself. On the other hand, anonymity may bring some conflict over trusting an identity holder within an online domain.

Identity fraud is a rapidly growing crime both online and offline, due to the lack of trust management and the right to be anonymous. For instance, the film *Catch Me If You Can* (with Leonardo DiCaprio and directed by Steven Spielberg), shows the effect of the pseudonym and multiple identities in real life situations. This film is based on the true story of Frank William who played a number of anonymous identities to deceive the system **[Danylak & Edmonds, 2005]**.

Early work on pseudonym detection, for example **[Huffaker & Calvert, 2005]**, investigates the use of language, identity disclosure and gender differences

among weblogs looking for pseudonym behaviour. Due to the relationship between accountability and anonymity online, researchers such as **[Ford & Strauss, 2008]** also proposed a schema for detecting virtual personas using both accountability and anonymity to ensure that a person can only operate one accountable pseudonym at a time. The authors believe that online anonymity often appears to undermine accountability.

2.2.3 Trust Implication and Honesty

Identity and trust are two related concepts: according to **[Seigneur & Jensen, 2004]** *"identity is a central element of computational trust"*. In fact when we trust other people or communities, we are prepared to reveal more honest information about ourselves. Today's Internet has many problems with trust establishment: as **[Steiner, 1993]** states, "*on the internet nobody knows if you're a dog*". **[Park *et al.*, 2002]** suggest that "*trustworthiness refers to the truthfulness of a website's contents and the site's reputation*". In the shared environment of social networking, it is essential to build a trusted environment in order to support collaborative activities while protecting sensitive information.

There are alternative solutions for the implementation of identity in a trusted environment **[Ying & Chris, 2009]**. According to **[Abdul-Rahman & Hailes, 2000]**, "*trust is a social phenomenon and any trust model must be based on the type of society*". In this paper the authors have carried out a survey of the social sciences and identified different characteristics of relevant trust. The goal of their work is to discuss a trust model based on real-world trust characteristics in online communities to help users in identifying trustworthy entities. Other work, such as the two-tiered approach proposed by **[boyd, 2002]**, also provides users with appropriate mechanisms for presenting themselves in a trusted community. Researchers such as Kim Cameron in his white paper 'Seven Laws of Identity' search for a solution to prevent the loss of online trust by defining a unifying identity Metasystem **[Cameron, 2005]**. His Metasystem architecture is based on a set of principles called the "Laws of Identity", which are proposed and universally adopted through continuing dialogue on the Internet.

From a psychological perspective, there are some theoretical aspects about honesty in the physical world. **[Donath, 2007]** defines honesty as one of a quality signal in identity representation. This work is based on 'Signalling Theory', which learns from animal communication and applies to human social interactions. In **[Somanathan & Rubin, 2004]** the authors study honesty

behaviour in market societies by focusing on the employment relationship and the behaviour of workers. This work has examined the factors that affect the accumulation of honesty in growth models and suggests that capital and honesty are co-determined in the long run. Similar work, such as **[Mazar *et al.*, 2007]**, also proposed and tested a theory of self-concept maintenance, which allows people to engage in dishonest behaviour to some level.

2.2.4 MySpace Identity Issues

While the MySpace network creates a broad new set of opportunities in personal and social areas, it also creates new threats and issues for its users. Since there are many concerns about disclosing personal information on this site, online profiling introduces a certain level of uncertainty and, as a result, inaccuracy in self-representation. These concerns, such as loss of privacy and fake identity, will gradually grow as the use of this social network grows. There are several guides for users such as the *Rough Guide to MySpace* **[Buckley, 2006]**, which describes how to play safe on this site, focusing on how to prevent identity theft, predators and false identities.

In the world of online networking it is difficult to identify people who construct a false persona. There is no limit to the number of profiles people can create on MySpace and there is no accurate verification when people join this site; it is therefore easy to create a false persona and connect to other recognized profiles. Faking a known identity is used to position oneself in a status hierarchy, for instance by claiming a connection to celebrities or well-known people. According to **[Donath & boyd, 2004]**, while public displays of connections can verify an individual's identity, they can also help someone else to establish a false identity. At what point does one decide to create a profile of 'Bin Laden' complete with photos, and start posting comments on other fake members such as 'George W Bush'? For some this means just having fun, while others may undervalue the meaning of people's connections, and even become intimidating towards other network residents. Some people also obtain multiple profiles and control different false identities to increase trust between their networks of friends; these malicious users are also called Sybil attackers **[Douceur, 2002]**. Another problem is related to identity manipulation such as in the case of switching gender. Women for instance receive more attention and are usually better supported online **[Nabeth, 2005]**.

Recently, the number of members on MySpace has declined due to a number of security issues compared with other social networking sites. For instance, there is deficient use of email verification as MySpace sends an email to verify each email address but does not check the email verification. It seems that people can register under any email address, only if the address is not used for another account. In other social communities, such as Facebook, college email addresses were initially used as technical authentication; this would then connect the online persona to the real person and makes their network more trustworthy. According to **[Dwyer *et al.*, 2007]**, which conducts a survey to compare MySpace and Facebook sites, MySpace has a poor reputation of trust compared with the Facebook network. The authors indicate that MySpace users have less trust in the site itself, and they and their friends are less willing to reveal identifying information about themselves.

On the other hand, MySpace allows younger people (age 14 to 16) to join as long as they change their profile setting to private. However, this conveys a lot of negative views towards this site such as the fear of sexual predators. Online predators take advantage of simple access to personal information published in profiles to trace their target. In particular, this alerts leading social networking companies to protect their site from 'sexual' predators. Accordingly, some schools and public libraries in the countries where MySpace is most used considered having restricted access to the site to protect their youngest members from sexual predators and malicious users **[MySpace Wikipedia]**. Researchers such as **[Lee *et al.*, 2008]** have concerns about users' safety online. They looked at how to motivate users to engage in self-protection behaviour and defend against predators and malicious viruses. Additionally, there are some hackers, such as the popular profile on MySpace called 'Samy' **[Lai, 2005]**, who created a self-propagating script to automatically make anyone who viewed his page his friend without obtaining any permission. In October 2005 the worm spread by duplicating itself into each friend and friends of friends, rapidly increasing Samy's friends and therefore overloading the MySpace servers. Though the Samy's worm was friendly, other hackers might use their skills to destroy or steal personal information, even from private profiles.

2.3 Related Work

This section aims to review some related works that appear to follow similar ideas for detecting identity representation in digital social applications.

The majority of these solutions follow the privacy **[Acquisti & Gross, 2005]** and trust implication **[Dwyer *et al.*, 2007]** approaches. Several smaller projects, such as **[Dokas *et al.*, 2002]**, **[Airoldi & Malin, 2004]** and **[Thongtae & Srisuk, 2008]**, have utilized data mining approaches in order to control the verification of identities online. Recently there have been some academic papers and studies in the same field of identity and social networking, such as **[Ying & Chris, 2009]** and **[Stutzman, 2006]**. A considerable amount of research has focused on social networking sites, including the social network structure, such as **[Petroczi *et al.*, 2006]**, **[Mislove *et al.*, 2007]** and **[Spertus *et al.*, 2005]**. However, there is little similarity between these studies and our approach since none of them propose a method to directly solve the problem of detecting and verifying the reality of identity.

This section briefly highlights some of the related publications and projects. These related works include identity management systems (Section **2.3.1**), evaluation of online social communities (Section **2.3.2**), social network analysis (such as friendship, similarity and personality analysis) (Section **2.3.3**), data mining (Section **2.3.4**), deception detection (Section **2.3.5**), and recommendation system (Section **2.3.6**).

2.3.1 Identity Management Systems

Many systems propose strong end-user controls over how identity information is distributed and managed. Two well-known identity management approaches are Passport by Microsoft (www.passport.com), and Federated Identities by the Liberty Alliance Project (www.projectliberty.org), both providing complete identity management architecture. There are some other projects that propose new identity management models to improve issues of online identity. For instance Microsoft InfoCard is a set of technologies that aims to overcome the current problems with digital identity management by allowing people to use their identities as easily and securely as in the physical world **[Gutierrez & Feigenbaum, 2006]**. One potential problem with InfoCard is that it is partly up to users to deal with the trust issues.

Other work, such as **[Pato, 2003]**, defines identity management as a set of processes, tools and social contracts, which enables secure access to an expanding set of systems and applications. This paper primarily categorises three models for deploying identity management systems: *Silo*, *Walled Garden* and *Federation*. Silo is the predominant model on the Internet today, while

Walled Garden represents a closed community of organizations and Federation includes systems such as Microsoft.NET Passport and the Liberty Alliance project.

On the other hand, **[Koch, 2002]** argues that a user-centric global identity management system is required to personalize identity. The author proposed a specific identity management solution called an IDRepository, which focuses on user empowerment to control profile attributes. His technical approach is to store the user's identity in a central place where it can be maintained by the user and accessed by different services. However, this central system requires strong privacy protection and must provide complete control to the user over deciding what information to disclose and share with which parties. Also recent systems, such as OpenID, introduced a centralised system to access the entire identity account into one single location **[Recordon, 2006]**.

When we decide on what to reveal about ourselves, we are performing identity management. An efficient identity management system would allow people to decide how to give data and when to act anonymously. In addition, the system should be able to validate a profile's identity and perform a trust decision based on an individual's representation, which we are aiming within this investigation.

2.3.2 Evaluation of Online Communities

The success of online social networking sites has attracted the attention of many researchers. A number of papers evaluating online social community sites have been published in recent years mainly focused on privacy protection and security issues. Although there is not sufficient academic work examining the validity of identity, previous research in this field shows the extent of research interest, and the lack of a solution to the social networking problems.

There are few academic studies on digital relationships and the structure of social interaction using MySpace. For example, **[Caverlee & Webb, 2008]** studied the characteristics of MySpace profiles based on facets of this social network such as sociability and the use of language. The authors in this paper analysed the language used in profiles to find the distribution terms in a large-scale database. They used a demographic model to represent the probability of each attribute within gender. This study has a slight similarity to our work; however, the focus was only on two identity traits (such as age and gender). Works such as **[Dwyer, 2007]** have studied the structure of social networking

using the MySpace generation as a case study, focusing on a framework to capture user's attitudes that influences interaction with others.

Other work such as **[Perkel, 2006]** argues that MySpace could be an informal learning environment to promote the development of new literacy. The author believes that although MySpace is not a perfect environment for learning and expressing languages, it may still provide new forms of literacy practice. Online community researcher danah boyd (she prefers her name in lower case), in a talk called 'Why youth heart MySpace', also indicates youths' desire to learn social culture through their online space by exploring their identity formation **[boyd, 2006]**. In particular, the author discusses issues of trust and intimacy on online networking using Friendster as a case study **[boyd, 2003]**.

The majority of researchers within this field are examining the Facebook community with the focus on identity presentation and information sharing. For instance, **[Acquisti & Gross, 2005]** examined the pattern of information exposure on Facebook profiles. The authors looked for identifying attributes, such as name, email and image, and focused on the visibility of profile information and possible forms of online attacks. In other works, such as **[Acquisti & Gross, 2006]**, the authors discuss the demographic differences of student's behaviour in regards to their privacy implication. Other works, such as **[Stutzman, 2006]**, have studied identity-sharing behaviour in online communities including the protection of information disclosure. **[Ellison *et al.*, 2006]** also analysed the Facebook network structure and the role of identity management in relation to social college life. Similar to our work, they analysed profiles' identity elements; however, they mainly focused on whether users are aligned and the positive outcomes associated with the use of this social network.

There are some other related works in the field of identity disclosure and privacy management in online social communities. For instance, **[Tufekci, 2008]** examined the disclosure behaviour on MySpace profiles and the relationship between disclosure as well as the issue of privacy. The author found little or no relationship between online information disclosure and privacy concerns. However, we show later that profiles who feel more concerned about their privacy (such as private profiles) disclose more honest information about themselves. Tufekci states that the "*disclosure behaviour on these websites was more heavily influenced by demographic characteristics of the participants*", believing that a user's characteristics can influence identity representation. This work has a comparable approach to our study, but the author has only focused on the disclosure aspect of identity by following Gofman's theory **[Goffman,**

1959] and Altman's theory **[Altman, 1977]** of privacy, and produces some statistics of how much information people reveal about themselves online.

Other works, such as **[Berman & Bruckman, 2001]**, have conducted some research on the different ways in which men and women behave online, and if people's communication patterns can help to determine their identity. The authors conducted a survey on online gaming to find out what motivates people to have such behaviour. The authors also look at whether peoples' communication patterns can help determine information about an individual's age, race or national origin. This work is related to ours in terms of analysing online identity elements by looking at gender difference and examining the deceptive information. Early work on the theory of personality, for example **[Casciaro, 1998]**, believes that personality has a strong correlation with accurate perceptions of networks that depends on both individual differences and situational factors. Looking at the accuracy of friendships, the author argues that an individual's position in the social structure and their personality traits are potential determinants of the expected accuracy in network perception.

Many researchers have focused on users' personalities. For example, for a better understanding of the personal and social implication of weblog authorship, **[Marlow, 2006]** investigated a large-scale survey on a weblog network. This work examines the user's personality, reflecting on the social effects of weblog authorship. The author shows that the community rewards the authors who put time into their network and give them more accountability credit. **[Bechar-Israeli, 1995]** also indicates that there is a strong link between one's nickname and personality.

There is a research gap in the area of the transformation of identity over time within online social networking. For instance, **[Kumar *et al.*, 2006]** presented a simple model of network growth and measured the evolutionary structure of a large social network. **[Hill *et al.*, 2006]** assessed the correlation between past and future representation of nodes behaviour to formalize a dynamic network representation. On the other hand, **[Holms *et al.*, 2004]** analysed a large number of Internet dating communities and examined multi-scaling behaviour on online social networks. In particular the authors focused on the time evolution and degree function of the network, believing that online interaction creates an unstructured network.

2.3.3 Social Network Analysis

The enormous expansion of online networking has motivated many researchers to produce a variety of literature on social network analysis. Many studies on social network analysis attempted to understand the graph of social networks and measured the properties of the network structure. For example, **[Mislove *et al.*, 2007]** observed four popular networks (Flickr, LiveJournal, Orkut, and YouTube) on a large scale and discussed that the in-degree (a count of the number of links directed to each node) correlated to the out-degree (the number of links that each node directed to others). The authors believed that connecting to high degree nodes in social networks is commonly used, while high degree nodes connected to low degree nodes shows the opposite behaviour. The authors suggested that understanding the structure of online networks can lead to an algorithm to detect trusted users by looking at the network properties and distribution rather than the identities themselves.

Previous studies, such as **[Hsu & Helmy, 2006]**, measured the relationship between close nodes and examined the influence of network connectivity on each relationship. This work analysed a small model to help understand the characteristics of network structure and user behaviour. In another work, researchers analysed the community structure from mobility traces and evaluated the different community detection approaches to identify both static and temporal communities **[Hui *et al.*, 2007]**. Early work such as **[Petroczi *et al.*, 2006]** aimed to develop a tool that provides a quantitative and continuous measure of strength of ties in virtual communities. Using a Danish social network site, **[Ryberg & Larsen, 2008]** also argue that the notion of exploring weak and strong ties is a valuable contribution to network structure, although this perspective does not automatically reflect on the individual's understanding of their influential or central positioning.

Other work, such as **[Cameron, 2004]** and **[Russo & Koesten, 2005]**, looked at the psychological aspect of social identity and proposed a social identity representation using three factors to examine the efficiency of these models within different studies. They examined the efficiency of three factors of social identity, centrality (being in the group), in-group effect (belonging to the group) and in-group tie (similarity to the group), and hypothesised that social identification can be represented based on these factors. **[Yoneki *et al.*, 2008]** also define centrality as an important property of network structure. Within our study, we will show how identity can be validated partly based on network structure such as centrality and similarity factors. This section highlights

similar works on friendship in Section **2.3.3.1** and similarity analysis in Section **2.3.3.2**.

2.3.3.1 Friendship Analysis

An identity can often be defined as *"You are who your network is"*, which emphasises a particular significance of the social network and its relationship to others **[Nardi *et al.*, 2000]**. Therefore, people reveal information about themselves depending on their audience. The theoretical and design implication of predicting friendship links from Facebook profiles is discussed by **[Lampe *et al.*, 2007]**, and suggests that the information provided by individuals has an impact on who interacts with them, and identifies the effects of identity information on users' interactions with others.

According to **[Donath & boyd, 2004]** there are different types of connections between people, which apply to both offline relationships as well as online connections. These include:

- **Friend:** someone known with possible interaction in the offline word, who is trusted for online communication and sharing identity.
- **Familiar stranger:** someone not necessarily known but one may communicate with such as friends of friends, who are less trusted than direct friends.
- **Stranger:** someone not known but have some link for some reasons such as gaining more links and popularity through connection.
- **Community:** anyone in the community, even those who are not similar or trusted.

Other social network researchers have observed how people make friends and how people rely on their friends for social support. For instance, 'Socialize This!' by Andrew Zolli discusses how certain social networking users deal with the feeling of dissatisfaction over the low number of friends on their friends list **[Zolli, 2004]**. Researchers such as **[Fono & Raynes-Goldie, 2006]** also examined user understanding of the term 'friend' on online social networking. The authors in this work analysed user behaviour and public articulation with the effect on social conflict, focusing on the LiveJournal network. They defined core keys of understanding friendship as: content, an offline facilitator, online community, trust, a courtesy and a declaration. The authors believe that online friendship is weak, and the higher number of friends in the friend list is equal to experiencing more social conflict or drama.

Analysing the type and quality of online friendships is out of scope of our research. However, it would be interesting to study this further and examine the correlation between the type of friendship and an individual's identity representation.

2.3.3.2 Similarity Analysis

There are many studies on similarity measurements, such as **[Adamic & Adar, 2003]**, which examine the personal homepage information of a university to predict relationships between individuals. This study uses a similarity ranking method to predict the possibility of one being a friend of another based on their text, links and mailing lists. Previous works, such as **[Abbasi & Chen, 2008]**, also applied a stylometric analysis to online texts with a novel pattern disruption mechanism; this can be used for identification and similarity detection of authorship**.**

Recent work, such as **[Brzozowski *et al.*, 2008]** and **[Hogg *et al.*, 2008]**, shows the impact of friend's similarity, by examining the influence of friendship and voting behaviour using a social network group (essembly.com). By analysing ideological social networks and distinguishing between 'friend', 'allies' and 'nemeses', they found that people have a greater similarity to their allies than friends, though users are more influenced by their friends. **[Maia *et al.*, 2008]** also proposed a methodology for clustering and identifying similar user behaviour in online social networks using YouTube data. The authors used a k-means (a clustering algorithm) to group users with similar behavioural patterns, believing that their model has the potential to improve recommendation systems in online social networks. In addition, **[Spertus *et al.*, 2005]** presented a comparison of six different similarity measures based on users' self-reported text for recommendation by using the Orkut social networking site.

Some other studies, for example **[Strauss *et al.*, 2001]**, examined the effect of similarity on rating different datasets and measured how similarity can impact on the relationship between individuals. The authors aim to answer how an individual's personality is related to their similarity by analysing different models of personality such as extroversion and emotional stability. Other work, such as **[Sherif *et al.*, 2000]**, uses the user's typing behaviour and a similarity measure, based on a simple matching methodology, to authenticate users as a pattern to signify a graphical representation. **[Casciaro, 1998]** also argues that the

situational factors and individual similarity and differences have an effect on the accuracy and performance of social networking.

2.3.4 Data Mining and Machine Learning

Many recent works have employed machine-learning techniques to find patterns in users' behaviour on online social networking. For instance, a previous study on prediction used a Bayesian framework to predict a user's identity such as age and gender based on users web browsing behaviour history **[Hu *et al.*, 2007]**. Their proposed algorithm improved when compared to the baseline algorithm, as their study shows 79.7% accuracy on gender prediction and 60.3% on age prediction. The authors are planning to predict other entities such as location and education. This work is comparable to our study; however, their method is not able to analyse if the predicted age and gender are real or false.

Previous studies, for example **[Hsu *et al.*, 2007]**, define a set of machine-learning approaches to predict and classify friend relations and profile information using LiveJournal data. This study documented attribute features, which are dependent on graph properties and an individual's demographic attributes. **[Galloway & Simoff, 2006]** focused on human-centred network data mining methods looking for the correlation between data attributes and the value for each attribute, using NetMap for visualisation and analysis of data relationships.

Other studies used data mining frameworks to construct a detection model by uncovering important patterns in data. For instance, **[Dokas *et al.*, 2002]** and **[Stolfo *et al.*, 2000]** constructed a class prediction model to identify attacks from both known and unknown intrusion, based on two categories of misuse and anomaly detection in network intrusions. Work such as **[Thongtae & Srisuk, 2008]**, has shown the efficiency of data mining techniques for the analysis of crime detection. **[Airoldi & Malin, 2004]** also introduced a filtering system for capturing email spam focusing on text classification. This approach examines the text used in email using data mining methods and classifies them into groups of frauds and non-frauds. Similar work, such as **[Badaskar *et al.*, 2005]**, used the characteristic of real text as a feature to distinguish real articles from fake, by creating a classification approach based on a language model. **[Agrawal *el al.*, 2003]** also used the same method of examining the text-based

and link-based profile information from a newsgroup site to develop a link detection algorithm.

Concentrating on unsupervised methods, **[Malin, 2005]** studied on deciding when two pieces of data correspond to the same entity. This work relies on name similarity and employs a hierarchical clustering method using an Internet movie database. The author evaluated several unsupervised methods such as hierarchical clustering and random walks for disambiguating names (where the same name references multiple entities). A novel approach also projected finding pseudonyms by automatically generating lexical patterns using a set of real-world name-alias data **[Bollegalla *et al.*, 2008]**. They used vector machine learning for ranking and evaluating the confidence of an alias for a name using anchor texts and page counts.

2.3.5 Deception Detection

Many recent works have prototyped the idea of deception detection based on user attributes. For instance, **[Zinman & Donath, 2007]** focused on deception detection on social networks by developing a research prototype to categorize spammer and non-spammer by inserting trust values into the system. Their categorisation is based on many factors such as sociability, events, actions, emotions, social relationships, and so on. Other studies, such as **[Toma *et al.*, 2008]**, offer some important insights into the practice of deception in the arena of online dating. This work addresses the self-presentation issue by comparing the information presented by daters to establish the truth about the information on online dating profiles. **[Burgoon *et al.*, 2005]** also proposed an automated tool to identify deception in the non-verbal communication environment by introducing a four-dimensional profile that reveals an individual's emotional and cognitive status, such as active/passive and tense/relaxed behaviours.

Other works, such as **[Mundinger & Le Boudec, 2005]** introduced a mathematical model to investigate the impact of liars who were trying to influence their social network and gain reputation. The authors assume that *"liars either have extremely positive or extremely negative behaviour to achieve maximum impact"*. Whitty, in her paper 'Liar, Liar!' also focused on examining liars on chat rooms **[Whitty, 2002]**. The author examined the gender differences in terms of online dishonesty, stating that: "*men are more likely to lie about their socio-economic status while women usually lie for safety reasons*". She highlights

the differences of interpersonal interaction in chat rooms and suggests that the active users are more likely to be expressive about themselves in chat rooms.

In addition, researchers have demonstrated how the revealed information in social networks can be exploited for social phishing and other attacks. For instance, **[Shrivastava *et al.*, 2008]** extracted a social network structure to identify a class of potential attacks on the network by proposing two algorithms, 'GREEDY' and 'TRWALK'. Other related works, such as the SybilGuard project, detect Sybil attacks in a distributed social network with the knowledge that Sybil users do not generate many trust links to non-Sybil users **[Douceur, 2002; Yu *et al.* 2006]**. Similar work also implemented a protocol as a simulation and tested against extracted data from the Orkut social networking service, which uses the social links between users to identify a Sybil attack **[Lesniewski-Laas, 2008]**.

2.3.6 Recommendation Systems

Trust enforcement systems have become a focus of the research community over time. Current recommendation systems such as eBay have serious effects on the user experience due to the rapid increase of online communities. Previous work on trust systems, such as **[Shand *et al.*, 2004]**, believes that "*people categorise people they know according to the type of trust they place in them*". For instance, close friends are more trusted than neighbours and colleagues. The authors proposed a trust framework using a consistent recommendation system, to allow users to share and exchange sensitive information.

Previous works on recommendation systems, for example **[Hsu *et al.*, 2006]**, also presented an approach based on collaboration recommendation on weblogs using Livejournal as a case study. **[Felt *et al.*, 2008]** proposed a browser modification to distinguish trusted and un-trusted online content. Believing that user-side filtering is less complex and cost effective, this could be protected by a browser enforcement policy. In addition **[Berkovsky *et al.*, 2007]** examine users' opinion about the impact of rating systems and their privacy.

2.4 Literature Summary

Online social networking and identity representation are active research areas with input from computer sciences, statistics, sociology, and psychology. Researchers in these fields propose different hypotheses, which can help to understand how an identity verification system should be built online. The previous and current studies mentioned in this chapter are mainly focused on the demographic results of network structure and privacy issues of online social networking. These papers have studied different metrics for evaluating online social communities and network analysis. However, there is little academic work towards determining the reality of identity in such online communities. None of the past studies determined if the self-reported identity in social networking is accurate, therefore, there has been little evaluation of the quality of online identity metrics. These studies fundamentally differ from our work as we mainly concentrate on determining the validity of identity within an online social community using data mining techniques. We believe that our approach is the first attempt to find the validity of identity on a large-scale data.

However, these studies help us to understand the wider picture of our research problem by discovering the research gap on the validation of online identities. Within this chapter we highlighted the most important papers, which are related to our research questions as follows:

We described the definition of identity described by **[Goffman, 1956]**, **[Baier *et al.*, 2003]**, **[Donath *et al.*, 1999]**, **[Pato, 2003]** and **[Gutierrez & Feigenbaum, 2006]**. In more common terms they defined identity as the means of distinguishing one from other. Studies on the psychological aspects of social identity representation examined the social implication of displaying identities publicly, for example **[boyd, 2004]**, **[boyd & Heer, 2006]** and **[boyd *et al.*, 2004]**. We argued that, according to **[Schilit *et al.*, 1994]** and **[Dey & Abowd, 1999]**, identity representation is based on context. Therefore, in the context of online social networking people sometimes maintain multiple accounts that represent different facets of their internal identity. However, there are systems such as OpenID to centralize identity accounts and enable users to gain some control and privacy **[Recordon, 2006]**.

We addressed online identity issues such as identity disclosure and privacy (such as **[Coates *et al.*, 2000]**, **[boyd *et al.*, 2002]** and **[Zarandioon *et al.*, 2009]**), the effect of anonymity **[Jøsang & Pope, 2005]**, and related works on trust and honesty (such as **[Cameron, 2005]**, **[Donath, 2007]** and **[Park *et al.*,**

2002]). We also tackled some of the issues related to MySpace identity disclosure (such as **[Donath & boyd, 2004]** and **[Lee *et al.*, 2008]**).

We reviewed some related works that appear to follow similar ideas for detecting identity representation in digital social applications. The majority of these solutions follow the evaluation of online communities, such as **[Caverlee & Webb, 2008]** and **[Ellison *et al.*, 2006]**, privacy issues, such as **[Acquisti & Gross, 2005]** and **[Tufekci, 2008]**, behavioural analysis, such as **[Marlow, 2006]** and **[Holms *et al.*, 2004]**, trust implication approaches **[Dwyer *et al.*, 2007]** and recommendation systems, such as **[Hsu *et al.*, 2006]** and **[Felt *et al.*, 2008]**.

We described previous studies on social network analysis, for example **[Hsu & Helmy, 2006]**, **[Hui *et al.*, 2007]** and **[Ryberg & Larsen, 2008]**, including friendship analysis, **[Fono & Raynes-Goldie, 2006]**, and similarity analysis, such as **[Brzozowski *et al.*, 2008]** and **[Hogg *et al.*, 2008]**.

Many recent studies such as **[Thongtae & Srisuk, 2008]** and **[Dokas *et al.*, 2002]**, have shown the efficiency of data mining techniques and analysis on pattern discovery. We highlighted previous studies, such as **[Hsu *et al.*, 2007]** and **[Hu *et al.*, 2007]**, which defined a set of machine learning techniques to predict and classify friend's relationships and user profiles. In addition **[Badaskar *et al.*, 2005]** used the characteristic of real text as a feature to distinguish real articles from fake.

The idea of deception detection based on user attributes and behaviour, such as **[Zinman & Donath, 2007]** and **[Toma *et al.*, 2008]**, are also highlighted. Other studies have also engaged on detecting liars and deception online, for instance **[Mundinger & Le Boudec, 2005]** and **[Whitty, 2002]**. In addition, we addressed studies that demonstrated how the information revealed in social networks can be exploited for social phishing and other attacks, for example **[Shrivastava *et al.*, 2008]**, **[Douceur, 2002]** and **[Yu *et al.*, 2006]**.

Chapter 3

Research Approaches

"The art of being yourself at your best is the art of unfolding your personality into the person you want to be. . . . Be gentle with yourself; learn to love yourself, to forgive yourself, for only as we have the right attitude toward ourselves we can have the right attitude toward others."

