

DESIGN AND ANALYSIS
OF SOCIAL NETWORK
SYSTEMS (SNS)

SYED MUHAMMAD ALI ABBAS

A thesis submitted in partial
fulfillment of the requirements
of the
Manchester Metropolitan
University for the degree of
Doctor of Philosophy

Centre for Policy
Modelling
Manchester Metropolitan
University Business
School

2016

To My Beloved Parents

ACKNOWLEDGEMENTS

I would like to thank my supervisor and Director of Studies, Bruce Edmonds, for your invaluable support, enthusiastic encouragement and useful critiques during the entire tenure of my PhD. Without your guidance this thesis would not have been possible. I have learnt a great deal from you. You have always stood by me and have helped me solve many technical and administrative issues. I also have a deep gratitude for the whole CPM (and its affiliates) family: Stefano, Ruth, Emma, Richard, Magnus, Pablo, Claire and Chris. You have all been nothing but amazing throughout my stay there. I am really grateful and thankful to the whole [MMU](#) staff for their support, in particular the research administrative staff, especially Diane Smith and Paul Duffy.

I would like to thank the Dean of the [MMU](#) Business School for offering me a prestigious scholarship to carry out my doctorate.

I would like to thank my collaborators and helpers without whom I would not have been able to complete my work. Firstly, I would like to thank Pablo Mateos, who was based at the Department of Geography UCL and is now an Associate Professor in Mexico, for your collaboration towards ethnic classification of my data. You have been extremely helpful. Secondly, I would like to thank Alan Mislove, who is now an Assistant Professor at the Northeastern University in the US, for your help and guidance towards data collection from Facebook.

I would like to thank my whole family for their immense encouragement, support and love, especially my brother, Adeel, and my sister-in-law, Shaherbano, for tolerating me and mainly providing for me throughout my stay in Preston. It was such a lovely time.

I would also like to thank my friend, Rameez. You have been a great source of inspiration and your influence has had a great effect on me. Time and again you have been there to support and also guide me. I have much respect for you.

Abstract

In the last few years, online Social Network Systems (SNSs) thrived and changed the overall outlook of the Internet. These systems play an important role in making the Internet social, a hallmark of Web 2.0. Various such systems have been developed to serve a diverse set of needs. [SNSs](#) provide not only a space for self-representation, but also mechanisms to build and maintain one's social network online. A lot of studies have been carried out on such systems to identify how people develop cultures of communication, sharing and participation and also to identify the network structure of such systems. In this thesis, we carry this line of research forward. Our aim is the identification of some key user characteristics and social processes which result in the emergence of a social network. These might help future platform and application developers in creating better, more efficient and more open and user-friendly [SNSs](#). Specifically, we make the following three major contributions:

a) One of the distinct features of an [SNS](#) is the public listing of friendship links - social network. Most of the personal details such as hometown and workplace information have been hidden from non-friends, but the list of friendships remains open. Being a true representation, people use their real names as their screen names. Such names alone contain detailed cultural information about their ethnicities, religion and even their geographical origins. Our first contribution is that we have made good use of such information by inferring ethnic classification of users of Facebook. We identified how clustered and segregated the overall social network is when users' inferred ethnicity is taken into account. Different cultures have different behaviours with distinct characteristics. This rich information can be used to develop an understanding and help create diverse applications catering for specific ethnicities and geographical regions; covering both the dominant and non-dominant groups. We have identified ethnicities of a subset of Facebook users with their friends and studied how different ethnicities are connected among and within each other. A large social network dataset of four thousand Manchester Metropolitan University ([MMU](#)) students have been selected from Facebook. We have extensively analysed this dataset for its network structure and also its semantic and social structure. Our work suggests our dataset is clustered and segregated on ethnic lines.

b) To develop a user liberating [SNS](#) where the control and the ownership of rich personal data is in the hands of [SNS](#) users, a clear understanding is required of how such systems on an individual and group level are developed and maintained. Never before in Social Sciences was it possible to study society on such a large scale. These systems have facilitated the study of individuals both at a local and global scale. However, at the moment very little knowledge is available to identify how people develop their friendship in reality.

So for example, it is not known whether in [SNSs](#) people meet others based on their attributes and interests, or if they simply bring online their real lives' social networks. And more specifically, what processes does one go through to develop her social network. To fill this knowledge gap in this thesis, as our second contribution, we have used a computer simulation technique known as Agent-Based simulation, to develop four simulation models based on both individuals' affinities and environmental aspects. Specifically, we have developed models of student interaction to develop social networks. Three University's datasets which include Caltech (Nodes 762, Edges 16651), Princeton (Nodes 6575, Edges 293307) and Georgetown (Nodes 9388, Edges 425619), have been used to check the performance and rigour of the model. Our evidence suggests that 'friend-of-a-friend' (FOAF) best represents social interactions in Caltech University. In the case of Princeton and Georgetown, we found a multitude of social and structural processes involved, which

are: attribute based (same dormitory, major or high school etc.), social interaction, random meet ups (through parties or other social events) and current friends introducing new friends.

c) We observe that in the main, [SNSs](#) are centralised, and depend solely on central entities for everything. With huge personal data on such [SNSs](#), advertising and marketing agencies have made very sophisticated systems to gather information about people. It is a goldmine for them for personalised advertisement. Also various governmental agencies have been using [SNSs](#) as an excuse to curb potential threats both legally and illegally, to obtain information on numerous users (people). In order to deal with such issues inherent in centralised client-server architecture, as the third contribution of this thesis, we have proposed and implemented a completely decentralised [SNS](#) in a peer-to-peer fashion. Our implementation is done in an open source Peer-To-Peer ([P2P](#)) client Tribler. To handle the dynamicity of users in a [P2P](#) system – their availability, we have developed mechanisms to deal with it. This [SNS](#) has been evaluated on a deployed system with real users. This prototype establishes the feasibility of a totally distributed [SNS](#), but its practicality when scaled to a full system would require more work.

Table of Contents

1	Chapter: Introduction	1
1.1	Research Questions.....	3
1.2	Thesis Outline	5
2	Chapter: Literature Review	7
2.1	Introduction	7
2.2	Agent-Based Modelling	7
2.2.1	Alternative Approaches	9
2.3	Social Networks	10
2.3.1	Regular Lattice	10
2.3.2	Random Networks.....	11
2.3.3	Power-Law Networks	12
2.3.4	Small-World Networks.....	13
2.3.5	Preferential Attachment.....	14
2.3.6	Comparison.....	14
2.4	Social Network System (SNS).....	15
2.4.1	History	15
2.4.2	Research Themes.....	17
2.4.2.1	Identity and Representation.....	17
2.4.2.2	Privacy.....	18
2.4.2.3	Growth.....	22
2.4.2.4	Segmentation	24
2.4.2.5	Trust	25
2.5	Distributed Social Network	25
2.6	Conclusion	27
3	Chapter: Analysis of Complex Network Datasets.....	28
3.1	Introduction	28
3.2	Social Network Analysis Measures	28
3.2.1	Node Degree Distribution	30
3.2.2	Assortativity Mixing.....	31
3.2.3	Cluster Coefficient.....	31
3.2.4	Geodesic Distance	31
3.2.5	Direct Similarity Measures.....	32
3.2.6	Eigenvalues and Eigenvector	32
3.2.7	Feature Extraction	32
3.2.8	Community Detection	33
3.2.9	Affinity	33
3.2.10	Silo Index	34

3.3	Cross-Sectional Datasets.....	34
3.3.1	Caltech.....	35
3.3.2	Princeton.....	38
3.3.3	Georgetown.....	41
3.3.4	Discussion.....	45
3.4	Critical Analysis.....	47
3.5	Conclusions.....	47
4	Chapter: Models of Student Interaction.....	49
4.1	Introduction.....	49
4.2	Overview.....	50
4.2.1	Purpose.....	50
4.2.2	Entities, State Variables, and Scales.....	50
4.2.3	Process Overview and Scheduling.....	51
4.2.3.1	Strategy 1 – Preferential Strategy.....	52
4.2.3.2	Strategy 2 – Friend of a Friend (FOAF) Strategy.....	53
4.2.3.3	Strategy 4 – Random Strategy.....	53
4.2.3.4	Strategy 4 – Hybrid Strategy.....	53
4.3	Design Concepts.....	53
4.3.1	Basic Principles.....	53
4.3.2	Emergence.....	54
4.3.3	Adaptation.....	54
4.3.4	Objectives.....	54
4.3.5	Learning.....	55
4.3.6	Prediction.....	55
4.3.7	Sensing.....	55
4.3.8	Interaction.....	55
4.3.9	Stochasticity.....	56
4.3.10	Collectives.....	56
4.3.11	Observations.....	56
4.4	Details.....	56
4.4.1	Initialisation.....	56
4.4.2	Input data.....	56
4.4.3	Sub-Models.....	57
4.4.4	Calibration and Validation.....	57
4.5	Limitations.....	59
4.6	Results.....	59
4.6.1	Caltech Results.....	59
4.6.1.1	Global Results.....	60

4.6.1.2	Attribute Level Results.....	63
4.6.1.3	Summary.....	64
4.6.2	Princeton Results.....	64
4.6.2.1	Global Results.....	64
4.6.2.2	Attribute Level Results.....	66
4.6.2.3	Summary.....	68
4.6.3	Georgetown Results.....	69
4.6.3.1	Global results.....	69
4.6.3.2	Attribute Level Results.....	70
4.6.3.3	Summary.....	72
4.7	Discussion.....	73
4.8	Conclusions.....	74
5	Chapter: Diversity and Clustering of MMU Students on Facebook.....	75
5.1	Ethics and Purpose.....	76
5.2	Facebook Graph (Reference Graph).....	77
5.3	Random Graph.....	81
5.4	Ethnic Classification.....	84
5.4.1	Onomap.....	84
5.5	Shortcomings.....	88
5.6	Data Sharing.....	90
5.7	Results.....	90
5.7.1	Affinity.....	90
5.7.2	Contracted Graphs.....	91
5.7.3	Ethnic Group.....	91
5.7.3.1	Comparison.....	94
5.7.4	Religion.....	95
5.7.4.1	Comparison.....	97
5.7.5	Geography.....	98
5.7.5.1	Comparison.....	100
5.7.6	Language.....	101
5.7.6.1	Comparison.....	103
5.8	Normalised Results.....	104
5.9	Summary of Results and Discussion.....	106
5.10	Conclusion.....	109
6	Chapter: Distributed Peer to Peer Social Network System.....	110
6.1	Introduction.....	110
6.2	Peer To Peer Networks.....	110

6.3	Background.....	111
6.4	Requirements of our System.....	112
6.4.1	Functionalities	112
6.4.2	Tribler.....	112
6.5	Detailed Design.....	113
6.5.1	Basic Request-Reply Protocol.....	113
6.5.2	Unavailability of the Peers	114
6.5.3	Scenarios for Establishing Friendship Links	115
6.5.3.1	Scenario 1	115
6.5.3.2	Scenario 2	116
6.5.3.3	Scenario 3	117
6.6	Experiments	118
6.6.1	Number of Friendship Link Establishment Requests	119
6.6.2	Total Time Taken for Receiving Friendship Replies	120
6.7	Prevention of Possible Attacks.....	120
6.8	Critical Analysis.....	121
6.9	Business Model.....	121
6.10	Future Work.....	122
6.11	Conclusions	123
7	Chapter: Conclusion	125
7.1	Contributions.....	125
7.2	Future Work	128
7.3	Recommendations for Future SNS Developers	128
7.4	Future SNSs	129
8	References	130
9	Glossary of Terms.....	141
Appendix A	Student Interaction Model	142
Appendix A.1	Main Step Function.....	142
Appendix A.2	Personal Preference Algorithm	144
Appendix B	MMU Facebook Dataset	146
Appendix B.1	Ethnic Group Distribution	146
Appendix B.2	Sub-ethnic Group Distribution	146
Appendix B.3	Geography Distribution	149
Appendix B.4	Religion Distribution.....	149
Appendix B.5	Language Distribution	150
Appendix C	MMU Facebook Graph vs. Random Graph.....	153
Appendix C.1	Ethnic Group Distribution.....	153
Appendix C.2	Sub-ethnic Groups.....	153

Appendix C.3	Religion	156
Appendix C.4	Geography.....	157
Appendix C.5	Language	158
Appendix D	MMU Facebook Graph Normalised Silo Index	162
Appendix D.1	Religious Group Normalised Silo Index	162
Appendix D.2	Geography Group Normalised Silo Index	162
Appendix D.3	Language Group Normalised Silo Index	163
Appendix E	R Code	166
Appendix E.1	Silo Index.....	166
Appendix E.2	Affinity Measure	167
Appendix E.3	Contract Graph	168
Appendix F	Annual Review Documents and Ethical Committee's letter.....	169
Appendix F.1	Progress Report 2009-10.....	169
Appendix F.2	RDF Form	171
Appendix F.3	Letter from the chair of the Ethics Committee	174

-

List of Figures

Figure 2-1 - Regular Lattice	11
Figure 2-2 - Random Network	12
Figure 2-3 - The degree distribution of a network formed via preferential attachment	13
Figure 2-4 - History of Social Network Systems (Boyd and Ellison 2007)	16
Figure 2-5 - Privacy Settings in 2005 (McKeon, n.d.)	21
Figure 2-6 - Privacy Settings in 2010 (McKeon, n.d.)	22
Figure 3-1 - An illustration of the problem of comparing and validating a class of simulated networks (left) to the available social networks of a target social system (right) (Abbas et al., 2014).	29
Figure 3-2 - Undirected Graph	30
Figure 3-3 - Caltech Social Graph with clusters	37
Figure 3-4 - Total Degree Distribution (log-log plot) of Caltech Social Network	38
Figure 3-5 - Princeton Social Graph	40
Figure 3-6 - Total Degree Distribution (log-log plot) of Princeton Social Network	41
Figure 3-7 - Georgetown Social Graph	44
Figure 3-8 - Total Degree Distribution (log-log plot) of Georgetown Social Network	45
Figure 4-1 Log-log plot of Total Degree Distribution of all the four simulation strategies and the reference dataset	62
Figure 4-2 Silo Index for Dorm, Major, Year and High School attributes for FOAF strategy and the Caltech reference network (the blue triangles represent the reference dataset (Caltech), and the red plus signs represent the FOAF strategy)	63
Figure 4-3 Log-log plot of Total Degree Distribution of all the four simulation strategies and the Princeton reference dataset.	66
Figure 4-4. Silo Index for Dorm, Major, Year and High School attributes for Hybrid strategy and the Princeton reference network (the blue triangles represent the reference dataset (Princeton), and the red plus signs represent the FOAF strategy)	67
Figure 4-5. Degree Mixing of Hybrid mode (bottom) and the Princeton reference dataset (top)	68
Figure 4-6 - Log-log plot of Total Degree Distribution of all the four simulation strategies and the Georgetown reference dataset.	70
Figure 4-7. Silo Index for Dorm, Major, Year and High School attributes for Hybrid strategy and the Georgetown reference network (the blue triangles represent the reference dataset (Georgetown), and the red plus signs represent the FOAF strategy)	71
Figure 4-8. Degree Mixing of Hybrid mode (bottom) and the Georgetown reference dataset (top)	72
Figure 5-1 - Total Degree Distribution (log-log plot) of MMU Social Network	80
Figure 5-2 - Simple Graph	82
Figure 5-3 Random Graph	82
Figure 5-4 - Total Degree Distribution (log-log plot) of Reference Graph and Random Graph (null model)	83
Figure 5-5 Onomap Group Social Graph	93
Figure 5-6 Normalised Links of Ethnic Groups	94
Figure 5-7 Ethnicity based Silo Index comparison between the reference and the null model	95
Figure 5-8 Religion Based Contracted Graph	96
Figure 5-9 Normalised Weights of Links Based On Religion	97
Figure 5-10 - Religion based Silo Index comparison between the reference and the null mode	98
Figure 5-11 Geography Based Contracted Graph	99
Figure 5-12 Normalised Links Based On Geographical Area	100
Figure 5-13 - Geography based Silo Index comparison between the reference and the null model	101
Figure 5-14 - Language Based Contracted Graph	102
Figure 5-15 Normalised Links Based On Languages	103
Figure 5-16 - Language Silo Index comparison between the reference and the null model	104
Figure 6-1 - Peer-To-Peer Network	111
Figure 6-2 - Friendship request retry mechanism	114
Figure 6-3 - Friendship request scenario 1: Both the source peer and the target peer are online	115

Figure 6-4 - Friendship request scenario 2: Both the source peer and the helpers try to contact the target peer.	116
Figure 6-5 - Friendship request scenario 3: On behalf of the source peer, the helpers relay the friendship request to the target peer	117
Figure 6-6 - Screenshot of friendship link establishment	118
Figure 6-7 - The number of total and successful friendship link establishments.	119
Figure 6-8 - Histogram of the time taken for each successful friendship requests.	120

List of Tables

Table 2-1 - Graph Comparison	15
Table 3-1 – Caltech Affinity Measures.....	35
Table 3-2 - Caltech University’s Attribute Spread	36
Table 3-3. Caltech Dormitory Distribution	36
Table 3-4. Princeton University’s Attribute Spread	39
Table 3-5 Princeton Affinity Measures	39
Table 3-6 - Georgetown University’s Attributes Spread	42
Table 3-7 - Georgetown Dormitory Distribution.....	42
Table 3-8 - Georgetown Affinity Measures	43
Table 3-9 - SNA Measures for all the three datasets	46
Table 4-1 - State Variables and Scales	51
Table 4-2 - Agent Level Variables	51
Table 4-3 - Algorithm to calculate "Personal Preference"	52
Table 4-4. Affinity values, for Caltech dataset, of the four attributes for all the strategies of interactions	57
Table 4-5 - Values of the four attributes for all the strategies of interactions (for Caltech University’s dataset)	57
Table 4-6 Calibration Inputs	58
Table 4-7 Validation Variables	58
Table 4-8 Modularity of Preferential and FOAF strategies with varying Dorm Preference (DP)	60
Table 4-9 Fitted centrality degree distribution with varying Dorm Preference (DP)	61
Table 4-10 - Reference Dataset (of Caltech) and Simulation Output Comparison.....	62
Table 4-11 - Reference Dataset (of Princeton) and Simulation Output Comparison	65
Table 4-12 - Reference Dataset (of Georgetown) and Simulation Output Comparison.....	69
Table 5-1 - Algorithm to crawl Facebook.....	78
Table 5-2 MMU Dataset Description	79
Table 5-3 Edgelist of Small Graph.....	81
Table 5-4 Edgelist of Small Random Graph.....	82
Table 5-5 Random Dataset Description	83
Table 5-6 Onomap Ethnic/Subethnic groups	86
Table 5-7 Onomap Classification.....	87
Table 5-8 - Ethnic Classification of MMU Students and the Reference datasets	87
Table 5-9 - MMU Ethnicity and Onomap Ethnic Classification	88
Table 5-10 - Onomap Case Classification.....	90
Table 5-11 Affinity Measures of Node Attributes.....	90
Table 5-12 - Normalised Silo Index for Ethnicity.....	105

List of Relevant Publications

Book Chapters:

- **SMA Abbas**, SJ Alam, and Bruce Edmonds (2014) Towards Validating Social Network Simulations, Advances in Intelligent Systems and Computing Volume 229, pp 1-12.
- SJ Alam, **SMA Abbas** and Bruce Edmonds (2014) Validating Simulated Networks: Some Lessons Learned, Multi-Agent-Based Simulation XIV, Lecture Notes in Computer Science 2014, pp 71-82.

Journals:

- **SMA Abbas** (2013) An Agent-Based Model of the Development of Friendship links within Facebook, Journal of Computational and Mathematical Organization Theory, June 2013, Volume 19, Issue 2, pp 232-252.

Conferences and Workshops:

- **SMA Abbas** (2013) [homophily](#), Popularity and Randomness: Modelling Growth of Online Social Network, Twelfth International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2013), Saint Paul, USA, 6-7th May, 2013.
- **SMA Abbas**, SJ Alam and Bruce Edmonds (2013) Validating Social Network Simulations, The 14th International Workshop on Multi-Agent-Based Simulation (MABS), Saint Paul, USA, 6-7th May, 2013.
- **SMA Abbas** (2011) An Agent-Based Model of the Development of Friendship links within Facebook, 7th European Social Simulation Association Conference (ESSA), Montpellier, France, 19-23, September, 2011.
- **SMA Abbas** (2010) A Segregation Model of Facebook, 6th UK Social Networks Conference (UK-SNA), Manchester, 2010.
- **SMA Abbas**, Johan Pouwelse, Dick Epema and Henk Sips (2009) A Gossip-Based Distributed Social Networking System, Proceedings Wetice 2009, pp. 93-98, IEEE CS Press.

Technical Reports:

- **SMA Abbas** (2011) Ethnic Diversity in Facebook, Centre for Policy Modelling (CPM) technical report CPM-11-213.
- **SMA Abbas** (2011) An Agent-Based Model of the Development of Friendship links within Facebook, Centre for Policy Modelling (CPM) technical report CPM-11-212.
- **SMA Abbas**, David Hales, Johan Pouwelse, and Dick Epema (2009) A gossip-based distributed social networking system, PDS technical report PDS-2009-001
- **SMA Abbas** (2009) Social Networking in the Virtual World. TU-Delft, PDS Technical report.

Posters:

- **SMA Abbas** (2012) Popularity and similarity among friends: An agent-based model for friendship development, European Conference on Complex Systems (ECCS), Brussels, Belgium, 3-7th, September, 2012.

1 Chapter: Introduction

One of the most powerful ways in which to understand humans is to view them as entities embedded in complex structures of social relationships and interactions. Since the time of Plato, philosophers have grappled with the problem of order in a society: how autonomous individuals can combine to create enduring and functioning societies (Borgatti et al. 2009). The studying of such complex networks can provide insight and explanation of various social phenomena – for instance creativity, information flow, power structure and even corporate profitability (Borgatti et al. 2009).

Social Networks are often represented using two basic building blocks: *nodes* and *links*. Nodes tend to represent people while links represent the connections between pairs of people. The concept of embeddedness (Granovetter, 1985), which says that humans are part of a larger social network, which is generally not visible to individuals (Burk, Steglich, and Snijders 2007), helps us understand the dynamics involved in the maintenance and development of social networks. Embeddedness states that individuals are part of a complex web of social relations and interactions (Borgatti et al. 2009). If we only look at the relationships between a few nodes and their effects on each other, we are missing out the complex impact upon the rest of the network. Each one of us is embedded in such networks. The importance of individual ties is far greater when a broader view is applied to it. Social Networks are so elaborate, complex and ubiquitous that one has to wonder what purpose they serve (Christakis and James H. Fowler. 2009). Studying them enables us to identify the following questions. How do they form and evolve over time? How is each individual affected by them? What network positions are important the ones?

In order to answer such difficult and yet interesting questions, especially when it comes to a macro or meso-level of understanding of society, sociologists have developed Social Network Theory. Social Network theory (which deals with such complex structures), focuses on the relationships among people, but pays less attention to their attributes. In this thesis, however, we focus not only on relationships, but on individuals and their behaviours. In terms of individual actions and relationships and how it is observed, Radcliffe-Brown (Radcliffe-Brown 1940) summarized the complexity of Social Networks by:

We can observe the acts of behaviour of these individuals, including, of course, their acts of speech, and the material products of past actions. We do not observe a 'culture', since that word denotes, not any concrete reality, but an abstraction, and as it is commonly used a vague abstraction. But direct observation does reveal to us that these human beings are connected by a complex network of social relations.

Humans are social animals, meaning we achieve our goals collectively. It has been argued that we have an innate tendency to develop and manage large and complex social groups. Psychologists and sociologists have shown that humans have an affiliation motive (Murray 1938), which means a sense of belonging. Also, we need to share and gather information about the people and the environment around us (Festinger 1950). This motivates us to form groups which can be referred to as *social groups*.

In terms of representing self and one's social networks on the Internet, we often use a Social Network System ([SNS](#)), which Boyd (Boyd, 2010) defined as *web-based services that allow individuals to:*

- (1) *Construct a public or semi-public profile within a bounded system;*

(2) *Articulate a list of other users with whom they share a connection;*

(3) *View and traverse their list of connections and those made by others within the system”.*

In the literature, however, there is another term used for such systems which is Social Networking Systems. Boyd deliberately did not use the word 'networking' in [SNSs](#), as she claims it emphasises relationship initiation, which is mostly between strangers. However, an [SNS](#) allows an individual to meet strangers too, but primarily it is used to articulate and make their current social network visible (Boyd & Ellison, 2007). We are using the [SNS](#) term defined by Boyd. In our work, we focus on social networks in an [SNS](#) which studies have shown, are found to be proxies of real life social network (Boyd & Ellison, 2007; Ellison, Steinfield, & Lampe, 2007).

In the last few years [SNSs](#) have thrived and changed the overall outlook of the Internet. These systems play an important role in making the Internet social – Web 2.0, which means a more social, collaborative, interactive and responsive web (Nations, n.d.). Various [SNSs](#) have sprung up to serve a diverse set of needs. [SNSs](#) provide not only a space for self-representation but also the mechanisms to build and maintain one's social network online. The ubiquity and popularity of such systems provide an interesting and unique opportunity to study not only humans at such a great scale, but the structure itself gives us insight into developing better and user-centred systems for future [SNS](#). A lot of studies have been carried out on such systems to identify how people develop cultures of communication, sharing and participation, and also to identify the network structure of such systems.

Before the creation/invention of [SNSs](#), the Internet was mostly just about content (A Mislove, 2009). Previously, the users of such content and even the creators, generally speaking, were not the focus. When [SNSs](#) came into being, ordinary Internet users became the central entity rather than a set of information publishers (A Mislove, 2009). Users join such systems to publish content and maintain and develop their dyadic relationships – which is generally represented by the 'friend' relationship. Such systems, being focussed on their users, provide a means to interact with other users. On top of that, they allow users to search and even befriend others who share similar interests, but along with these benefits [SNSs](#) provide some significant challenges, which we now review.

In terms of privacy and openness [SNSs](#) pose quite a few challenges. The big danger is that the publicly available information in [SNSs](#) will be misused. In certain cases the information provided may be very extensive and intimate. This poses risks ranging from identity theft to online and physical stalking; from embarrassment to price discrimination and blackmailing (Gross, Acquisti, & Heinz, 2005). At one end, [SNSs](#) encourage their users to share more personal information so their advertisement business model can thrive. On the other hand, users of such [SNSs](#) would like to have more control over their information, in particular, to protect their privacy. For instance, some users would like their personal information, such as political preference and sexual orientation, to remain private (Hanselmann & Hamprecht, 2012). Studies (Jernigan & Mistree, 2009; Alan Mislove, Viswanath, Gummadi, & Druschel, 2010) show that such details can nonetheless be inferred with high probability if a sufficient number of friends in a social network chooses to reveal their details (Hanselmann & Hamprecht, 2012).

Just like the real world but in an online setting, when using an [SNS](#) like Facebook if people tend to deviate from established social norms, they are either ridiculed or punished (Zhao & Grasmuck, 2008). This results in people wearing a mask which is compatible with the established norms, and pretend that is who they actually are. This becomes quite a

major issue for minority groups based on gender, religion, language or any other factor. Instead of having a self-defined identity, the conformity towards well established norms restricts not only individual representations but also freedom of expression. In particular, sexual diversity is often stigmatised and often, users' sexual identity disclosure decisions are shaped by both the social conditions of their online networks and the technological architecture of [SNSs](#) (Duguay, 2014).

An example of a privacy issue is a general policy adopted by an [SNS](#), or for any system which works globally, may have an unintended adverse effect on specific sub-groups. For example, in order to tackle fake profiles, Facebook developed a real name policy, which was implicitly in place since the beginning (Phillip, n.d.). With this policy if someone reports another user's profile as 'fake', then they need to provide some sort of identification which authenticates their name – for instance a state I.D., a library card or a piece of mail (Phillip, n.d.). This policy however, creates problems for members of the Lesbian, Gay, Bisexual, and Transgender ([LGBT](#)) community, who would prefer to be anonymous, and also Native Americans, whose names are difficult to authenticate by Facebook (Phillip, n.d.). There are subtle differences between male and female users as well. Tang et al. (Tang & Ross, 2011) looked at profiles of over 1.6 million users and found that females and males exhibit contrasting behaviours when it comes to revealing personal attributes such as gender, age and sexual preference. Also they found that females are more conscious than males about their online privacy. In terms of individual self-portrayal and behaviours, there is a wide variety of factors which diversifies each user. Nadkarni et al. (Nadkarni & Hofmann, 2012) conducted a systematic review on the psychological factors contributing to Facebook use. They break down identity creation into demographics (gender, ethnicity etc.), personality characteristics (extraversion, introversion, and neuroticism) and cultures (collectivism and individualism).

A study conducted on the [LGBT](#) community (Duguay, 2014) summarises quite well what users expect from a user-friendly [SNS](#): the protection of private information. The users consider this to be of utmost importance to their privacy and as such, more control needs to be provided to them. Also since privacy settings in an [SNS](#) keep changing, such as with Facebook, it becomes quite hard to keep track of them (Paul & Puscher, 2011). A clearer and simpler policy which empowers users, needs to be provided. Paul et al. (Paul & Puscher, 2011) have proposed colour-based privacy settings for Facebook to deal with such complex issues. Ideally, [SNSs](#) would cater for not only minority groups, but more importantly, an individual's preference for privacy and control. Our third contribution is towards tackling such issues. Depending upon how comfortable each user is, a targeted set of services will empower users and build confidence in such systems. We would like to make users the central entity of our [SNS](#), where the data is owned by users not our [SNS](#) and they have full control over it. If users are informed that their information will be solely used for personalised services, they might share their information.

In summary, SNSs provide significant benefits to represent and maintain social networks on the Internet, however, they also raise new issues of identity, privacy and control.

1.1 Research Questions

This thesis is geared towards better understandings of current Social Network Systems ([SNSs](#)) from the ground up, by focussing on individuals and their position in a social network. In order to identify how an [SNS](#) can fulfil a variety of needs such as growth, trust, openness and self-portrayal, we need to study how current successful systems behave. Our major contribution is to identify some key user characteristics and social processes,

which might help future platform and application developers in creating better, more efficient, more open and more user-friendly [SNSs](#).

Set out below are our research questions:

1. *Does an [SNS](#) represent a clustered and segregated network; and on what factor(s) does it cluster and segregate?*

In order to understand how an [SNS](#) is structured, we surveyed what research has already been carried out in this regard. We also identified if such studies can be generalised or whether they are focussed on one particular [SNS](#). For a better insight, we collected several datasets. We analysed the structural properties of these datasets and identified how many, if any, disconnected subgroups (sub-networks) they contain. We then identified whether there was a strong correlation between attributes, which results in very strong communities. Chapter 3 focusses on the three datasets which we have analysed. For a large social network, an in-depth diversity analysis, based on ethnic lines, is carried out in Chapter 5. We looked for evidence as to whether such networks were segregated, clustered or uniformly mixed. This analysis helped us identify the structure of the social network in an [SNS](#), providing us with insights on how to capture the interest of users and classify them, so that a set of targeted services can be offered to them. Based on personal traits and behaviours of users, a tailored set of friendship links can be suggested to them.

2. *How do individuals decide to link with others, for example to what degree is this local to their links or due to global influences?*

We wish to understand how individuals decide whom to befriend. In particular, we would like to understand what social and educational social processes drive the development of students' social networks – which are then translated into online representations.

To answer this we developed simulation models to allow a better insight of individual level behaviours; in particular, to evaluate to what extent each of the four specific mechanisms might be responsible for characteristics of some observed networks. These mechanisms take individual level attributes and behaviours and also global level constraints into account. These included how individuals interact and behave and on what factors they are more likely to concentrate when it comes to friendship development. These behaviours and structural constraints then result in the emergence of a social network. To test our several hypotheses, we then identified which is the best hypothesis for a particular dataset and checked whether it can be generalised to others. To identify if local preferences and/or personal and environmental attributes are responsible for social network development, we made use of a set of agent-based models to enhance our understanding. Also one of the most difficult aspects of an [SNS](#) is its growth. This can be addressed here where we identified what local and environmental aspects might play their role to develop a social network. This question is addressed in Chapter 4.

3. *How a completely distributed [SNS](#) could be developed and how it might perform?*

[SNSs](#) allow users to create identities and link them to friends who have also created identities, are highly popular. Such systems such as Facebook utilise a traditional client-server approach to achieve this, which means that all identities and their social links (the entire social network) are stored and administered on central servers. This high dependence on centralised systems results in the possible exploitation of private data. This also poses unintended and sometimes discriminatory issues for minority groups, based on, such as, sexuality and race/ethnicity. The public display of private information in a typical SNS may result in online stalking, employment discrimination and sometimes

even blackmailing (Gross et al., 2005). In terms of privacy settings, there is a diverse set of preferences across different cultures, gender groups, personality characteristics and ethnicities (Nadkarni & Hofmann, 2012; Phillip, n.d.; Tang & Ross, 2011).

Instead of relying on centralised systems to develop and maintain one's social network, how do we design a completely distributed [SNS](#)? We have made use of a completely decentralized architecture called peer-to-peer network, to establish a self-administered [SNS](#). This is proof of a concept solution which shows how individuals can develop and maintain their social network on their own, without sharing their personal or social information to any third party. The users control their personal information and set their own privacy settings. This work is built over our previous work (covered in the previous two research questions): analysing, categorizing and identifying users of SNSs who have a diverse set of behaviours. Once we empower users to control their own individual privacy settings, and gain insights on how to capture and classify them, we can offer a set of targeted services to them. This focusses on the practical side of things, i.e. what steps does one have to take to establish a link with his/her friend. In order to determine how it actually works in real life, we have designed and then deployed our solution. At the conclusion we share our results and also talk about general ideas that we can take away from our work. We answer this question in Chapter 6.

1.2 Thesis Outline

Chapter 2. Systematic Literature Review

This thesis covers several different [SNS](#) issues. This includes how local interactions and environment factors in a typical university settings would help students to develop their social network over time. This is later used as a basis for our agent-based models. Firstly we review work that has already been carried out in this vein and identify the other methodologies that can be used to study such dynamics. Secondly, we deal with [SNSs](#) in general and review the studies that have already been carried out. There are two strands of such studies: one looking at the macro level structure of [SNSs](#), and the other looking at what local constraints and individual affinities are responsible for friendship development. Finally, we look at how such a dynamic [SNS](#) can be developed where users are solely responsible for storage, data management and privacy by using distributed architecture. We compare our distributed [SNS](#) with other systems.

Chapter 3. Analysis of Complex Network Datasets

In this chapter, we deal with the standard set of measures used taken from Sociology, Computer Science and Physics, to analyse static and evolutionary networks and determine how we can best use them according to the underlying research problems. We also mention how and where we have used such measures. Thereafter, we describe the different datasets we have used in our work: how we collected them, what are the general outlook and measures of them and how are they being used in our work. After describing all our datasets in the previous section, we analyse them.

This chapter relates to the *Research Question 1*, where we describe structural and semantic measures to identify the structure of social networks in our studied datasets. In specific, it looks at the macro level affinities and communities presented in the studied datasets.

Chapter 4. Models of Student Interaction

We describe four agent-based models which we developed. These models look at the social dynamics involved in a typical university setting and then compare them against the empirical evidence (the datasets). We then describe which models best describe such datasets. We also highlight the shortcomings of our models, and compare our methodology ([ABM](#)) against other methodologies. The series of measures described in the previous chapter (Chapter 3) are used to compare model outcomes to the data. This work pertains to the *Research Question 2*, where we identify the roles of attributes and environment in the evolution of social networks.

Chapter 5. Diversity and Clustering of [MMU](#) Students on Facebook

In this chapter, we describe how we have collected a social network dataset. A large social network dataset of four thousand Manchester Metropolitan University ([MMU](#)) students has been selected out of Facebook. We have then inferred each individual's ethnic classification by using a name-based classifier Onomap. The main aim is to identify the ethnic classification of users. This includes identifying the diversity of the users' base and describing how clustered and segregated the social network is when users' ethnicity is taken into account. We also identified how different ethnicities are connected among and within each other.

This also relates to the *Research Question 1*, where we not only infer individual ethnic, religious, language and geographical based classification, but also affinities across the four inferred groups.

Chapter 6. Distributed [SNSs](#)

This chapter describes how to design and setup a completely distributed [SNS](#). We start off with a discussion concerning how a typical distributed [SNS](#) differs from a traditional centralised [SNS](#), and then describe how we have developed one such system. We highlight the architecture of it and share some of the results to illustrate how it might perform with real users. Finally, the results and some of its implications are discussed. The *Research Question 3* is answered in this chapter.

Chapter 7. Conclusion

The conclusion summarises our findings, major results, contributions to knowledge and what future developments could be made. We also describe what is not covered in our work and how such issues that were addressed could have been dealt with more comprehensively.

2 Chapter: Literature Review

2.1 Introduction

In this chapter we will review some of the literature relevant to this thesis. There are four main subsections: agent-based modelling, social networks, Social Network Systems ([SNSs](#)) and distributed social networks. Initially we will explain why we used Agent-Based Social Simulation (ABSS) to understand micro and macro level processes, and how this methodology differentiates from others (mainstream in social sciences) on two bases: empirical data and realism. In our work, we did not have access to longitudinal data of social network evolutions but we relied on limited empirical cross-sectional data. Agent-based models help us understand how emergent local processes translate into macro processes. In the next section, we cover what evolutionary algorithms and mechanisms of social networks offer us and which could have been used, such as mathematics, social science, computer science and physics, and why we did not use them.

There is a lot of analysis carried out on the structure of a social network in a typical [SNS](#), which helps us develop our models and which we extensively used when designing our models. However, there is a dearth of models which mimics how such social networks might have been developed. Our work tries to contribute solutions to this problem area.

Most of our work revolves around one particular [SNS](#), Facebook. We cover Facebook's history in some detail including its evolution in terms of its feature set and we also review the major studies carried out on it which helped, influenced and informed our overall work. We will particularly focus on how these relate to social networks and their evolution.

Our fourth contribution is towards distributed [SNSs](#). In particular we set out how a completely decentralised [SNS](#) is designed and what recent developments have been made in this vein and compare our solution with others.

2.2 Agent-Based Modelling

An agent-based model ([ABM](#)) is a computer program that creates a world of heterogeneous agents in which each agent interacts with other agents and with the environment (Hamill 2010). Agents are either separate computer programs or, more commonly, distinct parts of a program that are used to represent social actors—individual people, organisations such as firms, or bodies such as nation-states (N. Gilbert, 2008). In [ABM](#), it is assumed that the agent is an individual who has intentions or goals and makes choices that affect other agents whose choices in turn affect that individual, and this is reflected in the workings of most [ABMs](#). It is considered as 'generative social science', which assumes micro level specifications (of agents, environments and rules) and generates macro structures of interest (Epstein, 1999). An agent makes a decision on the basis of individual assessment of their situation (Bonabeau, 2002). The basis of agent-based simulation is in Artificial Intelligence (AI) and non-linear dynamics (N. Gilbert & Troitzsch, 2005).

[ABMs](#) tend to have the following four key assumptions (Macy & Willer, 2002b):

1. Agents interact with little or no central authority or direction. Global patterns emerge from the bottom up, determined not by a centralised authority but by local interactions among autonomous decision makers.
2. Decision makers are adaptive rather than optimising, with decisions based on heuristics, not on calculations of the most efficient action (Fukuyama, 1996). These

heuristics include norms, habits, protocols, rituals, conventions, customs, and routines.

3. Decision makers are strategically interdependent. Strategic interdependence means that the consequences of each agent's decisions depend in part on the choices of others. When strategically interdependent agents are also adaptive, the focal agent's decisions influence the behaviour of other agents who in turn influence the focal agent, generating a 'complex adaptive system' (Macy n.d.)
4. Agent-based models take into account not only an individual's preferences and his/her interactions but also geographical space into account. Additionally agents are adaptive and backward-looking (Macy & Willer, 2002a).

Agents have a complex set of modalities, which include behaviours based upon their memory and unique historical path (path dependence). They exhibit time based dependents (hysteresis) and follow non-markovian behaviours (Bonabeau, 2002). [ABM](#) is not a panacea for all problems, but it can be applied to sets of issues which have the following characteristics (Bonabeau, 2002):

- Interactions: Agents interact with each other in a non-linear fashion, resulting in a complex outlook.
- Position: The positions of agents are not fixed but tend to change over time. This could mean both in geographical terms and/or in social networks.
- Heterogeneity: Agents are not all the same and potentially all agents are different.
- Learning and Adaptation: Interactions within agents and also the environment, often involve learning and adapting to new situations/environments.

The major challenge in building [ABMs](#) is deciding how to specify how agents behave and what rules are required to enable agents to make decisions. Most of the times these rules are developed based on theories, hunches and common sense, mainly due to lack of data. If a one-to-one mapping of reality is replicated in an [ABM](#) (and supported data is also available), its complexity becomes hard to handle. For instance, it becomes really hard to identify what causes what. In order to deal with such issues, the following guidelines should be kept in mind when designing an [ABM](#): proceed systematically avoiding arbitrary assumptions, carefully grounding and testing each piece of the model against reality and introducing additional complexity only when it is needed (Farmer & Foley, 2009). In [ABM](#), agents have limited information (they do not have perfect knowledge), and they make their decisions based on their perceptions (Janssen, 2005). These perceptions do not have to include correct representations of reality and may vary among agents (Janssen, 2005) – this is what is called bounded rationality (Simon, 2000) in the literature. To provide more detail about it, agents, unlike rational beings, have the following characteristics (Edmonds, 1999):

- they do not have perfect information about their environment, in general they will only acquire information through interaction with the dynamically changing environment;
- they do not have a perfect model of their environment;
- they have limited computational power, so they cannot work out all the logical consequences of their knowledge;
- they have resource limitations (e.g. memory).

[ABMs](#) are used to show how simple local interactions of agents can generate familiar but emergent global patterns, such as the diffusion of information and emergence of norms and the development of social order/structure (Macy & Willer, 2002a).

In our work we have used agent-based models for the exploration of social processes and also environmental constraints imposed in university settings. Over the last five years or so, a lot of work has been carried out on [SNSs](#) to identify what the structure of social networks is and how people are connected to each other; both structurally, which takes into account the position of people (nodes) in a network, and semantic analysis, which takes the inter-attribute [homophily](#) into account. This work has been carried out on all levels from a small static college network to a large global social network. However, [ABM](#) has hardly been used to explore how such a network might have evolved, which is one of the distinctive properties of [ABM](#) as it helps in exploratory analysis.

We have used an agent-based simulation approach to design an interactive and realistic mechanism. The strength of this approach is the ability to design its fundamental entities and agents accurately to describe the behaviour and interactions among real actors and its ability to capture episodic and unpredictable volatility of social processes evidenced by data at fine grains of time (Moss, 2008).

2.2.1 Alternative Approaches

In mainstream social science, however, there are two major approaches other than agent-based simulation. The first is microsimulation which is a bottom up strategy for modelling the interacting behaviour of decision makers within a larger system. It uses data on representative samples of decision makers, along with equations and algorithms representing behavioural processes, to simulate the evolution through time of each decision maker and hence of the entire population of decision makers (S.B. Caldwell, 1997). However, the models do not permit individuals to directly interact with each other or to adapt (Macy & Willer, 2002a). Also typically, in microsimulation, there is no notion of social or physical space (like geography)(N. Gilbert, 2008).

The second alternative approach is the Stochastic Actor Oriented Model (SAOM) (Snijders, 1996) which relies on longitudinal or panel data and is used to study both selection and influence in social networks, thereby giving insights into whether behavioural processes emerge from or contribute to network formation (Opsahl, 2010). Also, SAOM uses strong assumptions as it is a 'surprise free' modelling approach and is good for null models. Both of these approaches rely heavily on the data. The first one does not offer an interactive or adaptive environment and the second one relies heavily on panel data. With a single snapshot of the data, neither of these are applicable to our research problem.

In computer science, there are many 'mechanistic and yet tractable' (Kim & Leskovec, 2010) network models, such as Preferential Attachment (A. L. Barabási & Albert, 1999) which specifies a link creation mechanism, resulting in a network with power-law degree distribution. These models, however, do not take node attributes into account. In machine learning and social network analysis, the emphasis has been more focused on the development of statistically sound models that consider the structure of the network as well as the features of nodes and links in the network (Kim & Leskovec, 2010). Examples of such models include Exponential Random Graphs (Pattison, 1996) and Stochastic Block Model (Airoldi, Blei, Fienberg, & Xing, 2008). These network models are generally intractable and do not offer emergence (Kim & Leskovec, 2010).

There is some work on models on dynamic social network as well. For instance Gulyás et al. (Gulyás, Kampis, & Legendi, 2013) applied simple rules at random networks in order

to produce networks meeting with properties, such as density and clustering. However such models are unable to measure the contribution of a group of actors to the network evolution (Uddin, Khan, & Piraveenan, 2015). For a longitudinal social network, Carley's work (Carley, 2003) combines SNA and ABM which incorporates probabilities and uncertainties into the structure information (Berger-Wolf & Saia, 2006). Similarly, to review longitudinal or temporal networks, Holme et al. (Holme & Saramäki, 2012) reviewed methods to analyse topological and temporal structure and models for elucidating their relation to the behaviours.

2.3 Social Networks

Before looking at the network generation process, we need to briefly review social networks. By looking at the structure of a social network, we learn a few things. There are a couple of unifying structural properties of social networks: [homophily](#) (love of the same), clustering (friend of a friend), the small-world effect, the heterogeneous distributions of friends, and community structure (Mark Newman, 2010a; Ugander, Karrer, Backstrom, Marlow, & Alto, 2011). Hamill et al. in their study (L Hamill & Gilbert, 2009), have summarised, more or less, the same characteristics. The one thing which is missing in the previous list is assortativity by degree of connectivity, which means the similar the number of links, the higher the chances of getting connected. It is a sort of degree [homophily](#).

In terms how [homophily](#) as an empirical fact (as measured on networks of nodes with characteristics) can clearly result from a number of processes/biases: (1) sheer prejudice - people don't make friends with the "wrong" kind of person or otherwise actively (2) implicit biases – it is easier to make friends and talk to someone with whom one shares culture, experiences, habits, beliefs etc. (3) meeting clustering - due to cultural differences, one tends to only meet certain kinds of people at certain events (e.g. at a place of worship obviously, but also you will find far fewer Muslims in pubs or other events with alcohol).

When we compare social networks to other networks, for example technological networks such as the food web and the world-wide-web, as explained above, we find a positive degree of correlation, which suggests that social networks may be robust to intervention and attack while technological networks are not (MEJ Newman, 2002). A good technique to grow a graph which satisfies all these characteristics of social networks. In terms of characteristics like age and nationality, Facebook found a very strong [homophily](#) between its users (Ugander et al., 2011). For instance, overall 84.2% friendship links are within same countries, illustrating that the community structure of Facebook is mostly based around local geographical area (Ugander et al., 2011).

In the previous section, we have covered some of the models like Exponential Random Graphs and Stochastic Actor Oriented Model (Snijders, 1996), used in social sciences to generate networks by evolution (and inter-node interactions too), but in this section we are going to focus on purely mechanistic approaches which have helped us develop realistic agent-based models of social networks.

2.3.1 Regular Lattice

One of the simplest networks is a regular lattice. It is developed when a ring of ' n ' nodes is formed and where each node is connected to its ' k ' nearest neighbours by undirected links (Watts & Strogatz, 1998). An example of such a network with 20 nodes, and where each node is connected to its 2 nearest neighbours is shown in Figure 2-1. Its whole network density is low. The size of personal networks is limited and as many of one node's neighbours will also be neighbours of each other, there will be high clustering. But it fails

to meet the other criteria such as heterogeneous distributions of friends (as all nodes have equal links), and community structure (L Hamill & Gilbert, 2009).

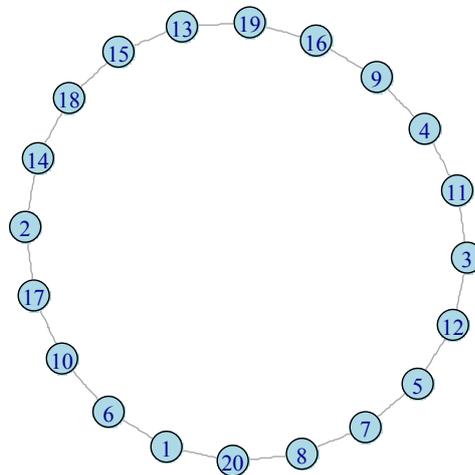


Figure 2-1 - Regular Lattice

2.3.2 Random Networks

Random networks, as the name says, define a set of networks where links are added to nodes in random fashion. One of the most seminal works on random networks has been carried out by Erdős and Rényi (Erdős & Rényi, 1959). Random network models assume that the probability that two nodes are connected is random and uniform (A. L. Barabási & Albert, 1999). Figure 2-2 shows a sample random network. Generally speaking these networks are created where random links are added on a static number of nodes. These networks have very short average path lengths between any two nodes. There are many models available like the ER model, to generate a random network with a specified number of nodes and the probability of connecting two nodes with each other. In a random

network the nodes follow a Poisson distribution with a bell shaped degree distribution, and it is extremely rare to find nodes that have significantly more or fewer links than the average (A.-L. Barabási & Bonabeau, 2003). One study (MEJ Newman, 2002) found out that the correlation (or assortativity) of nodes in such networks can be analytically shown to be zero, which is certainly not the case with social networks.

2.3.3 Power-Law Networks

Power-law networks are networks where the probability that a node has degree 'k' is proportional to $1/k^n$, for large 'k' and $n > 1$. Thus, the degree distribution of a power-law network follows an exponential decay. The parameter is called the power-law coefficient. Unlike standard random networks where all the nodes have more or less the same connectivity (degree), in power-law networks there exist a few nodes with very high connectivity (high degree) and the rest has a very low connectivity (low degree) (Erdős & Rényi, 1960). In power-law network, the probability that any node is connected to 'k' other nodes is proportional to $1/k^n$ (A.-L. Barabási & Bonabeau, 2003). Figure 2-3 shows how a preferential networks' degree distribution looks in a log-log scale. When developing a new friendship or any other social relationship, people generally do not know how many links others have and whether the target person would reciprocate their relationship (L

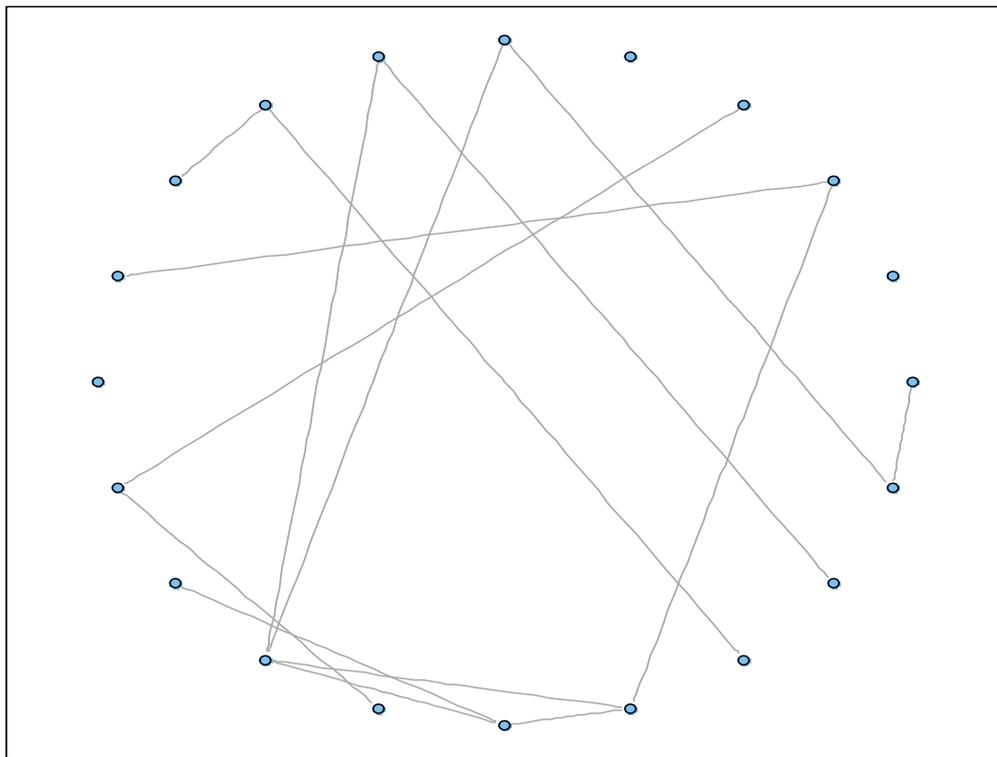


Figure 2-2 - Random Network

Hamill & Gilbert, 2009). This means it is hard to expect a realistic model based on power-law networks, where agents have limited information, to form links.

