

# **Online Dating in a Social Media Framework**

by

Allen Thomas Varghese, B.Tech

A dissertation submitted to the University of Dublin, Trinity College

in partial fulfillment of the requirements for the degree of

**Master of Science in Computer Science**

**University of Dublin, Trinity College**

August 2014

## Declaration

I declare that the work described in this dissertation is, except where otherwise stated, entirely my own work and has not been submitted as an exercise for a degree at this, or any other University.

Signed : \_\_\_\_\_

Allen Thomas Varghese

28 August 2014

## **Permission to Lend and/or Copy**

I agree that Trinity College Library may lend or copy this dissertation upon request.

Signed : \_\_\_\_\_

Allen Thomas Varghese

28 August 2014

## **Acknowledgements**

Many thanks to Dr Mukta Prasad for the support and guidance in completing this dissertation. Thanks is also due to the Mobile and Ubiquitous Computing course director - Dr Ciarán McGoldrick for timely assistance and advice. I am also grateful to the participants of the online study that was carried out as part of this research work, for taking time out of their busy schedules to use the application and provide feedback. The technical staff from the School of Computer Science and Statistics (SCSS) was very helpful with the administration aspects of the web server for the application. I am very grateful to my family for the moral support and encouragement.

ALLEN THOMAS VARGHESE

University of Dublin,  
Trinity College  
August 2014

# **Abstract**

## **Online Dating in a Social Media Framework**

University of Dublin, Trinity College,

2014

*Supervisor* : Dr Mukta Prasad

The aim of this research is to create a dating algorithm in the context of social networks such as Facebook. User preferences and behaviour are planned to be learnt from a social network instead of asking a user to fill a detailed manually-designed questionnaire. Various levels of user information would be captured automatically based on permissions set by the user. Additionally, information is also captured from exemplars i.e. desirable people on Facebook and celebrities, that the user feels they relate to. There would be a web based application to implement the dating algorithm. The UI design would incorporate Human-Computer-Interaction aspects to engage the user effectively and to capture the maximum amount of data possible. Psychological studies like Myers-Briggs evaluation, would be used to find a quantitative representation for people and for creating a baseline to compare different users. To validate findings, feedback from past dates of different users would be sourced and the necessary changes would be made to the dating algorithm.

# Contents

<b>List of Figures</b>	<b>ix</b>
<b>List of Tables</b>	<b>x</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Goal . . . . .	3
1.2 Challenges . . . . .	4
1.3 Motivation . . . . .	5
1.4 Contribution . . . . .	6
1.5 Overview . . . . .	7
<b>2 State of the Art</b>	<b>8</b>
2.1 Introduction . . . . .	8
2.2 Academic research . . . . .	10
2.2.1 Dataset analysis . . . . .	10
2.2.2 Recommender systems . . . . .	13
2.2.3 Analysis of user profile and activity . . . . .	16
2.2.4 Psychological analysis and evaluation . . . . .	24
2.3 Online dating websites and apps . . . . .	26
2.4 Patents . . . . .	29
2.5 Summary . . . . .	30
<b>3 Design</b>	<b>31</b>
3.1 Overview . . . . .	31
3.2 Personality Type . . . . .	31
3.2.1 Overview . . . . .	32
3.2.2 Model Definition . . . . .	34
3.2.2.1 Myers-Briggs personality model . . . . .	34
3.2.2.2 Big Five personality model . . . . .	37
3.2.2.3 Fisher personality model . . . . .	38
3.2.2.4 Personality Similarity Score . . . . .	39
3.2.3 Summary . . . . .	39

3.3 Exemplars . . . . .	40
3.3.1 Overview . . . . .	40
3.3.2 Model Definition . . . . .	40
3.3.3 Summary . . . . .	45
3.4 Social Information . . . . .	45
3.4.1 Overview . . . . .	46
3.4.2 Model Definition . . . . .	46
3.4.3 Summary . . . . .	49
3.5 Collaborative Filtering . . . . .	49
3.5.1 Overview . . . . .	49
3.5.2 Model Definition . . . . .	49
3.5.3 Summary . . . . .	51
3.6 Combined Model . . . . .	51
<b>4 Implementation</b>	<b>55</b>
4.1 Introduction . . . . .	55
4.2 Architecture . . . . .	56
4.2.1 Project Organization . . . . .	57
4.2.2 Accessing Celebrity Information . . . . .	59
4.2.3 Accessing Facebook Information . . . . .	59
4.2.4 Model Implementation . . . . .	62
4.3 UI Design . . . . .	63
4.3.1 User Interaction . . . . .	63
4.3.2 Layout . . . . .	64
4.4 Functionalities . . . . .	65
4.5 Challenges and Solutions . . . . .	73
<b>5 Evaluation</b>	<b>77</b>
5.1 Users . . . . .	77
5.2 Experiments . . . . .	77
5.2.1 Experiment 1 - Celebrity Choices . . . . .	78
5.2.1.1 Overview . . . . .	78
5.2.1.2 Hypothesis . . . . .	78
5.2.1.3 Procedure . . . . .	78
5.2.1.4 Analysis . . . . .	78
5.2.1.5 Alternate Approach . . . . .	83
5.2.1.6 Conclusion . . . . .	86
5.2.2 Experiment 2 - Similarity with Facebook friends . . . . .	86
5.2.2.1 Overview . . . . .	86
5.2.2.2 Hypothesis . . . . .	87

5.2.2.3	Procedure . . . . .	87
5.2.2.4	Analysis . . . . .	87
5.2.2.5	Conclusion . . . . .	89
5.3	Summary . . . . .	90
<b>6</b>	<b>Conclusion</b>	<b>91</b>
<b>7</b>	<b>Future Work</b>	<b>93</b>
<b>A</b>	<b>Research Proposal</b>	<b>i</b>
<b>B</b>	<b>Ethics Approval</b>	<b>iv</b>
<b>C</b>	<b>Big Five Personality Test</b>	<b>v</b>
<b>D</b>	<b>Myers-Briggs Personality Test</b>	<b>viii</b>
<b>E</b>	<b>Fisher Personality Test</b>	<b>xiv</b>
<b>F</b>	<b>Dating Websites in Ireland</b>	<b>xvii</b>
<b>G</b>	<b>International Dating Websites</b>	<b>xxi</b>
<b>H</b>	<b>Application Data Storage Format</b>	<b>xxv</b>
<b>I</b>	<b>Abbreviations</b>	<b>xxx</b>
	<b>Bibliography</b>	<b>xxxix</b>



# List of Figures

3.1	Myers-Briggs personality type and scores . . . . .	36
3.2	Cognitive functions . . . . .	41
3.3	Celebrity matching . . . . .	45
3.4	Collaborative Filtering . . . . .	50
3.5	Combined Model . . . . .	52
4.1	Architecture Diagram . . . . .	57
4.2	Project Organization . . . . .	58
4.3	Graph API Explorer . . . . .	60
4.4	Javascript Disabled . . . . .	64
4.5	Cookies Disabled . . . . .	64
4.6	UI Layout Structure . . . . .	65
4.7	Introduction . . . . .	66
4.8	Privacy and Ethics . . . . .	66
4.9	Contact information and application feedback . . . . .	67
4.10	Personal Information . . . . .	69
4.11	Big-Five Score . . . . .	69
4.12	Myers-Briggs Score . . . . .	70
4.13	Fisher Score . . . . .	70
4.14	List of Facebook Friends . . . . .	71
4.15	List of Celebrities . . . . .	71
4.16	Recommended Profiles . . . . .	72
4.17	User Ratings . . . . .	74
4.18	Date Feedback . . . . .	74
4.19	Application Invite . . . . .	75
5.1	Celebrity Data . . . . .	83
5.2	Experiment 1 - New approach . . . . .	86
5.3	Experiment 2 - Facebook friends data . . . . .	89
7.1	Mobile version . . . . .	94

# List of Tables

2.1	Dating applications that use Facebook or available on a mobile device . . . . .	27
2.2	Categories of dating websites . . . . .	28
3.1	Myers-Briggs personality types . . . . .	33
3.2	List of matching personality types . . . . .	43
3.3	Personality Type Compatibility Matrix . . . . .	44
3.4	Facebook Features and Personality Factor . . . . .	47
5.1	Experiment 1 - Celebrity Data . . . . .	79
5.1	Experiment 1 - Celebrity Data . . . . .	80
5.1	Experiment 1 - Celebrity Data . . . . .	81
5.1	Experiment 1 - Celebrity Data . . . . .	82
5.2	Experiment 1 - New approach . . . . .	84
5.3	Experiment 2 - Facebook data . . . . .	87
5.3	Experiment 2 - Facebook data . . . . .	88
5.4	User data statistics . . . . .	90
D.1	Scoring table for Myers-Briggs test . . . . .	xii
F.1	Dating websites in Ireland . . . . .	xvii
G.1	International Dating Websites . . . . .	xxi
G.1	International Dating Websites . . . . .	xxii
G.1	International Dating Websites . . . . .	xxiii
G.1	International Dating Websites . . . . .	xxiv

# Chapter 1

## Introduction

The emergence of internet has changed the world. It can be compared with the impact the invention of “wheel” had on the ancient civilizations. Ever since the internet became mainstream, there has been great changes to the way society and economy works on a day-to-day basis. Connectivity is everywhere and knowledge is power. Having access to the right data at the right time has become very crucial to the proper working of the economy. Content is king and is being consumed by everyone in a staggering scale. It is now expected to see a digital version of almost anything that happens in real-life. With access to a camera through low-cost smartphones, anyone can shoot a video and post a report of anything that happens almost instantly. In fact almost anyone who is well connected within the data economy of today, has an active digital profile. The everyday activities and events that happen in their life may be reproduced in the digital form to be seen by anyone they permit or give access to. Such groups of people are termed “social networks” and there has been many platforms that provide facilities and easy access to such networks the prominent being Facebook [19], a web based software platform that was started in 2004. It reached a membership of 1 billion users in 2012 [20]. This platform allows almost any kind of information to be shared with anyone including emotions, brand preferences and places visited. This shows how much real-life incidents can be translated to the digital world.

With 71% of people today using social networking websites [41], it is only a matter of time that dating would be woven around social networking information. Most of the websites today allow user to login to their websites using social login APIs. There are two advantages to this. This is an authentication mechanism of sorts which can be traced back in case of any misuse or fraud. Secondly, some information about the activities and events related to the user can be downloaded and analyzed in the digital form with the data received from social APIs. Users have also taken up on social logins as it helps them to avoid having to remember many login credentials for different websites and also they control the extent of personal information that is shared by changing the access levels for different categories of information like personal, work,

checked-in locations, movies watched, books read, photographs, group memberships etc. The social information gathered through social APIs can be used to match people based on common interests or by analyzing the sentiment in posts made to the personal social network as done by Five Labs [31]. They also create a personality profile of the user and allow them to compare it with any celebrity or any friend in their social network to see how different they are. Social information can also be used to auto-populate information in the user profiles so that the user saves time in filling details their profile in one or more websites.

The wide reach of social networks has enabled people to be in touch with others irrespective of spatial and temporal boundaries, all within a click or touch on an electronic communication device. As many aspects of our lives have gone digital, relationships have also taken that path. Dating as a socializing activity has also gone the digital route. People are now turn to dating channels like websites, smartphone apps and dating services in much bigger numbers as compared to the 2000s [42]. With internet connectivity and smartphones becoming ubiquitous, people have chosen to use online dating websites and smartphone apps to discover people whom they can go on a date with. With the search functionality, a huge list of profiles can be found that match exact preferences. The wide number of choices available has empowered people to be selective and thus go only on dates which they feel are really good for them. Today, it is estimated that the dating industry is worth \$2 billion [13] world-wide.

Due to the large number of people going the digital route, the cost of customer attention is also low which has led to growth of dating websites all over the world irrespective of language or country. It is estimated that there are 1400+ [39] dating websites across the world. Then there is the question of who is a match for someone and what is the extent of match between two people. This question is still debated as there is no explanation that is accepted by everyone. People have tried to solve this problem in different ways, some by manual intervention by offering personalized services catering to individuals, some by automating the matching process by having a matching algorithm that pours through the profiles and comes up with matches based on the preferred partner attributes specified in the dating profile and some have tried to create a hybrid of these two processes. Each of these approaches have had varying levels of success as per results of the surveys conducted by different dating websites. In all the mentioned scenarios, people have to fill out long questionnaires about themselves, their reaction in different scenarios, answers to specific questions etc for better matches. The number of questions typically range from 50 to 400+ depending on the website. This has been found out by creating a profile in many dating websites. Few of the websites like eHarmony [17], zoosk [55], chemistry [9] etc which claim to match people using scientific methods or dimensions of personality, have questionnaires.

Finding the right match for an user and reducing the complexity of the process for the user to find a match in a dating website are the two major issues that any commercial dating website tries to address which is evident by the steps they have taken such as limiting the number of profiles that can be seen per day [55][38] and automatic profiling of users to find recommendations [55][38]. Matching algorithms try to learn user preferences from their profile features and activity

on the website to propose matches, which may or may not be acceptable to the user. In this case, there also arises the problem of too many matches being shown to the user who then has to filter or go through the long list of matches to shortlist the preferred profiles. This process can be cumbersome and very tiring. This has been solved to some extent by using faceted search where a filters can be applied real-time on all possible matches. Keeping the number of filters to a required limit is necessary as humans have a sensory threshold for comprehension and too much information is off-putting and can force the user to leave the website altogether. Limiting the number of recommended matches per day also increasing the attention span of the user and reduces cognitive fatigue. Dating websites have included these considerations from time to time by studying the browsing habits of users and then incorporating feedback on such changes.

## 1.1 Goal

There are three goals that are planned to be completed.

*(a) Create an online dating algorithm that uses machine learning for improved prediction*

The machine learning component would learn from data of the subscribed users to improve the matching process and thus make better predictions that result in long-term relationships. The mathematical model for representing people would be chosen such that it is easy to implement and improve upon. The model would undergo change as the number of users increases and more input parameters are considered. There is also the possibility of using more than one mathematical model for user representation so that more features of a person can be analyzed and used in the matching process.

*(b) Use “exemplars” to understand preferences*

People are traditionally used to filling out long questionnaires in dating websites. Exemplars<sup>1</sup> reduce effort for the users by asking them to use their cognitive and intuitive perceptual capabilities to choose examples of people who have the personality features of their desired partner.

---

<sup>1</sup> Online Merriam Webster dictionary defines “exemplar” as a person or thing that deserves to be copied.

Exemplars can be of different types like people they know, celebrities, work colleagues, lifestyle brand preferences etc. The choice has to be made depending on what data is available for different exemplars and how that can be used within the algorithm in an easy and extensible manner.

*(c) Utilize social APIs for a holistic view of a person*

Most people have a significant digital presence today with a large chunk of their real-life interactions reflecting in their digital profiles. Such social information that gives a more accurate definition of an user's activities can be used to deduce the preferences of a person to a certain degree of accuracy. For example. If a person checks into a chinese restaurant, an assumption can be made about the person having a preference for chinese cuisine. But if there are multiple check-ins in different chinese restaurant, then it can be confirmed that the user has a preference for chinese cuisine. Such deductions can be also be used to validate the personal information entered by the user in their dating profiles. Matches can thus be proposed by finding users that have the same interests based on the assumption that people prefer to hang out with others who have the same tastes.

## **1.2 Challenges**

The main challenges faced in developing an online dating algorithm are :

*(a) Proprietary dating algorithms and locked datasets*

To develop any algorithm, the main requirement is to have enough data to test all possible scenarios including the boundary conditions, if any. Building a new data set takes a lot of time and effort. User data from existing dating websites are also locked away. So the entry barrier for validation of new techniques is very high. Also to compare the performance or design the algorithm, having a reference model speeds up the design process. Almost all of the matching algorithms used in commercial websites are proprietary, the information is not available. This has an impact on the time taken for design and validation of a new algorithm.

*(b) Biased user information*

Online dating websites provides more choices to users for finding the right partner. Because of the wide range of choices available, it becomes very hard to find out the profile that suits individual preferences. So sometimes users tend to exaggerate about themselves often creating the “perfect” version of themselves and also about their partner preferences. This inaccurate representation about themselves would prevent a new dating algorithm from proposing the right matches as validated later when they actually rate the dating experience. Such data can be hard to find and curate unless it has been reported. The most common features that are lied about are height weight and age [33]. Privacy concerns can also motivate people to withhold information or in some cases distort information as in most dating websites.

*(c) Choosing the “correct” representation of a person*

To create a dating algorithm, the model to represent a person has to be correctly chosen to fit the dating environment. There is no one right way to do it but would depend largely on the data available and other parameters. For example. In a dating website that has sparse user data, collaborative filtering (CF) approach for recommending matches would not work whereas a direct matching process based on partner preferences would work. When the amount of data increases, the model could be changed to work on CF or machine learning methods.

### **1.3 Motivation**

Finding the “right” partner is still a problem if we just consider the number of dating website available. As per a 2012 US survey by Harris Interactive on eHarmony [17] - one of the prominent dating websites with a membership of 21 million users, accounts for only around 4% of new US marriages. Considering Ireland, the number of divorced people in Ireland has increased by 150 per cent since 2002 [8]. As of April 2011 [8], single people over the age of 15, men account for 44 per cent and women account for 39 per cent of that figure. It is evident that people are actively searching for the right partner in spite of setbacks and thus worth trying to solve the “right” partner problem.

In terms of online dating, 59% [42] of Americans are open to online dating with 23% [42] having met a long term partner or spouse through dating websites. This is a positive sign in terms of adoption of new trends in the realm of online dating. In most of the dating websites, matches are proposed in a semi-automatic fashion or in a completely automated manner. Many algorithms

has been developed to find the best match as claimed by many of the large dating websites which is usually based on a proprietary method or technology. Most of these algorithms work on the premise of mutual interests leads to a match. There are also theories that dissimilar people are more preferable for a match. Each of the dating websites have their own method of predicting a match based on one of these theories. The most common form of recommending matches is by checking for matching profiles based on partner preferences given in the user profile. All attributes in the target user profile has to satisfy the partner preferences before it is recommended for a match. Since there could be cases where a comparison for an exact match would yield no results, there would be a category of possible matches where 10 out of 13 attributes matches. For such recommendations to work, it is required of the users to fill in lengthy questionnaires and also share a lot of personal information before a match is recommended.

Even though there has been attempts to include machine learning techniques to reduce the burden on the user and to provide better matches based on the profile activity of the user, these are not definitive approaches. The user would not always be surfing desired profiles and in such cases, the recommendation algorithm would learn a wrong set of user choices. Having the user provide choices by not being too restrictive would help tap into the range of choices that is acceptable. This can be done by requesting the user to highlight friends or celebrities that they admire or have traits that they desire in their date/partner. Such an approach removes the user from the constraint of having to think too much in an analytical manner and help them make natural choices as they would while describing their desired partner to another human in a conversation.

## 1.4 Contribution

The proposed contribution is to develop an online dating algorithm that has :

1. *Psychological profile similarity measurement* - Psychological profiles of users would be created from results of a personality test or from the details given in the user profile. The features from the psychological profile would be used to generate a similarity score to propose a match between users.
2. *Exemplar based prediction* - Suitable exemplars have to be defined which can be easily recognised by users and that can also be used to generate a similarity score for comparing two users. Celebrities is an exemplar that has wide appeal and has information readily available.
3. *A recommendation system that includes social information for prediction* - A system that can obtain social information about an user from the Facebook platform and creates a profile for computing similarity score for comparing users. The recommendation system takes into account the similarity scores for the psychological profile, celebrity choices and Facebook profile to calculate the final similarity score. This is then used to propose matches in the decreasing order of similarity score.



## 1.5 Overview

This report presents the research that has been conducted to create a new online dating algorithm based on exemplars and social information. A web application would be created to gather information from users, propose matches and gather feedback on the quality of proposed matches.

*Chapter 2* presents the academic and commercial research available in the area of online dating. This chapter also discusses the ups and downs of various approaches.

*Chapter 3* presents the design of the dating algorithm and the various conditions that have been considered.

*Chapter 4* presents the implementation aspects of the algorithm and how the application was developed. Screenshots of the application are shown and functionalities explained.

*Chapter 5* presents the evaluation of the algorithm. Two experiments are conducted to evaluate the effectiveness of the algorithm to recommend matches.

*Chapter 6* presents the conclusion of the research and the findings from the evaluation section.

*Chapter 7* discusses what can be done to improve the research done so far.

# Chapter 2

## State of the Art

This section is a discussion on the existing research and progress achieved in the area of online dating. The research that has been investigated has been divided into different categories for better classification and their impact discussed. The list of existing websites and their categorization based on customer model has been listed. Finally, a look at the patents that has been granted for online dating related activities on a website is also discussed.

### 2.1 Introduction

The problem of proposing matches is one which has many solutions. The solution can change based on the data available, the environment in which the matching process is done and then based on the commercial opportunity. Each of the solutions available today is influenced by these factors. Recommendation systems have been around for some time now with a strong focus in the retail domain, aimed at increasing the purchasing options for users. Companies like Netflix and Amazon have pioneered this approach and has had considerable success. With the rapid growth of online dating, the recommendation techniques that have been used in the retail domain were tailored for online dating. The approaches listed below have been found by researching how major dating websites work :

- *Attribute Matching* : This is the most widely used of all methods. In online dating websites, each person goes through a registration process where they provide personal details and partner preferences. The features of a user thus collected are divided into categories like must-have, binary or range based. An implicit feature vector is then created for each user and this set of values are used to compare different people and come up with recommendations. A cosine or Karl Pearson based similarity function is usually used for calculating a similarity score for proposing matches.

- *Activity Profiling* : Once a person has registered on a online dating website, they then have the capability to search for profiles, initiate a conversation, view more details, get recommendations etc. The various activities are profiled to understand the user preferences automatically since not all attributes/features can be specified explicitly. The information thus obtained is fed back to the implicit model and used for improving the quality of recommendations.
- *Collaborative Filtering (CF)* : This is a method that has seen wide applications in retail e-commerce websites like Netflix<sup>1</sup> and Amazon<sup>2</sup> for driving sales by recommending products that has been purchased by similar people. This approach can be applied for dating websites where there are a large number of users with ratings or some other factor that links different users is available. The idea here is that people that someone likes would be interesting to a person with similar tastes. A similarity function is defined based on a statistical model and the computed values are then ranked to come up with a list of recommendations. This model yields accurate results when the amount of data available is large. For small datasets, the prediction fails. Also the person under consideration should have sufficient nodes surrounding it and have enough incoming/outgoing links for being a person of interest. This method has found much success in recommending items on e-commerce website like Amazon, eBay etc.
- *Machine Learning* : This method is usually used as the second step of CF to automatically learn from the data available thereby providing an element of dynamicity in predictions compared to generating a recommendation only when a request is received. The data collected from users is usually divided into a training set, testing set and a validation set. The algorithm parameters that are generated using the training set are tested using the testing set and then improved using the validation set. This is a continuous improvement process as the data sets keep changing when new users join the network or the existing users update their information from time to time.

The above mentioned approaches have been analyzed by different researchers (*section 2.2*) and they have found different methods of exploiting these individually or with a combination of these approaches to yield results. The optimization or new approach would be usually tied to the data set used and the particular environment under consideration.

---

<sup>1</sup> The “Netflix Prize” is a contest organized by Netflix to find methods to improve their recommendation algorithm. Source : <http://www.netflixprize.com/>. Last accessed on 27-Aug-2014.

<sup>2</sup> Amazon recommendation system works using user ratings for products. Source : <https://www.amazon.com/gp/help/customer/display.html?nodeId=13316081>. Last accessed on 27-Aug-2014.

## 2.2 Academic research

Academic, individual and commercial research in the area of online dating have taken different approaches to understand and analyze the perception and responses of people when they indulge in dating activities online. This is done by either analyzing anonymized datasets from dating websites, exploring effectiveness of existing recommender systems by using them on sample datasets, checking data available on profile, monitoring user activity and then define different stages in dating, and finally applying psychological research to match people based on compatibility in their personality profiles. Some of the research is discussed in sections 2.2.1 and 2.2.2.

### 2.2.1 Dataset analysis

Researchers have been provided anonymized data sets from different dating websites for analysis and have come to different conclusions. They have also put forward suggestions on how to improve the user acceptance rate using statistical methods.

Alsaleh et al (2011) [1] explains an approach to give a better match for online dating websites utilizing the level of communication between different members by analyzing exchanges of e-mail, private messages and chat history. Links between different members that maintain contact are constructed in a graph like manner to identify whether communication between members are unidirectional or bidirectional. Similarity in profile features of members are taken into consideration during the initial stages to create local clusters of people. These clusters are then analyzed for the level of link quality and recommendations generated. There is also a confidence factor associated with recommendations to give a level of certainty. A recall measure is used to validate the correctness of matches recommended. Using the bi-link quality approach, the success rate is higher compared to the unidirectional approach where a match is defined on an invalidated request from the other end.

The dataset used contains information on 2 million members in which connection between nodes is very limited resulting in a sparse graph. This is a problem as it would not yield suitable matches due to the large number of nodes available in the data set. Creating local clusters can solve this problem to an extent as it helps create a better connected graph.

Chen et al (2011) [10] discusses about analyzing data of an Australian based online dating site using social network analysis methods. Graph theory is used to analyze relationships and heterogeneous relationships are only considered as around 97% of relationship preferences, is of straight nature. The level of communication exchange and the directionality is used to segregate users into various section of “bow-tie” structure. The core contains people with the highest levels of activity. The leafs signify users that have a very directional communication activity i.e. either

receiving messages or sending out messages. This kind of clustering is done using an algorithm for detecting strongly connected components. Connection between nodes in a social graph is very sparse and can be identified to some extent by the popularity factor of a person. When the number of nodes goes up like in the range of millions, it becomes very difficult to perform a walk through of the whole graph. The small clusters do not span beyond 2 or 3 connections and so looking out for outliers would not work. Content filtering based on predefined attributes and activity on the dating website was compared and found that there is a huge mismatch as people are not very sure of what they actually want.

The paper explored the analysis of the data set using graph theory and tried to come up with conclusions that can be used by an online dating organization to optimize their computing resources. User activity is an indicator of how serious a person is about finding a partner and bidirectional acceptance is taken forward by meeting up in the real world. The data found supports this fact. Recommendations are not effective to most of the users as they prefer to do a random walk and initiate a connection with a person that they feel suit their interest. This requires figuring out active nodes and building connections with these as origin.

Kunegis (2012) [34] uses complex number theory to recommend people to each other in the context of online dating. A dataset available from a Czech dating website was used and the ratings available between users was used. The rating value was used to deduce whether they were liked or disliked. Similarity between different users were defined if the like was reciprocated. A similarity function was defined by considering the connections between different users as a link in a graph. The levels of connections between different users were computed using an adjacency matrix, applying theorems from graph theory for unweighted graphs. By property of complex numbers, the weighted path count can be used to find similarity and likes by grouping the powers of the adjacency matrix based on whether they are even or odd. Another possible approach discussed was to use eigen vectors to arrive at a singular value decomposition of the adjacency matrix. The data from users was divided into a test set and a training set. To have sufficient data, the largest connected component was identified and users that have a path were used so that edges that are not connected do not turn up for the algorithm to rate. Standard algorithms for prediction like polynomials, hyperbolic sine, newman kernel, rank reduction and spectral extrapolation were used to test the effectiveness of the approach and it was found that in all cases irrespective of gender, this approach showed better results.

This approach has experimentally shown to work irrespective of gender which means it can be applied across both heterosexual and homosexual relationships. The only data point used to arrive at a conclusion of a match is the rating for an user that is given by others. The compatibility criteria between people cannot be restricted to one particular point and has to be explored at multiple levels. If this approach can be extended and adopted for other data points like personality types, with a pre-determined baseline for interpretation of results, then it holds scope for mass adoption. The link quality i.e. having a bipartite nature is essential for success which doesn't usually happen in online dating networks as it takes a lot of time for a normal user to gain popularity. Since most

of the users are passive in an online dating network, the recommendations would work only for a very small subset of users who are active.