Wilfred Peterson

3.1 Introduction

As the preliminary step to answer our research questions we first designed a customized crawler to gather data from MySpace profiles. Founded in 2003, MySpace is one of the largest social networking sites; it offers members the ability to customize their profile and control privacy and creativity of their pages. Members are free to express themselves and make the page their own by embedding music, video clips, Flash content and photographs. In December 2007 the total number of MySpace users was about 176 million; this number had grown to over 253 million in December 2008. The massive increase of 43% (76 million) in a year without considering the number of users, who left the site, indicates the high popularity of this social site. The MySpace population increases rapidly, although recently it had a massive decline in the number of members and visitors (256 million members as of December 2009) due to competition with other established social networks, such as Facebook **[Torkjazi *et al.*, 2009]**.

The MySpace community was chosen as a case study for our research due to its widespread population that accommodated a variety of people. By examining a profile's content we were able to construct a classifier model that can distinguish between different types of identity representation. The data was first accumulated by a robust web crawler that gathered the personal, professional and relationship information of connected profiles. Using a qualitative approach we then gathered some profiles with a known identity (such as real or fake). By

distributing an email survey we acquired an individual's views about their identity representation by asking them to rate the level of honesty of themselves and their friends. We also set up an online form and asked participants to generate a fake identity. These profiles with a known identity were used in our data mining process as a training dataset, so we learnt from data, evaluated their predicted accuracy and built up our classifier in parallel.

The collected data was used to implement a model that built a personality classifier for the determination of the type of identity, such as real or fake. We proposed a model of seven personality features, such as *expressive/anonymous*, *valid/fantasy*, *traceable/untraceable*, *active/inactive*, *popular/isolated*, *positive/offensive* and *sociable/unsociable.* These personality factors were chosen based on several different types of studies; these include a literature study, data mining, social network analysis and principal component analysis (see Section **3.3.1** for more detail). The values for these personality factors were discovered by methods such as text mining and social network analysis. Using text-mining techniques, we compared each text-based profile's content against several lists of known terms and classified them into different personality attributes. For instance, to check the validity of a location, the described address was compared to a database of known cities/countries, as well as a list of fantasy and offensive terms. Social network analysis, such as centrality analysis, was applied to explore the network structure, and examine '*popularity'* and 'sociability' characteristics by comparing the relationship between these attributes and the type of profile. The personality classifier is validated by data mining and principal component analysis. In addition, to measure how online identities are transformed over time, we applied evolution analysis for the same set of profiles over a period of time.

We have used several approaches to overcome our research problem, including the knowledge gained from the literature review. This chapter aims to describe the overall research approaches, including data collection and modelling the classifier. Section **3.2** describes data collection methods (such as crawling), and the survey study together with brief descriptions of data. In Section **3.3** we explain the procedure of modelling the classifier by defining personality factors and describing why they are chosen (Sections **3.3.1** and **3.3.2**). Sections **3.3.3** and **3.3.4** describe the process of text mining and network analysis methods used to examine and rate each personality factor. Section **3.4** also describes the evolutionary features of identity representation by examining profile information of the same person over time. We show how profile attributes change over a

period of time. This chapter will conclude with a discussion and summary of our research approaches (Section **3.5**).

The following chart (**Figure 3.1**) illustrates the research procedure from crawling data and classification model to data mining evaluation and identity transformation analysis.

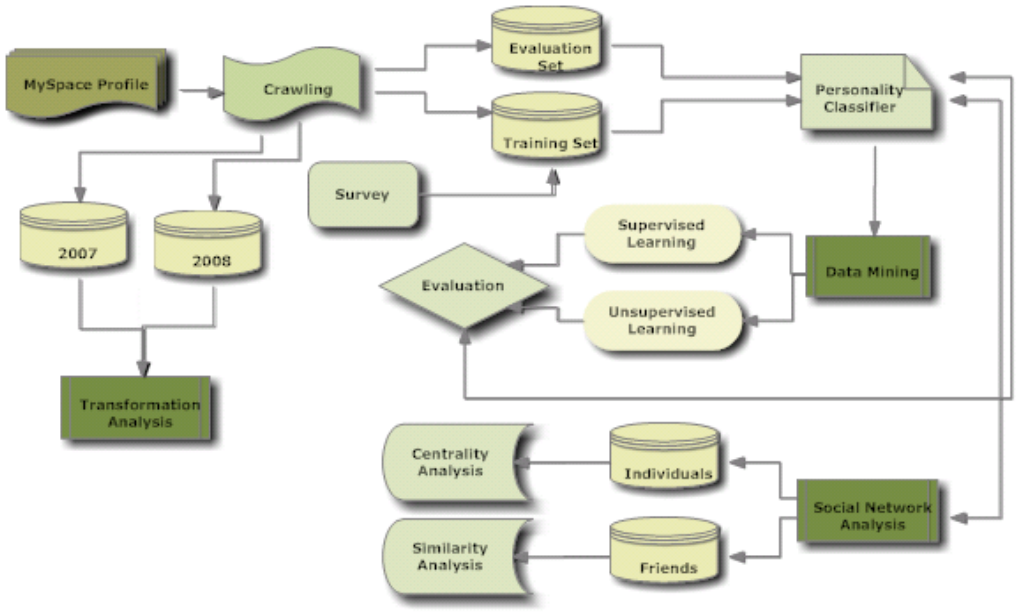


Figure 3.1 The research methods procedure

3.2 Data Accumulation

In this section, we describe the data collection process, including a description of our dataset. The first step for our research was to collect profiles that are rich with personal information. We first employed a quantitative study for mass downloading of MySpace profile content. A crawler was designed to accumulate information, such as personal and professional information as well as relational information (such as number of friends, their comments, and who are friends with whom). The data was extracted by the selection of 'FriendID' (MySpace members' unique number) to crawl pages up to two degrees of separation, targeting their top 40 friends (Section **3.2.1**). We also applied a qualitative study, such as an email survey, online form and manual search to gather a number of profiles as a training set with a known identity (Section **3.2.2**).

3.2.1 Robust Crawler

A custom crawler with time efficiency was implemented to collect a wide range of information required for this study. The crawler was written in PHP (hypertext pre-processor) to extract specific information from profiles and store the data within several MySQL databases. The PHP language was chosen because of its flexibility to limit and fetch the required information from each profile using 'regular expression', rather than storing the entire profile's content. Due to the use of embedded HTML code and major personalization on MySpace profiles, the crawler was customised and carefully designed to collect the required information automatically.

In the first stage of crawling, the automated script started by generating a list of a random selection of 50 seeds and accumulated 202,835 profiles, saved into the 'Seed' database (see **Figure 3.2**). The crawler used a unique FriendID to fetch the profile pages up to a depth of two degrees (friend and the friends of friends). After obtaining the seed's profile, we further collected the top 40 friends (the listed friends on the first page with no order) of each profile. Top 40 friends are selected for two reasons: firstly, friends on the first page may represent a real friend as people generally list their close friends toward the top of their friend list. Secondly, due to the large number of friends (thousands and millions) listed in the majority of profiles; it was less feasible to follow one stream of connection as we were looking for different groups of connected people. In the last stage of crawling, the collected FriendIDs of top friends were used to retrieve the profile information of 2,008,398 friends and saved into the 'Track' database. This technique is known as snowball sampling **[Goodman, 1961]**, which starts with some random seed and increases by the number of links to each seed. In total, we have accumulated about 2.2 million profiles, which represent only 1% of the entire MySpace population (176 million registered members at the time of crawling in December 2007).

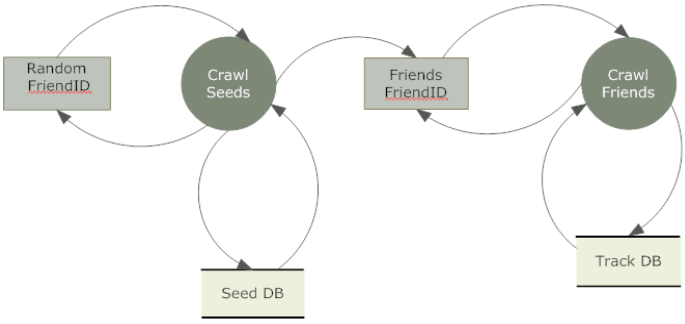


Figure 3.2 The crawling procedure

The number of collected profiles within three main categories is described in **Table 3.1** as '*Public*' (personal profiles), '*Private*' (limited profiles), and '*Bands*' (with music related information). The profile information is explained in more detail in Section **3.2.3.** In addition, the same profiles were collected a year later to undertake a transformation analysis on the identity representation over time (see Section **3.4**).

Table 3.1 The number of collected profiles by each category

Type of profile	Number of profiles		Total
	seeds	friends	
Public	113,969	1,101,032	1,215,001
Bands	19,908	181,371	201,279
Private	68,958	725,995	794,953
Total	202,835	2,008,398	2,211,233

The crawling algorithm is efficient and reasonably fast as the computation time for accessing and storing profile information, together with friends' connections, into a database is less than three seconds on a single processor. After improving our crawler and applying a multiple scripts program on a server (using a cron job), we were able to download almost one million unique profiles per day, including the time to identify mutual friends. Therefore, every profile on MySpace could be collected in a maximum of approximately seven months, taking into consideration new members (note that this social site increases by almost 5 million members each month at the time of crawling in 2007).

3.2.2 Qualitative Study

The information collected using our crawler was used as a '*validation*' dataset (profiles with unknown identity type), where the nature of identity, i.e. whether it is real or fake, is unclear. The test set is used for clustering data more effectively by pattern discovery and grouping each identity element. For instance, clustering data helped us to decide which personality factors are more important when we applied a data-mining algorithm such as unsupervised learning (see Section **4.3.3**). However, a set of profiles with known identity (such as real or fake) is required. To provide profiles with known identity types ('*training*' set) we applied a qualitative study to manually identify the true identity of profiles. We utilized a survey to contact a number of people via email at the University of

Sussex to confirm the ownership of a profile and accumulate the level of honesty for both participants and their friends based on self-rating. We also collected a number of fake invented profiles by asking participants to generate a fake identity. The training set is used to find a pattern in data when applying data mining algorithms such as supervised learning (see Section **4.3.2**) and predict the type of identity. The number of collected groups for each training set is represented in **Table 3.2**. We identified four types of users; *'real-celebrity'*, *'real-local*', *'fake-celebrity'* and *'fake-invented*', described as follows:

- **Real-celebrity:** Official profiles representing famous people, such as celebrities who have their name listed in the directory of official profiles on MySpace. As their page is recognized by MySpace as an official page, we assume that they are representing a real person. These profiles are obviously well connected and have very large numbers of friends that might affect our results; therefore, we needed to collect the profile information of local users.
- **Real-local:** Current students at the University of Sussex who responded to our email survey questionnaire (118 responses from 2019 emails) verified that the profile belonged to them and rated their level of honesty. Students from the University of Sussex were selected because, by comparing participants' information on the University directory website, we could check their true identity. We asked participants to confirm whether their page belongs to them; how they rate their level of honesty according to their identity representation; and if they think their friends (on average) are trusted and represented their true identity.
- **Fake-celebrity:** Those who fabricated known profiles (such as celebrities) with almost the same information; for instance, many profiles claim to be 'Britney Spears' or 'Osama bin Laden', which are not authorized in the official directory website. We determined these profiles manually, for instance by knowing of another real profile for the same person. Those who impersonate other people might be a fan of celebrities, or intend some political view, or are just having fun by making a fabricated profile.
- **Fake-invented:** We set up an online form and asked people, such as the University of Sussex FOSS group of Facebook friends, to generate a fake profile. We asked them to create an imaginary profile to represent any identity except their real identity or known people (such as celebrities, politicians, etc.).

Table 3.2 The number of participants in each training and validation dataset

Type of Identity	Known profiles (training set)	Unknown profiles (validation set)
Real-celebrity	417	not known
Real-local	118	not known
Fake-celebrity	457	not known
Fake-invented	308	not known
Total	1300	2,211,233

In our email survey at the University of Sussex, 100% of the 118 responses received confirmed that the profile belonged to them. Participants were asked to rate their level of honesty from 1-100%, where 100 is the highest level of honesty. They were also asked to rate the level of honesty of their friends. This survey helped us to understand how honest people are in terms of their identity representation and how much they trust their friends. For instance, **Figure 3.3** illustrates that the majority of participants rated their level of honesty higher than their friends' honesty. Of the total, 38% claimed that they are 100% honest, and many of participants fall into the range of 80-99% honesty value. Almost 18% of participants have set their profile visibility to private. On average, participants with a '*private*' profile rated their level of honesty higher than the '*public*' users. Through transformation analysis, we later found that '*private*' users altered their profile information less, which may indicate that they are more honest about their identity representation (see Section **5.3**). Also, through similarity analysis we found that '*real-local*' users are more similar to their network of friends (see Section **5.2.1.2**).

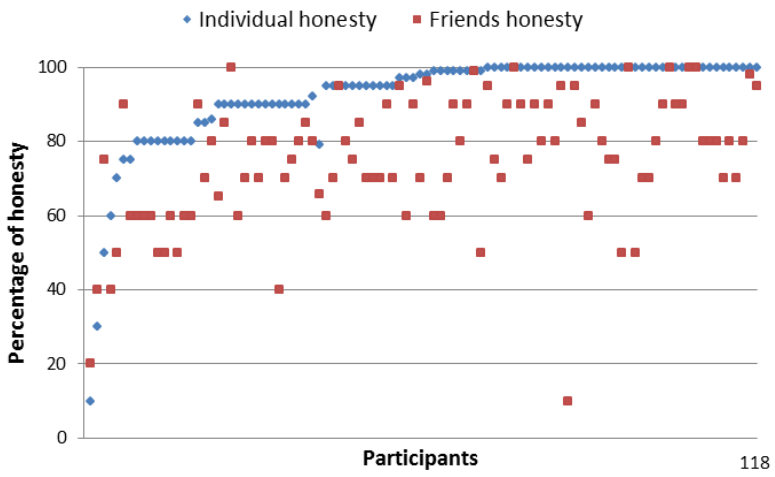


Figure 3.3 Self-rating honesty survey of both individuals and their friends

3.2.3 Description of Data

The collected dataset has a simple structure with three main categories, and each record can be identified with a unique FriendID number acting as a primary key:

- **Public Profiles:** These profiles are accessible to all friends and online visitors and contain some personal information, such as name, age, gender, location, last login date, status, orientation, children, smoke/drink preferences, zodiac sign, education, occupation, income, religion, body type, group membership, school, number of pictures, number of blog entries, number of friends and number of comments. We also generated information about the age of the profile by calculating the "date of profile creation" and "date of crawling". The age of the profile helped us to examine each profile fairly based on how long they have existed.
- **Bands' Profiles:** These are promotional profiles for musicians and bands. The profiles in this category contain a name, location, last login, band website, record label, date since member, and number of views, comments, friends and blog entries.
- **Private Profiles:** These profiles used privacy settings to control their privacy and make their page accessible to their own network of friends only. Accordingly we only had access to their basic information, such as name, age, gender, location and date of last login.

Appendix A describes each identity trait with their data range in more detail. The profile content and friendship information is categorized as follows:

- Personal information, such as age, gender and location.
- Professional information, such as education and occupation.
- Relational information, such as friends and comments details.
- Profile visibility type, such as public, private or bands.
- Activity information, such as last login, age of profile and blog entries.
- Type of profile such as real or fake (for training set only).

In this investigation 2.2 million representative profiles were harvested in a relational database. Of these profiles, 68% contained personal information (public profiles) and the remainder used privacy settings to control the visibility of their profiles (private profiles). Hence, limited personal information was available for private profiles (such as name, age, gender and location): relational

and professional features (such as education, profession and who is a friend of whom) were not visible to our crawler. It should be noted that MySpace was initially intended to be used by those over the age of 18 **[MySpace Wikipedia]**. However the age limit was reduced in 2006 to allow anyone over the age of 14 to use MySpace. Users in the age range of 14 to 15 years old are automatically visible and searchable only within their network of friends, and so profiles in this age range do not appear in this study.

We have not eliminated any information as obscure data may reveal some information about profile's identity. Photos are also not collected due to space and image processing requirements; thus we identified the number of enclosed photos for each profile, which may reveal some pattern about the type of identity. Furthermore, we manually observed a subset of profile photos to see how people present themselves through their images (see Section **5.1.3.2**). This would be an interesting further study to evaluate user identity using image identification.

3.3 Modelling Classifier

At this stage, adequate information was available through crawling and surveys, which gave us grounding on which to propose our classifier model. Our classifier program was developed based on a set of rules, which are generated in a development cycle using both data mining techniques **[Richardson & Domingos, 2002]** and social network analysis **[Agrawal *et al.*, 2003]**.

This section describes our assumption of how self-described identities can construct a classifier to detect the types of identity. In Section **3.3.1** we explain how the idea of personality factors was developed. We then define each personality factor used to classify identity elements through this study (Section **3.3.2**). In Section **3.3.3**, we explain the text mining methods that were applied to examine profile contents in order to extract some personality factors. Network mining approaches, such as centrality and similarity measurements, are also explained in Section **3.3.4**, which helped us to find some other personality factors.

3.3.1 Why Personality Factors

We propose seven opposite personality factors such as expressive/anonymous, valid/fantasy, traceable/untraceable, active/inactive, popular/isolated, positive/offensive and sociable/unsociable. These factors were chosen in parallel through several studies such as a literature study, data mining (such as clustering and classifying), social network analysis (such as centrality analysis), and principal component analysis (such as examining the correlation between each personality factor).

The idea of personality factors was based on a literature study. For instance, according to **[Roccas *et al.*, 2002]** personality can be defined using five big personality factors such as extraversion (e.g., activity and positive emotions), agreeableness (e.g., trust and modesty), conscientiousness (e.g., achievement and deliberation), neuroticism (e.g., anxiety and vulnerability), and openness (e.g., fantasy, feelings and ideas). We selected some of these personality factors, such as fantasy, active and positive (as they are more searchable from profiles), and examined their efficiency on detecting the type of identity through data mining and principal component.

Other work, such as **[Suler, 2002]**, describes how people manage their identity representation online with some factors such as the level of fantasy, reality, positive and negative attributes. Other studies on social network analysis, such as **[Russo & Koesten, 2005]**, also hypothesized that social identification can be represented based on three factors; centrality (being in the group), in-group effect (belonging to the group) and in-group tie (similarity to the group). Therefore, we applied a social network analysis to measure the centrality or '*popularity*' value for each connected profile, and later examined its effect on detecting the type of identity representation. In most cases, such as criminology studies, personalities are used for tailoring and making decisions about the type of people's behaviour **[Thongtae & Srisuk, 2008]**.

Personality factors such as '*expressive*', '*traceable*', '*valid*', and '*sociable*' were selected based on several text mining and data mining techniques. By clustering and classifying our large set of data (2.2 million profiles), we examined every iteration process and improved our classifier. Using data mining techniques (see Section **4.3**) and feeding variable, we decided which identity elements were more likely to define a personality factor. We also examined which personality factors are most important in detecting the type of identity (see Section **4.2**). For instance, through principal component analysis we found that

'*offensive/positive*' use of language has less influence on distinguishing the type of identity compared with other personality factors.

We also applied further studies such as profile customization and photo observation to identify more personality factors, which, due to manual observation, we decided to not include in our classifier (see Section **5.1.3**). Alternatively, other personality factors such as likeable, cooperative, professional, fanatical and so on, can be identified from profile contents, which are interesting for further research.

The following chart (**Figure 3.4**) illustrates our personality model, in which we will examine each personality factor to determine the type of identity representation, such as real or fake.

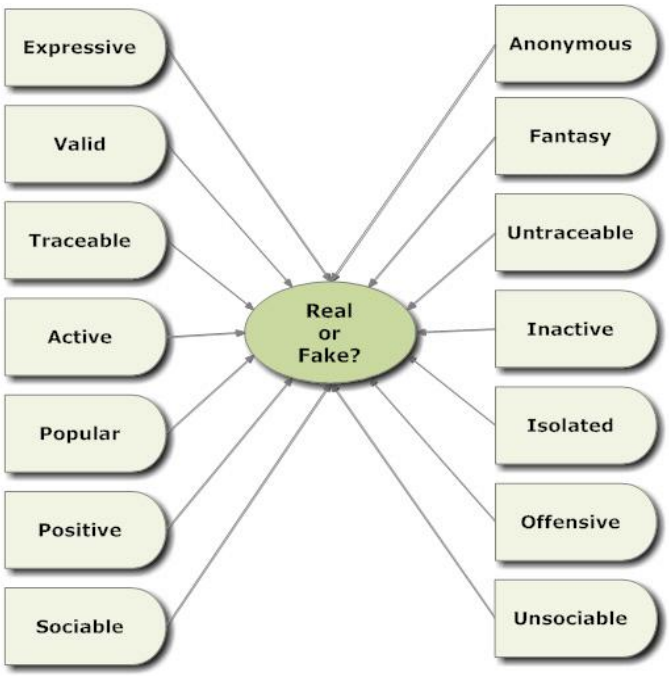


Figure 3.4 Identity model based on personality factors

3.3.2 Personality Factors Definition

This section aims to define each personality factor, although there is no standard definition for these characteristics of identities. We automatically classified each identity element into seven opposite pairs of personality terms. These personalities were chosen after examining and clustering both texts and links within the profile information in a development cycle.

- **Expressive/Anonymous:** Expressive is someone with a more communicative and open personality, who is prepared to share more information about himself/herself with other people in the network. We defined someone as '*expressive*' if he/she responded to the questions listed in the profile. Profiles that minimize their identifiable information are classified as '*anonymous*', which simply refers to the amount of undisclosed identity information on profile.

- **Valid/Fantasy:** We defined a profile as '*valid*' if each identity element (such as age range, existing location, occupation and school information) exists and is reasonably accepted. For instance, we compared each location (such as city, county and country) with a database of valid locations gathered from online to determine the existence of each location. We classified profile information into the '*fantasy*' group if the validity of the entity was not clear, or humorous, or not related to the subject.

- **Traceable/Untraceable:** The digital trace of identity can be found on profiles, including an email address, web link, employment company and school. This information can be used to find further perceptible information about the person. As well as checking the validity of this information, we calculated the number of traceable links from profiles to the outside world. For instance, we defined a profile as '*traceable*' if there was a valid list of schools he/she attended.

- **Active/Inactive:** We defined a profile as '*active*' if someone participated in his/her network activity, such as blog posts or group membership. We also calculated if someone actively logs into his/her profile based on a timestamp for the last login. On the other hand, profiles with less activity on their page were classified as '*inactive*'.

- **Popular/Isolated:** The popularity attribute is supported from other users, such as friends and visitors. We defined a '*popular*' profile as one that has a higher number of friends with more views or hits on the page. The level of popularity examined is based on centrality analysis (average in-degree, out-degree, closeness, and between-ness). This attribute is calculated based on the age of each profile. For instance, we do not expect new members to be as popular as the older members; we can calculate their '*popularity*' by knowing when someone joined the network and averaging against the number of friends. On the other hand, the '*isolated*' attribute determined when the centrality value is low.

- **Sociable/Unsociable:** Sociability is defined based on the level of communication with others in a social network. We defined someone as '*sociable*' according to the total number of comments sent or received on each profile. Also the timestamp and recipient of each comment shows the frequency of comments to each friend.
- **Positive/Offensive:** These attributes are based on the language used in the profile to describe the individual. For instance, we collected a number of databases, such as offensive and positive terms from online, and compared each text within a profile to determine the use of language. The optimistic and encouraging terms, which are used to describe an identity profile, were detected to define the level of the '*positive*' attribute. We also compared the self-described text with a list of offensive (unpleasant or hateful) terms to determine the level '*offensive*' language.
- **Real/Fake:** We also defined '*fake*' as someone who provides a false identity to mislead people into believing that this is a profile of a real person, who has no connection to the profile. A '*real*' profile may present a true identity of a real person.

Applying different methods, such as text and network analysis, led us to implement a personality classifier that examined each identity trait and calculated personality values. We built a weighting schema to rate each individual and their group of friends based on our personality model (see **Figure 3.4**). To rank each individual, we cast all profiles into different categories of positive or negative attributes, and automatically rated each personality attribute. Each attribute was awarded a normalized score between 0-100% based on their appearance.

The classifier effectively detects profile information and decides on the personality factors. However, to find the misclassification error rate, such as false positive and false negative, the accuracy of this model was evaluated through some data mining methods (see Section **4.3.4**). Later in the empirical chapter, we look to see if we can determine the type of identity by learning from known data (with real or fake identity tags) and improving the classification. We will show (in the results Section **5.2.3**) that the reality of the identity can be predicted with higher confidence using the pre-classified personalities rather than using original data.

3.3.3 Text/Content Mining

The key elements for classifying personalities are through profile content and network connections. The text classification provides an insight into the characteristics of each of the entities **[Gill & French, 2007]**. First we examined the content-based identifiers to classify the '*validity*' (truth about information), '*expressive*' (amount of revealed information), '*positive*' or '*offensive*' (use of language), and '*traceability*' (trace of identifying information). Using text-mining techniques we categorized each self-described text into a different group of personalities. For this purpose a classification program was developed to examine the data and classify each attribute within opposite groups of personalities. The features are determined using a mixture of *ad hoc* automated techniques, ranging from checking the validity of the entity to comparing the terms and language used against a list of known terms. In addition, the levels of '*popularity*' and '*sociability*' attributes are checked using a social network analysis (see Section **3.3.4**).

This section briefly describes how attributes, such as 'name' and 'location' were identified based on a profile's contents. The first step towards text mining was to create and adopt several databases, such as a list of valid cities and countries, schools and occupations, and a list of positive/offensive words. These databases were collected from online resources and employed to compare with our dataset by searching for patterns within each text using 'regular expression' methods. Although there is no standard way of finding out if some identity elements (such as age or gender, or indeed if the person really went to that school, etc.) are true, the validity of some attributes (such as city, country, occupation and school) have been checked by comparing to these collected databases.

Each identity trait is mined and classified into opposite pairs of personality factors. The value of each personality is increased or decreased based on the number of appearances of each attribute, and the ratios are measured within a range of 0-100%. This normalization of data into categorized fields is more understandable and faster to calculate by machine. Due to spare usage of terminology, however, mining and comparing a profile's content was not a straightforward process. For example, the pre-defined options in a drop-down menu are easy to categorize, but those entities with open-ended fields and descriptions needed some more attention to examine and classify. Therefore, from an efficiency perspective for our text classification technique, we had to consider an error rate for both input errors and processing errors in the use of language. However, the accuracy of the classification can be improved by further

text mining techniques. The following are some examples of the text classification process to classify 'name' and 'location':

Verifying Name

The username on social networking sites is not necessarily the real name of the user. However, we examined each username associated with a profile, comparing them with several different databases such as offensive terms and the name of known people; these were collected from online sources. We classified four different groups of representation as follows:

- **Related:** the name associated to a person that may be the real name of the person.
- **Fake:** the name appears to be fabricated and may relate to celebrities and known characters (such as 'Homer Simpson', 'Saddam Hussein', 'God', etc.).
- **Fantasy:** the name is not related to a person and it may not have a meaning (such as 'A.P.E.L', 'It's me', 'M@U', etc.).
- **Offensive:** the name contains some offensive language, which is detected when compared to the list of offensive words in a collected database.

Verifying Location

To validate location, the data were first grouped with SQL query language to see which cities or towns were used more within profiles, and then classified as a valid location. For instance, some terms, such as 'l.a.' or 'Lost Angeles' are commonly used by Los Angeles users and classified as valid information. This method saved computation time for later comparisons. Next, we compared the rest of the locations with a list of offensive language to detect any offensive term. Following this, a database of 250,000 valid cities was compared with the rest of the data to detect valid and fantasy locations. For instance, locations such as 'middle of nowhere', 'here', 'near you', etc., were not matched with our database and were classified as fantasy locations.

3.3.4 Network Mining

It is well known that social networks are examples of networks, and the properties of social networking, such as the 'small world', were studied by many sociologists **[Milgram, 1967]**. Observing and analysing the features of social

structure has the advantage of seeing a bigger picture about the structure of a community. According to **[Katona *et al.*, 2009]** a community is shaped by an individual's characteristics; individuals can also be influenced by a community. Therefore, we applied some social network analysis (such as centrality and similarity) to find the position and characteristics of each profile in relation to others in our sample network. We examined the structure of our network sample to score the level of '*popularity*', '*sociability*' and '*similarity*' attributes. These features helped us understand the network property and its relationship to the type of identity representation.

We defined community as a subset of a friend's connection. It can therefore be modelled as a graph *G={I, F},* where '*I*' represents an individual or node and '*F*' represents a friends' link or edge. Within this study, we examined the most obvious type of clique as the group of top 40 friends. This is because, since there is no classification of friends in MySpace (such as best friend, family or stranger), every connection is considered as a friend, while the list of friends may not indicate a strong tie among them. Therefore, when we mention 'friend' we really mean 'links' rather than a strong tie. It would be a better understanding of friendship if social networking sites provided an explicit type of friendship. According to **[Donath & boyd, 2004]** there are different types of connections between people, such as friends, familiar stranger, stranger and community. Some people tend to link to anyone through their friends or friends' of friends. For others who connect to strangers, the motivation could be that they are possibly seeking more links to obtain more popularity and attention: this can become addictive behaviour over time. This means that people are often listed as friends even though they do not particularly know or trust the cyber friend. Weighting the type of friendship is out of the scope of our study, however, it would be an interesting further study to discover the correlation between the strength of friendship and the type of identity representation.