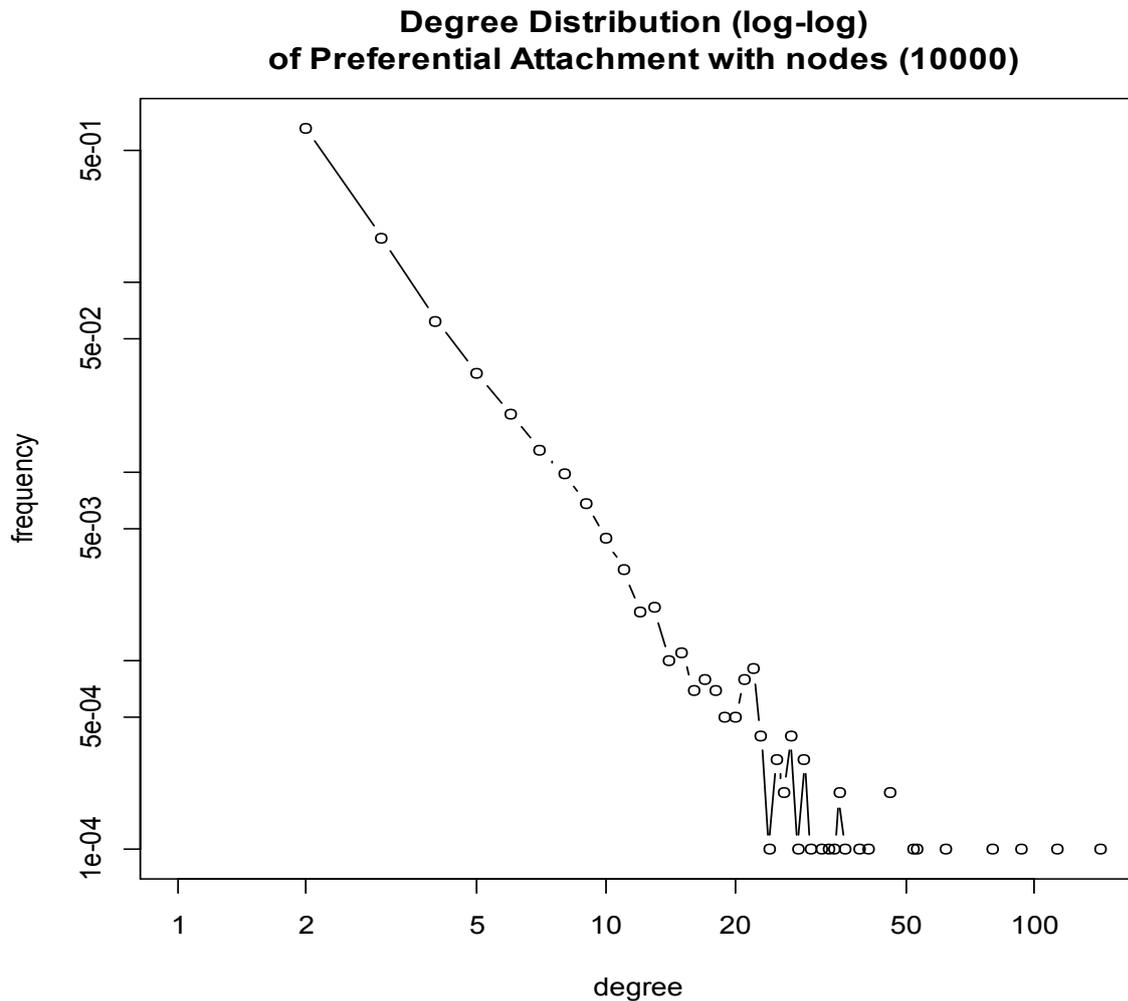


Figure 2-3 - The degree distribution of a network formed via preferential attachment

2.3.4 Small-World Networks

Small-world networks have a small diameter, on average the shortest paths and exhibit a high clustering coefficient (Adamic & Adar, 2003; A Mislove, 2009). In 1967 Milgram and his students did an experiment to identify how many acquaintances are required to develop a connection between any two people in a population (Milgram, 1967). They selected a few initial people to send a letter to a pre-identified broker in Boston, where they had very little information about him. The idea was that the starters would send a letter to a target person on a first name basis. In case they did not know him, they would send a letter to an acquaintance that they believed to be more likely to know the target (Schnettler, 2009).

The result from Milgram's experiment has fascinated people, indicating that any two people in the world, who apparently have nothing in common, at least theoretically, are connected through a small number of acquaintances. In 1998, Watts and Strogatz (Watts & Strogatz, 1998) developed a model incorporating high clustering, small average geodesic distances and the tendency of people to interact in groups (Adamic & Adar, 2003). They illustrated that by randomising a small number of links in a regular lattice, it

can be transformed into a small-world network. Also, they showed that social network of film actors – who acted with whom in a film, is actually a small-world network. After their work, Adamic (Adamic, 1999) showed that the World Wide Web (WWW) is also a small-world network. Similar studies on homepages of two universities, MIT and Stanford and their linkages among them, proved their social network to be of a small-world network (Adamic & Adar, 2003). When such networks are compared with social networks, we find that the small-world network does not produce nodes with high degrees of connectivity or display assortativity (L Hamill & Gilbert, 2009).

2.3.5 Preferential Attachment

In a growing network, preferential attachment is a property of link formation where the likelihood of the source node being connected to the target node, depends on the degree of the source node. This phenomenon is also called the rich-gets-richer (A. L. Barabási & Albert, 1999). This process was first determined by Yule in 1925 (Yule, 1925), often credited to his name by calling it the Yule Process, who used it to explain why the number of species per genus of flowering plants have a power-law distribution. Preferential attachment in a given network can be characterised as linear, if the probability of a node receiving a link is in linear proportion to the node's degree, or sub-linear, if the probability of a node receiving a link is, for example, in proportion to the log of the node's degree (A Mislove, 2009). This process has been quite successful to explain the existence of networks with power-law degree distribution (Vazquez, 2002).

2.3.6 Comparison

None of the evolutionary models discussed immediately above capture the properties of a typical social network of a [SNS](#). Hamill et al. (L Hamill & Gilbert, 2009) illustrated how their algorithm called *Social Circle* captures key aspects of large social networks such as low density, high clustering and assortativity of the connectivity degree. This model covers the overall characteristics of a social network, but does not capture how interaction and attributes could play their roles in meeting and then developing relationships. Thus being a general model it does not focus on individual behaviour and attributes which play a vital role in developing links based on [homophily](#). In the table below, inspired from (L Hamill & Gilbert, 2009), we have summarised a set of criteria of a typical social network for each of the studied networks. Regular networks meet only low density heterogeneous personal networks and high clustering. Random networks have just low density and short path lengths. Small world networks have only limited heterogeneity personal networks, but fail to produce fat tailed degree distribution. As for the preferential attachment, it does not have assortativity, clustering and short path lengths. *Social Circle*, as just discussed, meets almost all characteristics. However, it does not have any provisions of inter-agent (or node) interaction.

Table 2-1 - Graph Comparison

Characteristic	Regular	Random	Small-world	Preferential Attachment	Social Circle
Low density	√	√	√	√	√
Variation in size of personal network	√	Limited	Limited	√	√
Fat-tail	×	×	×	√	√
Assortative	×	×	×	×	√
High clustering	√	×	√	×	√
Communities	×	×	×	√	√
Short path lengths	×	√	√	×	√
Interactions	×	×	×	×	×

2.4 Social Network System ([SNS](#))

In this section we look at the history of [SNSs](#) and how they become part of everyday life. Most of our work revolves around one particular [SNS](#), Facebook – which is the most prominent [SNS](#). We cover Facebook’s history in some detail including its evolution in terms of feature set and also cover the major studies carried out on it which helped, influenced and informed our overall work.

2.4.1 History

In the year 1995, it was classmate.com which allowed users to connect with their school mates from the schools they had attended. It did not however, allow users to develop ‘friendships;’ or any other kind of links with each other (A Mislove, 2009). As for the first [SNS](#) which qualifies for the definition described earlier given by Boyd et al. (Boyd & Ellison, 2007), it was SixDegrees.com created in the year 1997. It allowed users to create their profiles and link with other users. Initially it attracted many users, but then it failed in 2000. The reasons being, according to Christakis et al. (Christakis & Fowler, 2009), was that it was ahead of its time, the market was not ready for such a system. In 2002, Friendster.com came into being. It was launched as a competitor to a match-making website, Match.com. Unlike the introduction of strangers for suitable match-making, Friendster envisaged that friends of friends are a better resource pool to find a suitable romantic partner (Christakis & Fowler, 2009) as the inherent trust of Social Network will bolster confidence among users unlike developing new relationships with complete strangers.

In In 2003, a rival of Friendster came onto the scene, MySpace.com. According to its

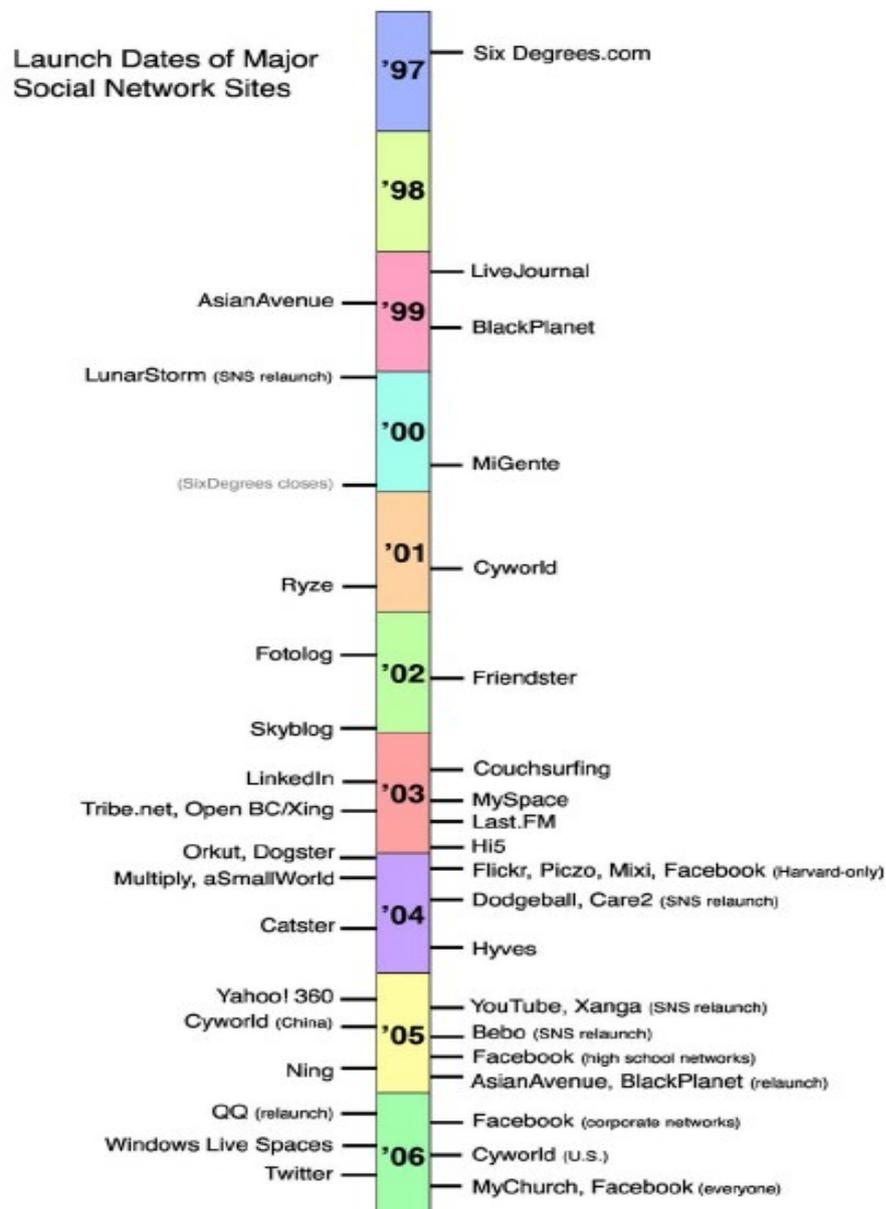


Figure 2-4 - History of Social Network Systems (Boyd and Ellison 2007)

founder, MySpace attracted users by cashing in on rumours saying that Friendster would adopt a fee-based system (Boyd & Ellison, 2007). Out of this confusion and panic, the current Friendster users started posting information about alternate SNSs which were free such as MySpace (Boyd & Ellison, 2007). According to Boyd and Ellison (Boyd & Ellison, 2007), MySpace also exploited the indie-rock bands which failed to comply with Friendster profile regulations and were then banished from using it. These bands were taken with open arms by MySpace. It became a popular tool for musicians to connect with their fans and also for fans to follow their favourite musicians. To cater for user needs, MySpace was unique in the sense that it allowed its users to suggest features which it would incorporate in itself, a user centric approach (Boyd, 2006). Also it provided mechanisms that allowed users to modify their profiles by adding content with HTML

scripts which resulted in a unique experience, whereby users can modify their profiles by simply copying and then pasting HTML code snippets (Perkel, 2006) to alter their default feature set.

Contrary to the common belief that Facebook was designed and then developed in Harvard in the year 2004, there are some very important historical facts on which it was based. According to Christakis et al. (Christakis & Fowler, 2009), the term ‘facebook’ predates the Internet and which comprised of a directory of students of each class with their photos and their accommodation information. Similar directories were common in various Ivy League universities and in high schools too. According to Christakis et al. (Christakis & Fowler, 2009), one of the first mentions of Facebook as a directory was made in 1979 (Faludi, 1979). The idea was to use it to evaluate potential mentors for first year students: “We used the facebook to see what people were like.... Sometimes you can tell from a picture” (Faludi, 1979). In 2004, after twenty five years of its inception, Mark Zuckerberg, a sophomore student at Harvard, took Facebook and made it online as thefacebook.com. We will discuss Facebook in detail in the next section. Since 2006 or so, [SNSs](#) have gone mainstream. We have talked about just the most prominent [SNSs](#) [but](#) for a complete and up to date list of popular [SNSs](#), visit the Wikipedia entry at (“List Of Social Networking Websites,” n.d.).

2.4.2 Research Themes

We have briefly described how Facebook, as it is known today, came into being in the previous section. In this section, we will see how it has evolved over the last ten years (in 2006, Facebook became mainstream) in various aspects. When Facebook started becoming popular, it did not have very distinctive features. The success of it, among many other aspects, is due to its user centric developments over the years.

2.4.2.1 Identity and Representation

Identity on the Internet arises from a social environment, and varies from one context to another. A very detailed analysis using a longitudinal data of students was analysed by Lewis et al. (Lewis, Kaufman, & Gonzalez, 2008). They collected data with the help of Facebook itself and the administration of the studied university. Such a close relationship with Facebook allowed them to have access to students’ profiles in detail and on request they accessed these profiles twice. From health to tastes, Lewis et al. analysed friendship strengths on many levels. They also manually identified the ethnic classification of each participant based on their profile pictures and the ethnic groups they were members of. This amount of unprecedented information of not only personal but social information, allowed them to study students’ behaviour in a very detailed manner.

In terms of race, nationality and also ethnicity, Facebook lacks any explicit fields which allows users to specify them, as it was found by Ginger (Ginger, 2008). This study tackles the racelessness issue of it, a kind of colour-blind perspective of its users, which is contentious. Race, which defines our identity, should be part of an [SNS](#). Ginger proposed that a new field with ‘race/ethnicity/nationality’ be introduced into Facebook, however, it should not be compulsory to complete it. As for self-representation, Zhao et al. (Zhao & Grasmuck, 2008) talked about *hoped* and *actual* self-portrayal in Facebook with the help of a profile in Facebook. Their claim is that online identity in [SNSs](#) is used to portray the hoped representation of self which lies somewhat between the true and ideal self, but is real and is mainly driven by social norms; hence it represents a socially desirable self. It

also increases self-image for the offline world. Identity on the Internet arises from a social environment and varies from one context to another.

According to Vitak (Vitak, 2008) unlike the offline world, where verbal and non-verbal cues are present, the online world provides an anonymous outlook to them which fosters stronger group interactions and improved impressions. They have looked closely at how the role of online identities from Facebook, in specific, relate with their offline selves. These cues are based on social categories of communicators, such as group membership, rather than physical features and appearances. According to another study made by Joinson (Joinson, 2008), which studied Facebook users with the help of an online survey, identified that the main usage of Facebook was to 'keep in touch' with your social circle.

A study carried out on two [SNSs](#), Facebook, and StudiVZ, which is a popular [SNS](#) in Germany, came to the conclusion that people do not present their idealised self; rather they use [SNSs](#) for a better representation for both expressing and communicating their real personality (Back et al., 2010). Applications, which are third party toolkits, such as SuperPokes, Visual Bookshelf and 'Are you interested' represent the multiplicity of self, showing multiple faces and negotiating different facets to a variety of audiences (Papacharissi, 2009).

As for the effects of using Facebook on one's subject of well-being, a study (Kross et al., 2013) found that there was a strong negative impact on young adults both moment-to-moment and how satisfied they were with their lives, over time with Facebook. The more lonely people felt at one time point, the more people used Facebook over time. Another study (Tiggemann & Slater, 2013) looked into the complex relationship of internet exposure (especially of Facebook) and body image concern in adolescent girls. Tiggemann et al., found out that across their whole sample (1,087 girls in Years 8 and 9 with the mean age of 13.7 years and SD of 0.7), there was a high correlation of internalisation of the idealistic thin body, which also derives them towards skinniness, with the use of the Internet (especially of Facebook).

Facebook has certainly become the dominant [SNS](#) in the world and especially in the US with almost two thirds of the US adults using Facebook (Rainie, Brenner, & Purcell, 2012). In an interesting study done by Pew Research to identify how many of the Facebook users had taken a break from it, it turns out there is a substantial number of users (61%) who stopped using Facebook for a period of a week or more (Rainie, Smith, & Duggan, 2013). A number of reasons were shared by the users, but the main reason (by 21%) shared was due to being busy with other activities and a shortage of time. Also, 20% of the adults who do not use Facebook, were once members of it.

2.4.2.2 Privacy

As far as the amount of information posted on Facebook is concerned, there was a study on its members which showed that members were publicising a lot of information about themselves and were generally unaware of the privacy options Facebook affords for them (Acquisti & Gross, 2006). There was a high level of trust in Facebook by its users, when compared with users of MySpace or Friendster (Acquisti & Gross, 2006).

The year 2006 brought about Facebook's first major redesign. A 'News Feed' was added to a user's homepage and a 'Mini Feed' appeared on individual profile pages (Forbes, n.d.). These two features were introduced in September of 2006 which, according to Ryan (Ryan, 2008), altered the Facebook users' perceptions of privacy and feelings of security. The 'News Feed', which is an aggregation tool for Facebook activities, was

displayed on users' homepages; whereas the 'Mini Feed' was made part of users' profiles containing what new changes they and their friends had made to their profiles. Neither of them could be altered by other users which resulted in one of the most unanimous displays of protest ever seen on the Internet (Ryan, 2008). A Facebook Group emerged out of this anger, called "Students Against Facebook News Feed (Official Petition to Facebook)". Within just three days it accumulated over 750,000 members, prompting substantial changes to the News Feed by Facebook immediately (Ryan, 2008). This huge protest was promptly noticed by Facebook itself, after which Mark Zuckerberg himself offered a public apology on this issue by saying:

We made the site so that all of our members are a part of smaller networks like schools, companies or regions, so you can only see the profiles of people who are in your networks and your friends. We did this to make sure you could share information with the people you care about. This is the same reason we have built extensive privacy settings — to give you even more control over who you share your information with.

Somehow we missed this point with News Feed and Mini-Feed and we didn't build in the proper privacy controls right away. This was a big mistake on our part, and I'm sorry for it. But apologizing isn't enough. I wanted to make sure we did something about it and quickly. So we have been coding nonstop for two days to get you better privacy controls. This new privacy page will allow you to choose which types of stories go into your Mini-Feed and your friends' News Feeds, and it also lists the type of actions Facebook will never let any other person know about. If you have more comments, please send them over.

This may sound silly, but I want to thank all of you who have written in and created groups and protested. Even though I wish I hadn't made so many of you angry, I am glad we got to hear you. And I am also glad that News Feed highlighted all these groups so people could find them and share their opinions with each other as well.

(Mark Zuckerberg, 8th September 2006)(Ryan, 2008; Zuckerberg, n.d.)

Ellison et al. (Ellison et al., 2007) have done a study on the Michigan State University students, where it was found that students primarily used Facebook to maintain existing offline relationships or to solidify what would otherwise be ephemeral, temporary acquaintanceships.

Valenzuela et al. (Valenzuela, Park, & Kee, 2009) explained the two main features which Facebook provides to its users. Their main purpose is to keep users updated about their friends and their activities. Each user has a personal homepage and a profile. The two features are: 'News Feed' and 'Mini Feed'. The 'News Feed' feature captures stories about one's friends, for instance, if someone befriended a new person, it will be displayed here. As for the 'Mini Feed', this revolves around the changes and props (applications) one's friends have added to their profile. These two features are really vital in re-enforcing the ties by keeping each other updated.

To improve the privacy of users by giving them more control of their settings, a user-friendly approach has been proposed by Liu et al. (Liu, Gummadi, Krishnamurthy, & Mislove, 2011) which tackles the users' perception and the actual Facebook privacy settings. According to them, only 37% of the time do they find that users' expectations about privacy settings are valid. The default privacy settings which are carefully chosen by Facebook, to control and safeguard one's information, do not represent what users actually want. Along the same line, in order to improve the privacy settings and ideally

meet with the users' expectations of it, Paul et al. (Paul & Puscher, 2011), proposed a very simplified three colour-based setting, which is not only easy to implement but overcomes the problems of complicated privacy settings. They also highlighted that, in spite of users getting more concerned about privacy settings over the years, Facebook privacy settings had become more relaxed. They implemented their system and when comparing with Facebook how long would it take to overcome privacy settings, their system reduced the time by manifolds.

By studying features of Facebook such as 'Likes' (an act of liking a Facebook item), the exposure of personal information to friends and 'friends of friends' (FOFs), the friends' list, network (educational/professional or geographical) and Wall posts, McKeon (McKeon, n.d.) developed a series of info-graphics that covered how, over the years, default privacy settings on Facebook have evolved, or perhaps relaxed. He took data from 2005 to 2010, being the two pictures of extreme years and the rest of them can be seen below (McKeon, n.d.) in Figure 2-5 and Figure 2-6. These charts illustrate that users are demanding more control over their privacy settings but, in terms of default settings however, over time, more information about users has been made public.

Figure 2-5 - Privacy Settings in 2005 (McKeon, n.d.)

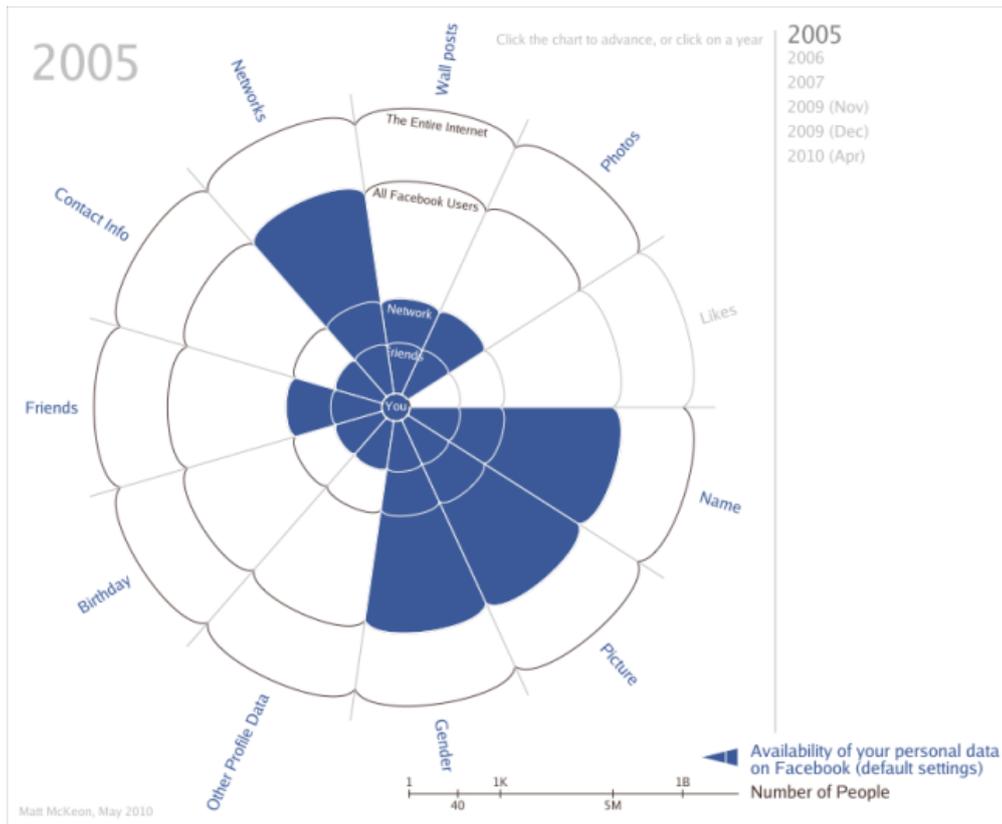
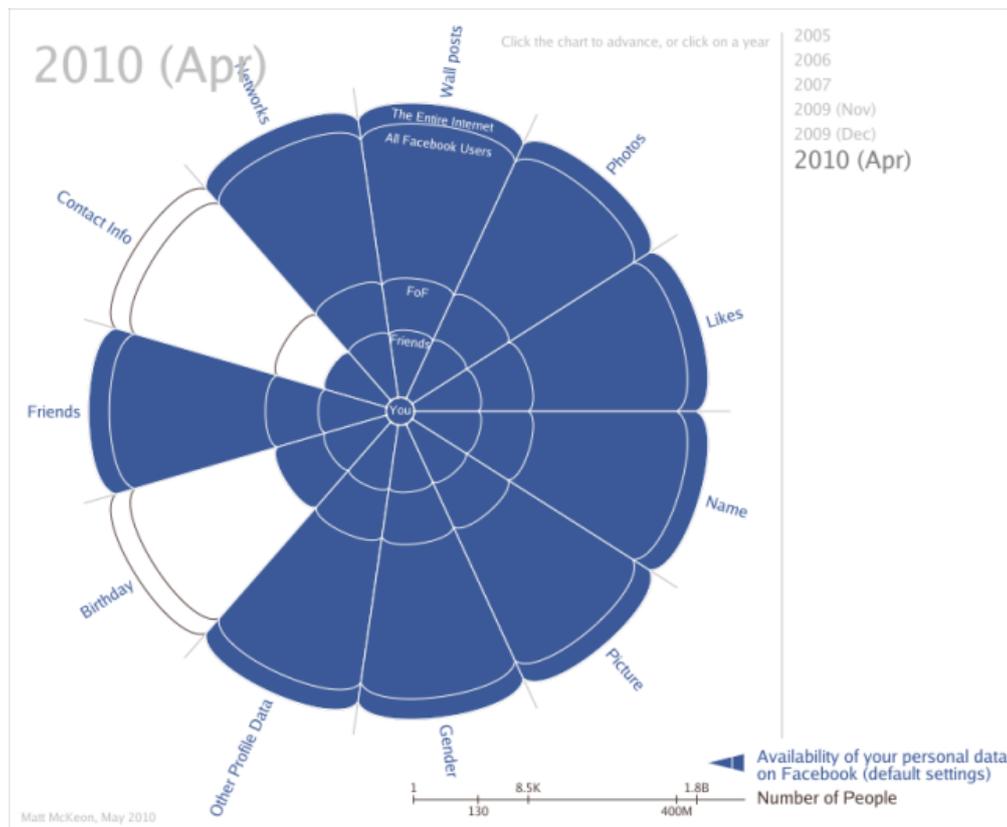


Figure 2-6 - Privacy Settings in 2010 (McKeon, n.d.)



2.4.2.3 Growth

Facebook started off as an [SNS](#) for US educational institutions. Users needed to have an academic email address ('edu') to register. From 26th September 2006 Facebook removed this restriction (Abram, n.d.), which allowed anyone with a registered email address to become its member. With Facebook membership open to the general public, at the end of 2006 the number of Facebook users grew to twelve million (Facebook, n.d.-c; Ryan, 2008). In one study (Lampe & Ellison, 2006) carried out in a university setting on freshmen students, it was found that more than 95% had heard of Facebook and 84% were already members of it before they joined their university. The same study points out that Facebook users do not use it for social browsing, but rather for connecting with already known people from real life instead of searching for new 'friends'. As far as the number of Facebook friends is concerned, according to another study, again involving students (Stutzman, 2006), the number of average friends grew from 46 to 111 by the end of the first semester. This might give us a picture of how many friends a freshman might make in the first semester of college: 65.

Many studies were carried out to understand how much users of Facebook trust other users and also their trust in Facebook as a service itself, e.g. (Dwyer, Hiltz, & Passerini, 2007). Profile fields that help users share common referents are more highly associated with numbers of friends than fields that share personal likes and dislikes (Lampe, Ellison, & Steinfield, 2007). Also Lampe et al. made a remark specific to Facebook in that it is a

different type of [SNS](#) which is closely tied with the offline self – through educational institutions mainly, thus the results should not be generalised for all other [SNSs](#) such as MySpace or Orkut. Similar remarks about not generalising results of one [SNS](#) to all other [SNSs](#) were made by Hogan (Hogan, 2009). Ellison et al. (Ellison et al., 2007) found out that the new students, unlike the juniors and seniors, had a greater tendency to meet new people. In fact, social capital was highly correlated with the intensity of Facebook usage.

In May 2007, Facebook opened its gate for third party applications (Joinson, 2008) which represents additional functionalities primarily in terms of games. The feature to add applications is one of the distinguishing features of Facebook from its competitor [SNS](#). Some [SNSs](#) tried to copy the same model, but the number of applications in Facebook was no match for them (Vitak, 2008).

Gilbert et al. (E. Gilbert & Karahalios, 2009) studied friendship strength in order to classify them into strong and weak ties. Thirty five participants from fifteen different departments participated in their study. In total, they found seventy four variables as predictors of tie strength in Facebook. From theoretical literature they developed a system of seven dimensions quantifying friendship strength. These dimensions included the communication activity through messaging based on duration, frequency and content analysis (positive and negative words usage), the amount of shared contacts and applications installed; the educational compatibility and the links shared. Based on these insights, they developed a linear model to predict tie strength. They found that the level of intimacy (closeness) is more of an important dimension than the level of intensity (frequency of communication) between two Facebook friends.

Delving into the experiences of using Facebook for keeping up with one's social network, Young (Young, 2011) investigated activities of adult users by conducting an online survey on 758 people and doing an in-depth interview with 18 of those people. Specifically Young targeted how adults socially engage on Facebook and what tools they use. Overall everyone found Facebook a convenient, cheap and economical tool to help connect with one's social circle. According to one of their respondents, Facebook offers 'Facestalk' which is the act of reviewing in detail another person's Facebook page to follow their activity without necessarily engaging in any form of communication with the person. This includes learning about others through their Facebook tools such as wall posts, status updates, photos and events. Facestalking is claimed by 67% of the population in their study. Facebook is considered a crucial part of social life with quite close attachments and is used to amplify offline relationships.

Based on a meme which started in Facebook back in July 2009 on 'how you met me' where users on Facebook shared how they have met their friends in real life or even in virtual life, Lada et al. (Adamic, 2012) collected a huge dataset which contained 2,570,182 posts from the meme, with 24,199,921 comments posted during July 2010 and November 2011. In this dataset, they found a very high number of female participants, 79.7%. They not only collected the text containing such a meme, but also personal information of the users who participated in this meme. According to their analysis, the majority of the people were from English speaking countries. Here is the breakdown in geographic terms: United States 70.1%; Great Britain 11.4%; Canada 5.8%; Australia 3.7%; New Zealand 0.7%; South Africa 0.6%; and Ireland 0.5%. The main result of their study was that a significant proportion of Facebook friendships originated at school, even for individuals who had not been in school for decades.

Hampton et al. (Hampton & Goulet, 2012) talked about the activity of Facebook users both on an individual and average level. They found that between 20% and 30% of users who are really active, have a great impact on the rest of the users. Due to these active

users the average Facebook user receives friend requests, receives personal messages, is tagged in photos, and receives feedback in terms of 'likes' at a higher frequency than they contribute. During their month of analysis of Facebook users, they found that around 40% of Facebook users made a friend request in the month whereas 63% of users received a friend request. As for the number of Facebook friends each user has, an average respondent underestimated by eighteen friends. Unlike real social networks which generally include just the active links, Facebook's social network has a lower density (0.36 vs. 0.12). The explanation given for this is that since Facebook affords to connect ties that otherwise might have gone dormant, it therefore has a larger social network as opposed to real life social networks.

With the ubiquitous presence of smart phones, especially in the western world, the use of social media has increased a lot. One of the startling features of Facebook is to provide news to its users; it shows how big and complex Facebook has become over the years. A recent study by Pew Research found that 30% of Americans got their news stories through Facebook (Pew & Project, 2013).

2.4.2.4 Segmentation

Stutzman (Stutzman, 2006) looked at the political affiliation and orientation of students and found that that during the first semester, with so many new friendships, the political orientation hardly changes. They also studied the number of friends for both liberal and conservative students and found that freshmen who were liberal had an average of 115.4 friends; while conservative freshmen had 117.6 friends.

To identify not only trust but also life satisfaction of Facebook users, Valenzuela et al. (Valenzuela et al., 2009) found a positive relationship between students' life satisfaction and their social and civic engagements such as political participation with Facebook usage. They found that on average, everyday students spend between ten minutes and an hour on Facebook. The usage of Facebook is more popular with young users as compared to older ones. To discern between different ethnic/racial groups, Seder et al. (Seder & Oishi, 2009) studied students' subjective well-being, by dividing them into European Americans and non-European Americans. They found that this subjective well-being is quite different between these two groups. Where European American students, having a homogenous Facebook friendship network was associated with higher life satisfaction and positive affect, as well as lower felt misunderstanding, non-European American participants, having a heterogeneous friendship network, no such well-being relationship could be established. So, according to them, diversity has a negative impact on one's well-being.

Facebook itself (Lars Backstrom & Bakshy, 2011) tried to understand how Facebook users divide their attention across their friends or contacts. Based on the activity of users, they selected 16 million Facebook users who had used Facebook at least 80% of the days in 2009 and 2010. To determine the attention of a user given to their friend, the following attributes were taken into consideration: profile views, photo views, messages, comments and wall posts. They identified messages on users' birthdays as an outlier, so they've removed it from their analysis. Female users of Facebook have a larger personal network than males. They have dealt specifically with both genders, male and female, and as to whether they both are in relationships or not, even same gender relationships are analysed (Lars Backstrom & Bakshy, 2011).

2.4.2.5 Trust

When Facebook was compared with MySpace to evaluate how trustworthy it was, Dwyer et al. (Dwyer et al., 2007) found that it was far ahead of MySpace. In Facebook, people tend to reveal more information about themselves like their home town and email address, which is generally not shown in profiles of other [SNSs](#). According to the same study, if we talk about percentage of users in Facebook and MySpace, it was shown that every (100%) Facebook users revealed their real name, as opposed to 66.7% of MySpace users, of course this is based on the sample population of the study. It was identified that , 94% of Facebook members included their email address, compared to about 40% of MySpace members. In terms of relationship status though, MySpace users were more transparent than Facebook users.

2.5 Distributed Social Network

The [SNS](#) such as Facebook utilises a traditional centralised system, a client-server approach in other words, which means that all identities (or profiles) and their social links (the entire social network), are stored and administered on central servers. This approach provides high mobility to users as they can log-in from any computer, but it also implies high dependence on predefined central server(s), which results in the possible exploitation of private data. Our approach is somewhat different. We use a completely decentralised system where users themselves are responsible to create and maintain their social networks. In this section, we cover what research has been carried out in decentralised (or distributed) [SNSs](#) and how does it relate with our work (see Chapter 6).

Since our work was carried out (which is covered in Chapter: Distributed Peer to Peer), a number of new systems have emerged with a much more advanced feature set than ours. We are going to cover some of them in this section. Some have focussed on the existing centralised [SNSs](#); some are geared towards privacy and some are concerned with accessibility. With completely decentralised systems a user faces a dilemma which is succinctly covered by Feldman et al. According to Feldman et al. (Feldman & Blankstein, 2012), if there is a completely decentralised system, then a user sacrifices availability, reliability, scalability, and convenience by storing his/her data on his/her own machine, or even trust his/her data to one of several providers that he/she probably does not know or trust any more than he/she would a centralised provider. DECENT (Jahid, Nilzadeh, & Mittal, 2012), which is based upon Distributed Hash Tables (DHTs), assigns privacy policies with three tier architecture. Each object in their system have either attribute-based (AB) or identity-based (IB) or a combination of both types. Read and append policies are AB, while write policies are IB. This system deals with encryption of the stored data by cryptographic policies and is not concerned with attacks based on routing. Jahid et al. (Jahid et al., 2012) compared two systems, Peerson (Buchegger, Schiöberg, Vu, & Datta, 2009) and Safebook (Cutillo, Molva, & Onen, 2011) which provide access via encryption but are not as detailed as their own system DECENT (Jahid et al., 2012). Also Peerson relies on a central entity by using OpenDHT which does not ensure robustness (had been down for quite some time) (Buchegger et al., 2009).

A number of [SNSs](#) exist which are of a distributed nature, but none of them are purely distributed. There are always some central entities involved in one way or another. Skype, with latest figures from ("Skype," n.d.) show as of 21 January 2013, it has more than fifty million concurrent users online. Skype was bought by Microsoft in 2010. It is a tw-level

[P2P](#) based chat and VoIP system, where normal users (or clients) are considered as 'peers' and specialized clients as 'super-peer'. Any peer with a public IP address having sufficient CPU, memory, and network bandwidth is a candidate to become a super node (Baset & Schulzrinne, 2006). Normal peers connect to super-peers to join the system. Being a closed source system, it is difficult to analyse, but some analysis has been done (Baset & Schulzrinne, 2006). Skype includes only one centralised server for user authentication and to keep usernames unique throughout the system. Each Skype client locally saves information about its friends (Buddy List). It also contains a Host Cache (HC) which bears a list of Super Nodes (SN), in which, at least one SN has to be valid in order for Skype to function. All peers with public IP can potentially become an SN. Searching a user is also possible via Global Index, provided that the target user has logged on in the last seventy two hours.

Maze (Hua, Mao, Jinqiang, Haiqing, & Xiaoming, n.d.) uses a social network to communicate and discover files. It uses a centralised ticketing server known as Ticket Grant Server (TGS), which issues tickets to all peers to identify them. This ticket is then served as a form of legitimate communication/transaction between peers. The ticket is only valid for a single communication. Also, Maze uses another centralised server called a Heartbeat server which, apart from holding a directory of peers, also checks the online status of each of them. In social maze, where friends can help discover new peers via their friends, it can run without involving this server; however, the TGS would still be required.

P-Grid (Aberer, Datta, & Hauswirth, 2004) uses a structured network. For dealing with the identity of peers keeping dynamicity of [P2P](#) in mind, it, just like Tribler (Pouwelse et al., 2008), establishes unique ID locally, which is generated via a hash function of current date and time, IP address, and a large random number. In case of a change of IP of a node, either because it has re-joined the system or DHCP has assigned a new IP, so called replicators help identify that node. Also, based on a structured network, Symphony (Manku, Bawa, & Raghavan, 2003) and SPROUT (Marti, Ganesan, & Garcia-Molina, 2005) target routing strategies based on the social links. SPROUT defines a trust model based on social distance (in terms of hops) between two nodes. The farther the node is in the social network, the lower the trust would be. For a particular key 'k', it forwards the query to one of the online friends whose node ID is closest to 'k'. Look-ahead is also possible with the distance of friends-of-friends id to the target key 'k' taking into account.

SybilGuard (Yu & Kaminsky, 2008) presents a solution to minimise Sybil attacks. It exploits social networks by stating that Sybil nodes can be detected and ignored, since they will not have many trusted links with genuine or trusted nodes. It however, does not describe how the social network would actually be bootstrapped and is left as a future work.

Based on semantic routing, Borch in his work, Social [P2P](#) (Borch, 2005), formed groups in a peer-to-peer fashion. It searched content and then formed implicit groups, leading to the formation of a social network. Also, depending on manual user preferences, over a period of time, a node can become closer to those who are nearer to its interests.

For dealing with an estimation of how many peers' information should locally be saved for gossiping, without knowing the actual number of peers in the system, SCAMP (Ganesh, Kermarrec, & Massoulié, 2003) adjusts local partial views of peer membership accordingly.

2.6 Conclusion

In this chapter, we learned what Agent-based models ([ABMs](#)) are, and what they offer. [ABMs](#) present a more realistic and intuitive approach, whereas macro level phenomenon arise through local interactions between agents and their environment. Agents are autonomous decision makers who learn not only from their history, but from other agents and their behaviours. We also learned about mechanistic approaches to develop social networks by studying random networks (where links are randomly created); preferential attachment (links are created with high degree nodes); scale-free networks (a power-law network with a selection bias for nodes with high links); and small-world networks (a pseudo random network with a low-average path length than a completely random network). After looking at the various studies carried out on [SNSs](#), we learned that none of these mechanistic approaches are good enough to describe and capture their structure. Specifically they do not produce a power-law degree (# of links) distribution with low density, high clustering and high assortativity of the connectivity degree. This is where our work fits, using [ABM](#), which we will cover in more detail in Chapter 4.

We also looked specifically at online social networks, which helped us identify social groups in the online environment, where social networks can be developed and maintained through an [SNS](#). In order to see how [SNSs](#) have evolved over time, we discussed the history of it, starting from SixDegrees.com (inspired from Milgram's experiment) developed in 1997, to Facebook. We briefly looked into various [SNSs](#) (such as SixDegrees which failed to succeed as it was ahead of its time), and how and why they become a success. The key finding we came across was that an [SNS](#) needs to be user-centric, giving more power to users to determine what needs to be changed in it. MySpace, another [SNS](#), took this approach.

We then learned a great deal about Facebook broken down into five research themes: identity and representation, privacy, growth, segmentation and trust. Starting from how it came into being from a student directory system to an online system. Initially, by being a closed system just for the Ivy League universities, gave Facebook a very good reputation and to some extent a status symbol too. When Facebook users were studied to determine how much they trust it and then compared this with users of other [SNSs](#) such as MySpace and Friendster, it stood out (Acquisti & Gross, 2006). We also looked into the activity of users where power users, like any other system, are crucial to maintain a system. The growth of Facebook too was also covered. Although Facebook users are demanding more control over their privacy settings, in terms of default settings however, over time, more information about them has been made public. The key feature of Facebook's success has been innovating interesting features.

In the last section, we highlighted some of the recent developments made in the field of distributed [SNSs](#) since our work and also compared our work with other existing works. We learnt that there is a dearth of completely decentralised [SNSs](#) comparable with our own solution. Successful systems, such as Skype, are not completely decentralised. Although it is a closed system some analysis has been carried out on it. Other systems such as SybilGuard do not give details on how their social networks are bootstrapped. Also, unlike our system, we did not find any such system which was deployed and then used by real users. This chapter has focussed on the methodologies and techniques we have used in our work and why we selected those for our analysis.

3 Chapter: Analysis of Complex Network Datasets

3.1 Introduction

We begin this chapter by outlining major measures – which are sometimes called Social Network Analysis ([SNA](#)) measures, used to determine the structural and semantic relationships in a typical network. These measures are used to validate social network models. In particular, they help us analyse and compare graphs, for example the reference graph to the model output graph. These have been taken from Computer Science, Physics, Maths and Social Sciences. We describe where they are applicable and to what purpose they serve. It is essentially a literature review of network measures to analyse, classify and compare networks.

In order to identify which mechanisms might have been involved in social interaction and then eventually friendship developments among university students, we relied on empirical data to validate our models. These datasets include three US universities, Caltech, Princeton and Georgetown and their Facebook network of students, which were thankfully, shared with us by Mason A. Porter of Oxford University and have been studied by him and others (Traud, Kelsic, Mucha, & Porter, 2008). These datasets are used in our agent-based model (whose details follow in the Section Cross-Sectional Datasets), which helped us identify how students in a university setting engage in various social, educational and recreational activities which lead to friendship development. In this chapter, we are going to highlight what sort of dimensions (nodes and links) these three datasets have and what individual information (about students) they contain. We will also cover some of the [SNA](#) measures of them as well.

3.2 Social Network Analysis Measures

As explained in (Abbas, Alam, & Edmonds, 2014; Alam, Abbas, & Edmonds, 2014), consider a model which generates a social network. Let us assume that we are only interested in the ‘final’ network(s) (i.e., we do not consider transient changes in a network that may have had occurred during a simulation run). The objective then is to compare this ‘final’ synthetic network with an empirical network that is obtained from a target social system as illustrated in ‘right end’. This section tackles the problem of comparing networks. As we have discussed in our work (Abbas et al., 2014), the problem of comparing all possible networks is huge. For example, there are $2^{n(n-1)/2}$ ways of connecting n nodes with undirected links and thus the complexity of the problem grows more than exponentially with the size of a given network. A full comparison therefore remains infeasible for large networks. For example, if you have two networks where twenty five people are connected with each other, the size of the problem of comparing these two networks is that there are more patterns of links between twenty five such people than the number of atoms in the universe¹.

¹ Since $2^{25 \times 24 / 2} > 10^{80}$ (which is an estimate of the number of atoms in the universe https://en.wikipedia.org/wiki/Observable_universe)

Validating simulated social networks

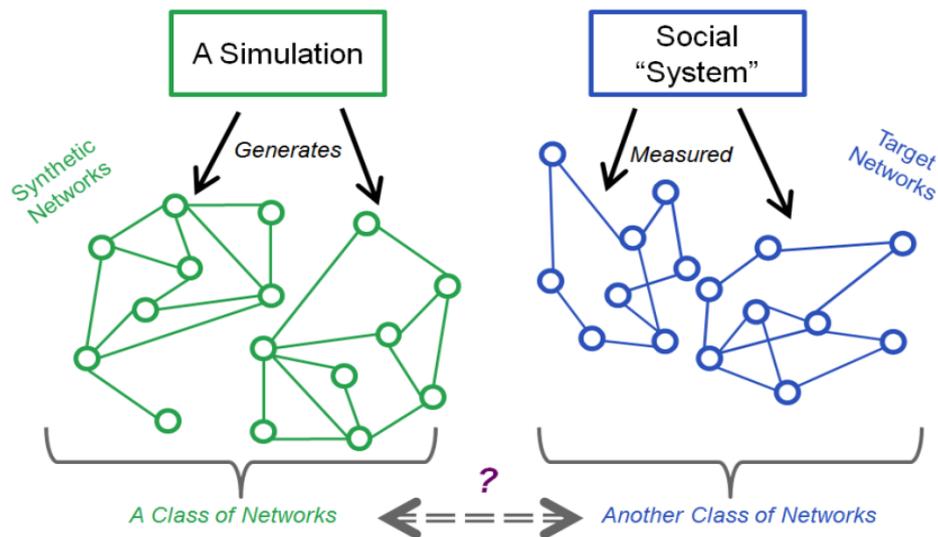


Figure 3-1 - An illustration of the problem of comparing and validating a class of simulated networks (left) to the available social networks of a target social system (right) (Abbas et al., 2014).

Using social network measures to evaluate and compare graphs is common. Given that a single measure (or small set of measures) is not going to establish that one has approximately the correct graph, which means how similar a graph is when compared with the reference graph, the point of this is presumably that a chosen measure captures some important aspect of a graph that is crucial to the kind of process being investigated. The implicit assumption here is that the closer the measure on the synthetic graph is to the same measure on the reference graph, the more similar the two graphs will be in this respect. Thus, [SNA](#) measures are often used to compare how 'close' different synthetic graphs are from a reference graph and thus determine which of several graphs are to be preferred.

How well this works depends upon the extent to which closeness with respect to a measure implies similar networks. Although there has not been much work in general, on this issue, (Kermack & Mckendrick, 1927) Kermack et al. report on how centrality measures degrade with the introduction of random changes in the network, however this result might depend on the topology of the network (IA McCulloh, Johnson, & Carley, 2012). Here we briefly preview some of the most used approaches.

illustrated by (Papadopoulos, Kitsak, Serrano, Boguñá, & Krioukov, 2012) which gives an example of when a model matches the degree distribution well but is shown inadequate when a distribution of node distances is plotted.

3.2.2 Assortativity Mixing

This measure identifies, in general terms, if the nodes with similar degrees are connected with each other, calculated by working out the correlation of the degrees of all connected node pairs in the graph. It ranges from -1 to 1: a positive value indicating that nodes tend to connect with others with similar degrees and a negative value to the contrary (MEJ Newman, 2002). Usually the averaged value of all node pairs is calculated for the whole graph for comparison – known as the global clustering coefficient. This measure identifies degree [homophily](#), love of the same degree. A value near 1 would result from a graph where there are uniformly densely interconnected nodes and other areas that are uniformly sparsely connected. A negative value might result from a graph where nodes with a high degree are distributed away from each other, which is generally how technological networks such as the world-wide-web are structured (MEJ Newman, 2002). Thus this measure is useful when the local uniformity of degree distribution (or otherwise) is important.

3.2.3 Cluster Coefficient

This is defined as the ratio of the number of edges that exist between a node's immediate neighbours and the maximum number of links that could exist between them. In other words, it identifies the proportion of triangles compared with the maximum possible. For directed graphs, the same measure is calculated by ignoring the direction of links.

$$\text{Clustering Coefficient} = \left(\sum_{i=1}^n \frac{\text{number of triangles connected to node } n_i}{\text{number of triplets centred on } n_i} \right) * \frac{1}{n} \quad \text{3-1}$$

, where n represents number of nodes.

If all one's friends know each other this would result in a cluster coefficient of 1, if none knew each other a value of 0. This measure might be important, for instance, if one had hypothesised that the process of making new friends via a 'friend of a friend' mechanism was in play. The local cluster coefficients can be displayed as a distribution or a simple average. This measure might be useful when the empirical evidence being compared against was collected in the form of 'ego nets' (Wellman & Potter, 1999).

3.2.4 Geodesic Distance

The general degree of node separation in a graph is measured through three metrics: average path length, network radius and network diameter. A geodesic distance is the shortest path between any two nodes (the least number of 'hops'). The average path length is the average of all-geodesic distances on the graph (Sala, Cao, Wilson, & Zablit,

2010). The diameter is the largest geodesic distance within the graph. For each node the eccentricity is the geodesic distance to the node furthest from it, the radius is then the minimum eccentricity in the graph. Geodesic distances are important in the presence of 'flood fill' gossip mechanisms where messages are passed on to all of a node's neighbours. The radius is a lower bound for the 'time' (measured in network jumps) for a message to reach all other nodes, the average path length the average 'time' for nodes to receive the message, and the diameter giving an upper bound.

3.2.5 Direct Similarity Measures

If the set of nodes is the same in two different graphs we can calculate the hamming distance between their adjacency matrices, which is how many links are different in each of them (I McCulloh & Carley, 2009; Wasserman & Faust, 1994). This gives a direct count of how many links would need to be changed to make them identical. A similar approach is correlating the columns (or rows) in each adjacency matrix (using, say, the Pearson Correlation Coefficient) and then using these numbers to indicate how much has changed. The big disadvantage of these approaches is that it requires the nodes to be the same, and it may be that small changes affect the critical topology (e.g. disconnecting sections of the graph), however in cases where this is unlikely (e.g. substantially random graphs with high minimum degree) these might be useful ways to proceed.

3.2.6 Eigenvalues and Eigenvector

The eigenvector approach is an effort to find the most central actors (i.e., those with the smallest farness from others) in terms of the 'global' or 'overall' structure of the network, and to pay less attention to patterns that are more 'local'. A node's importance is determined by the sum of the degree of its neighbouring nodes – representing its global centrality.

3.2.7 Feature Extraction

In this section we turn to very large graphs. Many of the above measures require a lot of computational power, making their use infeasible. Bagrow et al. (Bagrow, Boltt, Skufca, & Ben-Avraham, 2007) have introduced a technique to extract a rather small feature representation matrix from the structure of a large graph. For a graph, a geodesic distance based matrix, B-Matrix, is calculated where the n th row contains the degree distribution separated by n hops. For instance, the first row would contain the usual degree distribution of all nodes; while the last row with the highest n , would contain the network diameter. The distance between the nodes is calculated using the Breadth-First Search (BFS) algorithm. The dimension of this matrix is the total number of nodes \times network diameter. For isomorphic graphs, the calculated B-Matrices would be exactly the same. If we want to compare large graphs, this technique can be used to extract featured matrices from both of them and then compare them. For instance, calculating all linked pairs with the shortest distance to identify the average path lengths is computationally infeasible. This method allows one to compare large matrices (and hence graphs).