Fiore (2010) [29] discusses the behaviour of people in an online dating website. The data was gathered from the users of an online dating website through the information in user profiles and from questionnaires. The messages exchanged between different users were analyzed and number of distinct users present in the communication was noted. This was used to determine the popularity ranking of the user. As the users could leave the site at any moment, the message exchange to activity ratio of the users were calculated on a per-day basis so that the inactive days were not included. This approach eliminates any bias for passive users due to inactivity. Contacts were paired and analyzed on six areas - reciprocation, number of messages sent, distance between users by postal code, duration of interaction, time taken for first reply and finally the number of other interactions initiated by the pair. The questionnaire was used to inspect online dating behaviour for around 1,100 users and had well-validated psychometric aspects that throw light into the personalities of the users. Big Five test was used here as it has been shown to be related to both number and quality of personal relationships. A general trust and caution scale was also used to measure the tendency of a person to act in a guarded manner with others. People with high trust have a low threshold for interaction with others in risky or uncertain environments. Various parameters were analyzed like matches made based on age, ethnicity, filtering of profiles based on keywords that signify emotion, trust and optimism, communication, initial contact and reciprocation. All measures were analyzed within  $p < 0.01$ . It was found that women generally tend to be more pickier and tend to stay within most of the factors given in the partner preferences. Older women tend to reply earlier than younger women. Communication between popular men and women took more time for a response compared to others.

Median was used instead of mean many scenarios to interpret results as outliers do not have an effect on the median. Also the hypothesis test had very high levels of comparison  $< 1\%$ . A binary logistic regression model was used for prediction which would work in this approach as the focus was on measuring whether there has been a response from both parties. The prediction was fairly accurate and could give the probability of response from the receiver on a day-to-day basis. The study was mostly analysis of the results based on predefined hypothesis and validated some of the established notions in evolutionary psychology. From a mathematical standpoint, no new approach was suggested nor any elaborate statistical model discussed. Few of conclusions that the authors arrived at are - women tend to be contacted much more than men, the rejection rate for messages sent to people who match the profile preferences tend to be low, the textual content in profiles have to be added carefully as people form assumptions depending on the keywords used along with the context, there is a preference for the same ethnicity and finally, slightly less popular people would be the right bet to elicit a response.

Chun et al (2008) [11] seek to analyze the various trends in an online social network that showed much promise in South Korea called Cyworld. The users are only from Korea and they had around 16 million users by 2007. This online social network had the usual profiles features along with

a guest book that was open to non-members. Anyone could comment on a guest book and do not require login credentials. This was one of the reasons why this online network became very popular. The interactions between members were classified based on incoming message, outgoing message and the number of friends. The members were all presumed to be real users as they had to give a personally identifiable code that was accepted in South Korea. The activity volume was considered as a network and analyzed using graph theory. Each of the users were considered as nodes and the edges representing messages shared between users. The graph was a weighted graph with weights representing the number of messages between two users. There could be in-degree and out-degree associated with each node depending on the level of interaction. Analysis of logs regarding user activity was classified into tuples of 3 - owner, writer and message. It was found that around 39% of posts were self-posts which could be a general message to all or a reply to one of the posts. The possibility of spammers cannot be avoided. Cyworld has put in controls to detect and blacklist such users. Activity in logs was predominantly influenced by the friend network and the pressure to be online was dependent on the number of friends. This was found to taper off after 200 friends. Disparity factor of the network for different nodes showed that users tend to interact with a certain number of users more frequently than others. The rest of friends tend to be passive and have a very small activity.

The paper tries to draw parallels between the offline friend network and the activity of the same users on an online social network. Authenticity is tried to be enforced by using validation of government documents. The validation process of users is not detailed in the paper though. So there could be chances of fake profiles being created due to fake data. In order to increase the user reach, an open guest book functionality was included which allows anonymous posting of information. This can be a source of security vulnerability and also an opportunity for spammers to increase popularity ratings since the volume of posts has an impact on the network weights. Dividing the user activity into 3 factor tuples for analyzing the user activity graph was a good approach as it helps in groups of nodes during clustering and thus shows disparity very clearly by virtue of spacing between different nodes. The filtering of posts made by the user in the guest book based on the nature of the post is not done. This might not have been done as context based analysis of the message is required.

### **2.2.2 Recommender systems**

Recommender systems try to propose recommendations based on clustering, collaborative filtering or machine learning techniques. People with similar tastes to a user are first found and their matches are then analyzed for compatibility to the source user. In this case, a form of second level matching process is used to find out compatible matches. This form of approach has found better success in predicting better matches rather than direct comparison of user preferences. This

approach suffers from the “cold-start” problem<sup>3</sup> as a recommendation can only be done if and only if there is a critical mass of users.

Yu et al (2013) [51] takes a different approach to the usual collaborative filtering technique. The user interactions are taken into account for determining reciprocity. Binary values – 0 or 1 are used to represent this interaction. The response to an interaction is not analyzed. What matters is whether there has been an interaction. Each interaction is mapped onto a contact matrix and each value would be a vector that signifies whether a request of interest has been sent from one user to another and vice-versa. Having a set of values helps to deduce whether interactions had been one sided or whether there has been a complete rejection. Even when the response is one-sided, the data is still significant as it would be used to give recommendations at different levels. The possible combinations are pre-defined, ranked and then a similarity score is computed using a cosine function. Normalization is applied to avoid any bias. A penalty factor is also used to ensure that partial matches are penalized and a high level of match and reciprocity is rewarded. Different metrics were used for evaluating the performance. The value of the penalty factor was varied to see the impact on performance. A value of 0.6 was chosen as it dominates the pure collaborative filtering method and also gives great improvements when interests are reciprocated.

The proposed algorithm works in an undirected and unweighted graph. Even when collaborative filtering is applied, there is no mention about the cold-start problem that is usually faced by this method. For passive users, This method has the problem of clustering of recommendations around popular users since high levels of activity on a dating network like messaging and viewing profiles, can lead to more recommendations as the data set is rich for these users. The introduction of penalty score to stabilize the variation of results can vary from one data set to another. This value in the paper was chosen and corrected to fit results. A self-learning mechanism has to be available to counteract this manual process. More users leads to more data for unilateral predictions and this can be used for passive users or new users to do the initial set of predictions. Once sufficient level of data is achieved i.e. a pre-determined threshold based on the minimum data set required for collaborative filtering, the switch to the actual algorithm can be done.

Nayak et al (2010) [37] explores a combination of methods to provide a recommendation rather than just using the approach of “one-method-fits-all”. The profile information of the user, the expected partner profile, the activity on the network and the willingness to keep up activity on the network is tracked. A “successful” relationship can be defined if a potential match reciprocates an invitation of interest. For the proposed algorithm to provide recommendations, it requires input from multiple layers. There has to be sufficient number of users that have made successful reciprocation in order to apply the similarity function. The attributes from the personal vector for each user are assigned weights based on the frequency of occurrence in profiles. The weighted sum is then used to find out the similarity score. The user for which matches are to be provided is compared to the dataset of users and similarity is computed. If the value is sufficiently high, it can

---

<sup>3</sup> Cold start refers to the situation when a recommendation system fails to generate recommendations due to insufficient number of ratings or links available in the dataset.



be concluded that the user would also be interested in the relationships that are currently formed with the matched user. When there is insufficient data like at the onset of an online dating site, a different approach is used. K-means clustering is done based on gender and then comparisons are done using the centroid profiles for initial matches. A cosine based similarity function is used in this case. Once similarity values are computed for similar users and related matches, these values are combined together by applying weights that are determined empirically to arrive at the final match scores.

In this paper a dataset from an online dating website in Australia was used to arrive at conclusions. No actual data was collected from users on whether the recommendations provided by the algorithm, actually matched the expectations of the user. The weights assigned to different attributes depends on the number of users that have indicated these preferences in their profile. Thus there is a consideration of the common interest of a large group and these weights could change from time to time resulting in varied recommendations. To arrive at the final match, weights are assigned which are determined empirically. This can change from network to network and has not been defined clearly in the paper. Also the proposed algorithm would work well in small relationship based networks to generate recommendations in a reasonable time frame. Even though matches are proposed, the compatibility aspect between the proposed matches and the user is not evaluated.

Pizzato et al (2013) [44] deals with reciprocal recommenders. These work by taking into account details from the explicit profile, the implicit profile and the user profile. The explicit profile contains the list of desirable features as specified by the user. The implicit profile contains details about the user activity in the dating website which are mapped to pre-determined user behaviours. Then there is the user model that encompasses the various parameters of the reciprocal recommender, analyzes features from both the explicit and implicit model thus making the final decision for a match. The success of recommendation thus depends on how strong these three models are. In an online dating scenario, there are two kinds of users - active and reactive. Active users are those who are proactive in communication and are not deterred by rejection whereas reactive users respond only to incoming communication. Rest of the time they just browse through profiles. For active users (stereotypically men), the explicit profile and implicit profile would be rated higher compared to the reactive users (stereotypically female) as their online activity is very high. Profile features also play an important role in engagement of users. The more complete a profile, the more interested others would be. Also long textual profiles receive more communication from other users. The proactiveness of a user can be defined as a ratio of difference between expression of interest (EOI) sent and received to the total number of interests. Users can then be divided into 5 categories - highly reactive, mildly reactive, equally proactive/reactive, mildly proactive and highly proactive based on the proactiveness quotient. The acceptance or rejection of EOIs can be used to fine tune recommendation. There is a positive compatibility score between two users when there is a similarity in the people that both users like and a negative compatibility score when there is a similarity in the people that both user dislike. If the overall score is greater than 0.5, it signifies

that the users tend to be compatible with each other, else not.

The proposed reciprocal recommender works by analyzing the number of expression of interest (EOI) received and the acceptance rate for different users. Using this information, users are divided into different categories and then analyzed. No new solution has been proposed for the cold-start problem of collaborative filtering. Either direct feature match of explicit profiles are done or content based filtering is used till there is sufficient data. Even when matches are found, unless both users reciprocate positively, the match is not acknowledged. Such positively matched profiles are then used as the basis for further matching and generation of recommendations. Also mismatch of important attributes between users that have occurred due to oversight can also increase the rate of EOI rejection. Targeted inclusion of reactive profiles and ensuring that recommendations to the same user are not repeated, are some of the extra steps taken to ensure that recommendations are correct. The time span available for deducing information on users is very limited. User activity decreases drastically after the first 4 weeks with the first week, the most active. This could be because they might have found a date or found a suitable person whom they married. This phased user backouts further aggravates the cold-start problem.

Pizzato et al (2010) [45] discusses the RECON recommender that was developed as part of work done for an Australian dating website. The algorithm predicts a match in two steps : matching based on the provided attributes (explicit profile) and then based on the level of activity like viewing the profile and sending messages. The acceptance of a message is duly noted as a preference and the number of rejections is also used to tailor the recommendations. The number of preferred attributes are deduced from the activity and added to the implicit profile of the user. If no communication has occurred between two users (cold-start problem), the match of preferences and attributes of two users are computed called reciprocal score. It is assumed, a new user can be paired with anyone in the system. Success and recall for recommendations are computed based on the number of recommendations that had exchanged messages and were reciprocated positively.

The required to run the algorithm varies on the type of user. If it is a new user, then the time taken would be comparatively lower than generating recommendations by running the full algorithm. Then there is the problem of some popular users having a very high reciprocal score compared to the passive users, thus skewing the number and quality of recommendations. The cold-start problem has been discussed thus ensuring that everyone gets a set of recommendation. The quality will then vary based on activity on the dating website. Each user is assigned an implicit model that is representative of the internal working of the system. Any modifications can be easily made and thus provide a personalized recommendation for an user.

### **2.2.3 Analysis of user profile and activity**

User profile and activity analysis is an interesting way of finding out preferences of someone as it requires minimum user input in most cases. The users of the system are given full freedom

to explore the network of users and then rate them based on their preferences. The user trail, choice of profiles, ratings given out, communication established etc are used to create a profile automatically by the system. This profile can change over time depending on the nature of the user activity. In addition, the user can also make changes to their profile based on changing interests or circumstances. These facets of user profile is then used to come up with suitable recommendations. The drawback of this approach is that it requires sufficiently large amount of user activity and profile information to create a profile as pieces of data at some point of time might not reveal the true nature of the individual. Here more of machine learning principles come into force.

Zhao et al (2008) [53] aims at exploring the motivation of people in creating or “living out” an imaginary personality in the digital world within the context of a social networking platform called Facebook. The physical barriers are no longer relevant in the digital space. Anyone can create a persona of whatever fantasy they want to be. Introverts can become extroverts, slightly deformed people can become beautiful, shyness can turn into bravery etc. There is an element of anonymity that provides a liberating feeling for people to express themselves. The facebook platform does not provide a space for complete anonymity. It provides an anchored relationship with the offline world and ensures as much ground truth is covered as possible by tying in users with verifiable offline entities like education institutions, friends, places etc. There is the option to describe themselves through a summary and also to provide contact personal details like date of birth, contact number, sexual orientation, preferred books, movies, sports, music etc. In spite of the facility to provide a bundle of information about themselves, users also have the option to restrict the visibility of certain information for others depending on groups like friends and public. This enables people to create certain personas that are targeted to only a certain group of people. In this study, the participants profiles were thoroughly studied and different features available in profiles ranked manually. Only the publicly available information was scraped off the profiles. There has been cases where information was missing. It could either be because the user did not provide those information or the visibility of such information was restricted. Each of the participants were then interviewed to understand how far the digital personality projected through facebook represented the true identity of the user. It was found that user tend to suppress some information due to social and peer pressure. Also there was tendency to project implicit traits rather than expressing them explicitly.

This study was aimed at discovering and measuring how much the identity of a user was represented both online and offline. It was proved that users tend to migrate to a preferred identity in the digital world but in the process try to implicitly promote many features of their new identity. Validating different features can only be done with the help of anchoring with offline or real world entities. Extending this to the online dating scene can be a problem as such anchoring is not welcome as people prefer to share personal information during face-to-face meetings. The extent of information that is shared in an online dating profile is mostly explicit. These days many online dating websites ask users to include features that validate their claims. Those users who are aware

of such social personification, can easily detect such exaggerated additions to the personality of someone from the pictures posted on the profile, the keywords used in the description, their choices in book, music, movies, sports etc and from related social media profile linkages. The extent of damage to the social identity of a person in case of finding out the deception is not covered much in this study.

Terveen et al (2005) [48] is an extensive study about the various social matching approaches for online dating. Various claims are put forward regarding proposing a match for online dating like willingness of users to share sensitive personal information, social matching with explicit preferences is the right way, use of social networks to perform social matching, maintaining a balance with introduction between users and disclosure of varying levels of information, number of registered users in a network, methods to promote user interaction with better design of the website, capturing user feedback that can be used effectively to improve the algorithm and a strong focus on achieving user goals. The mentioned eight claims were investigated thoroughly and research questions for each of the claims were identified that could be further analyzed to detail. In the light of the various claims that were investigated and the questions that came up for further review, social matching can greatly improve the way in which online dating as an experience can be improved and scientific data collected. The experiences across a community can be used to generate or shape personalized user interactions for greater user retention, user choices and of course, finding the preferred partner. While recommenders give more freedom to users by helping to filter out potential choices, social matches gives the user a way to explore information in the context of their social circle and in the process consume information that they couldn't have thought of in the first place.

The analysis mostly follows a Q & A format. A context is defined, few questions are put forward and then a conclusion is drawn from the existing evidence on whether the premise was right or not. There was no mathematical approach taken but rather seen through the filter of a philosophical debate to arrive at a conclusion. There was a lot of facts and figures used to arrive at various conclusions though. Another aspect is the lack of research done in the field of social matching even when there is a lot of potential for such systems in terms of spreading collaboration and improving interactions between different groups within an organization and beyond. The inputs for social matching can be anything related to a particular context and can be determined based on the type of user interacting with that particular environment. This flexibility enables people to make choices based on specific goals and then try to push social matching to achieve those goals.

Diaz et al (2010) [14] proposes an approach that involves matching a set of predefined features and then ranking them based on the similarity level of values. Each profile of the user contains a description of preferred attributes and then a set of preferences that are binary or range based. The features available on profiles differs from person to person and thus has to be checked every time a match is being performed. For binary features, the values are XORed and for scalar attributes, the absolute difference is used for computation. Match queries can be run on binary as well

as scalar values. The attributes would be represented in binary values wherever possible and a category assigned to each to ensure that a match with higher compatibility is recommended. User activity on the online dating website is taken into account for assigning a relevance factor to the final recommendation. Using machine learning with a linear regression model, the values of some attributes can be predicted when they are not given in the profile. Ranking is then decided using decision trees that are constructed using the results from running the feature matching methods. Thus when two users have the same features highlighted in their profiles, the chances of them being matched is higher.

Since the algorithm works using features detected in the user profile, a more expressive profile has a higher chance of being matched with a suitable profile than otherwise. Also if the user is a paid member, then more data about the user would be made available for viewing indirectly stating that the user is interested in a relationship and is ready to spend time and effort. Scanning text description and other text for subtleties, is novel as it can highlight interests that can be overlooked. Having a large amount of features, assigning them as binary features and then querying them could result in an overhead in a real-life scenario. Also unless semantic querying is applied, marking the data cannot yield useful information for better prediction except for creating a linked query.

Fiore et al (2008) [28] deals with assessing how attractive an online dating profile is to people and what conclusions are drawn by checking out a profile. Sample data was extracted from Yahoo! personal website with an equal proportion of men and women. The entire profile was extracted as one image and then were split into different sections using adobe photoshop. These images were then showcased to the participants invited from different ethnic backgrounds using a custom website that anonymized the information about the profiles and users. Participants were asked to rate the images on a scale from zero to four and an associated confidence level from one to three. This ensures that no one is forced to make an uncertain decision at any point of time. They also had the option to not answer a question if they choose to. The sections shown to the participants included the profile picture, description, personal attributes, lifestyle preferences etc. There was a time limit of 50 minutes within which they had to rate as much profile entities as possible. The data set consisted of 50 unique profiles. A linear regression model based on least squares was created to analyze the results on different dimensions, attractiveness of the complete profiles from just the dimensions and the overall attractiveness based on the sections. The confidence ratings were more than 70% across participants indicating that they were moderately confident of their choices. It was concluded that in evaluating potential partners through online dating profiles, physical attributes are considered highly but equal importance was also given to the textual sections like description and other attributes as well. Also having a profile picture is very important as 85% responded that they would not even consider a profile that does not have a profile picture.

The ethnicity of the participants were predominantly Asian but the data set of profiles had very few Asians. Also the participants were not active in the dating scene. There arises questions on whether the participants did actually understand what it meant to be in a relationship and is in a position to evaluate someone from that standpoint. There is also the possibility that they

were doing a comparison with themselves on various aspects rather than an objective judgement. Then there is the question of how can attractiveness be defined by a few pre-determined attributes since dating or building a relationship is a very personal experience and is different for each and every person. Moreover attractiveness has to be reciprocated. So calculating the average across the dimensions would be the stepping stone towards further analysis of how to define more compatibility parameters. In this study all profiles had profile pictures. Without the pictures, would the participants be interested in a profile if the details of the person were given in a constrained environment is an interesting question and was not answered in detail. There is evidence that people tend to fill in optimistic thoughts about someone in a constrained scenario.

Hancock et al (2007) [33] in their study compares the profiles of 250 online users by comparing the ground truth with the self provided information in their online dating profiles. Participation was solicited through ads in a classifieds portal and out of the 479 people who responded, 251 people were chosen for the study. Only those involved in a heterosexual relationship were included. These people had profiles in 4 major online dating networks - Match.com, American Singles, Yahoo Personals and Webdate. There is an over representation of young people and under representation of older people (above 60). The participants were given a printed record of their online dating profile and asked to rate the accuracy of information especially about their height, weight and age on a scale of 1 to 5 with 5 being the most accurate. If they choose not to give an answer, then they are asked whether they would give an answer if questions on these 3 features were mandatory. After recording their feedback, the ground truth is established by checking their height, weight and age by direct measurement or validation from records like driver's license. A linear regression model was used to analyze the variations in these parameters across gender. It was found that irrespective of gender, shorter participants tend to overestimate their height. A similar trend was also found for heavy participants. Discrepancies in age was more common in older people but was not significant in this study due to under representation. The deviation between self-assessed values and the actual values was found to be intentional irrespective of gender. The study concludes that the concern about online deception is only partly justified as the level of deception in height, weight or age is not easy to detect face-to-face. Exceptions were also found in this study where there was more than acceptable deviations in the three factors.

The number of users for the study is very small to derive any meaningful significance from a statistical perspective. Even then, it helped prove some of the common assumptions about lying prevalent in online dating. It was proved that women lie are more likely to lie about their weight and men about their height. Age was pretty much the same. There is a bias in the results as there were much more younger people compared to older people. Other variables present in an online dating profile like education, ethnicity, family background, personal habits etc were not used for the study.

Fiore et al (2004) [27] gives a detailed overview about online personals. These ads have shed their image of only being only for those who are socially backward. Online matchmaking ads are increasing day by day. Some use it for seeking friendship, some for casual relationships and

mostly for a long term relationship maybe even marriage. The design of these systems have an influence on how we approach relationships over a period of time and slowly spreads to others who are not currently seeking a relationship. People tend to be attracted to those who have similar interests and tastes. As all traits of a person cannot be replicated online, there is a trade-off on what kind/number of features can be treated as a match. These ads can also be characterised as a sort of selling where we describe what we are with a bit of exaggeration, in some cases. These ads carry a description of the person which can tend to have very predictable answers over a period of time as more and more people jump on to the bandwagon. Some of the websites provide a search option to ferret out profiles that match exact characteristics that people are looking for. There are a variety of filters like age, height, weight, color, ethnicity etc. After a few profiles have been shortlisted, the next step is to initiate communication. many websites offer the opportunity for communication by either paying for them upfront or giving a few messages free and then pay for them. Some online dating websites have included personality tests and then try to match people based on the results. There are social network based sites where matches are made between members of the same network. These functions are available in all types of online personal systems being it mainstream or targeted for specific population.

Traditional meetups happened within a context which carried a certain degree of accountability and validation that came along with. So people tend to be more careful with the projection of their personality since relationships tend to sour if later found out that lies were spread by a person. Online dating websites spans across different social networks and can connect with anyone in the network. There are no security restrictions for visibility of information whatsoever. Details of dating websites that have been trying to include controls to restrict the amount of misrepresentation has not been given in the paper. It is mentioned that users might have to exaggerate a bit in order for the algorithm to filter out the profile. These days it has become common that people tend to write/generate content that can be parsed much more easier by machines rather than people. There is also no thorough discussion about how information presentation and retrieval plays an important role in user retention and promoting activity in the dating website. Also since different networks use their own proprietary algorithms, people tend to have multiple online profiles. A comparative study on this aspect is also missing from this paper.

Bachrach et al (2012) [3] conducted a study to use a large set of users and data from their facebook profiles to provide an unbiased and more generalized view of how different features in a facebook profile has a direct correlation to the personality of a person. The user base consisted of 180,000 users who participated in the study through a facebook application. A personality questionnaire based on the Five Factor Model had to be filled. The profile data of the users were accessed through the application. Due to security restrictions, access to some data from the profiles were blocked. In spite of this, at least 15,000 data points were available for each data point. Six profile features were considered - number of friends, groups, likes, photos uploaded, status messages and tags in pictures of friends. The participants were arranged in descending order of the count of various features. After sorting, the users are then divided into 10 equal and disjoint sets. Thus

the group average would be a much closer representation of individual users personality of those falling into the particular category. Clustered scatter plots were created to present the relationship between facebook features and personality. The correlations were tested using a t-distribution test and all correlations were found to be at  $p < 1\%$ . Mann-Whitney-Wilcoxon test were used to check if the top and bottom thirds of the participants differed in terms of mean personality score. All relations were found to be significant at  $p < 1\%$ . For predictions, coefficient of determination was used. Thus the user would be assigned a percentile based result instead of an absolute value. On analysis it was found that the trait Openness had a positive correlation to number of likes, groups and status updates. Conscientiousness is negatively related to number of likes and groups but have a positive relation to number of uploaded photographs.

Multivariate linear regression has been used to provide a recommendation. Root Mean Square Error (RMSE) was also calculated. It is mentioned that output from other machine learning methods did not change the outcome significantly. It has been proved that by combining more than one feature can give a better prediction of the user personality rather than relying on just one feature. In spite of a large dataset and set of feature points, extraversion was the easiest to predict and agreeableness the most difficult. Security settings on the user profiles may block access to data relevant to make a prediction. Self-curation can also affect the prediction as users can be biased and create a profile with incorrect information. Activity on facebook has a significant impact on the variables involved in analysis. One of the assumptions is online activity in social media is a reflection of what happens in the real world. For passive facebook users, the predictions would have to be taken with a grain of salt as these users have relatively no activity at all. Non-participation in social media arising out of personal choices need not necessarily translate into a socially inept person.

Zhao et al (2012) [54] describes how people in a relationship manage personal information on social networking sites. Facebook is a social entity that reflects the offline behaviour of its users to a great accuracy. In the real world, when two people are in a relationship, the nearest friend circle gets to know about it first and then others. On facebook, linking another person's profile in a status message that states they are in a relationship, sends out an update to the complete network of friends/acquaintances immediately. Once a relationship is publicly announced, there is a sense of pressure on the members in the relationship to be careful in what they do on facebook like the nature of wall posts, the photos they upload, the people they are in contact with, the events they attend, the places they visit etc. Scrutiny has become much easier and this leads to frequent Q & A sessions in the relationship. Contextual information can be obtained from facebook very easily as all the information is tied to a profile. The links for the profile can be used to construct a graph of interconnections and provide references to various situations.

This paper clearly highlights the various issues issues faced by people especially those in a relationship and the dilemma that they are being forced into because of the need to mimic offline behaviour in the online world. Precision scrutiny of activities on facebook has become easier by the facility of tagging a person. This feature even though meant for viral marketing, has brought about



tension in relationships because of perception issues by the other person in the relationship. The paper also discusses means of resolving tensions and mitigating the effects of conflicts. The issues highlighted in the paper can be avoided to a great extent by controlling the level of information exposure from the offline world to the online world. Peer pressure and higher levels of introversion has lead to a sense of trust in information available online. Users have now started curating their information that is shared by changing the security measures for data that is shared thus shielding from prying eyes.

Ellison and Hancock [18] discusses the various trends prevalent in online dating websites regarding presentation of self. People have been found to lie about their physical features the most compared to others. The ability to easily create a digital identity and the need to be socially appealing pushes people into the habit of lying to others. Trying to find out someone in a restaurant when on a date and not being able to find that person because the profile picture and how they actually look are very different, is becoming common these days. Few methods of dealing with these kinds of inaccuracies has been outlined. Having people to put down their signature associates their identity to the information that they have supplied and this in turn would trigger a strong feeling to be true to oneself. Having images of eyes gives a mental impression that people are being watched and unwittingly forces people to give out the truth. Looking for clues that can be cross-referenced for authenticity is another way. This would take time and might require networking in the offline world in some cases. The subtleties in the presentation style outlined in a dating profile also has an impact on how people are perceived. Some nuances are grammar, style of writing, the references available for the profile and the quality of references, all lend a level of authenticity to a profile and are often used to filter out profiles.

Online dating profiles is a place where people lie about their physical features in the hope of getting a partner that fulfills their desire of an attractive companion. Most of these users when they create their profile, do not think about the consequences of being unmasked as a liar. There is also an impact on the person who has invested time and money only to find out what they have come to believe is a pack of lies. This article analyzes from a psychological standpoint, the various attributes that people lie about and the methods to filter out effectively the truthful profiles. The ideas outlined can be used by people to have a better online dating profile and to be truthful as possible.

Fox and Warber (2013) [32] studied the impact of Facebook as a social networking site has on relationships for the generation that grew up along with the rise of Facebook. People are very active on SNSs today compared to 10 years ago. Since everyone is connected to large group of people in these networks, information spreads very fast. This applies to the status of a relationship as well. Before the rise of SNS, if someone is in a relationship or got married, it would take quite some time for the information to reach many people depending on the speed of post or occurrence of social event where information is shared face-to-face. It has been found in the study that women tend to take the relationship status on Facebook much more seriously than men. On the other hand, men did not consider a Facebook status as a strong sign of commitment and so are typically

slow in changing their relationship status.

The differences in perception about how much of a role SNSs have in the life of men and women can create problems in relationships with women tending to be suspicious if men do not respond with a Facebook status confirming the relationship. This can be due to the fear of women about men pursuing other partners. Thus monitoring profile activity of a person in conjunction with others can yield a much better picture of a person's life than just analyzing a person's social profile. This is especially true of events that are posted in SNSs about a person's life.

#### **2.2.4 Psychological analysis and evaluation**

Psychological analysis is an interesting approach taken by researchers to deep dive into the fundamentals of what makes a person tick. Even though this branch of study is still considered as speculation by many critics, it is considered to give a holistic view of the personality of anyone. Recommendations can be proposed based on similarities in personality profiles. There are many standard tests to assess personality and a general interpretation of the different types of personalities have been commonly agreed on. There is also a list of compatible matching personality types which are usually used by different online dating websites to match people. Many variants to the list of matching personality types have also been used to predict matches and such approaches have been considered to be the unique value proposition of these websites or apps.