In this section we have examined our sample network structure in term of centrality and similarity between participants and their friend's connectivity. We have identified the network features and metrics by tracing a friend's link to find out the centrality value (the mean of in-degree, out-degree, closeness and between-ness) (Section **3.3.4.1**). We also built a similarity measurement to detect a group of similar identity characteristics (Section **3.3.4.2**). We determined the similarity criteria for both self-described data and extracted personality factors between individuals and their network of friends. The

similarity analysis aimed at determining if a friend's similarities had any influence on deciding whether someone is real or fake.

3.3.4.1 Centrality

The concept of centrality features within networks has been discussed for many years. According to **[Russo & Koesten, 2005]** centrality is "*a measure of potential influence and popularity based on who an actor seeks to interact with within the social network".* There are many approaches to analyse the network and define centrality features, such as in-degree, out-degree, between-ness, and closeness between each node. We analysed our sample network based on an individual's position in the network and visualized a social graph of connection. Within this analysis we aimed at improving the performance of our classifier, and in particular at determining the characteristics such as '*popularity*' and '*sociability*'. Note that our network is directed, which encodes significant information about each node. This means that each tie has direction: for instance if A is a friend of B it does not mean that B is a friend of A.

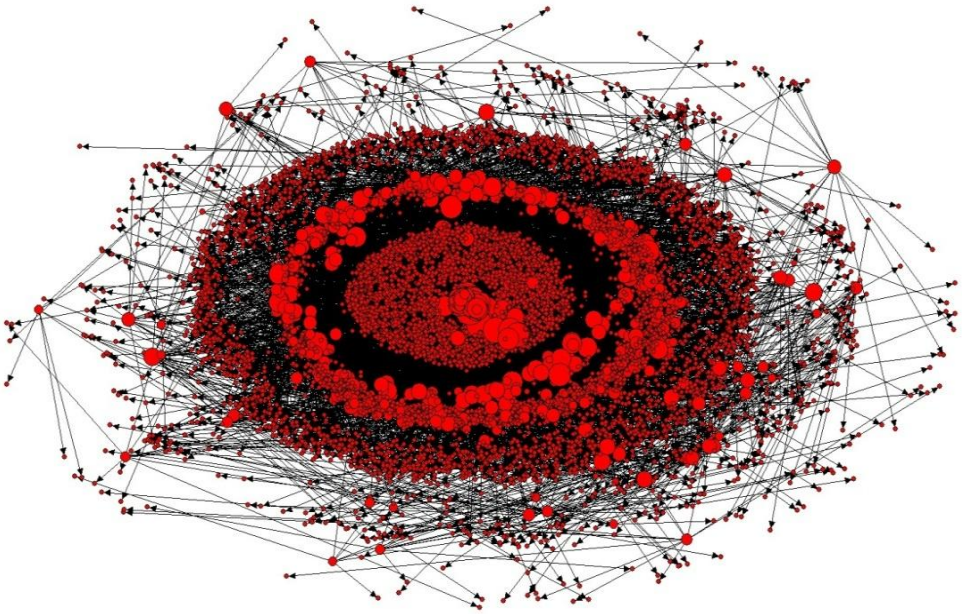
There are many tools for analysis and visualization of social networks, such as UCINET **[Borgatti *et al.*, 1999]** and Pajek **[Batagelj & Mrvar, 1998]**, which are both considered as the main tools to analyse the structure of our sample network. We applied the centrality measurement based on the Linton Freeman method **[Freeman, 1979]** (one of the authors of UCINET), which measures the centrality of nodes based on their degree distribution. We applied the following social network analysis to measure centrality value:

- **Out-degree:** The number of links of each node directed to others (number of people that 'A' has in their friend list).
- **In-degree:** A count of the number of links directed to each node (number of people that have 'A' in their friend list).
- **Between-ness:** Nodes between important groups of connection, which act as a broker (have influence on the information flow).
- **Closeness:** Nodes with shortest paths to other nodes (monitor the information flow in the network).

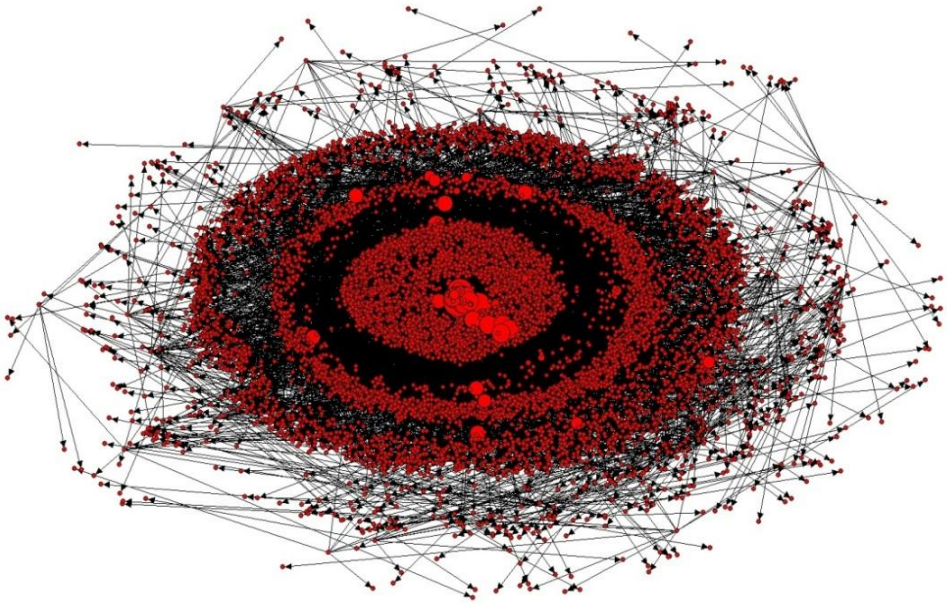
We calculated the centrality feature (average degree distribution, between-ness and closeness) of each profile as a '*popularity*' attribute: the scale ranges from 0-100%. It should be noted that new profiles were measured based on the age of their profile (the date of profile creation against the date of crawling). Therefore,

we did not expect recent profiles to be as popular and sociable as the older profiles.

The following graph (**Figure 3.5 (a)**) demonstrates a sample network of 11,635 connected profiles (with unknown identity). Due to computational complexity, it should be noted that it is difficult to illustrate the entire network with its high diameter. As seen in the graph the majority of nodes with a higher centrality value are located in the centre and the isolated nodes are at the edge. The size of each node indicates the value of centrality (the larger size represents the higher centrality value). The most central node with highest out-degree is called 'Tom', who is the founder of MySpace and automatically becomes a friend to anyone who joins the network. The clustered nodes indicate the strength and power of a specific group of friends. The high degree links imply the importance of the node and represent the core element of the social structure. Our statistical analysis shows that this network employs many high degree connections, which are strongly clustered in contrast to a very low degree connection. **Figure 3.5 (b)** also shows the nodes with the between-ness feature, which have great influence over information flow in the network. The size of each node indicates the level of between-ness value. Often, if these nodes are removed, the network splits into unconnected sub-cliques. For more results on centrality measurement, see the results Section **5.2.1.1**.



(a)



(b)

Figure 3.5 (a) Centrality (out-degree and in-degree distribution **(b)** Between-ness

3.3.4.2 Similarity

In the real world, people tend to connect with those who are demographically and behaviourally similar to them **[Brzozowski *et al.*, 2008]**. Similar attributes and values between people indicate the stronger link between them. This also replicates on online social networking as linkage relies on some similarity between connected people. According to **[McPherson *et al.*, 2001]**, "*homophily is the principle that a contact between similar people occurs at a higher rate than among dissimilar people"*. **[Mesch & Talmud, 2006]** also argue about the quality of a social relationship and indicate that the closeness to a friend is a function of social similarity as well as content, activity and duration of friendship.

The similarity measurement can reveal some information about the context of links between profiles and the correlation between their identity types. After all, 'we are who we are with', and this is a significant evaluation for both offline and online identity. There is no standard method to weight and verify the friends' connection in our sample network. The observation of an individual's choice of friends indicates that the fundamental factor in connecting profiles is similarity in identity (such as age, location, marital status, and education, together with the similarity of interests). We measured the similarity between an individual and their friends' characteristics based on both identity traits (such as age,

location, education, orientation, and so on) and personality (such as expressive, valid, sociable, popular and so on). We are aiming at understanding if similarities between group of friends in both self-described identity and classified personalities can decide on the type of identity.

For this purpose, we generated an identity code for each profile. The identity code is tagged with each profile's information as an identifier. For example, in the identity code 'P_V3MVVA-H11NKNNN33V', each character in the code represents the classification rating. The first character represents the type of profile (such as 'P' for public, 'R' for private and 'B' for bands), the second character represents the name classification (e.g. 'V' for valid name), the third character represents the age range (e.g. '3' for the 30s age range), gender (e.g. M for male), and so on. Generating an identity code speeds up the process of searching for any similarity between groups of friends with certain identity attributes. In addition, a group 'identity code' was obtained by calculating the average friend's personalities within a group, which facilitates the clique similarity in finding the structural equivalence of subgroups.

We applied the following formula to compare 'I', which represents the individual attributes, with 'F', which implies the average friends' attributes of top 40 friends. The average similarity is calculated by dividing the minimum value of each identity and personality elements into the maximum value between individuals and their friends. For instance, if a profile attribute, such as '*traceability*', has a value of 80%, and the average friends' '*traceability*' score is 60%, then the similarity value among that group of friends for being traceable would be 75%. We assumed any value greater than 70% as high with a similar correlation in the friends' network. We excluded profiles with no friends (where *f=0)* or one friend from our similarity analysis. There are some other similarity measures, such as Nearest Neighbours **[Malin, 2005]** and Euclidean based on similarity distance **[Elmore & Richman, 2001]**. However, due to computational efficiency we used a simple algorithm of min/max to compare the similarity value among a group of friends. In addition, it would be interesting to further examine the similarity correlation within friends' of friends (fof) **[Smarr, 2001]**. The abbreviation for each acronym is shown in **Table 3.3.**

$$Similarity(I,F) = \frac{min(I,F)*100}{max(I,F)}$$

$$I = \sum_{a=1}^{n} \mathrm{i}(a) \qquad F = \frac{\sum_{i=1}^{f}(\sum_{a=1}^{n} \mathrm{f}(a))}{f}$$

Table 3.3 Acronyms used within similarity algorithms

Acronym	Description
I	Individual
F	Friend
n	number of attributes
a	attribute
f	number of friends

Figure 3.6 illustrates the correlation between some identity traits and their similarities. The more clustered nodes show a higher correlation, while distributed and scattered nodes indicate less similarity between each pair of identity elements. The results from the similarity analysis show that friends are more similar in identity elements, such as age and location, rather than other identity traits, such as education, marital status, religion and orientation. In particular, participants who are in the same age group are more likely to be similar in other identity elements and the similarity rates are growing in parallel.

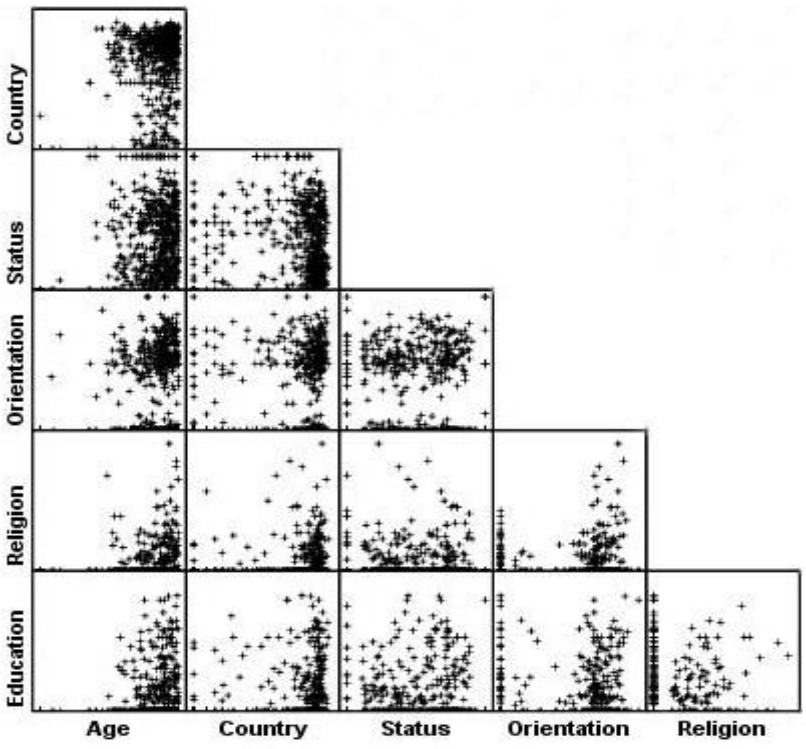


Figure 3.6 The correlation between similar identity elements

We measured what fraction of similarity people may have within their group of friends and learnt that similarity in entities plays a major role in deciding the type of identity (see results in Section **5.2.1.2**).

3.4 Transformation of Identity

This section describes the algorithm we implemented to measure how identity and personality representation transformed over time. To determine the transformation of self-presentation over a period of time, we measured the differences in previous and recent MySpace profile representation. This study will help us understand the evolutionary features of identity representation together with the direction of online social networking. For instance, several studies used confirm the instability of information in profile contents and personality features, with an overall transformation in identity representation of about 29% of '*static*' and 45% of '*dynamic*' information. We will see that a profile's content becomes less expressive, valid and traceable information, while other characteristics, such as popularity, sociability and activity increase over the period of a year (see Figure **5.18**). Also, measuring the identity transformation over our training dataset indicates that real profiles transformed their identity information less compared with the fake group (see Figure **5.16**).

Using our customized crawler we collected the same set of profiles, including their friends' information, during both 2007 and 2008. We describe the data collection process and the differences in data volume for the same profiles in Section **3.4.1**. We employed our classification model to determine a profile's characteristics for each dataset. To measure the evolution of online representation, we implemented a transformation algorithm by ranking both '*static*' and '*dynamic*' features of identity elements over time (see Section **3.4.2**).

3.4.1 Data Collection over Time

We employed a quantitative study for mass downloading of MySpace profile content within two different periods of time (the years 2007 and 2008) (see Section **3.2.1**). Our customized crawler accumulates information, such as personal and professional information as well as relational connections between groups of friends. The information was collected first in December 2007, and the same sets of seeds were used to crawl the same profiles again in December 2008.

Table 3.4 describes the number of participants within the different categories of '*public*' (personal pages), '*private*' (limited pages), and '*bands*' (with music bands related data) profiles. The difference in the number of profiles over time shows that the fraction of '*public*' and *'bands*' profiles has decreased over time by 10.8%

and 1.8% respectively, while the number of '*private*' profiles has increased. The differences in our dataset are small, but on a larger scale, this may indicate that users are more willing to change their profile setting from '*public*' to '*private*'. The difference in the number of our sample dataset shows that '*public*' profiles are six times more likely to leave the site than '*bands*', and 63 times more likely than '*private*' profiles. Based on the age of the profile (the date when a user becomes a member), we also found that older profiles are more likely to change to '*private'* settings.

Table 3.4 The number of collected profiles in both year 2007 and 2008

Type of Profile	Year 2007	Year 2008	Difference %
Public	1,215,001	1,083,623	-10.81
Private	793,995	795,319	+0.17
Bands	201,297	197,670	-1.80
Total	2,210,293	2,076,612	

3.4.2 Evolutionary Analysis

The same profile can represent the same person without having identical contents at different times. According to Leibniz's law **[Stevenson, 1972]**, if 'p' is identical to 'q', then every quality of 'p' will become the quality of 'q'. So if profile 'p' transforms to profile 'q' then some property in 'p' and 'q' is true at time 't1' and not true at time 't2'. As we do not know the truth about both identities in 't1' and 't2', we aimed to measure the profile's attributes for both 'p' and 'q'. We then compared both datasets to see how much 'p' transforms to 'q' in terms of identity representation over a period of time.

We first applied our classification program over both year 2007 and 2008 collected profile data and measured the personality factors for both datasets. To examine the transformation of identity, we compared both raw identity data (such as age, gender, location, etc.) and pre-classified personality attributes (such as expressive, valid, traceable, etc.). We then identified data as being either '*static*' or '*dynamic*' information. Static refers to the information that is less likely to change over time (such as age with constant change, gender, orientation, ethnicity and zodiac), while a '*dynamic*' attribute (such as location,

marital status, number of friends and interactions) is more changeable over time.

The formula below is proposed to compute the transformation of identity based on both static and dynamic information. We examined the transformation of identity for both individuals and their networks of friends using original and pre-classified data. If the same persona 'p' changes to persona 'q' over time 't' then the difference in representation can be calculated by dividing the minimum value to the maximum value between 'p' and 'q', whereas 'p' is the summation of '*static*' information for both original and pre-classified dataset, accumulated to the summation of '*dynamic*' attributes. We assumed that dynamic attributes have less impact on the overall transformation value due to the variable nature of these attributes (see **Table 3.5** for acronyms).

$$Transformation(p \rightarrow q) \quad t1 < t < t2$$

$$p = \sum_{a=1}^{n} s(a) + \frac{\sum_{a=1}^{n} d(a)}{2}$$

$$q = p \pm T$$

$$T(p,q) = 1 - \frac{min(p,q)}{max(p,q)}$$

Table 3.5 Acronyms used within transformation algorithms

Acronym	Description	Acronym	Description
T	transformation	n	number of attributes
p	persona (old)	t	time
q	persona (new)	s	static
a	attribute	d	dynamic

We illustrate the findings in the results chapter, Section **5.3**. We will discuss the results of how both individual and social networks evolve during a time frame and what impact it may have in determining the taxonomy of identity representation.

3.5 Summary

In this chapter we described our methodology for collecting data and implementing a customized classifier to detect and categorise profiles and their connected friends based on their self-representation. We employed two methods to accumulate data from MySpace profiles. We first downloaded profiles via automated crawling, which has the advantage of collecting a greater number of participants. We started by collecting a group of friends, rather than random profiles, to have a minimum cut in our sample network. Our crawler identified three types of profile; '*public*', '*private*' and '*bands*'. We were able to harvest approximately 2.2 million profiles over the course of two months. We then collected a number of profiles with known identities, such as official profiles and impersonators. We also conducted an email survey by asking participants to verify their online identity and rate their level of honesty.

We implemented a personality classifier to automatically categorise profiles according to our seven personality factors (expressive, valid, traceable, active, popular, sociable and positive). By creating a breakdown of profile information into these characteristics, we generated a database of profile attributes, where this may correspond to an actual image of user. A ranking score was integrated to associate with each profile's attributes and their network of friends. Our classifier aims to identify the profile's characteristics and examine the correlation between what is real and what is fake identity.

The classifier was generated in two steps. We first defined a classifier using a text mining approach and analysed the profile's content. Through text mining analysis we extracted the terms and language used against known terms. The terms were rated within different classifications of '*valid*', '*fantasy*', '*expressive*', '*offensive*' or '*positive*'. However, examining the identity content and the use of language was not a straightforward task, as the terminologies used are diverse when an open question is asked, and people tend to use different expressions to describe themselves.

Next, we examined network content by applying some social network analysis, such as centrality and similarity measurement. We explored community structure using social network techniques and portrayed our sample network and their relationship as a graph. The degree of distribution showed the popularity and power for central nodes as the majority of nodes had a significant high degree of connection forming a core element of a social structure. We

calculated the '*popularity*' attribute for each profile based on the centrality feature.

Through similarity analysis we also found that mutual friendships are a significant input in the determination of the type of identity representation. This is because people normally choose friends with similar identity traits and interests. Therefore, we investigated similarity measurement by clustering profiles with certain identifiers using our generated '*identity code*'. The identity code works as a signature for each profile compared with the group of friends' identity code. We learned that social network analysis reveals some property of the network, which can be used to find the property of a network of friends more effectively. Centrality and similarity have definite effects on determining the type of persona. We then employed our classifier for the same set of profiles within different time frames. By assigning data into '*static*' and '*dynamic*' categories, we measured the same profile in both 2007 and 2008 (see result in Section **5.3**).

The next chapter will describe the empirical methods used by examining the importance of each personality factor using principal component analysis. We will evaluate the prediction accuracy and performance using data mining algorithms. Through both data mining and principal component methods, we observed some patterns in data and improved our classifier in parallel.

Chapter 4

Empirical Techniques

"The truth is sustainable! A lie is unsustainable, and is eventually exposed through Time. A lie can't 'Face' the truth because of guilt and the liars will try to detach themselves through avoidance."

Carl Stoynoff

4.1 Introduction

In the previous chapter we used data accumulated from MySpace profiles to develop a classifier that categorizes profiles based on their personality attributes. We rated profiles based on how '*expressive*', '*valid*', '*traceable*', '*popular*', '*sociable*' and '*positive*' the individuals and their friends are. These personality traits were based on several text mining and network analyses. We used both known and unknown datasets to uncover patterns in the data in a development cycle and improved the classifier in parallel. We also examined the same set of profiles at a different period of time; this allowed us to measure how identity representation evolves through time focusing on the extracted personality factors.

Within this chapter we aim to evaluate the accuracy of our developed classifier to predict and verify represented identity. We first examined the correlation between each personality factor, so principal component analysis was applied to uncover which entity and personality elements are more significant in determining the validity of identity. We observed each personality factor to identify their importance by extracting the main component in our dataset. This analysis will improve the prediction accuracy further when we apply further data mining techniques.

Next we applied several machine learning techniques to automatically uncover significant patterns in our dataset. We used these patterns to build our classifier heuristically in a cycle by learning from the training set (known identity) and

applying to the validation set (unknown identity). Both supervised and unsupervised learning were employed to cluster and classify the training and validation datasets. Using different learning methods, we aimed to evaluate the performance of our classifier and validate the different types of identity.

This chapter is structured as follows: Section **4.2** explains the procedure for principal component analysis, such as component and rotation analysis, measuring the correlation between each personality and the influence on verifying identity. Section **4.3** describes the employed data mining techniques, such as supervised and unsupervised learning, where we take advantage of existing machine learning techniques to identify significant patterns in data. By analysing different algorithms, we are able to explain the prediction accuracy through a confusion matrix table.

4.2 Principal Component Analysis

Principal Component Analysis (PCA) is a common technique used to find a pattern in data with a high dimension **[Smith, 2002]** and **[Qu *et al.*, 2002]**. We took advantage of principal component features using the SPSS application by simplifying the personal attributes and generating a set of component variables **[Jae-On & Mueller, 1978]**. Each principal component is a linear combination of the variables with a minimal loss of information. We used this method to discover a simple model of relationships within variables and explain them using a smaller number of components. Accordingly, we used both raw data (original identity representation) and pre-classified data (personality factors) for both training and validation sets as an input for our principal component analysis. This analysis indicates which factors or components are more significant when examining online identities. Some identifiers are better indicators for deciding the type of identity while some are strongly associated with each other. Within this analysis we aim to answer the following questions:

- How many components are required to explain the patterns and relationships among variables?
- What is the correlation between each identity attribute? What are the characteristics of these components and how do they explain the observed data?
- In terms of the size and structure of our dataset, how efficient is it to use PCA? How much information is lost by using this analysis?

- How can PCA improve the prediction accuracy of our classifier model?

In Section **4.2.1** we first explain our primary analysis, which was used to discover the correlation between data and exclude any entities or personalities that are not correlated with others. We extract the main component and measure the amount of information that we lost as a result. We then discuss rotation component analysis in Section **4.2.2**, where we identify the common theme found in each component. We analyse each personality and identity attribute according to the dimension of extracted principal components.

4.2.1 Component Analysis

We began our primary analysis by looking at the inner correlation matrix between each identity attribute. We used the SPSS program to exclude any attribute that did not correlate with others or those that were highly correlated with each other **[Liu *et al.*, 2003]**. These distinct correlations may influence the determination of components in our analysis. According to **[Smith, 2002]**, PCA requires correlations between variables to be greater than 0.3: our correlation coefficients show that most identity elements are correlated reasonably well with each other (greater than 0.5).

We then examined if component analysis is a sufficient study using Kaiser's measure (KMO) **[Kaiser, 1960]**. KMO measures the adequacy of component analysis where its value varies between 0 and 1, and any value greater than 0.5 is adequate. The overall KMO for our set of variables within different types of identity ranges from 0.63 to 0.68. This range exceeds the minimum requirement and generally indicates that a component analysis may be useful within our dataset.

The eigenvalue (covariance value of two dimensional data) of each component is plotted using a Scree plot (**Figure 4.1**), which demonstrates the curve of seven personality factors. For the initial solution, there are as many components as input attributes, where those with a high eigenvalue are more significant and were retained in the analysis. The components on the shallow slope have a lower contribution to the solution. Therefore, the first three components on the steep slope, which have an eigenvalue greater than '1', are extracted as an optimal number of components. We exclude the tail and lose 17.32% of data in order to classify and identify our personality factors within three main components (see **Table 4.1** for more detail).

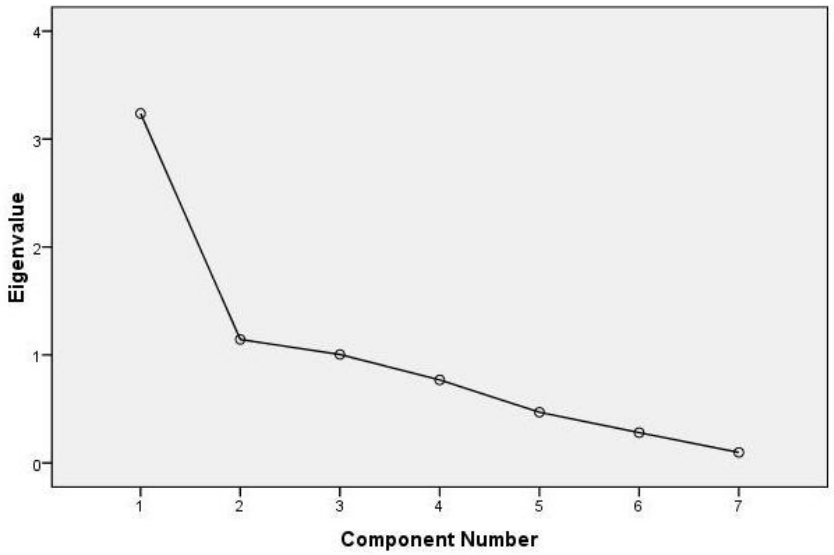


Figure 4.1 Scree plot to extract the main components

Table 4.1 Primary principal component analysis (KMO, loss of information and main components)

(Average accuracy: 82.68%)						
Component	Eigenvalues (**Real**)			Eigenvalues (**Fake**)		
	KMO:0.68 **Loss:** 15.09			**KMO:** 0.63 **Loss:** 19.56		
	Total	Variance %	Cumulative %	Total	Variance %	Cumulative %
1	3.24	46.24	46.24	1.96	31.04	31.04
2	1.14	20.33	66.57	1.366	26.51	57.55
3	1.00	18.34	84.91	1.11	22.89	80.44

Table 4.1 describes the eigenvalue associated with each extracted component for our dataset. The 'total' column indicates the eigenvalue and the amount of variance in the original variables for each component. The 'variance' column shows the ratio (as a percentage) for each component to the total variance of the variables. The 'cumulative' column shows the percentage of variance for the sum of first three components. The cumulative variability explained by these three components ranges from 31.04% to 84.91%. This suggests that there is redundant information across each identity type within the '*real*' and '*fake*' categories. For instance the first three components for '*real*' data demonstrate 84.91% of the total variance. Therefore, we have a 15.09% loss of information by reducing the number of identity variables into these three factors. There is

slightly lower loss of information for '*real*' data compared with the '*fake*' group, which means the possibility of determining someone as '*real*' is higher than '*fake*' using these main components. Within this component analysis we achieved 82.68% accuracy on average by defining our personality factors within three main components.

4.2.2 Rotational Component

To make various decisions on the importance of selected identity variables, it is important to identify common themes in each component. To do this, we utilized the rotational component that was used to find the correlation between each identity attribute in relation to the main three extracted components.

The results from the selected components' score over our training dataset are represented as a rotated matrix in **Table 4.2**. The differences between components within the '*real*' and '*fake*' identities indicate which attributes are more likely to determine the type of represented identity. For instance, the value for '*popularity*' and '*traceability*' for the '*real*' group is higher compared with the '*fake*' group. The table shows that attributes such as '*positive*' and '*active*' have a lower rate for the corresponding component and are not as significant as other personality types. This rotational analysis shows the order and importance of each personality within both the '*real*' and '*fake*' groups, which may help to distinguish between both types of profile. It also improved the data structure by reducing unnecessary data dimensions. Further prediction results from our principal component analysis can be found in the results (Section **5.2.2**).

Table 4.2 Extracted rotation components for each attribute using training dataset

Rotated Component Matrix						
Attributes	**Real**			**Fake**		
	Components			Components		
	1	**2**	**3**	**1**	**2**	**3**
Expressive	.825				.746	
Valid		.451				.504
Active			.097		.182	
Positive		.027				.067
Popular	.733			.674		
Sociable		.944			.790	
Traceable	.872					.776

4.3 Data Mining

In recent years, data mining techniques have attracted a great deal of research attention. According to **[Hand, 1998]**, data mining can be described as "*The process of secondary analysis of large databases aimed at finding unsuspected relationships that are of interest or value to the database owners*". Data mining played an important role in improving our classification model through the use of both supervised and unsupervised learning methods. These learning methods involve applying a model to determine some knowledge from data and evaluate our classifier in parallel. Some data mining techniques are more reliable than others in providing more accurate prediction results. Therefore, we used different algorithms to evaluate the performance of our classifier and analysed the efficiency of each learner.