A comparison of global and subgraph network characteristics is necessary when a generative mechanism is introduced into the agent-based simulations. As observed by Milo et al. (Milo et al., 2002), networks sharing similar global characteristics could exhibit

varying local structures. Global properties of the dynamic network may inform about the robustness of the underlying processes. On the other hand, local properties can show the variability that may occur for different settings for the same processes. Such clusters, commonly called communities, can be defined and identified in two ways: via the pattern of links and by the attributes of the nodes in the network.

3.2.8 Community Detection

Community Detection (Mark Newman, 2010b) is a technique to identify subgraphs where nodes are closely linked together, roughly that there are more links between the nodes of the subgraph than external links to the rest of the nodes (Newman, M. E. J. and Girvan, 2004). Hence, they represent a cluster, a community in other words. There are many algorithms available to identify such clusters, e.g. (Blondel & Guillaume, 2008; Reichardt & Bornholdt, 2006). To identify how closely connected these communities are, the concept of modularity was introduced. This is the fraction of the links that fall within the groups (or communities) minus the expected fraction if the links were randomly distributed. This ranges from 0 to 1, the higher the value, the more cohesive the community. An average value may be used for comparison. This measure, generally speaking, does not deal with overlapping communities, where individual nodes may belong to several communities. Also, the detection mechanism is not very robust. For instance, in a network, if we re-order the underlying edgelist (source and target pairs), which does not change the structure of the network; the community detection mechanism identifies different communities.

For graphs where some attributes of nodes are known, several techniques can be used to match the semantic structure of the graphs. Two are described below:

3.2.9 Affinity

[Affinity](#) (Alan Mislove et al., 2007) measures the ratio of the fraction of links between attribute-sharing nodes, relative to what would be expected if attributes were random. It ranges from 0 to infinity. Values greater than 1 indicate a positive correlation; whereas values between 0 and 1 have a negative correlation. For an attribute of nodes, such as dormitory for instance, we firstly calculate the fraction of links having the same dormitory. It is represented by:

$$S_a = \frac{|\{(i, j) \in E : s. t. a_i = a_j\}|}{|E|} \quad 3-2$$

where a_i represents the value a for a node i . In other words, we are identifying the total number of matched nodes with the same attribute values for an attribute a . E represents the total number of links. Next, we calculate E_A which represents the expected value when attributes are randomly assigned. It is calculated by:

$$E_a = \frac{\sum_{i=0}^k T_i (T_i - 1)}{|U|(|U| - 1)} \quad 3-3$$

where T_i represents the number of nodes with each of the possible k attribute values and U is the sum of all T_i nodes, i.e., $U = \sum_{i=0}^k T_i$. The ratio of the two is known as *affinity*:

$A_a = S_a/E_a$. This measure is then used to discover 'attribute level communities', that is subgraphs with high affinity. Either the affinity of a whole network could be compared with another or this is used to identify communities and then the presence of these is compared. This measure, however, deals with discrete attributes. It will be interesting if attributes with continuous measures (such as transitivity), can too be applied. This can be achieved by slightly modifying S_a (in the above equation) to deal with ranges.

3.2.10 Silo Index

This is an index which identifies the proportion of links between nodes with the same attribute value in a network (Krackhardt & Stern, 2011). If all the nodes that have a value Y for an attribute, only have links to other such nodes and not to nodes with any other values of that attribute, that means a very strong community exists, which is totally disconnected from the rest of the network. Such a set of nodes would have a maximum value of this index. In short, this index helps us identify how cohesive inter-attribute links are. It ranges from -1 to 1, representing the extreme cases (no in-group links to only in-group links respectively). It can be written as:

$$\frac{I - E}{I + E} \quad 3-4$$

where I represents the number of internal links and E the number of external links. In other words, it is the ratio of the difference in internal and external links, to the total links. It is quite similar to E-I index (Krackhardt & Stern, 2011), but with the opposite sign. An E-I index gives a value 1 when all links are external, while Silo Index has value 1 when all are internal. Hence, the Silo Index could be written as an I-E index. One of the most attractive features of the I-E index (just like E-I index), is that as it is a ratio and not dependent on the density of the network (Everett & Borgatti, 2012). So Silo Indices calculated on various networks of different sizes can be compared, if need be. Also the underlying principle behind the measure (of taking the ratio between internal and external counts of measures) can be applied to other centrality measures such as degree centrality, betweenness and eigenvector centralities (Everett & Borgatti, 2012). For a grouped node, from our work, we have identified that this measure biases nodes with a bigger size than others. Details of which will be covered in the subsection Contracted Graphs.

To identify the underlying processes of a network generation, these techniques can be used to identify which parameters may be more important than the rest. See (Abbas, 2011a, 2013) as examples.

3.3 Cross-Sectional Datasets

From now on we will concentrate on different types of dataset. These included three universities which are Caltech, Princeton and Georgetown. The information includes a cross-sectional data of social networks of students which have been taken from Facebook. In particular, it contains the attributes of students (for instance what major of studies they have; what dormitory they live in etc.), and their friendship network within the same university. It does not cover friendships (or any other kind of relationships) students might have outside their university. These datasets are bases for our [ABM](#) model, which helps

us identify how students in a university setting engage in various social, educational and recreational activities which then lead to friendship development. We will cover our model in more detail in the next chapter (Chapter 4). We also briefly cover the structural and attribute spread of these datasets. Also a few of the [SNA](#) measures will be discussed at the end.

3.3.1 Caltech

We have used the data of students of Caltech who use Facebook. This was provided to us by Mason A. Porter of Oxford University, and has been studied by him and others (Traud et al., 2008). The dataset includes both the attributes and social structure for 769 students. On average, each person has almost 43 friends. This dataset only represents intra-institute relationships, which may be the reason why we do not observe the average number of friends, as stated by Facebook (Facebook, n.d.-a) (130 friends). We note that it is a snapshot; it represents only links and attributes present at one single point of time. The data is completely anonymised where simple integer values represent each attribute.

Each student has the following four attributes: their dormitory; their year (1st, 2nd, 3rd, etc.); their 'major' (their main subject of study); and the high school they came from. In the dataset, three more attributes are mentioned for each individual. This includes: gender, status (student or faculty) and 'second_major' (second major of studies). However we have not considered them. Gender was dropped because its [affinity](#) is closer to 1 (1.08 to be precise), deeming it not very useful, because this means it has a very weak correlation for friendship development between any two students when same gender is considered. As for status, it has four values (1, 2, 5 and 6), without any description of what those values mean. In our work, we mainly focus on students, hence we considered all of them to be students. In terms of second major, we found it as the least helpful as it has the highest number of missing information. Out of 769 students, it is missing for 573 of them, almost amounting to 75%. The [affinity](#) measure of the four attributes is shown in Table 3-1. Apart from high school (0.36), all the attributes have a positive [homophily](#) (values greater than 1). This means that students who are friends (linked together) tended to have the same dorm, major or year attribute but not the same high school. We see that the dorm attribute has the highest [affinity](#), which corresponds to the community structure of it and where it was shown that the dorm is the most dominant attribute (Traud et al., 2008). Also similar to Traud et al.'s work where the goal was to identify communities, we used the same four underlying attributes as they did.

Table 3-1 – Caltech Affinity Measures

Attribute	Affinity
Major	1.48
Dorm	3.34
Year	2.45
High School	0.36

Table 3-2 - Caltech University's Attribute Spread

Attributes	Dorm	Major	Year	High School
Missing (%)	22.36	10	14.82	17.42
Unique	9	31	18	501
Average population	85.44	24.80	42.72	1.53
Proportion of average population of total population	11.15	3.22	5.55	0.19

In the dataset, there are a total of 501 high schools and 31 majors, showing the diverse background of the population. Out of the 769 people in the dataset, 501 have all the values for each attribute and the total number of links between them is 16656. Missing (%), in Table 3-2, summarises the extent of missing data (indicated by a '0'). Given this data includes only 769 students, we see a very high number of high school (501). In terms of calculating proportion of average students of total population, for each of the four attributes (see the last row of Table 3-3), we took these two values into account: the average population, and the total student population. So for dorm, it is calculated by taking an average population (85.44) and then dividing it by the total population (769) and then multiplying it with 100. This shows, at an attribute level, what proportion of total population falls in a unique attribute value (say a proportion of students on an average living in a dorm). It shows that the highest proportion of students is in dorm (11.15), then year (5.55), major (3.22) and lastly high school (0.19) follows. This means the dorm attribute seems most populated, hence the higher the chances of suitable friendship development between students exist. In the dataset, there are nine dormitories (including one unknown), whose spread can be seen in Table 3-3.

Table 3-3. Caltech Dormitory Distribution

Dorm	165	167	171	166	168	170	172	169	Unknown
Distr.	44	63	67	70	76	87	91	99	172

As for the social network of the Caltech dataset, we have shown this in Figure 3-3. For visualisation of the network, we have used the OpenOrd algorithm (Martin, Brown, & Klavans, 2011) which uses a force directed algorithm to identify close clusters of nodes. Nodes which are connected to each other by a common connection (of another node) are pulled together, while those which are not close are pushed apart (Herdağdelen, Zuo, Gard-Murray, & Bar-Yam, 2013; Martin et al., 2011). The group of nodes which are linked tightly together form a cluster. Also, for a better visualisation, this algorithm preferentially cuts long links between nodes of high degree (Boyack & Klavans, 2010). You can see several clusters shown in the figure, which shows there are strong communities in the graph. We cannot, however, infer how those communities were formed and what factors might be most important when such a graph is formed. However this shows that only a handful of communities exist in the network which gives an impression that students have tightly knitted groups. Also the number of communities are almost equal to the unique number of dorms. Hence it shows that there is a high clustering within each dorm. Out of

the whole population of 769 people, only 7 are disconnected from the main component (isolates in other terms), and are in three groups of 2, 2 and 3 people. During our analysis, we have not removed these three isolated groups.

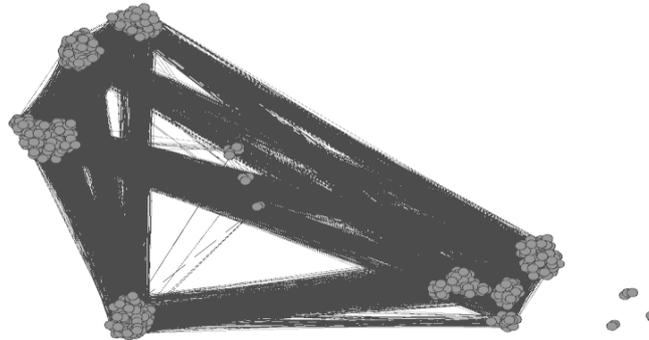


Figure 3-3 - Caltech Social Graph with clusters

We have also plotted the degree distribution of the social network in Figure 3-4. It illustrates that the power-law effect kicks in around the degree 100. This effect illustrates an exponential number of links, which means there exist a few nodes with very high connectivity (high degree) and the rest have a very low connectivity (low degree). This is further strengthened by running Kolmogorov-Smirnov test. It identifies whether the power-law outlook is statistically significant or not. If the probability is greater than 0.05, then it is. For Caltech we found that it was indeed greater than 0.05 ($p > 0.05$). Our understanding is that since this graph only contains inter-university links, not all students link with higher probability to the popular students (students with a very large number of friends), which is represented by a straight line in a log-log plot.

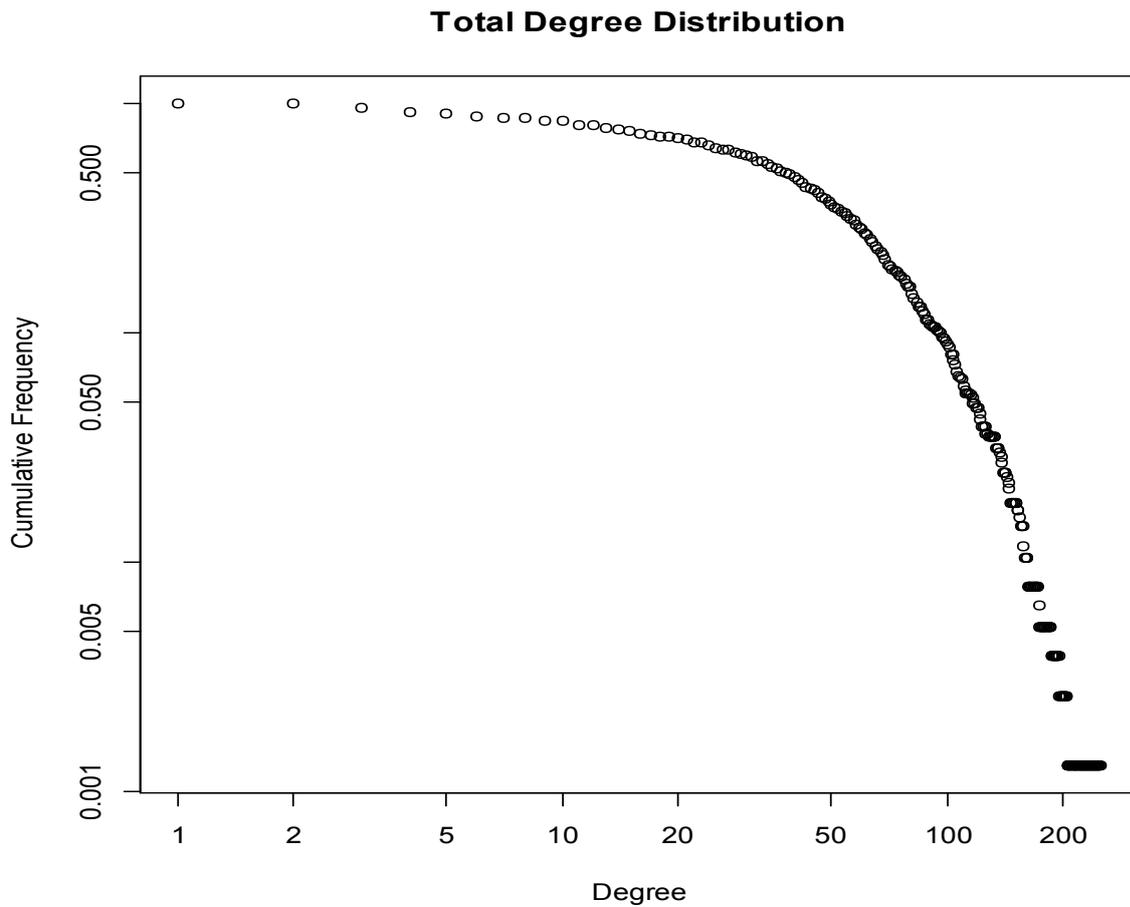


Figure 3-4 - Total Degree Distribution (log-log plot) of Caltech Social Network

3.3.2 Princeton

The underlying anonymous dataset of Facebook includes both the attributes and social structure for 6575 students of Princeton University. In total there are 293307 links – averaging to 89.2 friends. Each person has four attributes, which are: major course of study (major); their place of living (dorm); the year they joined the university, and their high school information. As for the spread of each attributes and their missing values, we have summarised this in Table 3-4.

Table 3-4. Princeton University's Attribute Spread

Attributes	Dorm	Major	Year	High School
Missing (%)	33.76	24.86	11.77	20.7
Unique	57	41	26	2235
Average population	115.72	160.88	244.30	2.95
St. Dev.	293.06	268.68	399.13	29.21
Proportion of average population of total population	1.76	2.44	3.71	0.04

Compared to Caltech covered earlier, also as Abbas (Abbas, 2011a) found, we see a very diverse population shown by the number of various high schools (see Table 3-4). Missing information in the dataset has been coded by 0. We have dealt with it carefully in our model, which will be covered in Chapter 4. Proportionally, unlike Caltech, the year attribute has the highest value, meaning the highest proportion of students share the year attribute. This means that any two random students would have the highest probability of sharing the year attribute. We have also calculated the [affinity](#) measure (explained earlier in Section 3.2.9) for the underlying dataset. In Table 3-5 you can see the [affinity](#) measures for the four attributes. Let us just repeat what this measure means: the values between 0 and 1 (less than 1) show a negative correlation and 1 and above show a positive correlation. It seems the highest [affinity](#) is of the year attribute with 4.07 value. This means that students connected by a link are 4.07 times more likely to share the year attribute than would be expected if attributes were random. The major and the dorm attributes have somewhat closer values (1.32 and 1.48), whereas the high school [affinity](#) is below 1 (0.89), which shows that it is negatively correlated with friendship link development. If we look at the number of dorm in Table 3-4, it is fairly large when compared with Caltech dataset (57 versus 9).

Table 3-5 Princeton Affinity Measures

Attribute	Affinity
Major	1.32
Dorm	1.48
Year	4.07
High School	0.89

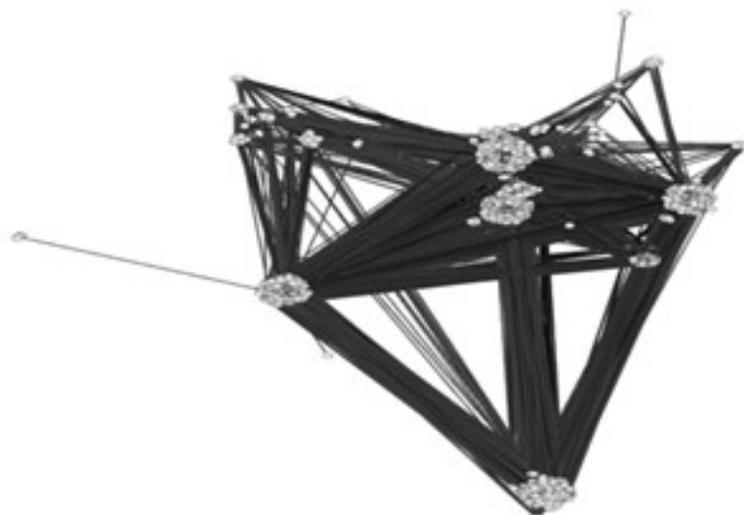


Figure 3-5 – Princeton Social Graph

As for the network structure, Figure 3-5 shows what it looks like. We can clearly see groups/communities in the network structure, which identifies the hub structure of the graph. From this figure alone, we cannot ascertain how those communities might have developed. However, if we look at Table 3-5, we can estimate how the year attribute (with the highest [affinity](#) of 4), might have a role in inter-year [homophily](#), in terms of causing strong communities. The number of communities seems around to the unique number of years attribute (26) in the dataset. Unlike Caltech's network, we do not see the number of communities comparable to unique dorms (57); meaning that these clusters cannot be dorm-based. In terms of the underlying degree distribution shown in Figure 3-6, just like Caltech University's graph, this too has a power-law outlook (straight line in a log-log plot) after a specific degree (200 in this case). This means that, after the degree 200 or so, there is a limited number of students who have a high number of friendship links; the rest of the students have very few friendship links. The power-law effect is statistically confirmed using the Kolmogorov-Smirnov test which produced p-value > 0.05.

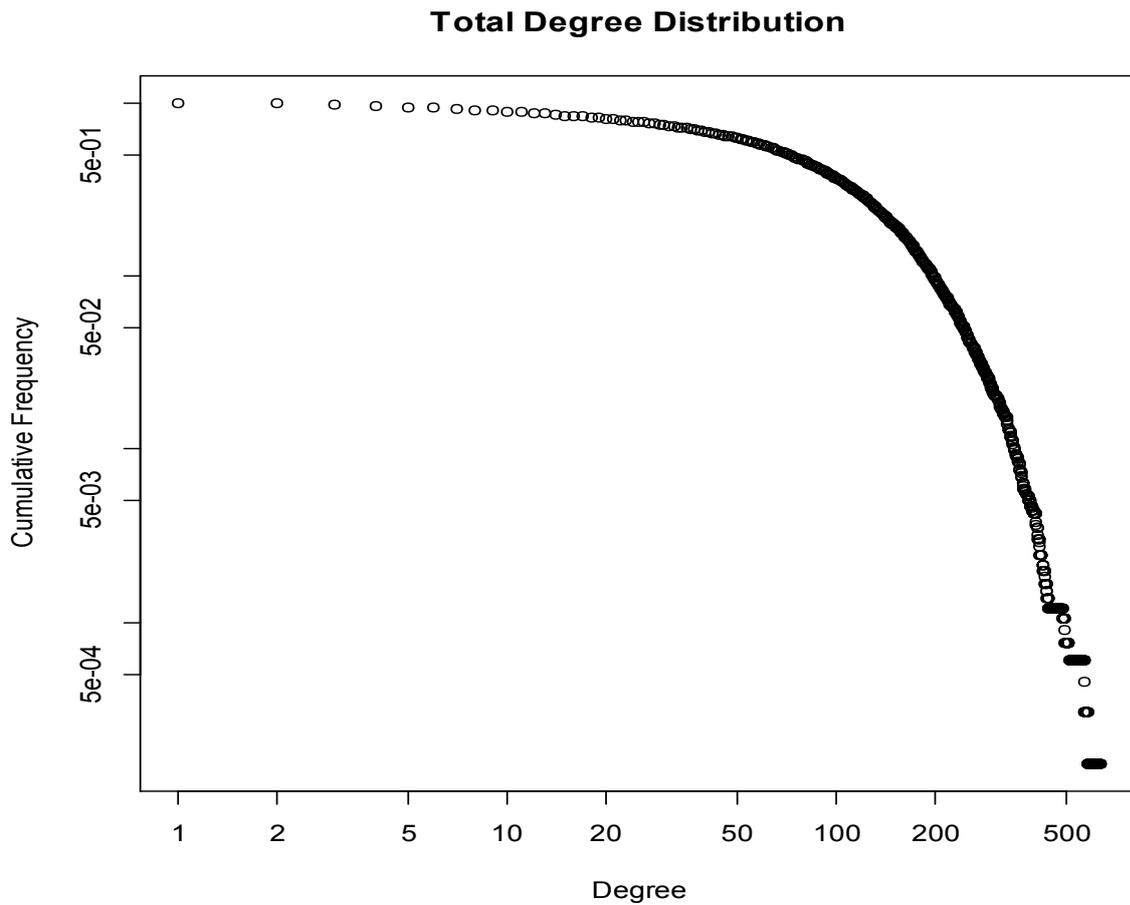


Figure 3-6 - Total Degree Distribution (log-log plot) of Princeton Social Network

3.3.3 Georgetown

Just like Caltech's and Princeton's datasets, this underlying anonymous dataset of Facebook includes both the attributes and the social structure for 9414 student of Georgetown University. In total there are 425638 links – averaging to 90.4266 friends. Each person has four attributes, which are: major course of study (major); their place of living (dorm); the year they joined the university, and their high school information. As for the spread of each attributes and their missing values, we have summarised it in Table 3-6.

Table 3-6 - Georgetown University's Attributes Spread

Attributes	Dorm	Major	Year	High School
Missing (%)	29.85	20.15	10.92	19.55
Unique	16	90	23	2874
Average population	588.37	104.6	409.3	3.27
Proportion of average population of total population	6.25	1.11	4.34	0.03

As you can in Table 3-6, just like Princeton University's dataset, the highest number of missing information – represented by 0, is in dormitory attribute (almost 30%). There are 16 dormitories in total. When compared to Princeton, it seems there are many fewer dorms. Proportionally the dorm attribute has the highest value (6.25), and then the year attribute comes. However looking at the missing information of the two attributes (30 versus 11), the year attribute seems more representative of the students. We have shown the distribution of dormitory population in Table 3-7. Apart from one dormitory, 83, the distribution of students is quite even. On average each dormitory has 588 students. In the case of major (major course of study), there are 90 majors. 20% of this information is missing. The year attribute has the lowest number of missing values (almost 11%). There are 23 different year values ranging from 1955-2010. As most of the people in this dataset are students, the concentration of them lies mostly in 2004 onward in the year attribute.

Table 3-7 - Georgetown Dormitory Distribution

Dorm	69	70	71	72	73	74	75	76	77	78	79	80	81	82	83	Unknown
Distr.	444	406	639	328	294	307	400	248	283	269	178	205	433	247	1923	2810

We have also calculated the [affinity](#) measure on all the four attributes (dorm, major, year and high school). Just to iterate, in [affinity](#) measure, the values between 0 and 1 (less than 1) show a negative correlation and 1 and above show a positive correlation. These can be seen in Table 3-8. The year attribute, again same as the Princeton's dataset, seems the most important factor when it comes to developing relationships based on [homophily](#) – 4.26 showing positive correlation between similar year friendship links. All of the attributes have a positive correlation with least valued attribute being high school (Table 3-8). This means that the students connected by a link are more likely to share these four attributes than would be expected if attributes were random.

Table 3-8 - Georgetown Affinity Measures

Attribute	<u>Affinity</u>
Major	1.36
Dorm	1.62
Year	4.26
High School	1.02

As for the graph, we have shown the greatest common component (gcc) of the Georgetown's social graph in Figure 3-7. It covers 99.7% of the total graph (26 nodes have been removed which were not connected), so 0.3% of the network is of nodes unconnected with this. This visualisation helps us capture the bigger picture: we can see both strong and weak communities. However, we will not be able to tell from this figure alone how those communities developed. In terms of visually investigating the communities (close knitted nodes), sum of them seems to be associated with the unique number of year attributes (23) in the dataset. In terms of population, the Georgetown dataset is the biggest one (9414 students) among the three datasets; hence we see more communities there. In the Georgetown dataset the more central communities are bigger than those which lie towards the edge of the figure.

In terms of the total degree distribution we have plotted it in Figure 3-8. Similar to the Princeton's figure, it seems that the overall distribution does not have a full power-law outlook. It is observed around 200 and onwards, when it becomes a straight line, representing a power-law effect. This means that, after the degree 200 or so, there is a limited number of students who have a high number of friendship links; the rest of the students have a low number of friendship links. We statistically confirmed this effect using the Kolmogorov-Smirnov test, where it was found to be significant ($p > 0.05$). As we are only looking at inter-university links, we might not get the full social network of each student. Hence the overall distribution does not look like a straight line. In terms of size of communities, we find quite a diverse set of them.

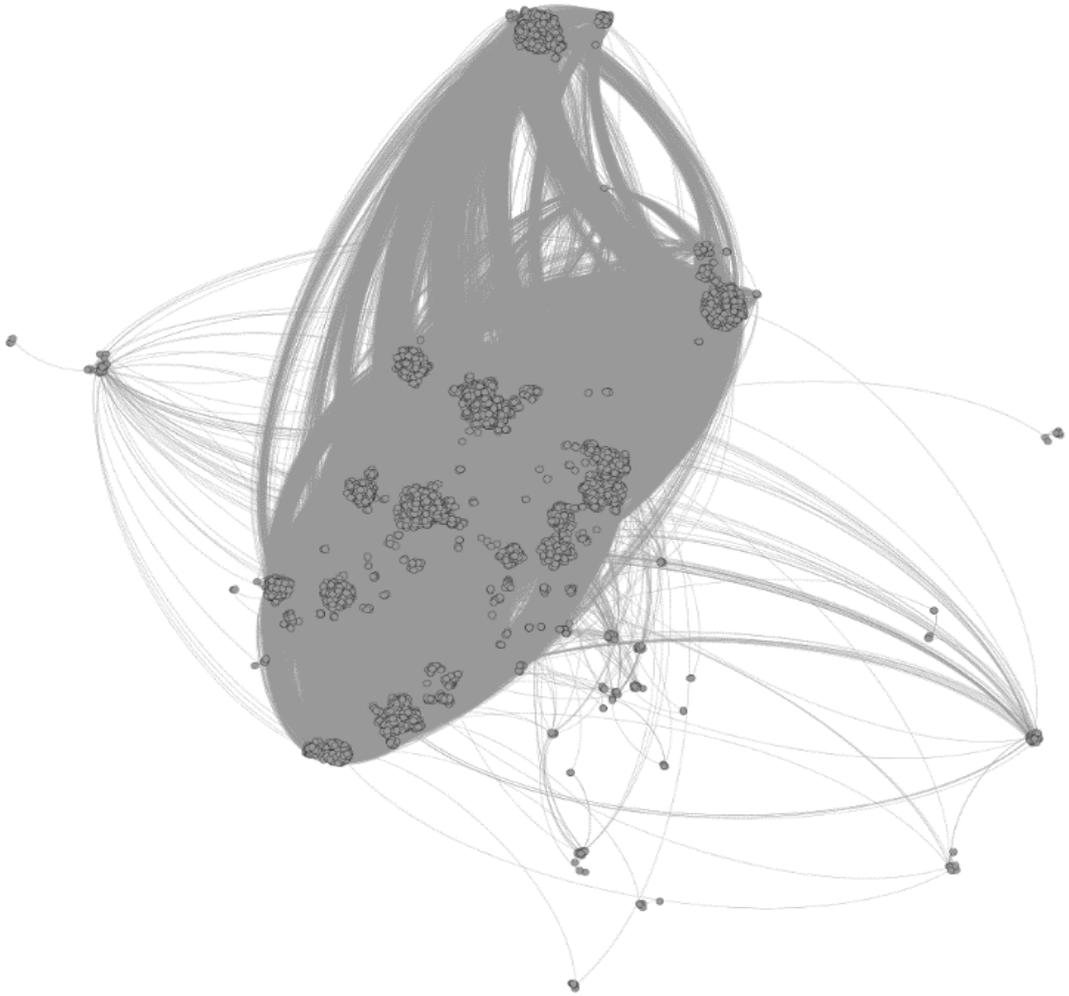


Figure 3-7 - Georgetown Social Graph

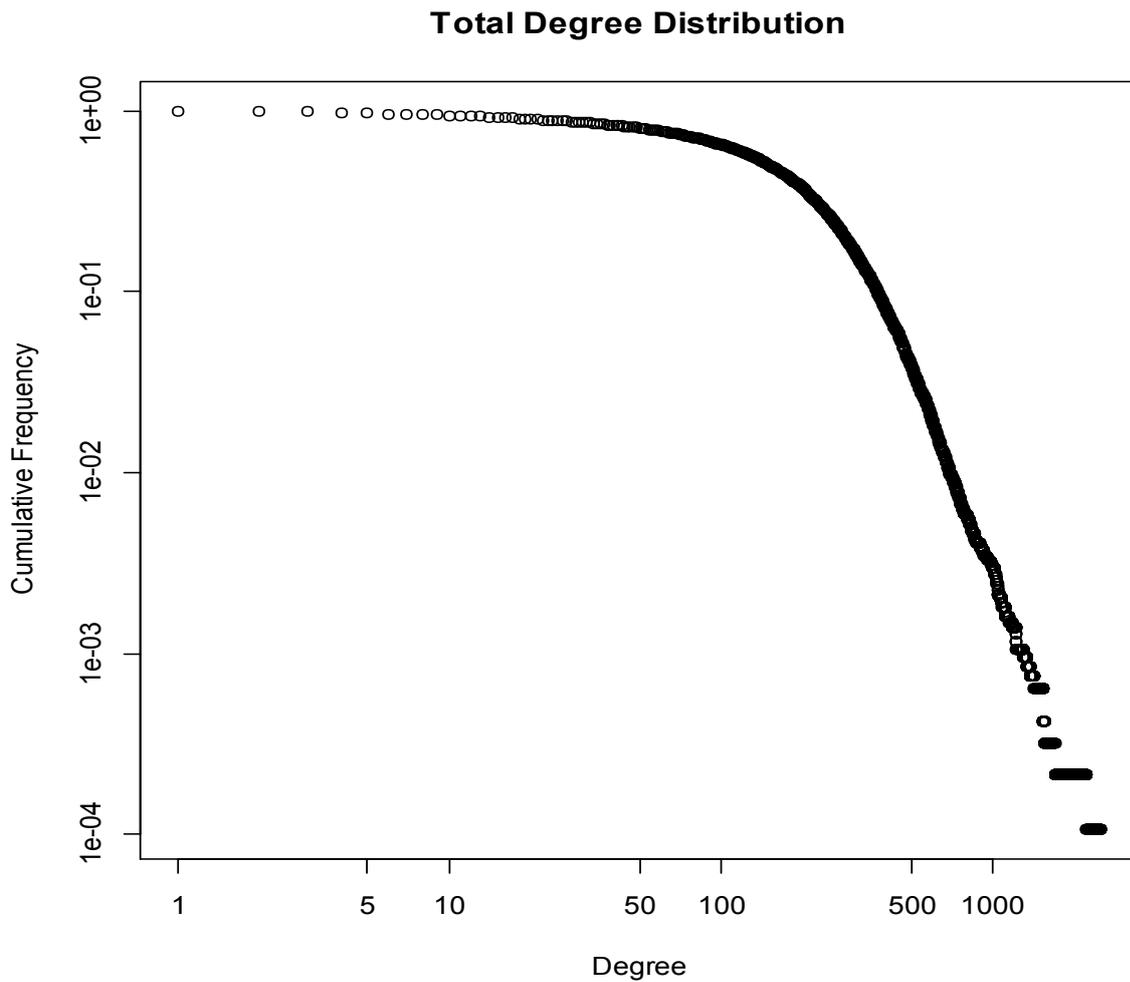


Figure 3-8 - Total Degree Distribution (log-log plot) of Georgetown Social Network

3.3.4 Discussion

Social Network Analysis ([SNA](#)) measures for all the three datasets is set out in Table 3-9. From only the number of links (degree) in the Caltech, Princeton and Georgetown datasets, we can classify them into small and big universities. Caltech having merely 43 friends on average represents a small university, whereas the other two represent big universities, where students have 89.2 and 90.4 friends on average. These datasets, one should clarify, do not represent all the students in a university, but it can be used as a proxy of the population of students. Also the links within each dataset contain only inter-university links. We believe this is the reason why we do not see a straight line in the overall degree distribution (which is an indication that degree distribution is of a power-law distribution) in the output graphs (see Figure 3-4, Figure 3-6 and Figure 3-8), which is typical for social networks.

Table 3-9 - [SNA](#) Measures for all the three datasets

Measures	Caltech	Princeton	Georgetown
Mean Degree	43	89.2	90.4
Cluster Coefficient	0.29	0.16	0.15
Assortativity (Degree homophily)	-0.07	0.09	0.08
Community Modularity	0.31	0.37	0.30
Highest Affinity	Dorm with 3.34	Year with 4.07	Year with 4.26

If we look at the cluster coefficient for the three datasets in Table 3-9, it shows that Caltech with 0.29 represents a more cohesive social network, when compared with 0.16 for Princeton and 0.15 for Georgetown universities. Meaning that students at Caltech tend to form more friendship links with their friend-of-friends. To rephrase, cluster coefficient captures an average of local clustering of all nodes' immediate neighbour connected to themselves – (the ratio of the number of edges that exist between a node's immediate neighbours and the maximum number of links that could exist between them). By merely looking at the cluster coefficient value for the Caltech dataset, we can say that the tendency of students to make friends with the friends of their friends, which works on triangulation of friendship development, should be the best suitable model. Princeton and Georgetown, being bigger datasets, have a positive assortativity (positive correlation) for degrees (0.09 and 0.08), whereas Caltech has a negative assortativity (-0.07). It shows, comparatively with the Caltech dataset, there is a stronger degree [homophily](#) (nodes connected with other nodes with similar number of friendship links) for Princeton and Georgetown. If we assume the number of friends represents popularity, then it means in Caltech popular students are not solely connected with other popular students but with other students with lower popularity, and vice versa – there is a mixture in friendship. However in Princeton and Georgetown, although somewhat with a weak affinity, the most popular students are connected with other popular students and the students with low popularity connected with other students with low popularity. In terms of affinity (fraction of links between attribute-sharing node), dorm for Caltech, and year for Princeton and Georgetown are the highest values. This means for Caltech students living in a same dorm are more likely to form a friendship, whereas for Princeton and Georgetown, the highest probability of two random students to form friendship is when they share the year attribute (the year of enrolment). Proportionally if we look at the concentration of students against unique attributes, we see that the dorm attribute is the most populated (11.15) for Caltech. Controlling for the missing information of attributes, then year is the most populated attribute for both Princeton (3.71) and Georgetown (4.34) universities. This affirms that affinity measure is a good indication of the most important attribute involved in the development of friendship links. In terms of community modularity, which determines the difference between the fraction of the links within communities (cohesive subgraphs) and the fraction of links randomly distributed, there is hardly any difference among the three datasets (0.31, 0.37 and 0.30). It shows that this measure cannot be used as a proxy to determine the complexity of the overall graph structure. However it shows that there is a strong community structure present in all three datasets.

3.4 Critical Analysis

As we demonstrated, there are many measures which help us compare graphs. The central question is, however, how does one determine which measures to use? To answer this, we first need to clarify the main goal of the comparison; also these measures can be sensitive to the overall size of the graph. Generally speaking most of these measures have been designed for a rather small network, where an individual's importance can be assigned a specific role. For instance if a node, with just two undirected links is between two almost disjointed communities, we can assign a broker role to it. For large networks, such as those here, due to their complexity, it is difficult to assign a role to a node's position.

In terms of datasets, we have used datasets of rather small universities; Caltech being a small technical university, whereas Princeton and Georgetown are private universities. It will be interesting to identify a public university's structure, both in terms of network structure and attribute spread.

There is a fairly high level of missing variables in these datasets. The lowest is 10% (major in the Caltech dataset), and the largest 33% (dorm in the Princeton dataset). Nodes with missing values might be forming communities in the network structure for the datasets. Also, when calculating [affinity](#) measure, we do not discard nodes with missing values. They might be playing a bigger role than is indicated by these values, a factor that needs to be further investigated.

In terms of assortativity, which represents degree [homophily](#) (nodes connected with other nodes of the same degree), we find the Caltech dataset to be different when compared with the Princeton or the Georgetown datasets, as it has a negative assortativity (-0.07), meaning there is a mixture of friendship between popular and not-so-popular students. As we have summarised earlier (see Section 2.3), one of the distinctive characteristics of typical social networks are their assortativity. However, Caltech does not display this. It may be due to the small size of the network, which might indicate that it is not a good representative of the bigger and actual social network of students at Caltech. If we look at the current number of students (which might not be that different from that in 2005, when the Facebook data was collected), we see that in total Caltech has 2255 (1001 undergraduate and 1254 graduate students) (University, n.d.), which is almost thrice of 769. It seems Caltech might have a power-law outlook after a degree 100 (see Figure 3-3), however, when we find the best suitable distribution, normal distribution is a better candidate, which is atypical for a social network. We will cover more details on this in Section 4.7.

3.5 Conclusions

In this chapter we discussed all the datasets which we have used in our analyses. We also covered some of the major Social Network Analysis ([SNA](#)) measures used in our work. We described how and when they are applied and how they are useful. In the end we described macro level measures ([SNA](#) measures) of all the datasets of Facebook used in our agent-based model. We will cover them again when we cover the chapter on Agent-Based Model.

To evaluate graphs, the common measures, known as [SNA](#), have been described which are not only used to analyse a graph, but compare them with each other. We list the prominent such measures taken from Computer Science, Maths, Physics and Social Sciences.

After analysing the [SNA](#) measures and the structural properties of three datasets: Caltech, Princeton and Georgetown, we can divide them into small and big universities; Caltech being the smaller university, with Princeton and Georgetown being bigger universities. From just the cluster coefficient measure, we can already see the difference between them, where Caltech has almost double the cluster coefficient (0.29) than that of Princeton's (0.16) and Georgetown's (0.15). It means there is a double probability of one's neighbours (friends) connecting with other neighbours for Caltech. In terms of [affinity](#), we found similar results for Princeton and Georgetown universities, where the year attribute is the most important one with almost 4, implying that students connected by a link are 4 times more likely to share the same year than would be expected if attributes were random. This is further strengthened by looking at the highest proportion of students falling under a unique attribute for all the four attributes. In terms of degree distribution, in all three cases, we did not find a straight line for the overall distribution in the log-log plots, due to the datasets having only inter-university links (no outside links are captured). However, after a certain degree, there is a power-law outlook in all of them, which is considered typical for a social network.

4 Chapter: Models of Student Interaction

4.1 Introduction

In Chapter 2 we explained how, since the advent of online Social Network Systems ([SNSs](#)), the Internet has become a ubiquitous presence in most people's everyday life. Billions of people have a presence on the Internet via an [SNS](#) 'profile', which is a publicly articulated webpage describing a virtual self. According to Pring (Pring, 2012), as of 2012, there are now over 2.8 billion social media profiles, representing around half of all internet users worldwide. Not only can people present themselves, they can also present their social network. Since 2004 when Facebook, (currently the most popular [SNS](#)) came into being, there has been a lot of research on how people form friendships and interact within it, e.g. (Dekker, 2007; Lewis et al., 2008; Marmaros & Sacerdote, 2006) As of June 2014, Facebook has 1.32 billion monthly active users to its credit (Facebook, n.d.-b), and it would not be an exaggeration to say that for many of these people, Facebook serves as the gateway to the Internet.

The aim of this chapter is to reconstruct the development of the social network, of a target [SNS](#), with the help of an agent-based methodology, so that a possible history of the social network and an understanding of it could be developed.

A lot of social network based models have been proposed, from a general, but realistic social network (e.g. see (Lynne Hamill, 2010; Lynne Hamill & Gilbert, 2008) to a data-driven students' social network (Singer, Singer, & Herrmann, 2009). However, these works do not address the process by which such a network might develop in an online environment. From previous chapters, especially from Chapter 2 where we covered the literature work on [SNSs](#), we learned that none of the mechanistic approaches are good enough to describe the process by which social networks and their structure come into being. We also learnt how realistic modelling techniques such as Agent-Based Modelling ([ABM](#)) can help us understand and capture the relationship between the emergent structure of these networks, linked to the micro-level processes that bring them into being. This chapter attempts to address this point. First, we simulate some possible strategies of how students meet and develop their social network. Then we compare the obtained results with the underlying target real-world dataset we have used, and in this way are able to make some inferences as to the strategies that the students used.

The main motivation behind this chapter is to understand, realise and explain how students interact in their social life and then develop social links with each other. We analyse various factors that impact upon friendship development, including the role of inherent attributes such as dormitory and the network position. We develop an agent-based simulation which helps us to model the students' interaction within university environments more realistically. This is an explanatory model. It aims to see which of several hypotheses best explain the social structures that are observed. Students meet and interact not only in their university, such as lecture halls, but also outside it, such as at parties and in the dormitories they live in. Keeping such a social life in mind, we hypothesised a set of possible interaction strategies that are present in students' lives. [SNSs](#) such as Facebook allows the sharing of the virtual-self including details such as gender, age and school affiliation, but this information alone is not enough to identify the exogenous social settings which resulted in any two people's friendship. A mechanism is required that includes both endogenous and exogenous aspects of friendship formation. From the perspective of [SNSs](#), such as Facebook, link prediction and recommendation (as it is commonly known in computer science) is quite a challenging, and lucrative feature. This model tries to explain the social interactions between people might result in their

social network. From information sharing to future business partnerships, the relationships which are developed at university have a significant impact on one's subsequent life (Mayer & Puller, 2008). Also, the friendships made during the earlier university years are crucial for not only staying (continuing education) in the university, but remains a crucial place for the meeting up of people. This is reflected by the friendship development in Facebook (Adamic, 2012). A large-scale analysis of the spread of a particular meme was studied (Adamic, 2012), and this suggested that the majority of people on Facebook had met their friends in school settings, for instance, due to being in the same class/grade. The study also shows that this is true for people regardless of age and gender.

We will be using three Facebook datasets of student networks from three US universities, Caltech, Princeton and Georgetown, which were, thankfully, shared with us by Mason A. Porter of Oxford University, and have been studied by him and others (Traud et al., 2008). The key focus is on analysing how students interact and build their social network over time. For each of the datasets, we can then see which friendship formation process best explains the observed social network structure as judged by a comparison with the reference dataset. In this chapter, the term *agent* will be used to refer to a student. In order to explain our model, we have used a standard protocol, known as Overview, Design concepts, and Details (ODD) (Grimm et al., 2010). We have provided the ODD description and also the whole model at the link below². This protocol was specifically developed to describe the details of an agent or an individual based model.

4.2 Overview

4.2.1 Purpose

The purpose of this agent-based model is to understand and explore the friendship development processes in Facebook based around four hypotheses of student link building. The scenarios of interactions have been drawn from real life interactions of students, details of which follow in Section 4.2.3. The model is used as a *search* mechanism to identify which interaction strategy captures the best representation of a Facebook social network of three different Universities (Caltech, Princeton and Georgetown Universities). The ultimate aim is to see which of a set of plausible processes (all consistent with what is known about student life) best explains the social networks that emerge.

4.2.2 Entities, State Variables, and Scales

In Table 4-1 we set out the state variables and their ranges in our [ABM](#). Before running our [ABM](#), we kept the total number of links for the underlying social network fixed, depending on the dataset we are using. Later that we discuss the fixed preference for the four attributes we are focussing on, which are **dormitory**, **major course of study**, **year of enrolment** and the **high school information** of each student. The values of these attributes come from calculating [affinity](#) measures on the dataset.

² <https://www.openABM.org/model/4350/version/1/view>

Table 4-1 - State Variables and Scales

Variable name	Brief description
Target number of links	Total number of links when the simulation is to stop – the total number of links in the reference dataset (16656 for Caltech, 293307 for Princeton and 425638 for Georgetown)
Random Seed	Dynamic random seed for simulation
DormPref (DP)	Preference of inter-dorm homophily (0-100)
MajorPref (MP)	Preference of inter-major homophily (0-100)
YearPref (YP)	Preference of inter-year homophily (0-100)
HighSchoolPref (HSP)	Preference of inter-high School homophily (0-100)
SimulationMode	Identifying friendship formation process (labelled 1-4) is being used (see Section 4.2 Process Overview and Scheduling for details)
Cluster Coefficient	At each simulation step, the overall cluster coefficient (number of triangles) is calculated.
Mean and Standard Deviation	Mean and Standard Deviation in the number of links of each agent is calculated and then recorded in a file.

For each agent in our model, the characteristics that we assign them are shown in Table 4-1. We assign a unique ID to each agent. We then initialise the population of our agents from the dataset we are using. It includes assigning four attributes (dorm, year, major and high school) from each student to an agent, so that an agent represents a student. Also each agent keeps track of the number of links it has made in each simulation step (tick). This information is stored in the *Friends Count* variable.

Table 4-2 - Agent Level Variables

Variable name	Brief description
ID	Identity of Agents – an auto increasing number starting from 1.
Dorm	The dormitory/hostel of an agent – an integer number
Year	The year of joining the university of an agent
Major	The major course of study of an agent
High School	The high school number of an agent
Friends Count	Total friends count

4.2.3 Process Overview and Scheduling

We explore four different hypotheses about link formation which we call “agent strategies”. Each strategy involves matching agents using their attributes but in different ways. These four strategies have been formulated so that they capture personal,

environmental and social aspects of student life. Personal preference, which shall be explained in the next section, (Section 4.3) is not taken into account when there are missing values for all the four attributes. Hence, in this case, we totally neglect the preference of both the source and the target agent. All of the four interaction strategies use the attribute values defined in Section 4.6. These strategies represent four hypotheses as to how individuals choose to link to other individuals.

4.2.3.1 Strategy 1 – Preferential Strategy

For this strategy, all agents have a predefined preference, which we term 'Personal Preference', for each of the four attributes described above. The idea is to implement a form of [homophily](#) – the love of the similar (McPherson, Smith-Lovin, & Cook, 2001), in the model. It is a probabilistic match based on the similarity of attributes between the source and the target agents. We illustrated the process in Table 4-3 for the dormitory attribute. First, a value between 0 and 100 is randomly selected in a uniform fashion – line 4. If it is under the predefined preference value (90 in case of dormitory preference for the Caltech University's reference dataset) and the attribute values of both the source and the target agents are known and match with each other, then the dormitory preference is satisfied; and we set the dormitory flag to true – line 6. Also, if the chance is greater than the preference value, it is satisfied as well – line 8. We repeat the same process for the remaining three attributes. If all the four attributes' conditions are satisfied, we make a friendship link between the source and the target agents – line 13.

Table 4-3 - Algorithm to calculate "Personal Preference"

<ol style="list-style-type: none"> 1. Pick two agents - target and source 2. Get dormitory preference (DP) – a fixed value 3. Set boolean sameDorm to False 4. Set chance a random value (random integer from 0 to 100) 5. IF (chance < DP) then: // 0<chance<=DP 6. IF (Dorm of Source = Dorm of Target) AND 7. (Dorm of Source != 0 AND Dorm of Target != 0)) then: 8. Set the Boolean sameDorm to True 9. ELSE then: 10. Set the Boolean sameDorm to True 11. ... 12. //repeat the same evaluation for the rest of the attributes (Major, Year, and High School) 13. IF (sameDorm AND sameMajor AND sameYear AND sameHighSchool are all TRUE) then: 14. // If all conditions satisfy, then: create a friendship link between the two 15. Form a link between the Source and the Target Agents
--

Each source agent selects a randomly chosen target agent after every simulation tick. The target agent is selected using a uniform probability distribution. After the selection, the source agent determines if the target agent satisfies its personal preference. If it does, an undirected link is created among them, which shows that they are friends.

4.2.3.2 Strategy 2 – Friend of a Friend (FOAF) Strategy

In this strategy, there are two phases for each agent. In the first phase, all agents make only limited random friends selected in a uniform distribution. This should satisfy both the source and target agents' preferences. If these are not satisfied, they do not form a link. In other words, the *preferential strategy* is employed by every agent during this initial period. After this first phase, personal preferences are not taken into account. From then on, in the second phase, new friends are selected in a 'friends-of-friends' manner. During this phase, starting from the first friend of a friend (FOAF) – whose degree is considered as the reference point, in chronological order, we search its friends and continue searching until we find a suitable agent. As soon as we find an available FOAF which has a greater degree than the reference FOAF – showing the popularity, we select it and then form a friendship link between the two. The rationale behind this strategy comes from a study carried out by Facebook itself (Lars Backstrom & Bakshy, 2011). This study, carried out on Facebook users in Iceland, found that 92% of all links created on Facebook have a path length of two, i.e., a triangle. This suggests that FOAF was mainly used to form friendship links.

4.2.3.3 Strategy 4 – Random Strategy

In this strategy personal preferences are not taken into account. All students arrange a small party which is held on a regular basis. The number of participants in a party is ten. The selection of the party participants is totally independent and unbiased towards any attribute. Also there is no bias in who develops a link with whom. At each party, a maximum of thirty new (random) friendships are made. Due to the random selection of party participants, there is a chance of selecting nodes which are already connected to each other. In that case, no new link is established.

4.2.3.4 Strategy 4 – Hybrid Strategy

This strategy is a combination of the above three strategies. At every simulation time step, a simulation strategy between Preferential and FOAF is chosen on a uniform basis. In order to preserve an element of randomness, the random strategy is run in every 20th time step. We have identified this particular value for the random strategy by running our model against the three reference datasets. This is almost similar to using a random strategy, where thirty new friendship links are developed at once, but allowing this to happen at every 20th tick. In other words, if we allowed one random friendship at roughly every 3rd tick, we would have achieved, more or less similar results. For an explanation of our model using ODD protocol, we will be using Caltech University's reference dataset as an example.

4.3 Design Concepts

4.3.1 Basic Principles

In this section, we cover why we selected the four interaction strategies covered in the previous section and whether there is evidence in the literature to support their choice. Before we discuss the evidence for each strategy, let us first focus on the structure of a social network. If we look at the structure of a social network, we learn a few things. For instance, after studying several social networks, the following unifying structural properties

of social networks were evident: [homophily](#), clustering (friend of a friend), the small-world effect, heterogeneous distributions of friends, and community structure (Mark Newman, 2010a; Ugander et al., 2011). Hamill et al. in their study (L Hamill & Gilbert, 2009), have summarised, more or less, the same characteristics. The one thing which is missing in the previous list is assortativity by degree of connectivity, which means the similar the number of links, the higher the chances of getting connected. It is a sort of degree [homophily](#). In terms of characteristics like age and nationality, Facebook, for instance, found a very strong [homophily](#) between its users (Ugander et al., 2011). In Facebook overall, 84.2% of friendship links are within the same countries, which also shows that the community structure of Facebook is mostly around local geographical areas (Ugander et al., 2011).