Amichai-Hamburger et al (2010) [2] aimed to improve or re-validate the work that was done by Ross et al (2009) in analyzing the personality of users based on various features of their Facebook profiles. In the proposed model called the Five-Factor-Model (FFM), personality of any person can be defined on the basis of five traits - Neuroticism, Extraversion, Openness to experience, Agreeableness and Conscientiousness. The initial results were based on self-reports from the participants which could be biased. The proposed approach is to directly analyze the Facebook profile of different users and draw conclusions. The participants for this study were a group of 237 undergraduate students as they had higher levels of usage. Their personality features were first captured using a questionnaire which scored on a five-point scale. In the next step, the information from the facebook profiles of the participants were captured under four dimensions so that classification would be easier. The information was encoded using a scheme that had coverage over all the major facets in a facebook profile. Different numbers were allocated to different features depending on the value of the feature on the facebook profile. Labels were also provided for easy understanding. The encoding and labelling of information was done manually. The summation of all information uploads was also calculated. The contact information dimension was rejected as only four participants provided information. Statistical analysis using ANCOVA, Chi-square test and regression analysis was done on the top and bottom third of all personality domains. The dependent and independent variables were chosen according to the domain. In contrast to the previous study, it was found that for those who score high on conscientiousness, have more friends and picture

uploads than individuals who scored lower.

The participants of this study were undergraduate students from the same university and they almost knew each other. There could be an element of social or peer pressure that could have an influence on how they behave. The environmental factors that condition the participants behaviour cannot be easily captured as it requires contextual data over a period of time as people tend to also change their facebook profile depending on their social interactions. Approaches or methods to cross validate information available on facebook profile was not discussed in the paper. The authors formulated five hypotheses - one for each dimension of FFM and then compared it with the results of Ross. et al (2009) for any deviation.

Ross et al (2009) [46] has made an attempt to find the impact of how the personality of a person is projected in a social context over a computer mediated communication medium like social networking sites. Motivation to be part of a social circle and competency in different forms of technology based interaction, has in various degrees influenced the levels of engagement of various people. An abstraction of personality was put forward using a personality model called Five-Factor-Model (FFM) that encompasses the traits of a person within five categories - neuroticism, extraversion, openness to experience, agreeableness and conscientiousness. These expected behavior from various levels of these traits were hypothesized in the context of a social networking website particularly facebook where the engagement model tapers from offline to online mode. Metrics were defined based on the user utilization data of different features available in facebook. The study that was done over a 2 week period, had a skewed ratio of men and women - 1 to 5.5 and a participant group size of around 97, all undergraduate students from a Southwestern Ontario university. The statistical corrections like splitting the data set into three divisions and comparing trends between these sections are done to compensate for any potential bias in data. The results from the study show that the traits defined by the FFM are demonstrated to a large extent in the online world compared to the offline world. Thus a measure of online social activity can give a very reasonable picture of the personality of a person.

This paper tries to associate personality types to an user based on activity exhibited in a social network. There are many theories of personality types and each has its own merit based on the nature of the study. This approach adds a sense of context rather than just inferring attributes about an user based on statistical models. The data collected for this study, severely limits the ability to generalize over a broad spectrum of demography as results for this group can have statistical analysis limitations because it is representative of an age group from 21 – 27. If this is combined with information available on an online dating website, then generating a personality type in addition to the implicit profile created by the system, would add another dimension/attribute for comparison and thus better recommendations.

Paunonen and Hong (2013) [40] researched the extent to which actual personality-related traits matched with the assumed personality traits. They carried their study on a group of university undergrads who had known each other for over 6 months. Each of the participants were asked to take the NEO-PI-R test for assessing their traits. Later they were asked to take the test in

a peer review mode to see how much they project their personality to their peer. By comparing the scores for the same set of users both way, it was found that, with acquaintance, the level of assumed similarity between friends increases. There is also a tendency to choose people who are similar to themselves.

Since this study was done for a group of undergraduates, the sample is not representative of the population when it comes to online dating. A more extensive study with thousands of participants are required to validate these proposals. This study shows that validation from external entities is an important aspect in ensuring truthfulness of information about a person. Linking multiple datasets of information like social and professional, might provide insight into many aspects of a person's personality. A peer rating score could serve as the indicator of trustworthiness of a person.

### **2.3 Online dating websites and apps**

The core research ideas discussed in sections 2.2 have been commercialized through different channels mainly - websites and smartphone applications. There are more than 1400 [39] online dating related websites in the world today under different sub-categories. The below categories have been compiled by visiting the various websites available in Appendix F and Appendix G :

- *Casual encounters* which includes one-night stands and mature affairs
- *Age based* websites which deal specifically with a pre-defined age category like 30+, 40+, 50+, 60+ etc
- *Situational* based like single parents, smokers only, vegetarians, animal lovers etc
- *Community* based like jews, blacks, whites, christians etc. There is some form of curation either manually or automatically using image recognition. Users can also block or raise an issue with a profile if required.
- *Country* based. Most of the major online dating website like eharmony, zoosk and match have an international footprint and thus have websites specific for different countries where the website would be tailored to suit the culture and outlook of the residents. The language would also be different based on where the website is being accessed with the website displaying content based on the widely used language by default. Language and locale can also be switched if required. Usually the front-end of the website would be hosted in a country specific server, the middleware and backend would be running from a common location but serving location specific content.
- *Local dating services* : These are dating services that cater to a very specific section of people who meet stringent criteria and offers a completely personalized service ensuring the

subscribers have value for money. They usually have money back guarantees. Such services do not have a wide reach and are mostly local i.e. serving very specific regions of a country.

A list of dating websites in Ireland is available in Appendix F and some of the international dating websites are given in Appendix G. Some of these website claim to match people based on scientific and mathematical principles, with some also including information from social media like favorite books, TV shows etc to find same interest areas for matching people. None of these websites use exemplars and social information together and then match people based on personality profiles.

Smartphone apps have been developed for different platforms like Android, iOS, Blackberry and Windows Phone. There are other platforms like Symbian OS and BadaOS which does not have a mature application market gauging from the number of applications available for download. Some of them have a social media connection. The list of dating applications is available in table 2.1.

Table 2.1: Dating applications that use Facebook or available on a mobile device

Dating Application	Facebook/Web application*	Mobile application
Hinge	No	Yes
Tinder	No	Yes
CoffeeMeetsBagel	Yes	Yes
OkCupid	Yes	Yes
Zoosk	Yes	Yes
Match	Yes	No
MatchMachine	Yes	Yes

\* Has a website or Facebook canvas application that users can login using Facebook credentials

The huge popularity and reach of social media has caught the attention of dating companies and they have created apps targeting these platforms to leverage on the network reach of a person. SNSs like Facebook have opened up their platform through APIs and development kits to enable rapid development of apps. These apps can be published through a portal of discovery called the “app store”. To make the application more visible, notifications about various activities can be sent to the friend network of person which would instantly appear in their alerts stream. Friends can also comment on their activities as well as join the application.

The dating websites and apps can be divided into categories depending on their approach to dating. The categories are given in table 2.2 along with the strengths and weaknesses.

Table 2.2: Categories of dating websites

Type	Strengths	Weaknesses
<p>Long Term e.g. EHarmony, Match, OkCupid,...</p>	<ul style="list-style-type: none"> <li>– Strong user base for deploying machine learning,</li> <li>– Prior research in psychology/neuroscience used for matchmaking</li> </ul>	<ul style="list-style-type: none"> <li>– slightly static, changes rely on new findings in Psychology which are often conflicting.</li> <li>– sticking to a safe business model prevents innovation</li> <li>– tedious forms and isolated systems resulting in biased and incomplete view of user, his life and relationships.</li> </ul>
<p>Short-term e.g. Tinder, Hinge, Zoosk,...</p>	<ul style="list-style-type: none"> <li>– ability to grow viral quickly and gather a large user base</li> <li>– intuitive user interfaces that are easy to understand</li> </ul>	<ul style="list-style-type: none"> <li>– attraction based on superficial beauty based cues yield unstable, non-bijective mappings</li> <li>– visual similarity between people as a good date-ability indicator, is controversial.</li> <li>– unclear whether these apps are a front-end to only gather data or whether they intend to seriously improve matchmaking, also resulting in other ethical dilemmas.</li> </ul>
<p>Emerging applications eg : MatchMachine, Five Labs</p>	<ul style="list-style-type: none"> <li>– application claims to observe Facebook data for matching</li> </ul>	<ul style="list-style-type: none"> <li>– continues to have fixed hypotheses, that is based purely on some facebook usage statistics, instead of observing relationships or accounting for examples.</li> <li>– new, unclear how successful this is.</li> </ul>

Type	Strengths	Weaknesses
Research approach (SomeoneLikeThat)	<ul style="list-style-type: none"> <li>– social network as best online proxy to real life.</li> <li>– allow detailed information as well as example based intuitive input , from larger complex database</li> <li>– active, continuous learning about a user from feedback</li> <li>– Bayesian approach: priors from psychology + data-driven information per user.</li> </ul>	<ul style="list-style-type: none"> <li>– to address hesitations about privacy, we will need to make our system transparent and explain with clarity, to convince a user of the definite benefit.</li> <li>– will need critical mass of users to take off.</li> <li>– stiff competition from big companies which could produce similar, rival apps (before we patent), but as this is an alternative model rather than a mainstream one, we can carve our niche.</li> </ul>

The current research is targeted at the long term category by trying to find matches that are similar or are compatible on multiple levels thereby increasing the accuracy of the match.

## 2.4 Patents

Three patents were found assigned to online dating companies, which are given below :

- Zoosk [52] - Identifying potential matches based on their location and compatibility factor. Two people are said to be compatible if there is a match in the profile characteristics and there exists some interaction through messages.
- E-Harmony [6] - Use of a satisfaction estimator which indicates the level of satisfaction that a person has in relationships and using a neural network to then match people based on this indicator. The levels of communication are also considered while proposing a match.
- jDate [47] - This patent deals with sending notifications to people when there exists a match between two people in feelings or interests and maintains anonymity about each user till a match is found. Both the users have to reciprocate feelings or interests for a match to be identified by the system.

There could be more more patent filings that happened but not made public. The details of such filings would be considered when more work is done for this research. The patents listed here deal with specific tasks around users and is not related to exemplars or social information as envisioned in chapter 3.

## **2.5 Summary**

This chapter looked into the various research available in the field of online dating. Many approaches have been proposed that involves a combination of manual intervention and automated processing. Each of the approaches discussed has had varying levels of success. Criticism on each of the academic research papers is available in section 2.2. The relevance of online dating websites and apps is highlighted in section 2.3. A list of dating websites and apps, within Ireland and on an international level is available. This is by no means the complete list but highlights the websites that has huge membership numbers.



# Chapter 3

## Design

This chapter presents the design aspects of the online dating algorithm and the various factors that were considered. A mathematical model for representing users and a method of matching users is also discussed.

### 3.1 Overview

To create an online dating algorithm, the first step is to decide the method by which two users can be matched. The method chosen for this research is “personality type” or “personality model” as described in section 3.2. A match is decided by the similarity score which is defined as the squared euclidean distance between dimensions of the personality models for two users. This process of computing a similarity score can be then extended to exemplars and for social information.

### 3.2 Personality Type

This section discusses how a personality model can be defined to match two users and how it can be developed for quantitative analysis. Some background into the different personality tests and their interpretations is also provided.

### 3.2.1 Overview

For proposing a match between two participants, there has to be some common ground for a match. The personality of a person is the driving force in influencing others to associate with that person. So being able to represent personalities of different people in a quantifiable manner is a good way to approach matching. It has been found over and over again that people with similar tastes and choices in life, have a greater chance to be together. Internationally accepted personality tests like Big Five, Myers-Briggs and Fisher, provide insight into various aspects of someone's personality which can be quantitatively expressed as a multi-dimensional vector. A 100% accuracy of personality predication is not guaranteed by these tests but they do offer a very good idea of how someone would be.

The personality types of registered users can be captured/deduced by requesting the users to answer standard personality questionnaires - Big Five, Myers Briggs and Fisher. There are different versions of these tests available on the internet that have evolved across the years depending on various factors. There are often a short version and a long version of these tests. The original version of these tests are administered by different organizations and they offer detailed personality assessments on payment. Some of them even offer training and certification on how to administer such tests. There has been wide adoption of such tests in different industries to create project teams that get along very well with each other and also when people are chosen to head leadership positions where a lot of money is at stake.

The three personality type tests chosen for this study are :

1. **Big Five :** In psychology, the Big Five personality traits are five broad domains or dimensions of personality that are used to describe human personality. The theory behind the Big Five factors is called the Five Factor Model (FFM). The Big Five factors are Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism. The Big Five model is able to account for different traits in personality without overlapping. During studies, the Big Five personality traits show consistency in interviews, self-descriptions and observations. Moreover, this five-factor structure seems to be found across a wide range of participants of different ages and of different cultures. A summary of the factors of the Big Five and their constituent traits:

Openness to experience : It can be described as the comparison between creativity over consistency in approach to activities done in a routine manner. It is also called "Intellect" as it refers to the intellectual capabilities of a person..

Conscientiousness : It can be described as the level of efficiency and organization demonstrated compared to a careless or spontaneous attitude to life.

Extraversion : Described as the trait which shows whether someone is a people person or not. People high in this trait prefer to be in company of a group and seem to draw energy

from them. The opposite is called an introvert.

Agreeableness : The trait that measures how much a person trusts their logic compared to trusting their emotions in a particular situation when faced with making a decision.

Neuroticism : This trait shows how much a person is susceptible to nervousness when faced with different situations. It is also referred to as “Emotional Stability”. Those who score high in this trait tend to show signs of depression.

2. **Myers-Briggs** : The purpose of the Myers-Briggs Type Indicator® (MBTI®) personality inventory developed by Isabel B. Myers and Katharine Briggs[49] was to make the theory of psychological types described by C. G. Jung understandable and useful in people’s lives. They state that variation in a person’s behavior is quite orderly and consistent, due to basic differences in the ways individuals prefer to use their perception and judgment. Perception is about awareness of things, people, happenings, or ideas. Judgment deals with conclusions about what has been perceived. Based on the choices made by a user in 4 categories :

- (a) Preference for being outgoing which signifies *Extraversion(E)* or *Introversion(I)*
- (b) Information which deals with how data is interpreted by someone and signifies *Sensing(S)* or *Intuition(N)*
- (c) When faced with making decisions, having a preference for logic over situation signifies *Thinking(T)* or *Feeling(F)*
- (d) Openness to external information or options during a situation signifies *Judging(J)* or *Perceiving(P)*

The above 4 dichotomies gives rise to 16 personality types as given in table 3.1.

Table 3.1: Myers-Briggs personality types

ISTJ	ISFJ	INFJ	INTJ
ISTP	ISFP	INFP	INTP
ESTP	ESFP	ENFP	ENTP
ESTJ	ESFJ	ENFJ	ENTJ

3. **Fisher** : Dr Helen Fisher [16] distinguishes between four broad biologically-based styles of thinking and behaving which she associates with four broad neurochemical systems. Four personality types are defined for each neurochemical. They are :

*Explorer* for dopamine who tend to be curious and seek novelty.

*Builder* for serotonin tend to agree more with social norms and respect authority.

*Director* for testosterone tend to be good at engineering and spatial/mathematical tasks.

*Negotiator* for estrogen/oxytocin tend to be intuitive and good with people.

In a study [5] done by Dr Fisher across two different groups of users, it was found that these biologically influences personality traits are replicable. She has designed a personality test questionnaire that has been validated by the 13 million users of dating websites - Match.com and Chemistry.com. In 2013 [16], it was tested on 17 newly weds and 17 long-married people and the results confirm her theories.

### 3.2.2 Model Definition

The results from each of the personality tests described in section 3.2.1 have a personality type and the corresponding value for each dimension of the personality type. The values for each dimension ranges from 0 to 1. To compare two users, each of the personality test results is considered as a vector with multiple dimensions. Each of the type and value represent a dimension. Thus the Big Five test has 10 dimensions - 5 dimensions representing the types and 5 dimensions for the values, Myers-Briggs test has 8 dimensions - 4 dimensions representing the types and 4 dimensions representing the corresponding values and Fisher test has 8 dimensions - 4 dimensions representing the types and 4 dimensions representing the values. Also since the number of dimensions are different for different personality tests, normalization would be applied.

#### 3.2.2.1 Myers-Briggs personality model

For Myers-Briggs test, there are 2 possible values for each of the type dimension which is represented by the following type vector.

$$f_t^M = \begin{bmatrix} \mathbf{E} \text{ or } \mathbf{I} \\ \mathbf{S} \text{ or } \mathbf{I} \\ \mathbf{T} \text{ or } \mathbf{F} \\ \mathbf{J} \text{ or } \mathbf{P} \end{bmatrix}$$

There is only one corresponding value for each of the types represented by the following value vector.

$$f_v^M = \begin{bmatrix} \mathbf{v}_1 \\ \mathbf{v}_2 \\ \mathbf{v}_3 \\ \mathbf{v}_4 \end{bmatrix}$$

Based on the type, the value has to be treated positive or negative at each indice 'n'

$$f_{s_n}^M = K_n * f_{v_n}^M$$

where

*n = 1 to 4, the indices of various dimensions*

$$K_n = \begin{cases} \mathbf{1} & \text{where } f_{t_n}^M \in \{E, S, T, J\} \\ -\mathbf{1} & \text{where } f_{t_n}^M \notin \{E, S, T, J\} \end{cases}$$

The possible range of each of the values  $v_1, v_2, v_3, v_4$  without considering the sign, is on a scale from 0 to 1. To standardize with values of other test scores, the value is converted from a scale of -1 to 1 into the scale of 0 to 1.

This is done by the formula given below :

$$v_1^* = \frac{1 + v_1}{2}$$

$$v_2^* = \frac{1 + v_2}{2}$$

$$v_3^* = \frac{1 + v_3}{2}$$

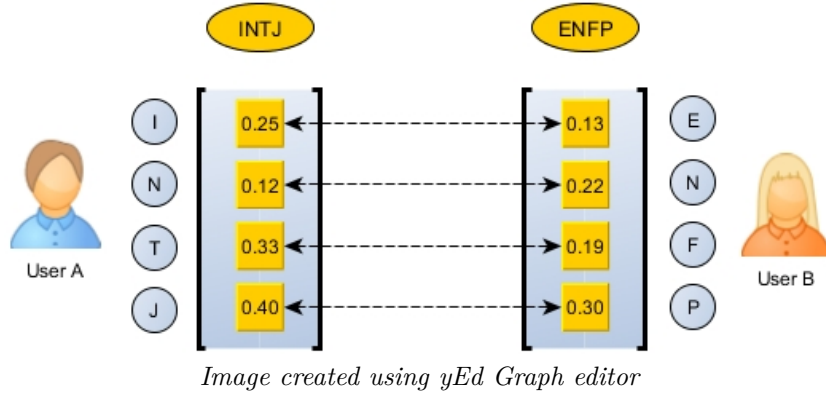
$$v_4^* = \frac{1 + v_4}{2}$$

Thus, the Myers-Briggs value vector becomes

$$f_v^M = \begin{bmatrix} v_1^* \\ v_2^* \\ v_3^* \\ v_4^* \end{bmatrix}$$

For two users say 'i' and 'j', the Myers-Briggs score vectors are defined as :

$$f_i^M \text{ and } f_j^M$$



Match between two users is defined in terms of the similarity score. Similarity can be calculated as the average squared euclidean distance between each of the dimensions in the value vector (*refer figure 3.1*). The distance between the dimensions is given by :

$$D^M(i, j) = \frac{\sum_{n=1}^{n=N} (f_{ivn}^M - f_{jvn}^M)^2}{N}$$

where

n = index of the dimension in the score vector

N = 4 since Myers Briggs has four dimensions for values

Thus the Myers-Briggs similarity score can be defined as :

$$S^M(i, j) = \exp(-D^M(i, j))$$

### 3.2.2.2 Big Five personality model

For the Big Five test, the scale of the values are in the range of 0 to 1. So no scale conversion is required. The type vector for Big Five is given by :

$$f_i^B = \begin{bmatrix} E \\ A \\ C \\ S \\ I \end{bmatrix}$$

There is only one corresponding value for each of the types represented by the following value vector.

$$f_v^B = \begin{bmatrix} \mathbf{v}_1 \\ \mathbf{v}_2 \\ \mathbf{v}_3 \\ \mathbf{v}_4 \\ \mathbf{v}_5 \end{bmatrix}$$

Match between two users is defined in terms of the similarity score. Similarity between two users 'i' and 'j' can be calculated as the average squared euclidean distance between each of the dimensions in the value vector. The distance between the dimensions is given by :

$$D^B(i, j) = \frac{\sum_{n=1}^{n=N} (f_{ivn}^B - f_{jvn}^B)^2}{N}$$

where

n = index of the dimension in the score vector

N = 5 since Big Five has five dimensions for values

The Big Five similarity score can thus be defined as :

$$S^B(i, j) = \exp(-D^B(i, j))$$

### 3.2.2.3 Fisher personality model

For Fisher test, the scale of the values are in the range of 0 to 1. So no scale conversion is required. The type vector for Fisher is given by

$$f_t^F = \begin{bmatrix} \mathbf{E} \\ \mathbf{B} \\ \mathbf{D} \\ \mathbf{N} \end{bmatrix}$$

There is only one corresponding value for each of the types represented by the following value vector.

$$f_v^F = \begin{bmatrix} \mathbf{v}_1 \\ \mathbf{v}_2 \\ \mathbf{v}_3 \\ \mathbf{v}_4 \end{bmatrix}$$

Match between two users is defined in terms of the similarity score. Similarity between two users 'i' and 'j' can be calculated as the average squared euclidean distance between each of the dimensions in the value vector. The distance between the dimensions is given by

$$D^F(i, j) = \frac{\sum_{n=1}^{n=4} (f_{ivn}^F - f_{jvn}^F)^2}{N}$$

where

n = index of the dimension in the score vector

N = 4 as Fisher has four dimensions for values

The Fisher similarity score can thus be defined as :

$$S^F(i, j) = \exp(-D^F(i, j))$$



### 3.2.2.4 Personality Similarity Score

In sections 3.2.2.1, 3.2.2.2 and 3.2.2.3, the similarity scores for different personality models have been found. An user can attend one or more personality tests. For a pair of two users, they need not necessarily attend the same test or in fact attend any test at all. The check for availability of scores for the same test for both users makes sure that users are being compared on the same scale. If a pair of users have not attended any test in common, then their combined similarity score would be zero. In all other cases, the combined personality similarity score can be defined as :

$$S^P(i,j) = \frac{\sum_{a \in (M,B,F)} e(a,i) * e(a,j) * S^a(i,j)}{\sum_{a \in (M,B,F)} e(a,i) * e(a,j)}$$

where

$$e(a,i) = \left\{ \begin{array}{l} 1 \text{ if user 'i' attended test 'a'} \\ 0 \text{ if user 'i' did not attend test 'a'} \end{array} \right\}$$

and

$$e(a,j) = \left\{ \begin{array}{l} 1 \text{ if user 'j' attended test 'a'} \\ 0 \text{ if user 'j' did not attend test 'a'} \end{array} \right\}$$

The individual similarity scores range from 0 to 1. So averaging the scores as above would ensure the combined personality similarity score lies between 0 and 1.

### 3.2.3 Summary

In section 3.1, the different personality tests that can be used were identified. In section 3.2, a mathematical model for representing the personality types and its corresponding values were discussed. A match between two users is based on a similarity score and defined in terms of a power of the mathematical constant  $e$ . For pair of users attending one of more tests, the match between them can be found by averaging the common personality similarity scores. The similarity score always has a value between 0 and 1.

### **3.3 Exemplars**

In section 3.2, matching users based on personality profiles has been discussed. In this section, the focus would be on choosing the right exemplar that can be used for the dating algorithm.

#### **3.3.1 Overview**

The choice of exemplar should satisfy two conditions :

- The exemplar should be widely recognized.
- There should be enough public data available on the exemplar so that people can verify the identity and validity of information.

One of the possibilities that satisfies the above conditions is “celebrities”. These are people who have made a mark in their chosen profession be it politics, entertainment, art, science, medicine etc. A peculiarity about the list of celebrities is that it is growing everyday with new people added in new professions. There is a lot of information about their personal lives available on the internet through biographies, news reports and self released social media information. People tend to associate themselves with celebrities, aspiring to be like them. They see some qualities in them which they try to achieve. Thus the choice of celebrities provides an indirect indication of the qualities that they might look for in their partner.

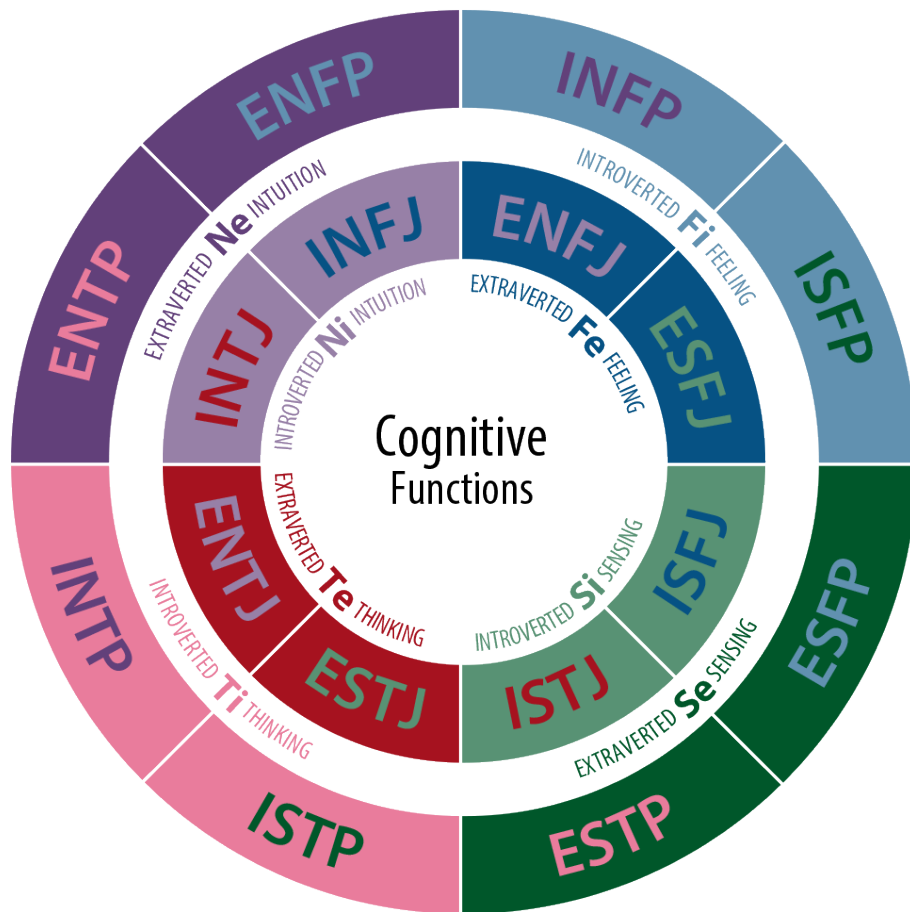
#### **3.3.2 Model Definition**

To capture and analyze the celebrity preferences, a list of celebrities [7] was compiled, that has the following information :

- Name - First, Middle and Last
- Alias
- Profession
- Myers-Briggs personality type : The personality types is determined by the information available on celebrities about their cognitive functions. This information is updated if a case is made that is backed by solid evidence.

Users are asked to choose the celebrities that they strongly associate with. The users can choose one or more celebrities. To find if a match exists between two users, the personality types of chosen celebrities are checked for compatibility. The various types can be represented as a graphic [12] shown by figure 3.2.

Figure 3.2: Cognitive functions



Extensive research and in-depth analysis about personality types has been done in this area by Paul D. Tieger and Barbara Barron-Tieger[50]. An easy to use list of matching personality types has been compiled by Jess Dods[15], a master administrator of the Myers-Briggs test. It is shown in table 3.2 where for each personality type, a corresponding list of best matches and compatible matches is available. The rest of the personality types out of the total 16 personality types, fall under the category of non-compatible types.