We took advantage of an existing data mining tool called Rapidminer (formerly known as Yale) **[Jungermann, 2009]**. Rapidminer is a graphical user interface environment for machine learning with many schemes for classification, including Support Vector Machines (SVM), Decision Trees, Bayesian, Association Rules and clustering. This application uses XML as a standard format to describe structured data and model a Knowledge Discovery (KD) process. In addition, the attribute evaluations and clustering schemes from the Weka library (another machine learning application) were integrated **[Holmes *et al.*, 1994]**.

Within this section we explain the four main steps involved with data mining approaches. Each process is explained and demonstrated without going into any algorithmic details. The results can be found in Section **5.2.3**. The four key steps are as follows:

- **Data pre-processing:** such as data formatting and outlier detection.
- **Supervised learning (classification):** such as Naïve Bayes, Lazy Learner (Nearest Neighbours), rules learning and tree learning (Decision Tree).
- **Unsupervised learning (clustering):** such as clustering (Agglomerative, K-means), similarity comparator and Association Rule generator.
- **Performance validation:** such as X-validation using confusion matrix.

4.3.1 Data Pre-processing

Data pre-processing is the procedure of preparing and cleaning data for further analysis. The pre-processing operations are necessary since particular learning schemes may not handle attributes of certain value types. After dividing

our data into different groups (such as training, test and validation datasets), we applied some data cleaning and pre-processing operations, such as outlier detection to filter any noise in data. These methods helped us to isolate some obscure features for further analysis.

4.3.1.1 Data Formatting

Data mining techniques were applied to both original data (such as age, gender, location, etc.) and pre-classified data (such as expressive, valid, etc.). Our dataset with known identity attributes, such as '*real-celebrity*', '*real-local*', '*fake-celebrity*' or '*fake-invented*' are labelled within our training set. We used both the 'one-third' technique (two-thirds for training and one-third for test set) for selecting the number of training and test datasets **[Klösgen & Zytkow, 2002]**. Therefore, from our 1300 known profiles, two-thirds are held back for the training set and the remaining one-third are allocated for the test set. The unknown data are used as the evaluation set to find some grouping attributes when unsupervised learning is applied. According to **[Guillaumin *et al.*, 2009]**, the larger the training set, the better the performance. However, we used only one set of data as rapidly changing data can make discovering patterns confusing. The selected datasets can be described as follows:

- **Training dataset:** The training set is used for building a model. Our training set contained 867 records where the initial value of our class variable '*identity*' was set to '*real-celebrity*', '*real-local*', '*fake-celebrity*' and '*fake-invented*'.
- **Test dataset:** The test set is used to measure the performance of the model. Our test set contained 433 records, where the class or '*identity*' value is ignored. However, the class value will later be exposed in the evaluation process to measure the performance accuracy.
- **Validation dataset:** The validation set is used for tuning the model, such as clustering. This dataset contains the entire unknown profiles of approximately 2.2 million records. So, we do not know the correct value for the class of '*identity*' in this dataset.

So that the data could be read by Rapidminer, all the selected datasets were converted into the XML integrated application. Additionally, we had the choice of using other data formats, such as .csv (comma separated values) and .xls (Microsoft Excel), but it is faster to use an XML Meta data format, such as:

- .aml – (attribute description file) this file is a simple XML document defining the properties of the attributes, such as name, range and the data source files.
- .dat – (dense file) understandable data format, which lists every record of our dataset.
- .arff – (attribute relation file) the Weka application uses this format for further techniques.

4.3.1.2 Outlier Detection

The outlier detection process identifies outliers in the given dataset based on the distance of points to their nearest neighbours **[Malin, 2005]**. This technique helps us to find individuals with the largest distance to neighbours according to their personality attributes. For instance, profiles that represented their identity very differently from others are most likely to be detected as outliers. This process takes an example set and passes each record with a Boolean status indicating true (outlier) or false (not outlier) (see **Figure 4.2**). The detected outliers were removed with the filter operator. This process eliminates invalid and noisy data and increases the processing speed.

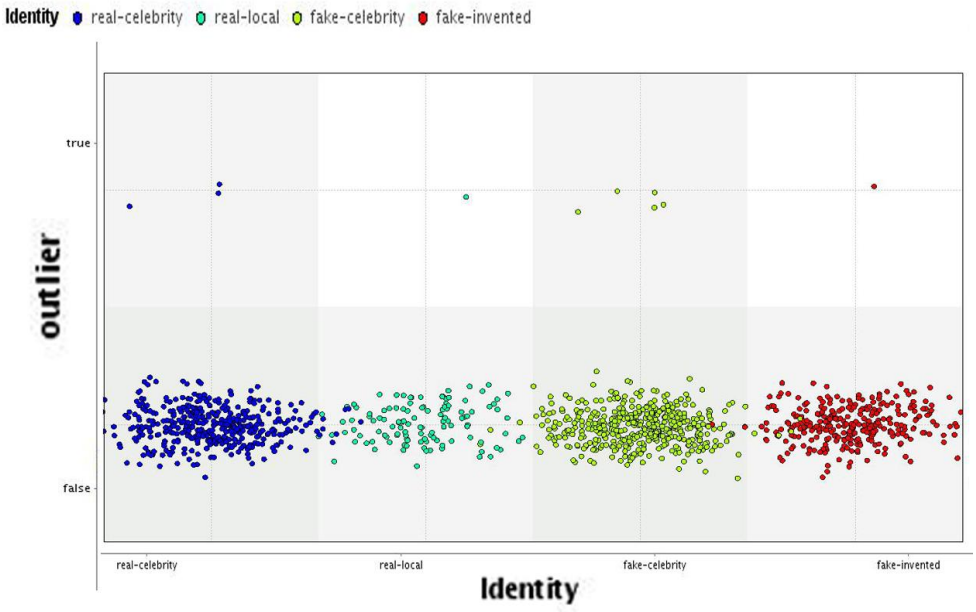


Figure 4.2 A scatter plot shows outliers based on the type of identity

4.3.2 Supervised Learning (Classification)

In order to classify our training data we mainly focused on supervised learning, also known as classification **[Chakrabarti, 2000]**. Unlike unsupervised techniques, supervised learning uses labelled data, so the class attribute is the target. This classification method is significant; by learning from the discovered knowledge our classifier was able to significantly improve its prediction accuracy. In this learning process our training set was associated with a label or class '*identity*' that determines the type of identity, such as '*real-celebrity*', '*real-local*', '*fake-celebrity*' or '*fake-invented*'. We applied both original and pre-classified data from our training set and trained multiple classifiers, such as the Decision Tree, Naïve Bayes, and Lazy Learner and later compared their performance (see results chapter, **Table 5.2** and **Table 5.3)**.

4.3.2.1 Decision Tree

Decision Trees are powerful classification techniques, which can be easily understood. They are primarily used for analytical modelling and classification, which visually brings up the hidden patterns on data. Using this method, we built a hierarchical classifier of our training dataset and classified data into different attributes. As featured in **Figure 4.3**, the Decision Tree presents rules that were learned over classification. Each branch on the tree is a classification question and the leaves reflect the probability of a specific identity type classification. It is important to decide which part of the rules corresponds to data and whether the data are reliable. Failing to ensure correctness of rules can result in an inability to classify data accurately. As seen in the graph '*sociable*', '*popular*' and '*traceable*' attributes are the main characteristics to distinguish different types of profile and '*real-celebrity*' is the most identifiable group, while attributes such as '*positive*' and '*active*' do not have much influence when using Decision Tree learner.

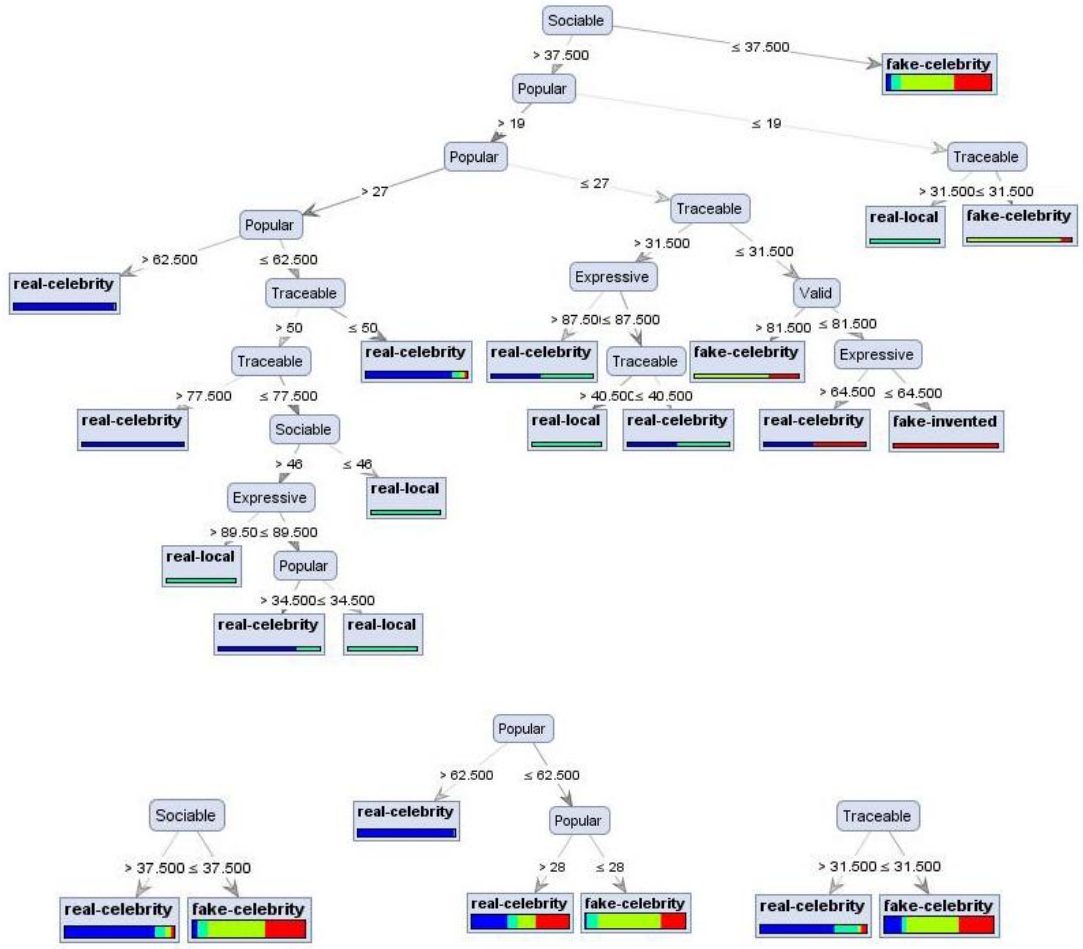


Figure 4.3 Decision Tree learner

4.3.2.2 Naïve Bayes

Additional supervised models have also been studied. These include the Naïve Bayes classifier, which weights each attribute for classification based on Bayes theorem **[Keogh & Pazzani, 1999]**. This learner computes the possibility of each feature in determining the class variable. The Naïve Bayes classifier was trained for each of our personality factors. The advantage of this classifier is that it generally performs well even with a small training set, but if there are many properties to check, the number of observations increases in order to estimate the probability. A Naïve Bayes classifier can be comparable to a Decision Tree, but it has lower accuracy and works faster with large datasets.

4.3.2.3 Nearest Neighbours

The Nearest Neighbours prediction is one of the oldest techniques used in data mining, and classifies attributes based on their close neighbours **[Klösgen & Zytkow, 2002]**. We used the Lazy Learner as a simple Nearest Neighbours classifier based on an explicit similarity measure. This technique uses similar close attributes in order to determine the classification of data. The distance between neighbours shows their similarity in terms of identity representation. **Figure 4.4** shows our training profiles based on their similarity distance. Clustered nodes in the centre represent the '*real*' profiles and the more distributed nodes are the '*fake*' group. Therefore, there is more similarity in attributes within '*real*' profiles in comparison to '*fake*' profiles. For instance, *'real-celebrity*' profiles have less distance between them as they have more in common, such as having greater number of friends (popularity), and have more valid and expressive information on their page; the profile attributes and identity traits for '*fake-celebrity*' are not similar to each other.

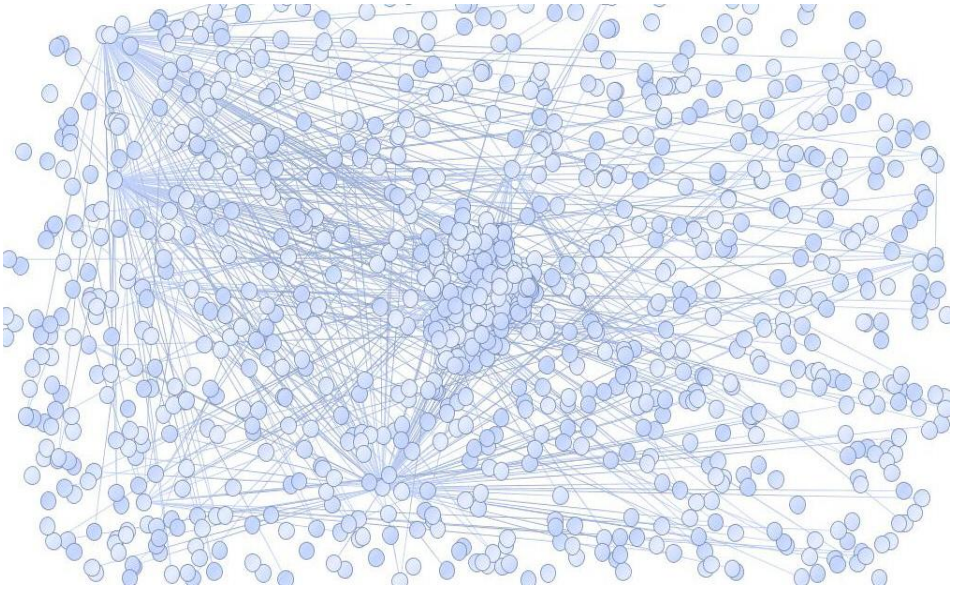


Figure 4.4 Nearest Neighbours similarity-based classification

4.3.3 Unsupervised Learning (Clustering)

Unsupervised learning methods are designed to observe hidden patterns in data **[Malin, 2005]**. The clustering method is a fundamental technique for the visualisation of classified data, which groups profiles with similar identity attributes. In general, the clustering process is faster to generate than classification methods, however, it takes more time to decide on the results. We

used the validation dataset with an unknown class attribute of '*identity*' to cluster similar personalities and entities, examining both original and pre-classified data. Some techniques such as agglomerative similarity and association rules were applied to our dataset. Some of the methods that were used to classify identity properties into different clusters are described below.

4.3.3.1 Agglomerative Clustering

This clustering technique performs generic agglomerative clustering based on a set of attributes and their similarities. For instance, the similarity between node A and node B is weighted according to the value of each personality factor from our rating model. Our validation dataset is clustered to quantify the neighbourhood connection and their similarities. This clustering of similar attributes categorised different identity representations within different groups. For instance, profiles with higher '*popularity*' values are clustered closer to one another that indicates that they have other shared values as well as popularity, such as their age range; scattered nodes are mainly similar in terms of orientation, income or gender.

4.3.3.2 Association Rules Generator

Rule generation is one of the major factors in data mining for knowledge discovery in unsupervised learning. The rules are in the form of 'if this … then that…' and can be used for understanding the relationship between entities based on their attributes. We have generated a set of rules by applying the association rules algorithm over our validation dataset. We then searched for significant rules in accordance with each personality factor. **Appendix C** lists a sample set of rules generated from '*public*' profiles. Each rule is assigned based on their frequency of appearance. These rules helped us to decide on the importance and frequency of each entity in relation to each personality attribute.

4.3.4 Performance Validation

In many cases the learned model is not of particular interest, but the accuracy of the model achieved in the evaluation process is of most importance. There are several validation methods that can be used to estimate the accuracy of learning models and measure the prediction performance, such as simple

validation, regression performance and T-test. We applied a validation operator called X-validation **[Jungermann, 2009]** and a confusion matrix **[Klösgen & Zytkow, 2002]**. Our aim is to achieve higher accuracy when using personality factors as an input in validation process. The result shows that we achieved 64% accuracy when using original data (such as age, gender, location, etc.) compared with 83% when using personality factors (see Section **5.2.3**). Within this section, we explain the predictive model applied along with the table of the confusion matrix.

The X-validation operator evaluates the learning method from the previous model applier. The model applier predicts labels for the test dataset and the performance evaluator compares them to the known labels. Over the iteration process, the cross-validation returns the average absolute and squared errors. **Appendix D** demonstrates the process of the X-validation operation, providing an XML sample description. For example, as seen in the graph, first the training dataset is used as an input for the learner to convey a model. Then our generated Decision Tree learner is used to generate a model and this is applied to the model applier. Next, the test dataset is loaded to predict the class of the '*identity*' value. The X-validation operator is then applied to deliver the performance of possible predictions for our unknown test set by one-third validation. Finally, the performance accuracy and classification error produces the results as a confusion matrix.

The confusion matrix is a well-known evaluation technique that factors a matrix of true-positive (TP), true-negative (TN), false-positive (FP) and false-negative (FN), and presents the performance based on precision and recall measurements. For example, if our model applier predicts an attribute as true-positive and our model suggests this as false-positive then the error rate will increase.

- **TP** is a correct classification of correct data
    - e.g. '*real*' correctly tagged as '*real*'
- **TN** is a correct classification of incorrect data
    - e.g. '*fake*' correctly tagged as '*fake*'
- **FP** is an incorrect classification of incorrect data
    - e.g. '*fake*' incorrectly tagged as '*real*'
- **FN** is an incorrect classification of correct data
    - e.g. '*real*' incorrectly tagged as '*fake*'

The metrics for performance evaluation are described in **Table 4.3**, where the overall accuracy can be calculated as:

$$\text{Accuracy} = \frac{a+d}{a+b+c+d} \quad \text{Precision} = \frac{a}{a+b} \quad \text{Recall} = \frac{a}{a+c}$$

- **Accuracy:** proportion of the total number of predictions that are correct.
- **Precision:** proportion of the predicted positive cases that are correct.
- **Recall:** proportion of positive cases that are correctly identified.

Table 4.3 The table of Confusion Matrix

	Actual (real)	**Actual (fake)**
Predicted (real)	TP (a)	FP (b)
Predicted (fake)	FN (c)	TN (d)

To examine the accuracy of our classifier model we computed the correlation between the predicted data and the actual data. To find the error rate, we were interested in the probability of when someone falls into an incorrect identity type. For instance, 'false-positive' where the '*real*' profiles are classified as '*fake*' and the '*fake*' profiles are tagged as '*real*'. **Table 4.4** demonstrates the performance of pre-classified data using the nearest neighbour learner. The precision ranges from 61% to 91% with the best performance for the '*real-celebrity*' group, where the precision for '*fake-invented*' is remarkably low. The highlighted values (diagonal line) represent the 'true-positive' and 'true-negative' predictions, with an average prediction accuracy of about 82%.

Table 4.4 The confusion matrix of Nearest Neighbours learner

Accuracy: 82.43%		**Actual Identity**				
		real-celebrity	**real-local**	**fake-celebrity**	**fake-invented**	**class precision**
Predicted	**real-celebrity**	358	16	12	6	91.33%
	real-local	2	42	0	0	95.45%
	fake-celebrity	25	47	384	132	65.31%
	fake-invented	32	13	61	172	61.87%
	class recall	85.85%	35.59%	84.03%	55.48%	

By applying different learning methods and adjusting our personality classifier, we expected to achieve less false detection in classification prediction. Applying different methods has the advantage of comparing and selecting the more efficient learner. As a result, we compared the accuracy obtained from different data mining learners using both original and pre-classified data. In order to achieve greater accuracy it is good practice to apply different performance operators, although observing many different results may be confusing. We will show in the results chapter, Section **5.2.3**, that the prediction rate using pre-classified data performs more accurately than using the original data. In addition, **Appendix E** describes the confusion matrix performance across different machine learning techniques for both inputs (original and pre-classified data) in more detail.

4.4 Summary

This chapter reflects on the validity and reliability of our classification model. We discussed the overall assumption about the techniques we used to analyse and verify data using both principal component analysis and data mining methods. Following the implementation of our personality classifier, we first applied some principal component techniques to extract the main components and measure the correlation of each component with the type of identity representation. Within this process we determined which identity elements are more important to identify the type of identity. Applying component analysis to our dataset confirmed that the advocated seven personality factors are a good indication that can be used further within our study. However, analysing the use of language, such as '*positive/offensive*', did not bring new insight to identifying the type of profile's identity (see **Table 4.2**). The computation time for running the component analysis sped up by removing the less significant dimensions in the data. We achieved a higher prediction performance based on fewer dimensions of data.

Next, we applied some data mining methods to extract significant patterns within identity information. Data mining refers to a set of techniques that uncover hidden patterns in data. Therefore, we took advantage of these techniques as it is highly inefficient to analyse a large set of data manually. The data mining procedure first set out to understand the dataset more clearly by using some pre-processing approaches (such as outlier detection). We then applied a number of supervised learning methods (classification techniques) to

help examine the training set and predict the test set class target variable. We also applied unsupervised learning (clustering techniques) to identify the natural grouping of similar attributes. The collected data from the crawler (with unknown identity) was used for unsupervised learning to observe the patterns within the data, while the data from the survey (with known identity) were used in supervised learning. Within the supervised learning process, we applied several learning approaches, such as the Decision Tree and Association Rules techniques. This model was then applied to the test dataset for prediction of the label or the class '*identity*'. We then measured the re-substitution error on the training set (known data) to examine the incorrect performance in labelled data. Finally, the performance accuracy of the learner was described as a confusion matrix, which is an evaluation technique used to factor a matrix of true-positive, true-negative, false-positive and false-negative.

Data mining techniques caused a considerable improvement on the overall performance of our classifier model. The next chapter describes and illustrates the result from comparing the different learning methods. The performance results show that the proposed model cannot be used as an absolute prediction of identity, although we found 83% accuracy on deciding the type of identity.

Chapter 5

Results and Findings

"We will act consistently with our view of who we truly are, whether that view is accurate or not."

Tony Robbins

In the previous chapter we employed different methods to evaluate and measure the influence of each personality factor against the type of identity. We first examined the correlation between each identity trait using principal component analysis and focused on the main components to reduce the possible dimensions within data. We then used different data mining techniques to evaluate the prediction performance of our personality model. We found patterns within the data by training profiles with a known identity (such as '*real*' or '*fake*') and predicted the identity type for unknown profiles. We improved our classifier in parallel by finding patterns in data.

Within this chapter we present the results that correspond to the methods used. The chapter is organized as follows: Section **5.1** demonstrates some statistical results and explains the properties of our dataset. Through these results, we learnt the correlation between each personality factor in deciding the type of identity within our model. We present the statistical relationship between each entity, including the initial analysis, further personality factors and highlights of extracted patterns in data. Exploratory results are illustrated in Section **5.2**, including the results from our social network analysis, principal component analysis and data mining algorithms. Section **5.3** explains the evolutionary results, such as the transformation of a profile's identity and the evolutionary features of the social network. This chapter concludes with a summary of our findings in Section **5.4**.

5.1 Statistical Results

Within this section we aim to provide some demographic and statistical results from our data analysis that provides a clear picture about our dataset and the identified patterns. Through this analysis, a measure of how MySpace users are currently representing their identity and defining their profiles were developed. The results include several types of measurement, including:

- Profile visibility and privacy setting
- Network of friends and their similarities
- Profile preferences and photo presentation
- Validity and reality of information
- Patterns in identity and personality
- Identity representation overview

We first present some results from our primary analysis on the original data, and demonstrate the type and amount of published self-described information in Section **5.1.1**. In Section **5.1.2** we then describe extracted patterns in the data that we found through data mining and social network analysis. For these analyses, we used both training (known data) and validation (unknown data) datasets. The training set evaluates the relationship with each personality and the type of identity, while the validation set provides a larger picture to help us understand the extent of each identity feature. Additionally, in Section **5.1.3**, we describe the result of examining other possible ways of extracting more personality factors, such as the influence of photos and profile customization. These extra personality factors can be used for further research to determine the types of identity representation.

5.1.1 Initial Data Analysis

This section aims to describe and illustrate how people exposed their identity in our sample dataset. We observed and explored various demographic results from users' appearance online. This analysis will help us understand the extent of identity representation and facilitated our classification procedure.

Our dataset initially contained over 4.8 million profiles. After removing mutual friends, our dataset currently consists of 2.2 million nodes with 2.4 billion edges between them. **Figure 5.1** illustrates the number of friends within two groups of '*public*' and '*bands*' profiles (it should be noted that we have no knowledge about

the number of friends for '*private*' profiles as this information is not publicly available).

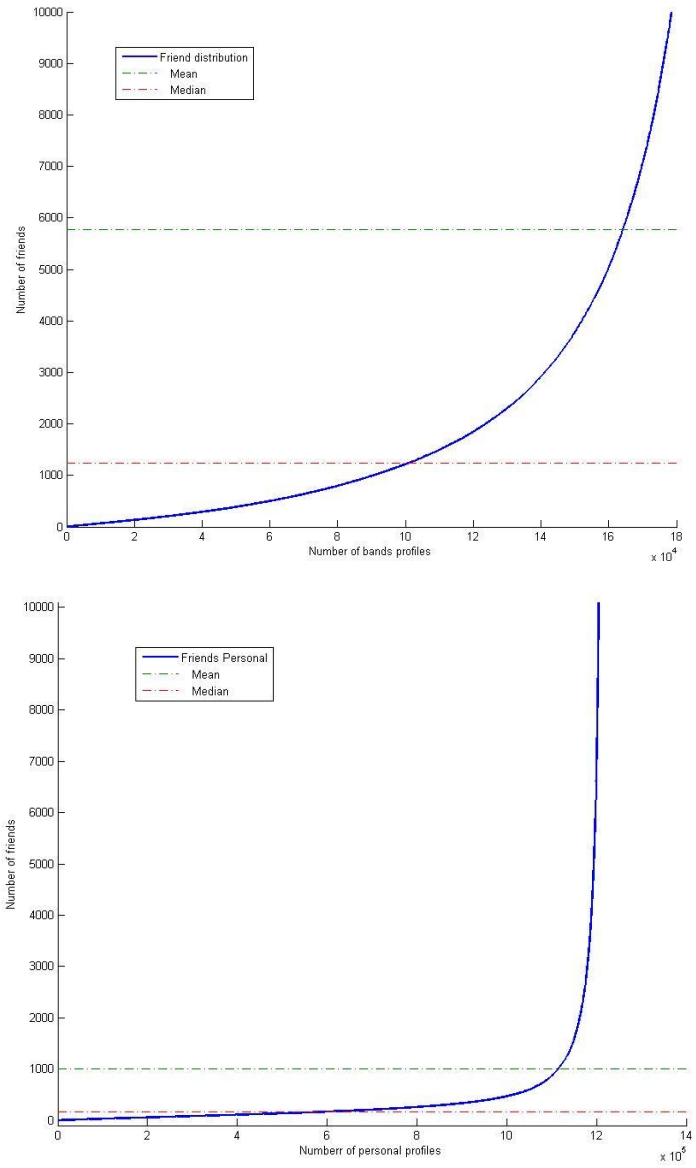


Figure 5.1 The degree distribution of friends for both public and bands profiles

Table 5.1 shows that our sample network employs many high degree connections, with an average of 1,010 friends for '*public*' profiles and 5,792 for '*bands*' profiles. Over 5% of our sample network has more than 10,000 friends and 2% have no friend or only have one friend. While the numbers of friends are almost the same for both male and female users, they vary in different age groups.

Table 5.1 The number of friends for both public and bands profiles

Friends Connection	Public Profiles	Bands Profiles
Number of participants	1,196,071	200,586
Total Friends	1,208,589,955	1,161,861,427
Maximum Friends	220,900,569	2,759,654
Average Friends (Mean)	1,010.46	5,792.33

However, the result shows that teens are more eager to have a greater number of friends (see **Figure 5.2**). We expected this result as from a psychological and sociological point of view **[boyd, 2006]**, teens and those in their early twenties are more eager to articulate and maintain their online social network of friends. Examining the age distribution between '*public*' profiles shows that published ages range from 16 to 107 years, while 12.1% of participants were not willing to reveal their age publicly. It appears that more than half of the participants (58.7%) were teenagers or people in their early twenties. Analysing gender also demonstrates that our sample network is more dominated by male users (58.1% vs. 40.5%), while 1.4% of users did not disclose their gender. **Figure 5.2** demonstrates the number of male and female users in our dataset with consideration to their age distribution. The mean age for females is 26.5 compared to the male age of 29.8. It can be assumed that, on average, more females are joining MySpace at a younger age than males.