Thus, it can be observed that we have made use of the general structure of social network when designing each of the interaction strategies, described in the previous section. In order to develop friendship links, in some cases we focussed on agent's attributes, whereas in other cases, we focussed on cluster coefficient based on links. For the preferential strategy, the [affinity](#) measures for the three underlying datasets (see Table 3-1, Table 3-5 and Table 3-8) re-affirm the importance of [homophily](#). Thus, we focussed on similar attributes of users when forming new friendship links, in this strategy. The FOAF strategy is a mixture of [homophily](#) of both attributes and node's degree. In the first phase, when FOAF operates on the same principles as the Preferential Strategy, we take attribute [homophily](#). Once this phase is over, FOAF operates on cluster coefficient and degree assortativity. Random strategy has been inspired from random networks. If neither the position of nodes in a network nor the attributes when forming friendship links mattered, would we be able to generate a similar graph to reference graph? For the hybrid strategy, we combined all of these strategies into one, which provided agents a flexible method to develop their friendship links.

4.3.2 Emergence

An agent is designed to develop its social network by interacting with other agents and then selecting the right target as its friend. This interaction results in the emergence of a social network where they are closely linked groups of agents, or communities in other words.

4.3.3 Adaptation

There is not any individual learning process designed in the model.

4.3.4 Objectives

Since there are four modes of interaction, there are four different objectives for agents, depending on the mode of interaction being used.

1. **Preferential Mode:** On interaction, the preferences of both the source and the target have to match, in order to form a link.
2. **FOAF mode:** In the first phase, when personal preferences are used, the target agent is randomly determined, then the preferences of both of them have to be satisfied. However, in the second phase, if the source has met the target agent via

a friend (target is a FOAF), then no fitness is required to develop a link between the two.

3. **Random Mode:** No fitness is required here - only within-party potential links are selected as possible. The links are made randomly among the party participants.
4. **Hybrid Mode:** Depending on the mode being run (preferential, FOAF, and random), the appropriate fitness mechanism applies and is then satisfied.

The overall goal of agents in all the interaction modes is to find suitable candidates with whom to create links.

4.3.5 Learning

There are no individual learning mechanisms in place.

4.3.6 Prediction

Agents do not have any predicting power. They make their decision based on their interactions with others – on the available information.

4.3.7 Sensing

Based on the employed interaction strategy, the agents sense each other and their attributes to evaluate their compatibility.

4.3.8 Interaction

In this section, we discuss how the agents might interact with each other in terms of making friends in real life. It is assumed that, by and large, these real life social links will then be duplicated within Facebook. It is worth noting that there is no interaction in the model other than link formation and the assessment of other agents. We do not claim that we present an exhaustive list of possible strategies; rather the idea is to explore *some* plausible ways that depend on the micro-level preference of agents and then evaluate them.

There are four modes, so there are four interaction strategies as well.

- **Preferential Mode:** Every agent comes across a random agent.
- **FOAF mode:** Depending on the number of links an agent has made. If it is lower than 30, then she will meet a random agent; otherwise she will interact with a FOAF.
- **Random Mode:** In this interaction strategy, agents interact with party attendees.
- **Hybrid Mode:** Depending on the mode being run (preferential, FOAF, and random), the appropriate interaction strategy applies here.

All the seemingly ad-hoc numbers in each of the interaction modes have been well-thought out, tried and tested. In the case of a random mode, the number of participants (30) really does not matter, as the whole mode could be termed as random. There is no [homophily](#) bias in this mode. The simulation is executed until the number of links is equal

to that in the underlying dataset. As for the FOAF mode, the maximum link of 30 has also been tried and tested. A larger or smaller value results in a bigger or smaller clustering coefficient than in the reference dataset. For the hybrid mode (mode 4), if the number of parties (mode 3) is set to minimal, it corresponds to the random mode (mode 2); which in itself is a hybrid mode of random and FOAF mode. In order to have a mixed mode, the parties (mode 3) had to be introduced to have a totally random influence in agents' interactions.

4.3.9 Stochasticity

There is a uniform randomness involved which allows any agent to meet any other (depending on the process of the strategy). And also due to a random seed, the order of interaction between agents is completely arbitrary – no ordering is defined. For instance, it is not defined that agent 1 is going to meet with agent 2 first, or with agent 3.

4.3.10 Collectives

There are no defined groupings or collectives. In an emerging fashion agents form communities but they are not explicitly assigned to a community.

4.3.11 Observations

The mean and standard deviation in the number of friends is calculated at each time step. Cluster coefficient which calculates the number of triangles in a network, is also calculated during each step of the simulation. There are numerous post-simulation measures (and statistics) we use, which will be covered in more details in Section 4.6.

4.4 Details

4.4.1 Initialisation

The number of agents in all simulation runs is dependent on the reference dataset we are using. For instance, in the case of Caltech University, the number of agents is 769. Each individual in the dataset provides the attributes for one agent in the simulation. All agents are created at the start. While initialising a simulation run, the agents are chosen in a random order. Interaction strategies for all the agents are set once in the beginning. It does not change. Each simulation runs until the number of links made is the same as in the reference dataset. No link is dropped or modified once it is created. This is due to the fact that Facebook links tend not to be dropped once made.

4.4.2 Input data

Following initialisation, environmental conditions remain constant over the course of the simulation run of the model. The pre-simulation calculated preferences for each of the interactions are hard-coded into the model.

4.4.3 Sub-Models

In order to identify the significance of attributes and their values used in the personal preference algorithm for the four attributes (dorm, year, major and high school), we relied on the [affinity](#) (Alan Mislove et al., 2010) measured from the datasets. This was used to initialise parameters in the model that affect the personal preference algorithm. Using the Caltech University’s reference dataset, here are the [affinity](#) measures of the four attributes in Table 4-4:

Table 4-4. Affinity values, for Caltech dataset, of the four attributes for all the strategies of interactions

<i>Dorm Affinity</i>	<i>Major Affinity</i>	<i>Year Affinity</i>	<i>High School Affinity</i>
3.34	1.48	2.45	0.36

We see that dormitory is the most important attribute here. This result matches previously published work (Traud et al., 2008). Hence, we used this attribute as a guide to have a parameter sweep of just the dorm value. In our initial work, we assigned random parameter values for the four attributes but that did not result in a good fit. Apart from the high school attribute, the rest is positively correlated. Each agent is initialised with the four attributes (major, dorm etc.) of a corresponding individual in the reference dataset (Caltech in this case). We have used these four attributes because of the conformity with the earlier studies done on students. Also, using the [affinity](#) measure, we found them to be highly significant. The values for each of the four attributes can be seen in Table 4-5. These values have been found to be the best fitted values when compared to the reference dataset (of Caltech University).

Table 4-5 - Values of the four attributes for all the strategies of interactions (for Caltech University’s dataset)

<i>Dorm Preference</i>	<i>Major Preference</i>	<i>Year Preference</i>	<i>High School Preference</i>
90	30	20	10

4.4.4 Calibration and Validation

In terms of calibration and validation we relied on various inputs and outputs. In this section, we describe in general terms how we carried out both calibration and validation processes. In the results section (see the Section 4.6), we have provided explicit examples for both calibration and validation processes for all of the three reference datasets (Caltech, Princeton and Georgetown). For calibration we initialised the overall preferences for the four variables: dorm, major, year and high school. Their values are calibrated according to the [affinity](#) measure. Also we calibrated the degree to switch from the preferential to the FOAF mode. This applies to both FOAF and Hybrid modes of interaction. These variables are displayed in Table 4-6.

Table 4-6 Calibration Inputs

Variable name	Brief description
Preference variables	Based on the affinity measure, we assign preferences for all the four variables: dorm, major, year and high school preferences.
FOAF specific variable	At what point should FOAF mode switch operating from the random mode to the friend-of-friends mode. This value is set specific to the studied reference dataset.

The overall process is as follows: calculate the affinities for the four variables (dorm, major, year and high school), which could range from 0 to infinity, and then convert them into a scale from 0 to 100. These then represent preferences for all the four variables. This step is done just once. Once the highest affinity measure is identified (dorm for Caltech and year for Princeton and Georgetown), create a parameter sweep only for it. For instance, in Caltech, dorm has the highest affinity with 3.34. Create a parameter sweep with dorm preference from 60 to 90, with 10 as an interval. Now create a parameter sweep for the node degree when preferential strategy switches to friend-of-friends in FOAF/random mode. This has a starting point, an ending point and an interval (for example, from 30 to 50 degrees, with an interval of 5). The best run was identified by matching following variables of the simulated network with the reference network: standard deviation in degree, degree [homophily](#) (assortativity) and overall cluster coefficient. The parameter configuration which had the least difference for these three variables were then selected.

As for the validation, we relied on a host of measures. In Table 4-7 we have summarised them. For the overall results we compared results of four of our models against the reference dataset. This includes looking at the: standard deviation in degree, global cluster coefficient, degree assortativity and the best fitted distribution. Once all these measures are calculated for the simulated dataset, we compare the same measures for the reference dataset. Once we shortlist one of the four modes of interaction (which fitted the best), we then analyse it in more details for attribute specific ratios using Silo Index. This involves calculating Silo Index for the four variables: dorm, major, year and high school, and then calculating the correlation between the best suited mode of interaction and the reference dataset. We have used a multi-dimensional fitting of many patterns (e.g. graphs or summary measures), which is a better way of validating for complex models than a single close fit to one data source (Grimm et al., 2005).

Table 4-7 Validation Variables

Variable name	Brief description
SD in Degree	Determines the standard deviation in number or links
Degree Assortativity	Homophily of the connected nodes
Cluster Coefficient	Global cluster coefficient of the overall network
Best Fitted Distribution	With the help of Least Square Error (LSE) method we identify the best fitted distribution.

Parameter Values for the distribution	By using Maximum Likelihood Estimation (MLE), we identified the parameter values of the best fitted distribution.
Correlation of Silo Index	For all the four attributes: dorm, major, year and high school, we calculated Silo Index for the reference and the best fitted networks, and then calculate the correlation between each pair of Silo Index

4.5 Limitations

Although we have developed this as an explanatory model, there are some limitations to our model and also in our approach. For our model, the foremost important limitation is that the entire population of agents is introduced in the beginning of the simulation – no change in population during the simulation is introduced. In actuality, however, this does not happen. Not everyone becomes part of an [SNS](#), or any other system, for that matter, at the same time. There are some early adopters and some are late; it is an evolutionary process. We did this because we lack the information on the chronological evolution of Facebook memberships, who joined when? If and when such data becomes available, we can address this. Also once a link is established between two agents, it is never changed. The rationale behind this condition is that people hardly ever delete their old friendship links. One of the distinctive features of [SNSs](#) is that it helps you connect with your old friends (say from your old neighbourhood or your friends from kindergarten) with whom you may not have anything in common anymore. Once you become 'friends' with them, it is a social norm not to delete them from you friends' list – even if you don't communicate with them at all. A further limitation in our [ABM](#) is that there is not much dynamicity in that agents cannot decide on their own mode of interaction. In other words, agents have limited control over the interaction strategies. We compare populations all of which have the same strategies to see which best causes the observed structure. We can then conclude that this kind of strategy probably predominates in the population. In reality there will be a mix of strategies in the population and, indeed, the same agent might use different strategies in different circumstances.

There is a general problem of defining the environment and/or boundary for externalities in agent-based models. We have also not accounted for it in our model. Finally, there is no learning/adaptive process defined in our model, during the course of a simulation, the preferences do not change. Before the simulation, statistics such as [affinity](#) have to be calculated in order to have an idea about an individual preference. This will be the subject of future research.

4.6 Results

In this Section, we compare the simulation results for each of the reference datasets (of Caltech, Princeton and Georgetown universities), against our model.

4.6.1 Caltech Results

In this section, we describe the results of our model using the Caltech University's dataset. First we compare the global (overall) results in Section 4.6.1.1 and then in Section

4.6.1.2 we discuss the attribute level comparison. To conclude, we summarise our findings in Section 4.6.1.3.

4.6.1.1 Global Results

In this section, we compare the structure and the community detection mechanism based on the overall network of the reference dataset with the various simulation strategies.

For the selection of the values for each attribute, we relied on statistical measures, which were correlations in this case. This is part of the calibration process. According to them, the parameter Dorm Preference (DP) plays a significant role in link development. Hence we aimed to understand the impact of varying this parameter on the network structure as well as on attribute based communities. We explored the parameter space for Dorm Preference, starting from 60% to 90% preference for the same dorm.

Table 4-8 Modularity of Preferential and FOAF strategies with varying Dorm Preference (DP)

<i>Reference</i>	<i>Modularity - 0.3</i>	
Dorm Preference	Preferential	FOAF
90	0.32	0.32
80	0.16	0.16
70	0.11	0.12
60	0.11	0.11

To identify the density of the links of the whole network, we use *community modularity* (Newman, M. E. J. and Girvan, 2004). It is defined as the fraction of the links that fall within the groups (or communities) minus the expected such fraction if links were randomly distributed. It ranges from 0 to 1, where 0 means there are no links within the identified community, and 1 means all links are within a community. As can be seen in Table 4-9 the closest modularity with the reference dataset is found when the Dorm Preference is set to 90. To calculate the community modularity, we have used the method described by Clauset et al. (Clauset, Newman, & Moore, 2004). So when the Dorm Preference is set high, the modularity correspondingly also becomes high. Also, in the FOAF strategy, just like the preferential strategy, the initial random network development which is based on both the source's and the target's preference, acts as a strong characteristic of a high modularity network.

Table 4-9 Fitted centrality degree distribution with varying Dorm Preference (DP)

Reference Dataset	Normal Distribution - Mean = 0.0282 and Variance = 0.0241			
Dorm Preference (DP)	Preferential Normal Distribution Parameter Values		FOAF Normal Distribution Parameter Values	
	Mean	Variance	Mean	Variance
90	0.028	0.0076	0.028	0.022
80	0.028	0.0055	0.028	0.022
70	0.028	0.0044	0.028	0.025
60	0.028	0.0039	0.028	0.024

In Table 4-9 we summarise the effects of the underlying distribution for the varying Dorm Preference of both preferential and random strategies. In order to identify the underlying degree distribution of the simulated networks, we used the method of Least Square Error (LSE) – the lower the value, the better the fit. Once the underlying degree distributed is identified, we used the method of Maximum Likelihood Estimation (MLE) to identify the parameter values for the distribution. Although the underlying distribution of the reference dataset and the FOAF strategy with DP being 90 was a Beta Distribution, when Least Squared Method (LSM) was applied to them, but with a very minor difference, the Normal Distribution was also a good fit. Since most of the simulation results of both the strategies reveal that they are normal in nature, we considered Normal Distribution as the best fitted distribution.

There is a major difference between the two strategies. In the case of preferential strategy, the variance decreases as the DP is decreased, while the FOAF strategy shows almost similar behaviour for all DP values. It can be said that there is a very low impact on network structures of initial friendships in the FOAF strategy which are based on personal preferences. We are focused on both community and network structures, hence we selected DP to be 90, as it is a better candidate for network modularity. Thus, for all the following results, the DP value is 90.

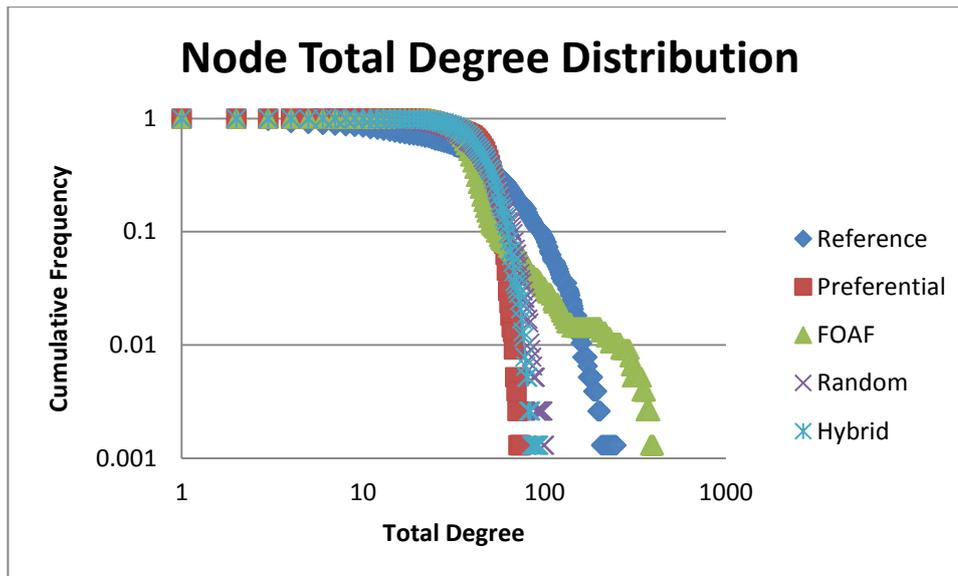


Figure 4-1 Log-log plot of Total Degree Distribution of all the four simulation strategies and the reference dataset

We have summarised in Figure 4-1 the node degree distribution of the Caltech dataset and the four interaction strategies. This only shows the final node degrees at the end of the simulation. The reference and the FOAF strategy's degree distributions show a power-law effect which suggests that most of the nodes have few links while only a few nodes have a higher number of links. The other three strategies, preferential, random and hybrid seem normally distributed in nature. Their links are more or less uniformly distributed.

If we consider various studies on the number of friends in Facebook (see (Alan Mislove et al., 2007; Panzarasa, Opsahl, & Carley, 2009; C. Wilson, Boe, Sala, Puttaswamy, & Zhao, 2009), almost all have found that it does have a power-law outlook, but there was a seminal work which proved this common belief wrong. According to Minas et al. (Minas Gjoka, Kurant, & Butts, 2009), Facebook has not one but two power-law regimes: one for node degrees less than 300 and one for greater degrees. We also found a similar pattern as described in Abbas's work (Abbas, 2011b). In this case, however, we do not see that for two reasons. Firstly, the dataset is too small and secondly the dataset just contains inter-school links. Hence we see just one power-law outlook.

We have concentrated on a few, but important factors of Social Network Analysis (SNA) in order to compare the reference datasets with the simulated network. The factors with their respective values can be seen in Table 4-10.

Table 4-10 - Reference Dataset (of Caltech) and Simulation Output Comparison

<i>Model Type</i>	<i>Avg. Distance</i>	<i>Connectedness</i>	<i>Cluster Coefficient</i>	<i>SD. of # of friends</i>	<i>Community Modularity</i>
Reference	2.47	0.98	0.23	37.03	0.3
Preferential	2.49	1	0.21	11.52	0.32
FOAF	2.61	1	0.22	35.05	0.32
Random	2.39	1	0.07	15.55	0.11
Hybrid	2.49	1	0.09	13.71	0.12

In Table 4-10 we can clearly identify that the FOAF strategy remains the best candidate when it is compared with the reference dataset. Although the reference dataset is not a fully connected network, the average distance, the standard deviation of number of friends, total cluster coefficient and even the overall modularity, is quite similar to the reference social network. The underlying distribution of both the reference and the FOAF strategy can be distinguished by their large standard deviation, which indicates there is a wide variation in node degree. This means that low and high node degrees exist, which is a typical characteristic of a social network.

4.6.1.2 Attribute Level Results

In this section, we compare the results of our simulation for each of the attributes with the reference dataset. We measured the results in terms of the Silo Index. Since the FOAF strategy has shown the best results, we are presenting Silo Indices comparisons of this strategy with the reference dataset.

Reference vs. FOAF Mode Silo Indices

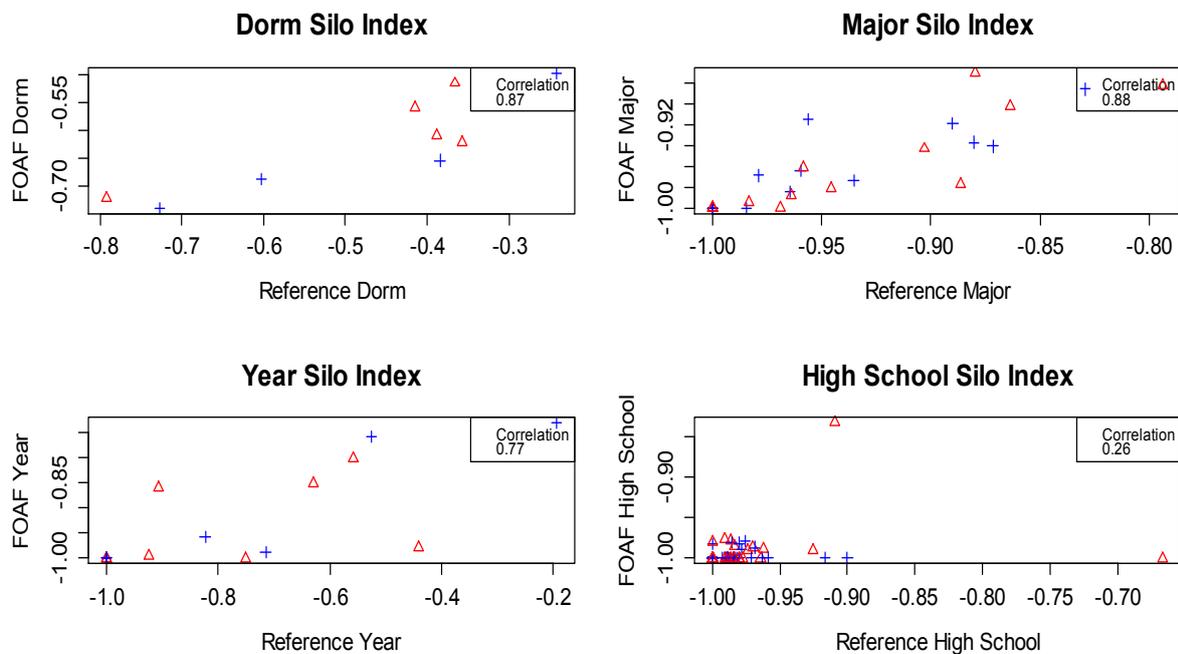


Figure 4-2 Silo Index for Dorm, Major, Year and High School attributes for FOAF strategy and the Caltech reference network (the blue triangles represent the reference dataset (Caltech), and the red plus signs represent the FOAF strategy)

In Figure 4-2 we calculate the Silo Index of the dorm, major, year and high school Silo Indices. The triangles represent the reference dataset (Caltech), and the plus signs represent the FOAF strategy. Apart from the high school Silo Index, the others have quite a high (≥ 0.77) correlation with the reference dataset. The number of High Schools in the dataset is quite high as shown earlier in Table 3-4. Overall, in all the four figures for Silo Indices, we find clusters. Most of the dorm values cluster around 0.4. Major has two distinct communities, one around -1 to -0.95, and the other one around -0.9 and -0.85. For

year, we see one group at -0.8. Most of the high school has a Silo Index closer to -1. This is found in the hybrid strategy as well.

4.6.1.3 Summary

In this section, we summarise our findings. We discover that in the hybrid strategy, the randomness of the random strategy has a major influence on it so we did not see much difference between these two strategies, be it general or attribute level comparison. By changing the occurrences of the random mode (by reducing and/or increasing parties), we tried to control random strategy selection in the Hybrid strategy, but the randomness of the preferential strategy also did not quite help to improve its fit to the reference dataset in the social network development. Although the attribute level communities produce comparable results to the reference network of the dataset, the totally random selection of target nodes in preferential strategy resulted in a low overall cluster coefficient and a low standard deviation in number of friends.

After analysing the results and comparing them with the reference dataset, we determined that the FOAF strategy (which initially takes local preferences into account but then works on a friend-of-a-friend basis) does the best. It captures the basic essence of the underlying network - from network level measures to the attribute level comparison, it presents itself as a good candidate for the understanding of students' interactions and social network development. The results of our FOAF strategy are also in line with a recent empirical study by Facebook itself (L. Backstrom & Leskovec, 2011) on its users in Iceland, showing that 92% of all links created on Facebook have a path length of two, i.e., a triangle. The initial setting of highly similar friends leads to a cohesive community structure and also the friends-of-a-friend process with a power-law outlook. The Random and Hybrid strategies, which are dominated by the random meeting of friends at events, did not explain the data well, showing a lower level of friendship triangles (cluster coefficient).

4.6.2 Princeton Results

In this Section, we compare the simulation results with the reference dataset from Princeton University. Firstly we compare the global or overall results in Section 4.6.2.1 and then, in Section 4.6.2.2 we discuss the attribute level comparison. In Section 4.6.2.3 we conclude our findings.

4.6.2.1 Global Results

In this Section, we compare the structure based on the overall network of the Princeton dataset with the various simulation strategies. In Table 4-11 we have summarised the basic Social Network Analysis ([SNA](#)) measures, over the reference dataset and the simulation results of the four interaction strategies.

Table 4-11 - Reference Dataset (of Princeton) and Simulation Output Comparison

Dataset/Model	St. Dev. Degree	Assortativity	Transitivity	Best Fitted Distribution
Ref.	78.55	0.09	0.16	Exponential (Alpha = 1.98) ³
Preferential	18.12	0.08	0.019	Normal
FOAF	93.76	0.11	0.09	Exponential (Alpha = 1.84) ³
Random	19.64	-0.002	0.03	Normal
Hybrid	79.97	0.105	0.07	Exponential (Alpha = 1.97) ³

Preferential and Random modes deviate most from the reference dataset. In terms of standard deviation in number of degrees (# of friends), they are not even close. Also the distribution of degree is in these cases *normal*, bell shaped, as opposed to exponential distribution. The FOAF mode has good results in terms of assortativity, transitivity and also has the best fitted distribution, however, it markedly differs from the reference in terms of the SD of the degree distribution, meaning that it has a wider variation in node degree distribution when compared with the reference dataset's (93.76 vs 78.55).

Hybrid mode captures the standard deviation (assortativity and connectedness and also the best fitted distribution) quite well, when compared with the reference dataset. However, in terms of transitivity, it is almost half of the reference dataset. In order to align it with the reference dataset, we ran a sensitivity analysis over the parameter space. We found better results when the parameters were changed, but that hampered the standard deviation and assortativity. Hence we focused more on the overall degree fitting and assortativity. The parameter values for the reference dataset, FOAF and Hybrid mode are also mentioned, where Hybrid mode has almost the same alpha value (the slope of the log-log plot) as the reference dataset, for the fitted distribution. The fitting of the degrees have been calculated by setting the minimum degree to 40 (which was identified by running the calibration process mentioned in the Section 0).

³ Statistically significant ($p > 0.05$), using Kolmogorov-Smirnov test

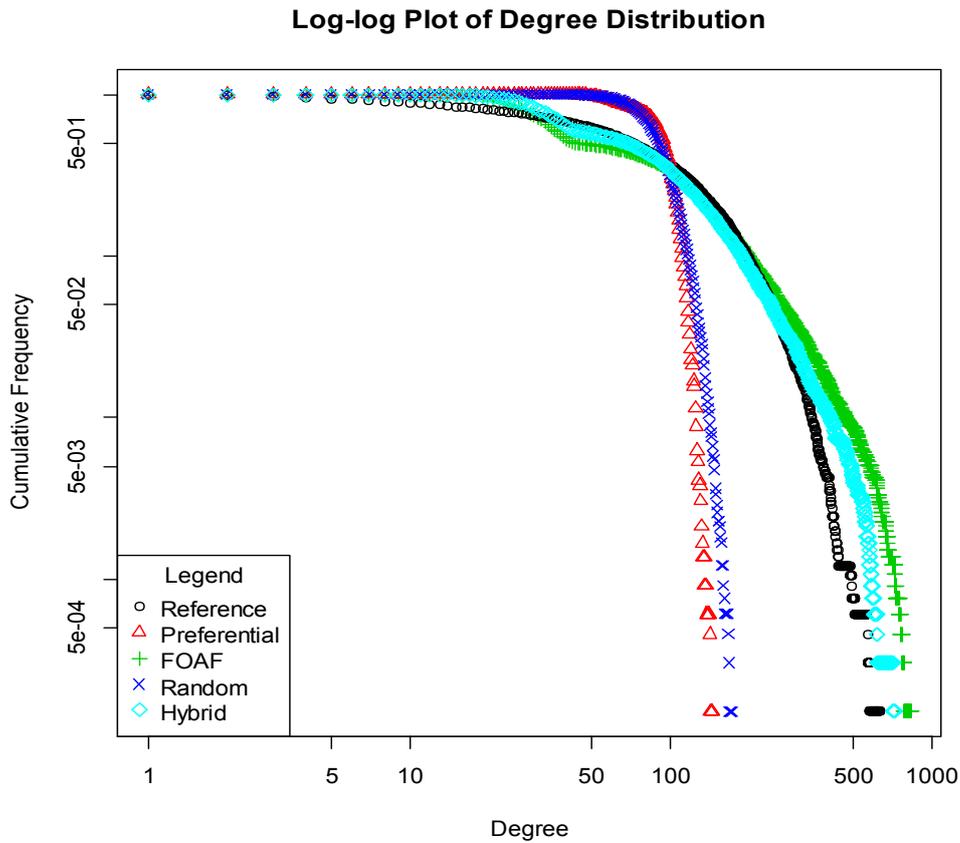


Figure 4-3 Log-log plot of Total Degree Distribution of all the four simulation strategies and the Princeton reference dataset.

We have summarised in Figure 4-3 the degree distribution of the reference and the four interaction strategies. This only shows the final node degrees at the end of the simulation. The reference, the FOAF and the Hybrid strategy's degree distributions show a power-law effect (with $p > 0.05$) which suggests that most of the nodes have few links while only a few nodes have a high number of links. The other two strategies, preferential and random, seem normally distributed in nature. Their links are more or less uniformly distributed, which is atypical for a social network.

In Table 4-11 we can clearly identify that the Hybrid strategy is the best candidate in comparison with the reference dataset. The underlying distribution of both the reference and Hybrid strategy can be identified by a huge standard deviation; which in turn reflects our earlier finding that both of these are in fact power-law distribution and roughly matches that of the reference dataset.

4.6.2.2 Attribute Level Results

In this Section, we compare our simulation results for each of the attributes with the reference dataset. We measured the results in terms of the Silo Index. Since the Hybrid strategy has shown the best results, we are presenting Silo Indices comparisons of this strategy with that of the reference dataset, in Figure 4-4.

Reference vs. Hybrid Mode Silo Indices

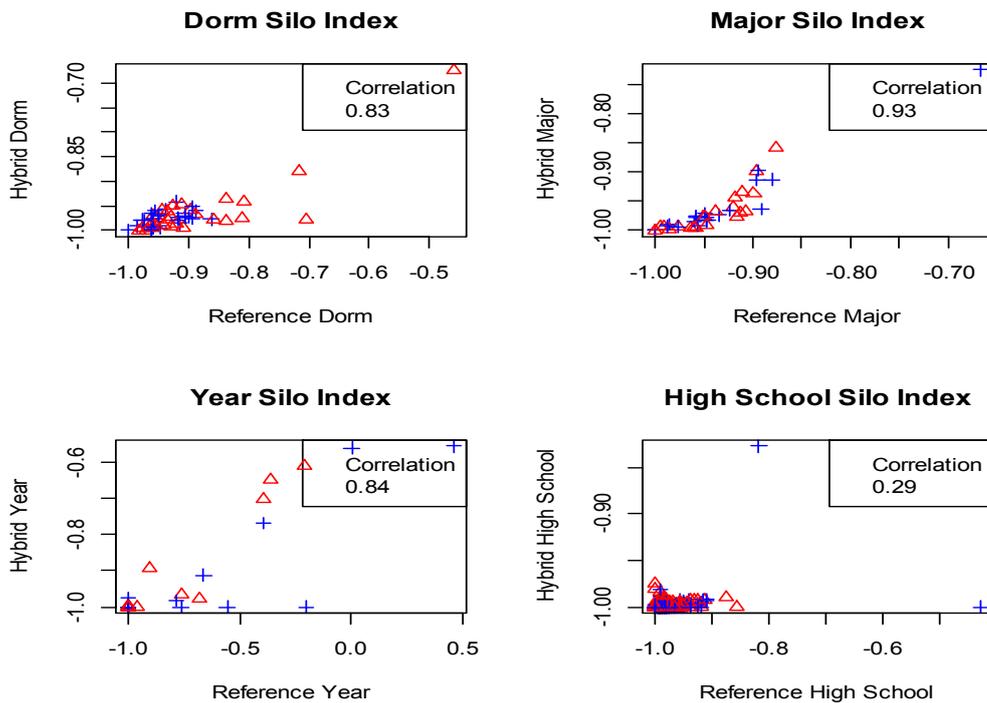


Figure 4-4. Silo Index for Dorm, Major, Year and High School attributes for Hybrid strategy and the Princeton reference network (the blue triangles represent the reference dataset (Princeton), and the red plus signs represent the FOAF strategy)

Apart from the high school Silo Index, the rest has quite a high (> 0.83) correlation with the reference dataset, meaning it fits the observed data quite well. Hence it is the preferred mode of interaction. Similar to the Caltech dataset, the triangles represent the reference dataset (Princeton) and the plus signs represent the underlying strategy (Hybrid in this case). The number of high schools in the dataset is quite high (2235) as shown earlier in Table 3-4, which makes it difficult to reduce the difference between the reference and the FOAF strategy, resulting in a lower correlation. For specifics, let us turn to the individual attribute based Silo Index. For dorm, we find most of the Silo Indices around -0.9, for both reference and the Hybrid strategy. In terms of major, the spread is wider for the majority (from -1 to -0.9). We have plotted degree mixing (MEJ Newman, 2003), which determines how nodes of similar degrees are connected with each other capturing degree homophily, in Figure 4-5.

Degree Mixing - Reference vs. Hybrid Mode

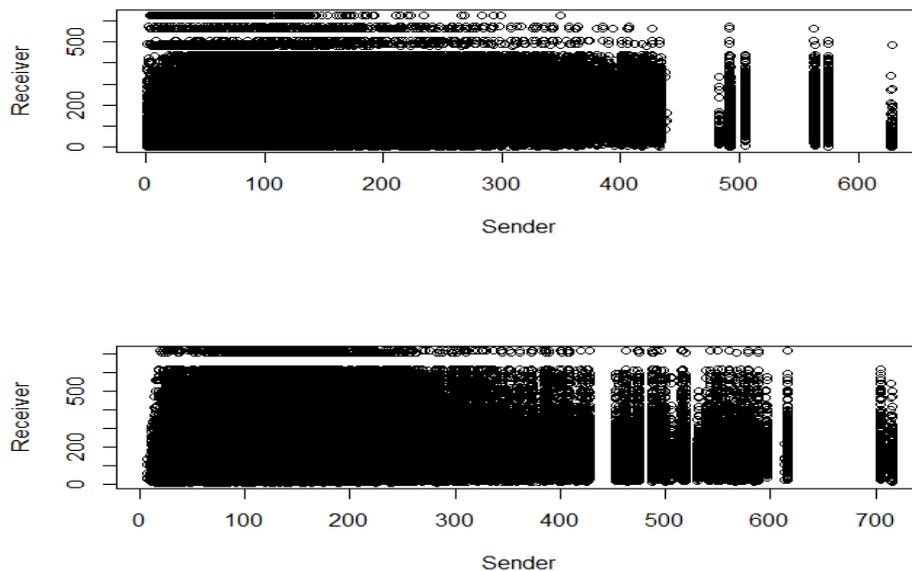


Figure 4-5. Degree Mixing of Hybrid mode (bottom) and the Princeton reference dataset (top)

In lower degrees (< 400), there is a high rate of similarity between the Hybrid and the reference dataset. In high degrees, the Hybrid mode is slightly different.

After comparing all the four attributes, the Hybrid strategy takes the lead when compared with the reference dataset; it presents itself as a good candidate for describing how students might have developed their social network.

4.6.2.3 Summary

After analysing the results and comparing them with the reference dataset, we determine that the Hybrid strategy, which is a combination of all three strategies: Preferential, FOAF and Random performs the best. It captures the basic essence of the underlying network. From network level measures to the attribute level comparison, it presents itself as a good candidate for understanding students' interactions and social network development. Also, FOAF mode captured most of the aspects, apart from the standard deviation in number of friends, which resulted in a different slope for the power-law outlook. The initial setting of highly similar friends leads to a cohesive community structure and also the friends-of-a-friend process with a power-law outlook. Preferential and Random strategies which are dominated by the random meeting of friends at events did not explain the data well.

We do not claim that we presented an exhaustive list of possible social processes, but rather analysed a few plausible variations. Focusing on personal preference, social structure with some randomness, presents itself as a promising strategy of interaction. While only pre-simulation statistics based on the underlying data, such as correlation, do not necessarily present the best parameter values for the initial friendship links, the parameter space has to be explored to find the best match.

4.6.3 Georgetown Results

In this Section, we compare the simulation results with the reference dataset of Georgetown University. Firstly, we compare the global or overall results in Section 4.6.3.1 and then, in Section 4.6.3.2 we discuss the attribute level comparison. In Section 4.6.3.3 we will discuss the mode that best captures the social dynamics when compared with the underlying reference dataset.

4.6.3.1 Global results

In this section, we compare the structure based on the overall network of the reference dataset with the various simulation strategies. In Table 4-12 we have summarised the Social Network Analysis (SNA) measures, over the reference dataset and the simulation results of the four interaction strategies.

Table 4-12 - Reference Dataset (of Georgetown) and Simulation Output Comparison

Dataset/Model	St. Dev. Degree	Assortativity	Transitivity	Best Fitted Distribution
Ref.	79.42	0.075	0.14	Exponential (Alpha = 2.217)⁴
Preferential	28.86	0.2	0.028	Normal
FOAF	107.92	0.35	0.174	Exponential (Alpha = 2.13) ⁴
Random	23.63	-0.002	0.013	Normal
Hybrid	89.86	0.28	0.103	Exponential (Alpha = 2.218)⁴

Similar to the results for Princeton University (see Section 4.6.2), in these results we also find that Preferential and Random modes deviate most from the reference dataset. In terms of standard deviation in number of degrees (# of friends), they are not even close. Also the distribution of their degree is *normal*, bell shaped, as opposed to exponential (for the reference dataset). The FOAF mode has good results in terms of assortativity, transitivity and even the best fitted distribution. In standard deviation, however, the difference is quite large when compared with the Hybrid mode.

Hybrid mode captures quite well the standard deviation, assortativity and connectedness and is also the best fitted distribution when compared with the reference dataset. In order to align it with that of the reference dataset, we ran a sensitivity analysis over the parameter space. We did find better results when the parameters were changed, but that hampered the fit concerning standard deviation and assortativity. Hence we focused more on the overall degree fitting and assortativity. The parameter values for the reference dataset, FOAF and Hybrid mode are also mentioned, where the Hybrid mode has almost the same alpha value as the reference dataset (2.218 versus 2.217), for the fitted distribution. The identified alpha values for the exponential distributions have been calculated by setting the minimum degree to 60.

⁴ Statistically significant ($p > 0.05$), using Kolmogorov-Smirnov test

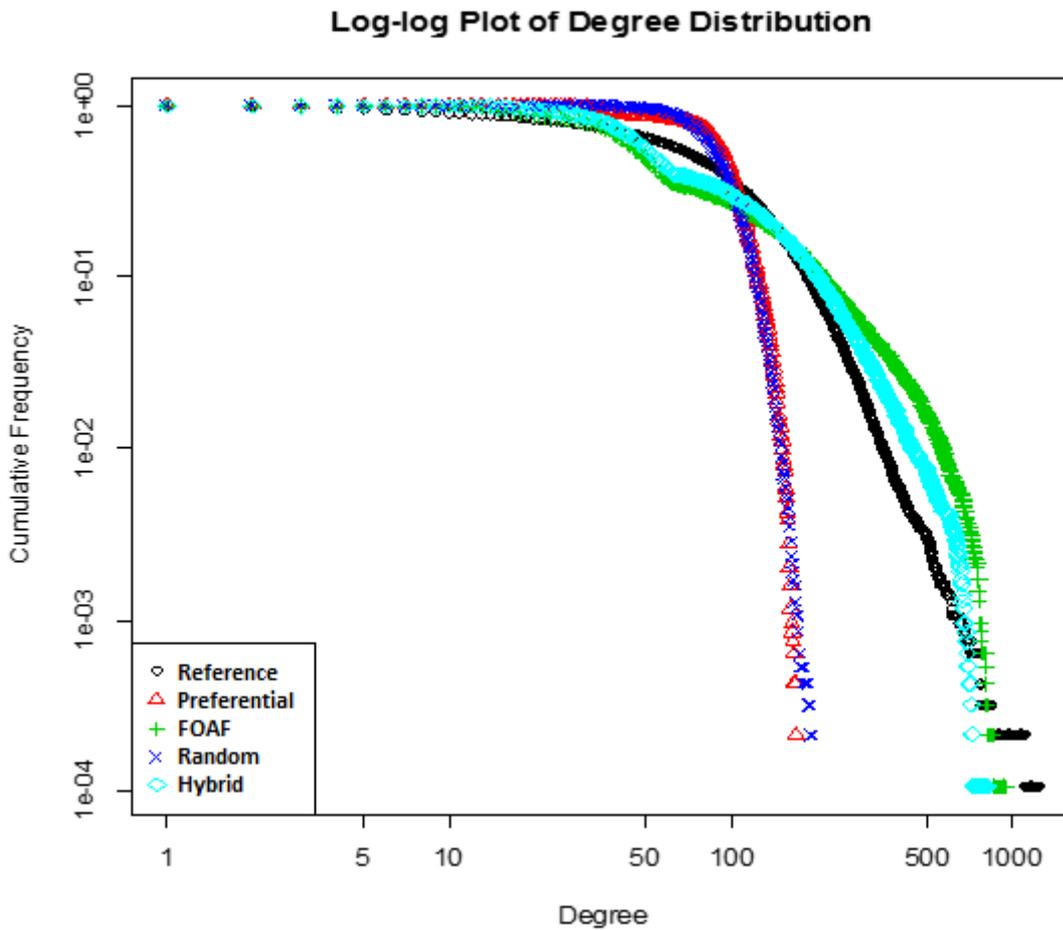


Figure 4-6 - Log-log plot of Total Degree Distribution of all the four simulation strategies and the Georgetown reference dataset.

Figure 4-6 shows the degree distribution of the reference and the four interaction strategies. This only shows the final node degrees at the end of the simulation. The reference, the FOAF and the Hybrid strategy's degree distributions show a power-law effect which suggests that most of the nodes have fewer links while only a few nodes have a lot of links. This is again confirmed by running Kolmogorov-Smirnov test where we found $p > 0.05$. The other two strategies, preferential and random seem normally distributed in nature. Their links are more or less uniformly distributed, unlike the reference dataset.

In Table 4-12 we can clearly identify that the Hybrid strategy remains the best candidate compared with the reference dataset. The underlying distribution of both the reference and the Hybrid strategy can be identified by a huge standard deviation; which in turn reflects our earlier finding that both of these are in fact power-law distribution, like the reference dataset (and social networks in general).

4.6.3.2 Attribute Level Results

Similar to the previous results of Princeton University, in the case of Georgetown University we also find the Hybrid strategy to be the best fit. Hence, in this section, we will focus on just the Hybrid strategy. We compare the results, which are based on the Silo

Index, against the reference dataset. In Figure 4-7 we have shown the comparison, along with the underlying correlation between the Hybrid and the reference dataset.

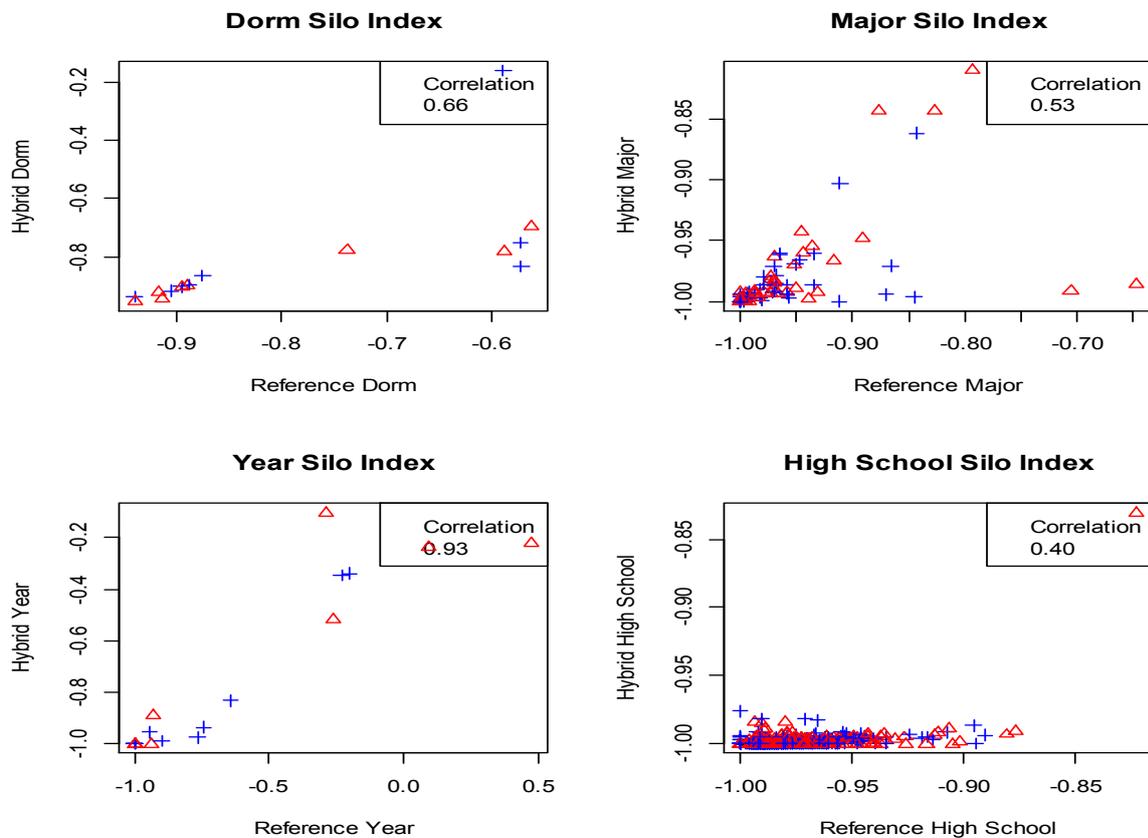


Figure 4-7. Silo Index for Dorm, Major, Year and High School attributes for Hybrid strategy and the Georgetown reference network (the blue triangles represent the reference dataset (Georgetown), and the red plus signs represent the FOAF strategy)

The Year attribute has the best correlation (0.93) – the dorm and the major attributes follow with 0.66 and 0.53. In the figure, the triangles represent an individual Silo Index of the reference dataset (Georgetown), and the plus signs represent that of the Hybrid strategy. In this reference dataset, the high number of high schools, 2874 in total, makes it really difficult to get a better correlation for the Silo Index. On average 3.27 students are in every high school. When comparing individual Silo Indices of each attribute, we find for the dorm Silo Index, two major clusters – one around -0.9 and the other one near -0.6. For major, however, most of the values lie at -0.975. The year attribute has two regimes as well – one near the -1 and the other one near -0.1. For high school, the values are between -0.9 and -1, however having so many unique values (2874), as explained earlier, the correlation does not match with those of other attributes (such as dorm).

We have plotted degree mixing (MEJ Newman, 2003), which determines how nodes of similar degrees are connected with each other, in Figure 4-8.

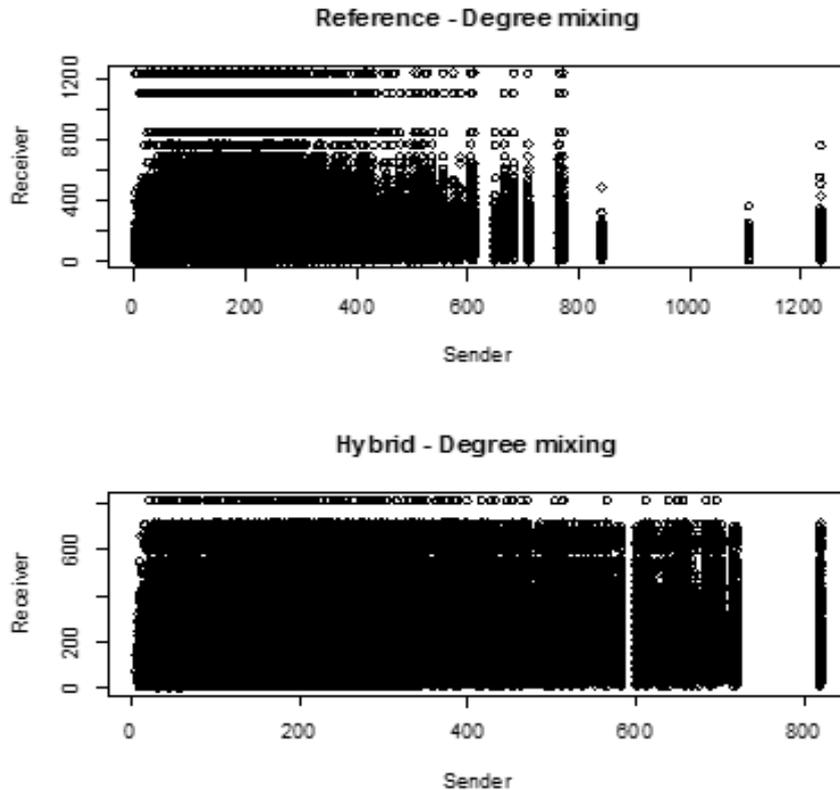


Figure 4-8. Degree Mixing of Hybrid mode (bottom) and the Georgetown reference dataset (top)

For lower degrees (< 600), there is a high rate of similarities between the Hybrid and the reference dataset. For the high degrees, the Hybrid mode is slightly different. For instance, the highest degree in the reference degree-mixing reaches up to 1200, whereas in the Hybrid mode it reaches 800. However, as shown in Table 4-12, the overall degree correlation (assortativity) of reference and hybrid modes, are fairly similar (0.14 versus 0.103).

After comparing all the four attributes, Hybrid strategy is the best fit to the reference dataset; it presents itself as a good candidate for describing how students might have developed their social network at Georgetown.

4.6.3.3 Summary

Georgetown, being a bigger University like Princeton (when compared with Caltech University), shows similar results to Princeton. In this case too, after analysing the results and comparing them with the reference dataset, we determine that the Hybrid strategy (which is a combination of all three strategies: Preferential, FOAF and Random) fits the reference data the best. It captures the basic essence of the underlying network. In terms of total distribution (see Table 4-12), we see that FOAF has quite a good match. Also, the FOAF mode captured most of the aspects, apart from the standard deviation in number of friends, which resulted in a different slope for power-law outlook. Preferential and Random strategies which are dominated by the random meeting of friends at events did not explain the data well.

4.7 Discussion

There are a couple of points which we learnt from our work. When we see the results produced by our [ABM](#), in all the three datasets, we come across following findings:

- All interaction strategies always produce a fully connected graph;
- Preferential and Random Strategy produce graphs with a normal distribution;
- The FOAF produces both normal and exponential graphs, depending on the size of the graph;
- The Hybrid strategy produces exponential graphs.

When dealing with social networks, we find both attribute [homophily](#) and degree [homophily](#) (assortativity), as was the case in two of the three reference graphs we have used, as typical characteristics. After knowing the reference graph's underlying degree distribution and also of graphs produced by all four strategies, we can already predict which strategy will compare the best to the reference graph. For a better comparison, we need to rely on getting useful measures from the reference graph.

In order to know input parameter values, such as values for personal preference for each attribute, we ran [affinity](#) measures on the reference graph. Also, in order to select the best candidate for the model, we ran a sensitivity analysis using three main measures, which are cluster coefficient (transitivity), standard deviation in total degree, and assortativity (degree [homophily](#)). This helped us find the best suitable candidate for each of the four interaction strategies. Although we have made sure that missing values are specially treated in our personal preference algorithm, it would be interesting to apply it on a dataset that does not contain any missing values, or at least one with lower levels of missing values. At the moment the percentage of missing values goes up to 33%. This also impacts the [affinity](#) measures which we have used to parameterise each of the four attribute values. To improve [SNA](#) measures, such as [affinity](#), provisions should be made for treating missing values differently. We have not, however, dealt with individual specific (or even group specific) attribute preferences. For instance, one group of agents might have a stronger dorm preference than others. If we know an individual's or even group level affinities, we might be able to improve our model to incorporate such subtleties. Ideally if we had a longitudinal or evolutionary dataset involving similar nodes over time, this would provide a greater deal of confidence to our methodology.

In the Hybrid mode, a better choice of random mode might be to change the number of individuals who develop friendship links. At the moment, preferential and FOAF strategies are executed randomly, and then at every 20th tick, the random strategy is run (to form at most 30 friendships). Another implementation, which would clarify the Hybrid mode, would be to run a randomly selected strategy in which only one new friendship is developed, instead of 30.