There are 16 personality types as per Myers-Briggs Type Indicator (*MBTI*) [49]. Each of the personality types are matched with the other types that are compatible. The data in table 3.2 can be translated in the form of a weighted matrix as shown in table 3.3. The best match is assigned a value of 1.0 and the next best match has an estimated compatibility of 50% less i.e. 0.5. The value of 0.1 has been given when a matching user type is not available. These values are arbitrary and has not been defined in any literature but chosen only for this research. To find compatibility between two personality types, a lookup function (*CC*) is defined which would fetch the compatibility score from table 3.3 for the specified pair of celebrity types. For similarity between two sets of points in space, Hausdorff distance states that the maximum value of distance should be considered. Thus for proposing a match between two users, the availability of best matches within the two list of celebrities is considered. If a match is found, then it is marked as compatible with a similarity score that is the maximum of scores for the best compatible match pairs available.

Let  $l_i^C$  and  $l_j^C$  be the list of celebrities for two users 'i' and 'j'. The celebrity vector  $l^C$  contains the list of personality types of the celebrities each user has chosen. Each of the personality types in  $l_i^C$  is checked for compatibility with each value in  $l_j^C$  (refer figure 3.3). The celebrity compatibility score can then be represented as

$$S^C(i, j) = \max_{a=1}^{a=n} \left( \max_{b=1}^{b=m} \left( CC(l_{ia}^C, l_{jb}^C) \right) \right)$$

where

$n$  = number of celebrities chosen by user  $i$

$m$  = number of celebrities chosen by user  $j$

$CC$  = Celebrity compatibility function defined as a one-to-many function whose value would

be  $\left\{ \begin{array}{l} 1.0, \text{ user } j \text{ type } \in \text{ best match} \\ 0.5, \text{ user } j \text{ type } \in \text{ compatible match} \\ 0.1, \text{ user } j \text{ type is not available} \end{array} \right\}$  as given in table 3.3

Table 3.2: List of matching personality types

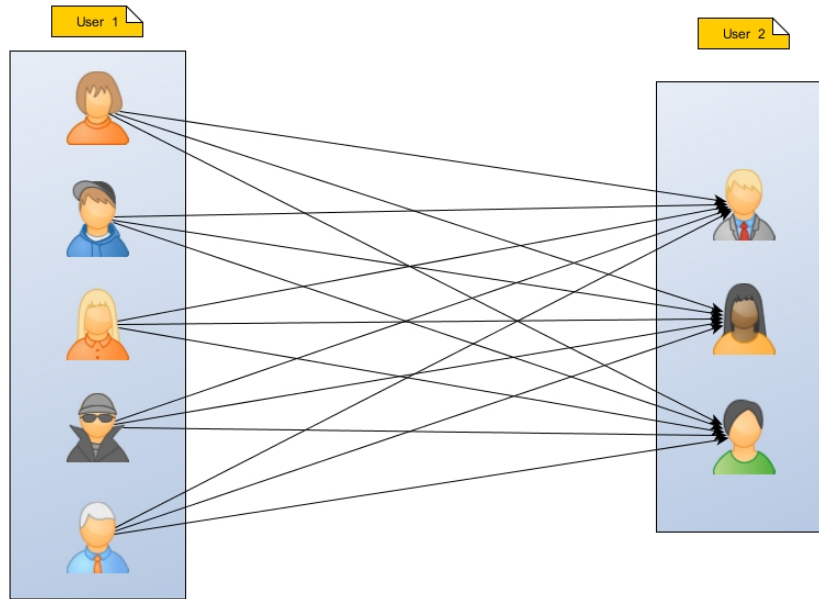
No.	Participant 1	Participant 2 (Best Match)	Participant 2 (Compatible Match)
1	ISTJ	ESTJ, ISTJ, INTJ, ISTP, ESTP	ENTJ, INTP, ENFJ, INFJ, ISFJ, ISFP, ENTP
2	ISTP	ESTJ, ISTJ, ENTJ, ESTP	ESFJ, ISFP, INTJ, ISFJ
3	ESTP	ISTJ, ESTP, ISTP, ESFP	ESTJ, ISFP, ENTJ, ENTP, INTP, ISFJ
4	ESTJ	ISTJ, ESFJ, ISFJ, ENTJ, INTJ, ISTP	ENTP, INTP, ESTP, ESFP, ISFP
5	ISFJ	ISFJ, ENFJ, ESTJ	ESFJ, ESTP, ISFP, INFJ, INFP, ESFP, ISTJ, ISFP
6	ISFP	ESFP, ISFP	ESTP, ESTJ, ESFJ, ISTP, ENFJ, INFJ, INFP, ISFJ, ISTJ, ENFP
7	ESFP	ESTP, ISFP	ESTJ, ESFJ, ISFJ, ESFP, ENTP, ENFJ, INFJ, ENFP, INFP
8	ESFJ	ESTJ, ENFP	ISFJ, ESFJ, ENFJ, INFP, ISFP, ISTP, ESFP
9	INFJ	ENTP, ENFP, INFJ, INFP, ENFJ	ISFJ, ESFP, ISFP, ENTJ, INTJ, INTP, ISTJ
10	INFP	ENFP, INFP, ENFJ, INFJ	ISFJ, ESFJ, ESFP, ISFP, ENTP, INTP
11	ENFP	INFJ, INFP, ENFJ, ENFP, ESFJ	ENTJ, ENTP, INTJ, INTP, ESFP, ISFP
12	ENFJ	ISFJ, ENFJ, ENTJ, INFJ, ENFP, INFP	ESFJ, ESFP, ISFP, INTP, ISTJ, ENTP
13	INTJ	ESTJ, INTJ, ISTP, ENTJ	INTP, INFJ, INFP, ENFP
14	INTP	ENTP, INTP, INTJ	ESTJ, ISTJ, ESTP, ENTJ, ENFJ, INFJ, ENFP, INFP
15	ENTP	ENTP, INTP, INFJ	ESTJ, ISTJ, ESTP, ESFP, ENTJ, ENFP, INFP, ENFJ
16	ENTJ	ESTJ, ISTP, ENTJ, ENFJ, INTJ	ISTJ, ESTP, ENTP, INTP, INFJ, ENFP

Table 3.3: Personality Type Compatibility Matrix

Type	ISTJ	ISTP	ESTP	ESTJ	ISFP	ESFP	ESFJ	INFJ	INFP	ENFP	ENFJ	INTJ	ISEJ	INTP	ENTP	ENTJ
<b>ISTJ</b>	1.0	1.0	1.0	1.0	0.5	0.1	0.1	0.5	0.1	0.1	0.5	1.0	0.5	0.5	0.5	0.5
<b>ISTP</b>	1.0	0.1	1.0	1.0	0.5	0.1	0.5	0.1	0.1	0.1	0.1	0.5	0.5	0.1	0.1	1.0
<b>ESTP</b>	1.0	1.0	1.0	0.5	0.5	1.0	0.1	0.1	0.1	0.1	0.1	0.1	0.5	0.5	0.5	0.5
<b>ESTJ</b>	1.0	1.0	0.5	0.1	0.5	0.5	1.0	0.1	0.1	0.1	0.1	1.0	1.0	0.5	0.5	1.0
<b>ISFP</b>	0.5	0.5	0.5	0.5	1.0	1.0	0.5	0.5	0.5	0.5	0.5	0.1	0.5	0.1	0.1	0.1
<b>ESFP</b>	0.1	0.1	1.0	0.5	1.0	0.5	0.5	0.5	0.5	0.5	0.5	0.1	0.5	0.1	0.5	0.1
<b>ESFJ</b>	0.1	0.5	0.1	1.0	0.5	0.5	0.5	0.1	0.5	1.0	0.5	0.1	0.5	0.1	0.1	0.1
<b>INFJ</b>	0.5	0.1	0.1	0.1	0.5	0.5	0.1	1.0	1.0	1.0	1.0	0.5	0.5	0.5	1.0	0.5
<b>INFP</b>	0.1	0.1	0.1	0.1	0.5	0.5	0.5	1.0	1.0	1.0	1.0	0.1	0.5	0.5	0.5	0.1
<b>ENFP</b>	0.1	0.1	0.1	0.1	0.5	0.5	1.0	1.0	1.0	1.0	1.0	0.5	0.1	0.5	0.5	0.5
<b>ENFJ</b>	0.5	0.1	0.1	0.1	0.5	0.5	0.5	1.0	1.0	1.0	1.0	0.1	1.0	0.5	0.5	1.0
<b>INTJ</b>	0.1	1.0	0.1	1.0	0.1	0.1	0.1	0.5	0.5	0.5	0.1	1.0	0.1	0.5	0.1	1.0
<b>ISFJ</b>	0.5	0.1	0.5	1.0	0.5	0.5	0.5	0.5	0.5	0.1	1.0	0.1	1.0	0.1	0.1	0.1
<b>INTP</b>	0.5	0.1	0.5	0.5	0.1	0.1	0.1	0.5	0.5	0.5	0.5	1.0	0.1	1.0	1.0	0.5
<b>ENTP</b>	0.5	0.1	0.5	0.5	0.1	0.5	0.1	1.0	0.5	0.5	0.5	0.1	0.1	1.0	1.0	0.5
<b>ENTJ</b>	0.5	1.0	0.5	1.0	0.1	0.1	0.1	0.5	0.1	0.5	1.0	1.0	0.1	0.5	0.5	1.0

The celebrity compatibility score thus calculated would be in the normalized form and hence have a value between 0 and 1 as the maximum value available in the compatibility type matrix is 1.0.

Figure 3.3: Celebrity matching



### 3.3.3 Summary

The exemplar to be used for the dating algorithm has been chosen as “celebrities”. A mathematical model has been defined for comparing the celebrity choices of users for calculating the similarity score between them. A lookup table and lookup function for fetching the values from this lookup table, would be used to find the pre-defined compatibility score. The calculated celebrity similarity score would have a maximum value of 1.0.

## 3.4 Social Information

In the previous section (3.3), the exemplar to be used was discussed. In this section, the focus is on how to use the social information associated with a user for proposing a match with another user.

### 3.4.1 Overview

To utilize social information, it is important to decide on the social networking website that :

- Has access to a lot of personal information
- Has a large number of connected users
- Has real-life information reflected accurately

The social network that satisfies the above criteria is Facebook, a social networking platform started in 2004 [19]. As of 2012, the membership has crossed 1 billion users. The profile information captured during registration is very extensive and covers all aspects related to the social life of a person. This extensive data set can be used by third-party applications through the Facebook API and SDK software interfaces. Privacy and security concerns have been taken into account and users have been granted the ability to specify different levels at which information would be made available.

### 3.4.2 Model Definition

Bachrach et al.[3] using a dataset of 5,000 users analyzed the dimensions of personality of Facebook users w.r.t. The Five Factor Model (or the Big Five personality test). The personality dimensions can be defined in terms of Facebook profile features like the number of friends, group memberships, likes, photos uploaded and number of photographs in which the user is tagged. The research paper defines the personality dimensions w.r.t. Facebook profiles features on a percentile scale. The equations used for plotting the graphs are not available. So extending upon this research, equations for smooth fit were created with the data available in graphs. The personality dimensions i.e. the big five personality type dimensions can then be calculated using percentile scores for all Facebook features in terms of all the users subscribed to the application.

The information on various facebook features for all participants are captured using Facebook Query Language (FQL). The percentile score for each feature is then computed with these values. Since the data set was not available from the research, the data sets that map facebook feature to personality factor were created from the graphs available in the research paper[3]. The percentile score for each factor for all users is computed by the following steps :

1. Store the value of each facebook feature for every user in a separate array
2. Sort all the arrays in ascending order of values
3. Find the first value in each feature array that is greater than the feature value for the user
4. The ratio of the array index of the matched value in step 3 to the total number of values in the feature array, is the percentile score



$$Percentile(p) = \frac{k}{N}$$

where

$k$  = array index of value which is the first value to be greater than the user feature value

$N$  = the total number of values in the array

The mapping between facebook features and personality factors is given in table 3.4.

Table 3.4: Facebook Features and Personality Factor

Personality Factor / Feature	Likes	Groups	Photos	Friends	Tags
<b>Openness</b>	✓	✓			
<b>Conscientiousness</b>	✓	✓	✓		
<b>Extraversion</b>	✓	✓		✓	
<b>Agreeableness</b>	✓				✓
<b>Neuroticism</b>	✓	✓		✓	

The percentile values obtained above are then used to compute the values for various personality dimensions in the Five Factor Model. Equations for best fit were created from the available data and range from 0 to 1. The equations are defined as :

- *Extraversion* : Three attributes of the user - likes, groups and friends, are used to calculate extraversion.

$$E_v = avg[(0.034 * L + 0.4595) + (0.0645796432 * G + 0.4439913952) + (0.1676363636 * F + 0.3927818182)]$$

- *Agreeableness* : Two attributes of the user - likes and tags, are used to calculate agreeableness.

$$A_v = avg[(0.1630736349 * T + 0.3353506443) + (-0.0351783343 * L + 0.4921188105)]$$

- *Conscientiousness* : Three attributes of the user - likes, groups and photos, are used to calculate conscientiousness.

$$C_v = avg[(-0.1074545455 * L + 0.5330272727) \\ + (-0.0677408177 * G + 0.5082930007) \\ + (0.0322825944 * P + 0.4616235133)]$$

- *Neuroticism* : Three attributes of the user - likes, groups and friends, are used to calculate neuroticism.

$$N_v = avg[(0.0740571532 * L + 0.4396714219) \\ + (0.0442467053 * G + 0.4571748677) \\ + (-0.0567552662 * F + 0.5064208778)]$$

- *Openness* : Two attribute of the user - likes and groups, are used to calculate openness.

$$O_v = avg[(0.1023885436 * L + 0.4210173347) + (0.075129056 * G + 0.4348862463)]$$

where

$L$  – percentile of likes

$G$  – percentile of groups

$F$  – percentile of friends

$T$  – percentile of tags

$P$  – percentile of photos

Match between two users is defined in terms of the similarity score computed from the Facebook features. The similarity between two users - 'i' and 'j', is defined as the average squared euclidean distance of personality dimension values given by :

$$D^{FB} = \frac{(E_{iv} - E_{jv})^2 + (A_{iv} - A_{jv})^2 + (C_{iv} - C_{jv})^2 + (N_{iv} - N_{jv})^2 + (O_{iv} - O_{jv})^2}{5}$$

The Facebook similarity score is then given by :

$$S^{FB}(i, j) = \exp(-D^{FB})$$

### 3.4.3 Summary

The section 3.4.1 discussed the reasons for choosing social information and why Facebook is a suitable choice. Section 3.4.2 discusses how a model can be defined for comparing two users based on features of their Facebook profile. Information from an existing research is used to create the equations for calculating the dimensions of Big Five personality. The formula for computing the similarity score is then expressed in terms of the distance between dimensions of personality between two users.

## 3.5 Collaborative Filtering

In the previous sections, two users were compared based their personality profiles, their choice of celebrities and features available from their Facebook profiles. In this section, a new model of finding matches between two users is discussed.

### 3.5.1 Overview

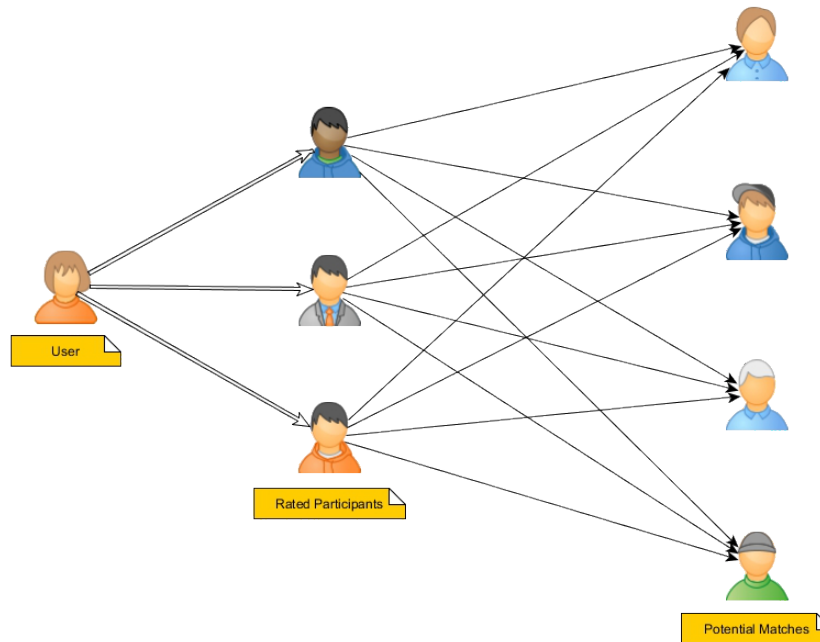
In this section, a new model of proposing matches to users is discussed that is based on the collective feedback of users for each other. The users are provided the option to rate other users on a scale of 1 to 10 according to prior acquaintance or based on their Facebook profile. The users are not required to rate anyone at all. Using these ratings as a reference, a match can be proposed by combining the rating and the similarity score between the users. This can also be interpreted as “people like people who are like themselves” indicating the two levels of interaction involved.

### 3.5.2 Model Definition

For proposing a match to an user, first the list of rated users are found. If there are none, then this approach cannot be applied. If there is atleast one person who has been rated by the user in focus, then collaborative filtering technique can be applied. The rating is done on a scale of 1 to 10 and

converted to the scale of 0 to 1 to ensure it is on the same scale as with the other similarity scores (refer sections 3.2, 3.3 and 3.4). All the users who have not received a rating from the current user would be considered a potential match. This is depicted in figure 3.4.

Figure 3.4: Collaborative Filtering



To finalize the matches for the current user, the following steps are required :

- Find the users from the list of potential matches who has attended atleast one personality test.
- Find the personality similarity score for each of the rated users and all the potential matches in a one-to-many manner provided the pairs have attended the same personality test.
- Use the ratings given by the user and find the normalized product of similarity scores with these ratings.

The above steps can be represented in the form of an equation as explained below.

Let  $u_i$  be the participant for whom the recommendation is generated

Let  $u_j$  be the ranked user

Let  $u_k$  be the user from the list of potential matches

Let  $r(u_i, u_j)$  be the ranking on a scale of 0 to 1 given by  $u_i$  for  $u_j$

Then, the ranked similarity score between  $u_i$  and  $u_k$  is given by :

$$R^{SS}(i,k) = \frac{\sum r(u_i, u_j) * S^P(u_j, u_k)}{\sum S^P(u_j, u_k)}$$

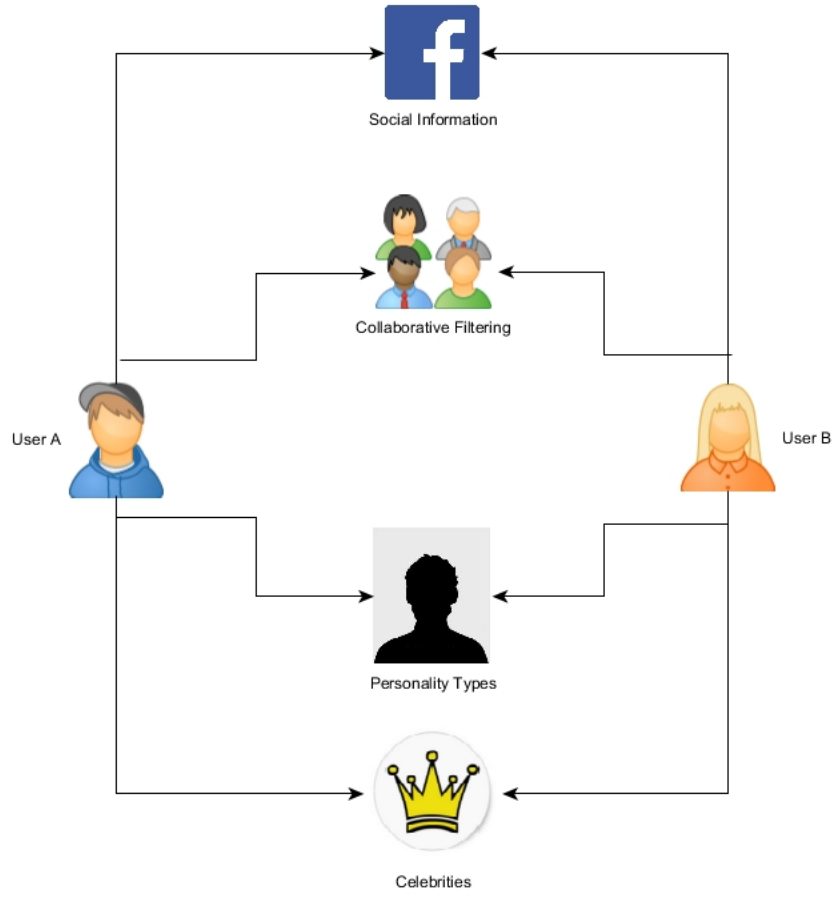
### 3.5.3 Summary

Collaborative filtering is applied to propose new matches to users based on their ratings for other users. The more users they rate, the better the model would perform as there are more people to compare with to reach a stable similarity score. This method fails when there are no ratings available for an user or has rated users who did not attend any personality tests and thus has no dimensions of personality to compare with others.

## 3.6 Combined Model

In sections 3.2 to 3.5, four models of finding similarity between people were discussed. If all four types of information were available - personality type results, list of preferred celebrities, list of Facebook friends who had the desired characteristics in your “dream” partner and ratings given to other users; then these models can be combined together to get the final similarity score. The combination can be visualized as shown in figure 3.5.

Figure 3.5: Combined Model



Thus the similarity score for two profiles in normalized form is given by :

$$SS(i, j) = \frac{\sum l^S(i, j)}{\sum n(l^S(i, j))}$$

where

$$l^S(i, j) \in \{ S^P(i, j), S^C(i, j), R^{SS}(i, j), S^{FB}(i, j) \}$$

$$n(l^S(i,j)) = \begin{cases} 1 & \text{where } l^S(i,j) \neq 0 \\ 0 & \text{where } l^S(i,j) = 0 \end{cases}$$

From the above equation there arise the following conditions :

1. *A new participant* : There is no preferences data available. Matches can be proposed only after preferences and profile information is captured. Personality tests and information from Facebook profile has to be sourced and analyzed.
2. *Ratings not provided by the participant* : The participant has taken one or more personality tests but has not rated anyone yet. In this case the ranked similarity score would be zero. The similarity score would completely depend on matching the list of celebrities between two participants.
3. *Celebrity list not created for the participant* : Ratings for other participants are available but the list for preferred celebrities is not available. In this case, the final similarity score would be the ranked similarity score as the celebrity score would be zero.
4. *Ratings and celebrity list not available for the participant* : The participant may have taken personality tests but the final similarity score would still be zero as no ratings or list of preferred celebrities are not available. In this case, matches cannot be proposed. Some filter criteria based on attributes would have to be used to propose matches. The possibilities are :
  - (a) Capture list of preferred attributes from the participant. These could be and not limited to age, ethnicity, salary, location, native language, social preferences, religion, children, career etc.
  - (b) Automatically generate a list of attributes from the data available in Facebook profile. Some information about the participant profile can be hidden due to security restrictions. Using the rest of data available, an implicit user profile can be created with a set of attributes and respective values. A personality type of the user can be deduced accurately based on the social activity of the user profile provided the social activity feed closely matches real-life events in the participant's life.
5. *Ratings and list of celebrities are available* : The participant has not attempted any of the personality tests. Comparison of two participants can only be done iff their personality type

information is available. Since ratings and list of celebrities are available, the similarity score can be calculated directly if the similarity score for personality types are available. Using the list of celebrities, the personality type of the participant can be deduced based on the assumption that people tend to identify with celebrities whom they aspire to be and who has similar personality traits, in most cases. Let  $l^C$  be the list of celebrities and  $l^T$  be the list of personality types that are a best match and also a possible match  $\forall l^C$ . Then the personality type for user  $i$  is given by

$$P_i = \max(\text{mode}(l^T))$$

If there are multiple types with the same count, the first type is chosen.



# Chapter 4

## Implementation

This chapter deals with the implementation details for the online dating algorithm that was discussed in chapter 3. A web application is developed to implement the dating algorithm and to gather feedback from users. This application would also be used to get anonymized aggregate information from the users for the evaluation phase as discussed in chapter 5.

### 4.1 Introduction

In sections 3.2 to 3.6, various models for comparing people were designed. To test these models with users, an application that facilitates interaction with users and that can implement these models is required. Also, celebrity choices and information from social media is also required. All these can be accomplished by creating a web application that can run on the Facebook platform. The platform provides an “Apps” API wherein anyone can create an application shell within the facebook platform and link the information to their own hosting services. The application meta-data has to be configured with parameters like site hosting URL, age restriction to prevent apps with mature content from reaching a young audience, english language support etc. Having an application on facebook platform also has the benefit of viewing analytics on the application through the Insights<sup>1</sup> dashboard without having to write extra code to track the administration of your application. Web hosting was on the School of Computer Science and Statistics (SCSS) server with public access granted to everyone for the application data. Redundancy and failover measures were managed by the Information Services (IS) team dedicated to SCSS.

---

<sup>1</sup> A tool released by Facebook in 2010 to provide developers metrics for different content. Refer : <https://developers.facebook.com/docs/insights>. Last accessed on 26-Aug-2014.

## 4.2 Architecture

The web application can be divided into three logical layers in terms of data presentation and processing. They are :

- *User Interface (UI) Layer* : This layer was designed using HTML5, CSS3 and jQuery framework. The new features defined in the HTML5 specification like default support for password fields, new tags with built-in validation, effects with CSS3 and finer control offered by jQuery, made the processing of developing the UI much easier. There was also a saving of around 20% in terms of the number of lines of JavaScript code written, when code was refractored using jQuery. Manipulation of HTML elements and changing CSS styles or attributes on the fly was much easier due to the built-in support in jQuery for multi-attribute selection functions. These files were zipped and minified<sup>2</sup> to reduce the file size and thus saving loading time of the webpage by cutting short the content required to be downloaded for displaying the page. The zipped files would be unzipped while rendering the webpage and can execute code without any loss of functionality.
- *Middleware Layer* : This layer handles the incoming requests from the user for displaying and saving content. The content is served from a web server running Apache<sup>3</sup> HTTP server installed in a machine running Linux operating system (OS). The server-side programming language called PHP v5.0 is used for development of this layer. It was chosen as the server technology since it powers around 244 million[43] websites around the world and is one of the most popular server scripting languages running on both windows and linux based platforms. It is also very versatile in terms of rapid development, external libraries and OOP support. Each completely contained processing task is written in one file called a PHP script which can have native file system commands, PHP syntax, plain text output or even a combination of one or more. These scripts are saved as plain text files and are invoked when the user requests for that particular script. The web server, executes the requested script and sends the output back to the user. Multiple processing tasks can also be combined together for execution in one go called “Batch Jobs” which can be easily written and invoked, either from the command line or from within another PHP script. Invoking a script from within another allows for chaining of processes to ensure data integrity and for creating rollback milestones for restoring the system to a previously stable state. These scripts can be created and edited in any text editor, with or without a GUI.
- *Data Storage Layer* : This layer manages the application data and makes sure it is saved and retrieved correctly. The data manipulation process is done using PHP with data stored and

---

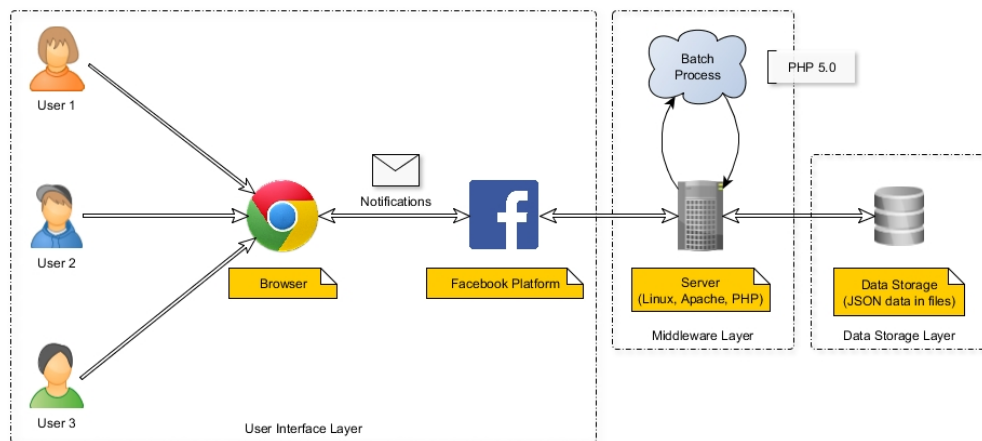
<sup>2</sup> Minification is a process in which contents of a file are removed to reduce the size without creating any problems with execution. Source : “<https://developers.google.com/speed/docs/insights/MinifyResources>” Last accessed on 26-Aug-2014.