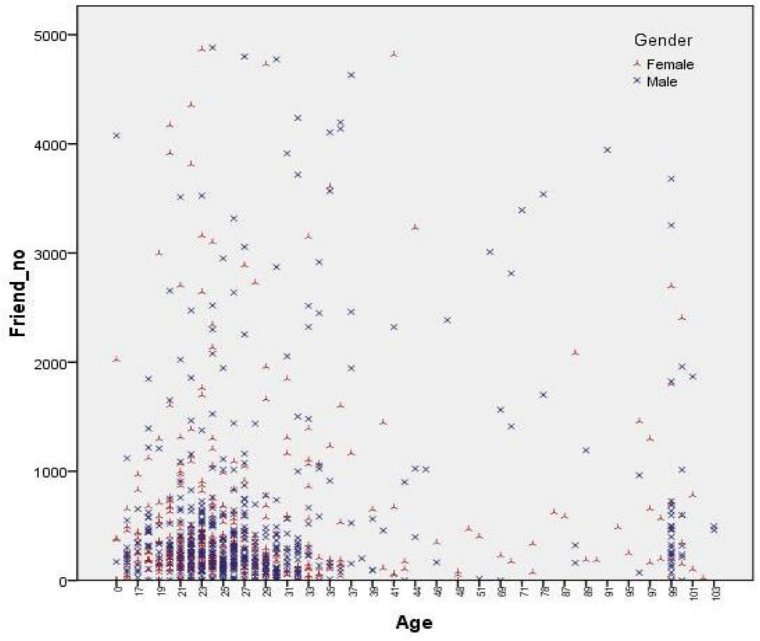


Figure 5.2 The age distribution between male and female users

Extracting profile descriptions of location (such as country, state and city) revealed that they are from very diverse geographical locations. Of the total, 75.4% claimed to live in the United States, which is almost three times more than the number of European users. The UK has the second largest MySpace population with 8.2%, and the third biggest population in our dataset is from the Philippines with 7.1% registered accounts.

In addition, we studied the number of activities and communication, such as comments, blog entries and the number of photos, which revealed the degree of communication and interaction between groups of profiles. **Figure 5.3** demonstrates the relationship between the number of friends, comments, blog entries and photos. The density in each block indicates the correlation within each interaction; for instance, the number of friends increases the frequency of comments and communication. Also, the number of photo and blog entries influences the number of comments and communication. These results help us to understand which elements are more important in order to classify each identity trait into different groups of personality factors.

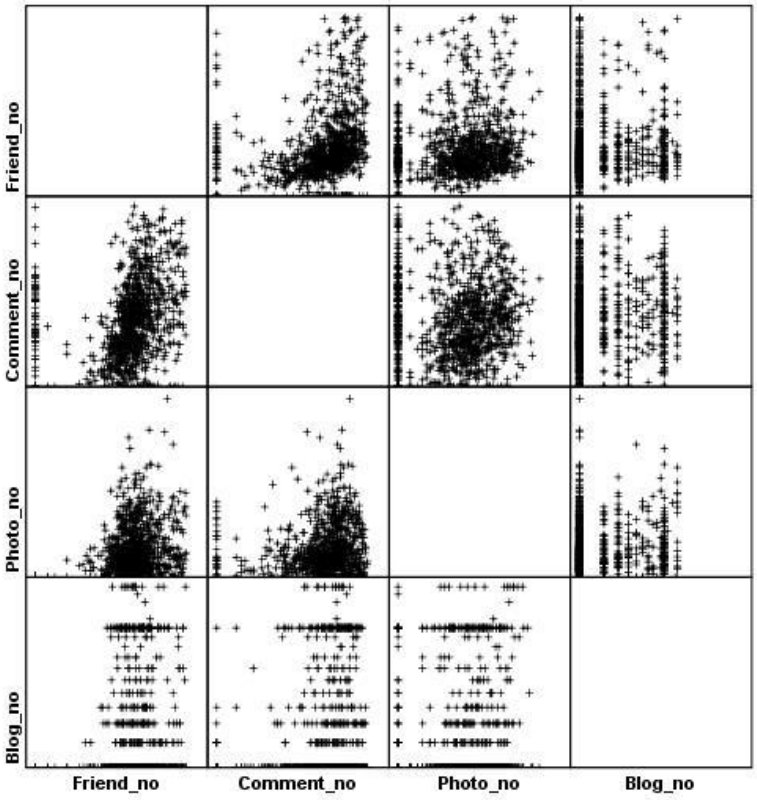


Figure 5.3 The correlation between represented identity traits

5.1.2 Pattern Discovery

Knowledge discovery is a process used to uncover patterns within data. The interpretations of each discovery are significant in order to facilitate our personality classification model. Therefore, we applied both text mining and network mining techniques to find patterns in our dataset. A large set of identity combinations were grouped to find the correlations between each identity element and to tune our personality classifier. The following describes some patterns we have identified, where some appear to be more expected.

- **Profile Visibility:** Comparing '*public*' and '*private*' profile visibility, it shows that one third of participants have modified their profile visibility to their group of friends only. There are more female users with '*private*' profiles than '*public*' profiles (63.6% vs. 36.4%); also female users change their privacy setting from '*public*' to '*private*' 9% more than male users. '*Private*' profiles are less active compared with '*public*' and '*bands*' profiles. '*Bands*' on the other hand are highly active, due to their music promotion, such as gigs, events and blogs. We also found 151 under-age users (aged between 14 and 15) with a '*public*' profile in our data sample; this category is not supposed to adjust their privacy to '*public*'. They disclosed all their personal information, including their photos.

- **Number of Friends:** Some identity elements, such as age, are strongly associated with the number of friends. Teenagers who fabricated their profiles had the most friends and created their own fame. The number of friends has a correlation with their geographical distance, for instance people tend to choose friends near to their location. Some other attributes are also noticeable, for example, people who spend more time writing comments on other profiles have a greater number of friends (see **Figure 5.3**).

- **Username:** Analysing the user name estimates that almost everyone in our data sample provides a name. Of these, 32% have disclosed both first and last name, 36% of the names are unrelated or fantasized, and 6% of the names are fabricated from celebrities or famous profiles associated with a fabricated photo. While females are more realistic about their name, male users seem to have used more fantasy names.

- **Gender Difference:** As with traditional gender difference, female users are more likely to hide their location, possibly due to their privacy protection. In contrast, male users seem to exaggerate their economic

status, such as income and occupation. Processing the use of language shows that male users use offensive language almost twice as much as female users (61% vs. 39%), while 58% of females use more positive language. In addition, males express themselves with a diverse language style, such as aggressive and offensive compared with the female group. Males therefore use more valid information to describe themselves compared with females (by 4%). Males have a higher number of friends; on average 976 compared with the female average of 632 friends. Females have more comments (an average 597 vs. 509), which indicates their sociability attributes. Females disclose more photos compared with male users (95 vs. 58). Male users are twice as likely to be members of a group activity as female users. Analysing the date of the last login shows that female users are more active on MySpace and login to their page more often, 62% compared with male users 53%. Males and females are equally as expressive, while males use more offensive words to describe themselves. Some patterns are correlated with other entities, such as gender and traceability attributes. For instance, females are less traceable than males, which are expected for psychology and sociology reasons.

- **Age:** Observing the correlation between age, occupation and income shows that profiles with higher incomes have more popularity and sociability characteristics, but on the other hand they provide less valid information, they are less traceable and used more offensive language. People in their 40s have a higher use of positive language, while people in their 30s are more traceable. About a quarter of participants hide both their age and gender. People who hide their gender are more likely to hide their age as well, and those who hide their age, gender and location have fewer friends. Those with a higher number of blog posts in their profile used fewer offensive words and have a correlation with the age of their profile. Also, profiles with a similar age and gender have a higher similarity rate in other interests and preferences.

- **Geographical Location:** Observing the profiles' geographical location confirms that, although the majority of profiles provide some information about their residency, they are selective about which part of their address should be disclosed: 94% disclosed their country of residence, 69% revealed their state/county and 82% disclosed their city/town, 8.6% named a place that does not exist. It appears that people feel more secure

disclosing their country than their city. They reveal their city/town more than their state/county in countries other than the US. A small percentage of users revealed their full address including the house number and their postcode. Overall more than half of the participants disclosed valid information about a location according to our classifier.

- **Marital Status and Orientation:** The result shows that two-third of participants disclosed both their relationship status and sexual orientation. Profiles that are looking for dating and serious relationships are more expressive and described themselves with more valid and traceable information. The married group and those in a relationship are more traceable than single and divorced, while swingers have the lowest traceability and the highest use of fantasy information. Bisexual users have a higher average number of friends, 1,325, and an average comment of 772, compared with straight users with 539 friends and 536 comments. Lesbians have more photos with an average of 115 than other sexual orientations. Gay users use more valid information to describe themselves; they are more traceable compared with other orientations.
- **Religions and Ethnicity:** Almost half of the participants indicate their religion: female users are more likely to disclose their religious views. Muslim users have on average 38 photos, while Christians have the highest average of 96 photos. Also, Muslim users have the lowest traceability of 24% compared with other religious groups. In term of ethnicity, black Africans are the most expressive. The result shows that Asian users are less traceable compared with white Caucasian (28% vs. 39%). Native Americans also have one of the highest values of using offensive language.
- **Profile's Photos:** On average married users have more photos compared with single users (93 vs. 65). Parents have on average 102 photos; they have higher rates for disclosing valid information and are less offensive, thus they have fewer comments on their page. It is interesting to see that profiles that did not reveal their age have on average more photos on their page. Asian, and especially Muslim, girls have fewer photos on their profiles. **Figure 5.4** shows the number of enclosed photos for our training profiles, and indicates that real profiles are more willing to enclose a photo in their profile compared with fake profiles.

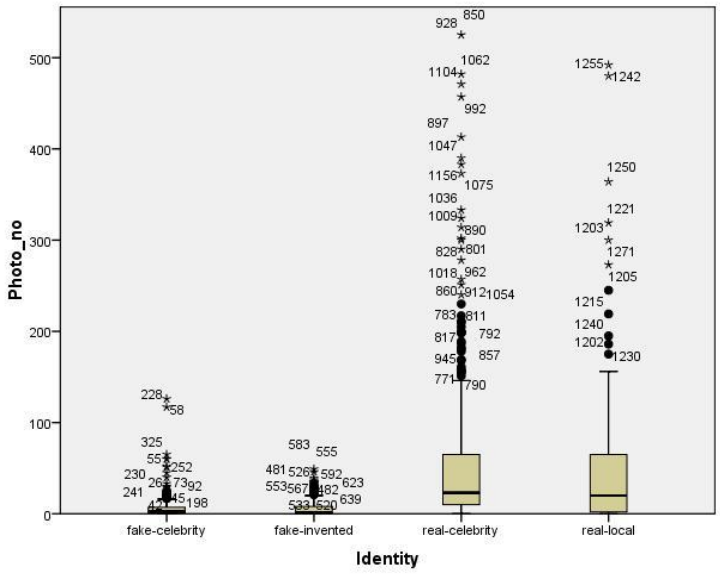


Figure 5.4 Comparing numbers of enclosed photos for each identity type

Overall, 69.2% of users state their age, gender, and their current residency, and almost 72% of participants disclosed all of their identity information. **Figure 5.5** demonstrates the disclosure of identity elements in accordance with our personality factors, such as '*anonymous*', '*fantasy*', '*offensive*' and '*valid*' attributes. For example, it shows that the username, city and occupation are the most fantasized pieces of information, while entities, such as the country of residence are the most valid information.

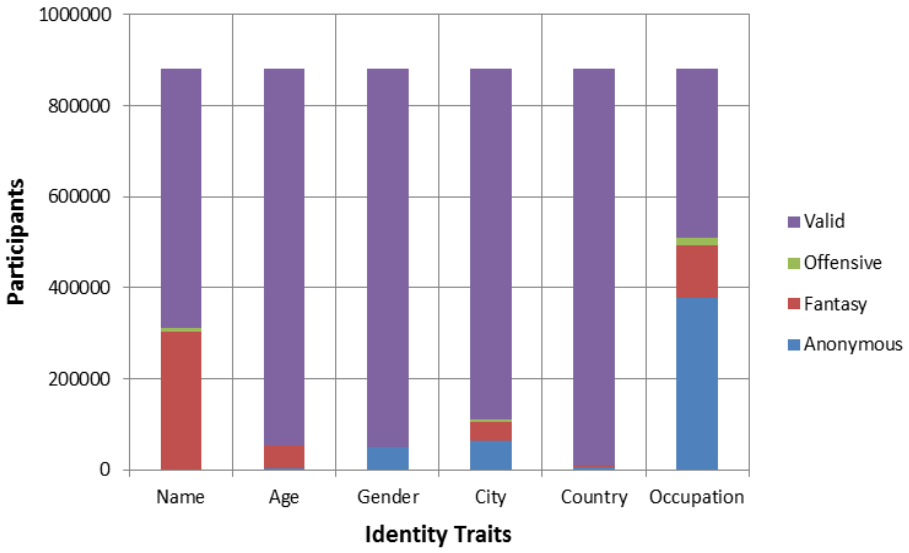


Figure 5.5 Frequency and the type of information disclosure

Figure 5.6 also shows the correlation between each personality attribute. For instance, '*fake*' identities are less '*valid*', while '*real*' nodes are less '*anonymous*', although there are some '*real*' nodes that have a higher value according to their '*fantasy*' attribute. Also we can see that the '*offensive*' attribute alone is not able to distinguish between the types of identity.

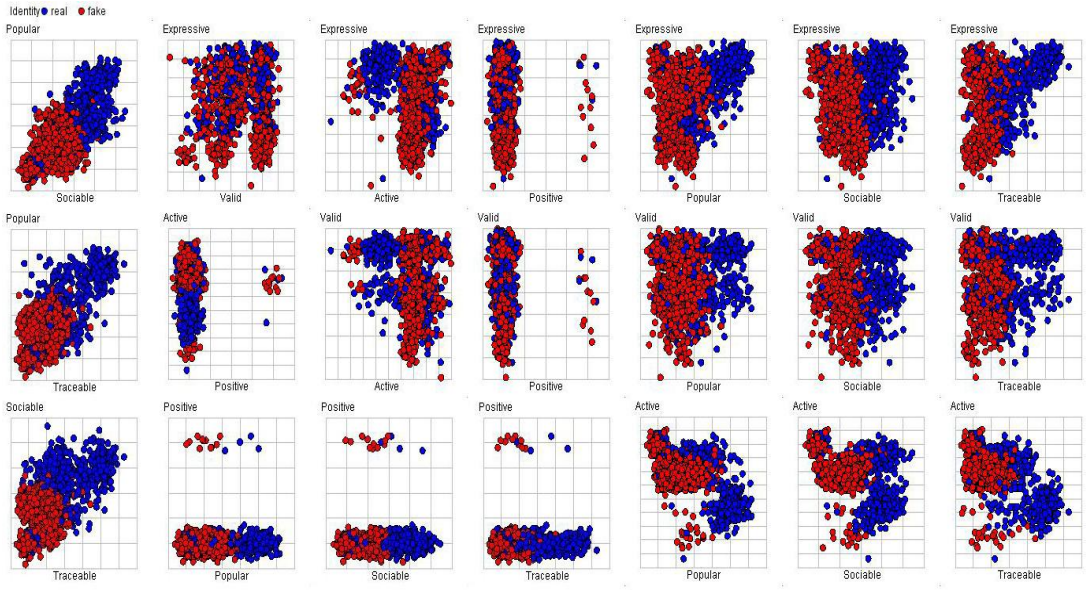


Figure 5.6 The correlation between each identity attributes

5.1.3 Further Personality Factors

A number of personality factors, such as ('*anonymous*', '*fantasy*', '*traceable*', etc.) were defined and analysed in Chapters **3** and **4**. In addition, we observed a small subset of profiles to measure the possibility of finding more users' attributes to distinguish '*real*' and '*fake*' personas. How people customized their profiles, and what type of photo (such as facial, group, fake and fantasy photos) they disclose, may reveal some information about the truth of the presented identity. Thus, these metrics are rather subjective and open to interpretation in identifying the validity of the identity. This would be interesting further research to analyse the look and feel of profiles together with an image processing approach to see if we can find any correlation within these attributes and the types of online identities. However, due to the time inefficiency of observing each profile manually, we decided not to include these attributes in our classifier. Within this section we briefly explain two different methods of examining identity traits, such as profile customization and photo observation.

5.1.3.1 Profile Customization

One of the reasons for the popularity of MySpace is that users have the freedom to alter and administer their pages. How much do people benefit from this ability to manage and customize their presentation of self? We selected our training dataset of 993 excluding *'fake-invented'* as there is no actual profile for this group. We manually observed if people customized their page and how much this customization demonstrated the owner's personality? We observed the embedded images, colour, songs, texture, background and other contents on their profiles. The result from this observation confirms that 64% of participants customized the look and feel of their pages by embedding code and other content into their profiles. Identity features, such as age, gender, marital status, etc., have considerable influence on how people define themselves to their audience through the choice of colour and background in their profiles. **Figure 5.7** compares the fraction of customizations in different identity groups with the influence of gender. It can be seen that '*real-celebrity'* are more likely to modify features on their page, while '*real-local'* are less likely to modify their page. Women are also more likely to customize their page compared with men. We analysed the profile customization by manual observation and decided to not include them in our personality classification.

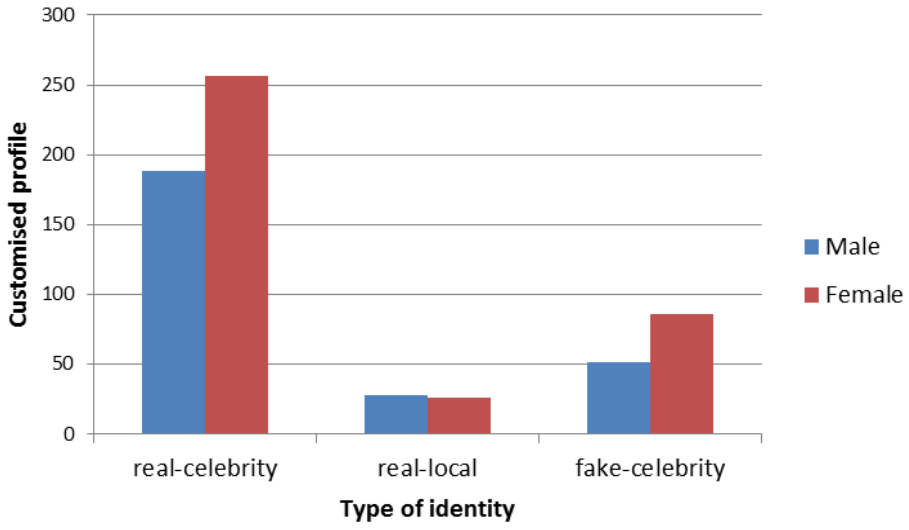


Figure 5.7 The number of customized profile within different type of identity

5.1.3.2 Photo Observation

Profiles often appear with photos, which give the highest evidence of identity performance. According to **[Riegelsberger *et al.*, 2003]**, "*the interpersonal cues given in a photo on the personal page can have a significant effect on trust in the whole site*". Manual examination of our training set of 993 profiles (excluding *'fake-invented'* as there is no actual profile for this group) indicates that the majority of profiles (82%) enclosed a photo and even a family album; these are classified into the following groups:

- **Facial Photo:** the photo related to a person and possibly a real image of the user.
- **Group Photo:** the photo contains a group of people and the user may not be identified in the photo.
- **Fake Photo:** the photo is apparently related to fame or celebrities.
- **Fantasy Photo:** the photo is not related to a person and could be any image.

Figure 5.8 demonstrates that the majority of participants disclosed a facial photo on their profile, 9% included a group photo, 6% of all images are clearly fabricated from celebrities, while 19% used fantasy and humorous images in their profile. The result shows that women disclosed personal photos more than men (62% vs. 38%). We also found that people who disclosed a photo have built a stronger network of friends and have more comments on their profiles. This study would be an interesting further research to analyse the profile's photo in order to verify the type of identity. However, due to image identification processing, we have not considered image identification for our entire dataset.

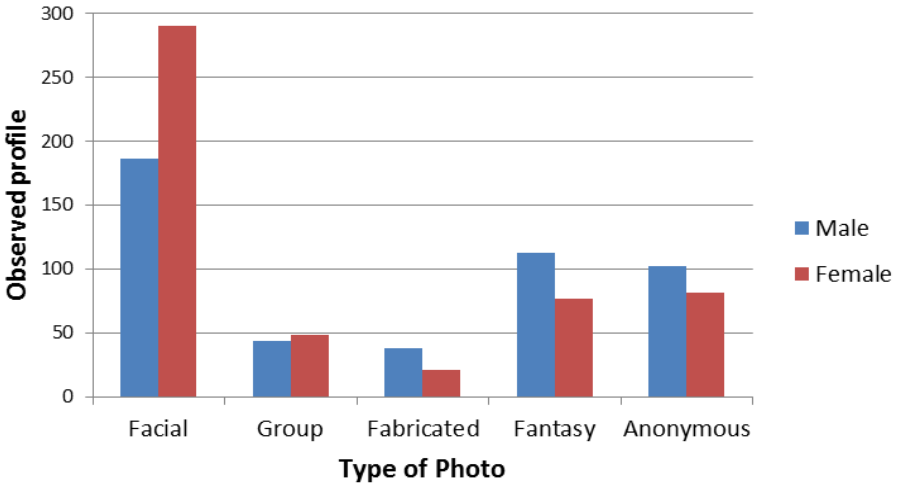


Figure 5.8 The type of published photos in observed profiles

5.2 Exploratory Results

This section provides some discussion and evidence of our findings through exploratory analysis. We have broken the results into three sections: social network analysis, principal component prediction, and machine learning comparisons.

In the next Section **5.2.1**, the results from the social network analysis show the structure and relationship between individuals and their friends. We have investigated two different types of measurements; centrality and similarity. Centrality analysis (such as in-degree, out-degree, between-ness, etc.) examines the position of each profile and models a social graph, where each node represents profiles and each edge represents the connection between profiles. The centrality analysis aims to find out the relationship between centrality and the type of identity. Similarity measurement, on the other hand, looks at how people are similar to their network of friends and examines the relationship between friends and the type of identity.

The prediction result from principal component analysis in Section **5.2.2**, demographically demonstrates the important factors of each personality and their relationship with each other. The evaluation results obtained from data mining are also included in Section **5.2.3**. By comparing and testing different data mining algorithms, we achieved some confidence on which algorithms are more accurate.

5.2.1 Social Network Analysis

In social networking the probability of spreading new characteristics depends on the influence of central nodes and the level of similarity and mutual interests between friends. The centrality and similarity analyses are important properties within a social group. The following results reflect on the effect of social network analysis in the determination of identity type. This examination provided evidence that social network measurements can reveal key properties of the network in order to classify the type of profile's information.

5.2.1.1 Centrality

As we discussed previously (in Section **3.3.4**), centrality analysis, such as out-degree, in-degree and between-ness, were applied to examine the

relationship between the type of identity and the position of each node within a network.

Measuring the centrality features of profiles with a known identity indicates that nodes with a higher centrality value have more influence over the network. **Figure 5.9** shows out-degree distribution of known profiles. It can be seen that the '*real-celebrity*' group has a high out-degree distribution: out-degree analysis alone is able to distinguish between the '*fake*' and '*real*' group, while it is difficult to indicate which sub-group (such as celebrity, local or invented) they belong to.

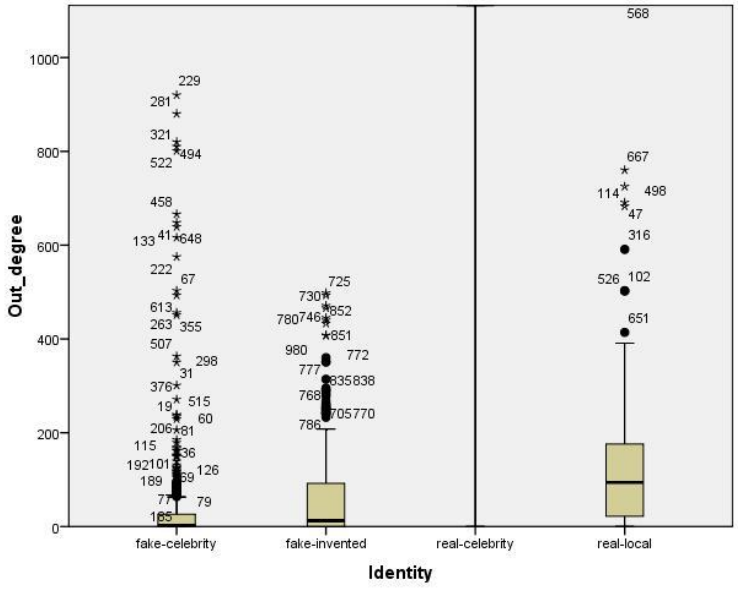


Figure 5.9 Out-degree distributions according to the type of identity

Figure 5.10 demonstrates the training nodes (known profiles) and their position within the network structure. Nodes are illustrated based on the type of identities and represented within different colours. The degree distribution indicates the strong association with centrality attributes, which has a significant high degree of connection forming a core element of a group structure. As seen in the graph, '*real-celebrity*' nodes are more closely tied to each other. The majority of nodes with zero out-degree (isolated nodes on the left side) are correlated to '*fake*' profiles. Although '*real-local*' nodes are distributed, there is little connection between '*fake*' and '*real-local*' groups. There is a weak connection between '*fake*' nodes as they seek to connect to '*real-celebrity*' group. According to **[Donath & boyd, 2004]**, this is because people would often like to connect to those who have a higher number of friends. We learnt that the higher centrality value between groups of nodes can distinguish the type of identity.

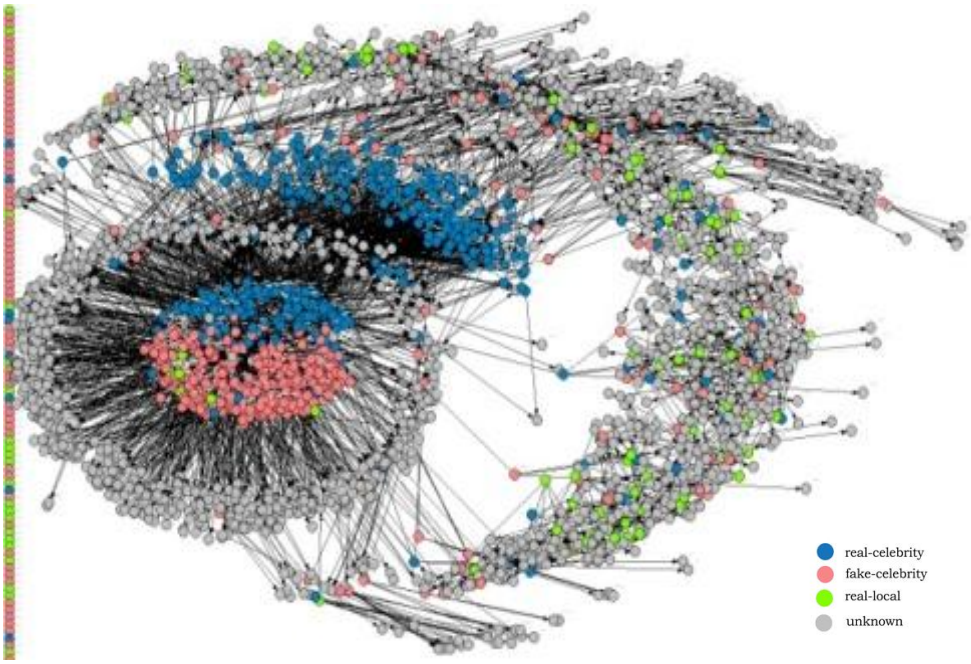


Figure 5.10 Network structure within known profiles and their friends

5.2.1.2 Similarity

Within our similarity examination, we first identified the notion of similarity within a group of users. We identified two identity elements as similar if they overlap between two profiles that have a connection as a friend. These similar attributes refer to both a user's identity traits and their personality factors, where we weighted each identity element based on our similarity formula (Section **3.3.4.2**). We measured similarity within both datasets of original identity elements (such as age, education, etc.), and pre-classified characteristics (such as valid, traceable, etc.).

By examining pre-classified personalities, we measured each personality factor to see if any of these personalities are important when examining the type of identity. **Figure 5.11** demonstrates the relationship between friends' similarity and their personality factors. For instance, as shown in the graph, similarity in attributes such as '*traceability*', '*validity*' and '*positive*' are not as significant as '*active*', '*sociable*' and '*popular*'. On average, participants are 67% similar to their group of friends.

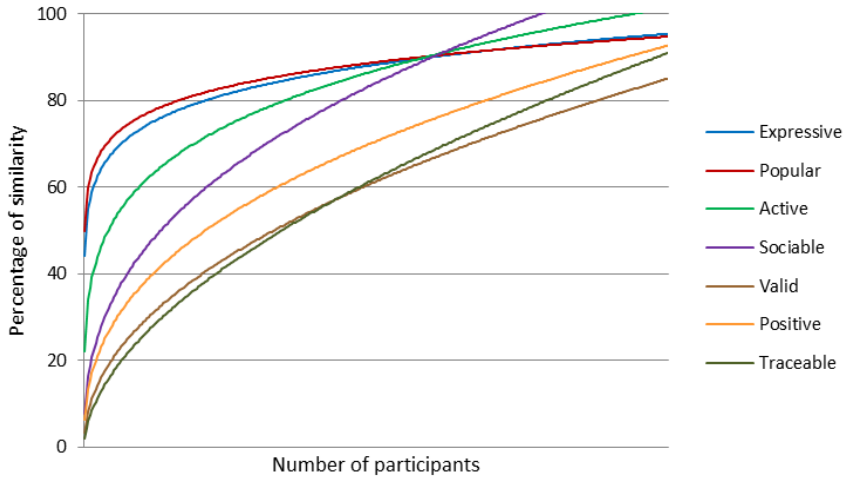


Figure 5.11 The similarity measurement between individuals (I=993) and their top 40 friends (F=17247)

The measurement of similarity over known profiles also provided some information about the relationships between the type of identity and the level of similarity in attributes. As shown in **Figure 5.12**, we are able to demonstrate that '*real-celebrity*' profiles are more similar to their friends than '*fake-celebrity*', while '*real-local*' fall between these two groups. In general, real profiles are more similar to their network of friends than fake personas. This may indicate that fake profiles have no standard for choosing a friend and they connect to anyone who responds to them. It should be noted that, as the *'fake-invented'* group were generated by online survey, we do not have any knowledge about their friends and therefore their similarity.