As we have seen in both Princeton and Georgetown results, the FOAF strategy produces fairly good results when compared with the reference datasets. In both cases however, the standard deviation in terms of degrees in particular, did not meet well, which was not the case for the Hybrid strategy. We believe that when FOAF mode is operating on a friend-of-friend mode, which is currently based on the popularity (higher degree) of a candidate, we might adjust it by using various other options. Searching for a popular candidate, we think, is causing a bigger divide between low and high degree nodes, which then results in a higher standard deviation in total degrees. The new mechanisms could

be, instead of popularity, a selection mechanisms based on random or on similar attributes (to some extent).

In all the generated graphs, we get a fully connected graph. If we would like to re-generate a disconnected graph from any of the four strategies, we cannot do so. Hence these models are only good for fully connected networks. In our case, there are a few disjointed nodes in the datasets (like in the Caltech and Georgetown datasets), but those were very few. When we look at the personal preference algorithm in the preferential strategy, we notice that it is grounded on the four attributes (major, high school, year and dorm) of students when it comes to finding a suitable student to develop a friendship link. There might be other important attributes but which are not present in the dataset. The main reason why we could not apply the findings from our ethnic analysis of Facebook users to our [ABM](#), was because of the lack of data on ethnicity (lack of either self-described or inferred ethnicity). Also there could be some hidden mechanisms which we did not capture. This work focusses on finding the best explanatory candidate for each of the three datasets. However, one should note that we are not proposing that these are the only plausible mechanisms which may result in the emergence of such networks.

4.8 Conclusions

An agent-based simulation has been described that supports explanation of how students make [SNS](#) links within three US universities, taking into account both endogenous and exogenous factors.

In this work we tried to understand how local preferences and structural factors might play a role in the development of a social network. We have devised and explored a limited number of strategies for student interaction. We compared our simulation results to reference datasets gathered from students' Facebook networks of Caltech, Princeton and Georgetown Universities. We relied on community detection methods and major [SNA](#) factors for comparison. The strategies of interaction varied from preferential attachment, based on the attribute values, to complete random interactions. In order to show that our model is flexible and generalisable, independent of the reference dataset, we have applied it to three different datasets (of Universities). We found the FOAF strategy, which focusses on personal preference and social structure, to be the best candidate for Caltech, and for Princeton and Georgetown we found the Hybrid strategy to be the suitable mode of interaction. It seems the interaction strategy depends upon the size of a university. For a smaller university, like Caltech, the FOAF mode is the most suitable but for a big and diverse university, like Princeton and Georgetown, there are a multitude of social and structural processes involved (captured by the Hybrid mode). These include: attribute based (same dormitory, major or high school etc.), social interaction, random meet ups (through parties or other social events) and current friends introducing new friends (Hybrid mode). Students meet and interact not only in their university, such as lecture halls, but also outside it, such as at parties and in the dormitories they live in. Also through current friends, students explore and develop further friendships from friends-of-friends we do not claim that we presented an exhaustive list of possible social processes, but rather analysed a few plausible variations. Focussing on only pre-simulation statistics based on the underlying data, such as correlation, do not necessarily present the best parameter values. For the initial friendship links, the parameter space has to be explored to find the best match. This shows that using several sensitivity analyses around macro-level measures such as transitivity and assortativity help determine the well suited parameter values.

5 Chapter: Diversity and Clustering of MMU Students on Facebook

We focus on a specialised service of Internet, where users are the central entities. This service is defined as a Social Network System ([SNS](#)). The users join such systems which provide one to one interactions among those who have similar interests. This involves articulating a virtual persona with the help of a 'profile', which becomes their online identity. It is essentially a web-page describing an individual by their attributes, such as age and gender, and also a list of interests, including hobbies, books and TV shows. Users then develop links with other users which usually reflect real life social links such as friends, family members and acquaintances, and also developing new links with those who share similar tastes and/or interests.

With the rise of [SNSs](#), a unique opportunity to study and understand social structure has arisen. Never before has it been possible to analyse society at such a huge scale with such details. An in-depth analysis of [SNSs](#) provides new insights into norms and cultures, and also becomes a vehicle for the future development of services on the Internet. Due to competitiveness and also privacy of users, providers such as Facebook, do not provide users' data and their social network to researchers (A Mislove, 2009). Our goal is to understand the structure of such a system, by focussing on the social network of users. One of the most important aims of this thesis is to identify how to capture complex social networks and then see how closely it is structured on ethnic lines. To achieve this, we have collected a large scale network of users from our University's ([MMU](#)'s) Facebook network. In this chapter, we will explain how we have estimated ethnicities of a subset of Facebook users with their friends and studied how such ethnicities are connected among and within each other. We collected our data, by accessing publicly available information by the crawling method. A large dataset of four thousand Manchester Metropolitan University ([MMU](#)) students has been collected. This dataset captures the diversity in terms of language, religion and geography of the Facebook users.

Since Facebook is projected as the representation of the real world, we would like to better understand how it is structured. Specifically, we would like to understand the complexity of a network in terms of ethnicity. Our hypotheses are:

H1 (a): The Facebook network is segregated on the ethnic lines;

H1 (b): The Facebook network is highly clustered on ethnic lines;

To test our hypotheses, we also have formed a null hypothesis, which is:

H0: The Facebook network does not segregate on ethnic lines and is not highly clustered on ethnic lines

These hypotheses, one should note, will only be tested against our collected data from the [MMU](#) University, and should not be considered generalisable findings for the whole Facebook social network.

A lot of research on [SNSs](#) has been carried out to identify the network structure and the hidden community outlook of an [SNS](#) (SA Catanese & Meo, 2011; Alan Mislove et al., 2007), but mainly these studies are carried out on social networks which are fully anonymised; they only contain nodes and their links without any individual level attributes like location or interests. Due to their focus on the structural attributes such as degree distribution and connectivity among nodes and also on privacy issues, personal information of individuals (like name, location, interests and profession etc.) are not collected. In our work, however, we were interested to determine how individuals are

connected with each other when their ethnicity and religious backgrounds are taken into account. In other words, we were interested in characteristics and also personal attributes of users.

[SNSs](#), like MySpace already suggests that users identify their race from a fixed list of predefined labels (like White, Black and Hispanic, etc.). In case someone does not have any of those ethnicities, they can select 'other', or simply leave the field empty. These limited options cannot be generalised for each country in the world, even in the broader sense. Facebook, for example, collects a lot of personal information like hometown, interests, email address, etc. There are no fields in Facebook where an individual can describe his/her ethnicity.

The reason for this, suggested by Ginger (Ginger, 2008) is, that Facebook, in two ways, serves to perpetuate inadvertently or covertly racist or discriminatory norms: the colour-blind mentality and the racialised visual classification of others. This is where our work tries to bridge the race/ethnicity self-identity needs of the users. Ideally, Facebook would provide a self-defining ethnicity/race feature to users. In our case, we relied on an inferred ethnicity. We collected a sizeable subset from Facebook and then applied a name-based ethnic classifier, Onomap (Lakha, Gorman, & Mateos, 2011), to the name information provided by the users. This tool helped us to estimate the ethnicity, language, religion and even geography of individuals in our dataset. For the whole dataset, we then analysed how various ethnicities, religious and language based groups are connected with each other. There is a dearth in the literature of research which studies such inter and intra ethnic linkages in an [SNS](#). Our work tries to bridge the gap. This allows us to test our hypotheses as to how users are inter-linked with each other on ethnic lines, and how closely structured they are in terms of clustering coefficient.

One of the principal contributions of our work is to analyse ethnic and racial classifications, linkages and preferences of Facebook users at a much greater level of detail. We start off by explaining how we collected our dataset from Facebook in detail. After that we describe the major [SNA](#) measures that we apply to this. We then describe how, with the collaboration of Lucas at UCL, we managed to do name-based ethnic classification of Facebook users on our dataset, by using Onomap (Mateos, P, Webber, R and Longley, 2007; Mateos, Longley, & O'Sullivan, 2011). This system allowed us to estimate individual level information on ethnicity, religion, language and even geography of users. After inferring such detailed cultural information, we analysed our dataset to see which groups are connected with each other and how much [affinity](#) they have for each of the other groups. These classifications give us a good insight into both individual and group level preference for a particular ethnicity, language or religious group. In the real world we observe that there are strong, and sometimes even segregated, communities based on religion, culture and language, this research seeks to determine the extent to which it is true with [SNSs](#). It also strengthens the argument that the reality of [SNSs](#) is a replica of real life social network.

5.1 Ethics and Purpose

Extracting social network data from Facebook is quite a complex issue. All these [SNSs](#), like Facebook, are reluctant to share users' information, even for research. In order to study social networks of an [SNS](#) the only method available (to those not working for the SNS) is to use a web crawler, which is what we have done here. Since the time we crawled Facebook for our research, improved privacy measures have been implemented. For instance at the present time, not only can Facebook users hide their profile from the

general public, they can also hide their social network. Several studies have looked into this matter in detail, such as (SA Catanese & Meo, 2011; Gross et al., 2005).

If Facebook provided a mechanism to explicitly ask each user, whether they are happy to share their information that would have served the purpose. However, that would only be applicable for a rather small study, not like ours which covers over half a million users. One thing should be clarified here, Facebook's terms of service, at the time of the research, only restricted the uses of extracted data; it did not restrict data derived from Facebook Properties (Hogan, 2009). Here, we are dealing with only the derived ethnic information of users, which is inferred from the name-based classifier Onomap. This information, we should clarify, is only guessed at based on the screen name of Facebook users. Also, in order to protect privacy of users, we are not going to share any data that would allow the identification of individuals and, having completed this research, have deleted any such personal data.

The overall purpose of our work is to identify inter and intra ethnic linkages of Facebook users at our university. We will test our hypotheses defined in the earlier section. Several studies on Facebook (SA Catanese & Meo, 2011; Lewis et al., 2008; Alan Mislove et al., 2010) have been carried out to understand how users develop their profiles and how taste and education level play a role in developing new ties, however not much work has been done on the role of ethnicity in this. This involves looking at the structure of social network and also the correlation between the derived attributes of users.

In accordance with the regulations at the Manchester Metropolitan University, we applied for an annual review process. This included a comprehensive review of all aspects of the underlying research. In specific, data collection methods, cleansing and storage was assessed. It was then accepted by the university's assigned reviewer. Subsequently, we have obtained a letter from the chair of the Ethics Committee, Professor Stephen Whittle, that as long as the data does not identify individuals and also not cause any distress to the subjects, this is consistent with the university's ethical guidelines. All of the relevant documentations have been provided in the Appendix F.

5.2 Facebook Graph (Reference Graph)

In this section, we describe the dataset we collected in order to test the above hypotheses. Our data collection method was as follows, which is summarised in Table 5-1. Firstly, we created a new Facebook account and then joined our university's network. To get ourselves registered into our university's network, we used our official email account with the domain *mmu.ac.uk*. We then started crawling the social network from the profile which we have added as a 'friend'. This involves capturing first and second names of the profile which belong to [MMU](#)'s network and then recording each of their friends' Facebook IDs. We continued this process until we had collected a substantial number of profiles, and hence could reconstruct the social network between the users crawled. We collected our data from Facebook during the period November 2009 to April 2010. At that time Facebook had almost 400 million users (Facebook, n.d.-c). Initially we wanted to crawl the whole location based network (regional network), such as Manchester, for all the profiles and their information but Facebook had removed that feature⁵. Instead of location we added a few people from our local network at the Manchester Metropolitan

⁵ The regional networks feature has been removed from Facebook around June 2009 - <http://www.facebook.com/blog.php?post=91242982130>

University ([MMU](#)). We then crawled the Facebook network through them. One profile was enough to get our crawler started.

Our methodology was to add a person from our network into our profile and then collect as much information as possible by crawling the network starting at that user. In total, we managed to collect a publically available dataset of almost half a million profiles. The web crawler we used is an adaptation of Alan Mislove's⁶ crawler, we would like to thank him for sharing the code with us. It was originally designed to get a location-specific network, the nodes and their links. Since we were interested in all the publicly available information - specifically racial and ethnic information - we modified it to suit our requirements by capturing complete names (first and second names) of users. We ran the crawler on one machine and then kept it running until we achieved a sizeable sample of the local Facebook network.

```
1. #Facebook Credentials
2. #Database Credentials
3. # Declare Sleep time
4. # connect to the database
5. # set up the browser state and cache
6. #Login to Facebook
7. # Process Users until all have been explored (the whole Facebook) – or we terminate
   the crawler
8. while (true) {
9.   #process a profile which hasn't been crawled yet (the first one according to its addition
   in the database)
10.  # open that profile
11.  # Parse its content: save their first and last name
12.  # set isCrawled = true in the database
13.  # get their number of friends' pages (a maximum of 400 is displayed in friends' list)
14.  # for each of them, save their ID, first and last name (for each page)
15.  # also for each of them, save their relationship with this profile (edge information)
16.  # repeat the process ONLY for those nodes which have MMU as their network
17.  # sleep for some milliseconds
18. }
```

Table 5-1 - Algorithm to crawl Facebook

For this crawl we used a (biased) Breadth-first-search (BFS) algorithm, which is a well-known traversal algorithm. It has been extensively used to crawl various [SNSs](#) (Alan Mislove et al., 2007; Panzarasa et al., 2009; C. Wilson et al., 2009). The algorithm starts from a single node, which is known as a *seed*, and then discovers its neighbours. In our case the agent crawler logs into Facebook with our account credentials; it fetches the profile page of every neighbouring user, scrapes data out of it, cleans it and then stores it in the local database. After that, it fetches the friends' list of the user and then the same

⁶ <http://www.ccs.neu.edu/home/amislove/>

process repeats for one of the friends of the user, based on first-in-first-out (FIFO) strategy and being a member of the [MMU](#) network. It means those Facebook profiles, which have [MMU](#) affiliation, are going to be traversed in the order they were identified and stored in the database. Thus all *ego* nodes are from the [MMU](#) network, while *alters* may not be. In Table 5-1 we have summarised each step of the algorithm. The protocol starts by firstly reading the Facebook credentials (ID and password) of our seed profile and then the database credentials where it stores the social graph and names of Facebook users. After this, it sets up a browser's state and then logs into Facebook. It then crawls through the students of [MMU](#), which we have added in our seed profile continuing to fetch their social network. It involves gathering both ego and alter profiles. The ego profiles belong to the [MMU](#) network, whereas for alter profiles, there is not a requirement. Nodes, the links of the social graph, are also recorded in the order they are found. The algorithm stops when all the nodes are visited. In our case, however, we selected the next neighbour only if it was from the [MMU](#) network.

For a large graph such as Facebook, the whole crawl is time and resource intensive. Even if we do not consider time constraints and also the technological constraints set up by Facebook, 44 terabytes of data would be required to be downloaded and processed for the whole Facebook network according to a Facebook study (M Gjoka, Kurant, Butts, & Markopoulou, 2010) carried out in 2010. Before any information is retrieved from any Facebook profile, it needs to be retrieved and then relevant information is scraped from it. If we consider a single friends list page, which is around 200kb (Salvatore Catanese, Meo, Ferrara, & Provetti, 2011), then take the current population of Facebook (Facebook, n.d.-a), a total 200KB x 1.23 billion = 293 terabytes of HTML data would have to be collected. For this reason we only collected a subset of it, obtaining only a relatively large dataset (half a million users' data) of Facebook users before terminating the crawler. So our strategy is to collect an incomplete BFS, but for our purpose of sampling [MMU](#) students' network, this is a sufficient strategy. In Table 5-2 we have summarised the structure of our dataset. The average number of friends in this set is higher than earlier reported Facebook statistics (130) (Facebook, n.d.-a), but it is quite comparable to another study done on Facebook (SA Catanese & Meo, 2011). Also the diameter of our crawl is almost similar to the same study.

Table 5-2 [MMU](#) Dataset Description

# of visited users	# of discovered neighbours	# of unique users	# of links	Avg. # of friends	Diameter
4601	568037	566012	1497443	326.28	6

Our dataset contains information from over half a million profiles (566012). Due to the limitation of the crawling strategy and the privacy settings of the users, which either protects users' public affiliation with [MMU](#) or restricts access from [MMU](#) network, the social network (complete ego-network) of 4601 people was gathered. The global Total Degree Distribution of this set can be seen in Figure 5-1.

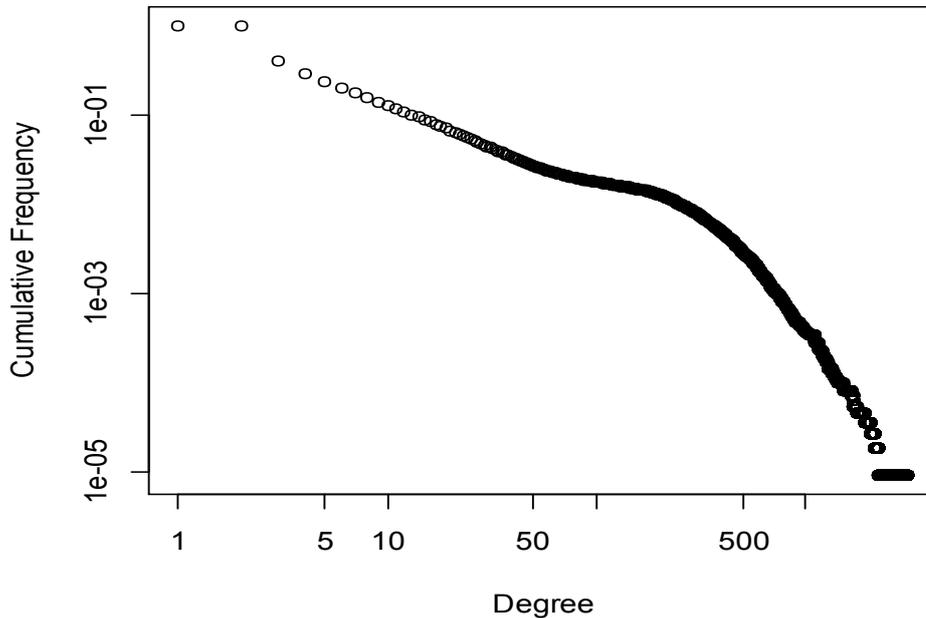


Figure 5-1 - Total Degree Distribution (log-log plot) of [MMU](#) Social Network

A seminal work regarding node degree distribution in Facebook has been done by Minas et al. (Minas Gjoka, Kurant, Butts, & Markopoulou, 2009). There, the authors showed that, unlike the established understanding of power-law distribution in degree distribution of nodes in an [SNS](#), Facebook's social network's distribution is different. Minas et al showed it has two different regimes of power-law outlook: one between $1 \leq k \leq 300$ and another $300 \leq k \leq 5000$, where 'k' represents node degree (or number of friendship links). The total degree distribution of our reference graph, plotted in Figure 5-1 shows a similar pattern for degrees smaller and greater than 300. There are clearly two identified regimes of power-law outlook as was found in Minas et al. (Minas Gjoka, Kurant, Butts, et al., 2009). For degrees less than 300, i.e. $1 \leq k \leq 300$, where 'k' is the degree, the fitted power law distribution has 2.06 as alpha, which was 1.32 (Minas Gjoka, Kurant, Butts, et al., 2009). And for the degrees over 300, it found out to be 2.88, which is not too far from 3.39, found in Minas et al. (Minas Gjoka, Kurant, Butts, et al., 2009). Both of these power-law outlooks have been confirmed by running Kolmogorov-Smirnov test over them, which showed that $p > 0.05$.

As of June 2014, Facebook has 1.32 billion monthly active users and, to its credit (Facebook, n.d.-b), Facebook has become a part of many people's everyday life. The amount of data within an [SNS](#) is huge, a holy grail for the research community and also for advertising companies. Since there are huge monetary and legal stakes involved, the data within [SNSs](#) is generally not shared; only a handful of researchers manage to get a subset of it and in turn, they are reluctant to share it with others. According to a study carried out in 2007, the amount of digital information created, captured, and replicated is 281 billion gigabytes (Gantz et al., 2008). Due to the relatively private nature of these [SNSs](#), the only means left to researchers is to use automatic tools such as a web crawler, or an [SNS](#) specific application to get users' information. Here we used the crawling strategy to get the data for our analysis.

There are a few caveats which should be explained here. Since we started from a single Facebook profile and then crawled from there onwards by exploring the social network to collect further information from other profiles, this methodology resulted in a fully connected network. Hence our results and findings cannot be generalised to the whole Facebook community, where there could be disconnected networks. As for the crawling strategy, although the underlying crawler can work in parallel, due to limited computational power and also relying on just one seed, it results in a biased network towards high degree nodes. In our case, as we collected the social networks of only those nodes which had an explicit [MMU](#) affiliation mentioned in their profile, our structural results may not be generalisable for the whole Facebook network.

5.3 Random Graph

To better understand the structure of the collected network, we compared it to a random network with the same nodes. We wanted to test whether the network structure did not matter and only the node attributes did, hence we developed a random network out of it. This could be considered as the null model. In order to identify how diverse our graph is, we are going to compare it against the null model, to determine whether the outcomes are inherently the same or not. This involves calculating and then comparing the diversity of nodes, in the reference and the random network, in terms of ethnicity, religion, language and geography. We have used the same nodes in the reference ([MMU](#) Facebook graph) and then randomised their links. This essentially means taking the reference graph's edge-list (source and target node pair), and then randomising them. For instance, let us consider an undirected graph with three nodes: 1, 2 and 3. If *node 1* is connected with two other nodes: 2 and 3, the edge-list would look like:

Table 5-3 Edgelist of Small Graph

Source	Target
1	2
1	3

This shows that node 1 has degree 2, whereas nodes 2 and 3 have degree 1. The graph can be seen in Figure 5-2. Please bear in mind that we consider the network as undirected. The number of visited nodes, which means the unique source nodes, is 1 in this case and the number of unique target nodes, or number of neighbours, is 2.

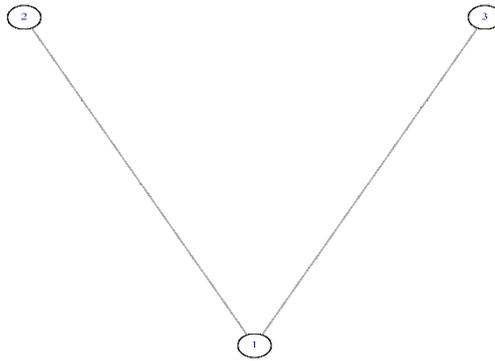


Figure 5-2 - Simple Graph

After we randomise the edge-list to form a random graph, the edge-list can now be:

Table 5-4 Edgelist of Small Random Graph

Source	Target
3	1
2	3

Now node 3 has degree 2, and node 1 and 2 have degree 1. For the random graph, the number of visited nodes is 2 (node 3 and node 2), and the number of unique neighbours is also 2 (node 1 and node 3). The random graph is shown in Figure 5-3. This is a simple case, where nodes do not have any attributes. If nodes had attributes like ethnicity, we could compare both simple and random graphs, to determine the diversity of the reference network when compared with the random graph. Also we can determine whether the difference is statistically different or not.

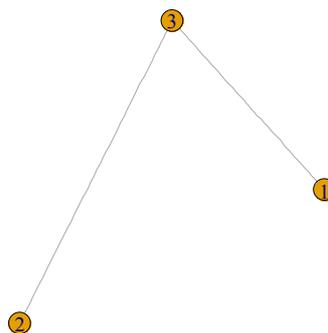


Figure 5-3 Random Graph

The global Total Degree Distribution of this network can be seen in Figure 5-4. The degree distribution, when compared with that of the reference graph, seems to have

drastically altered. Although there is still a power-law outlook for degrees greater than 7, the power of degrees is much lower. Also the highest degree is 19, which is 3796 for the reference graph.

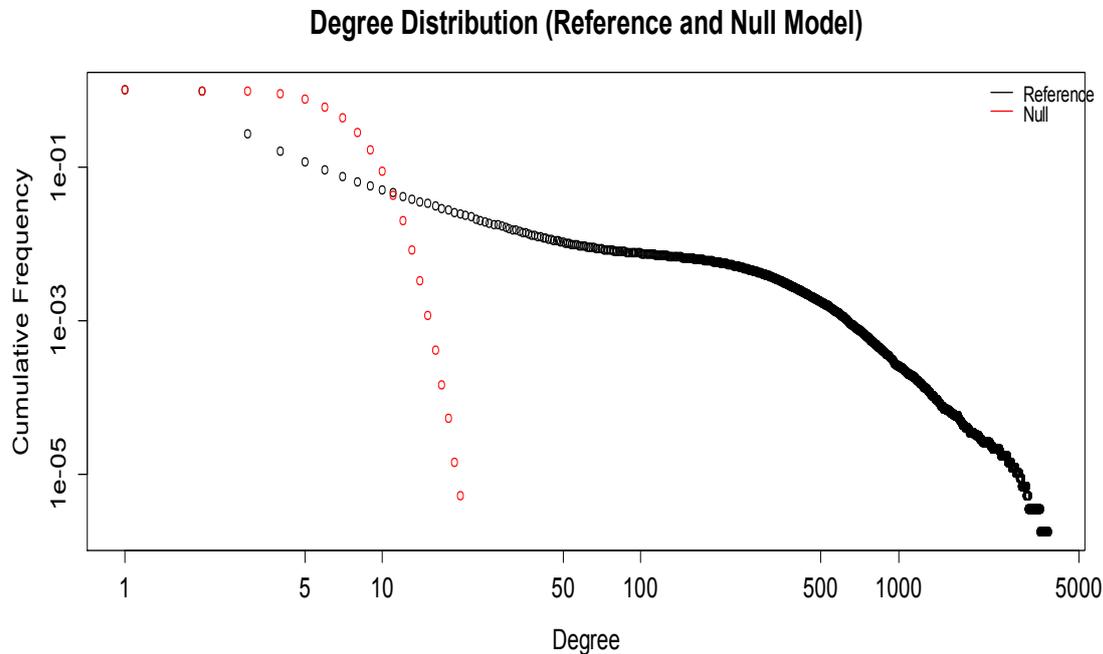


Figure 5-4 - Total Degree Distribution (log-log plot) of Reference Graph and Random Graph (null model)

As for the structural characteristics of the random graph, we have described them in Table 5-5. When we compare this table with that of the reference graph (see Table 5-2) we see that the number of visited users has increased manifolds (524511 versus 4601), which reduces the average number of friends from 326.28 to 2.64. Do consider that this measure only shows the number of friends of the visited users (source users).

Table 5-5 Random Dataset Description

# of visited users	# of discovered neighbours	# of unique users	# of links	Avg. # of friends	Diameter
524511	568037	566012	1497434	2.64	6

We have developed the random graph to have the same nodes as the reference graph. This means the nodes have the same attribute values for ethnicity, religion, language and geography classification. The main goal is that this would allow us to test our hypotheses: both covering the diversity of the reference graph and clustering, when compared with the random graph. Specifically, it will shed light on how, while maintaining the overall population which composed of dominant and non-dominant groups, inter and intra group affinities change and whether those differences are statistically different.

5.4 Ethnic Classification

In this section we describe what we mean by ethnicity and how we inferred an individual's ethnicity from their first and last name. Firstly, let us define what ethnicity stands for; according to Bulmer, ethnicity is a multi-faceted concept comprising the different dimensions that makes a person's identity, usually summarised as kinship, religion, language, shared territory, nationality and physical appearance (Bulmer, n.d.). For an individual, defining ethnic classification it is quite a contested process, as it is a subjective matter (Mateos, 2007). There are, however, positive aspects of it as well. For instance, in public health and demography literature, there seems to be a consensus that classification of population into distinct ethnic groups has proven useful to fight discrimination and entrenched health and social inequalities (Mateos, 2007; Mitchell, Shaw, & Dorling, 2000).

In this next section, we are going to talk about the name-based ethnic classifier we have used. The main idea behind this tool is to fill the gap for the missing ethnic, religious and language classifications of users. This allows us to identify the ethnic mix of our dataset, and also inter and intra ethnic propensities to develop friendship links between users.

5.4.1 Onomap

In this section, we discuss the Onomap (Mateos, P, Webber, R and Longley, 2007; Mateos et al., 2011) project which helped us identify, to an approximate degree, the ethnicity of nodes within our dataset. It helps us infer, probabilistically, the ethnicity of a Facebook user based on their first and second name. Having each Facebook profile's ethnic information allowed us not only the diversity of our graph, but the individual preference of link development. To estimate the ethnicity of each Facebook user, we collaborated with the geography department at UCL, London. This estimation was done on the basis of Facebook profile user names. According to a study carried out by Dwyer et al. (Dwyer et al., 2007) on a small dataset of sixty nine Facebook users, it was shown that all those users of Facebook revealed their real name, which was significantly higher than other [SNSs](#), such as MySpace. One can, of course, keep a fake name; there are no mechanisms to stop you from doing so. However, with the 'real name' Facebook policy, if someone reports a user's profile as 'fake', then they need to provide some sort of identification which authenticates their name – for instance a state I.D., a library card, or a piece of mail (Phillip, n.d.). This policy, however, creates problems for members of the [LGBT](#) community, who would like to be anonymous and also Native Americans, whose names are difficult to authenticate by Facebook (Phillip, n.d.). In general, however, the study by Dwyer et al. (Dwyer et al., 2007) tells us that people are more likely to trust and share more information on Facebook, than on any other [SNS](#). So the names used in Facebook become a very useful proxy for further estimation.

Onomap classification is based on surnames and forenames which help in estimating ancestral groups, producing valuable insights when ethnicity, linguistic or religious data are not available at appropriate temporal, spatial or nominal (number of categories) resolutions (Mateos, P, Webber, R and Longley, 2007). For our [MMU](#) dataset, where we only have first and second names of Facebook users, Onomap fulfils our requirement quite well, and provides inferred ethnic, religion, language and geographical approximations. In the Onomap scheme of classification, Mateos et al. used Cultural-Ethno-Linguistic ([CEL](#)) taxonomy which stands for Cultural, Ethnic or Language groupings. This [CEL](#) taxonomy summarises four main dimensions of an individual's

identity, which are: a religious tradition, a geographic origin, an ethnic background, (usually reflected by a common ancestry genealogical or anthropological links) and a language (or common linguistic heritage).

The Onomap system covers data from 28 countries with detailed information from the UK in particular. The data is accumulated through the UK electoral register and public telephone directories of 27 countries. It is thus suitable for inferring characteristics of a UK subpopulation, though it may have less fine-grained information on kinds of people that are rarer in the UK. There are 10.8 million unique surnames and 6.5 million unique forenames. It has its own system of classification with 185 Onomap types, aggregated into 66 ethnic subgroups and 16 groups. We have shown these ethnic and sub-ethnic classifications in Table 5-6. Onomap takes into account both first and second names. It has been evaluated against large population registers where the self-reported ethnicity is available next to a person's name and preliminary results show an overall specificity and sensitivity around 80–90% (Lakha et al., 2011). Its details can be found in (Mateos, P, Webber, R and Longley, 2007; Mateos et al., 2011). It has successfully been applied in Camden and Southwark Primary Care Trusts as well as other public and private organisations (Mateos et al., 2011), to approximate the ethnic background of registered people. Naming networks were constructed linking surnames through the forenames they share in 17 countries at the individual-person level drawn from the aforementioned UCL World-names database.

Onomap Group	Onomap Subgroup
African	African
	Black Southern African
	Congolese
	Ethiopian
	Ghanaian
	Nigerian
	Sierra Leonian
	Ugandan
Celtic	Celtic
	Irish
	Scottish
	Welsh
East Asian & Pacific	Chinese
	East Asian & Pacific
	Hong Kongese
	Malaysian
	South Korean
	Vietnamese
English	Black Caribbean
	English
European	Afrikaans
	Albanian
	Balkan
	Baltic
	Czech
	Dutch
	English
	European
	French
	German
	Hungarian
	Italian
	Polish
	Romanian
	Russian
	Serbian
	Ukrainian

Onomap Group	Onomap Subgroup
Greek	Greek
Hispanic	Hispanic
	Portuguese
	Spanish
International	International
Japanese	Japanese
Jewish And Armenian	Armenian
	Jewish
	Jewish And Armenian
Muslim	Bangladeshi
	Eritrean
	Iranian
	Lebanese
	Muslim
	Muslim Middle East
	Muslim North African
	Muslim Stans
	Pakistani
	Pakistani Kashmir
	Somalian
	Turkish
Nordic	Danish
	Finnish
	Nordic
	Norwegian
	Swedish
Sikh	Sikh
South Asian	Hindi Not Indian
	Indian Hindi
	South Asian
	Sri Lankan
Unclassified	Unclassified
Void	Void

Table 5-6 Onomap Ethnic/Subethnic groups

The Onomap system takes the first and last names of a person as an input and then after several iterations, (dependent upon whether subgroup ethnicity of first or second name is going to be used, or alternatively, the upper level ethnicity), an ethnic estimation is calculated based on the data collected. These iterations include the various Onomap classifications defined earlier. There is a score assigned to each output as well. In addition to the score, Onomap estimates, for each pair of first and last names, not only fine-grain ethnic classification by Onomap subgroup, but it also estimates geography, religion and

language classifications. Table 5-7 summarises the different inferred information, supplied for an individual.

Table 5-7 Onomap Classification

Onomap Group	Onomap Subgroup	Geographical Area	Religion	Major Language
--------------	-----------------	-------------------	----------	----------------

To our knowledge, name-based ethnicity recognition has never been applied to Social Network Systems ([SNSs](#)) on such a large scale. Sometime ago, the diversity team of Facebook released the trend of various ethnicities in the US, based on the US census data (Jackson & Rogers, 2007). Also, Facebook itself tried to understand different ethnic behaviours by classifying US users by the census data (Chang & Rosenn, 2010). Other than that, we are not aware of any study on [SNSs](#) with a dataset as large and diverse as ours.

Table 5-8 - Ethnic Classification of [MMU](#) Students and the Reference datasets

Ethnicity	Total Students 2004/5	% MMU	% in the reference dataset
White	24614	74.91%	68.32%
Black	1022	3.11%	1.41%
Asian	3697	11.24%	24.18%
Chinese	1153	3.51%	1.40%
Other (including Mixed Heritage)	870	2.65%	0.42%
Not known or refused	1504	4.58%	4.27%
Total	32860	100.00%	100%

We have also compared the total ethnic distribution of [MMU](#) students with that of our overall reference network, shown in Table 5-8, the data of which comes from an [MMU](#) report for the student population of 2004/5 ([MMU](#) University, 2007). Since the ethnic breakdown in the Onomap classification scheme differed from the ethnic classification used by [MMU](#)'s document, we combined the following groups of Onomap, to compare as best as possible with that of [MMU](#)'s.

Table 5-9 - [MMU](#) Ethnicity and Onomap Ethnic Classification

MMU Ethnicity	Onomap Ethnic Group
White	Celtic, Greek, Hispanic, Nordic, English, European and Jewish and Armenian
Black	African
Asian	Muslims, Sikh and South Asian
Chinese	East Asian and Pacific, and Japanese
Other (including Mixed Heritage)	International
Not Known	Unclassified and Void

When we see Table 5-8 (and also Table 5-9), we find that our reference dataset, when compared with the overall ethnic distribution of [MMU](#) students, has a fairly good representation. In terms of specifics, however, non-Chinese Asians amount to more than double the percentage of the population (24.18% versus 11.24%). As we have mentioned before, this is mainly due to the initial seed of our crawler being the profile of a Muslim student. We have mitigated the over-representation and also under-representation of each ethnic group (as well as other Onomap based groups), by normalising the groups, so that we may compare their propensities at the same level, details of which are covered in the Section 5.7.2.

5.5 Shortcomings

There are a few limitations in our dataset, starting with the crawling strategy we employed. Both empirical and theoretical research shows that incomplete BFS tends to favour highly connected nodes (Lee, Kim, & Jeong, 2009), resulting in a skewed degree distribution. Due to limited access to resources however, we relied on this strategy. In order to make sure our crawling strategy gathered somewhat representative dataset, we verified two similar distinct power-law regimes that are known to occur in larger datasets (Minas Gjoka, Kurant, Butts, et al., 2009).

For the last few years, the users of [SNSs](#), especially of Facebook, have become more aware of their privacy issues. The public display of one's profile and friends' list used to be common, but more recently a lot of the users have started to make their information private. In our data collection, we did manage to get more than a screen name for Facebook users (for example hometown), but for conformity and generalisation, we have not taken them into account.

We are unable to quantify (or cross-check) how much of this ethnic estimation by Onomap was correct for our whole dataset, but based on its internal classifications, we can shed more light on it. Additionally, as an internal check, though not systematic, we manually validated a few hundred records, and compared these with Onomap's automated classification, to see if it had been correctly estimated. This involved not only looking at the first and last names, but also looking at all the other information on the user's profile. Specifically to gather cues, such as the cultural and religious groups users are affiliated with and also their profile picture. We found Onomap to be quite a good estimator. This method of manual estimation is established in the literature, and has been used by

Lewis et al. (Lewis et al., 2008), where a combination of profile pictures and first and last names were used together to identify a person's ethnicity.

One important shortcoming of Onomap needs to be mentioned here. It does not work well with mixed ethnicities. According to the office of national statistics, the mixed population in the UK is around 2% (Office for National Statistics, 2013), which is not enough to interfere with the planned hypotheses testing. For the sensitivity of each estimate, a 'personal score' has also been calculated, but for simplicity, we did not take this into account in our analysis. Onomap has an internal mechanism to identify how good an estimation is, based on its prior knowledge of names. It assigns one of the eight cases to its classification. Below are their descriptions which have been taken from the help file of Onomap Software (Lakha et al., 2011):

- CASE 1: Both the surname and forename are unclassified or not found in the dictionaries;
- CASE2: The surname or forename is unclassified or not found in the dictionaries;
- CASE 3: Both Onomap types of surname and forename are the same: the person is assigned to that Onomap type;
- CASE 4: Both Onomap subgroups of surname and forename are the same: the person is assigned to that Onomap subgroup;
- CASE 5: If the absolute difference between Onomap scores of each name element is larger than 0.2: the person gets assigned to the Onomap type with the highest score;
- CASE 6: If the Onomap groups are the same: the person is assigned to that Onomap Group;
- CASE 7: If the absolute difference between Onomap scores of each name element is smaller than 0.2: the person gets assigned to the Onomap type with the highest score;
- CASE 8: If the forename and surname cannot be identified because of formatting problems: the person gets assigned to the Onomap group: 'Unclassified';

For our dataset we identified the distribution of all such cases, which can be seen in Table 5-10. The highest distribution of 35.55% falls into the case 3 classification, which means Onomap recognised the same Onomap type (high level of ethnicity) of both first and second names and then assigned the same to it. In case of low-level ethnicity (Onomap subgroup), merely 4423 (0.78) have been recognised with having the same sub-ethnic classification of both first and second names. Almost 25% of names (pairs of first and second names) have a biased classification from either the first or second name – Case 5. However, 15.17% of names have been assigned with a less biased classification based either on first or second names (the absolute point score difference between first and second names was less than 0.2). Almost 15% of the names have not been classified because either the first or second name was not found in Onomap's dictionary. Almost 3.5% of names could not be classified. Due to formatting issues, there was not a single instance of unrecognised pairs of first and second names in our dataset; hence we do not see any case 8 distributions in Table 5-10.

The significance of this shows how complex the whole Onomap classification can be. We did not, however, treat any of the cases defined here any differently.

Table 5-10 - Onomap Case Classification

Onomap Classification (%)	Case 1	Case 2	Case 3	Case 4	Case 5	Case 6	Case 7
	3.49	15.56	35.55	0.78	25.30	4.16	15.17

5.6 Data Sharing

During our work we really struggled to get hold of data and our thanks goes to the researchers who guided us and shared their datasets with us. Our work would not have been possible without their support. However, the privacy of users is at the core of our work and we will ensure that no personal information of any kind will be released. For the moment we have decided not to release any of our data. We already have deleted all personal information of users, and to further anonymize the dataset, we have randomized the order of the social network, so it is infeasible that anyone could reverse-engineer who individual Facebook users were.

5.7 Results

Before we start sharing the results, let us summarise some of the measures we used.

5.7.1 Affinity

This measure is used to discover 'attribute level communities', i.e., subgraphs with high [affinity](#). In Table 5-11 we have shown the affinities of each of the five attributes. The basis of this lies in the fact that users are statistically much more likely to be friends with other users who share their attributes (A Mislove, 2009). It shows that all the attributes are positively correlated with link development. Just to reiterate, values greater than 1 show positive correlation with link development. In the case of the null model, it appears that none of them are significantly different from any other.

Table 5-11 Affinity Measures of Node Attributes

Attribute	Affinity of Reference Social Network	Affinity of Null Model
Ethnic Group	1.52	1.02
Ethnic Subgroup	1.36	1.03
Geographical Area	1.38	1.03
Religion	1.44	1.02
Language	1.03	1.03

For the reference network, the most important node attribute is ethnicity, with religion following thereafter. We have also calculated the [affinity](#) measures of the same attributes for the null model. It seems none of them have a defining [affinity](#) when compared with

others. However, interestingly the [affinity](#) measure for the language attribute of the null model is the same as the reference social network (1.03). We believe this is due to the large number of groups presented in the language attribute. [Affinity](#) measure is a macro level measure and does not deal with individual group [homophily](#) among individuals. For inter and intra group affinities, instead of [affinity](#) measure, once we look at the Silo Index for each of the individual groups, we can see how different the outcomes are for the null model. More details will be covered in the subsection 5.7.6.1 under the results for the language attribute.

5.7.2 Contracted Graphs

In order to see how the four attributes: ethnic group, geographical area, religion and language, are useful to develop friendship links, we have contracted our underlying graph into four different graphs. For each group in the four attributes, we also calculated the extent to which they are linked internally with the help of the *Silo Index*. This is an index that identifies the proportion of links between nodes with the same attribute value in a network.

We have contracted the graphs based on the four attributes we are concentrating on. This involves merging all nodes having the same attribute (such as ethnicity) into one, while keeping track of their links with nodes with attribute values. For instance, all nodes which have ethnicity 'Muslim' will be merged into one single node called Muslim. This allows us to capture the diversity of nodes at an attribute level and also reduces the complexity of the large network. Let us explain this with an example:

....If we look at the [MMU](#)'s reference network, its adjacency is matrix A , where A_{ij} contains either 1 if there's a link between node i and j , otherwise 0. The dimension of this matrix would be: 566012 (number of unique nodes) rows x 566012 columns. When the same reference network is reduced by contracting ethnicity of users into groups, the number of nodes reduces to the number of unique ethnicities found in the reference network (16 nodes) and the dimension of the adjacency matrix changes into: 16 rows x 16 columns. Unlike the non-weighted links in the reference network, we have weighted links in the contracted graphs, where A_{ij} contains a count of the number of links between node i and node j .

For each attribute, we plot how groups are linked with each other. In each graph we plotted, we have shown the weights of developing links of each group with the help of the Silo Index. We have also shown how closely each group is connected with all the others with the help of a plot. Let us describe how we have contracted the underlying graphs.

5.7.3 Ethnic Group

As described in the subsection on Onomap (Section 5.4.1), there are 16 top level ethnic groups in total. In our contracted group of ethnic groups, we find all these groups which show the diversity of our dataset. In Figure 5-5 we show how each ethnic group is connected with each other and what their Silo Indices are. The node size represents their population size; the English (45.8%), Muslim (21.02%), and Celtic (16.42%) are what we call dominant groups, which are also linked closely with each other. The value followed by the ethnic name is its Silo Index in the figure. The Muslim group has the highest Silo Index (-0.21), followed by the English (-0.35) and Celtic (-0.74) groups. There is a very high correlation of Silo Index with that of node size (population), which one could safely say is

an artefact of the measure. Although the English group has the highest population, when it comes to inter-group [homophily](#), the Muslim group leads.

To identify how each ethnic group is linked together, we have normalised the weight of links by the Equation 5-1:

$$\text{weight} (A \rightarrow B) = \frac{\# \text{ of links between node A and node B}}{\text{Total \# of links of node A}} * 100 \quad \mathbf{5-1}$$

In order to know the weight from node A to node B, we calculate the percentage of links flowing from node A to node B, with that of the total number of links of node A. Note that although we are dealing with undirected graphs, this produces different weights between node A and node B. For instance, if node A has 10 links, out of those 7 flow to node B and 3 to node C, whereas node B has only 2 links in total which flow towards node A, the weights of weight (A -> B) and weight (B -> A), will be:

$$\text{weight} (A \rightarrow B) = \frac{7}{10} * 100 = 70 \quad \mathbf{5-2}$$

$$\text{weight} (B \rightarrow A) = \frac{2}{2} * 100 = 100 \quad \mathbf{5-3}$$

This means from the perspective of node A, that 70% of the link belongs to node B, but from node B's perspective, it is 100%. In other words, this weight calculation normalises the weights of each node, by assuming that dominant and non-dominant nodes can be compared at the same scale. This measure overcomes the population biasness inherent in the Silo Index, by comparing each group at the same level.

This produces a normalised adjacency matrix of the graph showing how each group is linked with others. If all links from an ethnic group *A* are with another group *B*, the normalised weight from *A* to *B* would be 100. We plot the adjacency matrix in Figure 5-6. The highest weight (60) is between Celtic and English. We can clearly see the dominant groups (English, Muslim and Celtic) are highly connected not only with themselves, but with the non-dominant groups as well. Most of the non-dominant groups are linked with the English group. The South Asia, Sikh, and Nordic groups are connected with the Muslim group, while the English group is heavily connected with Celtic. To identify the top 10% of the links in the adjacency matrix, we looked at the highest values in the whole matrix, where we found that the English group covers 64% of them, whereas the Muslim and the Celtic group cover 32% and 4%.

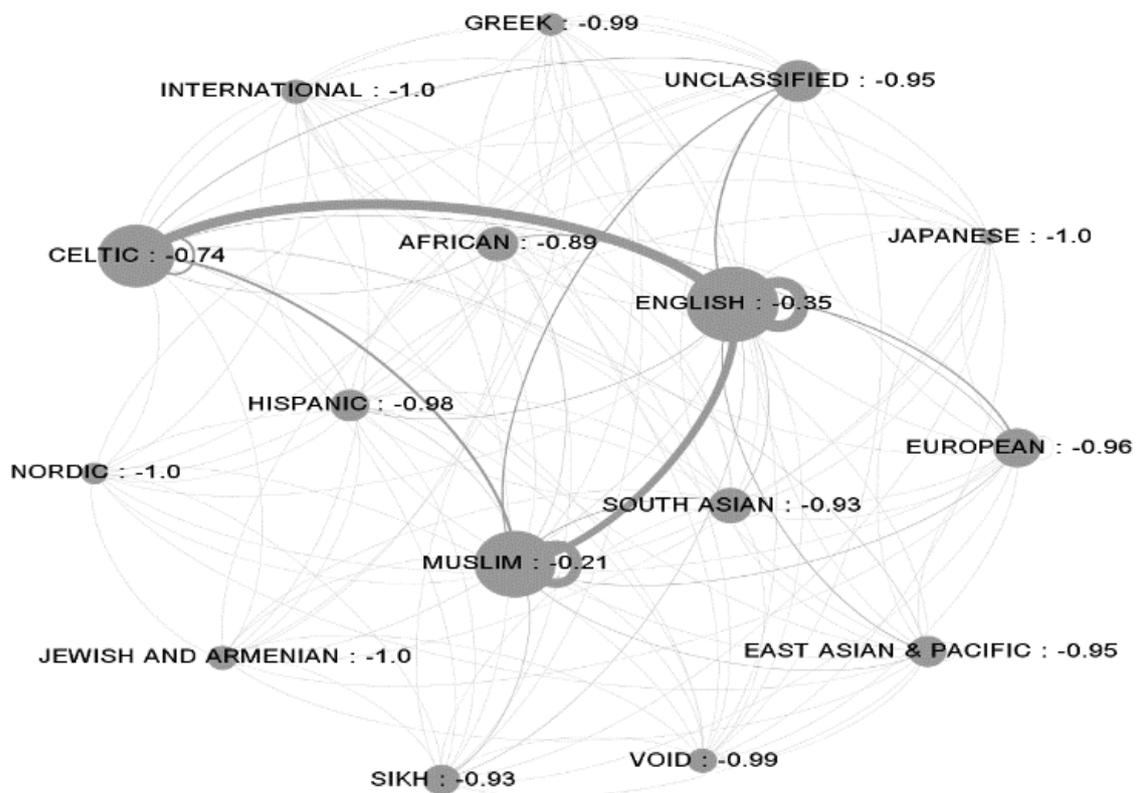


Figure 5-5 Onomap Group Social Graph

LevelPlot - Ethnic Group

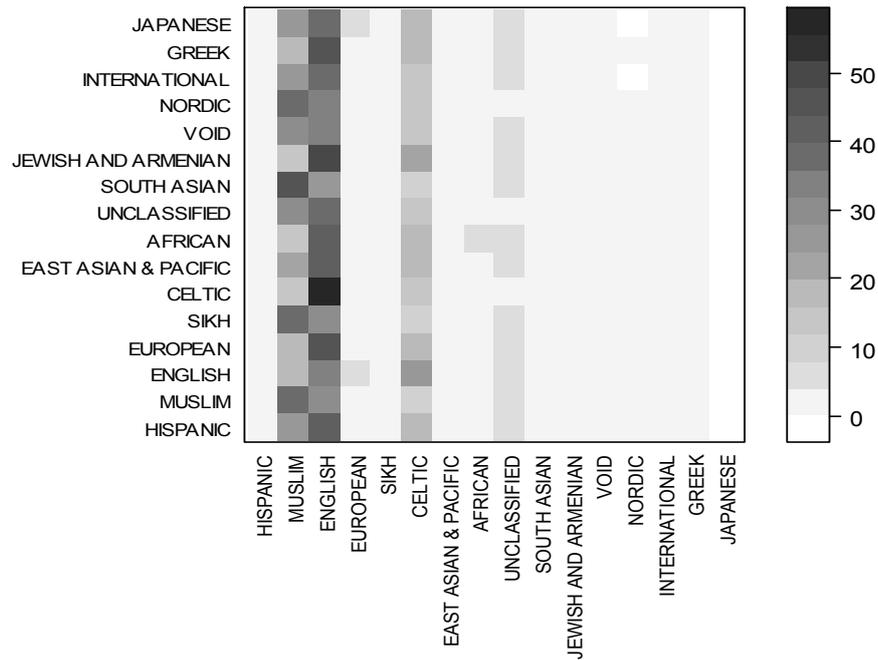


Figure 5-6 Normalised Links of Ethnic Groups

5.7.3.1 Comparison

To compare the underlying Silo Indices of the reference graph to that of the random graph (null model), we have plotted the Silo Index of both of them in Figure 5-7. The bigger the difference between the reference and the null model, the higher the weight of inter-group linkages. In the figure below, we have displayed three bars for each of the ethnic groups. The first bar (the red one) shows the Silo Index in the reference network. The green bar shows the Silo Index in the null model, whereas the blue bar represents the difference between the reference and the null model's Silo Index. The Muslim group has the highest difference (0.58), followed by the African (0.10) and Celtic (0.09) groups. The Muslim group is a dominant group (21.02%), but certainly not the most dominant (the English group represents 45.8%), but it has the highest difference between the reference and the null model. It means that the highest difference cannot be explained by merely the population size of an ethnic group.