<sup>3</sup> An open-source HTTP server available from 1996. Source : <http://httpd.apache.org/>, Last accessed on 26-Aug-2014

retrieved in JavaScript Object Notation (JSON) format. Using JSON has two advantages - data retrieval can be done by calling a “key” which is unique and the data is stored in a human readable format which makes it easier to make changes if required. Since data is stored in plain text format, it can be used across different OSs or file systems without changes. The data from JSON files can be easily consumed in the UI layer with javascript and in the middleware layer using PHP, without any changes as JSON support is provided by technologies used in both layers.

Figure 4.1 shows the pictorial representation for the architecture of the application which has the layout of various entities.

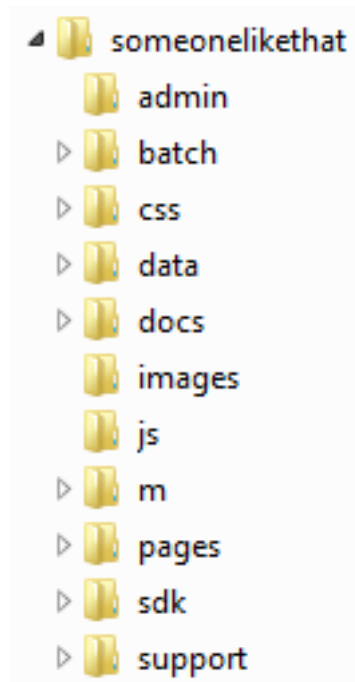
Figure 4.1: Architecture Diagram



#### 4.2.1 Project Organization

The web application requires different kinds of files or resources. A proper project organization is required to manage resources as well as to ensure data saved for one functionality does not affect the existing data. The resources are stored in hierarchial structure as shown in figure 4.2.

Figure 4.2: Project Organization



Details of resources available in various folders are :

- *admin* : Contains PHP scripts and supporting files that are used to anonymize and aggregate application data
- *batch* : Contains PHP scripts that are used to execute batch jobs
- *css* : Contains CSS files that are used for rendering the HTML elements in the UI layer
- *data* : Contains user data, celebrity data and social information about users
- *docs* : Supporting documentation for the application
- *images* : All images are stored here
- *js* : All javascript and jQuery script files are stored here
- *m* : Contains the resources for the mobile prototype
- *pages* : Contains the HTML and PHP web pages for the application
- *sdk* : Contains the SDK files provided by Facebook for PHP
- *support* : The source files of jQuery framework and the Chart library

For loading any of the resources mentioned in the above folder, the relative path from the invoking resource to the target resource is used so that if the root folder is renamed or moved to a different location, the application still would execute smoothly.

### 4.2.2 Accessing Celebrity Information

Celebrity information is collected [7] and saved in a comma separated file (CSV). For each celebrity, the following fields are available :

- Name - First, Middle and Last
- Alias
- Profession
- Myers-Briggs personality type

The data for all the celebrities (~300) is read from the CSV file using *fgetcsv* function available in PHP. There are no unique identifiers (IDs) for any of the celebrities. It is because when a new celebrity is added to the list in future, the only change to the file would be a new line at the end of the file. Having unique IDs might result in duplicates if not carefully scrutinized. But when the celebrity information is used in the UI layer, a unique ID is created on the fly for each celebrity by combining the first name, middle name and last name, without any special characters or spaces. This is on the assumption that all people can be uniquely represented by their name and if there are any conflicts, then new data cannot be added to the existing celebrity list. The auto-generated ID is then saved when the user marks one or celebrities as their choice.

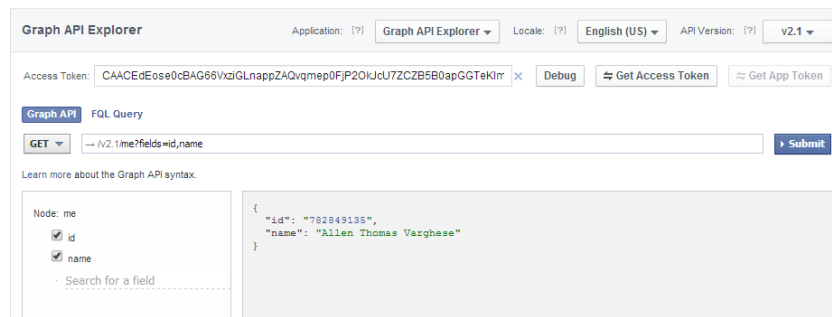
### 4.2.3 Accessing Facebook Information

Social information about users from Facebook can be obtained using the PHP SDK [26]. The request to Facebook platform is made in using REST over HTTP with the URL token of the Facebook API. The API returns the information in JSON format. If not available, then an error code and corresponding message is returned. Depending on the amount of data present, the time taken would also vary. Time can also differ with each call depending on the number of requests served by the Facebook platform and hence there is a possibility of a timeout after 10 seconds. If a timeout happens, then another request would be sent. This resending of requests would be tried for 3 successive times and after that user is informed about the occurrence of timeout. The JSON data can then be parsed using PHP and after processing, sent to the UI layer. The friends information is saved in the data storage layer for caching purposes aimed at reducing the number of calls to the Facebook API.

Information about different events or statuses within the application is sent to the users using the Facebook Notification API. The user ID and text message has to be specified for each notification. There is no facility for sending a delayed notification. The notification data is sent through the Facebook PHP SDK using REST over HTTP. The notification would be displayed in the notification stream of the user. When the notification is clicked, the user would be taken to the application.

Facebook has a developer console where the information available to an application can be tested and displayed before used by the developer. This utility - Graph API Explorer<sup>4</sup> (refer Figure 4.3), has been used throughout the development of the application to ensure that correct data is obtained and there has been no change in the API due to frequent updates happening to the Facebook platform. The developer need to choose the application ID for which the queries are requested and also ensure that an appropriate access token is obtained. This can be done by clicking on the “Get Access Token” button and choosing the list of permissions that is required for the application. The access token has an expiry time which is usually 24 hours. If required, the time can be increased when the access token is created programmatically from within PHP SDK by specifying the required time period in seconds.

Figure 4.3: Graph API Explorer



The Facebook API calls are HTTP REST requests which requires the access token with each request. The graph explorer provides the options to create queries/requests using a search feature that has the complete list of fields available. The query builder is very interactive and creates fields by just choosing the required fields in the query builder. The query string can then be submitted to the facebook platform and the result viewed in the results pane along with the time taken for processing the request. The auto-generated query can be edited manually or using the automatic query builder to view updated results. Once confirmed, the generated URL can be used in the code to create an HTTP GET or POST request and then parse the JSON response to extract the required data.

Social information about the users that is available on a facebook profile like the number of likes,

<sup>4</sup> Source : <https://developers.facebook.com/tools/explorer>. Last accessed on 26-Aug-2014.

group memberships, number of photographs uploaded, count of friends and tags in photographs are captured occasionally due to the time taken and number of API calls being made. This process is performed using a batch script that is written in PHP and runs from a command line shell. Facebook supports querying for information through a query language called Facebook Query Language (FQL) which are used to obtain social data. The queries used are :

- Get count of Likes - *SELECT likes\_count FROM user WHERE uid = me()*
- Get list of Groups - *SELECT gid FROM group\_member WHERE uid = me()*
- Get count of Photos - *SELECT photo\_count FROM album WHERE owner = me()*
- Get count of Friends - *SELECT friend\_count FROM user WHERE uid = me()*
- Get list of Tags - *SELECT object\_id FROM photo\_tag WHERE subject = me()*

The above information for each user is accessed and saved to the profiles of each user. The percentiles of each user for each attribute is then computed. This sub-process is memory intensive as the complete list of information has to be maintained and processed in memory. Once the calculation is completed, the percentile scores for each attribute is saved in each user profile. These scores are then used when the compatibility score is computed.

In addition to the above, there are few other calls also being made to access/obtain specific information like :

- Permissions for the facebook application : The list of permissions that are displayed to the user are :
  - 'basic\_info', 'email', 'user\_about\_me', 'friends\_about\_me', 'user\_activities',
  - 'friends\_activities', 'user\_birthday', 'friends\_birthday', 'user\_checkins',
  - 'friends\_checkins', 'user\_education\_history', 'friends\_education\_history',
  - 'user\_events', 'friends\_events', 'user\_groups', 'friends\_groups',
  - 'user\_hometown', 'friends\_hometown', 'user\_interests', 'friends\_interests',
  - 'user\_likes', 'friends\_likes', 'user\_location', 'friends\_location',
  - 'user\_notes', 'friends\_notes', 'user\_photos', 'friends\_photos',
  - 'user\_questions', 'friends\_questions', 'user\_relationships',
  - 'friends\_relationships', 'user\_relationship\_details', 'friends\_relationship\_details',
  - 'user\_religion\_politics', 'friends\_religion\_politics', 'user\_status', 'friends\_status',
  - 'user\_subscriptions', 'friends\_subscriptions', 'user\_videos', 'friends\_videos',
  - 'user\_website', 'friends\_website', 'user\_work\_history', 'friends\_work\_history',

- `'user_actions.music'`, `'friends_actions.music'`, `'user_actions.news'`,
- `'friends_actions.news'`, `'user_actions.video'`, `'friends_actions.video'`,
- `'user_games_activity'`, `'friends_games_activity'`.

The complete list of permissions is available in Facebook documentation[25].

- Obtaining the default profile information : This is done by using a Facebook GET request with query as `"/me"` which gives the fields and information which has `"public"` access status associated with it. Those fields which have a narrower privacy setting than `"public"`, would not be available in the response.
- Friends list : To obtain the list of friends, a Facebook GET request has to be issued with query as `"/me/friends"` which gives the list of friends with two fields - `name` and `id` for each user. This information is an array of values in the JSON response.

#### 4.2.4 Model Implementation

This section discusses how the model definitions given in sections 3.2, 3.3 and 3.4 are implemented in the application. They are given below :

- *Personality Type Model* : The personality type and its values have been modelled as vectors in section 3.2. These have been represented internally by a data structure for results from each test with a JSON key for the type and another JSON key for the value. Thus any dimension of the personality type can be compared.
- *Exemplar Model* : The celebrity model has been modelled with respect to a lookup table that has weights assigned for compatible personality types as seen in section 3.3. The celebrity choices of the user is represented internally as an array of celebrity IDs whose personality types are obtained from the main CSV file that has the complete list of celebrities. The weighted matrix is included in a lookup function that returns the corresponding weights based on the personality types that is passed as parameters to the function.
- *Social Information Model* : This model is represented in the application as one JSON data structure that has defined keys for both personality types and values. The social information obtained from Facebook platform is saved in a separate file and the computed values for dimensions of personality type is then added to the file of the user.

*Collaborative Filtering* mentioned in section 3.5 is a combination of the personality model and ratings given to users. The ratings given by an user for others are also stored as a separate JSON data structure for easy lookup and mapping. A typical data structure for an user who has provided all the information is given in Appendix H.



## 4.3 UI Design

Users who are already familiar with the Facebook platform, switching to a Facebook application is a different experience because of the content and how it is presented. The transition would be an smooth experience if the application has a similar experience to Facebook. With this goal in mind, the same color scheme and font types - Lucida Grande, Tahoma, Arial, Verdana; as in Facebook were used. The Facebook identity guide [23] was used during the development process to validate the compatibility in design<sup>5</sup>. There is a dedicated website [22] and developer group [24] for resolving any issues with application development on Facebook platform. The application is a Facebook Canvas application with the hosting of information and other assets on a server external to Facebook. The images used in the application and report were created/edited using GIMP<sup>6</sup> and yEd Graph Editor<sup>7</sup> software.

A hierarchical model was used for propagating display properties of elements in different screens. This ensured consistency in rendering of screen elements and their attributes like font, color, location, size etc. The pages were also loaded in a hierarchical manner to take advantage of the inheritance properties of CSS3 and jQuery. For screen changes, separate CSS and JavaScript files were loaded for each page which allowed for functionality of one screen to be isolated from the other screens and thus prevent accidental overwrite of properties or data during runtime.

### 4.3.1 User Interaction

Presenting data in an easy to understandable format is very important to increase the impact of the study and to make sure that users understand the requirements. Users should be able to have easy access to data at any point of time. This can be achieved by incorporating a certain level of responsiveness in the application without complicating user tasks and also by having the links within the application in the main interface.

For example. When the user chooses a rating for an user, the information is saved immediately rather than having the user click on another button to save the information. This automatic saving of information based on individual clicks ensures that information is not lost and the state of the application displayed to the user is consistent till the last known stable state, in case of any failure.

Navigation within the application is by using a section on the left side of the display that has all the links within the application. This pane side section is always visible. Auto-hide functionality was not added as users need to have direct access to the functionalities at any point in time. Indicators are available for the options under “Date Center” on the left-hand pane for capturing the user’s attention. It would a small rectangle displayed to the right end of an option with

---

<sup>5</sup> Changes are made to the Facebook platform from time to time. So there might be a difference in the application in terms of font types, sizes or color depending on the latest update available in Facebook. Development on the application has been stopped as this study is over.

<sup>6</sup> Available for download at [www.gimp.org](http://www.gimp.org). Last accessed on 26-Aug-2014.

<sup>7</sup> Available for download at [http://www.yworks.com/en/products\\_yed\\_about.html](http://www.yworks.com/en/products_yed_about.html). Last accessed on 26-Aug-2014.

a number indicating the number of profiles associated with the user in terms of recommended profiles or date matches.

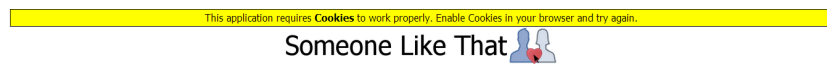
There are three personality tests in the application which when attended would give the personality type and the scores for each dimension of personality. This result is displayed in the form of a radar chart or a bar chart to visualize the variation in scores between different dimensions. The chart can also be saved for future reference by invoking the popup menu over the image.

The Facebook application can be accessed through Mozilla Firefox or Google Chrome. These two browsers has near complete support for HTML5 and CSS3 features. Some users disable JavaScript and the use of cookies during browsing. Such settings have an adverse effect on the application and would cause it to stop responding as the user side interaction is dependent on the availability of javascript. Also, Facebook API is dependent on both JavaScript and cookies to function properly. Cookies are also required to maintain user session data on the web application. So a check for Javascript support (figure 4.4) and cookie availability (figure 4.5) have been added to the application which would appear as a bar at the top of the main page.

Figure 4.4: Javascript Disabled



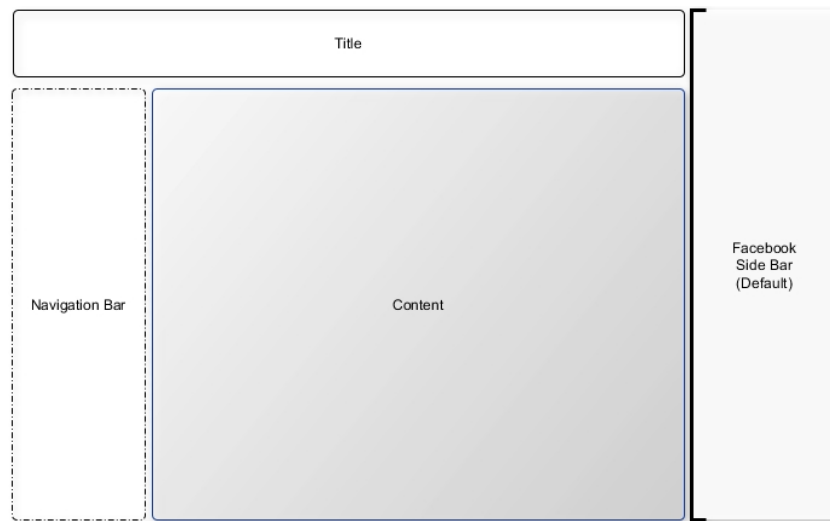
Figure 4.5: Cookies Disabled



### 4.3.2 Layout

The full screen size of the user browser is not available in the canvas application as Facebook has a side bar on the right hand side of the page that is used for promotion of other apps and content based on what is available in the canvas application. The application is opened in an HTML iFrame element with default security features like Cross Origin Resource Sharing (CORS) enabled. Such features prevent hacking of the application by third-party scripts and also ensures that only those resources from the server where the application is hosted can be accessed. HTML5, CSS3 and jQuery were used to design the layout and make sure the layout stays the same even when the screen is resized or when displayed in a resolution lower than 1280 x 960. The available screen area was divided as shown in figure 4.6 to segregate content based on its nature.

Figure 4.6: UI Layout Structure



If the content cannot be displayed correctly with the available display resolution, the browser automatically includes a horizontal scroll bar for the Navigation Bar + Content area and a vertical scroll bar for Title+Content area. The application has been designed for a fixed width of 1030px and so if the application is viewed in a display resolution more than that, the application area would be displayed with center alignment in the screen.

#### 4.4 Functionalities

The sections available in the web application are :

- Introduction : This section gives an introduction to the research being carried out. This would be the landing page when the user tries to access the application. A screenshot is available as figure 4.7.
- Privacy and Ethics : This section details the various privacy and ethical considerations approved by the Ethics committee <sup>8</sup> of SCSS. A screenshot is available as figure 4.8.

---

<sup>8</sup> Exact details of ethics approval and research proposal are available in Appendix A.

Figure 4.7: Introduction

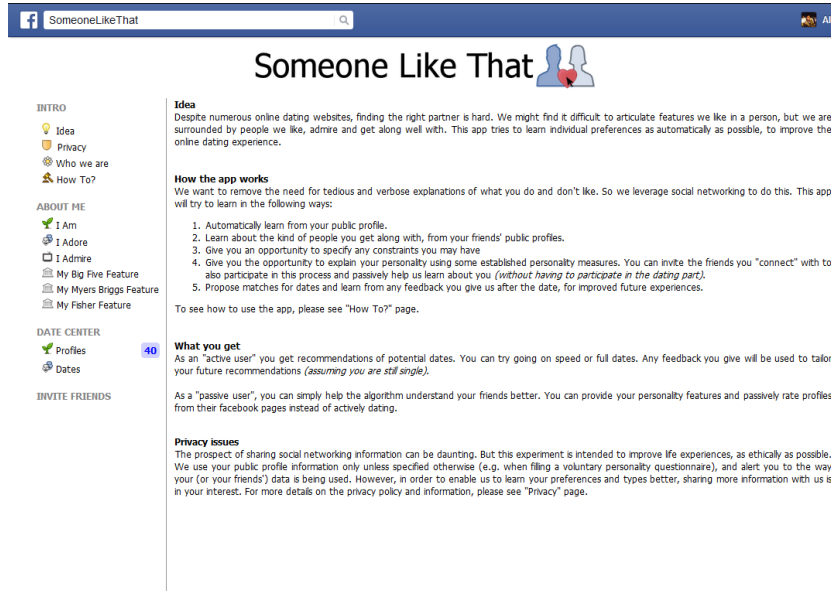
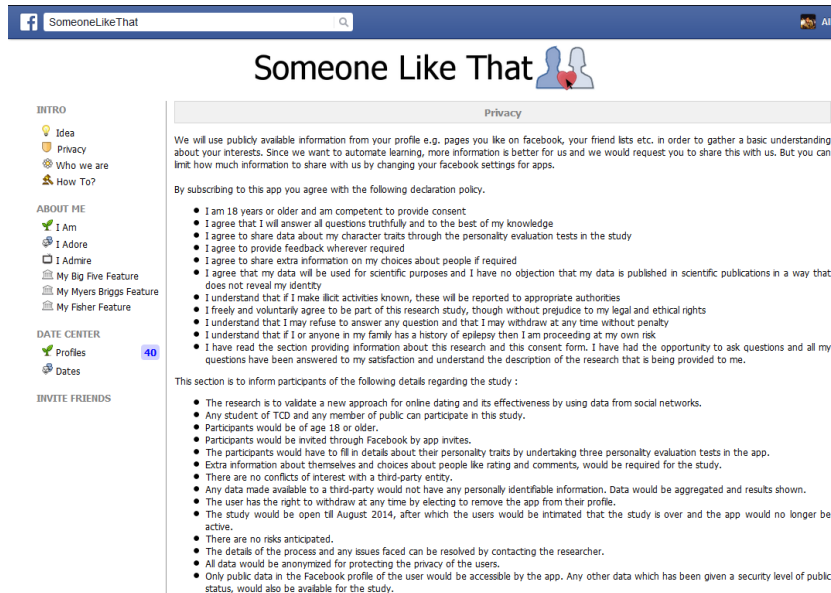
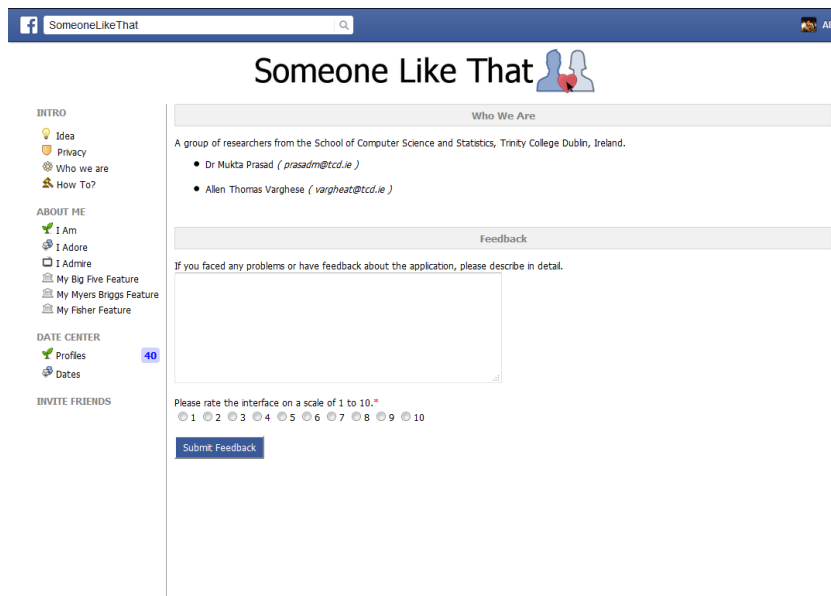


Figure 4.8: Privacy and Ethics



- Application Feedback : Any feedback on the application is captured here. Users can give a detailed description of any problem they faced and also give a rating for the application on a scale of 1 to 10. A screenshot is available as figure 4.9.

Figure 4.9: Contact information and application feedback



- Contact Information : If any of the users face any problems in using the application, they can reach out to the researchers through e-mail.
- Personal Information : In this section, the date of birth, gender, a self-description and status of participation (active or passive) is captured. None of these fields are required fields. These details was initially planned to be used during the evaluation phase but then avoided due to the low number of users who signed up for the application. A screenshot is available as figure 4.10.
- 1st Personality Test : The Big-Five personality test questions were taken from International Personality Item Pool[4], a scientific collaboratory for the development of advanced measures of personality and other individual differences. The marking instructions for determining

various features are also available on the website. There are 50 questions that users have to answer on a scale of 1 to 5 with 1 being “very inaccurate” to 5 being “very accurate”. After taking the test, the results are displayed as score<sup>9</sup> for each factor of the Big-Five test on a scale of 0 to 100. The combined result is also shown in the form of a radar chart that was created using Chart.js framework with the default options on. A screenshot is available as figure 4.11.

- 2nd Personality Test : The Myers-Briggs personality test questions were sourced from the personality test available for doctors joining internal medicine at the Dartmouth Hitchcock Medical Center[36]. There are 70 questions which represent different situations and the user has to choose from one of them. The final scores for different personality types are calculated<sup>10</sup> out of 100, based on the answers. The scores are calculated on a scale of 100 and are represented on a bar chart created using Chart.js, for easy comparison. The user can also save this chart if needed. A screenshot is available as figure 4.12.
- 3rd Personality Test : The Fisher personality test questions[30] were adapted from Dr Helen Fisher’s book - “Why Him? Why Her?”. There are 56 questions that have to be answered each having a value from 0 to 3 where 0 being “Disagree” and 3 for “Agree”. The scores are calculated<sup>11</sup> on a scale of 100 and are represented on a bar chart created using Chart.js, for easy comparison. The user can also save this chart if needed. A screenshot is available as figure 4.13.
- Facebook friends list : This section displays the list of friends available in the user’s social network each with a checkbox to indicate selection. There is also an option to search and a toggle button to view only the selected friends. The Facebook API for friends with URL - “/me/friends” is used. The complete friend information is obtained as a JSON data file which is parsed using PHP to display the information. Each time the user marks any checkbox, the information is saved immediately and a popup is displayed at the top of the screen indicating status of the operation. A screenshot is available as figure 4.14.
- Celebrities : This section has a list of over 350 celebrities with their Myers-Briggs personality scores and other attributes like first name, middle name, last name, alias and profession. The user can choose one or more celebrities by clicking on the checkbox next to each celebrity. Users can also search for celebrity by typing their partial names. For a quick view, there is a button to view the already selected celebrities. Users can also change their choice by unchecking the checkbox for the corresponding celebrity. A screenshot is available as figure 4.15.

---

<sup>9</sup> The questionnaire and scoring sheet is available in Appendix C

<sup>10</sup> The questionnaire and scoring sheet is available in Appendix D

<sup>11</sup> The questionnaire and scoring sheet is available in Appendix E

Figure 4.10: Personal Information

The screenshot shows the 'Basic Info' section of a user profile. The user's name is 'Someone Like That'. The gender is set to 'Male'. The birthday is '1985 May 29'. There is a text area for 'Describe Yourself' with the text 'Hi'. The 'User Type' is set to 'PASSIVE USER'. There are 'Save info' and 'Cancel' buttons at the bottom.