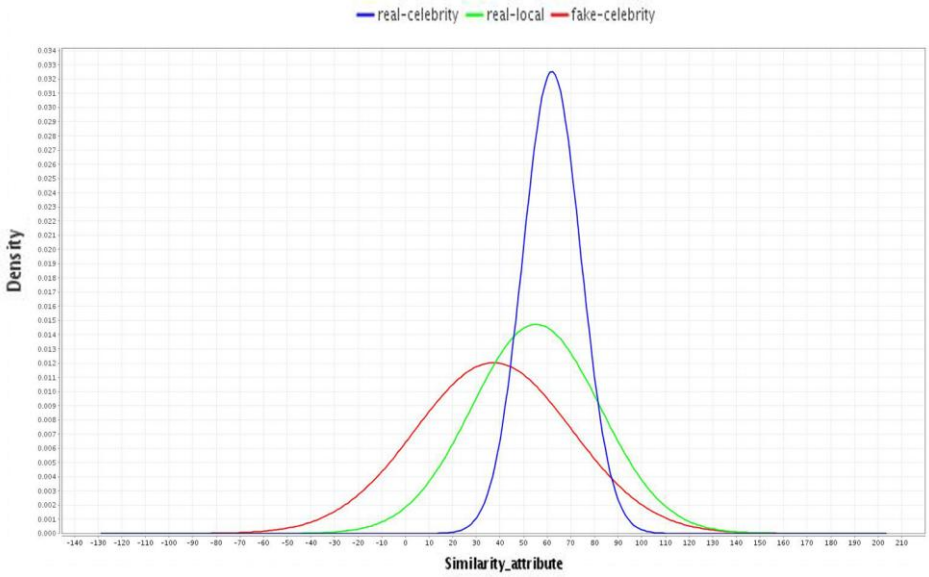


Figure 5.12 The density of similarity within different types of identity

5.2.2 PCA Prediction

The results from principal component analysis indicate the correlation between extracted principal components and its effect on predicting the type of identity. We analysed each personality and identity attribute using a component-loading graph, shown in **Figure 5.13**. The following graph illustrates the relationships between observed attributes and their dimension according to each extracted principal component. The plot projects the data along the directions where the identity varies the most. The principal components are located in a single axis in space, and the major direction of variability is neither along the 'dimension-1' nor the 'dimension-2' axis but somewhere in between them. Each identity element is represented by a vector, and the length and direction of each vector indicates how they contribute to the main principal component. These variables with long vectors are strongly associated with their dimension and therefore may provide more useful information about each entity. As seen in the graph, attributes with a longer vector such as '*traceable*' and '*sociable*' have more effect, while personalities such as '*positive*' or '*valid*' have less effect on deciding the type of identity.

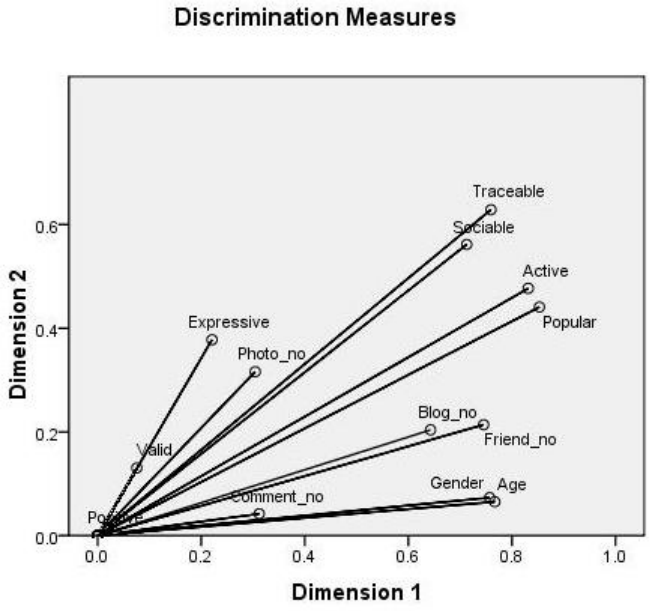


Figure 5.13 Principal component dimensions in accordance to different identity features

In order to understand the relationship between each component in relation to their identity type, the main factors for each identity variable can be illustrated in a tree graph. The tree graph of selected components, and the probability of how each principal component decides on the type of identity, is demonstrated

in **Figure 5.14**. By converting the value of each personality factor into three main components as 'pc1', 'pc2' and 'pc3', we are able to define each identity group. As shown in the graph, the '*real*' and '*fake*' group are identified within different dimensions, which describes the relationship between each factor and the type of identity. For instance, the top hierarchy of the graph shows that the majority of the '*fake*' group have higher correlations with 'pc1', while the '*real*' group are mostly distinguishable with 'pc2'.

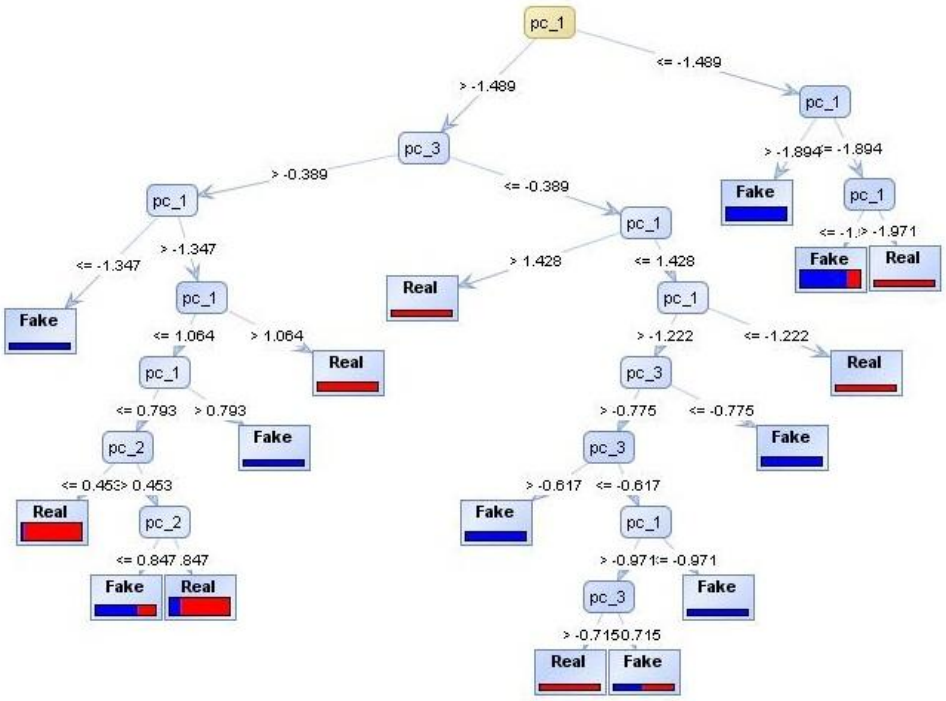


Figure 5.14 Decision Tree based on the principal components

Principal component analysis shows high confidence in distinguishing different types of identity. **Figure 5.15** illustrates the relationship between each component and the confidence rate achieved for each identity type. For instance, the confidence obtained by 'pc1' is higher than 'pc2' and 'pc3'. The higher numbers of clustered nodes indicate the higher confidence in the classification, while the distributed nodes have lower confidence rates. Those nodes which are misclassified with the opposite type are the error rate of false-positive and false-negative. We achieved an accuracy rate of 79% using PCA in predicting the type of identity.

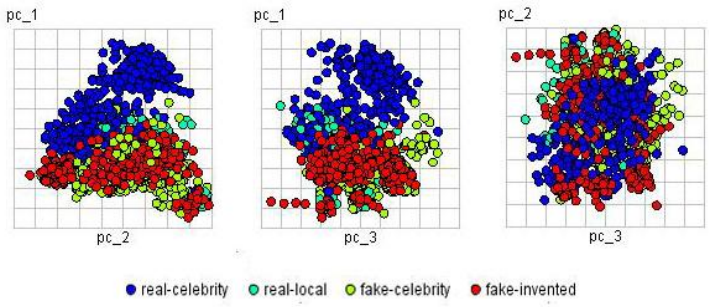


Figure 5.15 The correlation between each component and the types of identity

5.2.3 Machine Learning Comparison

In order to evaluate the accuracy of our classifier, we compared the confidence level on predicting the type of identity with different learners. We applied both types of data: original data (such as age, gender, location, etc.) and pre-classified data (such as valid, popular, traceable, etc.). The accuracy obtained for each learner is based on the confusion table (explained in Section **4.3.4**). We first examined the known data to evaluate the accuracy of identity prediction. By using a cross validation method, two-thirds of our dataset was selected as the training set and one-third as the test set.

We created different models by training data and evaluating the prediction performance on the test dataset using the confusion matrix. In order to compare each learner, we applied different data mining learners for both original data and pre-classified data, and each learning algorithm indicates different accuracy (see **Table 6.2**).

Table 5.2 Comparing different learners' accuracy using pre-classified data

	Decision Tree %	**Rule Learner %**	**Nearest Neighbors %**	**Naïve Bayes %**	**Average Accuracy %**
Public	89.58	90.07	88.15	83.20	87.75
Private	69.63	68.39	67.07	59.57	66.17
Band	99.09	99.21	98.55	91.32	97.04
Average Accuracy	86.10	85.89	84.59	78.03	**83.65**

Table 5.2 describes the comparison of the different learners applied to our training dataset. As seen in the table, '*private*' profiles have less prediction accuracy compared with '*public*' and '*bands*' profiles. This is because the '*private*' profiles have less identity attributes to use in each learner compared with other

types of profiles. The applied Decision Tree learner achieved the highest overall accuracy in detecting the type of identity by 86.1%, while the Nearest Neighbours was the fastest learner.

5.3 Evolutionary Results

After performing our classifier algorithm and identifying the personality factors for both sets of profiles (2007 and 2008), we formulated a transformation algorithm to measure both '*static*' and '*dynamic*' features of identity traits over time. As we explained in Section **3.4**, the identity traits were categorized into two groups: static (information that is unlikely to change over time, such as gender, Zodiac, etc.) and dynamic (information that may change over time, such as location, occupation, etc.). Analysing the profiles transformation of identity representation shows significant changes in profile contents for both '*static*' and '*dynamic*' features. Our results show rapid transformation over static identity traits. For instance, the average age changed from 25.36 to 30.24 for the same set of profiles during the period of one year, while we expected the average age range change to be 25.36+1. Also, gender modification from male to female increased by 1.19%, which is seven times more than shifting from female to male. This gender transformation is more widespread within the teenage group.

Through this analysis we found that, on average, a profile's content changed by 29% for static data and 45% for dynamic data over a one-year period. **Figure 5.16** shows that the *'real-local'* group have altered their profile information less compared with other groups. *'Real-celebrity'* group are more subject to transformation on their '*dynamic'* information, while '*fake-celebrity*' changed their '*static'* information more rapidly.

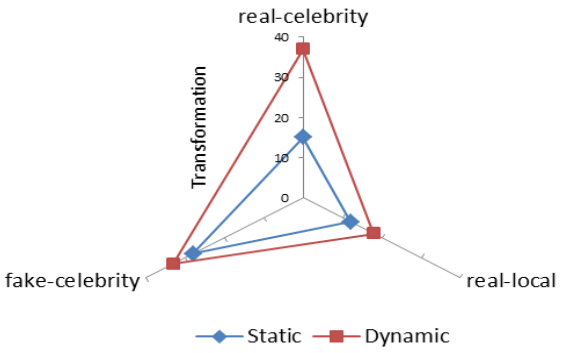


Figure 5.16 Static and dynamic transformation of identity over time

Understanding the impact of geographical, cultural and social status when constructing an identity profile may define some metrics that differentiate between the same profiles over a period of time. For instance, based on the residency of participants, US and UK users have a higher transformation on their identity representation, such as orientation, religion and ethnicity. Teenagers and females are more consistent at altering their '*static*' identity features, while male users increasingly transformed their '*dynamic*' identity representation (such as location, occupation and group membership). Those who are seeking dating and relationships also modify their online profiles more often. **Figure 5.17** compares the identity representation for both '*public*' and '*private*' profiles, and indicates that '*private*' profiles have made fewer changes in their self-described contents compared with '*public*' profiles.

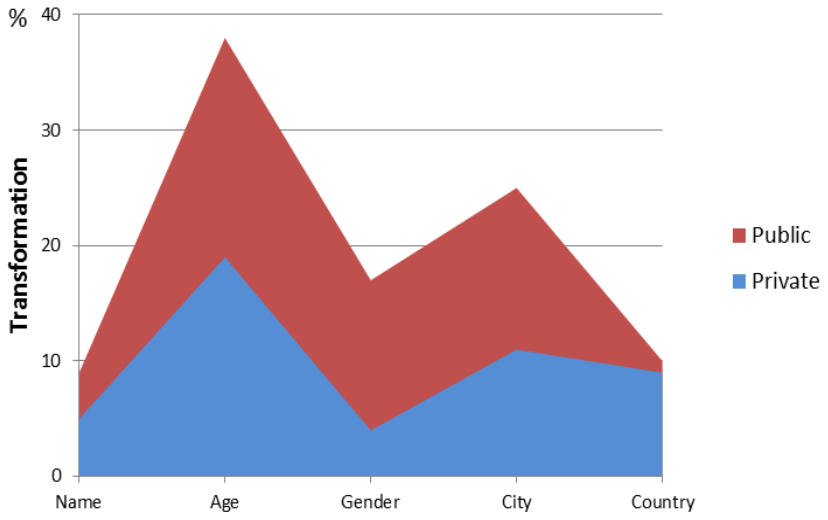


Figure 5.17 Transformation of identity for both public and private profiles over time

Figure 5.18 shows the identity representation transformation of average characteristics within the same profiles over a year. As seen in the graph, while people are becoming more '*active*' and '*sociable*' and acquiring more '*popularity*', they also become more '*anonymous*' and use more '*fantasy*' information to describe themselves.

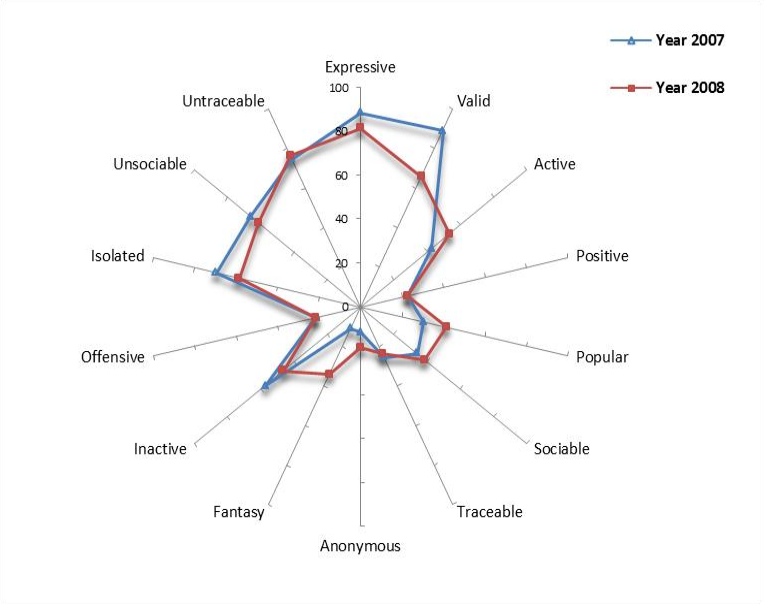


Figure 5.18 Transformation of personality attributes over time

In addition, we applied the similarity analysis in accordance with the level of identity transformation to see if similarity between friends has changed over time. **Figure 5.19** shows that over time profiles become less similar to their group of friends. It is interesting that those with a lower similarity rate have a higher transformation in their representation. On average, the similarity rate between groups of friends reduced by 8%. Dissimilarity in such a social network indicates that the linkage is less based on mutual interest, which will question the strength and meaning of friendship in online social networking.

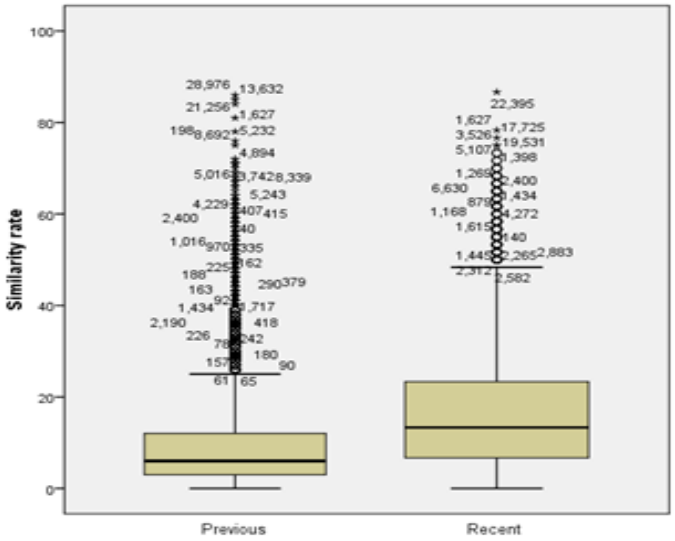


Figure 5.19 Transformation in similarity comparing both previous and recent profiles

5.4 Summary

Within this chapter, we described the demographical and statistical results of our study. The statistical results (Section **5.1**) show both initial and post analysis of data. We first examined the original data to describe our dataset and the fraction of identity representation. We then rated each personality factor using text mining and social network analysis to find out some useful patterns in data. These patterns helped us to improve our classifier in a development cycle. We also attempted to advance our classifier by searching for more personality factors, such as analysing photos and profile customization.

Some exploratory results from our study on social networks, principal component and data mining analysis are illustrated within the exploratory section (Section **5.2**). We examined two main properties of social networks; centrality and similarity. Centrality measurement (such as in-degree, out-degree and between-ness) proved to be a valuable method for distinguishing '*real*' and '*fake*' profiles. By training the known profiles, we learnt that profiles with a higher centrality value are more likely to be categorized within the '*real*' group. However, we have to consider an error rate for those who are carefully trying to gain popularity by faking to a real persona.

We then proposed a similarity algorithm (Section **5.2.1.2**) and discovered which identity elements and personality factors are more similar within a group of friends, and how the level of similarity correlated to the type of identity. For example, examining the pre-classified attributes shows that '*valid*' and '*positive*' attributes are not as important as being '*sociable*' and '*popular*'. This examination helped us understand if friendship (linkage) is based on trusting each other, and revealed more information about the context of links between people. However, being honest and trusting each other is an undirected property in a community, as friends are not necessarily honest with each other to the same degree.

The results from principal component analysis verified that, by extracting the main components, the information can be reduced and distinguished in a greater number of categorized groups. However, we lost some ineffective data during the analysis, which had an effect on the confidence of the prediction. The results from different learners are presented in the machine learning (Section **5.2.3**), where we investigated several data mining methods on both original and pre-classified data. This comparison suggests that the overall accuracy on the training dataset using the pre-classified data is more accurate than using the

original data. Although the diversity of data in both datasets is almost the same size, the pre-classified data is more reliable and has higher confidence in predicting the type of identity by 82.9% on average compared to 64.4%. After applying different data mining techniques, we achieved the best accuracy through the Decision Tree learner with 84.6% accuracy. To observe the relationship between entities, choosing the most useful data-mining algorithm is significant. For instance, some learners, such as Association Rules, are time ineffective; however they show higher performance and accuracy as a result. We found out that our classifier is more efficient in detecting and verifying identity representation compared with using the original data (see **Table 6.2**).

Furthermore, the analysis of identity evolution over a period of time for the same profile was presented in Section **5.3**, where both individual and network evolution of identity representation were examined. The results from profile's transformations confirm significant changes over both '*static*' and '*dynamic*' data. By examining the network evolutionary feature we found the relationship between the type of identity and the amount of identity transformation. In addition, the similarity measurement of both datasets (past and recent profiles) shows that people are gradually becoming less similar to their network of friends over time.

The next chapter discusses the overall results and findings. We will finalize the thesis with the conclusions chapter, including the limitation and a feasibility study over our proposed classifier, and discuss further possible extensions to this research.

Chapter 6

Discussion and Conclusions

"A real act of honesty is not enough to be honoured by everyone, but being witnessed by you and God alone."

Czeonollo

In the previous chapter we illustrated our main findings, including statistical, exploratory and evolutionary results. We presented the relationship between each entity, including the initial analysis, and discovered patterns in our dataset. Our exploratory results, including the outcomes from social network analysis, principal component analysis and several data mining approaches, were presented. We also explained the evolutionary results, such as the transformation of a profile's attributes and the evolutionary features of the social network over time.

Within this chapter we conclude our thesis with discussion on our findings and describe the correlation between the types of identity in accordance with our personality factors. In Section **6.1**, we discuss the overall efficiency of our classifier model and how a more sophisticated classifier could be implemented in the future. This chapter concludes with an overview of the future system in Section **6.2.1**, and highlights our research limitations in Section **6.2.2**. Interesting further studies are described in Section **6.2.3**, with the summary of the thesis in Section **6.2.4**.

6.1 Discussion of Results

Determining the type of identity is not a simple matter and cannot be modelled easily with a computational system. Considering the limitations of our study (see Section **6.2.2**), we proposed a classifier to distinguish between fact and fiction in relation to online identity representation. We employed several techniques, such

as social network analysis, principal component analysis and data mining, to implement and evaluate our personality classifier (see Section **3.3.1**). This classification model creates an image of a profile's characteristics, which may represent an actual identity of a profile holder. Our classifier is able to automatically examine each personality factor and therefore predict if the represented identity is real or fake.

Table 6.1 presents the fraction of each personality classification within different identity types over our training set (known data). As seen in the table, '*fake*' profiles have lower values for each personality factor, while the '*real-celebrity*' group have the highest values and '*real-local*' profiles fall in between.

Table 6.1 The percentage of each personality factors within known profiles

Training dataset (known profiles) N=1300							
Type of Identity	**Expressive %**	**Valid %**	**Active %**	**Traceable %**	**Popular %**	**Sociable %**	**Offensive %**
real-celebrity	85.1	89.96	87.96	48.19	60.77	54.83	0.38
real-local	84.77	89.66	71.31	36.07	25.72	28.35	0.17
fake-celebrity	75.26	84.42	78.97	15.37	15.77	22.81	0.13
fake-invented	72.28	75.15	81.86	09.65	17.92	12.30	0.00
Average	79.35	84.80	80.03	27.32	30.05	29.57	0.17

Our classifier is efficient in terms of usability, accuracy and computational time. Through different data mining methods, we found an average confidence rate to distinguish between real and fake profiles, while using original data from profile content will give us a less accuracy in predicting the type of identity representation. This means that we may be able to tag someone's profile as '*real*' or '*fake*' by examining a profile's characteristics. **Table 6.2** shows that by using the original data for each learner we achieved less accuracy in prediction (an average 64.4%), while using the pre-classified data is much faster to analyse and achieved 82.9%. Using original data in different machine learning proved to perform poorly in classifying and predicting the type of identities, while the pre-classified data improved the accuracy by almost 18%. Therefore, if we train profile's information based on their personality factors, the prediction accuracy would be much higher and the processing speed would be much faster.

Table 6.2 Average learner performance comparison when using both original and pre-classified data

	Decision Tree %	Rule Lerner %	Nearest Neighbors %	Naïve Bayes %	Average Accuracy %
Original Data					**64.42**
Personality Factor					**82.86**

We can conclude our findings as:

- Personality factors, such as '*expressive*', '*valid*', '*traceable*' and '*positive*' can determine the type of identity.
- An individual position on the network, such as centrality, has a correlation to the type of identity. For instance the more central nodes are more correlated with '*real'* group.
- The similarity between groups of friends can decide on the validity of identity. For instance '*real'* profiles are more similar to their friends than '*fake'* profiles.
- There is a correlation between the type of identity representation and the amount of transformation in self-described profiles over time. For instance, '*real*' profiles are less transformed over time than '*fake*' profiles.
- Existing methods such as data mining, social network analysis and principal component analysis can examine and predict the type of identity representation.

6.2 Conclusions

Determining the type of identity representation cannot be modelled easily with a computational system. According to **[Jøsang & Pope, 2005]** people present themselves differently within different contexts. People decide what type of online identity they want, revealing truth of self, keeping personal information to a minimum, fabricating an existing identity or creating a fantasy character. While some people are living in their fantasy world or fabricating other profiles, others might be at risk by disclosing a variety of personal information online. On the other hand, identity misrepresentation devalues the meaning of social networking. However, in all cases it is difficult to distinguish an online persona from the real person. This is because identity is a complex subject and only the

person who creates an online identity knows whether he/she has been honest or not.

Within this study we aimed to identify how people present themselves online and how to validate their profile identity. We examined several approaches in order to verify and distinguish between profile types. We investigated MySpace profiles on a large-scale and implemented a classification system to automatically examine each identity disclosure. Using different methods, we tried to answer our research questions by identifying some profile characteristics to identify the type of self-described identity online. For example, the principal component analysis helped us to reduce the data dimension and discover the important identity features for further data mining analysis. Data mining algorithms also helped us to evaluate our classification system in a development cycle. For instance, we presented a data mining framework that automatically found patterns in the known data and used these patterns to predict the type of identity for unknown data. We also measured the structure of the network by applying similarity and centrality analysis. The relationship between each node in our sample network provided insight into the individual's position in the network according to the type of identity representation. The relationship and interaction among the group of friends indicates the influence of friendship on online social networking. In terms of the efficiency of each technique we used (such as data mining, PCA and social network analysis), the confidence we achieved from our classifier would not be achieved without all these techniques.

In addition, we identified the evolutionary patterns in profile information over time and the influences on deciding the type of identity. Due to identity complexity, people are not expected to disclose one type of identity in the long term and their identity changes over time. We examined our dataset (approximately 2.2 million MySpace users) over two periods of time to see how profile 'p' transforms to 'q' over time. Although it is not obvious which one of the personas (p or q) represents the actual profile, we found some patterns in their identity transformation; this explains some evolutionary features in online social networking. This study proved that '*fake*' profiles are more subject to identity transformation compared with '*real*' profiles. Our transformation study indicates that, while profile attributes are becoming more active, sociable and have more friends, they also become more anonymous and use more fantasy information to describe themselves. Although the diversity of these attributes is not massive, over a longer period of time this shows the direction of online social networking.

Our initial classifier was difficult to adjust, but through a development cycle we learnt from the data and adjusted the classifier to achieve higher accuracy. There are sets of potential approaches and further work to improve the classifier model further. For instance, the efficient approach is to find different groups of profiles with known identity types and train them to achieve higher prediction and accuracy levels. Our classifier should be dynamic in order to detect transformation of identity, as through time people update and modify their identity representation. Furthermore, it would be more reliable if our classifier was able to trace multiple identities through different online communities. Although it is not an easy process to detect multiple identities, there needs to be a system, such as OpenID **[Recordon, 2006]**, to centralize and access the entire identity account into one single location.

In the end, implementing such a system to detect and evaluate online identity representation is a worthwhile goal that can be used to build a trust model within online social networking. Our personality classification system is powerful when combined with a recommendation model based on a human rating, as using users' interaction and a rating system is more reliable than the computer algorithm alone. Such an identity validation system can be used for many systems, such as firewalls, to detect and block unwanted connections. We predict that in the future the usage of social networking will be increasingly extended to trusted devices and systems that mediate interactions and transactions in the social world. Online social networking should rely on some form of identity management to secure underlying trust systems. Thus, the current state of the art in social networking does not properly address identity management within a trusted system, which remains the challenge for further research.

6.2.1 Future System

In this section, we evaluate the efficiency of our identity model if implemented as a future system. The property of our identity classifier should support at least the following requirements to assure higher feasibility:

- **Hardware and Software Resources:** Management tools are required to monitor and track disclosure of identity information in order to validate the type of identity. The network capability should be reconfigurable in real time to detect profile modification, rapidly address security threats, adapt to the context and support user's needs. Additionally, a high level

of software, such as data mining, social network analysis and language detection, is required to detect and validate profile identity in background. We are confident that technology development will take care of this issue by introducing more powerful software and devices.