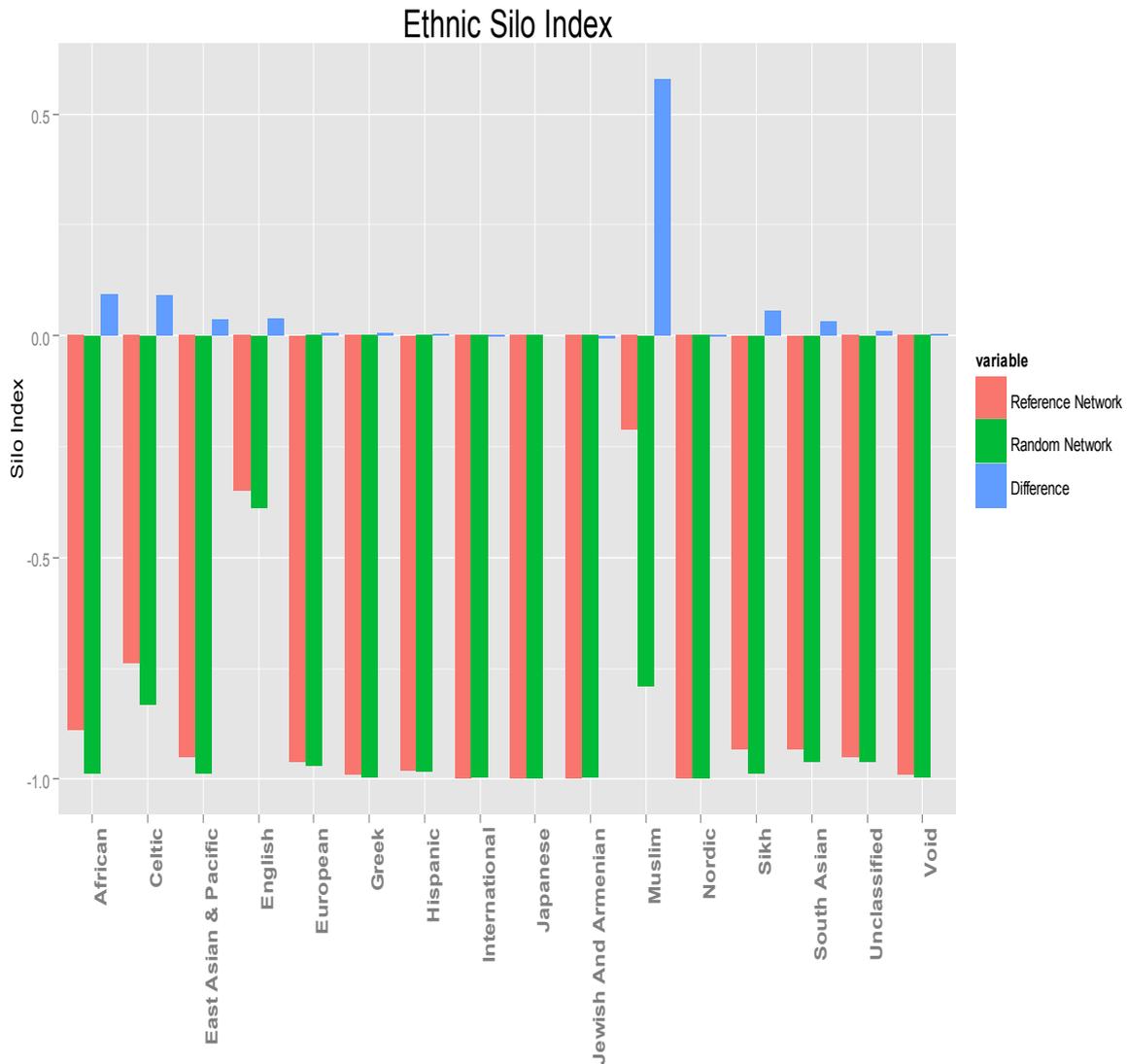


Figure 5-7 Ethnicity based Silo Index comparison between the reference and the null model

We have also performed a t-test to compare the difference between the underlying reference graph (Facebook graph) with that of the random graph (null model), to identify how significantly different they are. It turns out they are different with a p-value < 0.05.

5.7.4 Religion

As for the religious groups, we also have quite a diverse set of them based on the Onomap classification – 12 of them in total. The dominant groups are: Christian: Protestant (51.48%), Muslim (21.14%), Christian (9.28%) and Christian: Catholic (7.83%). We have plotted the contracted social graph, based on religion, from our dataset in Figure 5-8. The Muslim group has the highest Silo Index (-0.21), the same as it was found with the contracted graph based on ethnicity in the previous section. This means that Onomap classification for both religion and ethnicity for the Muslim group are the same, as is evident from their same population size (21%). The rest of the groups have closer to -1

Silo Index, which shows their links are mostly with external groups, and their internal links are almost non-existent.

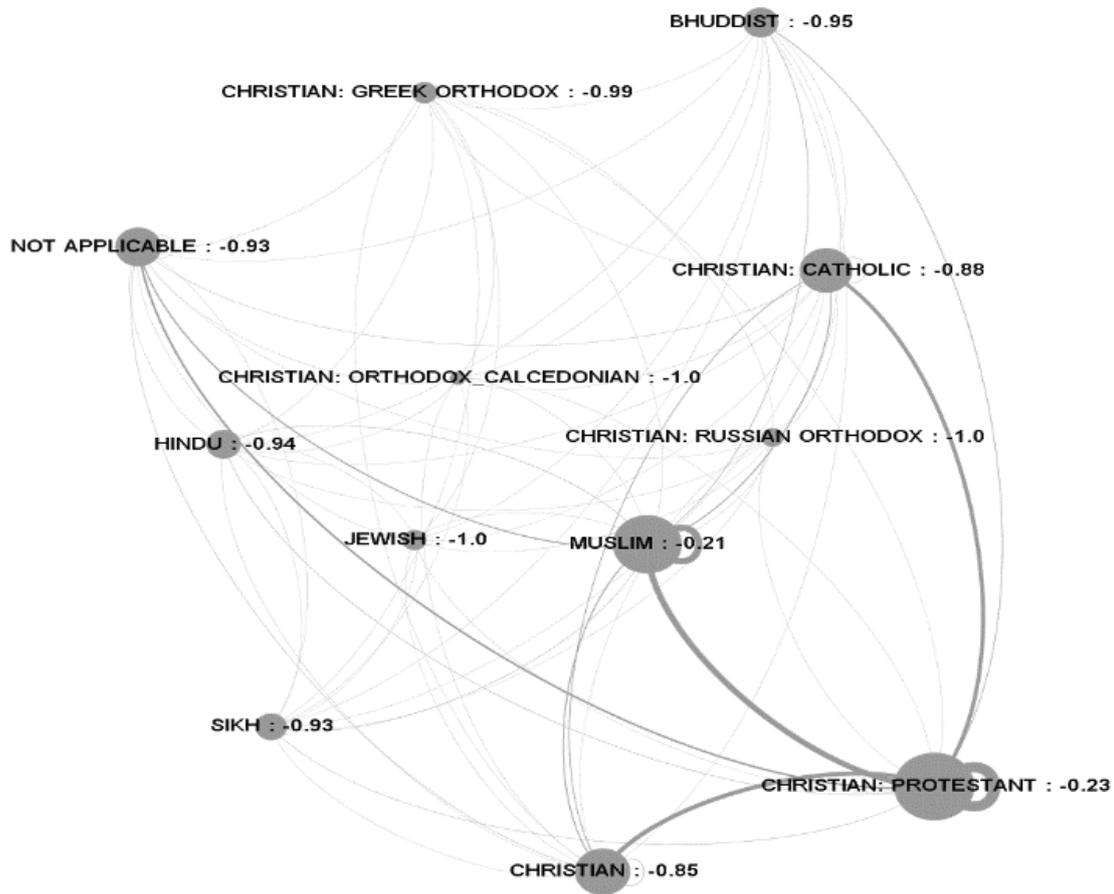


Figure 5-8 Religion Based Contracted Graph

To identify how inter and intra linked each religious group is, we plotted their weighted adjacency in Figure 5-9. The most populated group, the Christian Protestant, has the highest number of links with most of the groups. Overall the Christian group has the highest linkages with the Christian Protestant group (63.5). The scale on the right side of the figure shows the range (from 0 to 63.5). We have further investigated by calculating the top 10% of the links in the adjacency matrix. The linkage with Christian Protestant group covers 64.2% of the top links, whereas the Muslim group covers 28.5% of the links.

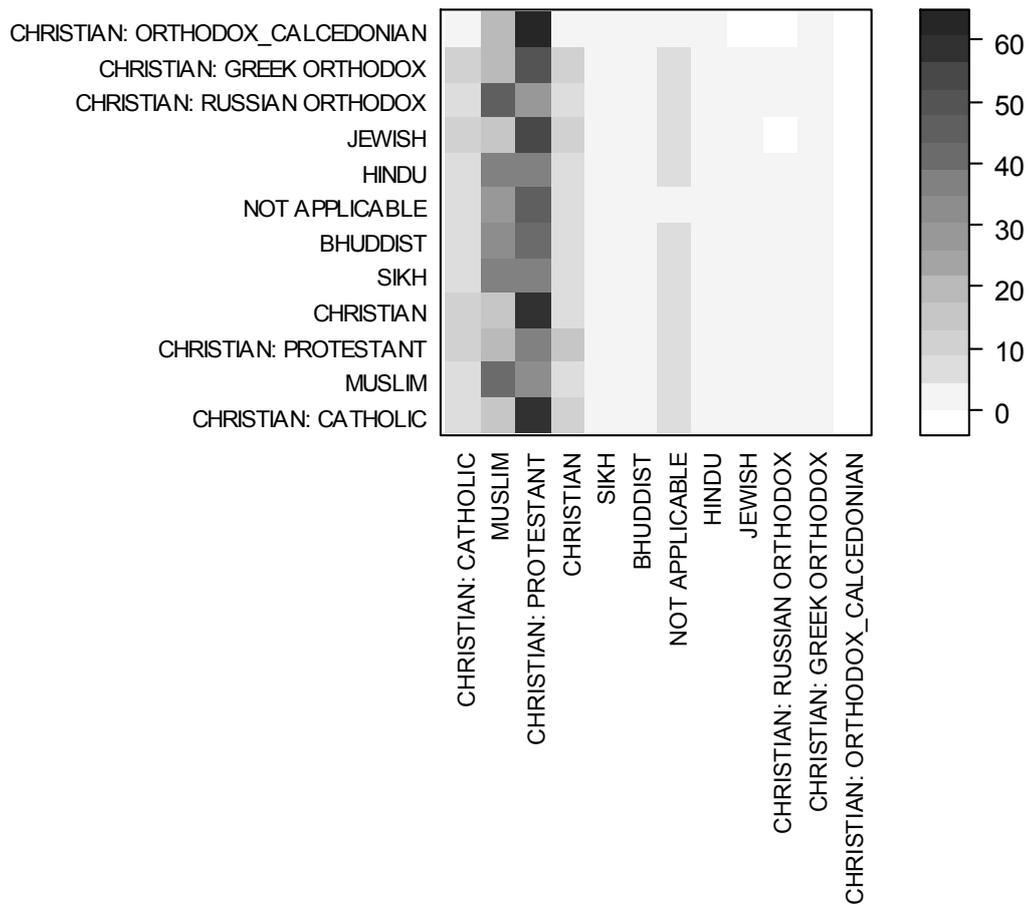


Figure 5-9 Normalised Weights of Links Based On Religion

5.7.4.1 Comparison

Similarly to the ethnic group, we have plotted the Silo Indices of religious groups for the reference graph and the random graph (null model) in Figure 5-10. In this case too, the Muslim group has the highest difference (-0.52), and then the Hindu (-0.3) and the Christian: Protestant (-0.19) groups come. Please note that just like Muslim as an ethnic group, Muslim as a religious group has the same high difference when compared with null model. It means Muslims, being one of the dominant groups, have the highest inter-link [homophily](#) than any other group. The least amount of changes or no change at all, was observed in the Christian: Orthodox_Calcedonian group. To see if there is a statistical difference between the two, we ran the t-test, and found the p-value < 0.05, which signifies that there is indeed a significant difference between the two.

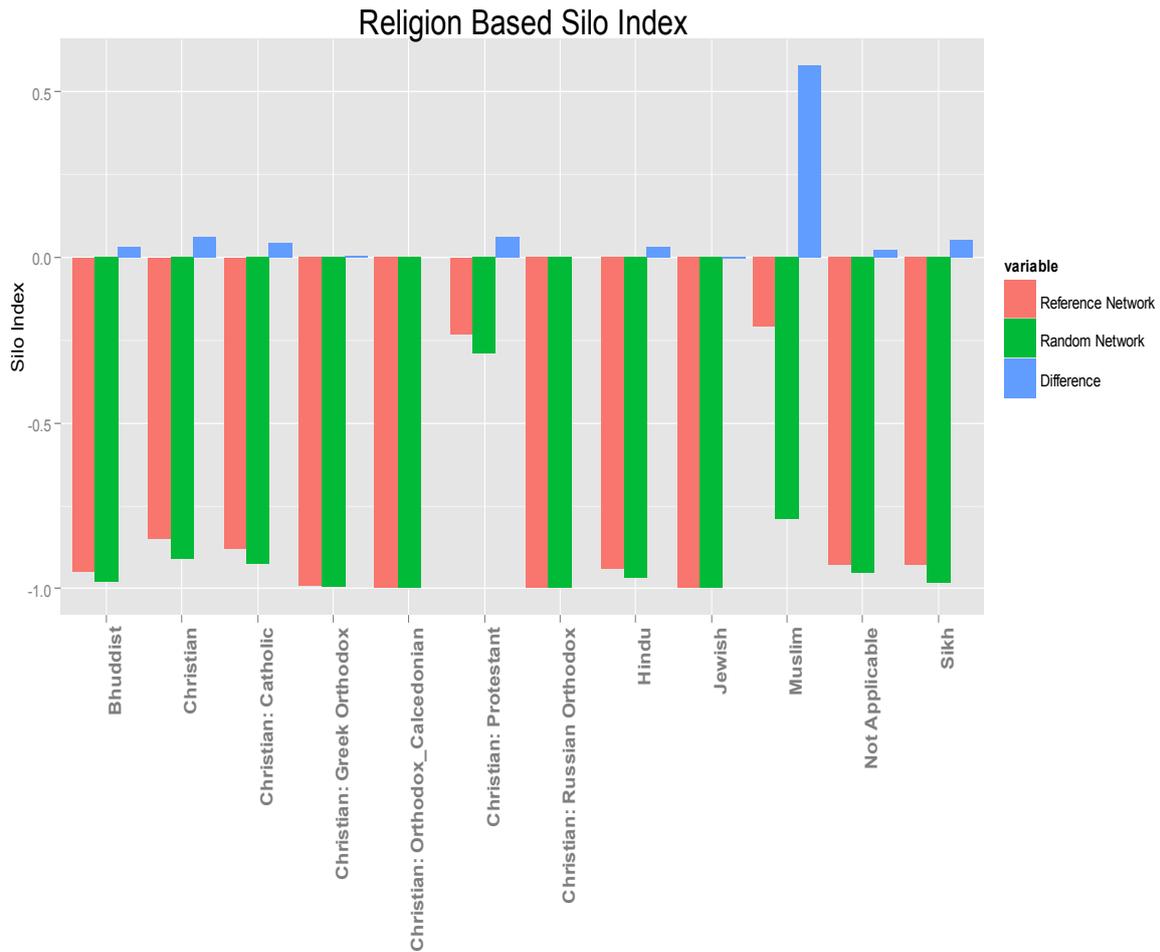


Figure 5-10 - Religion based Silo Index comparison between the reference and the null mode

5.7.5 Geography

For geography, we have plotted the same contracted graph. This attribute shows where the person might geographically be based according to their Onomap ethnic classification. The dominant groups are: British Isles (62%), South Asia (16.14%), and the Middle East (7.04%). In Figure 5-11 we have shown how the contracted graph (with inter and intra groups) are connected with each other. Again, the node size represents the population size. The Silo Indices for the three major groups are: British Isles (0.08), South Asia (-0.44) and the Middle East (-0.84).

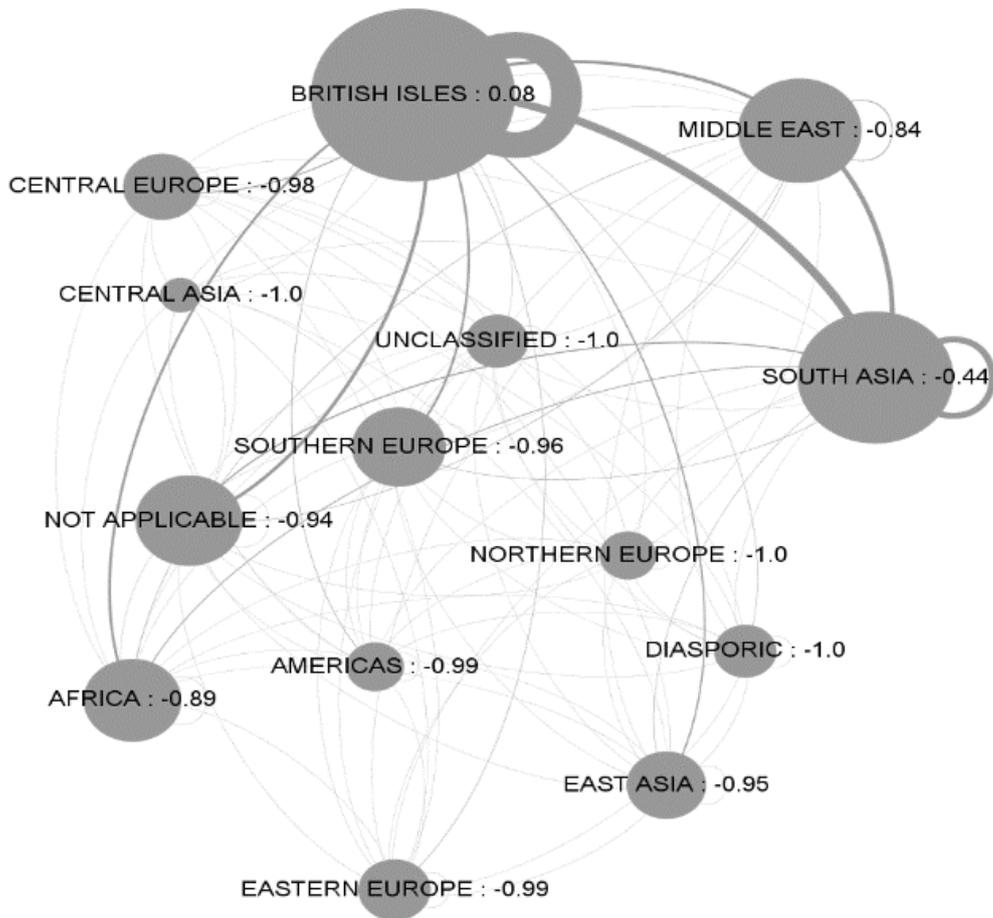


Figure 5-11 Geography Based Contracted Graph

To identify how closely each geography group is connected with each other, we have plotted the same weighted adjacency matrix in Figure 5-12. If we look at the scale in the figure (on the right side of it), we can see the highest value is somewhere around 70. It is actually 71.3, which is the link proportion from the American group to the British Isles group. For the top 10% of links, we learn that the British Isles group cover 73.68% (14 out of 19 top values) and the rest of the links (35.71) are covered by the South Asia group.

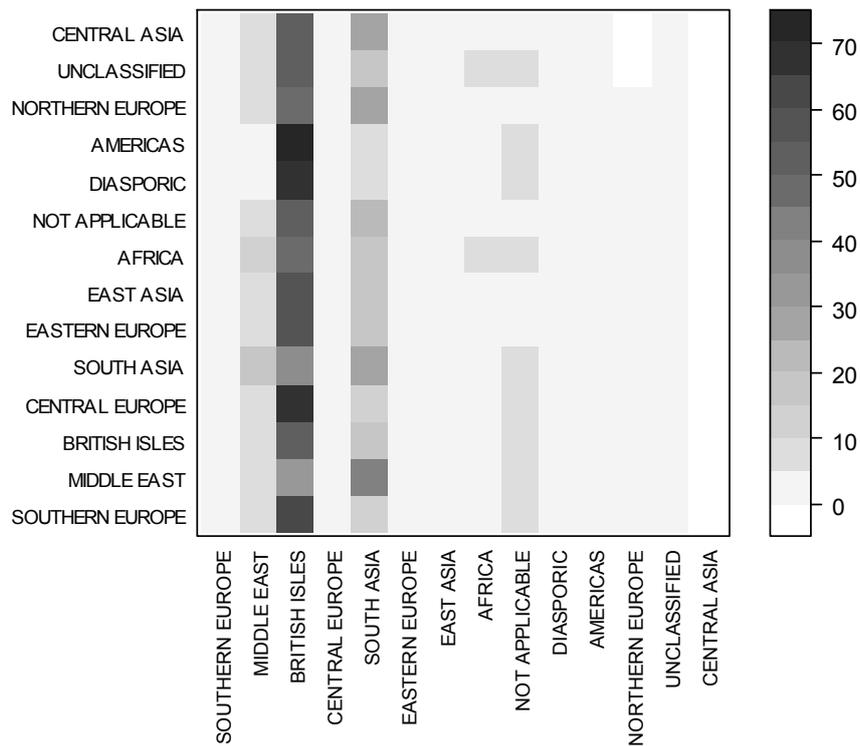


Figure 5-12 Normalised Links Based On Geographical Area

5.7.5.1 Comparison

In Figure 5-13 we have plotted the Silo Indices of the geography based groups for the reference and the null model. The South Asia group has the highest difference (0.38), and then the British Isles (0.17), the Middle East (0.09) and the African (0.09) come next in the list. These results rule out that the inter-group propensity for the high population (dominant) groups, such as the British Isles and South Asia, in both reference and null model, have the same propensity. There is a clear higher preference for inter-group linkages in the reference network. Using the same t-test, we found that there is a significant difference between these two cases with p-value < 0.05.



Figure 5-13 - Geography based Silo Index comparison between the reference and the null model

5.7.6 Language

Language classification is also supplied by Onomap. In our dataset, there are in total 74 language groups. The contracted group is shown in Figure 5-14:

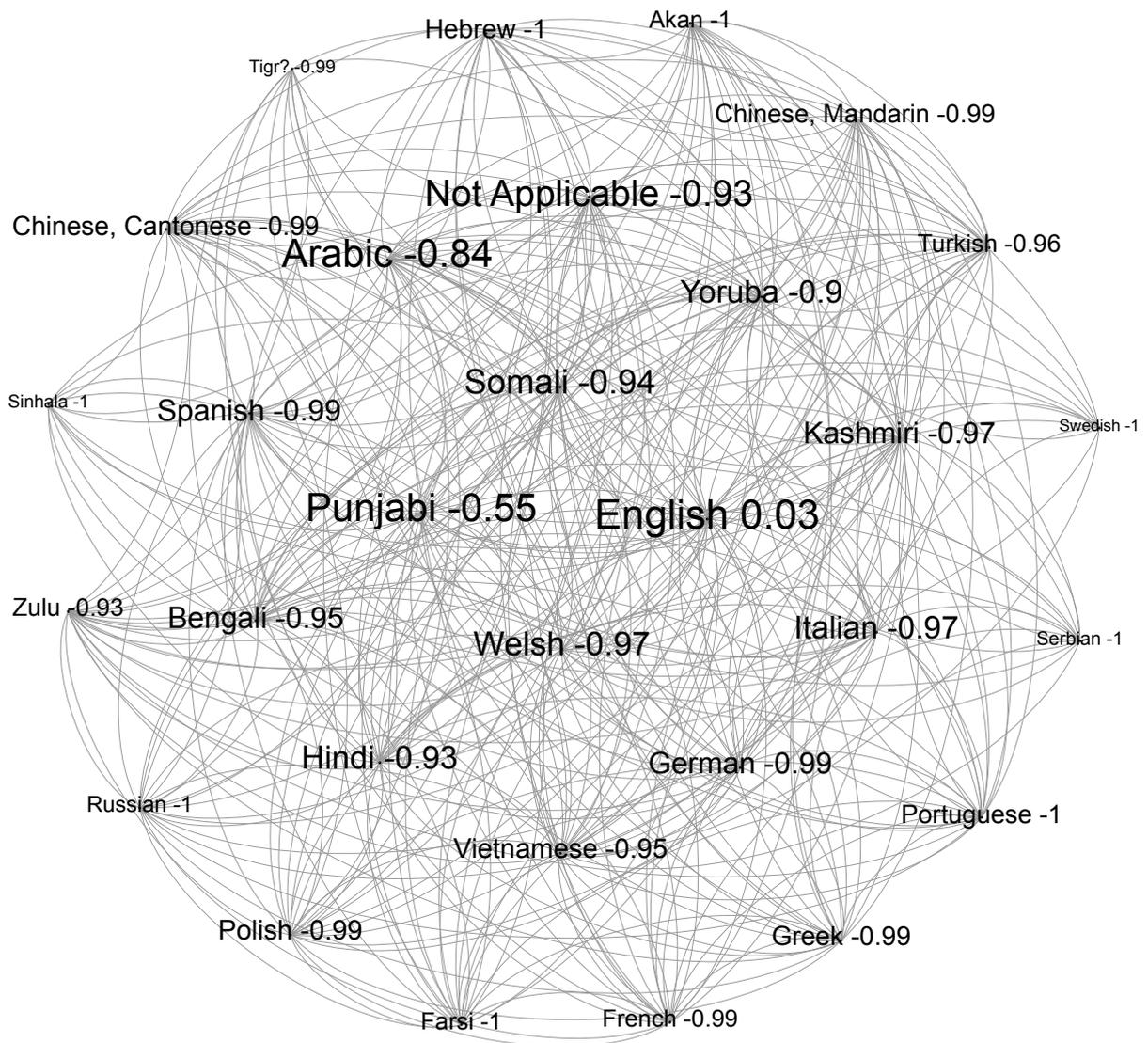


Figure 5-14 - Language Based Contracted Graph

Since there are a great many language groups, we have just shown the names of the groups with more than a weight of 1 in Figure 5-14. The dominant groups are: English (59.88%), Punjabi (11.51%) and Arabic (6.43%). The Silo Index of these three groups are: 0.03, -0.55, -0.84. The strongest inter-link weight (0.03) is between English speaking users.

To see how these languages are connected with each other, we plotted their links in Figure 5-15. Due to a large number of languages, we have removed the row/column names.

Like all previous plots of normalised links, we see that the dominant groups are clearly the most connected groups for all the groups. There are two languages, Slovenian and Nyanja, which have a 100% connection with the English language. These are due to them being the peripheral nodes in the overall datasets, and also due to their under-representation. There is only one user who has been classified as Slovenian and is connected with an English user. The same is true for the Nyanja users who are just two users and both are connected with the same English user. Ignoring these two extremely

under-represented groups, the Latvian group has the highest weighted link with the English language with a weight of 79.4.

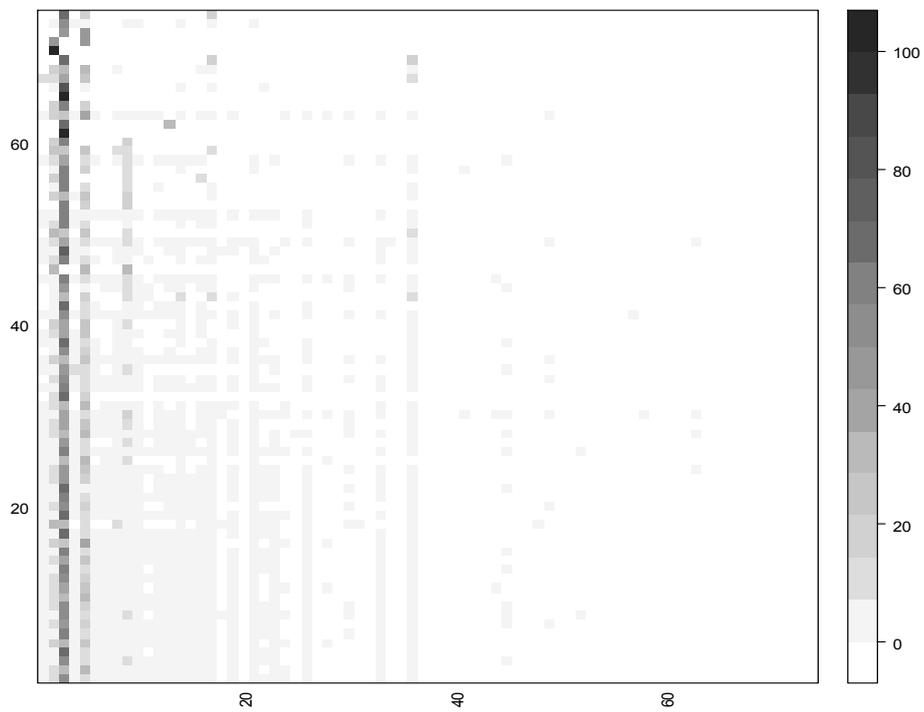


Figure 5-15 Normalised Links Based On Languages

5.7.6.1 Comparison

For the language attribute, similar to previous attributes, we have plotted the Silo Index difference between the reference and the null model in Figure 5-16. Since there is a large number of language groups involved which would clutter the plot, we have removed the names of the language groups. Forty two of those groups have no difference at all. In terms of statistical difference, we found that the p-value < 0.01 , when t-test is ran.

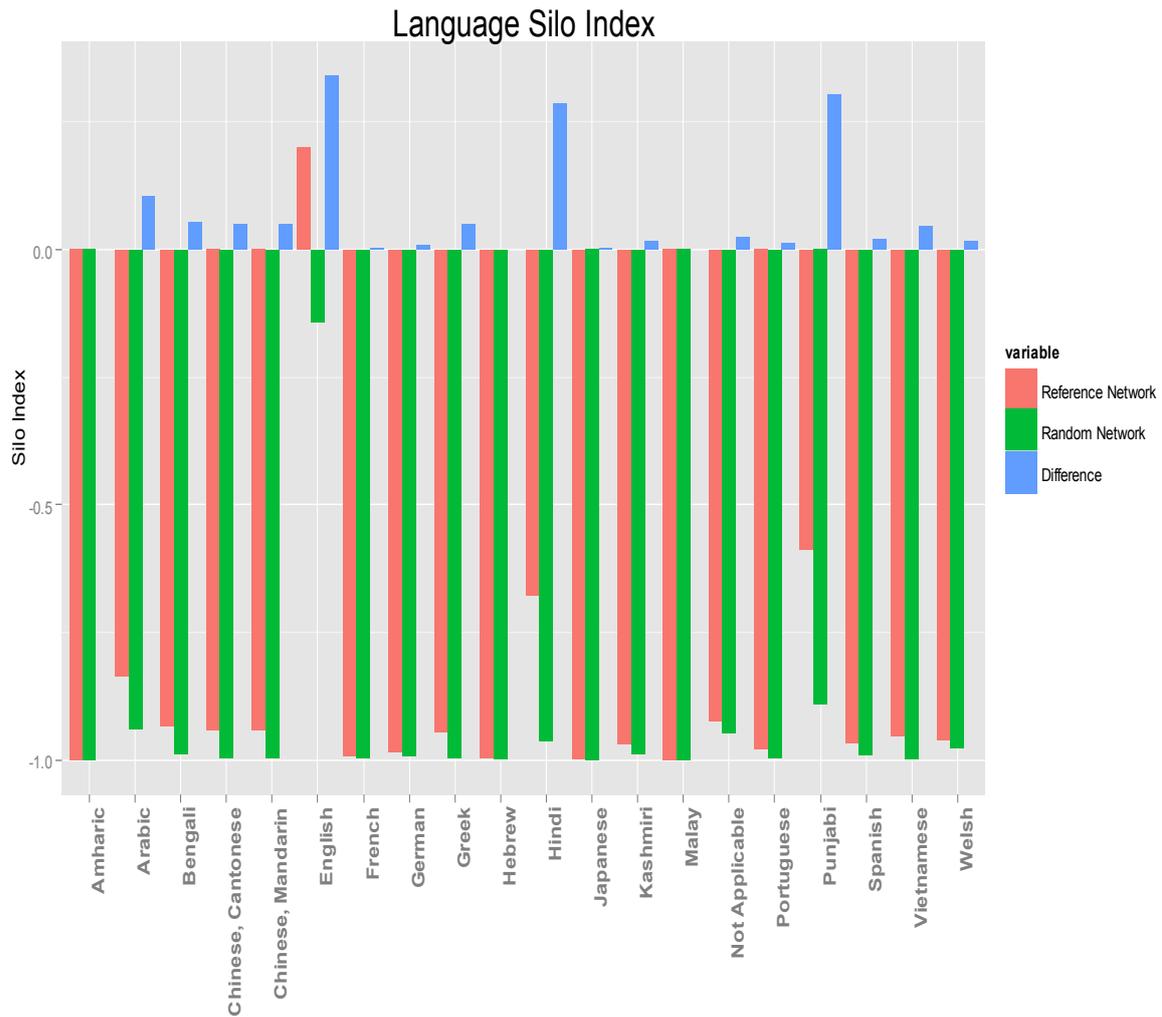


Figure 5-16 - Language Silo Index comparison between the reference and the null model

5.8 Normalised Results

We have already used a normalised adjacency matrix, at a group level, to establish how each ethnicity is connected with each other. This allows us to look at both dominant and non-dominant at the same level. We understand that the group size plays an important role in having opportunities. As an example, assume that there is a dominant group with over 90% of nodes and one minor group with 10% of links. If we randomise the links with an n average of links for all the nodes, the non-dominant nodes would not have the same opportunities as the dominant group to connect with each other. Their network is going to be more diverse than the dominant group. In order to deal with such issues, we have further investigated this matter, and came up with our own measure, called the *normalised silo index*, which takes into account the group size and the opportunities it has to develop inter and intra group links. In order to calculate the measure, let us firstly define how we calculate the number of internal and external links. For internal links, we take the count of internal links as the numerator and the total number of similar nodes as the denominator, as is shown in the Equation 5-4.

$$\text{Average internal links} = \frac{\# \text{ of internal links}}{\# \text{ of similar nodes}} \quad 5-4$$

Similarly we calculate external links by taking count of the external links as denominator in the Equation 5-5:

$$\text{Average external links} : \frac{\# \text{ of external links}}{\# \text{ of similar nodes}} \quad 5-5$$

Once both internal and external counts have been made, we take their difference as numerator and their sum as denominator to calculate a normalised Silo Index. The equation is shown below:

$$\text{Normalized Silo Index} = \frac{\text{Average internal links} - \text{Average external links}}{\text{Average internal links} + \text{Average external links}} \quad 5-6$$

This index takes into account both the size and opportunity of a group to develop links with other groups. Just like the Silo Index, the *normalised Silo Index* too, ranges from -1 to 1. -1 means all links are external (no same type links), and 1 means all same type links. We have applied this measure to all four dimensions: ethnicity, religion, geography and language. Below we have shown a normalised Silo Index for ethnicity in Table 5-12. The rest of the tables for religion, geography and language have been displayed in Appendix D. From all these tables we can conclude that most of the values in normalised Silo Indices are inward looking.

Table 5-12 - Normalised Silo Index for Ethnicity

Ethnic Group	Normalized Silo Index
Greek	0.89
Sikh	0.87
South Asian	0.87
African	0.86
East Asian & Pacific	0.84
Muslim	0.75
Hispanic	0.64
Japanese	0.56
Void	0.53
English	0.46
Celtic	0.42
European	0.39
Unclassified	0.34
Nordic	0.16
Jewish And Armenian	0.13
International	-0.04

After normalisation almost all groups, except internationals, have a positive Silo Index. It means there is more segregation than was apparent using a not-normalised Silo Index. With top three groups being: Greek (0.89), South Asian and Sikh (0.87), the Muslim group now is at the sixth position with 0.75. In terms of religious group, we find Hindu (0.9) to be at the top. The Muslim group has fallen to the fourth position with 0.75.

It can be observed that these results drastically change the amount of segregation inferred. Most of the groups based on ethnicity, religion and geography have a positive Silo Index, which shows segregation at a group level. For language, out of 82 groups, 36 of them had non -1 values, but are all positive (ranging from 0.99 to 0.18). This further strengthens the conclusion that most of the groups (ethnic, religious, etc.), are more linked with themselves than others. For language, in one of the cases (Vlaams), the normalised Silo Index is 1. This comes out as a rounded value. The actual value is 0.9984898. In all cases, the extreme values of 1 or -1 do not really exist. It is a by-product of rounding the values to two decimal points. To reiterate, Silo Index with values -1 and 1 represent the extreme cases (no in-group links to only in-group links respectively).

5.9 Summary of Results and Discussion

This analysis is used to evaluate and understand the diversity of Facebook users. In an [SNS](#), such as Facebook, the users are given the right to hide their personal information such as age, gender and even their friendship links, but the screen name remains visible. We found that Facebook is unique when it comes to names, for instance, when it is compared with MySpace. In a small Facebook study (of sixty nine Facebook users) carried out by Dwyer et al. (Dwyer et al., 2007), it was shown that every user (100%) of Facebook revealed their real name; also the same study showed that people are more trusting of Facebook than other [SNSs](#). This increased the confidence of our approach. We tried to make full use of this information by estimating user's ethnicities with a name-based ethnic classifier, Onomap. Based on these estimations and the social graph, we analysed the whole network. This included classification on religion, language and geographical area as well. For a better understanding of each identified group, we also identified the underlying communities and also calculated inter and intra links between various groups within each attribute.

Our dataset certainly does not represent the whole Facebook network, but the degree distribution of our visited nodes (those which have their ego network crawled) shows that there are two regimes of power-law effect, as was shown in Minas et al. (Minas Gjoka, Kurant, Butts, et al., 2009), for node degrees greater than and less than 300. The ethnic distribution, however, does not match with the whole Facebook network. As for the behaviour of various ethnicities, we see a clear difference between the dominant and non-dominant groups. Dominant groups not only have a high number of average links, but they are also closely linked with themselves, unlike the non-dominant groups. Apart from just one case in language classification, the English group has a positive Silo Index. The rest of the Silo Indices of all attributes are negative.

In terms of ethnic classification by Onomap, there are a number of points pertaining to non-European classes. The diversity and heterogeneity present in bigger countries is somewhat amiss. For Britain, if we look specifically at white Britons, there is a high number of diversity than say a non-white Muslim Briton. Groups such as Celtic refer to Irish, Scottish and Welsh, whereas there is a different English class too. For a non-white minority group in Britain, such as Pakistanis, there is only one group 'Muslim'. This issue can be reduced if we look into Onomap's sub-ethnicity 'Pakistani'. For Pakistan, being the

sixth largest country in the world and with fifty languages, it does not capture the multitude of heterogeneity present in the society. There is two major religious divides, and then there is a language divide. This issue of clumping over 1.6 billion Muslims all around the world into a Muslim group reduces the complexity of diversity and heterogeneity. Whereas we find that for both nationality and ethnicity, Europeans and Christians have been sub-categorised into various groups. This puts Muslim groups at peril. Also our analysis does not support multi-attribute comparisons. For instance, if we look at Pakistan and India, two of the most diverse countries in the world, whilst the Punjabi language is spoken in both countries Pakistan is primarily a Muslim country, where most speakers are Muslim, compared to India, where most of the Punjabi speakers are Sikh. When we look at the combined language group, we see that both Pakistani Muslims and Indian Sikhs are combined together into one group: Punjabi. Onomap, it is best to say, works very well for European populations where not only is there a country divide, but a religion divide as well (quite a few Christian denominations are present). This allows a more in-depth analysis of various groups at a finer scale. This level of analysis, for minorities in the UK who had come from all over the world, is not well-suited.

In terms of ethnic distribution, we provided confidence by comparing the reference network's ethnic distribution with that of [MMU](#)'s students. Before we consider specific details, let us focus on the [MMU](#)'s diversity. We found that there was almost a 5% difference of Asian students between [MMU](#)'s and other higher education institutions in the UK (11.24% versus 6.38%). This shows that [MMU](#) is doing fairly well to induct Asian students. Asian students, however, do not include Chinese students – it is a separate ethnicity. For our comparison, with a somewhat crude manner, we grouped the Onomap ethnic groups into similar groupings used by [MMU](#) (see Table 5-8 and Table 5-9).

The biggest over-representation of Asians in the dataset is more than double (24.18% versus 11.24%), which comprised of Muslims, Sikhs and South Asians. One should note that Muslims in Onomap classifications also involve non-Asians, such as Somalians and Eritreans. Both over and under representation has been adjusted by using contracted graphs by normalising weights across all groups. The ethnic distribution of our dataset includes that of the general public as well. Just to reiterate, our methodology was to obtain a social network of [MMU](#) students, which might include people not affiliated with [MMU](#) – either because they have not mentioned it in their profile or they belong to the general public. For [MMU](#) specific results, one way forward is to focus on an individual student, by focusing not only on their ethnicity, but also their social networks. In other words, instead of the whole network, we break this network into a unique number of visited student's social networks, (4061 to be precise) and then see both the inter and intra ethnic propensities. This would allow us to also capture heterogeneity, which might exist in each student's social network. Also this line of work would allow us to draw a few policy recommendations for [MMU](#).

In order to capture inter and intra group linkages of our dataset we contracted our overall graphs into attribute based graphs. This means we calculated ethnicity (Onomap Group), religion, language and geography based graphs and their adjacency matrices. In all the contracted graphs the preferences of non-dominant groups vary among dominant groups. We also calculated how each group is connected with itself with the help of the Silo Index.

For the ethnicity graph, we found overall sixteen groups, which is the same number of ethnic group Onomap operates with (please see the subsection on Onomap 5.4.1). This gives us a hint about the diversity of our dataset. We found two classes of groups: dominant and non-dominant. The non-dominant groups are mostly connected with the

dominant groups. As for the dominant groups, they are either connected with themselves or with other dominant groups. The three dominant groups are: English (45.8%), Muslim (21.02%), and Celtic (16.42%) groups. The rest of the groups are non-dominant groups. In terms of inter-group propensity, by using the Silo Index, we found that the Muslim group had the highest propensity (-0.21). However when we normalised the Silo Indices, the Greek group became the most inward group with 0.89, and almost all groups have an inward outlook. The Muslim group drops down at the sixth position with 0.75. There is clear cohesiveness, for instance, as the International, Nordic, Sikh, South Asian and Muslim groups have the most number of links with the Muslim group, while the rest of them are mostly connected with the English group. In the normalised matrix, the highest weight we found was 60%. This was from the Celtic group to the English group. As for the top 10% of the links in the adjacency matrix, we found that the English group covers 64% of them, whereas the Muslim and the Celtic group cover 32% and 4%.

As for the religion graph, we found twelve groups with four major groups: Christian: Protestant (51.48%), Muslim (21.14%), Christian (9.28%) and Christian: Catholic (7.83%). In terms of inter-link propensity, again the Muslim group stood out with the highest value of Silo Index (-0.21). The Christian group had the highest linkages with the Christian Protestant group (63.5). For the top 10% of linkages we found that the Christian Protestant group covers 64.2%, whereas the Muslim group covers 28.5% and the not-applicable group covers 7.14% of the links.

According to the geographic classification, most of the users (62%) are assigned to the British Isles group, followed by the South Asia and the Middle East group (16.14% and 7.04%). These three groups are major groups. The highest weight, 71.3%, was found from the Americas group to the British Isles group. The top 10% of links are covered by the British Isles (73.68%) and the South Asia (35.71%) groups.

For a language contracted graph, we found seventy four groups, with three major groups: English (59.88%), Punjabi (11.51%) and Arabic (6.43%). The Silo Index of these three major groups are: 0.03 (English), Punjabi (-0.55) and Arabic (-0.84).

Individual propensity when ethnic, religious, language and geographical attributes are quite strong for both dominant and non-dominant groups. This provides confidence for high clustering/segregation of individuals into strong communities. After comparing the reference network with the null model, we can easily identify that these inter-link [homophily](#) (or segregation) and clusterings are quite pronounced in the reference network because of its network structure. We can rule out that these results have a greater impact merely on the population distribution (for instance high population of English group in ethnicity), hence we rejected the null hypothesis which states:

H0: The Facebook network does not segregate on ethnic lines and is not highly clustered on ethnic lines.

Clearly, based on our reference network, Facebook is divided into strong communities when the ethnicity (or religion, language or geographical) attribute is taken into account. This confirms both of our hypotheses, which state that:

H1 (a): The Facebook network is segregated on the ethnic lines;

H1 (b): The Facebook network is highly clustered on ethnic lines;

5.10 Conclusion

Facebook is one of the mainstream [SNSs](#), which has become part of everyday life. In this chapter, we tried to disentangle the relationships which people have on Facebook, when their ethnic, religious, language and geographical groups are taken into account. This involves the approximation of one's ancestral group, producing valuable insights when ethnicity, linguistic or religious data are not available at appropriate temporal, spatial or nominal (number of categories) resolutions (Mateos, P, Webber, R and Longley, 2007). We analysed the diversity of [MMU](#) students.

We are confident about the size of our dataset when we look into the distribution of its social network. For instance the degree distribution of our visited nodes (those which have their ego network crawled) highlighted that there are two regimes of power-law effect, as was found by Minas et al. (Minas Gjoka, Kurant, Butts, et al., 2009), for node degrees greater than and less than 300.

In terms of our hypothesis, we clearly found clusters when ethnic, religious, language and geographical areas are taken into account, but the most clear divisive group is the English group, when the language attribute is considered. In terms of segregation, we do not find full segregation, but dominant groups do have very high numbers of inter-group friendships, which strengthens our confidence in inter-group [homophily](#). The [affinity](#) measure, which calculates [homophily](#) between nodes sharing the same attribute, has been applied to both the reference Facebook and the Random network (the null model). This measure provides confidence, at the macro level, for both clustering and segregation of the social network. Ethnicity comes as the most important attribute (1.5) with religion coming second (1.4). These values, for all the four attributes, normalise to 1 for the null model, supporting the fact that the network structure of reference graph shows greater personal preference of inter-ethnic friendship links. This was found in all four dimensions of users' attributes: religion, ethnicity, language and geographical area. Hence we reject our null hypothesis which states that neither clustering nor segregation exists in our dataset.

Our dataset is quite diverse in nature, but some of the ethnic groups are over and underrepresented. In order to deal with dominant and non-dominant groups, we normalised the weights between them, so that all groups can be compared on the same level. We also believe that our starting point of crawling has a vast impact. We started crawling from a Muslim user, hence we see so many Muslims present in our dataset. In order to overcome this issue, for all intra and inter-ethnic/religious/language and geographical groups, we normalised the propensities of each group to compare them on the same grounds.

There are many offshoots from our work. The first is to study the detailed ethnic classification represented by Onomap Subgroups in our dataset for a broader understanding. Also, we intend to analyse whether the ethnic breakdown of users can be divided on the identified geography or not. We would also like to do multi attribute comparisons for better insights as well.

6 Chapter: Distributed Peer to Peer Social Network System

6.1 Introduction

In this chapter we are going to cover technical details and challenges of a completely decentralised Social Network System ([SNS](#)). This is, one should clarify, not a sophisticated system like Facebook, but rather a proof of concept system. We will start off with the rationale behind our work and then go on to talk about the technical details involved in both the development and also the deployment of such a solution.

As discussed in previous chapters, Social Network Systems ([SNSs](#)), which allow users to create identities and link them to friends who have also created identities, are highly popular. Systems such as Facebook and Twitter⁷ utilise a traditional client-server approach to achieve this, which means that all identities and their social links (the entire social network) are stored and administered on central servers. Although this approach supports highly mobile user access, users can log-in from any computer, it also implies high dependence on predefined central server(s), which results in the possible exploitation of private data.

In this chapter we present an alternative approach which uses a completely decentralised peer-to-peer system to create and store the social network. Our approach is based on a gossip protocol for discovering potential peers as friends. Our system is self-administered and works in a highly transient environment of peer availability. It is not a sophisticated system, but rather a 'proof of concept' system. We propose the design and implementation of a distributed Social Network System ([SNS](#)) that is scalable and robust, allowing users to perform core social networking functions of establishing and removing social links without any requirement for centralised servers or administration.

Before we describe how our solution works, we need to define what a Peer-To-Peer Network is.

6.2 Peer To Peer Networks

Peer-to-peer networks are a type of distributed network with a decentralised architecture. The participating computing entities or nodes in other words, are called peers. They act as both *client* and *server* in the network. Schollmeir et al. (Schollmeier & Universitat, 2001) defined the nodes of [P2P](#) network as 'Servent' which has been derived from the first syllable of the term *server* ("Serv-") and the second syllable of the term *client* ("-ent"). Thus this term Servent represents the capability of the nodes of a Peer-to-Peer network of acting at the same time as a server as well as a client (Schollmeier & Universitat, 2001). We have shown its architecture in Figure 6-1 below. It shows that all participating peers are at the same level. There is no hierarchy.

⁷ www.twitter.com

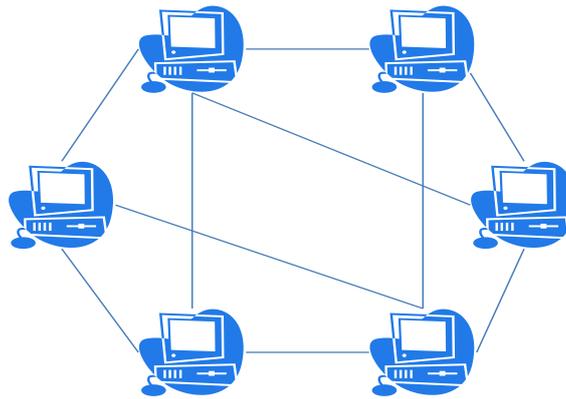


Figure 6-1 - Peer-To-Peer Network

6.3 Background

One of the recent trends in the cyber world is the emergence of Social Network Systems ([SNSs](#)). An increasing number of people are flocking towards these systems and engaging in new group-based social activities. With Facebook and Twitter being in the top ten of the most visited websites in the world (Alexa, n.d.), a huge potential and [affinity](#) of people towards social network can be seen.

The impact of online [SNS](#) has been tremendous. However, currently these systems depend on a centralised architecture and are therefore prone to become the victims of possible exploitation by central authorities. Also, being centralised systems, they are highly dependent on centralised entities with complete authority.

In this chapter, we propose a design and initial implementation of a decentralised [SNS](#) based on a gossip protocol, which means peers periodically pick another random peer from the network to exchange data with (Voulgaris, 2006). Gossip protocol is used because of its light-weight nature and also high scalability characteristics. Our system establishes friendship links among peers under dynamic conditions of peer availability. As has been observed in [P2P](#) systems, be it structured or unstructured, the rate at which peers join and leave the system, or in other words, the churn rate, is very high (“Handling Churn in a DHT,” n.d.). We have developed our system which establishes friendship links among them in such a dynamic environment. The possible peers which may become friends are discovered in a gossip fashion. Our system is self-administered and does not depend upon any central entity, which means that all of the social network is handled by the peers themselves. The notion of friendship link establishment in our system is the basic building block of forming the [SNS](#). On top of this structure, many applications can be developed. Currently, in our Tribler system (Pouwelse et al., 2008), cooperative downloading (Garbacki, Iosup, Epema, & van Steen, n.d.) is one of the applications using friendship links by making use of idle bandwidth of one’s friends to boost one’s download performance.

The structure of this chapter is as follows. In Section 6.4, we discussed the functionalities of our [SNS](#) provided to the users, along with the concepts already part of Tribler. In Section 6.5, our detailed design is presented. Section 6.6 discusses the evaluation of our [SNS](#) with experiments we carried out with the deployed system. Possible

attacks and their prevention in our system are discussed in Section 6.7. Critical analysis of our work is covered in Section 6.8. Future work is presented in Section 6.9 and related work in Section 6.10. The chapter ends with the conclusion in Section 6.11.

6.4 Requirements of our System

In this section, we will discuss the requirements for the [SNS](#) we want to design. First, a set of functionalities provided to the users are discussed. These functionalities have been taken after analysing five prominent [SNSs](#) which are: Friendster, Orkut, Facebook, MySpace and LinkedIn (Abbas, 2009). After that, we highlight the basic concepts of Tribler which are relevant to our solution.

6.4.1 Functionalities

In this section, the functionalities provided to the users of our [SNS](#) are going to be listed. As explained in the above section, these have been drawn after studying numerous [SNSs](#), such as Facebook and Myspace. Essentially they are the core requirements of a typical [SNS](#). Below, the terms 'user' and 'peer' are used interchangeably. In our [SNS](#), the peer who initiates a friendship request to another peer is known as the source peer, and the peer for whom this request is intended is known as the target peer. The functionalities provided to the users are the following:

- a) Adding new friends: In order to build a social circle, a peer can request other peers discovered by the underlying peer sampling service (PSS), which are potential candidates for being friends, to become their friends. The target peer has to reply to the friendship request sent by the source peer, and if the reply is positive, both the peers become friends.
- b) Removing friends: The source peer removes the target peer from its friends list. Also, it requests the target peer to remove it from its list.
- c) Maintaining status of friends: The system must keep peers up-to-date about the online status of their friends.

6.4.2 Tribler

One of the most prominent protocols for peer to peer ([P2P](#)) systems was developed in 2003. It is called Bittorrent (Cohen, 2003). The major success of it lies in it tackling two major issues (Rahman, 2011), which earlier [P2P](#) protocols suffered from, which are: freeriding (Adar & Huberman, 2000) and spam control (Liang, Kumar, Xi, & Ross, 2005).

Freeriding means that users do not contribute – they merely download data and do not share theirs (upload) with anyone. As for spam, it represents the spread of malicious content by some peers (Rahman, 2011). In the Bittorrent protocol, using Tit for Tat (TFT) strategy, peers upload to those others who reciprocate to them the most. This ensures that peers who upload less, get less in return (Rahman, 2011). That is how the freeriding problem is handled. As for the spam problem, Bittorrent solves the problem by avoiding it: content location and dissemination are not part of the Bittorrent protocol (Rahman, 2011). In order to download a file, one needs a metafile called torrent (.torrent files), which contains the information needed to start downloading. Dissemination of these torrents can be achieved by various means such as emails or a website. There are specialised,

centralised websites called trackers which publish torrents, and also act like a bootstrap server that provides newly arriving peers with addresses of other peers. Peers then share pieces of a big file with each other based on TFT, using which they prefer as the fastest uploading partners.