**Basic Info**

Gender:  Male  Female

**My Birthday**  
 year month day  
 1985 May 29

**Describe Yourself**  
 Write a short text about yourself. This text is shown to the people you are recommended to

Hi

**User Type**  
 ACTIVE USER  PASSIVE USER

Save info Cancel

Figure 4.11: Big-Five Score

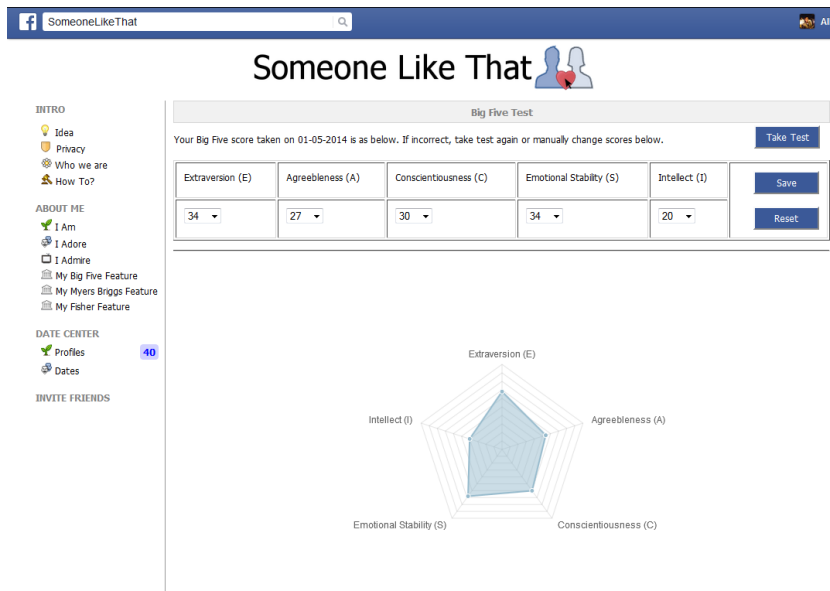


Figure 4.12: Myers-Briggs Score

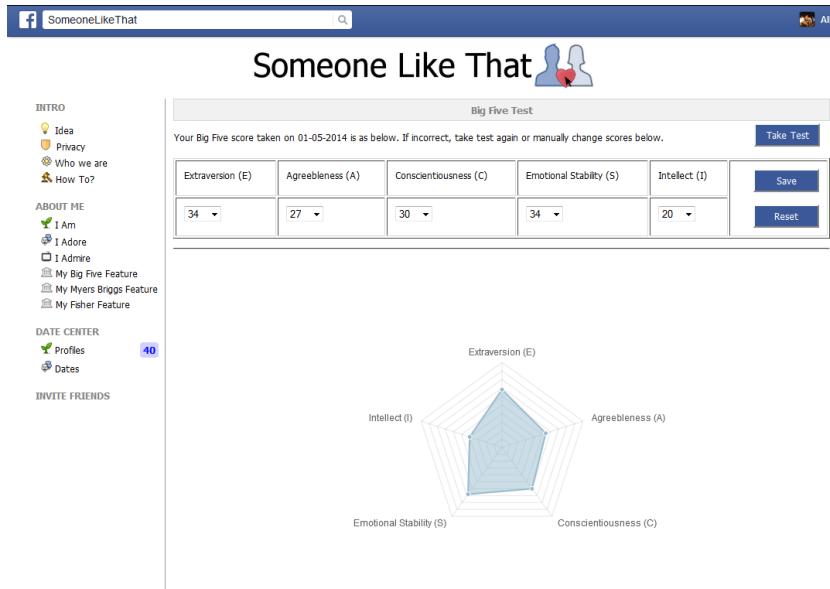


Figure 4.13: Fisher Score

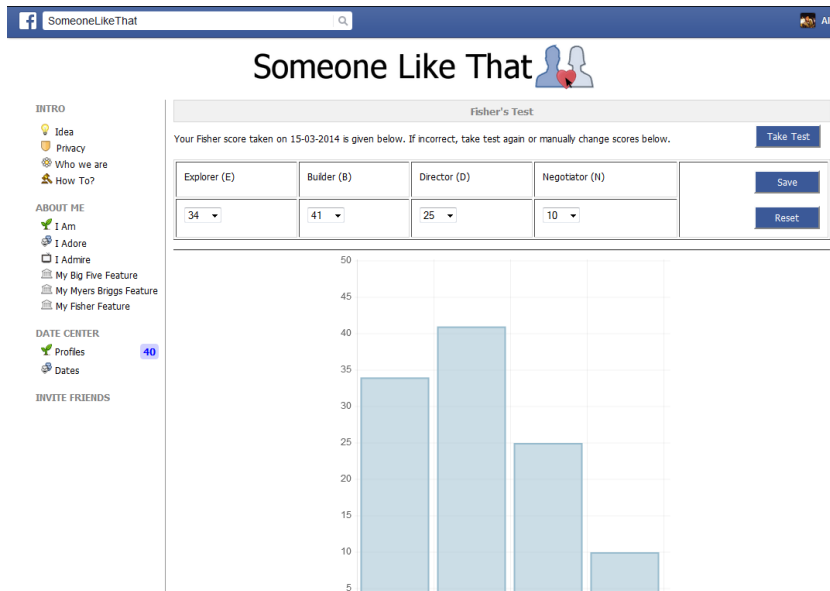
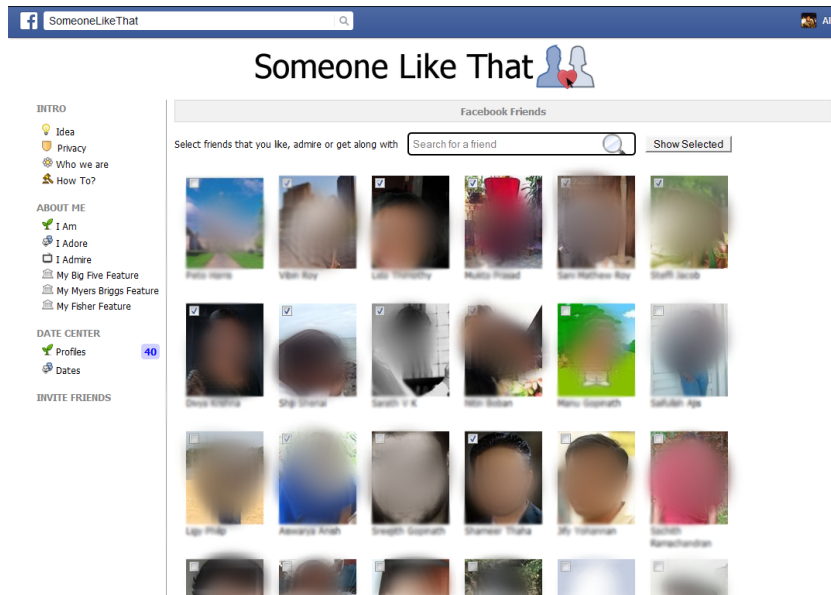




Figure 4.14: List of Facebook Friends



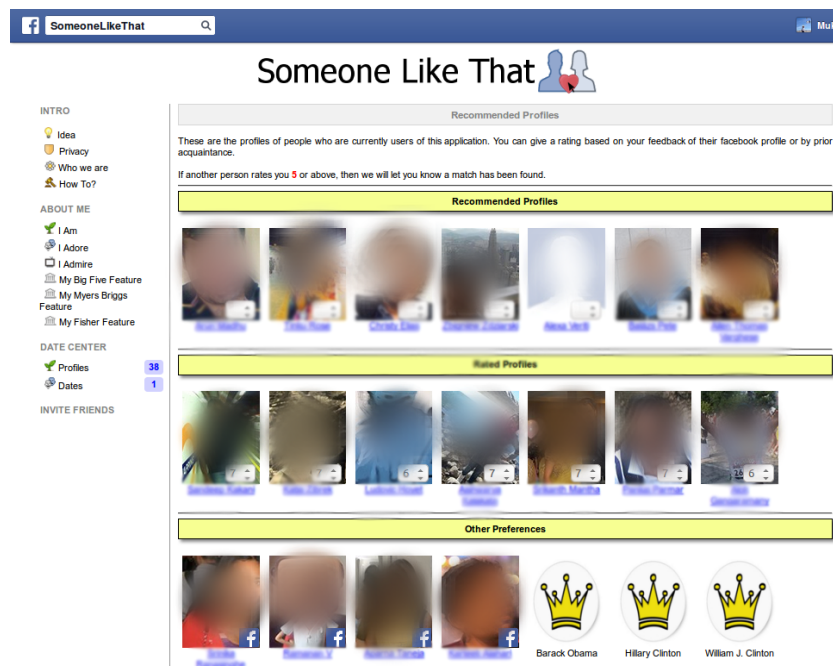
*The user images and names have been blurred to respect privacy*

Figure 4.15: List of Celebrities



- Recommended Profiles : This section gives a list of all users who have signed up for the application. Gender based filtering is not implemented as the amount of data available for each gender is small but together is large enough. There are three categories - “Recommended Profiles”, “Rated Profiles” and “Other Preferences”. Users can rate others on a scale of 1 to 10 based on their Facebook profile or based on prior acquaintance. Each user is represented by their profile picture, a drop down box for rating and a link to their Facebook profile. The recommended profiles section shows the list of users which have not been rated. The rated profiles section has the list of users that have been rated by the currently logged-in user. The other preferences section has the list of chosen Facebook friends and celebrities. A screenshot is available as figure 4.16.

Figure 4.16: Recommended Profiles



*User profile images and names have been blurred to respect privacy*

- Date Feedback : Users are matched for dates if they give each other a rating of 5 or more out of a scale of 10 in the recommendations section. People can go out on a real or virtual date and then give their feedback on a scale of 1 to 10. A screenshot is available as figure 4.17.

For this research, the users can also give their rating based on their previous experience or knowledge about the other person with or without actually meeting them. This flexibility was included because of the geographic distances between the vast majority of the users who had signed up. This data is used to validate the effectiveness of the algorithm. There are two sections - “Recommended Dates” and “Rated Dates”. The former section shows the list of profiles with a link to the profile along with a drop down field for providing the rating. The latter section has the list of users (*dates*) who have been rated. The data in this page is saved automatically whenever the user changes the value of any rating. Users have the freedom to change their data provided, at any given time. A screenshot is available as figure 4.18.

- Application Invites : People can be invited to be part of the research/application through the Facebook platform itself by using the “Notifications” API. Small text or HTML formatted messages can be sent to the user’s social network of friends with a back link to the application. The message is - “Hi. I would like to solicit your participation for a study required as part of my master’s program in Trinity. The study is to understand the impact that personality has on relationships. Your input is required in two steps. In the first step, you enter data into the application to create a personality profile. The estimated time for completing the study is 15 minutes. The details of your personality profile would be available at the end of the test. The data thus collected would be processed by an algorithm to make predictions. In the second step, we solicit your feedback on the effectiveness of our predictions and the data provided would be used to improve the algorithm. Please continue to <https://apps.facebook.com/someonelikethat>”. The “notification” would be made available in the notification stream section of the target user. Facebook by default, sends out an alert whenever there is a new item in the notification stream no matter which channel they have logged in, be it web or a mobile device. If the target user clicks on the notification, they would be taken to the Facebook application where they would be asked for confirming the permissions and then logged into the application. A screenshot is available as figure 4.19.

## 4.5 Challenges and Solutions

Below are the challenges faced during the development of the application.

- *Support for PHP libraries* : Some of the PHP5 packages were not enabled during the initial stages of application development. These were related to use of APIs for modification of files and under different permission groups in Linux. This was resolved by enabling these libraries

Figure 4.17: User Ratings

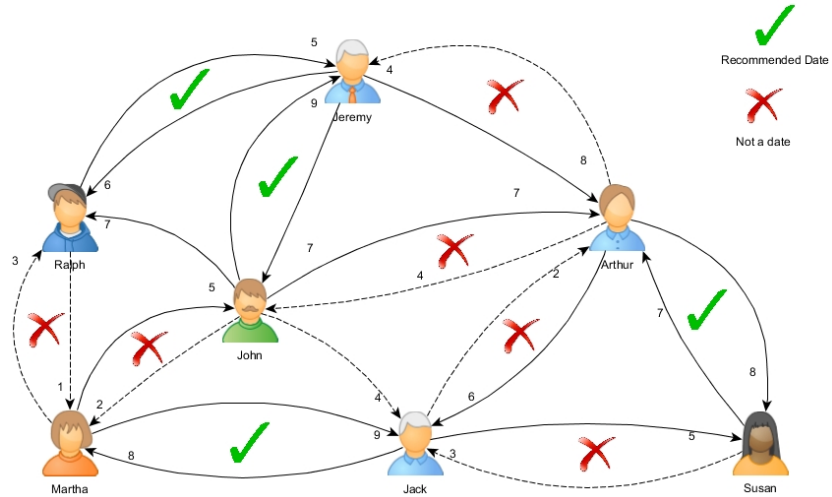
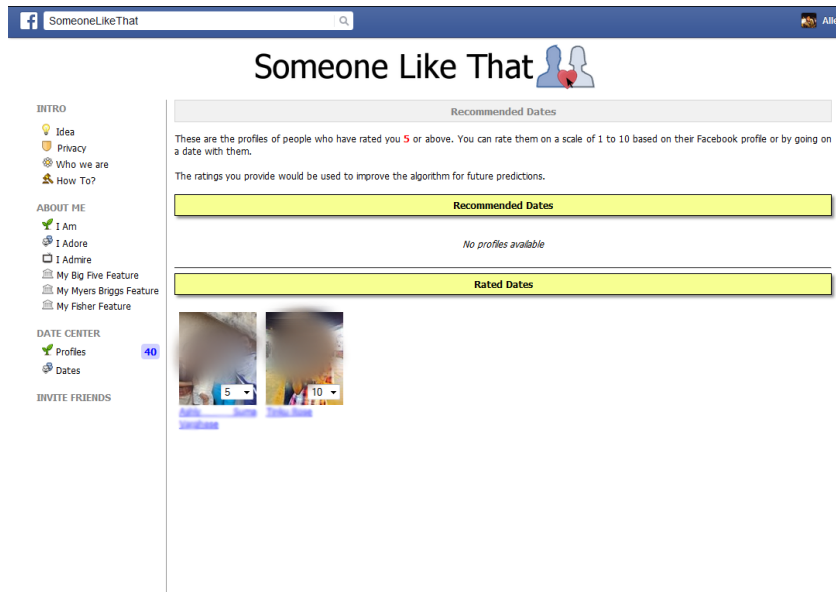
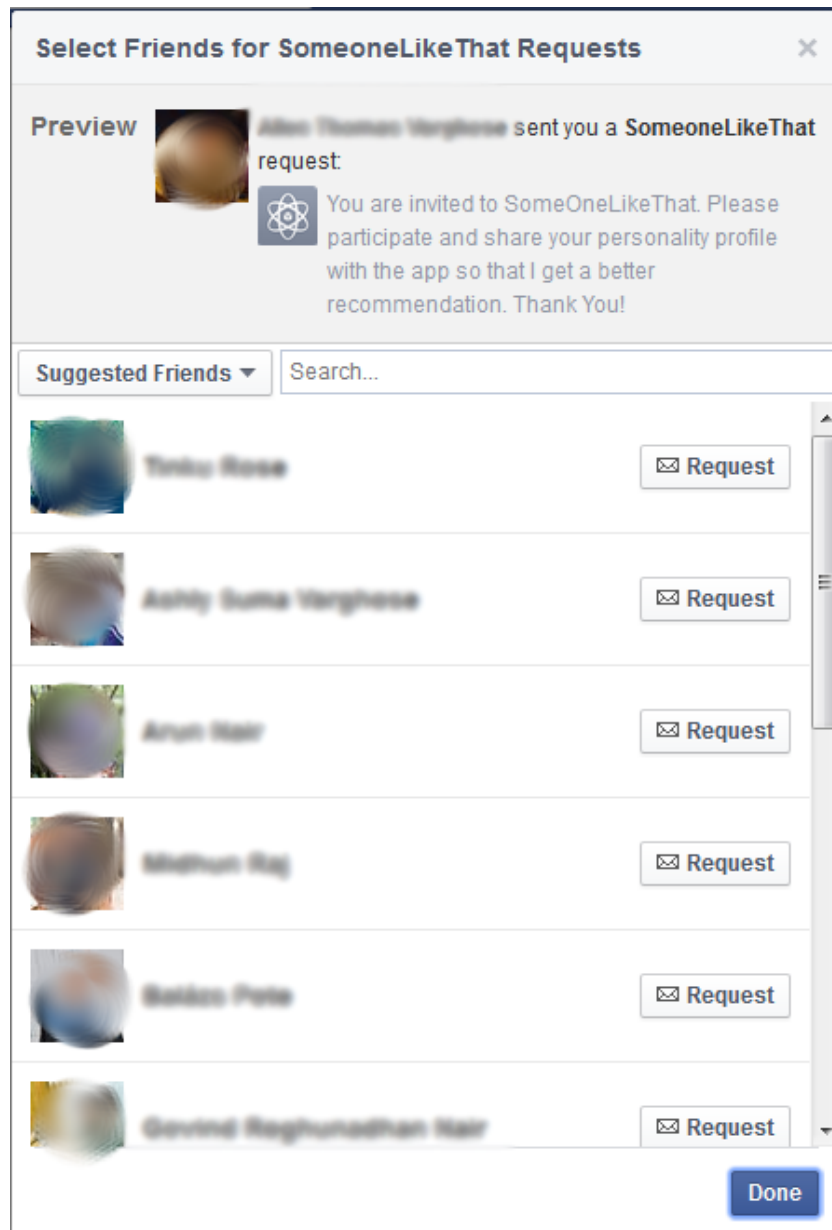


Figure 4.18: Date Feedback



*User images and names have been blurred to respect privacy*

Figure 4.19: Application Invite



*User images and names have been blurred to respect privacy*

in the settings file of PHP5 installation and also by granting sufficient permissions in Linux user groups.

- *Cross-Browser issues* : There were some rendering issues especially with using CSS across Firefox and Chrome. This was resolved by removing browser specific acceleration tips and using only HTML5 specific alternatives. The root cause is that the level of adoption of CSS and HTML standards is not consistent across browsers.
- *Facebook API updates* : The API version of Facebook platform has changed atleast twice in the past 1 year. Some of the updates deprecated support for some features or put an expiry date for few of the existing features in the coming 6 months [21]. Due to rapid updates happening, the documentation for PHP SDK is not updated at the same pace making it very hard to understand and make appropriate changes. Posting questions in the developer group on Facebook [24] gets answered by Facebook employees. This is the solution that majority of the Facebook API based developers have adopted.

# Chapter 5

## Evaluation

This chapter discusses the evaluation of the algorithm that has been discussed in section 3. Not all sections of the algorithm are tested though. The focus is on how information from exemplars and social networks like Facebook can be used accurately for calculating similarity between two users.

### 5.1 Users

Users were invited to join the Facebook application[35] through Facebook group posts, Facebook messages and over e-mail. Anyone of age 18 years or more can join the application and submit their data. Over a period of 4 months, around 60 people signed up. The participants were given identifiers when the data was analyzed so that they cannot be individually identified.

### 5.2 Experiments

Two experiments are planned in this section. These experiments are to validate the assumptions used during the design of the algorithm in chapter 3.

- The first experiment is to check the accuracy of the weighted matrix used for computing the similarity score between celebrity choices. An alternate approach is also discussed.
- The second experiment is to check the accuracy of the equations used in generating the personality dimensions from Facebook profile features. The equations were synthesized from the charts available in the research done by Bachrach et al (2012) [3].

## 5.2.1 Experiment 1 - Celebrity Choices

### 5.2.1.1 Overview

In this experiment, the celebrities chosen by the users are analyzed for similarity to the user's personality profile. If there are celebrities that are defined as 'best matches' to the user's personality type, then it means the user has a high preference for people who are similar to themselves.

### 5.2.1.2 Hypothesis

*Can the weights assigned for 'best match' and 'compatible match' scenarios in table 3.3, be used to compare two users?*

A weighted matrix was used in section 3.3.2 and shown in table 3.3, which assigned the value of 1.0 to a best match, 0.5 to a compatible match and 0.1 others. These values were chosen arbitrarily and assigned based on two levels. The value of 0.1 was given to other matches to avoid nullification of the average score when multiplied by zero.

### 5.2.1.3 Procedure

Participants are asked to choose celebrities that they look up to or those that have qualities which they desire in their partner. Participants are free to choose as many celebrities as they like. A restriction is that they are not allowed to add a celebrity that is not present in the application since the list of celebrities is prepared with personality profile information tagged with each celebrity. Only the list of users that have attended the Myers-Briggs personality test are used for analysis since the celebrities information has only Myers-Briggs personality type results.

### 5.2.1.4 Analysis

The personality types as per Myers-Briggs type indicator for each celebrity is listed. The compatibility value for personality types is taken from table 3.3 and also listed. The average of the compatibility values for each person with their celebrity list is then computed and shown as the similarity score on a scale of 0 to 1. The similarity score is computed from the equation defined in section 3.3.2. The anonymized data is available in table 5.1.



Table 5.1: Experiment 1 - Celebrity Data

#	User	Type	Celebrity List	Type	Match Value	Average Score	Similarity Score
1	U7835	ISFJ	Abraham Lincoln Abraham Maslow Alan Alexander Milne Albert Einstein Andy Griffith Carol Burnett Carol Moseley-Braun Chevy Chase Colin Powell Elizabeth II	ENFJ ENFJ INFP INTP ENFJ ENFP INFJ INTJ INTJ ISFJ	1.0 1.0 0.5 0.1 1.0 0.1 0.5 0.1 0.1 1.0	0.54	1.0
2	U1041	INFP	Abraham Lincoln Albert Einstein Alfred Hitchcock Arthur Conan Barack Obama Bruce Willis Diane Sawyer Donna Reed George Washington John F. Kennedy Mel Gibson Michael Jordan Mother Teresa, Calcutta Oprah Winfrey Robin Williams Shirley Temple Black Socrates St. John St. Teresa of Avila Steve Jobs Tom Cruise	ENFJ INTP ENTP ESTP ENFJ ESTP ENFJ INFP ISTJ INTJ INFJ ENFJ INFJ ENFJ ENFP INFJ INTP INFP ISFJ ENTJ ISTP	1.0 0.5 0.5 0.1 1.0 0.1 1.0 1.0 0.1 0.1 1.0 1.0 1.0 1.0 1.0 0.5 1.0 0.5 0.1 0.1	0.65	1.0

Table 5.1: Experiment 1 - Celebrity Data

#	User	Type	Celebrity List	Type	Match Value	Average Score	Similarity Score
3	U6895	ESFP	Albert Einstein	INTP	0.1	0.48	1.0
			Elizabeth II	ISFJ	0.5		
			Fred Astaire	ISFP	1.0		
			Hillary Clinton	INTJ	0.1		
			Isaac Newton	INTP	0.1		
			Jacqueline K Onasis	INFP	0.5		
			Jane Austen	INTJ	0.1		
			John F. Kennedy	INTJ	0.1		
			Johnny Depp	ENFJ	0.5		
			Madonna	ESTP	1.0		
			Marilyn Monroe	ISFP	1.0		
			Matthew Mc Conaughey	ENFJ	0.5		
			Michael Jackson	ISFP	1.0		
			Neil Diamondvocalist	INFP	0.5		
			Robert Downey Jr	ENFP	0.5		
			Robin Williams	ENFP	0.5		
			Shirley Temple Black	INFJ	0.5		
			Tom Cruise	ISTP	0.1		

Table 5.1: Experiment 1 - Celebrity Data

#	User	Type	Celebrity List	Type	Match Value	Average Score	Similarity Score
4	U1089	INFP	Alfred Hitchcock Ben Affleck Bruce Willis C. S. Lewis Clint Eastwood Dan Aykroyd David Eddie Murphy Ernest Hemingway Evander Holyfield Jane Austen Jerry Seinfeld Martin Luther King Jr. Rene Descartes Socrates St. John St. Luke St. Teresa of Avila Weird Al Yankovick Will Smith William Shakespeare	ENTP ENFJ ESTP INTJ ISTP INTJ ENFJ ESTP ESTP ISTJ INTJ INFJ INFJ INTP INTP INFP INFP ISFJ ENTP ENFP INFP	0.5 1.0 0.1 0.1 0.1 0.1 1.0 0.1 0.1 0.1 0.1 1.0 1.0 0.5 0.5 1.0 1.0 0.5 0.5 1.0 1.0 0.5 1.0 1.0	0.54	1.0
5	U1261	ISFP	Bill Gates Franklin D. Roosevelt Hillary Clinton Michelle Obama Mother Teresa, Calcutta Nelson Mandela Oprah Winfrey Patrick Stewart Pete Sampras Steve Jobs William Shakespeare	ENTJ ENTJ INTJ INTJ INFJ INFJ ENFJ ENTJ ENFJ ENTJ INFP	0.1 0.1 0.1 0.1 0.5 0.5 0.5 0.1 0.5 0.1 0.5	0.28	0.5
6	U3682	ESFP	Barack Obama Hillary Clinton William J. Clinton	ENFJ INTJ ESFJ	0.5 0.1 0.5	0.37	0.5

Table 5.1: Experiment 1 - Celebrity Data

#	User	Type	Celebrity List	Type	Match Value	Average Score	Similarity Score
7	U1173	ISFJ	Abraham Maslow Albert Einstein C. G. Jung Carl Rogers Emily Bronte Martin Luther King Jr. Socrates Weird Al Yankovick	ENFJ INTP INTP INFP INTJ INFJ INTP ENTP	1.0 0.1 0.1 0.5 0.1 0.5 0.1 0.1	0.31	1.0
8	U5863	INTP	Albert Einstein Bill Gates	INTP ENTJ	1.0 0.5	0.75	1.0
9	U1047	ESTJ	Alfred Hitchcock Alfred Lord Tennyson Alicia Silverstone Aristophanes Arnold Schwarzenegger Arsenio Hall Woodrow Wilson Woody Harrelson Yogi Berra Zachary Taylor	ENTP ISFJ ENFP INFJ INTJ ESFP INTJ ESFP ISFP ISTP	0.5 1.0 0.1 0.1 1.0 0.5 1.0 0.5 0.5 1.0	0.62	1.0
10	U1041	INFP	Alfred Hitchcock Arthur Conan Bill Gates Clint Eastwood Sean Connery	ENTP ESTP ENTJ ISTP ENFJ	0.5 0.1 0.1 0.1 1.0	0.36	0.5
11	U6337	ISTP	Jerry Seinfeld Martin Luther King Jr. Tom Hanks Will Smith	INFJ INFJ ENTP ENFP	0.1 0.1 0.1 0.1	0.1	0.1
12	U1049	INTP	Bruce Willis	ESTP	0.5	0.5	0.5

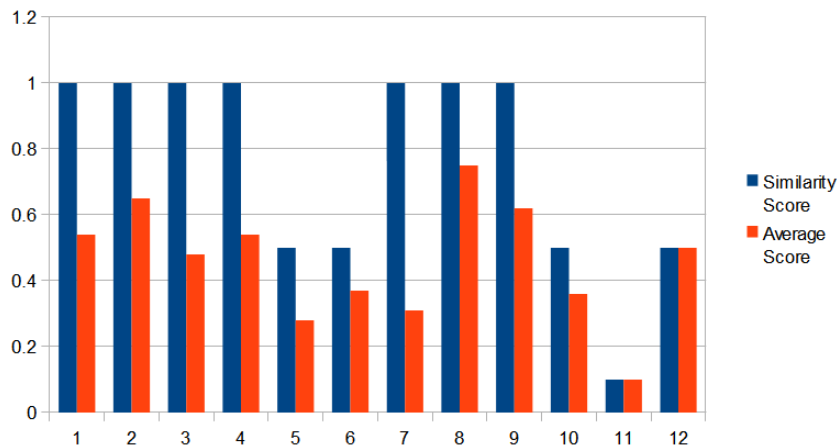
The similarity score is shown as a bar graph in figure 5.1 which shows the scores for different users. The mean for average score is 0.52 and the standard deviation 0.182. There are some outliers

in the data with few participants choosing around 20 celebrities and some less than 5. Majority of those who have chosen atleast 5 celebrities, have their average compatibility score within 2 sigma which confirms majority of users sub-consciously manage to choose celebrities who are similar to them.

Considering the similarity score computed as per equations in section 3.3.2, the mean is 0.75 and the standard deviation is 0.375 which includes almost all of the users except for the one which does not have any celebrities with personality types in the best match or compatible match personality types. Thus *the hypothesis is valid*. The values seem to follow a step-like pattern which is due to the presence of the maximum function used to compute the similarity score. Thus the highest values from the set of celebrity personality type matches would only be used for the similarity score.

The celebrity choices made by these users have not been influenced by any indicator in the application. In fact, the celebrities were displayed in alphabetical order with a facility to search by names. The personality types of these celebrities were not shown in the celebrities page to avoid a manual comparison of their personality type with those of the celebrities and thus prevent rigging of results.

Figure 5.1: Celebrity Data



### 5.2.1.5 Alternate Approach

The current method of computing similarity score based on celebrity choices is to compare the list of celebrities for two users and then compute the compatibility score by comparing the personality type of each celebrity with all the celebrities in the other user's list, assigning scores based on compatibility level (*from table 3.3*) and then finding the maximum of these compatible scores.

Instead of having arbitrary values for compatibility of personality types, the count of best matches, preferred matches and no matches can be taken. Then add the count of best and preferred matches, and obtain the ratio w.r.t the total count. This can be expressed as an equation :

$$S^C = \frac{n_{BM}(CC) + n_{PM}(CC)}{n_{BM}(CC) + n_{PM}(CC) + n_{NM}(CC)}$$

where

$n_{BM}(CC)$  – Count of output from celebrity compatibility function that checks only best matches

$n_{PM}(CC)$  – Count of output from celebrity compatibility function that checks only preferred matches

$n_{NM}(CC)$  – Count of output from celebrity compatibility function that checks only for no match cases

The anonymized data for the new approach is available in table 5.2

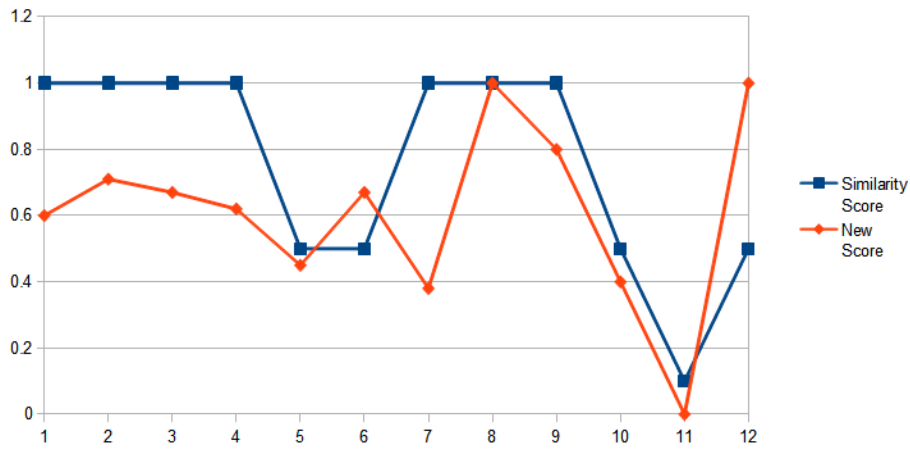
Table 5.2: Experiment 1 - New approach

#	User	Count	New Score	Similarity Score
1	U7835	BM : 4 PM : 2 NM : 4	0.6	1.00
2	U1041	BM : 11 PM : 4 NM : 6	0.71	1.00
3	U6895	BM : 4 PM : 8 NM : 6	0.67	1.00
4	U1089	BM : 8 PM : 5 NM : 8	0.62	1.00
5	U1261	BM : 0 PM : 5 NM : 6	0.45	0.50
6	U3682	BM : 0 PM : 2 NM : 1	0.67	0.50
7	U1173	BM : 1 PM : 2 NM : 5	0.38	1.00

#	User	Count	New Score	Similarity Score
8	U5863	BM : 1 PM : 1 NM : 0	1.00	1.00
9	U1047	BM : 4 PM : 4 NM : 2	0.80	1.00
10	U1041	BM : 1 PM : 1 NM : 3	0.40	0.50
11	U6337	BM : 0 PM : 0 NM : 4	0.00	0.10
12	U1049	BM : 0 PM : 1 NM : 0	1.00	0.50

Plotting the previously computed similarity score and the new similarity score side by side as a line graph (*refer figure 5.2*), it can be seen clearly that the values for the new approach tend to be greater than the previous approach. The new mean has increased to 0.8 and the standard deviation has jumped to 0.278 which is an increase of 6.67% and 12.58% respectively, compared to the previous approach. It can be seen from table 5.2, when the number of celebrities chosen are less than 3, the new score tends to be on the boundary i.e. either 0 or 1. Thus for users that have celebrity choices more than 5, these two approaches can be used interchangeably depending on performance requirements.