- **Cost Effectiveness:** Social networking systems are designed to be cost effective and scalable for both users and service providers. However, we have to consider the network traffic and a database to keep track of personalities. There are many other costs, such as machines and human resources required for the identity verification system. Users can also benefit from validating their potential friend's identity for no or little cost.
- **Computation Time:** Our identity algorithm is efficient in terms of computational complexity (time complexity O(n)). The allocated time to evaluate each online profile and decide on the type of identity is very small (a second per profile). However, applying this system for the entire network may result in slowing down the processing time.
- **Privacy and Legal Issues:** A major drawback about our identity detection approach is that very personal information is examined, which raises many privacy issues. The management of identity has to be tightly coupled with a privacy policy and legal legislation to support users' privacy. The enforceability of privacy has to be provided to users so that they have control over which identity traits to disclose and to what extent.
- **Adaptability:** Currently our system is implemented using data from the MySpace community. The system should be adaptable and able to perform within different platforms with minimum configuration. Preferably this should allow users to have greater support for accessing their network of friends on other social networking sites and evaluate online identities across different platforms.
- **Performance:** Our classifier model is capable of detecting identities with confidence of, on average, 82.9%. Increasing the number of training set (users with known identity), and introducing more personality factors, could potentially improve the performance. However, our classifier would be more effective when combined with an efficient recommendation system.

6.2.2 Limitations

A number of limitations should be taken into account, including the design, implementation and functionality of our classifier model as follows:

- Due to the computational complexity of collecting the entire MySpace network, our data sample represents a small proportion (1% at the time of crawling) of the entire MySpace population. The result obtained from collected data may not be applicable if the entire populations were examined.
- We also have only collected the information of the top 40 friends, as some profiles, such as celebrity profiles, have a large number of friends (thousands or even millions). However, limiting the number of friends made it more feasible to examine a population based on a diverse network. We only have access to some basic information about '*private*' profiles. Also, the friends' information for '*private*' profiles is excluded in our study, as we have no access to their list of friends.
- The MySpace network is growing rapidly and many features we examined may change over time. Logging the changes quickly will provide a clearer picture about the transformation of online identity. Such a system should be able to constantly examine the history of a profile in order to verify identity. However, due to time limitations, we have collected the information over two periods of time.
- Detecting the type of identity takes considerable effort. Although we attempted to detect different groups of users in our study, there is no guarantee that our system detects users who carefully designed their profile with or without deception in mind.
- In this study, due to image processing identification, the analysis of users' photos was limited to a number of profiles, which is beyond the scope of our investigation. Photos are the most identifying identity feature and can provide us with more information about the validity of the profile's identity.
- The numbers of training profiles that we collected are not sufficient. Additional training datasets will obtain more accurate results when using machine learning algorithms. However, using many different sets of training data may confuse the system.

6.2.3 Further Research

There are many challenges for future investigation on distinguishing different types of identities online. This section describes some further studies as follows:

- Many psychological and sociological factors are involved in why people choose to act using an honest or dishonest characteristic within different contexts. It would be interesting to conduct a further study and take advantage of the theory of criminal psychology and human social behaviour and embed them into an identity verification system.
- Studying other social networking sites and their differences in terms of identity disclosure within different social contexts, such as gaming, chat rooms, blogging and dating communities, are interesting topics for further research. Further study is required to examine identity across different online communities and compare the properties of a highly trusted community with a less trusted one. We relied on data obtained from a single network (MySpace) rather than analysing different online social networking sites. It would be interesting to study different social networking sites and compare our results. For example, Facebook members are more tied to real world friendship and their members are believed to be more honest in comparison to MySpace users **[Dwyer *et al.*, 2007]**.
- It would be interesting to explore if the accuracy of our classifier can be improved further. This would be possible by obtaining more personality factors (such as cooperative, entertaining, enthusiastic, friendly, knowledgeable, arrogant, fanatical and so on) and incorporating more training data with known identities. Furthermore, embedding Natural Language Processing (NLP) with an identity detection system has the potential of effectively identifying and potentially classifying each identity feature.
- One highly interesting area of further research would be to explore how a profile's attributes influences the group, and if the individual fits well within a community. It would be an interesting study to identify different types of friendship or linkage regarding people's interactions and attributes to see if they fit well within a group of friends. According to **[Katona *et al.*, 2009]** not only can individual behaviour be predicted

from community structure but also community can be shaped from users' behaviour.

- Human evaluation required is based on user experience on online social networking. An extensive user study can be performed to assess the users' view of online representation. For instance, setting up some online profiles (both fake and real) as a game or quiz and asking participants to rate them based on their own criteria.
- We used existing methods, such as data mining, principal component analysis and social network analysis. This is because we believe that there are enough existing technologies to carry out this kind of investigation. However, employing other algorithms, such as SybilGuard (detect multiple identities) would be an interesting examination for further research **[Yu *et al.*, 2006]**.
- Further study on friend analysis could identify an individual's motivation for selecting someone as a friend, such as seeking similar interests and interactions, popularity, or deceptive behaviour (such as spammer, predator and identity fraud). One approach would be to examine the strength of a friendship and the context in which people reveal their identity. For instance, by analysing the type of friendship (such as friend, familiar stranger, stranger and community) the correlation between the type of connection and validation of identity representation can be examined further.
- Ultimately, this research could lead to a study on how to build a trust model on online social communities. The combination of a recommendation system and our identity validation system works more effectively when based on both machine learning and human recommendation. Such a trust model can verify the type of identity, predict and filter any deceptive and spammer acting as a firewall, for instance, accepting or rejecting a friend's request.

6.3 Summary of the Thesis

In summary, within this thesis we discussed how personal information can be classified and analysed to determine the validity of identity. We investigated the MySpace social networking site using a dataset collected through a spider and personalized script. After obtaining the profile content,

different types of personality and the taxonomy of identity were examined. To examine the validity, use of language and creditability of disclosed information, a text classifier was proposed that classified profiles based on their personality factors. Later we constructed an algorithm to rate each profile based on their taxonomy of identity and examined the 'top 40 friends' of each individual to measure their similarity.

To validate and find a pattern in the observed data, several data mining algorithms were applied and trained from our model, which helped us to improve our classification model. Principal component analysis was implemented to reduce the dimension of identity variables and extract the main factors and components. Analysing the network of friends also provided a fundamental understanding of the structure of our sample network. We applied some social networking techniques, such as similarity and centrality analysis, to measure profiles in relation to their friends in the network. Centrality and similarity analysis proved to be a metrics to distinguish different types of identity representation. Further, we examined the evolutionary features of identity representation over the period of one year and presented the results and discussion through this thesis. The transformation analysis indicates how personality factors changed over time and the direction of social networking sites.

To restate our initial research problem, this thesis began with an introduction and overview of our research problem (identity representation, validation of identity, trust management and privacy implication) (Section **1.2)**. We explained the objectives and contributions in Section **1.3**, on which we focussed within this study, including the research questions, research methods and research ethics. The goal of our study was to provide a reliable way to establish an identity model in order to detect and validate identity representation within online social networking. Our main focus is to evaluate online identity representation, the amount of published information, the validity of profile information, community structure, and how similarity between friends affects the type of identity. Furthermore, the main goal for our research was to provide a more trusted environment for online networking by helping the users and service providers to decide on the level of trust by examining a profile's personality.

We conducted a further literature study on related subjects and highlighted the related works. Our research background in Section **2.1** included the theory of the notion of identity, digital representation, social and multiple identities, and

an overview of the MySpace social networking site and its features. We studied online identity issues, such as privacy, anonymity, and trust, which opened up a wide direction for our research. MySpace identity concerns, such as ownership and fake identities, are also highlighted in Section **2.2**. The related works described current and past approaches to overcome online identity issues. Previous works, such as identity management systems, evaluation of social communities, social network analysis (such as centrality and similarity analysis), data mining, deception detection, and recommendation systems, are also explained in Section **2.3**. The summary of the literature reviewed is included in Section **2.4**.

We employed many different research approaches to implement a model for detection and validation of the type of identity. Our research approaches are described in Chapter **3**, including data accumulation and modelling the classifier. We examined MySpace profiles using both quantitative and qualitative studies to collect personal information from online profiles. The popularity of MySpace gave us an opportunity to observe this online community as a case study. For our qualitative study we crawled approximately 2.2 million profiles over a two month period. Using robots crawler (Section **3.2.1**), we accumulated large-scale information; however the type of identity is unknown. Therefore, we utilized a qualitative survey (see Section **3.2.2**) to collect four types of profile: '*real-celebrity*', *'real-local'*, *'fake-celebrity*' and '*fake-invented*'. These types of identities were used and trained for data mining purpose.

We explained the procedure of modelling our classifier by defining personality factors and the reason for choosing these factors (Sections **3.3.1** and **3.3.2**). We processed and clustered our data sample into seven opposite personality metrics: expressive/anonymous, valid/fantasy, active/inactive, positive/offensive, popular/isolated, sociable/unsociable, and traceable/untraceable. Using these metrics we intended to identify the type of identity for each individual in accordance with their network of friends. Section **3.3.3** explained the process of text mining in order to classify data within different personality factors.

We also incorporated an analysis of the social network structure using both centrality (such as in-degree and out-degree) and similarity measurement (Section **3.3.4**). This analysis helped us to explore the structural properties of our sample network and therefore classify profiles' content into more identifiable personalities (such as active, popular and sociable). We examined the relationship between centrality, similarity and the type of identity, and learnt

that centrality has a correlation with the type of identity and similarities have influence on profile characteristics. By examining a profile's contents and an individual's position in the community, we were able to build and improve our personality classifier.

Furthermore, we measured the evolutionary features of identity representation by examining self-described identity profiles of the same persona over time (Section **3.4**). This analysis helps us to understand how profile attributes changed over time. For instance we learned that '*fake*' profiles alter their profile content more rapidly than '*real*' profiles. We presented an investigation into a timeframe of identity transformation by comparing two sets of profile content together with their connections. We proposed a transformation algorithm to measure the differences in past and current representation within two groups of '*static*' and '*dynamic*' features. We also observed social structure and measured the similarity matrix within a group of friends over different timeframes.

Empirical methods such as PCA and data mining, which evaluated our proposed classifier, are described in Chapter **4**. We explained the procedure for principal component analysis, such as component and rotation analysis, measuring the correlation between each personality and the influence on predicting the identity type (Section **4.2**). A set of data mining techniques (Section **4.3**) were employed to train known data and estimate the validity of online identity. We described the data mining techniques, such as supervised and unsupervised learning, where we took advantage of existing machine learning techniques and identified some patterns within our dataset. By analysing different algorithms, we were able to explain the prediction accuracy through a confusion matrix table (see **Table 4.3** in Chapter **4**), which shows our classifier performance in terms of predicting the type of identity.

We presented the results and findings of our proposed approaches in Chapter **5**. We presented several statistical results, including the initial analysis, further personality factors and extracted patterns in data, which helped to improve our personality classifier (Section **5.1**). Exploratory results were also illustrated in Section **5.2**, including the results from our social network analysis, principal component analysis and data mining approaches. The evolutionary results, such as the transformation of profiles identity and the evolutionary features of online social networking, were also illustrated in Section **5.3**.

Finally, we concluded this thesis with a discussion on our findings in Chapter **6**, and a conclusion, including the efficiency of our classifier model and how more sophisticated classifiers could be implemented (Section **6.2**). We included an

overview of the possible future system (Section **6.2.1**) and our research limitations (Section **6.2.2**), including the opportunities for further research in Section **6.2.3**.

Bibliography

[Abbasi & Chen, 2008]
Abbasi, A. and Chen, H. (2008), "Writeprints: a Stylometric Approach to Identity-level Identification and Similarity Detection in Cyberspace", ACM Transactions on Information Systems 26.

[Abdul-Rahman & Hailes, 2000]
Abdul-Rahman, A. and Hailes, S. (2000), "Supporting Trust in Virtual Communities", Hawaii International Conference on System Sciences 33, Maui, Hawaii.

[Acquisti & Gross, 2005]
Acquisti, A. and Gross, R. (2005), "Information Revelation and Privacy in Online Social Networks (The Facebook case)", Pre-proceedings version, ACM Workshop on Privacy in the Electronic Society (WPES), Carnegie Mellon University.

[Acquisti & Gross, 2006]
Acquisti, A. and Gross, R. (2006), "Imagined Communities: Awareness, Information Sharing and Privacy on The Facebook" Proceedings of the 6th Workshop on Privacy Enhancing Technologies, Cambridge.

[Adamic & Adar, 2003]
Adamic, L. and Adar, E. (2003), "Friends and Neighbours on the Web", Soc.Networks 25(3): 211–230.

[Agrawal el al., 2003]
Agrawal, R., Rajagopalan, S., Srikant, R. and Xu, Y. (2003), "Mining Newsgroups Using Networks Arising from Social Behaviour", In WWW, pages 529–535.

[Airoldi & Malin, 2004]
Airoldi, E. and Malin, B. (2004), "Data Mining Challenges for Electronic Safety: the case of Fraudulent Intent Detection in E-mails", In Proc IEEE ICDM-2004 Workshop on Privacy and Security Aspects of Data Mining.

[Altman, 1977]
Altman, I. (1977), "Privacy Regulation: Culturally Universal or Culturally Specific", Journal of Social Issues, 33(3), 66-84.

[Badaskar *et al.*, 2005]
Badaskar, S., Agarwal, S. and Arora, S. (2005), "Identifying Real or Fake Articles: Towards better Language Modelling".

[Baier *et al.*, 2003]
Baier, T., Zirpins, C. and Lamersdorf, W. (2003), "Digital Identity: How to be Someone on the Net", In: Proceedings of the IADIS International Conference of e-Society.

[Batagelj & Mrvar, 1998]
Batagelj, V. and Mrvar, A. (1998), "Pajek - Program for Large Network Analysis", University of Ljubljana.

[Bechar-Israeli, 1995]
Bechar-Israeli, H. (1995), "From <Bonehead> to <cLonehEad>: Nicknames, Play and Identity on Internet Relay Chat", Journal of Computer- Mediated Communication.

[Berkovsky *et al.*, 2007]
Berkovsky, S., Borisov, N., Eytani, Y., Kuflik, T. and Ricci, F. (2007), "Examining Users´ Attitude towards Privacy Preserving Collaborative Filtering", Proceedings of DM, UM.

[Berman & Bruckman, 2001]
Berman, J. and Bruckman, A. (2001), "The Turing Game: Exploring Identity in an Online Environment"; Convergence, 83-102.

[Bollegalla *et al.*, 2008]
Bollegalla, D., Honma, T., Matsuo, Y., and Ishizuka, M. (2008), "Identification of Personal Name Aliases on the Web", In Proceedings of the World Wide Web Conference 2008.

[Borgatti *et al.*, 1999]
Borgatti, S., Everett, M. and Freeman, L. (1999), "UCINET 5 for Windows: Software for Social Network Analysis", Analytic Technologies, Inc., Natick, MA.

[boyd, 2002]
boyd, d. (2002), "FACETED ID/ENTITY: Managing Representation in a Digital World", Computer Science, Brown University, September 2002.

[boyd *et al.*, 2002]
boyd, d., Jensen, C., Lederer, S., and Nguyen, D.H. (2002), "Privacy in Digital Environments: Empowering Users", Workshop abstract to appear in Extended Abstracts of the ACM Conference on Computer Supported Co-operative Work (CSCW 2002).

[boyd, 2003]
boyd, d. (2003), "Reflections on Friendster, Trust and Intimacy", In: Intimate (Ubiquitous) Computing Workshop – Ubicomp, Washington.

[boyd, 2004]
boyd, d. (2004), "Friendster and Publicly Articulated Social Networking", Conference on Human Factors and Computing Systems (CHI 2004).

[boyd *et al.*, 2004]
boyd, d., Chang, M. and Goodman, E. (2004), "Representations of Digital Identity", Information Management and Systems, University of California, Berkeley.

[boyd, 2006]
boyd, d. (2006), "Identity Production in a Networked Culture: Why Youth Heart MySpace", American Association for the Advancement of Science, San Francisco.

[boyd & Ellison 2007]
boyd, d. and Ellison, N. B. (2007), "Social network sites: Definition, history, and scholarship", Journal of Computer-Mediated Communication, 13(1), article 11.

[boyd & Heer, 2006]
boyd, d. and Heer, J. (2006), "Profiles as Conversation: Networked Identity Performance on Friendster", In Proceedings of the Hawai'i International Conference on System Sciences (HICSS-39).

[Brzozowski *et al.*, 2008]
Brzozowski, M., Hogg, T. and Szabo, G. (2008), "Friends and Foes: Ideological Social Networking", in Proc. of the SIGCHI Conference on Human Factors in Computing, ACM Press, New York.

[Buckley, 2006]
Buckley, P. (2006), "The Rough Guide to MySpace and Online Communities", Press Release: 2 November 2006.

[Burgoon *et al.*, 2005]
Burgoon, J.K, *et al.* (2005), "An Approach for Intent Identification by Building on Deception Detection", presented at Hawaii International Conference on System Science (HICSS'05).

[Cameron, 2004]
Cameron, J.E. (2004), "A Three-Factor Model of Social Identity", Self and Identity, 3:3,239 - 262.

[Cameron, 2005]
Cameron, K. (2005), "The Laws of Identity", White paper, Architect of Identity, Microsoft Corporation.

[Casciaro, 1998]
Casciaro, T. (1998), "Seeing things clearly: Social Structure, Personality, and Accuracy in Social Network Perception", Social Networks, 20: 331-351.

[Caverlee & Webb, 2008]
Caverlee, J. and Webb, S. (2008), "A Large-scale Study of MySpace: Observations and Implications for Online Social Networks", In Proc. of ICWSM.

[Chakrabarti, 2000]
Chakrabarti, S. (2000), "Data Mining for Hypertext: A Tutorial Survey", ACM SIGKDD Explorations, 1(2):1-11.

[Coates *et al.*, 2000]
Coates, D., Adams, J., Dattilo, G., and Turner, M. (2000), "Identity Theft and the Internet", Colorado University.

[Damiani *et al.*, 2003]
Damiani, E., De Capitani di Vimercati, S. and Samarati, P. (2003), "Managing Multiple and Dependable Identities", University of Milan, Published by the IEEE Computer Society.

[Danylak & Edmonds, 2005]
Danylak, R. and Edmonds, E. (2005), "The Interactive Game: Origins and Effects", Proceedings of the second Australasian conference on Interactive entertainment, p.65-69, Sydney, Australia.

[Dey & Abowd, 1999]
Dey, A.K. and Abowd, G.D. (1999), "Towards a better Understanding of Context and Context-awareness", Proceedings of the 1st international symposium on handheld and ubiquitous computing, London, Springer.

[Dokas *et al.*, 2002]
Dokas, P., Ertoz, L., Kumar, V., Lazarevic, A., Srivastava, J. and Tan, P.N. (2002), "Data Mining for Network Intrusion Detection," Proceedings of National Science Foundation Workshop on Next Generation Data Mining.

[Donath & boyd, 2004]
Donath, J. and boyd, d. (2004), "Public Displays of Connection", BT Technology Journal, 22:71–82.

[Donath, 2007]
Donath, J. (2007), "Signalling Identity", chapter abstracts with representative bibliographic references.

[Donath *et al.* 1999]
Donath, J., Karahalios, K. and Viegas, F. (1999), "Visualizing Conversation", Proceedings of the Thirty-second Annual Hawaii, International Conference on System Sciences (HICSS 32), Los Alamitos, CA: IEEE Computer Society Press.

[Douceur, 2002]
Douceur, J.R. (2002), "The Sybil attack", in Proceedings of IPTPS, Cambridge, MA, pp. 251–260.

[Dwyer, 2007]
Dwyer, C. (2007), "Digital Relationships in the 'MySpace' Generation: Results from a Qualitative Study", Proceedings of the 40th Hawaii International Conference on System Sciences (HICSS), Hawaii.

[Dwyer *et al.*, 2007]
Dwyer, C., Hiltz, S.R. and Passerini, K. (2007), "Trust and Privacy Concern within Social Networking Sites: A Comparison of Facebook and MySpace", Proceedings of AMCIS 2007, Keystone, CO.

[Eckel & Wilson, 2003]
Eckel, C.C., Wilson, R.K. (2003) "Conditional trust: sex, race and facial expressions in a trust game".

[Ellison *et al.*, 2006]
Ellison, N., Lampe, C. and Steinfield, C. (2006), "Spatially Bounded Online Social Networks and Social Capital: The Role of Facebook", Annual Conference of the International Communication Association (ICA).

[Elmore & Richman, 2001]
Elmore, K. and Richman, M. (2001), "Euclidean Distance as a Similarity Metric for Principal Component Analysis, Mon, Wea, Rev. 129:540–549.

[Fairhurst, 2003]
Fairhurst, M. (2003), "Document Identity, Authentication and Ownership: the Future of Biometric Verification", Proc. of 7th Int. Conf. on Document Analysis and Recognition (ICDAR 2003), Vol. II, IEEE Computer Society, Edinburgh, Scotland, pp. 1108–1116.

[Felt *et al.*, 2008]
Felt, A., Hooimeijer, P., Evans, D. and Weimer, W. (2008), "Talking to Strangers without Taking their Candy: Isolating Proxied Content", In Proc. of SocialNets '08, ACM.

[Fono & Raynes-Goldie, 2006]
Fono, D. and Raynes-Goldie, K. (2006), "Hyperfriendship and Beyond: Friends and Social Norms on LiveJournal", Internet Research Annual Volume 4: Selected Papers from the AOIR Conference (pp. 91–103).

[Ford & Strauss, 2008]
Ford, B. and Strauss, J. (2008), "An Offline Foundation for Online Accountable Pseudonyms", In Proc. of the 1st International Workshop on Social Network Systems (SocialNets), Glasgow, Scotland.

[Freeman, 1979]
Freeman, L.C. (1979), "Centrality in Social Networks: Conceptual Clarification", Social Networks, 1: 215-239.

[Galloway & Simoff, 2006]
Galloway, J. and Simoff, S.J. (2006), "Network Data Mining: Methods and Techniques for Discovering Deep Linkage between Attributes", In APCCM '06: Proceedings of the 3rd Asia-Pacific conference on Conceptual modelling, Australian Computer Society, Inc.

[Gill & French, 2007]
Gill, A.J. and French, R.M. (2007), "Level of Representation and Semantic Distance: Rating Author Personality from Texts", Proceedings of the Second European Cognitive Science Conference (EuroCogsci07).

[Goffman, 1959]
Goffman, E. (1959), "The Presentation of Self in Everyday Life", New York: Doubleday.

[Goodman, 1961]
Goodman, L.A. (1961), "Snowball Sampling", Annals of Mathematical Statistics, 32, 148–170.

[Grayson, 2002]
Grayson, T.R.D. (2002), "Philosophy of Identity", Part of the Identity Planet series.

[Guillaumin *et al.*, 2009]
Guillaumin, M., Verbeek, J., and Schmid, C. (2009), "Is that you? Metric learning approaches for face identification", In ICCV, 2009.

[Gutierrez & Feigenbaum, 2006]
Gutierrez, A.J. and Feigenbaum, J. (2006), "Towards Better Digital Identity Management", Sensitive Information in a Wired World.

[Hand, 1998]
Hand, D.J. (1998), "Data Mining: Statistics and More?", The American Statistician, 52, 112-118.

[Hegrat, 2007]
Hegrat, N. (2007), "Selling your MySpace Account on Ebay", http://www.associatedcontent.com/article/144349/selling_your_myspace_account_on_ebay.html?cat=35

[Hill *et al.*, 2006]
Hill, S., Agarwal, D., Bell, R. and Volinsky, C. (2006), "Building an Effective Representation for Dynamic Network", Computational and Graphical Statistics.

[Hogg *et al.*, 2008]
Hogg, T., Wilkinson, D., Szabo, G. and Brzozowski, M. (2008), "Multiple Relationship Types in Online Communities and Social Networks", in Proc. of the AAAI Spring Symposium on Social Information Processing.

[Holms *et al.*, 2004]
Holme, P., Edling, C. and Liljeros, F. (2004), "Structure and time-Evolution of an Internet Dating Community", Social Networks, 26:155.

[Holmes *et al.*, 1994]
Holmes, G., Donkin, A. and Witten, I.H (1994), "WEKA: A Machine Learning Workbench", Department of Computer Science, University of Waikato, Hamilton, New Zealand.

[Hsu & Helmy, 2006]
Hsu, W.H. and Helmy, A. (2006), "Capturing User Friendship in WLAN Traces", IEEE INFOCOM poster.

[Hsu *et al.*, 2006]
Hsu, W.H., King, A., Paradesi, M.S., Pydimarri, T. and Weninger, T. (2006), "Collaborative and Structural Recommendation of Friends using Weblog-based Social Network Analysis", In Proceedings of the 2006 AAAI Spring Symposium on Computational Approaches to Analyzing Weblogs.

[Hsu *et al.*, 2007]
Hsu, W.H., Lancaster, J., Paradesi, M.S. and Weninger, T. (2007), "Structural Link Analysis from User Profiles and Friends Networks: A feature construction approach", Proceedings of ICWSM.

[Hu *et al.*, 2007]
Hu, J., Zeng, H. and *et al.* (2007), "Demographic Prediction based on User's Browsing Behaviour", Proceedings of the 16th international conference on World Wide Web.

[Huffaker & Calvert, 2005]
Huffaker, D.A. and Calvert, S.L. (2005), "Gender, Identity, and Language Use in Teenage Blogs", Journal of Computer-Mediated Communication, 10(2), article-1.

[Hui *et al.*, 2007]
Hui, P., Yoneki, E., Chan, S. and Crowcroft, J. (2007), "Distributed Community Detection in Delay Tolerant Networks", MobiArch, Kyoto.

[Jae-On & Mueller, 1978]
Jae-On, K. and Mueller, C.W. (1978), "Factor Analysis: Statistical Methods and Practical Issues", Beverly Hills, CA: Sage Publications.

[Jøsang & Pope, 2005]
Jøsang, A. and Pope, S. (2005), "User Centric Identity Management", CRC for Enterprise Distributed Systems Technology (DSTC Pty Ltd), AusCERT conference.

[Jungermann, 2009]
Jungermann, F. (2009), "Information Extraction with RapidMiner", In Wolfgang Hoeppner, editor, Proceedings of the GSCL Symposium 'Sprachtechnologie and eHumanities', pages 50–61.

[Kagal *et al.*, 2001]
Kagal, L., Finin, T. and Joshi, A. (2001), "Trust-based Security in Pervasive Computing Environments", In IEEE Communications.

[Kaiser, 1960]
Kaiser, H. F. (1960), "The Application of Electronic Computers to Factor Analysis", Educational and Psychological Measurement, 20, 141-151.

[Katona *et al.*, 2009]
Katona, Z, Zubcsek, P. P. & Sarvary, M. (2009), "Network Effects and Personal Influences: Diffusion of an Online Social Network".

[Keogh & Pazzani, 1999]
Keogh, E.J., and Pazzani, M. (1999), "Learning Augmented Bayesian Classifiers: A Comparison of Distribution-based and Classification-based Approaches", Proc. 7th Intl. Workshop on AI and Statistics (pp. 225-230).

[Klösgen & Zytkow, 2002]
Klösgen, W. and Zytkow, J.M. (2002), "Handbook of Data Mining and Knowledge Discovery" Oxford University, Press.

[Koch, 2002]
Koch, M. (2002), "Global Identity Management to Boost Personalization", In: Proc. 9th Research Symp on Emerging Electronic Markets, p. 137 – 147.

[Kumar *et al.*, 2006]
Kumar, R., Novak, J. and Tomkins, A. (2006), "Structure and Evolution of Online Social Networks", In Proceedings of the 12th ACM SIGKDD Philadelphia.

[Lai, 2005]
Lai, E. (2005), "Teen Uses Worm to Promote Site: Manipulation Pushes MySpace Site to Record Hits, but Raises Security Concerns", Computerworld.

[Lampe *et al.*, 2007]
Lampe, C., Ellison, N. and Steinfield, C. (2007), "A Familiar Face(book): Profile Elements as Signals in an Online Social Network", Proceedings of the SIGCHI conference on Human factors in computing systems, San Jose, ACM.

[Lederer *et al.*, 2003]
Lederer, S., Beckmann, C., Dey, A. and Mankoff, J. (2003), "Managing Personal Information Disclosure in Ubiquitous Computing Environments", Technical reports CSD-03-1257.

[Lee *et al.*, 2008]
Lee, D., Larose, R. and Rifon, N. (2008), "Keeping our Network Safe: a Model of Online Protection Behaviour", Behaviour & Information Technology, 27:5,445 - 454.

[Lesniewski-Laas, 2008]
Lesniewski-Laas, C. (2008), "A Sybil-proof One-hop DHT", In Workshop on Social Network Systems.

[Liu *et al.*, 2003]
Liu, R.X., Kuang, J., Gong, Q. and Hou, X.L. (2003), "Principal component regression analysis with SPSS", Compute Methods Programs Biomed 2003; 71:141–7.

[Maia *et al.*, 2008]
Maia, M., Almeida, J. and Almeida, D. (2008), "Identifying User Behaviour in Online Social Networks", Proceedings of the 1st workshop on Social network systems, Glasgow, ACM.

[Malin, 2005]
Malin, B. (2005), "Unsupervised Name Disambiguation via Social Network Similarity", in: Workshop Notes on Link Analysis, Counterterrorism, and Security.

[Marlow, 2006]
Marlow, C.A. (2006), "Linking without Thinking: Weblogs, Readership, and Online Social Capital Formation", International Communication Association Conference, Dresden.

[Mazar *et al.*, 2007]
Mazar, N., Amir, O. and Ariely, D. (2007), "The Dishonesty of Honest People: A Theory of Self-concept Maintenance", Journal of Marketing Research, 45, 633-644.