The implementation of our [SNS](#) has been done in Tribler, which is a Bittorrent based file-sharing client. A little background on important and relevant concepts of Tribler is set out below.

In Tribler, peers have a permanent identifier (PermID), which is based upon public-private key pairs. A peer, the challenger, can challenge another peer, the challengee, for its identity by generating a large random number. The challengee encrypts it with its private key, and then the challenger decrypts the result with the public key of the challengee. If the result of this decryption is the same as the original random number, the authentication succeeds.

Tribler has an epidemic protocol called BuddyCast for peer and content discovery services. In Buddycast, peers exchange messages with random peers (exploration) and semantically close peers called Taste Buddies (exploitation). After a pairwise exchange, both the involved peers merge their lists of peers and then rank them according to their preference list similarities. They both retain only the top N best ranked peers. The notion of Taste Buddies, or semantically close peers, helps to reduce the randomness of peers, which eventually leads to better content searching results. Peers take care not to contact the same peer for the next four hours.

The contextual information based on the communication through Buddycast among peers is stored in each Tribler peer in a local database known as the Mega-Cache. It stores information about peers, torrents, and preferences (list of recent downloads). This is then used by Tribler to gossip, calculate similarities, and recommend torrents.

6.5 Detailed Design

In this section, we are going to explain how a friendship link between the source and the target peers is established in a very dynamic environment in which peers may go offline at any time. In 6.5.1, we will present the basic mechanism of the friendship link establishment, which is based on the request-reply concept, between the source and the target peers. In Section 6.5.2, we will discuss the underlying retry mechanism and also the notion of the helpers with their role. Based on the target and the source peer availability, we will present the friendship link establishment scenarios in Section 6.5.3.

6.5.1 Basic Request-Reply Protocol

For establishing a friendship link, the mechanism follows the request-reply notion. The source peer initiates it by sending a friendship request to the target peer. The target peer then takes its decision by accepting or rejecting the friendship request and sends its reply back to the source peer. If the reply is positive, both the source and the target peer become friends.

6.5.2 Unavailability of the Peers

In order to deal with the unavailability of both the source and the target peer, we have designed two mechanisms, which work for both friendship requests and friendship replies, which we discuss below.

1) Retry: If the target peer is not online, the source peer will retry to connect to it in five minute intervals thereafter, in case the target peer comes back online. Similarly for receiving the reply from the target peer, if the source peer is not online, or unconnectable for some reason, the same retry mechanism is adopted by the target peer to dispatch its reply to the source peer. The initial retry time interval of minutes is increased to twenty four hours, after one day has passed since the friendship request/reply was initiated. After a week of unsuccessful delivery of requests or replies, all pending friendship messages (requests and replies) are dropped from the source and the target peers.

In order to increase the chances of contacting the other peer, both the source and the target peers save messages that could not yet be successfully delivered, i.e., the pending messages (requests and replies), in case they are going offline. In their next session, both of them read these messages and then dispatch them. We present the retry mechanism in Figure 6-2:

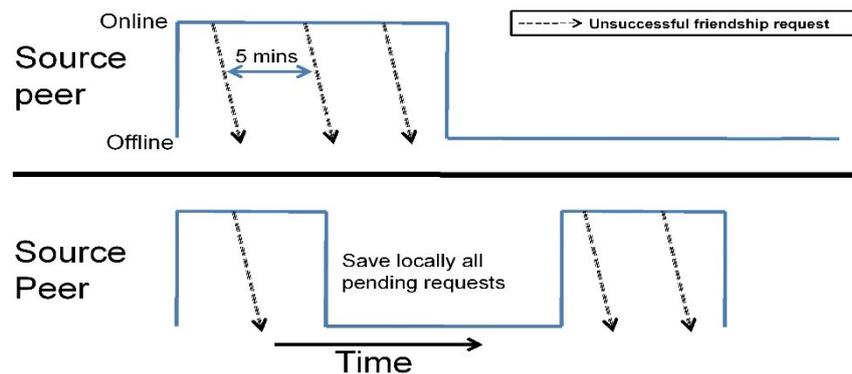


Figure 6-2 - Friendship request retry mechanism

2) Helpers: To increase the chances of establishing a friendship link between the source peer and the target peer, we have introduced the concept of helpers. Helpers are online friends and taste buddies of the source peer, in case of friendship requests. And in

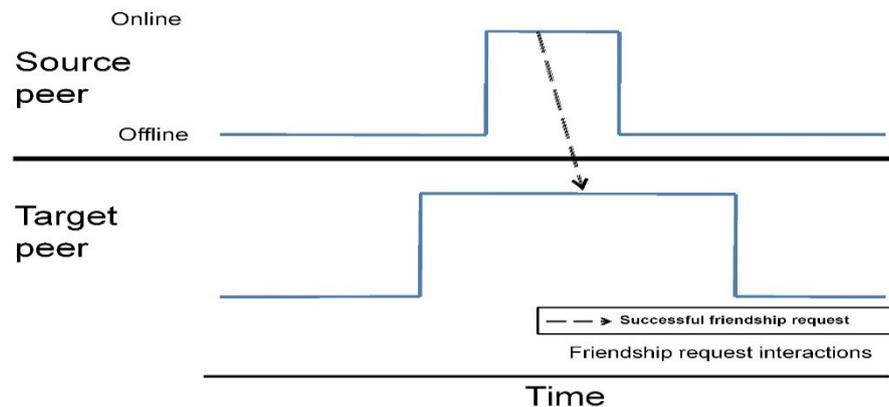


Figure 6-3 - Friendship request scenario 1: Both the source peer and the target peer are online

the case of a friendship reply, they are online friends and taste buddies of the target peer. When the source peer is unable to connect to the target peer for requesting friendship link establishment, it dispatches its friendship request to these helpers. Helpers then also try to contact the target peer every five minutes. Helpers also used by the target peer for forwarding its friendship reply to the source peer, in case it is unable to contact it. Helpers, just like the source and the target peer, also save the unsuccessful friendship requests/replies locally when they are going offline. On their next start up, they try to deliver them to the intended peer.

6.5.3 Scenarios for Establishing Friendship Links

Depending upon the availability of the source and the target peer, we distinguish different scenarios for establishing a friendship link between them. Note that these scenarios only show the friendship request part. The reply part follows the same scenario.

The possible scenarios of friendship link establishment between the source and the target peer are the following:

6.5.3.1 Scenario 1

Both the source and the target peers are online. The source peer directly sends the friendship request to the target peer. Depending upon the target peer's response, it is added to the source peer's friends list. Figure 6-3 above shows the interaction between the source and the target peer.

6.5.3.2 Scenario 2

The source peer is online, but the target peer is not. The source peer after an

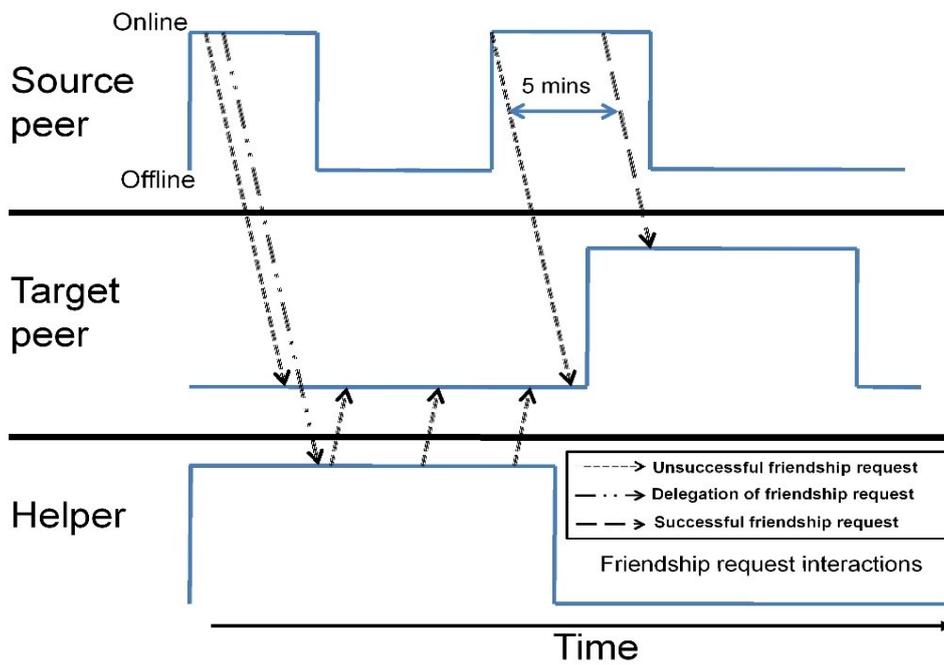


Figure 6-4 - Friendship request scenario 2: Both the source peer and the helpers try to contact the target peer.

unsuccessful attempt to connect to the target peer, employs the retry mechanism mentioned above, involving both itself and the helpers. This scenario is shown in Figure 6-4 above.

6.5.3.3 Scenario 3

The source peer has gone offline after initiating the friendship request, but the target peer is online. Since the source peer cannot connect to the target peer, it dispatches the friendship request to its helpers. The helpers then connect to the target peer and forward the friendship request to it. This interaction can be seen in Figure 6-5.

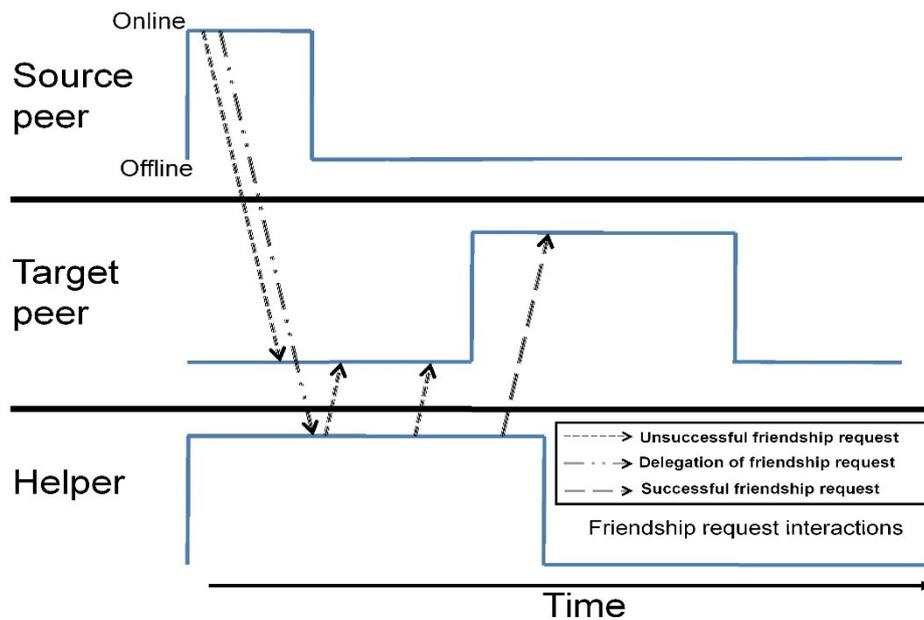


Figure 6-5 - Friendship request scenario 3: On behalf of the source peer, the helpers relay the friendship request to the target peer

Figure 6-6 below shows a screenshot of a friendship establishment request in Tribler.

6.6 Experiments

In this section, we will show the success rate of our system with the help of reliability experiments. Success here means that the source peer indeed gets the reply from the target peer on its friendship link establishment request, regardless of it being positive or negative. Our system was deployed in Tribler 4.5 released on November 11th 2008. The updated version was then rolled into our website, which was downloaded all over the world. We did not send out invites to anyone to use the [SNS](#) features in the latest version. People used our service of their own accord, on top of Tribler. To record friendship establishment related statistics of our [SNS](#), we have developed a crawler, which is a specialised [P2P](#) client, and which is being run on one of our servers at TU Delft. According to the crawler's statistics, the overall population of our system was 535 peers until 20th November 2008, which means that the collected data covers ten days. Based on the list of peers supplied by the underlying PSS (Buddycast, in this case), this crawler connects with every peer which comes online and asks it to supply all the friendship requests it has made so far. In case of stumbling upon the same peer after a while, it asks for new friendship requests since the last encounter and also updated friendship records which have now received replies. All clients save their friendship requests and replies are recorded in their local database. The analysis in this section has been made on this retrieved data. In the first result in Section 6.6.1, we present the number of friendship requests made by all peers we contacted and the fraction that were successful. In Section

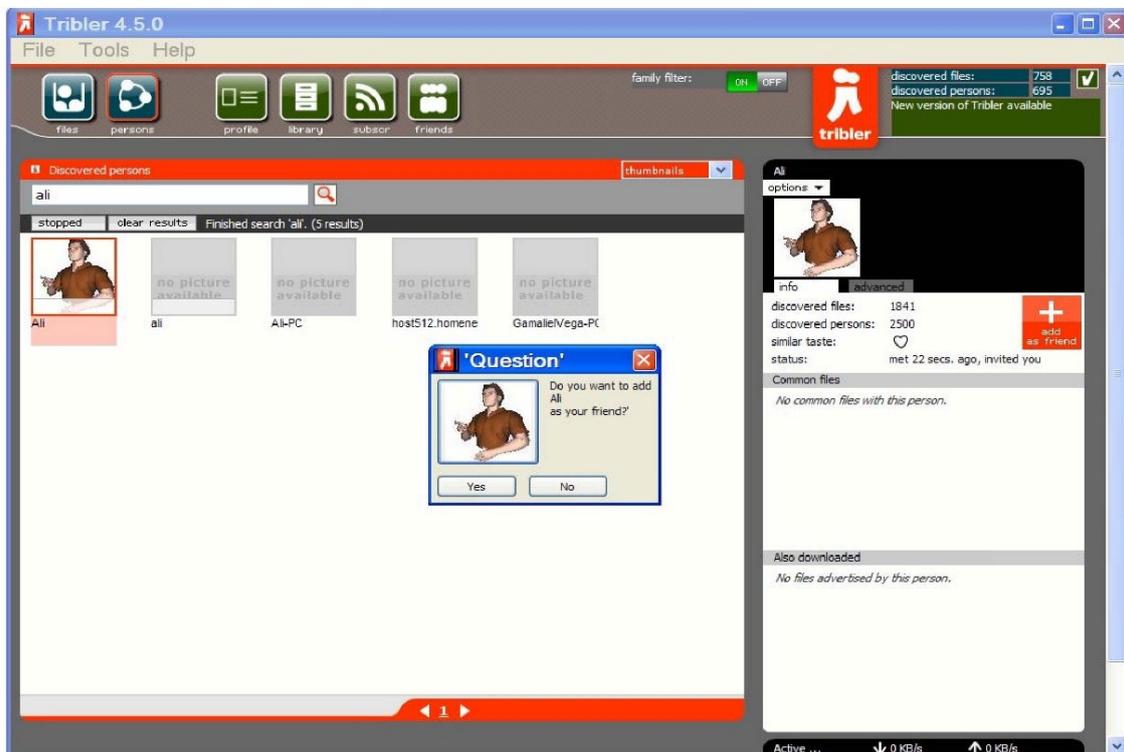


Figure 6-6 - Screenshot of friendship link establishment

6.6.2, we show for each of the friendship link establishment requests, how much time was taken for its reply.

6.6.1 Number of Friendship Link Establishment Requests

Over a period of ten days in total, there were 588 friendship link establishment requests made by 132 peers. Out of those requests, 191 were successful, resulting in a success rate of 32%. There are several reasons for such a low success rate. The target peers may not have come online after the source peer initiated the request. Even if they did, they were not online at the same time as the source or helpers were, or the request has expired. As mentioned in Section 6.5.2, after a certain time period, i.e., a week, all pending requests are expired. Figure 6-7 shows this result. Grey bars represent the total numbers of the requests made by peers and the blue bars represent how many of them were successful.

In order to know what is the actual success rate, we have parsed the log files on our super peers used by Tribler, which record information of all peers (Tribler clients) when they come online to get a list of fresh peers through Buddycast. Out of the 588 requests, the target peers mentioned in 298 requests never came online throughout the crawling phase. In addition to that, in 39 requests, the involved target peers were not seen during the complete period from when the source peer initiated the friendship request until the expiry of the request, i.e., after seven days. That means, out of 588 requests, only 251 (588 minus 298 minus 39) represent the total friendship requests. Keeping this figure in mind, now the overall success rate becomes 76%.

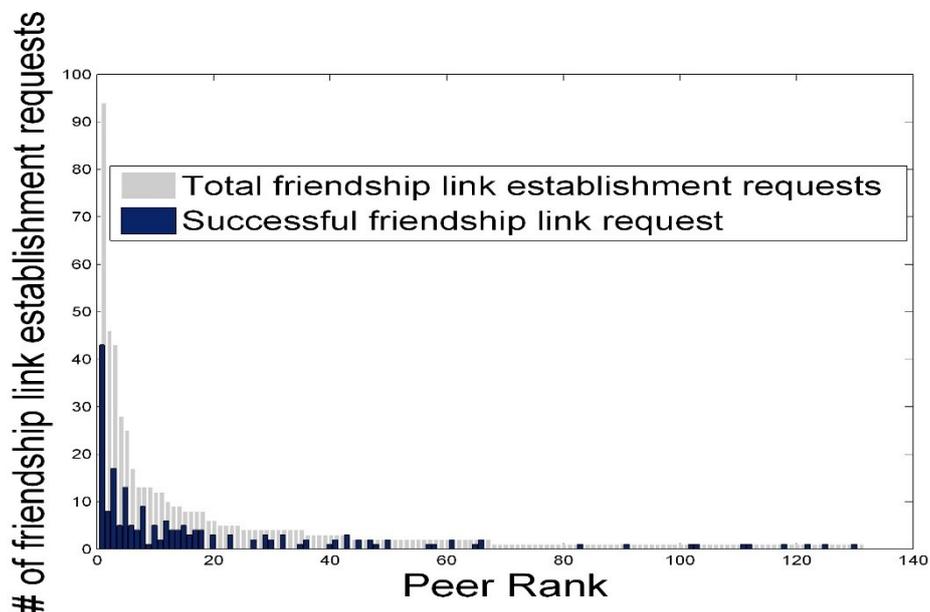


Figure 6-7 - The number of total and successful friendship link establishments.

6.6.2 Total Time Taken for Receiving Friendship Replies

The total time used in requesting and then receiving a reply on friendship link establishments can be seen in Figure 6-8. We show the results of friendship link establishment requests with both positive and negative replies. For all the successful 191 requests, it shows the histogram of the time in minutes taken to make them successful. On average, almost 128 minutes were taken to get the response.

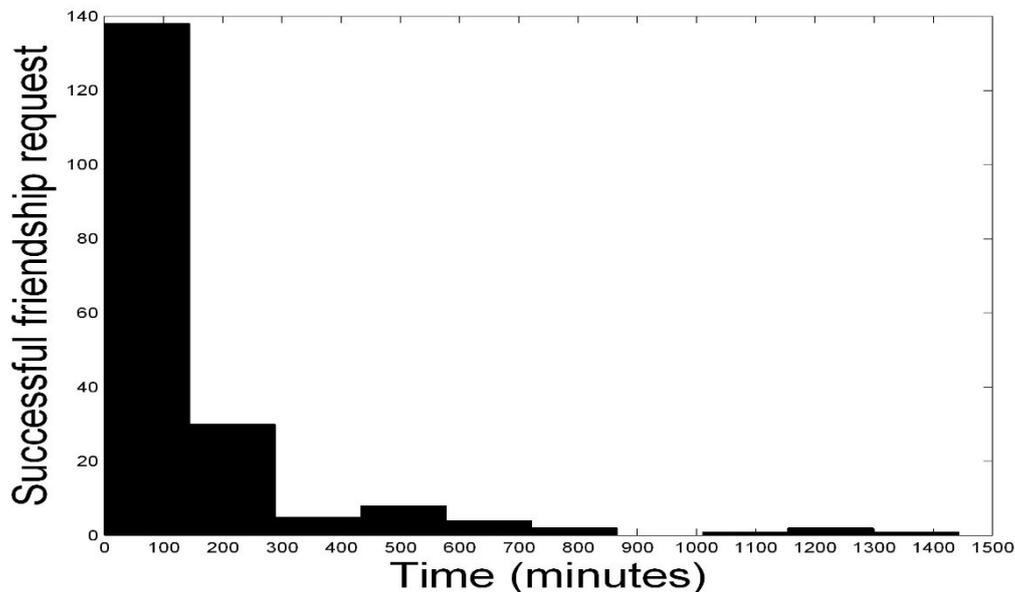


Figure 6-8 - Histogram of the time taken for each successful friendship requests.

6.7 Prevention of Possible Attacks

In the current design, it is possible for malicious peers to target and subvert the system. There are two main possible attacks, which are Distributed Denial of Service (DDOS) and a special DDOS, 'man in the middle attack'. In order to thwart such potential attacks, we have established certain safeguards which we shall detail below along with an explanation of the attacks.

DDOS is a type of attack where a peer is asked by a huge number of other peers for some service. The motive behind this attack is to overload a peer so that even legitimate peers are unable to access it and get its service. For the DDOS attack, we restrict a user, who is running the client from a binary, to make at most ten friends per day. We cannot, of course, overcome this problem if a user has modified our source code. This restriction can be accomplished fairly easily as all the friendship requests will be recorded by the system.

In the 'man in the middle attack', a helper tries to overload a peer, or a group of peers, with a huge number of illegitimate friendship requests. To counter this, we have devised a solution by incorporating the use of signed requests: the peer who initiates the request (source peer) first signs it with its private key. This would allow the receiver (target peer) to determine that it is indeed from the source peer. Since only one instance of a Tribler

client can run on a single machine, no malicious peer can fake or develop multiple instances, and thus multiple identities.

In the Tribler eco-system, each individual is holding their profile, which puts the responsibility of identity creation and maintenance of self (by means of a profile) to everyone. ID theft is not possible. It works better for users to have control over their information however, once a user is offline, his/her friends will not have access to it. Usually there is a very high churn-rate in a [P2P](#) system, but since we are focussing on development, interaction and maintenance of a social network by users, the behaviours of them will adapt to remain more active and connected to the system – indirectly solving the churn rate problem. Looking at the average amount of time people spend on Facebook (around thirty minutes a day), that will, we hypothesise, increase the availability of users.

6.8 Critical Analysis

We have developed a proof of concept system showing how, by using decentralised architecture, users can develop their social network. This is in no way a sophisticated system which could provide all the usual services provided by a typical [SNS](#) but it is a step towards a user liberating system by enlisting the technical details and challenges one might face. We have tried to show the success rate of our system on real users, in order to determine the best way our system would work. One of the most interesting ways forward is to simulate our system which will allow us to check the strengths of our solution for a much bigger user base and would also help us design a better and more secure system. This will also allow us to identify what the best time frame is for friendship retry (which is currently a week).

As we have discussed earlier, with completely decentralised systems, a user faces a dilemma which is succinctly covered by Feldman et al. (Feldman & Blankstein, 2012). Namely, if there is a completely decentralised system, then a user sacrifices availability, reliability, scalability, and convenience by storing his/her data on his/her own machine, or even trust his/her data to one of several providers that he/she probably does not know or trust, any more than he/she would a centralised provider. This tells us that a completely decentralised system might not be possible with the current advancements. For the moment, a system like diaspora (“Diaspora,” n.d.) might work best. Being a real-world decentralised social network, it operates on federated servers (Bielenberg & Helm, 2012). Users can either host their data on their own servers (by maintaining security, integrity and reliability of their data), or join existing servers. The users, general speaking, prefer the latter (Bielenberg & Helm, 2012), which puts them in a precarious position. For those preferring a server of their own, they are required to be quite technically advanced, which puts general users off.

6.9 Business Model

In terms of a business model, we believe the best way forward is to build on the freemium (a contraction of free and premium) model. Freemium means a basic version is provided free of charge and then a premium cost is paid on added-value versions. Although this model existed from the 1980s (Wikipedia contributors, n.d.) however, the term Freemium was coined by a venture capitalist Fred Wilson in a blog post in 2006 (Pujol, 2010; F. Wilson, 2006). This is how Fred explained the model:

Give your service away for free, possibly ad supported but maybe not, acquire a lot of customers very efficiently through word of mouth, referral networks, organic search marketing, etc., then offer premium priced value added services or an enhanced version of your service to your customer base.

Several successful internet based services like Flickr and LinkedIn use this model. Also large software companies (such as Linux, Firefox, and Apache) which operate in the open source marketplace use this model (Teece, 2010). An exclusive paid service will not work as it would not only create hurdles to attract initial users, but also future users as well. With so many free popular services out there now, users expect that the basic service for any new system should be free of charge (Teece, 2010). This will help us market our solution to a wider audience. We are not looking to introduce ads in our solution, as we believe that they are quite intrusive. Value-added or premium services, inspired from successful systems such as Skype and LinkedIn, might include:

- Live video chatting with multiple users;
- Strangers may access other profiles, provided that each individual has allowed strangers to access their profile.

All such added or even free services will respect each user's privacy settings. For instance if a user has not authorised a stranger to contact them, then under no circumstances will we provide access to their profile.

6.10 Future Work

Here we present three possible extensions of our work, which address the possibility of peers having a mobile identity and of real-life contact searches.

Binding a person to his peer identity which is independent of his IP address, computer and current location, can allow him to regain his social network no matter from where he joins the system. We will be focusing on the problem facing a peer whenever it wants to re-join the [P2P](#) system, either after having lost its data, or having changed its computer, or even its location. We call this the mobile identity problem of the peers. We want to enable such a peer to easily regain its social network. It would ask the system for its social network by supplying its credentials, i.e., username and password. Then the system will gossip around and try to find its friends, who had previously stored its social network. Its request would be directed to its friends, who would eventually help it regain its social network.

We would also like to extend our current [SNS](#) by incorporating, searching and then establishing friendship links with one's real-life friends. After getting the contacts list from a peer's associated email service, we will carry out a search of its real friends, in case they happen to already exist. The process would first search locally in the peer's mega-cache, and then expand its search by contacting highly connected peers (peers with a large social circle) and then ask them for the peer's friends. Once they have been found, our retry mechanism would come into the play and then try to establish friendship links with them.

In addition to that, we would like to develop mechanisms to diversify social networks. Link prediction and recommendation is one of the key features to grow an [SNS](#) and we would like to work towards it. This knowledge comes from our work on developing links among students when their attributes are taken into account (covered in previous chapters 4 and 5). This may include self-defined ethnicity, which would diversify the social network.

6.11 Conclusions

This chapter described a completely decentralised, self-administered, light-weight and scalable Social Network System ([SNS](#)). However, one should clarify that this is more of a proof of concept system. The peers, which are potential friends, are discovered through gossip based protocols. To overcome the dynamicity of peer availability in [P2P](#) systems, we have demonstrated our mechanisms which establish friendship links between peers in such a transient environment. Current implementation and deployment have been done in Tribler [P2P](#) client, but it can run on any gossip based [P2P](#) client. We have also carried out reliability experiments to show the behaviour of our [SNS](#) in real-life scenarios. Our data for the experiments has been collected with our deployed [SNS](#) under the Tribler 4.5 release. To thwart unwanted and malicious attacks of DDOS and ‘man in the middle attacks’, we have developed safeguards to restrict users overloading the system and to use signed friendship requests. Two further extensions of our work have also been presented.

This work is more of a proof of concept that without any central entity, users can develop and maintain their own social networks by themselves. It is by no means a powerful [SNS](#) which can run on multiple devices at once, providing several communication and sharing services which are commonplace in a typical [SNS](#).

As we have seen after studying centralised [SNSs](#), such as Facebook, users would like to have more control over their own data, in who can access what and to what extent. In order to have more innovation and to further complicate privacy issues, these [SNSs](#) offer multiple applications which are developed by third parties, such as games, birthday reminders etc. They pose additional security and privacy concerns, as these apps can also access private information. There is a general argument about data in that, who actually owns it?

The free service given by [SNSs](#) comes at the cost of sharing your personal information with them, so that personalised ads can be directed towards you. For instance, if you are starting an ad campaign to target 25-30 year olds living in New York, you can ask Facebook just to target them, based on the explicit information users have shared with it. It gets more complex for implicit information. For instance, it was recently shown that Facebook can predict when partners will most likely break up their relationship solely based on their interaction history. That is not the end of it. Facebook also stores the information you were about to share, but edited first before posting it. All the edits are stored by Facebook. The implications of [SNSs](#) is bigger than socialising. Since having a public profile in an [SNS](#) has become a norm, employers vet prospective employees and even keep tabs on current ones.

It was in early 2013 when the National Security Agency’s (NSA) scandal came to the mainstream media, thanks to the whistle-blower Edward Snowden. News reports revealed that NSA (and its international partners) had been involved in global surveillance program(s) around mobile, telephone and Internet communication systems (Wikipedia, n.d.-a). For more information, please see (Wikipedia, n.d.-b). Several mainstream [SNSs](#) such as Facebook, Google, Twitter and YouTube were compromised (either with their consent or secretly by the NSA itself). Facebook owner, Mark Zuckerberg, admitted that after this scandal, Facebook users’ trust has dampened:

The trust metrics for Facebook, Twitter, and Google have all gone down since the NSA scandal first broke.

(Mark Zuckerberg)(Grove, n.d.)

If anything, we have seen in the news media and also from this research that there is a growth in privacy seeking solutions, despite SNSs encouraging users to share more information with them (Stutzman, Gross, & Acquisti, 2012; Wang et al., 2014). These trends for more privacy prone systems have only been accelerated after the NSA scandal (Wang et al., 2014). Systems like ours may fulfil users' requirement for a self-administered service, but it needs further development both by providing more functionalities and also securing them against possible attacks.

7 Chapter: Conclusion

7.1 Contributions

In this chapter we assess the extent to which we achieved our aims and objectives and how we answered the questions set out in the beginning of our work. Also we indicate how the future of [SNSs](#) could look, according to us.

In this thesis, we explored the various mechanisms involved in the development and maintenance of relationships (friendship) in a typical Social Network System ([SNS](#)). Our aim was the identification of some key areas, which can help future platform and application developers in creating better, more efficient, more open and user-friendly [SNSs](#). We now evaluate our aims and objectives against what was achieved:

a) *To assess and analyse a dataset from an [SNS](#) if it represents segregated communities; and on what factor(s) does it cluster and segregate.*

We learned a great deal about Facebook usernames, which is somewhat an understudied aspect. Most of the users on Facebook use their real names as their username (Dwyer et al., 2007), which in turn reveals a lot about their ethnic, religious and even language background. By using a name-based ethnic classifier, Onomap, which estimates ethnic, religious and language classification on our dataset of Facebook data of [MMU](#) students, we determined on what lines the biggest [SNS](#), Facebook, is segregated. We have used such information along with social (friendship) networks to identify how various ethnic, religious, language and geography based groups are inter and intra linked with each other. In all the contracted graphs based on ethnicity, geography, language and religion, the preferences of non-dominant groups vary among dominant groups. Our underlying hypotheses were:

H1 (a): The Facebook network is segregated on the ethnic lines;

H1 (b): The Facebook network is highly clustered on ethnic lines,

which were tested against a null hypothesis:

H0: The Facebook network does not segregate on ethnic lines and is not highly clustered on ethnic lines

The analysis based on the Silo Indices of the reference dataset ([MMU](#)'s) and a random dataset, suggests the Facebook network is indeed clustered and segregated when the estimated ethnicity of users' is taken into account. Furthermore, our analysis highlighted the similar pattern (of clustering and segregation) for groups based on language, religion and also geographical area. Hence the null hypothesis is rejected. This is further strengthened when we look at the normalised Silo Index for all four attributes: ethnicity, religion, geography and language, as most normalised Silo Index values are positive, showing that these groups are inward looking.

In the case of ethnicities, we found that there is a clear cohesiveness, for instance, the International, Nordic, Sikh, South Asian and Muslim groups have the most number of links with the Muslim group, while the rest of them are mostly connected with the English group. Also, we found in terms of inter-group propensity, the Muslim group stood out with the highest propensity (-0.21). When compared with the reference network (the null model), we found the Muslim group had the highest difference (0.58) (African with 0.10 and Celtic with 0.09 come afterwards) in the Silo Index. This shows that although the Muslim group is a dominant group (21.02%), it is certainly not the most dominant (the English group

represents 45.8%), it cannot be explained by merely the population size. When we analysed the normalised Silo Index the Muslim group falls at the sixth position with 0.75. The most inward group became the Greek group with 0.89. Taking individual group's affinities for inter and intra groups opens up a great deal of applications for diverse users.

Our dataset of Facebook Data from [MMU](#) students certainly does not represent the whole Facebook network, but the degree distribution of our visited nodes (those which have their ego network crawled) shows that there are two regimes of power-law effect, as it was shown by Minas et al. (Minas Gjoka, Kurant, Butts, et al., 2009), for node degrees greater than and less than 300. It indicates that our dataset, in terms of degrees, is a good representation of Facebook's social network. Also, after comparing the student ethnic distribution of [MMU](#), with that of our reference network, we learnt that our dataset is fairly representative in this respect as well.

b) To build a series of evidence driven agent-based models to identify the micro to macro level social processes involved in friendship development

In order to explain and capture the dynamics involved in social networks and their development, we found Agent-Based Modelling techniques the best suited methodology, when it was compared with mechanistic approaches such as random networks (where links are randomly created); preferential attachment (links are created with high degree nodes); scale-free networks (a power-law network with a selection bias for nodes with high links) and small-world networks (a pseudo random network with low-average path length than a completely random network).

Based on social theories, and data collected from [SNSs](#) through self-reported surveys, memes and previous studies, we developed initial understandings of the preferences and affinities of individuals, which inter-played their role in friendship development. Based on the gained knowledge, we concentrated on the structure, maintenance and development of social networks in a typical [SNS](#). Unlike social network theories, which focus solely on relationships between individuals, we focused on attributes of individuals too. To calibrate and test our [ABM](#), we used datasets from three universities, which included attributes and relationships of individuals. These included Caltech, Princeton and Georgetown universities. Our evidence supports that in the case of Caltech, being a smaller university, we found new friend relationships are mainly introduced by current friends (friends of a friend mode). However, in the case of bigger and more diverse universities, such as Princeton and Georgetown, there is a multitude of social processes involved. These include occasions for meeting others with similar based on the same dormitory, major or high school etc., social interaction, random meet ups (through parties or other social events) and current friends introducing new friends.

We wanted to apply the gained knowledge of inter and intra ethnic preferences to our [ABM](#) models for student interactions (and link development), but were not able to do so. This was mainly due to lack of data in the underlying datasets about student's ethnicity so we could not establish further validation of our models and/or datasets of cross-sectional nature.

c) To build a distributed [SNS](#) which liberates the user to manage their social networks.

Since there is a large amount of personal data held within [SNSs](#), advertising and marketing agencies have created very sophisticated systems to gather information about people. It is a goldmine in terms of information about users that can be used to direct

personalised advertisements. Also various governmental agencies use this as an excuse to curb potential threats in the name of national security and have been using [SNSs](#), both legally and illegally, to obtain information about numerous users (people). In order to deal with such issues inherent in centralised client-server architecture, especially after the NSA scandal which reported that everyone using mainstream social media and email services was being tracked, we have proposed and implemented a complete decentralised [SNS](#) using a peer-to-peer approach. It is a self-administered system and we have explained how users can develop and manage their social network themselves. With real users using our system, we proved how effective it is as well as identified where it could be improved.

Our decentralised Social Network System ([SNS](#)) is only a proof of concept; hence it is not ready to be translated into a real-world system for a large user base. This prototype establishes the feasibility of a totally distributed [SNS](#), but it would take further work to ensure that it would scale up to a large system. Nonetheless, it is a step forward in designing a user-centric system which is secure and transparent.

In terms of knowledge contributions, our work has covered several different aspects of an [SNS](#). Firstly, we provided useful dynamic models for student interaction, leading to friendship development. These models, whose rationales come from the studies and also the general outlook of [SNSs](#), were found to be generalisable across three different datasets. It not only provides local mechanisms for individual friendship development, but it also provides a valuable contribution in terms of understanding network formation processes. To establish the validity of our models, we relied on established techniques from various fields, such as Computer Science and Social Science. We applied not only structural measures (such as clustering coefficient), but we combined attribute specific measures (such as the Silo Index) to our validation process, thereby enhancing best practices to validate one's model against the real-world datasets.

Secondly, in terms of specific inter and also intra preferences across ethnic, religious, geographical and language based groups, we identified how various groups behave and found them to be clustered and segregated into communities. This contributes to our understanding of cliques in our society. Also it provides useful information that might inform the provision of targeted services for each of those groups. This may include better friendship recommendations, but also a host of applications. Thirdly, in order to empower users we also delved into technicalities of a completely decentralised [SNS](#). This meant designing a self-administered and controlled system, which also provides necessary means against unwanted attacks. Without collaboration from the very helpful academics this work would not have been possible. We would deeply like to thank everyone for their guidance and helpful collaborations.

Our work can be broadly divided into two strands which we consider to be directly linked. One strand, broadly speaking, identifies the social and educational processes and ethnic preferences of students. The second strand talks about a proof of concept of a completely decentralised Social Network System ([SNS](#)). These are linked in several ways. Firstly, in order to understand the social interaction of students, which then leads to their friendship development, we have identified different mechanisms depending on the size of a university. Also when an individual's ethnic background is taken into account we can, with some confidence, offer individual ethnic preferences for inter and intra groups that an individual belongs to. This involves taking their ethnic, educational and network position, and then identifying what new applications they might be interested in. One such application could be to recommend new friendships to individuals, which is link prediction and recommendation; a highly lucrative feature for the growth of an [SNS](#). Secondly, as a proof of concept, we have identified how a completely decentralised social network could

be constructed, giving full power to users to manage their data as they wish. It is grounded in keeping the privacy of users intact.

7.2 Future Work

Here we briefly discuss some future research that I am interested in:

- 1) We have gained access to another dataset which is much bigger than ours, but with a limited ethnic classification of users. We would like to analyse this using the name-based ethnic classifier tool, Onomap and see if similar results can be reproduced using this dataset. Also, we would like to expand on our findings on ethnic classification to other [SNSs](#), such as Myspace, which already provides, albeit limited, ethnic classification of users. This would help validate and assess the quality of the classification that Onomap produces.
- 2) We would like to test our [ABM](#) against other datasets to determine its generalisability and flexibility. We have already done that with three datasets, but there is a potential for further validation. Although some of its algorithms are dependent on the underlying attributes of nodes (for instance the personal preference algorithm), with a slight modification it could be adjusted for any dataset. We would also like to produce a synthetic social network with the same characteristics as the ones we collected to allow its free distribution to other researchers without comprising the privacy of our subjects.
- 3) As for our distributed [SNS](#), there is a lot of room for further research. One of the biggest challenges is to have control over data. We would like to explore how data that is spread over a distributed system might be managed by its original creator. We would also like to test our proposed system with a larger user population and higher usage. So far, we have not carried out testing with real-time communication based applications, such as messaging between users. We would like to evaluate the efficacy of such processes using further simulations of the proposed system. Lastly we would like to devise a mechanism to predict link recommendations for a variety of users, based on their self-defined attributes and behaviours.

7.3 Recommendations for Future SNS Developers

The success of an [SNS](#) lies in keeping users and their preferences as the main focus. Based on our experiences and the literature review, we recommend some key features for a user-centric SNS. We determine success when users are satisfied and fully aware of the privacy settings that a system affords them. A single policy applied to all users can create issues for unintended groups. An [SNS](#) which offers flexible services to a diverse set of users, so that a customisable system can be developed, is more likely to succeed. The key, in our view, is to give users the control over their own data and offer privacy policies so that each individual can easily set out their own preferences. The data should ultimately be owned by users themselves, not by the system. Also a system should be transparent, so that users know at all times how their data might be used to infer hidden information about themselves.

7.4 Future [SNSs](#)

To cater for improved privacy and security concerns of users, systems like TOR (Sabatini & Sarracino, 2013b), which provides anonymous interface to Internet, do exist. TOR is a great tool for journalists and whistle-blowers all around the world, allowing them to collaborate on sensitive projects without being tracked by anyone. It relies on onion routing protocols which distributes your communication over several places on the Internet, so no single point can link you to your destination. The communication follows a random pathway through several intermediaries (known as relay nodes in TOR's vocabulary) that covers your tracks so no observer at any single point can tell where the data came from or where it is going (Sabatini & Sarracino, 2013b). Users of [SNSs](#) would like to have more control and privacy over their data (Duguay, 2014), which implies that what we need is an [SNS](#) which provides a private and more secure environment to users which is also easy and flexible for them to use. Tribler, the underlying system of our [SNS](#), has been dealing with anonymity issues for a while now. It has been tested with a subset of TOR's onion protocols to provide anonymous communication between users (Sabatini & Sarracino, 2013c), but it does not provide the same level of anonymity which is provided by TOR (Sabatini & Sarracino, 2013a). We see similar systems to be commonplace, once the technical barriers are removed.

8 References

- Abbas, S. (2009). *Social Networking in the Virtual World*. Delft. Retrieved from <http://www.tribler.org/trac/raw-attachment/wiki/VirtualCommunities/SNS-report.pdf>
- Abbas, S. (2011a). An agent-based model of the development of friendship links within Facebook. In *7th European Social Simulation Association Conference*. Montpellier, France.
- Abbas, S. (2011b). Ethnic diversity in Facebook. *Technical Report Series of Centre for Policy Modeling*. Manchester. Retrieved from <http://www.cfpm.org/~ali/FacebookReport/FacebookOnomapGroupReport.pdf>
- Abbas, S. (2013). Homophily , Popularity and Randomness : Modelling Growth of Online Social Network. In *Proceedings of the 12th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2013)* (pp. 135–142).
- Abbas, S., Alam, S., & Edmonds, B. (2014). Towards Validating Social Network Simulations. *Advances in Social Simulation*, 229(Springer (2014)), 1–12. http://doi.org/Doi.10.1007/978-3-642-39829-2_1
- Aberer, K., Datta, A., & Hauswirth, M. (2004). Efficient, self-contained handling of identity in peer-to-peer systems. *IEEE Transactions on Knowledge and Data Engineering*, 16(7), 858–869. <http://doi.org/10.1109/TKDE.2004.1318567>
- Abram, C. (n.d.). Welcome to Facebook, everyone. Retrieved from <http://blog.facebook.com/blog.php?post=2210227130>
- Acquisti, A., & Gross, R. (2006). Imagined Communities : Awareness , Information Sharing , and Privacy on the Facebook. *Privacy Enhancing Technologies*, 36–58.
- Adamic, L. (1999). The small world web. *Research and Advanced Technology for Digital Libraries, Springer B*, 443–452.
- Adamic, L. (2012). How You Met Me. *AAAI Conference on Weblogs and Social Media*, 371–374. Retrieved from <http://www.aaai.org/ocs/index.php/ICWSM/ICWSM12/paper/viewFile/4681/5009>
- Adamic, L., & Adar, E. (2003). Friends and neighbors on the web. *Social Networks*, 25(3), 211–230. [http://doi.org/10.1016/S0378-8733\(03\)00009-1](http://doi.org/10.1016/S0378-8733(03)00009-1)
- Adar, E., & Huberman, B. (2000). Free riding on Gnutella. *First Monday*, 1–20. Retrieved from <http://ptwich2.lib.uic.edu/ojs/index.php/fm/article/view/792>
- Airoldi, E., Blei, D., Fienberg, S., & Xing, E. (2008). Mixed Membership Stochastic Blockmodels. *Journal of Machine Learning Research : JMLR*, 9, 1981–2014. Retrieved from <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3119541&tool=pmcentrez&rendertype=abstract>
- Alam, S. J., Abbas, S., & Edmonds, B. (2014). Validating Simulated Networks: Some Lessons Learned. *Multi-Agent-Based Simulation XIV*, (Springer Berlin Heidelberg), 71–82.
- Alexa. (n.d.). The top 500 sites on the web. Retrieved from <http://www.alexa.com/topsites>
- Back, M. D., Stopfer, J. M., Vazire, S., Gaddis, S., Schmukle, S. C., Egloff, B., & Gosling, S. D. (2010). Facebook profiles reflect actual personality, not self-idealization. *Psychological Science*, 21(3), 372–4. <http://doi.org/10.1177/0956797609360756>
- Backstrom, L., & Bakshy, E. (2011). Center of attention: How facebook users allocate attention across friends. *Proc. 5th International Conference on Weblogs and Social Media.*, 34–41. Retrieved from <http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/viewFile/2899/3259>
- Backstrom, L., & Leskovec, J. (2011). Supervised random walks: predicting and recommending links

- in social networks. In *Proceedings of the fourth ACM international conference on Web search and data mining* (pp. 635–644). Retrieved from <http://dl.acm.org/citation.cfm?id=1935914>
- Bagrow, J. P., Bollt, E. M., Skufca, J. D., & Ben-Avraham, D. (2007). Portraits of Complex Networks. *EPL (Europhysics Letters)*, *81*(6), 68004. <http://doi.org/10.1209/0295-5075/81/68004>
- Barabási, A.-L., & Bonabeau, E. (2003). Scale-free networks. *SCIENTIFIC AMERICAN*, (May). Retrieved from <http://www.digitalunion.osu.edu/r2/summer06/sass/Articles/SciAm2003.pdf>
- Barabási, A. L., & Albert, R. (1999). Emergence of Scaling in Random Networks. *Science*, *286*(5439), 509–512. <http://doi.org/10.1126/science.286.5439.509>
- Baset, S. A., & Schulzrinne, H. G. (2006). An Analysis of the Skype Peer-to-Peer Internet Telephony Protocol. *Proceedings IEEE INFOCOM 2006. 25TH IEEE International Conference on Computer Communications*, 1–11. Networking and Internet Architecture; Multimedia. <http://doi.org/10.1109/INFOCOM.2006.312>
- Berger-Wolf, T. Y., & Saia, J. (2006). A framework for analysis of dynamic social networks. *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 523–528. <http://doi.org/10.1145/1150402.1150462>
- Bielenberg, a, & Helm, L. (2012). The growth of Diaspora-A decentralized online social network in the wild. *Computer Communications Workshops (INFOCOM WKSHPS), IEEE*, 13–18. <http://doi.org/10.1109/INFOCOMW.2012.6193476>
- Blondel, V., & Guillaume, J. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, *2008*, P10008. Retrieved from <http://iopscience.iop.org/1742-5468/2008/10/P10008>
- Bonabeau, E. (2002). Agent-based modeling: methods and techniques for simulating human systems. *Proceedings of the National Academy of Sciences*, *99*(suppl. 3), 7280–7287. <http://doi.org/10.1073/pnas.082080899>
- Borch, N. (2005). Social peer-to-peer for social people. *The Int'l Conf. on Internet Technologies and Applications*. Retrieved from <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.78.4155&rep=rep1&type=pdf>
- Boyack, K. W., & Klavans, R. (2010). Co-citation analysis, bibliographic coupling, and direct citation: Which citation approach represents the research front most accurately? *Journal of the American Society for Information Science and Technology*, *61*(12), 2389–2404. <http://doi.org/10.1002/asi.21419>
- Boyd, D. (2006). Friendster lost steam. Is MySpace just a fad. *Apophenia Blog*, 1–8. Retrieved from <http://www.danah.org/papers/FriendsterMySpaceEssay.html>
- Boyd, D. (2010). Social Network Sites as Networked Publics: Affordances, Dynamics, and Implications. *Networked Self: Identity, Community, and Culture on Social Network Sites*, 39–58. <http://doi.org/10.1162/dmal.9780262524834.119>
- Boyd, D., & Ellison, N. (2007). Social Network Sites: Definition, History, and Scholarship. *Journal of Computer-Mediated Communication*, *13*(1), 210–230. <http://doi.org/10.1111/j.1083-6101.2007.00393.x>
- Buchegger, S., Schiöberg, D., Vu, L., & Datta, A. (2009). PeerSoN: P2P social networking: early experiences and insights. *Proceedings of the Second ACM EuroSys Workshop on Social Network Systems*, 46–52. Retrieved from <http://dl.acm.org/citation.cfm?id=1578010>
- Bulmer, M. (n.d.). The ethnic group question in the 1991 Census of Population. *Ethnicity in the 1991 Census. Demographic Characteristics of the Ethnic Minority Populations*, 1.
- Carley, K. M. (2003). Dynamic Network Analysis. *Dynamic Social Network Modeling and Analysis Workshop Summary and Papers*, 133–145. <http://doi.org/10.1103/PhysRevB.81.041203>
- Catanese, S., & Meo, P. De. (2011). Crawling facebook for social network analysis purposes. *Proceedings of the International Conference on Web Intelligence, Mining and Semantics*, 0–7.