Figure 5.2: Experiment 1 - New approach



### 5.2.1.6 Conclusion

The celebrities chosen by the participants were analyzed for similarities with the user. It was found that majority of the users chose celebrities who had compatible personality profiles. Such choices were not observed when the number of chosen celebrities were less than 5. An alternate approach for calculating compatibility with celebrities was also analyzed and it was found that the new approach yielded higher similarity scores compared to the previous approach and had similarity score on the boundary values of 0 and 1 when the number of celebrities were 3 or less.

## 5.2.2 Experiment 2 - Similarity with Facebook friends

### 5.2.2.1 Overview

In this experiment, a similarity score is calculated for users compared to the chosen Facebook friends. These friends should have joined the application so that their Facebook profile information can be accessed and processed. This condition has to be satisfied as this step is a confirmation from the user for sharing their social information without any privacy concerns. The users chosen for this experiment should have taken the Big Five personality test as the dimensions of personality deduced from the Facebook profile features is represented in the five dimensions of Big Five personality model.



### 5.2.2.2 Hypothesis

The dimensions of personality from Facebook profile features were influenced and built upon from the research done by Bachrach et al (2012) [3]. The equations have been synthesized from the graphs provided in the research paper. So the hypothesis to be tested is :

*Are the personality dimensions deduced from the synthesized equations accurate enough to be used for computing a similarity score between users?*

### 5.2.2.3 Procedure

The users are asked to choose Facebook friends who they admire the most and those who possess the desired qualities of their future partner or date. These friends can be invited to join the application and share their data. If they did not join, their information is avoided. Only the comparison of those who are currently users of the application are considered. The similarity score between two users is calculated based on the squared euclidean distance between dimensions of the personality (*refer section 3.4.2*).

### 5.2.2.4 Analysis

The anonymized data for users and their chosen Facebook friends are available in table 5.3. Only those users who has a Big Five personality test result combined with Facebook friends who are users of the application developed for this research is available in the table. The cells marked with a hyphen represent no data. The hypothesis can be tested by checking how the similarity score varies with different combinations of users and Facebook friends.

Table 5.3: Experiment 2 - Facebook data

#	User	Friend	E*	A*	C*	S*	I*	Distance	Similarity Score
1	U7835		0.4349	0.4153	0.4993	0.4684	0.4308	0.00692	0.99310
		U3682	0.4212	0.3354	0.4854	0.4809	0.4349		
2	U3741		0.4344	0.4131	0.5000	0.4680	0.4298	0.00659	0.99342
		U3682	0.4212	0.3354	0.4854	0.4809	0.4349		
3	U7538	-	-	-	-	-	-	-	-
4	U1018	-	-	-	-	-	-	-	-
5	U1041	-	-	-	-	-	-	-	-

Table 5.3: Experiment 2 - Facebook data

6	U6895		0.4524	0.4131	0.5000	0.4497	0.4298	0.00638	0.99364
		U6337	0.4440	0.3354	0.4854	0.4572	0.4349		
7	U6012		0.4440	0.3354	0.4854	0.4572	0.4349	0.00108	0.99891
		U3682	0.4212	0.3354	0.4854	0.4809	0.4349		
8	U1089		0.4440	0.3354	0.4854	0.4572	0.4349	0.00656	0.99356
		U7584	0.4344	0.4131	0.5000	0.4680	0.4298		
9	U1261		0.4212	0.3354	0.4854	0.4809	0.4349	0.00000	1.00000
		U3682	0.4212	0.3354	0.4854	0.4809	0.4349		
10	U3682	-	-	-	-	-	-	-	-
11	U1023	-	-	-	-	-	-	-	-
12	U5863	-	-	-	-	-	-	-	-
13	U1041	-	-	-	-	-	-	-	-
14	U6609	-	-	-	-	-	-	-	-
15	U1080	-	-	-	-	-	-	-	-
16	U1049	-	-	-	-	-	-	-	-
17	U7584	-	-	-	-	-	-	-	-
18	U6337		0.4440	0.3354	0.4854	0.4572	0.4349		
		U3682	0.4212	0.3354	0.4854	0.4809	0.4349	0.00108	0.99892
		U6895	0.4524	0.4131	0.5000	0.4497	0.4298	0.00638	0.99364
19	U2289		0.4212	0.3354	0.4854	0.4809	0.4349		
		U3682	0.4212	0.3354	0.4854	0.4809	0.4349	0.00000	1.00000
		U7473	0.4344	0.4131	0.5000	0.4680	0.4298	0.00669	0.99343

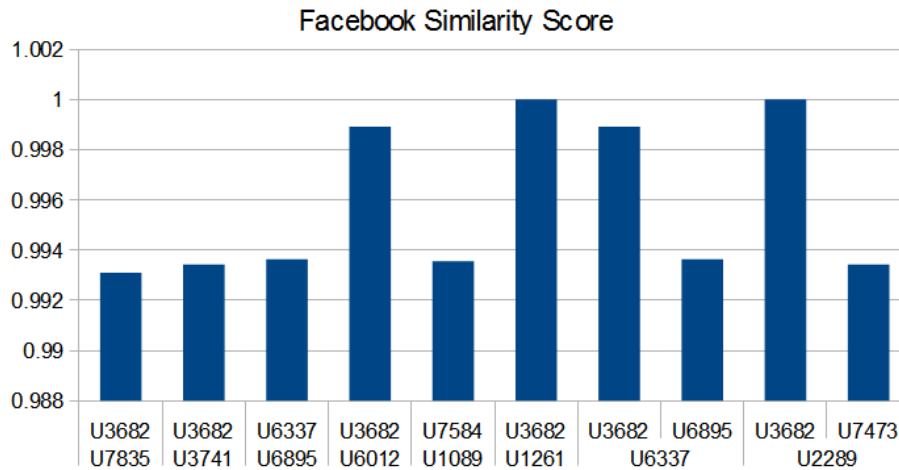
\* *E* - Extraversion, *A* - Agreeableness, *C* - Conscientiousness, *S* - Emotional Stability, *I* - Intellect

The user and Facebook friend combinations that have data in table 5.3 can be represented using a bar graph as shown in Figure 5.3. It can be found that all scores lie between 0.992 and 1.0 inclusive of the interval end points. The mean of the similarity scores is 0.995864 and standard deviation is 0.00312 representing a high level of concentration around the mean. Some of the reasons such behaviour is demonstrated are :

- The percentile computation gives very close values. Since the percentile depends on the values available for different Facebook features across all users in the application, the values of each of these features could be close by which gives very close percentile values and thus very low deviation from mean.

- The number of users is not big enough to provide a series of values that has enough distinct values to affect the percentile computation.
- The data for each of the Facebook profile features is not representative of actual life i.e. real-life events are not being accurately represented on Facebook and hence the values are very small or insignificant enough to make an impact.

Figure 5.3: Experiment 2 - Facebook friends data



The similarity scores are very high for the available data set from which it can be concluded there is a high level of similarity between the Big Five dimensions of personality of the user and the Facebook friends, thus confirming our hypothesis. But this observation cannot be stated as an absolute truth or as an objective statement due to lack of enough data for observing repeatability of results. Also some indicator regarding the extent of real-life translation of events to consumable digital social information, also has to be defined as a weighted measure to ensure that all ranges of values regarding Facebook features can be analyzed accurately within a certain threshold.

### 5.2.2.5 Conclusion

The Facebook similarity scores for users and their chosen Facebook friends are computed. The similarity scores were very close to the boundary value of 1.0 and also showed a tendency to be very close to the mean. The reasons for such clustering and possible solutions to introduce some sparsity to the data, were also discussed. It is assumed that when there is more social information about users available to the application, there would be a much graceful separation in similarity scores.

### 5.3 Summary

Two experiments were conducted to validate the assumptions during the design phase (*Chapter 3*). It was found that in case of celebrities, the majority of the similarity scores were above the mean. An alternate approach to calculate the similarity score was discussed. This approach is comparable to the existing approach in terms of the scores and could be a bit faster in terms of computation depending on the number of operations involved. In the case of Facebook based similarity scores, they showed high clustering around the mean. Thus it can be concluded that users tend to associate themselves very well with examples of people they know in real-life and also with their friends in social networks.

Aggregate information of the data collected through the application is given in Table 5.4 :

Table 5.4: User data statistics

SI No.	Description	Count
1	No of Users	57
2	User Ratings	87
3	Date Ratings	9
4	Attended 1 personality test	3
5	Attended 2 personality tests	4
6	Attended 3 personality tests	16
7	No of celebrities chosen	126
8	No of Facebook friends chosen	184
9	Average user rating	7.26
10	Average date rating	7.89

# Chapter 6

## Conclusion

This research involved the design of an online dating algorithm and development of a web application for implementing the algorithm. Three goals were outlined in chapter 1 which were planned to be achieved by the end of this research.

The first goal was to use machine learning for improved prediction. Even though a complete machine learning system with automatic feedback has not been implemented, still similarity between two users based on a personality model has been designed and implemented. Using this model, a similarity score is computed based on euclidean distance and then matches are proposed to users based on decreasing order of scores.

The second goal was to use exemplars for understanding user preferences without asking the user to fill in a long questionnaire about their preferences. This was accomplished using celebrities as the exemplar for this study. The personality types of the celebrities were used to find out which are the compatible personality types and then match between users were defined according to the similarity score computed using the mathematical model for celebrities. An assumption was made about a weighted matrix for compatible types and this was evaluated to be accurate enough to be used in the dating algorithm.

The third goal was to use social information available on social networking websites like Facebook to understand the holistic view of the person. This was implemented by using features from Facebook profiles like count of photographs, number of group memberships, number of tags, count of likes and number of friends; to create a Big Five personality model based on which similarity scores between two users were calculated. The accuracy using the representation of Facebook features as a personality model was evaluated and found to be accurate enough to be used in the dating algorithm. More validation in terms of repeatability has not been achieved though.

In addition to the above models, collaborative filtering technique has also been used generating recommendations. The users were asked to rate others on a scale of 1 to 10. These ratings and similarity scores of the users who received a rating, were then used to find potential matches thus utilizing a two-level approach for proposing matches.

There were challenges faced during the implementation with compatibility issues for the Facebook PHP SDK and in some cases with browsers. These took time to resolve but were done successfully. The translation of the design to the appropriate schematics of PHP language was easy due to pre-defined APIs available for processing of data structures and accessing files.

The application developed has been successful to capture user information, process it accurately and propose matches to users. The users were able to view the date suggestions and then provide a rating on how suitable they found the proposal to be. This date feedback information has not been used for this research but can be used at some point in future as a feedback mechanism for the dating algorithm to improve date matches.

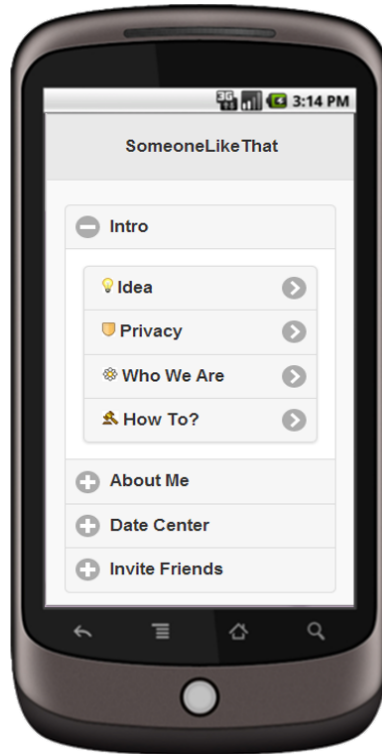
# Chapter 7

## Future Work

The number of participants for this study were around 60 which is not the ideal dataset to analyze characteristics of a dating algorithm. The effectiveness of the algorithm can be measured more accurately if a large number of users - around 100,000, sign up for the application similar to what Bachrach et al (2012)[3] has done in their study. This would allow statistical analysis to be done on large amounts of data even when data pruning has to be done. The users for this application were mostly from friends, fellow students and acquaintances. Due to various reasons many of those who agreed to be part of this study did not participate or backed out at the last minute. Partnering with local groups like “meetups” or other dating services might increase the number of participants and also help in gathering more data.

The current version of the application was hosted as a Facebook canvas application which required to be signed in using a large screen device having atleast 1280 x 960 pixels resolution. Since the number of smartphones has increased exponentially, having a smartphone application might increase the adoption rate. The design of the whole application would have to be changed for compatibility with various form factors of small screen devices. Integration with Facebook would have be done using Android SDK, iOS SDK or Javascript SDK depending on the operating system of the smartphone. One possible design is shown in figure 7.1.

Figure 7.1: Mobile version



In the mobile prototype, the various navigation options are shown in the landing screen. The different functionalities would be shown in a new window by clicking on the various options. The icons used have also been reduced in size to match the screen resolution. The content has also been reformatted by reducing the font size and changing the font type to fit the screen.

The next step would be to create a prototype specific for release to public, similar to the online dating websites available today. The website would not be running as a Facebook application as such but would connect to Facebook in the background to fetch social information thus retaining the user experience specific to the prototype. As more and more users sign up to the prototype, the existing algorithm can be tweaked to work with varying datasets and also to calibrate performance. Including information other social networking platforms like Google+ and Twitter to gain more information about the user, can also be looked into.



# Appendix A

## Research Proposal

Title : Online Dating in a Social Media Framework

Summary : The aim of this project is to learn dating recommendation algorithms in the context of social networks such as Facebook. We want to learn user preferences and behaviour from the social network instead of asking a user to fill a detailed manually-designed questionnaire. Additionally, we want to learn from exemplars i.e. desirable people on facebook.

The project will involve the following components :

1. Learning the best user interface required to gather important attributes for the purposes of learning this algorithm.
2. Using psychological studies such as Myers-Briggs evaluation to find a quantitative representation for people in addition to attributes such as their Facebook likes, sociability, interests etc.
3. Learning from parallel studies in metric learning, matching and recommendation based algorithms (Netflix etc.).
4. The implementation and timelines for this project must be appropriate in order to capture user interest and gather a substantial database, to enable us to quantitatively evaluate the methods we try.
5. Finally, we aim to learn from past "dates" to feed back into our system and improve recommendations on this basis.

Methods and Measurements Used : The user and Facebook graph data collected would be used to deduce the initial values of various personality traits of the user by their public activity on Facebook such as the number of groups they are a member of, the number of pictures they have uploaded, the number of likes, the number of friends etc. Users can also choose the celebrities that are similar in personality. The data would be collected and anonymized for privacy. To understand the personality of the user, three psychological tests are used – Myers Briggs, Big Five and Fisher.

The tests are adapted from free online personality tests that are publicly available. The sources would be credited in the web application. The test details and marking are provided with this application for reference. The data thus collected would be aggregated and using machine learning algorithms, predict possible matches. The algorithms would be improved over a period of time to increase the effectiveness.

Participants : Any student of TCD and any member of public can participate in this study. Participants are invited through facebook by sending application invites. All participants would be above the age of 18 years. Every Facebook application has a dashboard where various settings of the application can be controlled including age. Once the age limit has been imposed, users of age 18 and above only would be able to access the application. We are expecting 100 users to sign up for the application irrespective of gender or age. If less number of people sign up within the time, then we would use data from the signedup users.

Debriefing : The details of the experiment would be outlined in the Facebook application with links to details of the study. These would be displayed when the user joins the application in Facebook and the user can thus decide whether they would want to join this study or not, even before they accept the application invitation.

Ethical Considerations : The feedback from the participants about their dating experience (real or virtual) or feedback about someone, would have to be truthful and this would require confirmation that they have responded truthfully. We would have a confirmation section for this purpose.

Duration : This study would be open from the date when approval is received till 31st August 2014. At the end of this period, a notification/e-mail would be sent out to all the participants informing them that the study is over and that the application would be decommissioned.

Data Retention and Protection : A facebook application would be used to record input from the user and involves collecting the following data :

- Name
- Age
- Gender
- Brief Description
- Gender and age range for recommendations

- List of friends with details of name, gender and profile picture
- Number of likes
- Number of photographs uploaded
- Number of groups
- Rating and comments on users/recommendations
- Celebrities with similar personality
- Personality profile through personality evaluation tests

Any other publicly available information in the user's profile might be used for refining the algorithm. The user has complete control over the data displayed in the Facebook profile and can choose to hide any section as required. The data collected is required for constructing a statistical model for each person and then running algorithms to map people based on their personality traits information that is captured through the application. The users can provide their feedback on recommendations as ratings or comments which in turn would be used to improve the algorithms.

Before submission of user specific data, the user has to confirm access of their information using Facebook API. The data collected would be saved for further processing. Once the processing phase starts, the names would be anonymized for privacy and the rest of the details used for arriving at conclusions.

Any participant can opt out of the study at any point of time by disassociating the application from their apps list in Facebook. Any changes to the user's profile after opting out, would not be captured by the facebook application. Only data that has been accessed till that date would be retained. Information about the study would be provided in the form of a link or described within a section of the application.

There would be a feedback form for the user to fill in details of their experience in using the application like any problems in using the application, any changes recommend and a rating for the application. Response from the user for this form, is optional. Any data if received, would be analyzed and appropriate changes made to the web application.

Project Contact Details :

Researcher : Allen Thomas Varghese (vargheat@tcd.ie)

Course : MSc in Mobile and Ubiquitous Computing (2013/14)

ID : 13311840


Trinity College Dublin

Thesis Supervisor : Dr Mukta Prasad (prasadm@tcd.ie)

Course Director : Dr. Ciarán Mc Goldrick (Ciaran.McGoldrick@scss.tcd.ie)

# Appendix B

## Ethics Approval

Allen Thomas Varghese <vargheat@tcd.ie>

---

**RE: Re-Approval required for modifications? - 053/14**

**Tricia Fowler** <Tricia.Fowler@scss.tcd.ie>  
Reply-To: Tricia.Fowler@scss.tcd.ie  
To: Allen Thomas Varghese <vargheat@tcd.ie>  
Cc: Research Ethics <research-ethics@scss.tcd.ie>

11 February 2014 16:16

Hi Allen

Thank you for these revisions. You may now proceed with this study.

We wish you success in your research.

Kind Regards  
Tricia

Tricia Fowler  
Executive Officer – Research Unit  
School of Computer Science & Statistics  
O’Reilly Institute  
Trinity College  
Dublin 2

# Appendix C

## Big Five Personality Test

### Questionnaire[4]

1. Am the life of the party. (1+)
2. Feel little concern for others. (2-)
3. Am always prepared. (3+)
4. Get stressed out easily. (4-)
5. Have a rich vocabulary. (5+)
6. Don't talk a lot. (1-)
7. Am interested in people. (2+)
8. Leave my belongings around. (3-)
9. Am relaxed most of the time. (4+)
10. Have difficulty understanding abstract ideas. (5-)
11. Feel comfortable around people. (1+)
12. Insult people. (2-)
13. Pay attention to details. (3+)
14. Worry about things. (4-)
15. Have a vivid imagination. (5+)
16. Keep in the background. (1-)
17. Sympathize with others' feelings. (2+)
18. Make a mess of things. (3-)
19. Seldom feel blue. (4+)
20. Am not interested in abstract ideas. (5-)
21. Start conversations. (1+)
22. Am not interested in other people's problems. (2-)
23. Get chores done right away. (3+)

24. Am easily disturbed. (4-)
25. Have excellent ideas. (5+)
26. Have little to say. (1-)
27. Have a soft heart. (2+)
28. Often forget to put things back in their proper place. (3-)
29. Get upset easily. (4-)
30. Do not have a good imagination. (5-)
31. Talk to a lot of different people at parties. (1+)
32. Am not really interested in others. (2-)
33. Like order. (3+)
34. Change my mood a lot. (4-)
35. Am quick to understand things. (5+)
36. Don't like to draw attention to myself. (1-)
37. Take time out for others. (2+)
38. Shirk my duties. (3-)
39. Have frequent mood swings. (4-)
40. Use difficult words. (5+)
41. Don't mind being the center of attention. (1+)
42. Feel others' emotions. (2+)
43. Follow a schedule. (3+)
44. Get irritated easily. (4-)
45. Spend time reflecting on things. (5+)
46. Am quiet around strangers. (1-)
47. Make people feel at ease. (2+)
48. Am exacting in my work. (3+)
49. Often feel blue. (4-)
50. Am full of ideas. (5+)

The numbers in parentheses after each item indicate the scale on which that item is scored (i.e., of the five factors: (1) Extraversion, (2) Agreeableness, (3) Conscientiousness, (4) Emotional Stability, or (5) Intellect/Imagination) and its direction of scoring (+ or -).

#### Scoring[4]

For + keyed items, the response

"Very Inaccurate" is assigned a value of 1,

"Moderately Inaccurate" a value of 2,

"Neither Inaccurate nor Accurate" a value of 3,

"Moderately Accurate" a value of 4,

"Very Accurate" a value of 5.

For - keyed items, the response

"Very Inaccurate" is assigned a value of 5,

"Moderately Inaccurate" a value of 4,

"Neither Inaccurate nor Accurate" a value of 3,

"Moderately Accurate" a value of 2,

"Very Accurate" a value of 1.

Once numbers are assigned for all of the items in the scale, sum all the values to obtain a total scale score.

# Appendix D

## Myers-Briggs Personality Test

### Questionnaire[36]

1. At a party do you:
  - a. Interact with many, including strangers
  - b. Interact with a few, known to you
2. Are you more:
  - a. Realistic than speculative
  - b. Speculative than realistic
3. Is it worse to:
  - a. Have your “head in the clouds”
  - b. Be “in a rut”
4. Are you more impressed by:
  - a. Principles
  - b. Emotions
5. Are more drawn toward the:
  - a. Convincing
  - b. Touching
6. Do you prefer to work:
  - a. To deadlines
  - b. Just “whenever”
7. Do you tend to choose:
  - a. Rather carefully
  - b. Somewhat impulsively
8. At parties do you:
  - a. Stay late, with increasing energy
  - b. Leave early with decreased energy



9. Are you more attracted to:
  - a. Sensible people
  - b. Imaginative people
10. Are you more interested in:
  - a. What is actual
  - b. What is possible
11. In judging others are you more swayed by:
  - a. Laws than circumstances
  - b. Circumstances than laws
12. In approaching others is your inclination to be somewhat:
  - a. Objective
  - b. Personal
13. Are you more:
  - a. Punctual
  - b. Leisurely
14. Does it bother you more having things:
  - a. Incomplete
  - b. Completed
15. In your social groups do you:
  - a. Keep abreast of other's happenings
  - b. Get behind on the news
16. In doing ordinary things are you more likely to:
  - a. Do it the usual way
  - b. Do it your own way
17. Writers should:
  - a. "Say what they mean and mean what they say"
  - b. Express things more by use of analogy
18. Which appeals to you more:
  - a. Consistency of thought
  - b. Harmonious human relationships
19. Are you more comfortable in making:
  - a. Logical judgments
  - b. Value judgments
20. Do you want things:
  - a. Settled and decided
  - b. Unsettled and undecided
21. Would you say you are more:
  - a. Serious and determined
  - b. Easy-going

22. In phoning do you:
  - a. Rarely question that it will all be said
  - b. Rehearse what you'll say
23. Facts:
  - a. "Speak for themselves"
  - b. Illustrate principles
24. Are visionaries:
  - a. somewhat annoying
  - b. rather fascinating
25. Are you more often:
  - a. a cool-headed person
  - b. a warm-hearted person
26. Is it worse to be:
  - a. unjust
  - b. merciless
27. Should one usually let events occur:
  - a. by careful selection and choice
  - b. randomly and by chance
28. Do you feel better about:
  - a. having purchased
  - b. having the option to buy
29. In company do you:
  - a. initiate conversation
  - b. wait to be approached
30. Common sense is:
  - a. rarely questionable
  - b. frequently questionable
31. Children often do not:
  - a. make themselves useful enough
  - b. exercise their fantasy enough
32. In making decisions do you feel more comfortable with:
  - a. standards
  - b. feelings
33. Are you more:
  - a. firm than gentle
  - b. gentle than firm
34. Which is more admirable:
  - a. the ability to organize and be methodical
  - b. the ability to adapt and make do

35. Do you put more value on:
  - a. infinite
  - b. open-minded
36. Does new and non-routine interaction with others:
  - a. stimulate and energize you
  - b. tax your reserves
37. Are you more frequently:
  - a. a practical sort of person
  - b. a fanciful sort of person
38. Are you more likely to:
  - a. see how others are useful
  - b. see how others see
39. Which is more satisfying:
  - a. to discuss an issue thoroughly
  - b. to arrive at agreement on an issue
40. Which rules you more:
  - a. your head
  - b. your heart
41. Are you more comfortable with work that is:
  - a. contracted
  - b. done on a casual basis
42. Do you tend to look for:
  - a. the orderly
  - b. whatever turns up
43. Do you prefer:
  - a. many friends with brief contact
  - b. a few friends with more lengthy contact
44. Do you go more by:
  - a. facts
  - b. principles
45. Are you more interested in:
  - a. production and distribution
  - b. design and research
46. Which is more of a compliment:
  - a. "There is a very logical person."
  - b. "There is a very sentimental person."
47. Do you value in yourself more that you are:
  - a. unwavering
  - b. devoted



Scoring [36] is done by the following steps :

1. Count the number of checks in each of the A and B columns, and total at the bottom.
2. Copy the totals for Column 2 to the spaces below the totals for Column 3. Do the same for Columns 4 and 6.
3. Add totals downwards to calculate your totals.
4. The letter with the highest score is the type for each section.