[McPherson *et al.*, 2001]
McPherson, J.M., Smith-Lovin, L., and Cook, J.M. (2001), "Birds of a feather: Homophily in social networks", In J. Hagan & K. S. Cook (Eds.), Annual review of sociology, vol. 27: 415-444. Palo Alto, CA: Annual Reviews.

[Mesch & Talmud, 2006]
Mesch, G. and Talmud, I. (2006), "The Quality of Online and Offline Relationships: The Role of Multiplicity and Duration of Social Relationships", The Information Society,22:3,137 - 148.

[Milgram, 1967]
Milgram, S. (1967), "The Small World Problem", Psychology Today, 2(60).

[Mislove *et al.*, 2007]
Mislove, A., Marcon, M., Gummadi, K.P., Druschel, P. and Bhattacharjee, B. (2007), "Measurement and Analysis of Online Social Networks", In Proceedings of the 5th ACM/USENIX Internet Measurement Conference (IMC'07).

[Mundinger & Le Boudec, 2005]
Mundinger, J. and Le Boudec, J.Y. (2005), "The Impact of Liars on Reputation in Social Networks", In Proceedings of Social Network Analysis: Advances and Empirical Applications Forum, Oxford, UK.

[MySpace Terms, 2006]
MySpace.com, (2006), "Terms of Use Agreement", http://www.myspace.com/index.cfm?fuseaction=misc.terms

[MySpace Wikipedia]
MySpace Wikipedia, http://en.wikipedia.org/wiki/MySpace

[Nabeth, 2005]
Nabeth, T. (2005), "Understanding the Identity Concept in the Context of Digital Social Environments", INSEAD CALT (the Centre for Advanced Learning Technologies), CALT-FIDIS.

[Nardi *et al.*, 2000]
Nardi, B., Whittaker, S. and Schwarz, H. (2000), "It's Not You Know, It's Who You Know: Work in the, information Age".

[Pato, 2003]
Pato, J. (2003), "Identity Management: Setting Context", Trusted Systems Laboratory, HP Laboratories Cambridge.

[Park *et al.*, 2002]
Park, H.W., Barnett, G.A. and Nam, I.Y. (2002), "Hyperlink-affiliation Network Structure of top Web Sites: Examining Affiliates with Hyperlink in Korea", Journal of the American Society for Information Science, 53(7), 592-601.

[Perkel, 2006]
Perkel, D. (2006), "Copy and Paste Literacy: Literacy Practices in the Production of a MySpace Profile", paper presented at the DREAM-Conference: Informal Learning and Digital Media: Constructions, Context, Consequences, Odense, Denmark.

[Petroczi *et al*, 2006]
Petroczi, A., Nepusz, T. and Baszo, F. (2006), "Measuring Tie-strength in Virtual Social Networks", Connections, 27(2):49–67.

[Pfitzman & Hansen, 2008]
Pftzmann, F. and Hansen, M. (2008), "Anonymity, Unlinkability, Undetectability, Unobservability, Pseudonymity, and Identity Management", consolidated proposal for terminology, Technical report.

[Qu *et al.*, 2002]
Qu, Y., Ostrouchovz, G., Samatovaz, N. and Geist, A. (2002), "Principal Component Analysis for Dimensions Reduction in Massive Distributed Data Sets", In Proceedings of IEEE International Conference on Data Mining (ICDM).

[Recordon, 2006]
Recordon, D. and Reed, D. (2006), "OpenID 2.0: A Platform for User-centric Identity Management", In Proceedings of the Second ACM Workshop on Digital Identity Management, DIM '06. ACM, New York.

[Richardson & Domingos, 2002]
Richardson, M. and Domingos, P. (2002), "Mining Knowledge-Sharing Sites for Viral Marketing", In Proc. of the Eighth Intl, Conference on Knowledge Discovery and Data Mining (SIGKDD'02).

[Riegelsberger *et al.*, 2003]
Riegelsberger, J., Sasse, M.A. and McCarthy, J.D. (2003), "Shiny Happy People Building Trust? Photos on E-commerce Websites and Consumer Trust", Proceedings of the CHI2003, April, Ft. Lauderdale, FL.

[Roccas *et al.*, 2002]
Roccas, S., Sagiv, L., Shalom, H. S., Knafo, A. (2002), "The Big Five Personality Factors and Personal Values", Pers Soc Psychol Bull June 2002 28: 789-801.

[Russo & Koesten, 2005]
Russo, T.C. and Koesten, J. (2005), "Prestige, Centrality, and Learning: A Social Network Analysis of an Online Class", Communication Education, 54:3,254 - 261.

[Ryberg & Larsen, 2008]
Ryberg, T. and Larsen, M.C. (2008), "Networked Identities: Understanding Relationships between Strong and Weak Ties in Networked Environments", Journal of Computer Assisted Learning, 24, 103-115.

[Seigneur & Jensen, 2004]
Seigneur, J.M. and Jensen, C.D. (2004), "The Role of Identity in Pervasive Computational Trust", Trinity College Dublin, Department of Computer Science, TCD-CS-2004-48.

[Schau & Gilly, 2003]
Schau, H.J. and Gilly, M.C. (2003), "We are what we Post? Self-Presentation in Personal Web Space", Journal of Consumer Research, 30 (December), 385–404.

[Schilit *et al.* 1994]
Schilit, B.N, Adams, N. and Want, R. (1994), "Context-Aware Computing Applications", IEEE Workshop on Mobile Computing Systems and Applications.

[Shand *et al.*, 2004]
Shand, B., Dimmock, N. and Bacon, J. (2004), "Trust for Ubiquitous, Transparent Collaboration", University of Cambridge Computer Laboratory, J.J. Thomson Avenue, Cambridge, Wireless Networks 10, 711–721, Kluwer Academic Publishers.

[Sherif *et al.*, 2000]
Sherif, T.A., Magdy, A.S. and Marghny, H.M. (2000), "Identity Detection of Typist relying on Image Processing Techniques", ICGST-GVIP Journal, Volume 7, Issue 3.

[Shrivastava *et al.*, 2008]
Shrivastava, N., Majumder, A. and Rastogi, R. (2008), "Mining (social) Network Graphs to Detect Random Link Attacks", In Data Engineering ICDE 2008, IEEE 24th International Conference on, pages 486–495.

[Smarr, 2001]
Smarr, J. (2001), "Technical and Privacy Challenges for Integrating FOAF into Existing Applications".

[Smith, 2002]
L.I. Smith, L.I. (2002), "A Tutorial on Principal Components Analysis", Cornell University, USA.

[Somanathan & Rubin, 2004]
Somanathan, E. and Rubin, P.H. (2004), "The Evolution of Honesty", Journal of Economic Behavior & Organization, Elsevier, vol. 54(1), pages 1-17.

[Spertus *et al.*, 2005]
Spertus, E., Sahami, M. and Buyukkokten, O. (2005), "Evaluating Similarity Measures: A Large-scale Study in the Orkut Social Network", Proceedings of 11th International Conference on Knowledge Discovery in Data Mining (pp. 678–684). New York: ACM Press.

[Squicciarini *et al.*, 2009]
Squicciarini, A.C., Shehab, M. and Paci, F. (2009), "Collective Privacy Management in Social Networks", ACM World Wide Web Conference.

[Steiner, 1993]
Steiner, P. (1993), "On the Internet, nobody Knows you're a Dog", Cartoon in The New Yorker,
http://www.unc.edu/depts/jomc/academics/dri/idog.html

[Stevenson, 1972]
Stevenson, L. (1972), "Relative Identity and Leibniz's Law", The Philosophical Quarterly, Vol. 22, No. 87 (Apr., 1972), pp. 155-158.

[Stolfo *et al.*, 2000]
Stolfo, S., Lee, W. and Mok, K. (2000), "Adaptive Intrusion Detection: a Data Mining Approach", Artif. Intelli. Rev, 533–567.

[Strauss *et al.*, 2001]
Strauss, J.P., Barrick, M.R. and Connerley, M.L. (2001), "An Investigation of Personality Similarity effects (relational and perceived) on Peer and Supervisor Ratings and the Role of Familiarity and Liking", Journal of Occupational & Organizational Psychology, 74(5): 637–657.

[Stutzman, 2006]
Stutzman, F. (2006), "An Evaluation of Identity-sharing Behavior in Social Network Communities", In: Proceedings of the 2006 iDMAa and IMS Code Conference, Oxford.

[Suler, 2002]
Suler, J.R. (2002), "Identity Management in Cyberspace", Journal of Applied Psychoanalytic Studies 4:455–460.

[Thongtae & Srisuk, 2008]
Thongtae, P. and Srisuk, S. (2008), "An Analysis of Data Mining Applications in Crime Domain", Computer and Information Technology Workshops 2008, IEEE 8th International Conference on , vol., no., pp.122-126.

[Toma *et al.*, 2008]
Toma, C.L., Hancock, J.T. and Ellison, N.B. (2008), "Separating Fact from Fiction: An Examination of Deceptive Self-presentation in Online Dating Profiles", Personality and Social Psychology Bulletin, 34, 1023-1036.

[Torkjazi *et al.*, 2009]
Torkjazi, M., Rejaie, R., and Willinger, W. (2009), "Hot today, gone tomorrow: on the migration of MySpace users", In Proceedings of the 2nd ACM Workshop on online Social Networks (Barcelona, Spain, August 17 - 17, 2009).

[TrendMaker, 2006]
TrendMarker, (2006), "The Social Butterfly Effect", the Social Networking Web-World, Universal McCANN.

[Tufekci, 2008]
Tufekci, Z. (2008), "Can you see me now? Audience and Disclosure Regulations in Online Social Network Sites", Bulletin of Science Technology & Society, vol. 28, no. 1, pp.20–36.

[Watts *et al.*, 2002]
Watts, D.J. and Dodds, P.S. and Newman, M.E.J. (2002), "Identity and Search in Social Networks", Science 296:1302–1305.

[Whitty, 2002]
Whitty, M.T. (2002), "Liar, Liar! An Examination of how Open, Supportive and Honest People are in Chat Rooms", Computers in Human Behaviour, 18, 343-352.

[Windley, 2005]
Windley, P.J. (2005), "Digital Identity", Unmasking Identity Management Architecture (IMA), O'Reilly Media, Inc.

[Wisse & Jansen, 2006]
Wisse, P, Jansen, P. (2006), "Identity Management Distilled, a Comparison of Frameworks", University of Amsterdam, Sprouts: Working Papers on Information Systems, 6(12).

[Ying & Chris, 2009]
Ying, Z. and Chris, H. (2009), "Identity Construction and Trust Building in Developing International Collaborations", Journal of Applied Behavioral Sience, 45(2), 186-211.

[Yoneki *et al.*, 2008]
Yoneki, E., Hui, P. and Crowcroft J. (2008), "Distinct Types of Hubs in Human Dynamic Network", In EuroSys Socnet.

[Yu *et al.*, 2006]
Yu, H., Kaminsky, M., Gibbons, P.B. and Flaxman, A. (2006), "SybilGuard: Defending against Sybil Attacks via Social Networks", Technical Report IRP-TR-06-01, Intel Research Pittsburgh.

[Zarandioon *et al.*, 2009]
Zarandioon, S., Yao, D., and Ganapathy, V. (2009), "Privacy-aware Identity Management for Client-side Mashup Applications", In *Proceedings of the 5th ACM Workshop on Digital Identity Management* (Chicago, Illinois, USA, November 13 - 13, 2009). DIM '09. ACM, New York, NY, 21-30.

[Zinman & Donath, 2007]
Zinman, A. and Donath, J. (2007), "Is Britney Spears Spam?", Paper presented at the CEAS 2007, Fourth Conference on Email and Anti-Spam, Mountain View, CA.

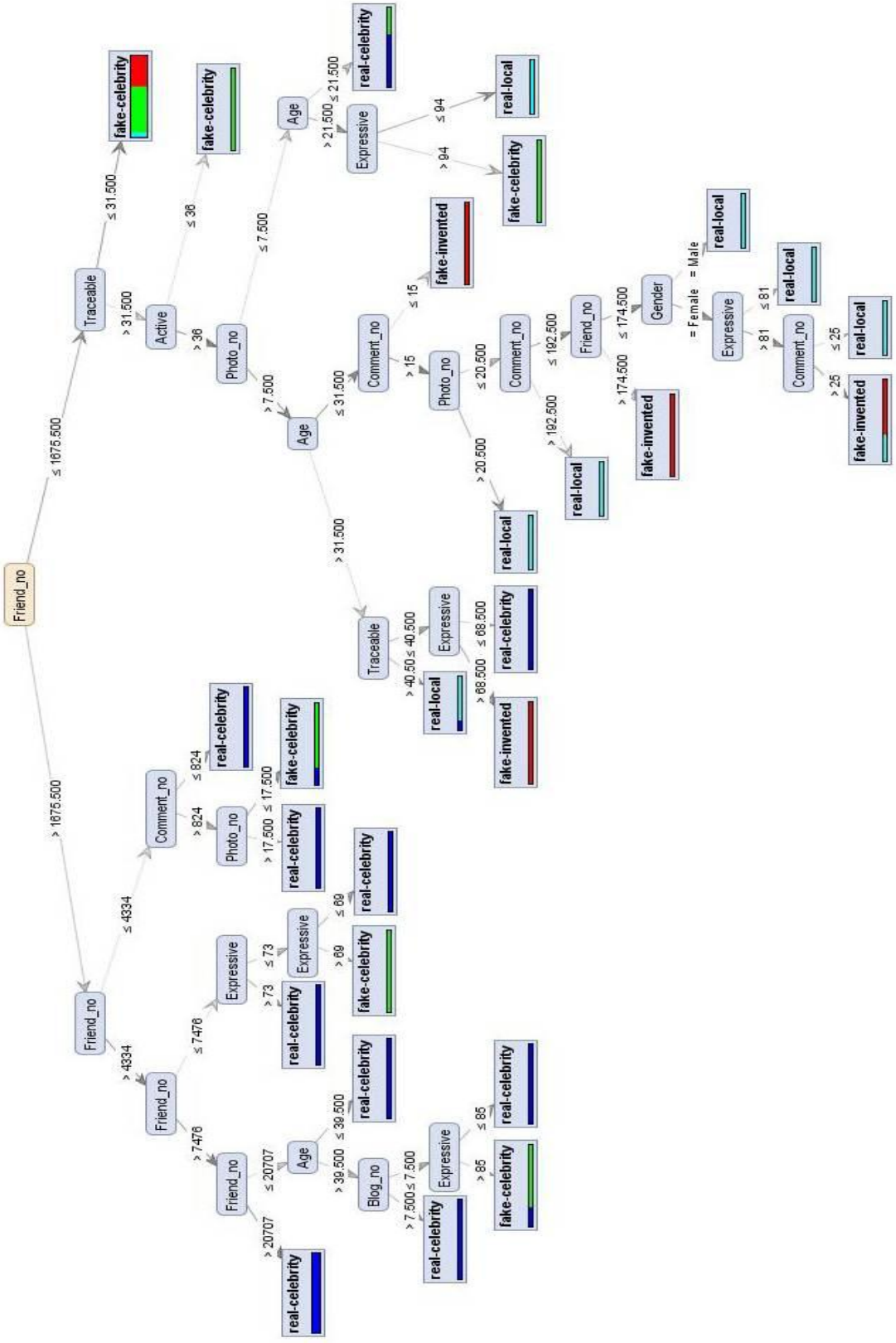
[Zolli, 2004]
Zolli, A. (2004), "Socialize This!" American Demographics, Vol. 26, Issue. 7.

Appendix

Appendix A: Grouping of each identity features

Attribute	Category
Profile	Public, private, bands
Friend-ID	the unique number for each profile
Username	hidden, valid, offensive or fantasy
Age	hidden, underage(14-15), teens(16-19), 20s(20-29), 30s(30-39), 40s(40-49), 50s(50-59), 60+(60-119) or exaggerated
Gender	male, female or hidden
City	hidden, valid, offensive or fantasy
Country	hidden, valid, offensive or fantasy
Last login	high active: logged in within a day, active: logged in within a week, moderate: logged in within two weeks, low active: logged in within a month, not active: logged in within or more than two months
View number	number of hits on the bands profile ranges 0 to millions
Member since	the age of bands profile as the year since they have join
Band URL	hidden, valid, offensive or fantasy
Record label	hidden, valid, offensive or fantasy
Here for	networking, dating, serious relationships, friends or the combination of these
Status	single, in relationship, married, divorced or swinger
Orientation	straight, gay, lesbian, gay lesbian , bi or not sure
Occupation	hidden, valid, offensive or fantasy
Education	High school, in college, some college, college graduate, professional school or post grad
Body type	hidden, fantasy (under estimate, over estimate , normal)
Zodiac	pisces, aquarius, libra, leo, cancer, taurus, gemini, capricorn, virgo, sagittarius or aries
Religion	catholic, protestant, Christian other, Jewish, Muslim, Buddhist, Hindu, scientologist, Mormon, Taoist , atheist , agnostic, wiccan, other
Smoke/Drink	no/no , no/yes, yes/no, yes/yes
Children	not for me, proud parent, someday or undecided
Ethnic	white Caucasian, black African, Asian, east Indian, Latino Hispanic, pacific islander, middle eastern, native American, other
Income	less than £30k, £30k_£45k, £45k_£60k, £60k_£75k, £75k_£100k, £100k_£150k, £150k_£250k, £250k+
Group	number of group activity, ranges from 0 to 10
School	number of school, ranges from 0 to 10
Blog	number of blog, ranges from 0 to 100
Photo	number of school, ranges from 0 to thousands
Comments no	number of comments, ranges from 0 to millions
Friends no	number of friends, ranges from 0 to millions

Appendix B: Decision Tree learner using both personality factors and original data

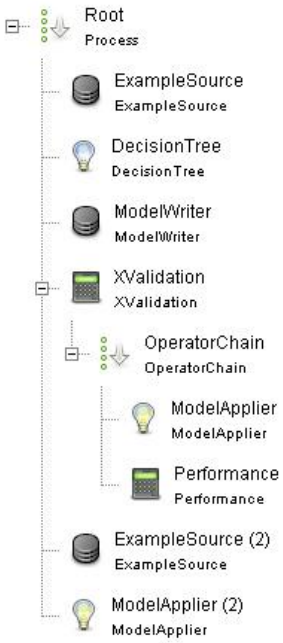


Appendix C: A sample of Association Rules learner

if Popular ≤ 19 and Active ≤ 80.500 and Friend_no ≤ 33.500 then fake-celebrity (1 / 1 / 232 / 22)
if Friend_no > 948 then real-celebrity (400 / 1 / 17 / 0)
if Expressive ≤ 52 then fake-invented (0 / 2 / 5 / 104)
if Comment_no ≤ 30.500 and Expressive ≤ 98 and Friend_no ≤ 7 and Age ≤ 24.500 then fake-celebrity (0 / 0 / 28 / 1)
if Traceable ≤ 31.500 and Comment_no > 9.500 and Age ≤ 26.500 and Comment_no ≤ 199 and Comment_no > 25.500 and Friend_no > 121 then fake-invented (0 / 1 / 1 / 26)
if Traceable ≤ 31.500 and Age > 36.500 and Age ≤ 67 and Age > 48.500 and Expressive ≤ 94 then fake-celebrity (0 / 1 / 20 / 2)
if Traceable ≤ 31.500 and Comment_no > 8.500 and Gender = Male and Expressive ≤ 73 then fake-invented (0 / 0 / 0 / 12)
if Traceable ≤ 31.500 and Gender = Female and Photo_no > 0.500 then fake-celebrity (2 / 1 / 80 / 45)
if Photo_no > 16.500 and Comment_no > 93.500 then real-local (2 / 46 / 0 / 2)
if Comment_no > 7.500 and Popular > 27 and Traceable ≤ 40.500 and Sociable ≤ 29 and Comment_no > 55.500 then fake-invented (0 / 1 / 0 / 15)
if Friend_no > 11 and Comment_no ≤ 50.500 and Valid ≤ 85.500 and Friend_no ≤ 77.500 and Traceable ≤ 22.500 then fake-invented (0 / 0 / 1 / 13)
if Traceable ≤ 22.500 and Active ≤ 86 and Friend_no ≤ 202.500 and Expressive > 73 and Age ≤ 27.500 then fake-celebrity (0 / 0 / 7 / 0)
if Valid ≤ 84.500 and Comment_no ≤ 59 and Expressive > 81 and Popular > 19 and Friend_no ≤ 74.500 then fake-invented (0 / 1 / 0 / 7)
if Traceable ≤ 22.500 and Age > 36.500 and Age ≤ 65.500 and Comment_no ≤ 29 and Active ≤ 86 then fake-celebrity (0 / 0 / 7 / 0)
if Comment_no > 48.500 and Photo_no > 21.500 then real-local (0 / 10 / 0 / 0)
if Valid ≤ 85.500 and Age > 27.500 and Friend_no > 172.500 then fake-invented (0 / 0 / 1 / 6)
if Traceable ≤ 22.500 and Age ≤ 18 and Comment_no ≤ 27 and Gender = ? then fake-celebrity (0 / 0 / 0 / 0)
if Sociable ≤ 12.500 and Age > 30.500 and Age ≤ 65.500 and Age > 34.500 and Age ≤ 46.500 then fake-celebrity (0 / 0 / 9 / 2)
if Age ≤ 25.500 and Age > 18.500 and Comment_no > 55.500 then real-local (0 / 11 / 1 / 0)
if Valid ≤ 85.500 and Age > 22.500 and Gender = Female and Valid > 81.500 then fake-invented (0 / 0 / 0 / 5)
if Traceable ≤ 22.500 and Age ≤ 18.500 and Comment_no ≤ 27 and Valid > 90 then fake-celebrity (0 / 0 / 4 / 0)
if Age > 25.500 and Active ≤ 67 then fake-invented (0 / 0 / 0 / 5)
if Age ≤ 17.500 and Friend_no > 61.500 then fake-celebrity (0 / 0 / 3 / 0)
if Valid ≤ 85.500 and Traceable ≤ 31.500 and Photo_no > 12 then fake-invented (1 / 0 / 1 / 4)
if Traceable > 31.500 and Valid ≤ 78 then real-local (0 / 6 / 0 / 0)
if Age > 25.500 and Age ≤ 27.500 and Profile = private and Gender = Male then fake-celebrity (0 / 0 / 1 / 0)
if Friend_no ≤ 4.500 and Expressive ≤ 98 and Expressive > 81 then fake-celebrity (0 / 0 / 4 / 0)
if Valid ≤ 81 and Expressive > 94 and Valid ≤ 69 then fake-invented (0 / 0 / 0 / 5)
if Age ≤ 26.500 and Gender = Female and Valid > 93 then real-local (1 / 16 / 7 / 2)
if Gender = Female and Age > 22 then fake-invented (0 / 3 / 10 / 13)
if Age ≤ 24 and Popular > 19 then fake-invented (0 / 0 / 0 / 5)
if Photo_no > 4.500 and Age > 29.500 and Comment_no ≤ 32 then fake-celebrity (0 / 0 / 7 / 1)
if Blog_no > 0.500 and Friend_no ≤ 79.500 then real-local (0 / 8 / 0 / 1)
if Photo_no > 8.500 and Age > 25.500 then fake-invented (0 / 0 / 0 / 3)
if Age ≤ 27.500 and Comment_no > 13.500 then fake-celebrity (0 / 0 / 3 / 0)
if Expressive ≤ 83 and Sociable > 12.500 then real-celebrity (4 / 0 / 0 / 0)
if Active ≤ 86 then real-local (0 / 3 / 0 / 0)
if Age > 65.500 and Age ≤ 98.500 then fake-invented (1 / 0 / 0 / 3)
if Age > 33.500 and Age ≤ 80.500 and Age > 57 then fake-celebrity (0 / 0 / 1 / 0)
if Age ≤ 17.500 and Gender = Male then fake-celebrity (0 / 0 / 1 / 0)
if Valid > 78.500 and Age ≤ 27.500 and Age > 20.500 and Gender = Female then fake-celebrity (0 / 0 / 1 / 0)
if Valid ≤ 90 and Age > 19 and Age ≤ 63.500 then real-local (0 / 4 / 0 / 0)
if Age ≤ 33.500 then fake-invented (2 / 2 / 3 / 6)
if Age > 50.500 then real-celebrity (2 / 0 / 0 / 0)
if Age > 41.500 then fake-celebrity (0 / 0 / 1 / 0)
if Profile = private then real-celebrity (1 / 0 / 1 / 0)

correct: 1152 out of 1300 training examples.

Appendix D: X-validation process with XML file




```
<operator name="Root" class="Process" expanded="yes">.
    <operator name="ExampleSource" class="ExampleSource">.
        <parameter key="attributes" value="\\home-fileserver\rf41\WindowsProfile\My Documents\train_private.aml"/>.
    </operator>.
    <operator name="DecisionTree" class="DecisionTree">.
        <parameter key="keep_example_set"   value="true"/>.
    </operator>.
    <operator name="ModelWriter" class="ModelWriter">.
        <parameter key="model_file" value="\\home-fileserver\rf41\WindowsProfile\My Documents\private1.mod"/>.
    </operator>.
    <operator name="XValidation" class="XValidation" expanded="yes">.
        <operator name="OperatorChain" class="OperatorChain" expanded="yes">.
            <operator name="ModelApplier" class="ModelApplier">.
                <list key="application_parameters">.
                </list>.
            </operator>.
            <operator name="Performance" class="Performance">.
            </operator>.
        </operator>.
    </operator>.
    <operator name="ExampleSource (2)" class="ExampleSource">.
        <parameter key="attributes" value="\\home-fileserver\rf41\WindowsProfile\My Documents\test_private.aml"/>.
    </operator>.
    <operator name="ModelApplier (2)" class="ModelApplier">.
        <list key="application_parameters">.
        </list>.
        <parameter key="keep_model" value="true"/>.
    </operator>.
</operator>.
.
```

Appendix E: Confusion Matrix comparing different learners over both original and pre-classified data

Pre-classified data

Decision Tree: Accuracy 84.64%

	true real-celebrity	**true real-local**	**true fake-celebrity**	**true fake-invented**	**class precision**
predicted real-celebrity	380	10	15	10	91.57%
predicted real-local	9	66	3	4	80.49%
predicted fake-celebrity	17	20	369	98	73.21%
predicted fake-invented	11	22	70	198	65.78%
class recall	91.13%	55.93%	80.74%	63.87%	

Rule learner: Accuracy 83.81%

	true real-celebrity	**true real-local**	**true fake-celebrity**	**true fake-invented**	**class precision**
predicted real-celebrity	383	28	16	12	87.24%
predicted real-local	6	45	9	15	60.00%
predicted fake-celebrity	18	31	379	116	69.67%
predicted fake-invented	10	14	53	167	68.44%
class recall	91.85%	38.14%	82.93%	53.87%	

Nearest Neighbours: Accuracy 82.43%

	true real-celebrity	**true real-local**	**true fake-celebrity**	**true fake-invented**	**class precision**
predicted real-celebrity	358	16	12	6	91.33%
predicted real-local	2	42	0	0	95.45%
predicted fake-celebrity	25	47	384	132	65.31%
predicted fake-invented	32	13	61	172	61.87%
class recall	85.85%	35.59%	84.03%	55.48%	

Naïve Bayes: Accuracy 80.58%

	true real-celebrity	**true real-local**	**true fake-celebrity**	**true fake-invented**	**class precision**
predicted real-celebrity	372	20	18	11	88.36%
predicted real-local	21	63	34	29	42.86%
predicted fake-celebrity	18	19	346	132	67.18%
predicted fake-invented	6	16	59	138	63.01%
class recall	89.21%	53.39%	75.71%	44.52%	

Original Data

Decision Tree: Accuracy 66.97%

	true real-celebrity	**true real-local**	**true fake-celebrity**	**true fake-invented**	**class precision**
predicted real-celebrity	333	28	36	17	80.43%
predicted real-local	18	46	31	42	33.58%
predicted fake-celebrity	11	25	281	86	69.73%
predicted fake-invented	14	5	84	145	58.47%
class recall	88.56%	44.23%	65.05%	50.00%	

Rule learner: Accuracy 64.98%

	true real-celebrity	**true real-local**	**true fake-celebrity**	**true fake-invented**	**class precision**
predicted real-celebrity	371	18	9	5	92.06%
predicted real-local	0	24	0	0	100.00%
predicted fake-celebrity	46	76	448	302	51.38%
predicted fake-invented	0	0	0	3	100.00%
class recall	88.97%	20.34%	98.03%	0.97%	

Nearest Neighbours: Accuracy 63.39%

	true real-celebrity	**true real-local**	**true fake-celebrity**	**true fake-invented**	**class precision**
predicted real-celebrity	324	25	9	3	89.75%
predicted real-local	23	49	47	50	28.99%
predicted fake-celebrity	24	9	241	89	66.39%
predicted fake-invented	5	21	135	148	47.90%
class recall	86.17%	47.12%	55.79%	51.03%	

Naïve Bayes: Accuracy 62.36%

	true real-celebrity	**true real-local**	**true fake-celebrity**	**true fake-invented**	**class precision**
predicted real-celebrity	352	26	13	5	88.89%
predicted real-local	14	18	1	4	48.65%
predicted fake-celebrity	50	74	441	300	50.98%
predicted fake-invented	1	0	2	1	25.00%
class recall	84.41%	15.25%	96.50%	0.32%	