# Appendix E

## Fisher Personality Test

### Questionnaire[30]

#### SCALE 1

- \* I find unpredictable situations exhilarating
- \* I do things on the spur of the moment
- \* I get bored when I have to do the same familiar things
- \* I have a very wide range of interests
- \* I am more optimistic than most people
- \* I am more creative than most people
- \* I am always looking for new experiences
- \* I am always doing new things
- \* I am more enthusiastic than most people
- \* I am willing to take risks to do what I want to do
- \* I get restless if I have to stay home for any length of time
- \* My friends would say I am very curious
- \* I have more energy than most people
- \* On my time off, I like to be free to do whatever looks fun

#### SCALE 2

- \* I think consistent routines keep life orderly and relaxing
- \* I consider (and reconsider) every option thoroughly before making a plan
- \* People should behave according to established standards of proper conduct
- \* I enjoy planning way ahead
- \* In general, I think it is important to follow rules
- \* Taking care of my possessions is a high priority for me

- \* My friends and family would say I have traditional values
- \* I tend to be meticulous in my duties
- \* I tend to be cautious, but not fearful
- \* People should behave in ways that are morally correct
- \* It is important to respect authority
- \* I would rather have loyal friends than interesting friends
- \* Long established customs need to be respected and preserved
- \* I like to work in a straightforward path toward completing the task

### SCALE 3

- \* I understand complex machines easily
- \* I enjoy competitive conversations
- \* I am intrigued by rules and patterns that govern systems
- \* I am more analytical and logical than most people
- \* I pursue intellectual topics thoroughly and regularly
- \* I am able to solve problems without letting emotion get in the way
- \* I like to figure out how things work
- \* I am tough-minded
- \* Debating is a good way to match my wits with others
- \* I have no trouble making a choice, even when several alternatives seem equally good at first
- \* When I buy a new machine (like a camera, computer or car), I want to know all of its technical features
- \* I like to avoid the nuances and say exactly what I mean
- \* I think it is important to be direct
- \* When making a decision, I like to stick to the facts rather than be swayed by people's feelings

### SCALE 4

- \* I like to get to know my friends' deepest needs and feelings
- \* I highly value deep emotional intimacy in my relationships
- \* Regardless of what is logical, I generally listen to my heart when making important decisions
- \* I frequently catch myself daydreaming
- \* I can change my mind easily
- \* After watching an emotional film, I often still feel moved by it several hours later
- \* I vividly imagine both wonderful and horrible things happening to me
- \* I am very sensitive to people's feelings and needs
- \* I often find myself getting lost in my thoughts during the day
- \* I feel emotions more deeply than most people
- \* I have a vivid imagination
- \* When I wake up from a vivid dream, it takes me a few seconds to return to reality

\* When reading, I enjoy it when the writer takes a sidetrack to say something beautiful or meaningful

\* I am very empathetic

Scoring[30]

Total of your answers for Scale 1 : This is how much you have of the Explorer personality type. They are attracted to other Explorers.

Total of your answers for Scale 2 : This is how much you have of the Builder personality type. They are attracted to other Builders.

Total of your answers for Scale 3 : This is how much you have of the Director personality type. They are attracted to Negotiators.

Total of your answers for Scale 4 : This is how much you have of the Negotiator personality type. They are attracted to Directors.



## Appendix F

# Dating Websites in Ireland

Table F.1: Dating websites in Ireland

Website	URL	Description	Paid
Connecting Singles	<a href="http://connectingsingles.ie/">connectingsingles.ie/</a>	All features are free which includes forums, blogs, ecards, flowers, polls etc. and is community driven. Flooded with ads.	No
Another Friend	<a href="http://anotherfriend.com/">anotherfriend.com/</a>	Claims to be the largest Irish owned and operated dating site.	Yes
Love Life	<a href="http://lovelife.ie/">lovelife.ie/</a>	Has a mail message service for registered users	No
Match	<a href="http://ie.match.com">ie.match.com</a>	Irish version of the main website match.com	Yes
Flirt Box	<a href="http://flirtbox.ie/">flirtbox.ie/</a>	Meet with singles	No
Single Parents	<a href="http://singleparents.ie">singleparents.ie</a>	Dating for single parents and 100% free	Yes
Spark Dating	<a href="http://sparkdating.ie">sparkdating.ie</a>	Members from Northern Ireland and Republic of Ireland. Free except for making a donation to a charity of choice for EUR 25. Has 10,000 members.	No

<b>Website</b>	<b>URL</b>	<b>Description</b>	<b>Paid</b>
Connecting Singles	<a href="http://connectingsingles.ie/">connectingsingles.ie/</a>	All features are free which includes forums, blogs, ecards, flowers, polls etc. and is community driven. Flooded with ads.	No
Get Out	<a href="http://getout.ie">getout.ie</a>	Singles dating and dating events	Yes
The Meeting Point	<a href="http://themeetingpoint.ie/">themeetingpoint.ie/</a>	Dating for heterosexuals and homosexuals. The Meeting Point is run by Global Personals Limited, a UK company. Free to join.	Yes
Zoosk	<a href="http://zoosk.com">zoosk.com</a>	Uses behavioural matching based on preferences and behaviour. Uses location based search to find members near to your location. Total membership is 27 million.	Yes
Dating In Ireland	<a href="http://datinginireland.com">datinginireland.com</a>	Uses online personals to display profile information.	Yes
Plenty More Fish	<a href="http://plentymorefish.com/ie_dating/">plentymorefish.com/ie_dating/</a>	Have all kinds of dating and has over 2.5 million single profiles available.	Yes
Thirty Flirty	<a href="http://thirtyflirty.ie/">thirtyflirty.ie/</a>	Only Irish website for people over 30 yrs.	Yes
Parship	<a href="http://parship.ie/">parship.ie/</a>	Supports dating for straight and gays. Uses personality test to match people. In 11 countries across Europe. Has 52% of members as graduates.	Yes
Dating Buzz	<a href="http://datingbuzzireland.com/s/">datingbuzzireland.com/s/</a>	Started in 1997.	Yes

<b>Website</b>	<b>URL</b>	<b>Description</b>	<b>Paid</b>
Connecting Singles	<a href="http://connectingsingles.ie/">connectingsingles.ie/</a>	All features are free which includes forums, blogs, ecards, flowers, polls etc. and is community driven. Flooded with ads.	No
EliteSingles	<a href="http://elitesingles.ie/">elitesingles.ie/</a>	77% of members are university graduates and between ages 30 - 55. Have profile verification and security.	Yes
Mature Free And Single	<a href="http://maturefreeandsingle.com/ireland/">maturefreeandsingle.com/ireland/</a>	Dating for those over 40s. Setup in 2008	Yes
Marital Affair	<a href="http://maritalaffair.ie/">maritalaffair.ie/</a>	Adult and affair dating from 2006. Has office in UK.	Yes
Twos Company	<a href="http://twoscompany.ie/">twoscompany.ie/</a>	Based out of UK. Has members between 25 to 75 years old, across 32 counties.	Yes
Older Dating	<a href="http://olderdatingonline.com/ie/">olderdatingonline.com/ie/</a>	Dating site for those over 50s. Based out of UK	Yes
Intro	<a href="http://intro.ie/">intro.ie/</a>	Matchmaking dating agency for lasting relationships. Provides support in setting up dates	Yes
Speed Dater	<a href="http://speeddater.ie/">speeddater.ie/</a>	Speed dating and singles parties	Yes
50 Plus Club	<a href="http://50plus-club.ie/">50plus-club.ie/</a>	50+ dating and senior singles	Yes
We Love Dates	<a href="http://welovedates.com/">welovedates.com/</a>	Spread across 25 counties	Yes
Irish Internet Dating	<a href="http://irishinternetdating.com/">irishinternetdating.com/</a>	Free signup and searching for members. Paid for unlocking features.	Yes
Northern Ireland Dating	<a href="http://nidating.com/">nidating.com/</a>	Created in 2000 and run by local volunteers	Yes
Woo	<a href="http://woo.ie/">woo.ie/</a>	Online friendship and dating community	No

<b>Website</b>	<b>URL</b>	<b>Description</b>	<b>Paid</b>
Connecting Singles	<a href="http://connectingsingles.ie/">connectingsingles.ie/</a>	All features are free which includes forums, blogs, ecards, flowers, polls etc. and is community driven. Flooded with ads.	No
Find Chat n Date	<a href="http://findchatndate.com/">findchatndate.com/</a>	Specifically for chatting and flirting activities.	No
Find Me Irish Dates	<a href="http://findmeirishdates.ie/">findmeirishdates.ie/</a>	Has over 61,000 people signed up.	No
Singles 365 Ireland	<a href="http://singles365.com/ireland/">singles365.com/ireland/</a>	Over 3.5 million members. Profiles checked for authenticity.	Yes
Gay Ireland 4u	<a href="http://simplythebest.ie/">simplythebest.ie/</a>	Flirt, chat and meet gay people	Yes
Match Affinity	<a href="http://matchaffinity.ie">matchaffinity.ie</a>	Compatibility based dating and relationship site. Uses an affinity questionnaire to propose matches.	Yes

## Appendix G

# International Dating Websites

Table G.1: International Dating Websites

Website	URL	Description	Paid
EHarmony	eharmony.com/	Have patented The eHarmony Compatibility Matching System.	Yes
Plenty Of Fish	pof.com	Uses POF relationship chemistry predictor measure to understand five broad dimensions of personality	No
Match	match.com	Most widely used website in the world and has nearly 18 million members.	Yes
Lovestruck	lovestruck.com	Targets potential partners according to location and covers many of the major cities across the world.	Yes
Dating Direct	datingdirect.com	Has 20 million members across Europe and merged with match.com in 2009.	Yes

Table G.1: International Dating Websites

<b>Website</b>	<b>URL</b>	<b>Description</b>	<b>Paid</b>
Parship	parship.com/	PARSHIP uses a patented test, this time called The PARSHIP principle®, which analyses 32 personality traits and is based on an algorithm of 136 rules.	Yes
Muddy Matches	muddymatches.co.uk/	Started in 2006, Muddy Matches is aimed at 'muddies', who by the girls' definition is 'any person who loves the countryside and is not afraid of a bit of mud'.	Yes
Beautiful People	beautifulpeople.com	Online dating for beautiful people only. Existing members have to vote for new members.	Yes
Tastebuds	tastebuds.fm/	Music based online dating site.	Yes
Doing Something	doingsomething.co.uk/	Matches are proposed based on date idea that has been suggested.	Yes
Chemistry	chemistry.com/	People are recommended based on the personality test devised by Dr Helen Fisher.	Yes
JDate	jdate.com/	Catering specifically to people of Jewish faith or heritage	Yes

Table G.1: International Dating Websites

<b>Website</b>	<b>URL</b>	<b>Description</b>	<b>Paid</b>
OkCupid	okcupid.com	Employs a mathematical algorithm based on your answers to a series of questions, and tells what percentage match, friend, and enemy are with any given user. Started in 2004 and has a membership of little over 1.2 million.	Yes
Perfect Match	perfectmatch.com	Uses Duet® Total Compatibility System that matches the "whole" you - your personality, lifestyle, values and preferences.	Yes
Adult Friend Finder	adultfriendfinder.com	Large user base of members looking for a casual one-night hook-up or friend with benefits.	Yes
Get It On	getiton.com	Matching based on sexual compatibility	Yes
Date	date.com	Matches are done based on over 50 profile attributes and among 10 million members	Yes
Christian Mingle	christianmingle.com	Large user base of Christian singles looking for a faith-centered relationship.	Yes
Black Singles	blacksingles.com	Started in 2002 and caters only to the black population	Yes
Skout	skout.com	Social discovery based on location and community	Yes

Table G.1: International Dating Websites

<b>Website</b>	<b>URL</b>	<b>Description</b>	<b>Paid</b>
Tinder	gotinder.com	Uses gestures to indicate preference for other users	No
Cupid	cupid.com/	Dating service that lets people meet with each other directly	Yes
Ashley Madison	ashleymadison.com/	Caters specifically to dating by married people. Started in 2001 and has 13.2 million members	Yes
Veggie Date	veggiedate.org/	Dating for Vegetarian Singles and Vegetarian Social Networking	Yes
The Big And The Beautiful	thebigandthebeautiful.com/	Dating for those who are plus sizes	Yes
Singldout	singldout.com/	Exclusive dating site connecting single professionals via LinkedIn	Yes
Hinge	hinge.co/	People are introduced only through friend networks	Yes
Coffee Meets Bagel	coffeemeetsbagel.com/	Displays one match per day at noon.	Yes
True Beginnings	truebeginnings.com/	Has over 30 million registered users. Uses compatibility questionnaires for matching people.	Yes



## Appendix H

# Application Data Storage Format

The below section shows how data is represented internally and saved in JSON format. Details about each field is given within “/\*...\*/”. This data structure is replicated for each user. Some users might not have all the fields since they have not submitted all information.

```
{
/*      User ID */
  "id ":" UserABC"

/*      Basic information on user having following fields      */
,
  "basic info ":{

/*      Gender */
    "gender ":" Female"

/*      Year of birth */
    ,
    "year ":" 1980"

/*      Month of birth */
    ,
    "month ":" November"

/*      Day of birth */
    ,
    "day ":" 18"

/*      Self description*/

```

```

    ,      "desc ":" Complicated "
  }

/*      Active or Passive user      */
,      "match_pref ":" P "

/*      Big Five personality test questionnaire      */
,      "big5_qn ":{
        "qn_1 ":" 4 "
      ,      "qn_2 ":" 1 "
      ,      "qn_3 ":" 4 "
      ,      "qn_4 ":" 4 "
        . . .
        . . .
      ,      "qn_48 ":" 5 "
      ,      "qn_49 ":" 4 "
      ,      "qn_50 ":" 5 "
    }

/*      Big Five personality test result      */
,      "big5_val ":{
        "E ":" 47 "
      ,      "A ":" 48 "
      ,      "C ":" 41 "
      ,      "S ":" 28 "
      ,      "I ":" 46 "
    }

/*      Date on which user attended Big Five personality test      */
,      "big5_date ":" 2014 04 24 15:42:36 "

/*      Myers Briggs personality test questionnaire      */
,      "mbti_qn ":{
        "qn_1 ":" A "
      ,      "qn_2 ":" A "
      ,      "qn_3 ":" A "
        . . .
        . . .
      ,      "qn_68 ":" A "
    }

```

```

    ,      "qn_69": "A"
    ,      "qn_70": "B"
  }

/*  Myers Briggs personality test result      */
,  "mbti_val": {
    "EIType": "E"
    ,  "EIValue": "50"
    ,  "SIType": "S"
    ,  "SIValue": "15"
    ,  "TFType": "F"
    ,  "TFValue": "40"
    ,  "JPType": "P"
    ,  "JPValue": "50"
  }

/*  Date on which user attended Myers Briggs personality test      */
,  "mbti_date": "2014 04 24 15:57:33"

/*  Ratings given by the user to others      */
,  "user_rating": {
    "User78": {
      "id": "User78"
      ,  "rating": "7"
      ,  "date": "2014 05 02 15:26:33"
    }
    ,  "User22": {
      "id": "User22"
      ,  "rating": "7"
      ,  "date": "2014 05 02 15:58:33"
    }
  }

/*  Facebook friends chosen by the user      */
,  "fb_friends": ["User36", "User22", "User52", "User65"]

/*  Fisher personality test questionnaire      */
,  "fisher_qn": {
    "q1": "2"
  }

```

```

    ,      "q2": "2"
    ,      "q3": "3"
    ,      . . .
    ,      . . .
    ,      "q54": "0"
    ,      "q55": "3"
    ,      "q56": "3"
  }

/* Fisher personality test result */
,  "fisher_val": {
    ,      "E": "9"
    ,      "B": "14"
    ,      "D": "21"
    ,      "N": "55"
  }

/* Date on which user attended Fisher personality test */
,  "fisher_date": "2014 05 01 14:53:51"

/* Date recommendations for the user */
,  "dating_recommendations": [
    ,      [" User100747 ", 7]
    ,      [" User100974 ", 7]
    ,      . . .
    ,      . . .
    ,      [" User37 ", 7]
    ,      [" User36 ", 7]
  ]

/* Date on which date recommendations were generated */
,  "dating_recommendations_date": "2014 06 18 19:35:02"

/* Celebrities chosen by the user */
,  "celebrities": [" BarackObama ", " HillaryClinton ", " WilliamJ. Clinton "]

/* Big Fiver personality type from Facebook profile */
,  "fb_personality_type": {
    ,      "Type1": "E"
  }

```

```

,      "Type1Val":0.42118054609333
,      "Type2 ":"A"
,      "Type2Val":0.3353506443
,      "Type3 ":"C"
,      "Type3Val":0.48540662636667
,      "Type4 ":"S"
,      "Type4Val":0.48085195164667
,      "Type5 ":"I"
,      "Type5Val":0.4348862463
,      "date ":"2014 06 15 22:45:00"}

/*      List of mutual matches      */
,      "mutual_matches":{
,          "User22":{
,              "rating ":"10"
,              "date ":"2014 03 22 19:14:41"
,          }
,          "User78":{
,              "rating ":"10"
,              "date ":"2014 05 04 01:12:34"
,          }
,      }
}

```

# Appendix I

## Abbreviations

API	-	<u>A</u> pplication <u>P</u> rogramming <u>I</u> nterface
CF	-	<u>C</u> ollaborative <u>F</u> iltering
CSS	-	<u>C</u> ascading <u>S</u> tyl <u>S</u> heets
CSV	-	<u>C</u> omma <u>S</u> eparated <u>V</u> alues
CORS	-	<u>C</u> ross <u>O</u> rig <u>I</u> n <u>R</u> esource <u>S</u> haring
GUI	-	<u>G</u> raphical <u>U</u> ser <u>I</u> nterface
HTML	-	<u>H</u> yper <u>T</u> ext <u>M</u> arkup <u>L</u> anguage
HTTP	-	<u>H</u> yperText <u>T</u> ransfer <u>P</u> rotocol
ID	-	<u>I</u> Dentifier
IS	-	<u>I</u> nformation <u>S</u> ervices
JSON	-	<u>J</u> ava <u>S</u> cript <u>O</u> bject <u>N</u> otation
OOP	-	<u>O</u> bject <u>O</u> riented <u>P</u> rogramming
OS	-	<u>O</u> perating <u>S</u> ystem
Q & A	-	<u>Q</u> uestions & <u>A</u> nswers
REST	-	<u>R</u> epresentational <u>S</u> tate <u>T</u> ransfer
SCSS	-	School of <u>C</u> omputer <u>S</u> cience and <u>S</u> tatistics
SNS	-	<u>S</u> ocial <u>N</u> etworking <u>S</u> ites
SSL	-	<u>S</u> ecure <u>S</u> ockets <u>L</u> ayer
UI	-	<u>U</u> ser <u>I</u> nterface
URL	-	<u>U</u> niform <u>R</u> esource <u>L</u> ocator
XOR	-	<u>e</u> Xclusive <u>O</u> R

# Bibliography

- [1] ALSALEH, S., NAYAK, R., XU, Y., AND CHEN, L. Improving matching process in social network using implicit and explicit user information. In *Web Technologies and Applications*. Springer, 2011, pp. 313–320.
- [2] AMICHAH-HAMBURGER, Y., AND VINITZKY, G. Social network use and personality. *Computers in Human Behavior* 26, 6 (2010), 1289–1295.
- [3] BACHRACH, Y., KOSINSKI, M., GRAEPEL, T., KOHLI, P., AND STILLWELL, D. Personality and patterns of facebook usage. In *Proceedings of the 3rd Annual ACM Web Science Conference* (2012), ACM, pp. 24–32.
- [4] BIGFIVETEST. Source : "[ipip.ori.org/New\\_IPIP-50-item-scale.htm](http://ipip.ori.org/New_IPIP-50-item-scale.htm)". Last accessed on 06-Aug-2014.
- [5] BROWN, L. L., ACEVEDO, B., AND FISHER, H. E. Neural correlates of four broad temperament dimensions: Testing predictions for a novel construct of personality. *PloS one* 8, 11 (2013), e78734.
- [6] BUCKWALTER, J., CARTER, S., FORGATCH, G., PARSONS, T., AND WARREN, N. Method and system for identifying people who are likely to have a successful relationship, May 11 2004. US Patent 6,735,568.
- [7] CELEBRITYTYPES. Source : [www.celebritytypes.com](http://www.celebritytypes.com). Last accessed on 11-Aug-2014.
- [8] CENTRALSTATISTICSOFFICE. Source : <http://www.cso.ie/px/pxeirestat/Staire/SelectVarVal/Define.asp?maintable=CDD05&PLangSource>. Last accessed on 22-Aug-2014.
- [9] CHEMISTRY. Source : <http://www.chemistry.com/>. Last accessed on 17-Aug-2014.
- [10] CHEN, L., AND NAYAK, R. Social network analysis of an online dating network. In *Proceedings of the 5th International Conference on Communities and Technologies* (2011), ACM, pp. 41–49.
- [11] CHUN, H., KWAK, H., EOM, Y.-H., AHN, Y.-Y., MOON, S., AND JEONG, H. Comparison of online social relations in volume vs interaction: a case study of cyworld. In *Proceedings of the 8th ACM SIGCOMM conference on Internet measurement* (2008), ACM, pp. 57–70.

- [12] COGNITIVEFUNCTIONS. Source : [upload.wikimedia.org/wikipedia/commons/e/e6/CognitiveFunctions.png](http://upload.wikimedia.org/wikipedia/commons/e/e6/CognitiveFunctions.png). Last accessed on 11-Aug-2014.
- [13] DATINGMARKETSIZE. Source : <http://www.datingsitesreviews.com/article.php?story=Some-New-Facts-About-eHarmony>. Last accessed on 27-Aug-2014.
- [14] DIAZ, F., METZLER, D., AND AMER-YAHIA, S. Relevance and ranking in online dating systems. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval* (2010), ACM, pp. 66–73.
- [15] DODS, J. Source : [www.massmatch.com/mbti-2.php](http://www.massmatch.com/mbti-2.php). Last accessed on 11-Aug-2014.
- [16] DRHELENFISHER. Source : <http://blog.chemistry.com/2013/11/14/blueprint-of-the-mind-by-dr-helen-fisher/>. Last accessed on 23-Aug-2014, 11 2013.
- [17] EHARMONY. Source : <http://www.eharmony.com>. Last accessed on 17-Aug-2014.
- [18] ELLISON, N. B., AND HANCOCK, J. T. Profile as promise: Honest and deceptive signals in online dating. *Security & Privacy, IEEE* 11, 5 (2013), 84–88.
- [19] FACEBOOK. Source : <https://www.facebook.com/facebook>. Last accessed on 22-Aug-2014.
- [20] FACEBOOK-MEMBERSHIP. Source : [https://www.facebook.com/facebook/info?ref=page\\_internal](https://www.facebook.com/facebook/info?ref=page_internal). Last accessed on 22-Aug-2014.
- [21] FACEBOOKBLOG. Source : <https://developers.facebook.com/blog/>. Last accessed on 27-Aug-2014.
- [22] FACEBOOKBRANDWEBSITE. Source : <https://www.facebookbrand.com/>. Last accessed on 11-Aug-2014.
- [23] FACEBOOKDESIGNGUIDELINE. Source : <https://www.facebookbrand.com/>. Last accessed on 26-Aug-2014.
- [24] FACEBOOKDEVELOPERS. Source : <https://www.facebook.com/groups/fbdevelopers/>. Last accessed on 11-Aug-2014.
- [25] FACEBOOKLOGINPERMISSIONS. Source : <https://developers.facebook.com/docs/facebook-login/permissions/v2.0>. Last accessed on 11-Aug-2014.
- [26] FACEBOOKPHPSDK. Source : <https://github.com/facebook/facebook-php-sdk-v4/>. Last accessed on 11-Aug-2014.
- [27] FIORE, A. T., AND DONATH, J. S. Online personals: An overview. In *CHI'04 extended abstracts on Human factors in computing systems* (2004), ACM, pp. 1395–1398.



- [28] FIORE, A. T., TAYLOR, L. S., MENDELSON, G. A., AND HEARST, M. Assessing attractiveness in online dating profiles. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (2008), ACM, pp. 797–806.
- [29] FIORE, A. T., TAYLOR, L. S., ZHONG, X., MENDELSON, G. A., AND CHESHIRE, C. Who’s right and who writes: People, profiles, contacts, and replies in online dating. In *HICSS* (2010), pp. 1–10.
- [30] FISHERTEST. Source : [idigitalcitizen.files.wordpress.com/2009/08/text-and-instructions-for-why-him-why-her-test.pdf](http://idigitalcitizen.files.wordpress.com/2009/08/text-and-instructions-for-why-him-why-her-test.pdf). Last accessed on 06-Aug-2014.
- [31] FIVELABS. Source : <http://labs.five.com>. Last accessed on 22-Aug-2014.
- [32] FOX, J., AND WARBER, K. M. Romantic relationship development in the age of facebook: An exploratory study of emerging adults’ perceptions, motives, and behaviors. *Cyberpsychology, Behavior, and Social Networking* 16, 1 (2013), 3–7.
- [33] HANCOCK, J. T., TOMA, C., AND ELLISON, N. The truth about lying in online dating profiles. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (2007), ACM, pp. 449–452.
- [34] KUNEGIS, J., GRÖNER, G., AND GOTTRON, T. Online dating recommender systems: The split-complex number approach. In *Proceedings of the 4th ACM RecSys workshop on Recommender systems and the social web* (2012), ACM, pp. 37–44.
- [35] MUKTA PRASAD, A. T. V. Source : [apps.facebook.com/someonelikethat](https://apps.facebook.com/someonelikethat). Last accessed on 11-Aug-2014.
- [36] MYERSBRIGGS. Source : [dev.im.org/meetings/past/2007/documents/2007](http://dev.im.org/meetings/past/2007/documents/2007) Last accessed on 26-Aug-2014.
- [37] NAYAK, R., ZHANG, M., AND CHEN, L. A social matching system for an online dating network: a preliminary study. In *Data Mining Workshops (ICDMW), 2010 IEEE International Conference on* (2010), IEEE, pp. 352–357.
- [38] OKCUPID. Source : [www.okcupid.com](http://www.okcupid.com). Last accessed on 27-Aug-2014.
- [39] ONLINEDATINGSTATS. Source : <http://visual.ly/uk-online-dating-stats-dating-friends>. Last accessed on 27-Aug-2014.
- [40] PAUNONEN, S. V., AND HONG, R. Y. The many faces of assumed similarity in perceptions of personality. *Journal of Research in Personality* 47, 6 (2013), 800–815.
- [41] PEWRESEARCHCENTER. Couples, the internet and social media, source : [http://www.pewinternet.org/files/2014/02/PIP\\_Couples\\_and\\_Technology-FIN\\_021114.pdf](http://www.pewinternet.org/files/2014/02/PIP_Couples_and_Technology-FIN_021114.pdf). Last accessed on 22-Aug-2014.

- [42] PEWRESEARCHCENTER. Online dating & relationships source : [http://www.pewinternet.org/files/old-media//Files/Reports/2013/PIP\\_Online%20Dating%202013.pdf](http://www.pewinternet.org/files/old-media//Files/Reports/2013/PIP_Online%20Dating%202013.pdf). Last accessed on 22-Aug-2014, October 2013.
- [43] PHPUSAGE. Source : [php.net/usage.php](http://php.net/usage.php). Last accessed on 11-Aug-2014.
- [44] PIZZATO, L., REJ, T., AKEHURST, J., KOPRINSKA, I., YACEF, K., AND KAY, J. Recommending people to people: the nature of reciprocal recommenders with a case study in online dating. *User Modeling and User-Adapted Interaction* 23, 5 (2013), 447–488.
- [45] PIZZATO, L., REJ, T., CHUNG, T., KOPRINSKA, I., AND KAY, J. Recon: a reciprocal recommender for online dating. In *Proceedings of the fourth ACM conference on Recommender systems* (2010), ACM, pp. 207–214.
- [46] ROSS, C., ORR, E. S., SISIC, M., ARSENEAULT, J. M., SIMMERING, M. G., AND ORR, R. R. Personality and motivations associated with facebook use. *Computers in Human Behavior* 25, 2 (2009), 578–586.
- [47] SUDAI, G., AND BLUMBERG, D. Method and apparatus for detection of reciprocal interests or feelings and subsequent notification, Sept. 7 1999. US Patent 5,950,200.
- [48] TERVEEN, L., AND McDONALD, D. W. Social matching: A framework and research agenda. *ACM transactions on computer-human interaction (TOCHI)* 12, 3 (2005), 401–434.
- [49] THEMYERS&BRIGGSFOUNDATION. Source : <http://www.myersbriggs.org/my-mbti-personality-type/mbti-basics/>. Last accessed on 23-Aug-2014.
- [50] TIEGER, P. D., AND BARRON-TIEGER, B. *Just your type: Create the relationship you've always wanted using the secrets of personality type*. Hachette UK, 2001.
- [51] YU, M., ZHAO, K., YEN, J., AND KREAGER, D. Recommendation in reciprocal and bipartite social networks—a case study of online dating. In *Social Computing, Behavioral-Cultural Modeling and Prediction*. Springer, 2013, pp. 231–239.
- [52] ZADEH, S., MEHR, A., AND GOTLIEB, C. System and method for identifying nearby, compatible users, June 6 2013. US Patent App. 13/706,182.
- [53] ZHAO, S., GRASMUCK, S., AND MARTIN, J. Identity construction on facebook: Digital empowerment in anchored relationships. *Computers in human behavior* 24, 5 (2008), 1816–1836.
- [54] ZHAO, X., SCHWANDA SOSIK, V., AND COSLEY, D. It's complicated: how romantic partners use facebook. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (2012), ACM, pp. 771–780.
- [55] ZOOSK. Source : [www.zoosk.com](http://www.zoosk.com). Last accessed on 22-Aug-2014.