Retrieved from <http://doi.acm.org/10.1145/1988688.1988749>

- Catanese, S., Meo, P. De, Ferrara, E., & Provetti, A. (2011). Extraction and Analysis of Facebook Friendship Relations. *Computational Social Networks: Mining and Visualization*.
- Chang, J., & Rosenn, I. (2010). epluribus: Ethnicity on social networks. *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*, 18–25. Retrieved from <http://www.aaai.org/ocs/index.php/ICWSM/ICWSM10/paper/viewPDFInterstitial/1534/1828>
- Christakis, N. A., & Fowler, J. H. (2009). *Connected: The Surprising Power of Our Social Networks and How They Shape Our Lives*. New York (Little, Br, Vol. 3). <http://doi.org/10.1111/j.1756-2589.2011.00097.x>
- Clauset, A., Newman, M., & Moore, C. (2004). Finding community structure in very large networks. *Physical Review. E, Statistical, Nonlinear, and Soft Matter Physics*, 70(6 Pt 2), 66111. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/15697438>
- Cohen, B. (2003). Incentives build robustness in BitTorrent. *Workshop on Economics of Peer-to-Peer Systems*. Retrieved from <http://pdos.csail.mit.edu/6.824-2010/papers/cohen-btecon.pdf>
- Cutillo, L., Molva, R., & Onen, M. (2011). Safebook: A distributed privacy preserving Online Social Network. *World of Wireless, Mobile and Multimedia Networks (WoWMoM), 2011 IEEE International Symposium on A., (IEEE)*, 1–3. Retrieved from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5986118
- Dekker, A. (2007). Realistic social networks for simulation using network rewiring. *International Congress on Modelling and Simulation*, (i), 677–683. Retrieved from http://www.mssanz.org.au/MODSIM07/papers/13_s20/RealisticSocial_s20_Dekker_.pdf
- Diaspora. (n.d.). Retrieved April 15, 2014, from <https://joindiaspora.com/>
- Duguay, S. (2014). “He has a way gayer Facebook than I do” : investigating sexual identity disclosure and context collapse on a social networking site. *New Media and Society*, 1–17. <http://doi.org/10.1177/1461444814549930>
- Dwyer, C., Hiltz, S., & Passerini, K. (2007). Trust and privacy concern within social networking sites: A comparison of Facebook and MySpace. *Proceedings of Americas Conference on Information Systems (AMCIS)*. Retrieved from <http://aisel.aisnet.org/amcis2007/339>
- Edmonds, B. (1999). Modelling Bounded Rationality in Agent-Based Simulations Using the Evolution of Mental Models. In *Computational Techniques for Modelling Learning in Economics* (pp. 305–332). http://doi.org/10.1007/978-1-4615-5029-7_13
- Ellison, N. B., Steinfield, C., & Lampe, C. (2007). The Benefits of Facebook “Friends:” Social Capital and College Students’ Use of Online Social Network Sites. *Journal of Computer-Mediated Communication*, 12(4), 1143–1168. <http://doi.org/10.1111/j.1083-6101.2007.00367.x>
- Epstein, J. M. (1999). Agent-Based Computational Models and Generative Social Science. *Generative Social Science: Studies in Agent-Based Computational Modeling*, 4(5), 41–60.
- Erdős, P., & Rényi, A. (1959). On random graphs I. *Publ. Math. Debrecen*, 290–297. Retrieved from http://ftp.math-inst.hu/~p_erdos/1959-11.pdf
- Erdős, P., & Rényi, A. (1960). On the evolution of random graphs. *Publ. Math. Inst. Hungar. Acad. Sci*, 286(1), 257. <http://doi.org/10.2307/1999405>
- Everett, M. G., & Borgatti, S. P. (2012). Categorical attribute based centrality: E-I and G-F centrality. *Social Networks*, 34(4), 562–569. <http://doi.org/10.1016/j.socnet.2012.06.002>
- Facebook. (n.d.-a). Facebook Statistics. Retrieved from www.facebook.com/press/info.php?statistics
- Facebook. (n.d.-b). Facebook Statistics. Retrieved from <https://newsroom.fb.com/company-info/>
- Facebook. (n.d.-c). Facebook Timeline. Retrieved from <http://www.facebook.com/press/info.php?timeline>

- Faludi, S. C. (1979). Help Wanted: Bass Tacks. Retrieved from <http://www.thecrimson.com/article/1979/9/28/help-wanted-pbwould-you-like-a/>
- Farmer, J. D., & Foley, D. (2009). The economy needs agent-based modelling. *Nature*, 460(7256), 685–686. <http://doi.org/10.1038/460685a>
- Feldman, A., & Blankstein, A. (2012). Social networking with frientegrity: privacy and integrity with an untrusted provider. *Proceedings of the 21st USENIX Conference on Security Symposium, Vol. 12*, 31–31. Retrieved from <https://www.usenix.org/system/files/conference/usenixsecurity12/sec12-final67.pdf>
- Forbes. (n.d.). The Evolution Of Facebook. Retrieved from <http://www.forbes.com/pictures/femf45jjk/2006-mark-zuckerbergs-profile-2/>
- Fukuyama, F. (1996). Hidden Order: How Adaptation Builds Complexity. *Foreign Affairs*, 75(4), 137. <http://doi.org/10.1162/artl.1995.2.333>
- Ganesh, A. J., Kermarrec, A.-M., & Massoulie, L. (2003). Peer-to-peer membership management for gossip-based protocols. *IEEE Transactions on Computers*, 52(2), 139–149. <http://doi.org/10.1109/TC.2003.1176982>
- Gantz, J. F., Chute, C., Manfrediz, A., Minton, S., Reinsel, D., Schlichting, W., & Toncheva, A. (2008). The diverse and exploding digital universe. *An IDC White Paper - Sponsored by EMC*. Retrieved from <http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:The+Diverse+and+Exploding+Digital+Universe#0>
- Garbacki, P., Iosup, A., Epema, D., & van Steen, M. (n.d.). 2Fast : Collaborative Downloads in P2P Networks. In *Sixth IEEE International Conference on Peer-to-Peer Computing (P2P'06)* (pp. 23–30). IEEE. <http://doi.org/10.1109/P2P.2006.1>
- Gilbert, E., & Karahalios, K. (2009). Predicting tie strength with social media. *Proceedings of the 27th International Conference on Human Factors in Computing Systems - CHI '09*, 211. <http://doi.org/10.1145/1518701.1518736>
- Gilbert, N. (2008). Agent-Based Models. *SAGE Publications*, 153(153), 98. <http://doi.org/10.4135/9781412983259>
- Gilbert, N., & Troitzsch, K. G. (2005). *Simulation for the Social Scientist (second edition)*. McGraw-Hill International. Retrieved from <http://books.google.com/books?hl=en&lr=&id=fBlaulpmNowC&oi=fnd&pg=PR1&dq=Simulation+for+the+social+scientist&ots=PBOZKCq1Wk&sig=RSXgVTM2IRFgJDFmfUAyfR30Ezk>
- Ginger, J. (2008). THE FACEBOOK PROJECT THE MISSING BOX : THE RACIAL POLITICS BEHIND. *Structure*. Retrieved from THEFACEBOOKPROJECT.COM
- Gjoka, M., Kurant, M., & Butts, C. (2009). Unbiased sampling of facebook. *Search*, 1–15. Retrieved from <http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Unbiased+Sampling+of+Facebook#0>
- Gjoka, M., Kurant, M., Butts, C., & Markopoulou, A. (2009). Unbiased sampling of facebook. *arXiv Preprint arXiv:0906.0060*, 1–15. Retrieved from <http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Unbiased+Sampling+of+Facebook#0>
- Gjoka, M., Kurant, M., Butts, C., & Markopoulou, A. (2010). Walking in Facebook: A case study of unbiased sampling of OSNs. In *INFOCOM, 2010 Proceedings IEEE* (pp. 1–9). Retrieved from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5462078
- Granovetter, M. (1985). Economic action and social structure: the problem of embeddedness. *American Journal of Sociology*. Retrieved from <http://www.jstor.org/stable/2780199>
- Grimm, Volker, Eloy Revilla, Uta Berger, Florian Jeltsch, Wolf M. Mooij, Steven F. Railsback, Hans-Hermann Thulke, Jacob Weiner, Thorsten Wiegand, and Donald L. DeAngelis (2005). *Pattern-*

- oriented modeling of agent-based complex systems: lessons from ecology*. *science* 310, no. 5750 (2005): 987-991.
- Grimm, V., Berger, U., DeAngelis, D. L., Polhill, J. G., Giske, J., & Railsback, S. F. (2010). The ODD protocol: A review and first update. *Ecological Modelling*, 221(23), 2760–2768. <http://doi.org/10.1016/j.ecolmodel.2010.08.019>
- Gross, R., Acquisti, A., & Heinz, H. J. (2005). Information revelation and privacy in online social networks. *2005 ACM Workshop on Privacy in the Electronic*, 707(November), 71. <http://doi.org/10.1145/1102199.1102214>
- Grove, J. Van. (n.d.). Zuckerberg: Thanks NSA, now people trust Facebook even less. Retrieved from <http://www.cnet.com/uk/news/zuckerberg-thanks-nsa-now-people-trust-facebook-even-less/>
- Gulyás, L., Kampis, G., & Legendi, R. O. (2013). Elementary models of dynamic networks. *The European Physical Journal Special Topics*, 222(6), 1311–1333. <http://doi.org/10.1140/epjst/e2013-01928-6>
- Hamill, L. (2010). Communications, Travel and Social Networks Since 1840: A Study Using Agent-based Models. *Social Networks*.
- Hamill, L., & Gilbert, N. (2008). A Simple but More Realistic Agent-based Model of a Social Network. In *Proceedings of European Social Simulation Association Annual Conference, Brescia, Italy*.
- Hamill, L., & Gilbert, N. (2009). Social circles: A simple structure for agent-based social network models. *Journal of Artificial Societies and Social Simulation*, 12(2). Retrieved from <http://jasss.soc.surrey.ac.uk/12/2/3.html>
- Hampton, K., & Goulet, L. (2012). Why most Facebook users get more than they give. *Pew Internet & American Life Project*. Retrieved from http://www.pewinternet.org/~media/Files/Reports/2012/PIP_Facebook_users_2.3.12.pdf
- Handling Churn in a DHT. (n.d.). Retrieved from https://www.usenix.org/legacy/event/usenix04/tech/general/rhea/rhea_html/usenix-cr.html
- Hanselmann, M., & Hamprecht, F. (2012). One Plus One Makes Three (for Social Networks). *PLoS One*, 7(4), 1–8. <http://doi.org/10.1371/journal.pone.0034740.g001>
- Herdağdelen, A., Zuo, W., Gard-Murray, A., & Bar-Yam, Y. (2013). An exploration of social identity: The geography and politics of news-sharing communities in twitter. *Complexity*, 19(2), 10–20. <http://doi.org/10.1002/cplx.21457>
- Hogan, B. (2009). A Comparison of On and Offline Networks through the Facebook API. *Social Science Research Network Working Paper Series*. <http://doi.org/10.2139/ssrn.1331029>
- Holme, P., & Saramäki, J. (2012). Temporal networks. *Physics Reports*, 519(3), 97–125. <http://doi.org/10.1016/j.physrep.2012.03.001>
- Hua, C., Mao, Y., Jinqiang, H., Haiqing, D., & Xiaoming, L. (n.d.). Maze: a social peer-to-peer network. In *IEEE International Conference on E-Commerce Technology for Dynamic E-Business* (pp. 290–293). IEEE Comput. Soc. <http://doi.org/10.1109/CEC-EAST.2004.44>
- Jackson, M. O. M., & Rogers, B. B. W. (2007). Meeting strangers and friends of friends: How random are social networks? *The American Economic Review*, 97(3), 890–915. <http://doi.org/10.1257/aer.97.3.890>
- Jahid, S., Nilizadeh, S., & Mittal, P. (2012). DECENT: A decentralized architecture for enforcing privacy in online social networks. *Pervasive Computing and Communications Workshops (PERCOM Workshops), 2012 IEEE International Conference on*, 326–332. Retrieved from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6197504
- Janssen, M. a. (2005). Agent-Based Modelling. *Modelling in Ecological Economics*, 155–172.
- Jernigan, C., & Mistree, B. F. T. (2009). Gaydar: Facebook friendships expose sexual orientation. *First Monday*, 14(10). <http://doi.org/10.5210/m.v14i10.2611>

- Joinson, A. (2008). Looking at, looking up or keeping up with people?: motives and use of facebook. *Proceedings of the Twenty-Sixth Annual SIGCHI*. Retrieved from <http://dl.acm.org/citation.cfm?id=1357054.1357213>
- Kermack, M., & Mckendrick, A. (1927). Contributions to the mathematical theory of epidemics. Part I. *Proc. R. Soc. A*. Retrieved from <http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Contributions+to+the+mathem+atical+theory+of+epidemics.+Part+I.#1>
- Kim, M., & Leskovec, J. (2010). Multiplicative attribute graph model of real-world networks. *Algorithms and Models for the Web-Graph*, 8(1–2), 113–160. <http://doi.org/10.1080/15427951.2012.625257>
- Krackhardt, D., & Stern, R. N. (2011). Informal Networks and Organizational Crises : An Experimental Simulation *. *Social Psychology*, 51(2), 123–140.
- Kross, E., Verduyn, P., Demiralp, E., Park, J., Lee, D. S., Lin, N., ... Ybarra, O. (2013). Facebook use predicts declines in subjective well-being in young adults. *PloS One*, 8(8), e69841. <http://doi.org/10.1371/journal.pone.0069841>
- Lakha, F., Gorman, D. R., & Mateos, P. (2011). Name analysis to classify populations by ethnicity in public health: validation of Onomap in Scotland. *Public Health*, 125(10), 688–96. <http://doi.org/10.1016/j.puhe.2011.05.003>
- Lampe, C., & Ellison, N. (2006). A Face (book) in the crowd: Social searching vs. social browsing. *Proceedings of the 2006 20th*, 167–170. Retrieved from <http://dl.acm.org/citation.cfm?id=1180901>
- Lampe, C., Ellison, N., & Steinfield, C. (2007). A Familiar Face (book): Profile Elements as Signals in an Online Social Network. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM.
- Lee, S. H., Kim, P., & Jeong, H. (2009). Statistical properties of sampled networks. *Network*.
- Lewis, K., Kaufman, J., & Gonzalez, M. (2008). Tastes, ties, and time: A new social network dataset using Facebook. com. *Social Networks*, 30(4), 330–342. <http://doi.org/10.1016/j.socnet.2008.07.002>
- Liang, J., Kumar, R., Xi, Y., & Ross, K. (2005). Pollution in P2P file sharing systems. *INFOCOM 2005. 24th Annual Joint Conference of the IEEE Computer and Communications Societies*. Retrieved from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1498344
- List Of Social Networking Websites. (n.d.). Retrieved from https://en.wikipedia.org/wiki/List_of_social_networking_websites
- Liu, Y., Gummadi, K., Krishnamurthy, B., & Mislove, A. (2011). Analyzing Facebook privacy settings: User expectations vs. reality. *Proceedings of the 2011 ACM SIGCOMM Conference on Internet Measurement Conference*, (ACM), 61–70. Retrieved from <http://dl.acm.org/citation.cfm?id=2068823>
- Macy, M. W., & Willer, R. (2002a). From Factors To Actords : Computational Sociology and Agent-Based Modeling. *Annual Review of Sociology*, 28(1), 143–166. <http://doi.org/10.1146/annurev.soc.28.110601.141117>
- Macy, M. W., & Willer, R. (2002b). From Factors To Actors: Computational Sociology And Agent-based modeling. *Annual Review of Sociology*, 28(1), 143–166. <http://doi.org/10.1146/annurev.soc.28.110601.141117>
- Manku, G., Bawa, M., & Raghavan, P. (2003). Symphony: Distributed Hashing in a Small World. *USENIX Symposium on Internet* Retrieved from http://www.usenix.org/event/usits03/tech/full_papers/manku/manku.pdf
- Marmaros, D., & Sacerdote, B. (2006). How Do Friendships Form? *The Quarterly Journal of Economics*, 121(1), 79–119. <http://doi.org/10.1162/003355306776083563>
- Marti, S., Ganesan, P., & Garcia-Molina, H. (2005). *DHT routing using social links*. (G. M. Voelker &

- S. Shenker, Eds.) (Vol. 3279). Berlin, Heidelberg: Springer Berlin Heidelberg.
<http://doi.org/10.1007/b104020>
- Martin, S., Brown, W., & Klavans, R. (2011). OpenOrd: an open-source toolbox for large graph layout. *Conference on Visualization, 7868*(May 2012), 786806-786806–11.
<http://doi.org/10.1117/12.871402>
- Mateos, P., Webber, R and Longley, P. (2007). *The Cultural, Ethnic and Linguistic Classification of Populations and Neighbourhoods using Personal Names. CASA Working Paper 116* (Vol. 44). Retrieved from <http://eprints.ucl.ac.uk/3472>
- Mateos, P. (2007). Classification Methods and their Potential in Population Studies. *Population, Space and Place, 13*(May), 243–263. <http://doi.org/10.1002/psp>
- Mateos, P., Longley, P. a., & O’Sullivan, D. (2011). Ethnicity and Population Structure in Personal Naming Networks. *PLoS ONE, 6*(9), e22943. <http://doi.org/10.1371/journal.pone.0022943>
- Mayer, A., & Puller, S. L. (2008). The old boy (and girl) network: social network formation on university campuses. *Journal of Public Economics, 92*, 329–347.
- McCulloh, I., & Carley, K. (2009). Detecting change in longitudinal social networks. *Journal of Social Structure*. Retrieved from <http://oai.dtic.mil/oai/oai?verb=getRecord&metadataPrefix=html&identifier=ADA550790>
- McCulloh, I., Johnson, A., & Carley, K. (2012). Spectral Analysis of Social Networks to Identify Periodicity. *The Journal of Mathematical Sociology, 36.2*, 80–96. Retrieved from <http://www.tandfonline.com/doi/abs/10.1080/0022250X.2011.556767>
- McKeon, M. (n.d.). The Evolution of Privacy on Facebook. Retrieved from <http://www.mattmckeon.com/facebook-privacy/>
- McPherson, M., Smith-Lovin, L., & Cook, J. M. (2001). Birds of a Feather: Homophily in Social Networks. *Annual Review of Sociology, 27*(1), 415–444.
<http://doi.org/10.1146/annurev.soc.27.1.415>
- Milgram, S. (1967). The small world problem. *Psychology Today, 1*(1), 61–67. Retrieved from http://measure.igpp.ucla.edu/GK12-SEE-LA/Lesson_Files_09/Tina_Wey/TW_social_networks_Milgram_1967_small_world_problem.pdf
- Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., & Alon, U. (2002). Network motifs: simple building blocks of complex networks. *Science (New York, N.Y.), 298*(5594), 824–7.
<http://doi.org/10.1126/science.298.5594.824>
- Mislove, A. (2009). *Online social networks: measurement, analysis, and applications to distributed information systems*. Retrieved from <http://hdl.handle.net/1911/61861>
- Mislove, A., Marcon, M., Gummadi, K. P., Druschel, P., Bhattacharjee, S., & Bhattacharjee, B. (2007). Measurement and analysis of online social networks. *Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement - IMC '07*, 29. <http://doi.org/10.1145/1298306.1298311>
- Mislove, A., Viswanath, B., Gummadi, K. P., & Druschel, P. (2010). You Are Who You Know : Inferring User Profiles in Online Social Networks. In *Proceedings of the third ACM international conference on Web search and data mining* (pp. 251–260).
- Mitchell, R., Shaw, M., & Dorling, D. (2000). *Inequalities in life and death What if Britain were more equal? Joseph Rowntree Foundation*.
- MMU University. (2007). *Equality and Diversity Monitoring Data for Staff and Students*. Retrieved from <http://www.mmu.ac.uk/equality-and-diversity/doc/revised-equal-ops-report-staff-students.pdf>
- Moss, S. (2008). Alternative Approaches to the Empirical Validation of Agent-Based. *Journal of Artificial Societies and Social Simulation, 1*, 5.
- Nadkarni, A., & Hofmann, S. G. (2012). Why Do People Use Facebook? *Personality and Individual Differences, 52*(3), 243–249. <http://doi.org/10.1016/j.paid.2011.11.007>

- Nations, D. (n.d.). What is Web 2.0?
- Newman, M. E. J. and Girvan, M. (2004). Finding and evaluating community structure in networks. *Phys. Rev. E*, 69(2), 26113. <http://doi.org/10.1103/PhysRevE.69.026113>
- Newman, M. (2002). Assortative mixing in networks. *Physical Review Letters*, 2(4), 1–5. Retrieved from <http://prl.aps.org/abstract/PRL/v89/i20/e208701>
- Newman, M. (2003). Mixing patterns in networks. *Physical Review E*, 67, 26126. Retrieved from <http://pre.aps.org/abstract/PRE/v67/i2/e026126>
- Newman, M. (2010a). *Networks: an introduction*. Oxford University Press.
- Newman, M. (2010b). *Networks: an introduction*. <http://doi.org/10.1007/978-3-319-03518-5-8>
- Office for National Statistics. (2013). Key Statistics and Quick Statistics for local authorities in the United Kingdom, 1–27.
- Opsahl, T. (2010). Modeling the evolution of continuously-observed networks: Communication in a Facebook-like community. *Arxiv Preprint arXiv:1010.2141*, 1–22. Retrieved from <http://arxiv.org/abs/1010.2141>
- Panzarasa, P., Opsahl, T., & Carley, K. M. (2009). Patterns and Dynamics of Users' Behavior and Interaction: Network Analysis of an Online Community. *Journal of the American Society for Information Science*, 60(5), 911–932. <http://doi.org/10.1002/asi>
- Papacharissi, Z. (2009). The virtual geographies of social networks: a comparative analysis of Facebook, LinkedIn and ASmallWorld. *New Media & Society*, 11(1–2), 199–220. <http://doi.org/10.1177/1461444808099577>
- Papadopoulos, F., Kitsak, M., Serrano, M. Á., Boguñá, M., & Krioukov, D. (2012). Popularity versus similarity in growing networks. *Nature*, 489(7417), 537–40. <http://doi.org/10.1038/nature11459>
- Pattison, P. (1996). Logit models and logistic regressions for social networks. *Psychometrika*, 61(3), 401–425. Retrieved from <http://www.springerlink.com/index/T2W46715636R2H11.pdf>
- Paul, T., & Puscher, D. (2011). Improving the Usability of Privacy Settings in Facebook. *Arxiv Preprint arXiv:1109.6046*. Retrieved from <http://arxiv.org/abs/1109.6046>
- Perkel, D. (2006). Copy and Paste Literacy: Literacy practices in the production of a MySpace profile. *Informal Learning and Digital Media*, 2009(April 28), 203–224. Retrieved from http://people.ischool.berkeley.edu/~dperkel/writing/perkel_copypasteliteracyDRAFT_August2007.pdf
- Pew, B. Y., & Project, S. J. (2013). The Facebook News Experience. *Econstor*, 1–40.
- Phillip, A. (The W. P. (n.d.). Online “authenticity” and how Facebook’s “real name” policy hurts Native Americans. Retrieved December 6, 2015, from <https://www.washingtonpost.com/news/morning-mix/wp/2015/02/10/online-authenticity-and-how-facebooks-real-name-policy-hurts-native-americans/>
- Pouwelse, J. A., Garbacki, P., Wang, J., Bakker, A., Yang, J., Iosup, A., ... Sips, H. J. (2008). TRIBLER: a social-based peer-to-peer system. *Concurrency and Computation: Practice and Experience*, 20(2), 127–138. <http://doi.org/10.1002/cpe.1189>
- Pring, C. (2012). 100 more social media statistics for 2012. Retrieved May 10, 2012, from <http://thesocialskinny.com/100-social-media-statistics-for-2012/>
- Pujol, N. (2010). Freemium: Attributes of an Emerging Business Model. *SSRN Electronic Journal*, (December), 1–4. <http://doi.org/10.2139/ssrn.1718663>
- Rahman, R. (2011). *Peer-to-Peer System Design: A Socioeconomic Approach*. Delft University of Technology.
- Rainie, L., Brenner, J., & Purcell, K. (2012). Photos and Videos as Social Currency Online shares of men and women. *Pew Research Center's Internet & American Life Project*.

- Rainie, L., Smith, A., & Duggan, M. (2013). Coming and going on Facebook. *Pew Research Center's Internet & American Life Project*. Retrieved from http://www.winthropmorgan.com/wp-content/uploads/2013/03/Education-Advocacy-Toolkit/PIP_Coming_and_going_on_facebook.pdf
- Reichardt, J., & Bornholdt, S. (2006). Statistical mechanics of community detection. *Physical Review E* 74.1., (16110), 1–16. Retrieved from <http://pre.aps.org/abstract/PRE/v74/i1/e016110>
- Ryan, J. A. (2008). The Virtual Campfire : An Ethnography of Online Social Networking. *I Can, Master of*(May), 200. Retrieved from <http://www.thevirtualcampfire.org/thevirtualcampfiresm.pdf>
- S.B. Caldwell. (1997). Dynamic Microsimulation and the Corsim 3.0 Model. *Ithaca, NY: Strategic Forecasting*.
- Sabatini, F., & Sarracino, F. (2013a). Specification of Tribler Anonymity. *Econstor*, 1–40. Retrieved from <http://hdl.handle.net/10419/88145>
- Sabatini, F., & Sarracino, F. (2013b). Tor: Overview. *Econstor*, 1–40. Retrieved from <http://hdl.handle.net/10419/88145>
- Sabatini, F., & Sarracino, F. (2013c). Towards Anonymity. *Econstor*, 1–40. Retrieved from <http://hdl.handle.net/10419/88145>
- Sala, A., Cao, L., Wilson, C., & Zablitz, R. (2010). Measurement-calibrated graph models for social network experiments. *Proceedings of the 19th International Conference on World Wide Web WWW 10 (2010)*, (May). <http://doi.org/10.1145/1772690.1772778>
- Schnettler, S. (2009). A small world on feet of clay? A comparison of empirical small-world studies against best-practice criteria. *Social Networks*, 31(3), 179–189. <http://doi.org/10.1016/j.socnet.2008.12.005>
- Schollmeier, R., & Universität, T. (2001). A Definition of Peer-to-Peer Networking for the Classification of Peer-to-Peer Architectures and Applications. *In Peer-to-Peer Computing, 2001. Proceedings. First International Conference On. IEEE*, 101–102.
- Seder, J. P., & Oishi, S. (2009). Ethnic/racial homogeneity in college students' Facebook friendship networks and subjective well-being. *Journal of Research in Personality*, 43(3), 438–443. <http://doi.org/10.1016/j.jrp.2009.01.009>
- Simon, H. a. (2000). Bounded rationality in social science: Today and tomorrow. *Mind & Society*, 1(1), 25–39. <http://doi.org/10.1007/BF02512227>
- Singer, H. M., Singer, I., & Herrmann, H. J. (2009). Agent-based model for friendship in social networks. *Physical Review. E, Statistical, Nonlinear, and Soft Matter Physics*, 80(2 Pt 2), 26113. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/19792206>
- Skype. (n.d.). Retrieved from <https://en.wikipedia.org/wiki/Skype>
- Snijders, T. a. B. (1996). Stochastic actor-oriented models for network change. *The Journal of Mathematical Sociology*, 21(1–2), 149–172. <http://doi.org/10.1080/0022250X.1996.9990178>
- Stutzman, F. (2006). Student Life on the Facebook, 14, 2006. Retrieved from <http://chimprawk.blogspot.com/2006/01/student-life-on-facebook.html>
- Stutzman, F., Gross, R., & Acquisti, A. (2012). Silent Listeners: The Evolution of Privacy and Disclosure on Facebook. *Journal of Privacy and Confidentiality*, (2), 7–41. Retrieved from <http://hdl.handle.net/10419/88145>
- Tang, C., & Ross, K. (2011). What's in a name: A study of names, gender inference, and gender behavior in facebook. *Database Systems for Adanced* Retrieved from <http://www.springerlink.com/index/P23X88V75664M167.pdf>
- Teece, D. J. (2010). Business models, business strategy and innovation. *Long Range Planning*, 43(2–3), 172–194. <http://doi.org/10.1016/j.lrp.2009.07.003>
- Tiggemann, M., & Slater, A. (2013). NetGirls: the internet, Facebook, and body image concern in adolescent girls. *The International Journal of Eating Disorders*, 46(6), 630–3.

<http://doi.org/10.1002/eat.22141>

- Traud, A. L., Kelsic, E. D., Mucha, P. J., & Porter, M. a. (2008). Comparing Community Structure to Characteristics in Online Collegiate Social Networks. *SIAM Review*, 53(3), 17. <http://doi.org/10.1137/080734315>
- Uddin, S., Khan, A., & Piraveenan, M. (2015). A Set of Measures to Quantify the Dynamicity of Longitudinal Social Networks. *Complexity*. <http://doi.org/10.1002/cplx>
- Ugander, J., Karrer, B., Backstrom, L., Marlow, C., & Alto, P. (2011). *The Anatomy of the Facebook Social Graph*. *Arxiv preprint arXiv* (Vol. abs/1111.4). Retrieved from <http://dx.doi.org/10.1006/jcss.1995.1065>
- University, C. (n.d.). Caltech at a Glance. Retrieved December 13, 2015, from <https://www.caltech.edu/content/caltech-glance>
- Valenzuela, S., Park, N., & Kee, K. F. (2009). Is There Social Capital in a Social Network Site?: Facebook Use and College Students' Life Satisfaction, Trust, and Participation. *Journal of Computer-Mediated Communication*, 14(4), 875–901. <http://doi.org/10.1111/j.1083-6101.2009.01474.x>
- Vazquez, A. (2002). Growing networks with local rules: Preferential attachment, clustering hierarchy and degree correlations. *Arxiv Preprint Cond-mat/0211528*. Retrieved from <http://arxiv.org/abs/cond-mat/0211528>
- Vitak, J. (2008). *Facebook Friends": How Online Identities Impact Offline Relationships*. *Aladinrcwrlcorg*. Georgetown University. Retrieved from <http://aladinrc.wrlc.org/dspace/handle/1961/4433>
- Voulgaris, S. (2006). Epidemic-Based Self-Organization in Peer-to-Peer Systems. *These de Doctorat, VU University, Amsterdam*. Retrieved from http://www.cs.vu.nl/en/Images/S_Voulgaris_03-10-2006_tcm75-258074.pdf
- Wang, G., Wang, B., Wang, T., Nika, A., Zheng, H., & Zhao, B. Y. (2014). Whispers in the Dark : Analysis of an Anonymous Social Network. In *IMC* (pp. 137–149). <http://doi.org/10.1145/2663716.2663728>
- Wasserman, S., & Faust, K. (1994). *Social network analysis: Methods and applications*. Retrieved from http://books.google.com/books?hl=en&lr=&id=CAm2DplqRUIC&oi=fnd&pg=PR21&dq=Social+Network+Analysis:+Methods+and+Applications&ots=HvFnviYGNh&sig=7539mD0kW_db48J2UTzfEaUNFPc
- Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of "small-world" networks. *Nature*, 393(6684), 440–2. <http://doi.org/10.1038/30918>
- Wellman, B., & Potter, S. (1999). The elements of personal communities. *Networks in the Global Village*. Retrieved from <http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:The+elements+of+personal+communities#0>
- Wikipedia. (n.d.-a). Global surveillance disclosures (2013–present). Retrieved from [https://en.wikipedia.org/wiki/Global_surveillance_disclosures_\(2013–present\)](https://en.wikipedia.org/wiki/Global_surveillance_disclosures_(2013–present))
- Wikipedia. (n.d.-b). Global surveillance disclosures (2013–present).
- Wikipedia contributors. (n.d.). Freemium. Retrieved December 20, 2016, from <https://en.wikipedia.org/wiki/Freemium>
- Wilson, C., Boe, B., Sala, A., Puttaswamy, K. P. N., & Zhao, B. Y. (2009). User interactions in social networks and their implications. *Proceedings of the Fourth ACM European Conference on Computer Systems - EuroSys '09*, 205. <http://doi.org/10.1145/1519065.1519089>
- Wilson, F. (2006). My Favorite Business Model. Retrieved December 20, 2016, from http://avc.com/2006/03/my_favorite_bus/

- Young, K. (2011). Social ties, social networks and the Facebook experience. *International Journal of Emerging Technologies and Society*, 9(1), 20–34. Retrieved from [http://www.swinburne.com/hosting/ijets/journal/V9N1/pdf/Article 2 Young.pdf](http://www.swinburne.com/hosting/ijets/journal/V9N1/pdf/Article%20Young.pdf)
- Yu, H., & Kaminsky, M. (2008). SybilGuard: Defending against Sybil Attack via Social Networks. *IEEE/CM Trans- Actions on Networking*, 16(3), 576–589. <http://doi.org/10.1145/1151659.1159945>
- Yule, G. U. (1925). A Mathematical Theory of Evolution, Based on the Conclusions of Dr. J. C. Willis, F.R.S. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 213(402–410), 21–87. <http://doi.org/10.1098/rstb.1925.0002>
- Zhao, S., & Grasmuck, S. (2008). Identity construction on Facebook: Digital empowerment in anchored relationships. *Computers in Human Behavior*, 24, 1816–1836. Retrieved from <http://www.sciencedirect.com/science/article/pii/S0747563208000204>
- Zuckerberg, M. (n.d.). Facebook blog: An Open Letter from Mark Zuckerberg. Retrieved from <http://blog.facebook.com/blog.php?post=2208562130>

9 Glossary of Terms

ABM: Agent Based Modelling

Affinity: It measures the ratio of the fraction of links between attribute-sharing nodes, relative to what would be expected if attributes were random

CEL: Cultural-Ethno-Linguistic

Degree Homophily: Love of the same degree. Nodes connecting with other nodes with similar number of links (degree).

Homophily: Love of the same

LGBT: Lesbian, Gay, Bisexual, and Transgender

MMU: Manchester Metropolitan University

P2P: Peer-to-Peer

SNA: Social Network Analysis

SNS: Social Network System

Appendix A Student Interaction Model

In this section, we will highlight the main piece of codes used in Student Interaction Agent-Based Model.

Appendix A.1 Main Step Function

We list down the main step function of our [ABM](#) model below. It shows how each strategy from the point view of the source agent is read out and then the suitable target agent is chosen to form a link.

```
// Run the step function at every tick
@ScheduledMethod(start = 0, interval = 1, priority = 0)
public void step() {

    //Find the context this person exists in.
    Context<StudentAgent> context = ContextUtils.getContext (this);

    //Network of students
    Network friends = (Network)context.getProjection("StudentNetwork");

    // Get an instance of parameter from the RunEnvironment
    Parameters param = RunEnvironment.getInstance().getParameters();
    // Get the simulation mode (1-4)
    simulation_mode = (Integer)param.getValue("simulation_mode");
    // Get the dorm preference for the personal preference
    double dormPreference = (Double)param.getValue("dormPreference");
    // Get the dorm preference for the personal preference
    double majorPreference = (Double)param.getValue("majorPreference");
    // Get the dorm preference for the personal preference
    double yearPreference = (Double)param.getValue("yearPreference");
    // Get the dorm preference for the personal preference
    double highSchoolPreference = (Double)
param.getValue("highSchoolPreference");
    // Get the initial number of links to be developed using Random
Strategy (1) for FOAF Strategy (2)
    int randomFriends = (Integer)param.getValue("initialRandomFriends");
    // Set the targetStudent to null
    StudentAgent targetStudent = null;
    // Check if simulation should proceed - the total number of links
hasn't reached the total links of the reference dataset
    if (checkMean()){
    // If strategy is Preferential Strategy (1)
    if (simulation_mode == 1){
        // Pick a random agent and call it targetStudent
        targetStudent = (StudentAgent)context.getRandomObject();
        // If random agent is the same as the self agent, then return
        if (targetStudent == this)
            return;
        // Else check the source and the target compatibility by the
personal preference algorithm. If it does not comply, then return
        if (!isCompatibleFriendship(this, targetStudent)){
            return;
        }else {
        // Link with targetStudent
        friends.addEdge(this, targetStudent);
        }
    }
}
```

```

// If strategy is FOAF strategy (2)
else if (simulation_mode == 2){
    // If it is second phase, pick a friend of a friend (FOF)
    if (getTotalFriendsCount() > randomFriends){//RANDOM_FRIENDS){
        // Search a new friend in foaf
        targetStudent = getFOAF();
        // Link up with the targetStudent
        friends.addEdge(this, targetStudent);
    }else{ // If it is the first phase, pick a random targetStudent as
the target agent
        targetStudent = (StudentAgent)context.getRandomObject();
        // Check if the target and the source agents are compatible. If
not, then return
        if (!isCompatibleFriendship(this, targetStudent)){
            return;
        }else {
            // Link with targetStudent
            friends.addEdge(this, targetStudent);
        }
    }
}
// If strategy is party strategy (3)
}else if (simulation_mode == 3){
    hasPartyStarted = true;
    // Do nothing - since startpartying is scheduled to run after a
predefined interval of ticks (10).
    // If strategy is hybrid strategy (4)
}else if (simulation_mode == 4){
    // Create a random object
    Random r = new Random();
    // Pick a random number between 0 and 2
    int randomSimulationMode = r.nextInt(2);
    // Since the mode runs from 0-2, we need to add 1 more to it to
adjust with the current simulation modes (1-3)
    randomSimulationMode ++;
    // Local variable to hold the current simulation mode
    subSimulationMode = randomSimulationMode;
    // Set a counter for strategy for the three strategies
    if (randomSimulationMode == 1){
        this.TOTAL_MODE_1++;
    }else if (randomSimulationMode == 2){
        this.TOTAL_MODE_2++;
    }

    }else if (randomSimulationMode == 3){
        this.TOTAL_MODE_3++;
    }
}
// If the random strategy is strategy 1 (random)
if (randomSimulationMode == 1){
    targetStudent = (StudentAgent)context.getRandomObject();
    if (!isCompatibleFriendship(this, targetStudent))
        return;
}
// If it is FOAF strategy (2), repeat the same logic of FOAF
strategy
else if (randomSimulationMode == 2){
    if (getTotalFriendsCount() > randomFriends){//RANDOM_FRIENDS){
        // Search a new friend in foaf
        targetStudent = getFOAF();
    }else{
        targetStudent = (StudentAgent)context.getRandomObject();
        if (!isCompatibleFriendship(this, targetStudent))

```

```

        return;
    }
} else if (randomSimulationMode == 3){
    // Start party
    startpartying();
}
}
}
}

```

Appendix A.2 Personal Preference Algorithm

Below we show how the personal preference algorithm, which is the core of our [ABM](#), is coded in Java.

```

// Check if source (a) and target (b) are compatible. If they are, return
// true, else return false
private boolean isCompatibleFriendship(StudentAgent a, StudentAgent b){
    Parameters param = RunEnvironment.getInstance().getParameters();
    // Get the dorm preference for the personal preference
    double dormPreference = (Double)param.getValue("dormPreference");
    // Get the dorm preference for the personal preference
    double majorPreference = (Double)param.getValue("majorPreference");
    // Get the dorm preference for the personal preference
    double yearPreference = (Double)param.getValue("yearPreference");
    // Get the dorm preference for the personal preference
    double highSchoolPreference = (Double)
param.getValue("highSchoolPreference");
    // To initialize set all booleans for each attribute to false
    boolean sameDorm = false, sameMajor = false, sameYear = false,
sameHighSchool = false;
    // Create a random object
    Random r = new Random();
    // Pick a first random number for Dorm Preference (DP)
    int randomType = r.nextInt(100);
    // See if randomType is less than or equal to dormPreference
    if (randomType <= dormPreference){
        // Check if both a's and b's dorm preference are same and are not
missing (0)
        if (a.getDorm().equals(b.getDorm()) && (!(a.getDorm().equals("0")
&& b.getDorm().equals("0")))){
            // If satisfied, set boolean sameDorm to true
            sameDorm = true;
        }

    } else{ // Does not matter - any dorm would do
        sameDorm = true;
    }
    // Pick a first random number for Major Preference (MP)
    randomType = r.nextInt(100);
    // See if randomType is less than or equal to majorPreference
    if (randomType <= majorPreference){
        // Check if both a's and b's major attributes are same and are not
missing (0)
        if (a.getMajor().equals(b.getMajor()) &&
(!a.getMajor().equals("0") && b.getMajor().equals("0"))){
            // If satisfied, set boolean sameMajor to true
            sameMajor = true;
        }
    } else{ // Does not matter - any major would do

```

```

        sameMajor = true;
    }
    // Pick a first random number for Year Preference (MP)
    randomType = r.nextInt(100);
    // See if randomType is less than or equal to yearPreference
    if (randomType <= yearPreference ){
        // Check if both a's and b's year attributes are same and are not
missing (0)
        if (a.getYear() == b.getYear() && !(a.getYear() == 0 &&
b.getYear() == 0)){
            // If satisfied, set boolean sameYear to true
            sameYear = true;
        }
    }else{ // Does not matter - any year would do
        sameYear = true;
    }
    // Pick a first random number for High School Preference (HSP)
    randomType = r.nextInt(100);
    // See if randomType is less than or equal to highSchoolPreference
    if (randomType <= highSchoolPreference){
        // Check if both a's and b's high school attributes are same and
are not missing (0)
        if (a.getHighSchool().equals(b.getHighSchool()) &&
(!a.getHighSchool().equals("0") && b.getHighSchool().equals("0"))){
            // If satisfied, set boolean sameHighSchool to true
            sameHighSchool = true;
        }
    }else{ // Does not matter - any high school would do
        sameHighSchool = true;
    }
    // If all conditions have satisfied for all the four attributes
    if (sameMajority && sameMinority && sameDorm && sameMajor && sameYear
&& sameHighSchool){
        // Return true
        return true;
    }
    // Else return false
    return false;
}
}

```

Appendix B [MMU](#) Facebook Dataset

In this appendix, we will give more details of the Facebook dataset that we have collected. For each of the inferred attribute: ethnicity, sub-ethnicity, religion, language and geography, we are going to show the overall population.

Appendix B.1 Ethnic Group Distribution

Ethnic Group	Population
English	324029
European	138042
Celtic	102790
Hispanic	67050
Muslim	54293
Unclassified	49580
East Asian & Pacific	22684
South Asian	21709
Nordic	12824
Jewish And Armenian	9639
Greek	9604
International	7357
African	5913
Void	4750
Sikh	3432
Japanese	1930

Appendix B.2 Sub-ethnic Group Distribution

Sub-ethnic Group	Population
English	323549
Italian	63987
Unclassified	49580
Spanish	41426
Celtic	32885
Irish	31330
Scottish	26943

Muslim	20284
Portuguese	17687
Pakistani	17146
European	15804
Indian Hindi	14672
Polish	12906
German	12228
Welsh	11632
French	11078
Greek	9604
Hong Kongese	8274
Hispanic	7937
Jewish	7734
International	7357
East Asian & Pacific	5562
Chinese	5301
Balkan	5290
Void	4750
Swedish	4680
Pakistani Kashmir	4065
Turkish	3982
Serbian	3965
Somalian	3825
Russian	3636
Sikh	3432
South Asian	3393
Bangladeshi	3148
Danish	2685
Vietnamese	2605
Nigerian	2280
Finnish	2164
Sri Lankan	2145
Czech	2126

Ghanaian	1974
Japanese	1930
Romanian	1864
Nordic	1776
Norwegian	1519
Jewish And Armenian	1509
Hindi Not Indian	1499
Hungarian	1225
Dutch	1107
African	999
Albanian	871
Black Caribbean	832
Iranian	695
South Korean	602
Afrikaans	561
Ukranian	558
Lebanese	515
Baltic	484
Armenian	396
Malaysian	340
Muslim North African	263
Black Southern African	225
Sierra Leonian	201
Eritrean	187
Ethiopian	180
Muslim Middle East	117
Muslim Stans	66
Congolese	45
Ugandan	9

Appendix B.3 Geography Distribution

Geography	Population
British Isles	425985
Southern Europe	137974
Not Applicable	54330
South Asia	51291
Central Europe	40213
Eastern Europe	32912
East Asia	25606
Middle East	23694
Northern Europe	12824
Africa	10861
Diasporic	9243
Unclassified	7357
Americas	2858
Central Asia	478

Appendix B.4 Religion Distribution

Geography	Population
Christian: Protestant	295482
Muslim	107738
Christian	49372
Christian: Catholic	42240
Not Applicable	27064
Hindu	16574
Bhuddist	12104
Void	3771
Christian: Greek Orthodox	2698
Jewish	1762
Christian: Russian Orthodox	1113
Christian: Orthodox_Calcedonian	144

Appendix B.5 Language Distribution

Language Group	Population
English	415394
Not Applicable	65790
Italian	63987
Spanish	47126
German	28012
Punjabi	22387
Hindi	19480
Arabic	19337
Portuguese	17687
Polish	12906
Welsh	11632
French	11078
Chinese, Mandarin	9861
Greek	9604
Chinese, Cantonese	8274
Serbian	7896
Hebrew	7694
Swedish	4680
Kashmiri	4065
Turkish	3982
Somali	3825
Russian	3619
Bengali	3243
Danish	2685
Vietnamese	2605
Yoruba	2280
Finnish	2164
Sinhala	2145
Akan	1974
Czech	1944
Japanese	1930

Romanian	1864
Norwegian	1519
Hungarian	1225
Filipino	962
Dutch	872
Albanian	871
Farsi	756
Basque	613
Croatian	602
Korean	602
Catalan	591
Afrikaans	561
Ukrainian	558
Chinese, Min Nan	433
Armenian	396
Malay	370
Maltese	352
Lithuanian	322
Burmese	320
Bosnian	307
Vlaams	235
Zulu	225
Macedonian	221
Slovenian	192
Tigrž	187
Slovak	182
Amharic	180
Javanese	106
Estonian	106
Gikuyu	71
Galician	71
Thai	67
Latvian	55
Luba-Kasai	45

Icelandic	44
Ladino	40
Nepali	40
Bulgarian	40
Wolof	21
Schwyzerdütsch	20
Tahitian	18
Bemba	15
Shona	11
Ganda	9
Georgian	9
Baoulž	8
Seselwa Creole	7
French	
Azerbaijani, North	7
Fulfulde	4
Turkmen	3
Swati	2
Belarusan	1
Kirghiz	1
Kazakh	1

Appendix C [MMU](#) Facebook Graph vs. Random Graph

In this appendix, we will give more details of the Facebook graph when compared with the random graph (null model). For each of the inferred attribute: ethnicity, sub-ethnicity, religion, language and geography, we are going to show Silo Indices.

Appendix C.1 Ethnic Group Distribution

<i>Ethnic Groups</i>	<i>Ref</i>	<i>NM</i>
English	-0.22015	-0.3885
East Asian & Pacific	-0.89127	-0.98719
Greek	-0.94472	-0.99588
Muslim	-0.26604	-0.78967
Hispanic	-0.94895	-0.98368
African	-0.86507	-0.98535
European	-0.94723	-0.96867
South Asian	-0.66069	-0.96146
Celtic	-0.74119	-0.83172
Unclassified	-0.94167	-0.96128
Jewish And Armenian	-0.99559	-0.99517
Japanese	-0.99833	-1
Sikh	-0.84369	-0.9849
International	-0.99709	-0.99718
Void	-0.99028	-0.9949
Nordic	-0.99668	-0.99772

Appendix C.2 Sub-ethnic Groups

<i>Ethnic Sub-Groups</i>	<i>Ref</i>	<i>NM</i>
English	-0.22567	-0.39025
Hong Kongese	-0.94256	-0.99428
Greek	-0.94472	-0.99588
Bangladeshi	-0.93124	-0.98773
Pakistani Kashmir	-0.96992	-0.98724
Malaysian	-1	-1
Pakistani	-0.62179	-0.91287
Spanish	-0.9749	-0.99076
East Asian & Pacific	-0.95138	-0.99639
Muslim	-0.82421	-0.93578
African	-0.98395	-0.99731
European	-0.98829	-0.99503
Hindi Not Indian	-0.98905	-0.99845
Irish	-0.90425	-0.9593
Ethiopian	-1	-1
Scottish	-0.96199	-0.97369
Unclassified	-0.94167	-0.96128
Jewish And Armenian	-0.9971	-0.99877
Portuguese	-0.97988	-0.99427
Vietnamese	-0.95389	-0.9984
Celtic	-0.90017	-0.93389
Jewish	-0.99614	-0.99696
Indian Hindi	-0.66473	-0.9718
German	-0.99353	-0.99697
French	-0.99323	-0.99619

Japanese	-0.99833	-1
Welsh	-0.96109	-0.97635
Sikh	-0.84369	-0.9849
International	-0.99709	-0.99718
Polish	-0.97589	-0.99407
Somalian	-0.94381	-0.99079
Chinese	-0.98932	-0.99688
Void	-0.99028	-0.9949
Hispanic	-0.99183	-0.998
South Asian	-0.98489	-0.99372
Nigerian	-0.87327	-0.99323
Russian	-0.99864	-0.99876
Black Caribbean	-0.9905	-0.99829
Ghanaian	-0.95762	-0.99776
Turkish	-0.92032	-0.99204
Afrikaans	-1	-1
Danish	-1	-1
Iranian	-0.98143	-0.99796
Sierra Leonian	-0.99191	-0.9973
Balkan	-1	-0.99878
Muslim Middle East	-0.99777	-1
Italian	-0.97165	-0.99091
Finnish	-0.9979	-1
Eritrean	-0.99296	-1
Romanian	-1	-1
Hungarian	-1	-1

Sri Lankan	-0.99335	-1
Serbian	-1	-1
Black Southern African	-0.94513	-1
Dutch	-0.9889	-0.99801
Swedish	-1	-1
Nordic	-1	-1
Norwegian	-0.99757	-1
Lebanese	-0.98928	-0.99858
Armenian	-1	-1
South Korean	-1	-0.99791
Ukranian	-1	-1
Czech	-1	-1
Muslim North African	-0.99472	-1
Albanian	-1	-1
Muslim Stans	-1	-0.99766
Congolese	-0.99823	-1
Baltic	-0.98502	-1
Ugandan	-1	-1

Appendix C.3 Religion

<i>Religious Groups</i>	<i>Ref</i>	<i>NM</i>
Christian: Protestant	-0.0996	-0.29095
Bhuddist	-0.90276	-0.97963
Christian: Greek Orthodox	-0.94713	-0.99483
Muslim	-0.26887	-0.78872

Christian: Catholic	-0.87686	-0.92357
Christian	-0.8683	-0.91075
Hindu	-0.66275	-0.96952
Not Applicable	-0.93067	-0.95093
Jewish	-0.99614	-0.99696
Sikh	-0.84369	-0.9849
Christian: Russian Orthodox	-0.99709	-0.998
Christian: Orthodox_Calcedonian	-1	-1

Appendix C.4 Geography

<i>Geography Groups</i>	<i>Ref</i>	<i>NM</i>
British Isles	0.259632	-0.09487
East Asia	-0.89331	-0.9857
Southern Europe	-0.92765	-0.97096
South Asia	-0.41386	-0.82064
Middle East	-0.82144	-0.9314
Africa	-0.88511	-0.97636
Central Europe	-0.98079	-0.98925
Not Applicable	-0.93561	-0.95706
Diasporic	-0.99549	-0.99566
Unclassified	-0.99709	-0.99718
Eastern Europe	-0.9728	-0.99072
Americas	-0.99082	-0.9984
Northern Europe	-0.99668	-0.99772
Central Asia	-0.99862	-0.99761

Appendix C.5 Language

<i>Language Group</i>	<i>Ref</i>	<i>NM</i>
English	0.19994	-0.14089
Chinese, Cantonese	-0.94256	-0.99428
Greek	-0.94472	-0.99588
Bengali	-0.93302	-0.98745
Kashmiri	-0.96992	-0.98724
Malay	-1	-1
Punjabi	-0.58752	-0.89201
Spanish	-0.96733	-0.98933
Chinese, Mandarin	-0.94269	-0.99496
Arabic	-0.83569	-0.93973
Not Applicable	-0.92431	-0.9477
German	-0.98533	-0.99299
Hindi	-0.67665	-0.96305
Amharic	-1	-1
Portuguese	-0.97988	-0.99427
Vietnamese	-0.95389	-0.9984
Hebrew	-0.99613	-0.99695
French	-0.99322	-0.99618
Japanese	-0.99833	-1
Welsh	-0.96109	-0.97635
Polish	-0.97589	-0.99407
Somali	-0.94381	-0.99079

Yoruba	-0.87327	-0.99323
Galician	-1	-1
Russian	-0.99862	-0.99937
Akan	-0.95762	-0.99776
Chinese, Min Nan	-1	-1
Turkish	-0.92032	-0.99204
Afrikaans	-1	-1
Danish	-1	-1
Farsi	-0.98496	-0.99737
Serbian	-1	-0.99856
Gikuyu	-1	-1
Italian	-0.97165	-0.99091
Finnish	-0.9979	-1
Tigr?	-0.99296	-1
Romanian	-1	-1
Hungarian	-1	-1
Sinhala	-0.99335	-1
Zulu	-0.94513	-1
Basque	-1	-1
Vlaams	-0.92188	-1
Bulgarian	-1	-1
Swedish	-1	-1
Norwegian	-0.99757	-1
Dzongkha	-1	-1
Armenian	-1	-1
Korean	-1	-0.99791

Burmese	-0.99263	-1
Dutch	-1	-1
Ukrainian	-1	-1
Czech	-1	-1
Albanian	-1	-1
Luba-Kasai	-0.99824	-1
Catalan	-1	-1
Maltese	-1	-0.99632
Slovak	-1	-1
Filipino	-1	-1
Shona	-1	-1
Lithuanian	-0.97927	-1
Croatian	-1	-0.99455
Bemba	-1	-1
Bosnian	-1	-1
Ladino	-1	-1
Thai	-1	-1
Latvian	-1	-1
Estonian	-1	-1
Georgian	-1	-1
Ganda	-1	-1
Macedonian	-1	-1
Slovenian	-1	-1
Javanese	-1	-1
Wolof	-1	-1
Tahitian	-1	-1

Nyanja	-1	-1
Baoul?	-1	-1
Schwyzerdütsch	-1	-1
Nepali	-1	-1
Azerbaijani, North	-1	-1
Seselwa Creole French	-1	-1
Fulfulde	-1	-1
Belarusan	-1	-1

Appendix D MMU Facebook Graph Normalised Silo Index

This appendix covers normalized silo index for religion, geography and language.

Appendix D.1 Religious Group Normalised Silo Index

<i>Religion</i>	<i>Normalised Silo Index</i>
Hindu	0.9
Christian: Greek Orthodox	0.88
Sikh	0.87
Muslim	0.75
Buddhist	0.73
Christian: Protestant	0.48
Christian: Catholic	0.39
Christian	0.36
Not Applicable	0.33
Christian: Russian Orthodox	0.32
Jewish	0.24
Christian: Orthodox_Calcedonian	-1

Appendix D.2 Geography Group Normalised Silo Index

<i>Religion</i>	<i>Normalized Silo Index</i>
East Asia	0.83
Americas	0.76
Africa	0.74
South Asia	0.72
Middle East	0.59
British Isles	0.58

Eastern Europe	0.57
Southern Europe	0.56
Central Asia	0.53
Central Europe	0.36
Not Applicable	0.35
Diasporic	0.17
Northern Europe	0.16
Unclassified	-0.04

Appendix D.3 Language Group Normalised Silo Index

<i>Language Group</i>	<i>Normalized Silo Index</i>
Vlaams	1
Lithuanian	0.99
Zulu	0.99
Burmese	0.96
Vietnamese	0.96
Tigr?	0.95
Yoruba	0.92
Akan	0.91
Greek	0.89
Farsi	0.88
Chinese, Mandarin	0.87
Hindi	0.87
Turkish	0.87
Chinese, Cantonese	0.86
Luba-Kasai	0.84
Somali	0.79
Bengali	0.75
Punjabi	0.73
Norwegian	0.71
Polish	0.7
Portuguese	0.7
Finnish	0.69
Sinhala	0.67
Spanish	0.63
Italian	0.62

Arabic	0.6
English	0.57
Japanese	0.56
Kashmiri	0.54
German	0.42
French	0.4
Welsh	0.37
Not Applicable	0.34
Russian	0.27
Hebrew	0.24

Appendix E R Code

For our work, we have relied on many statistical and visualizations tools such as ORA, Gephi and R. In this section we will share some of the code of our algorithms written in R.

Appendix E.1 Silo Index

```
# Get Silo Index
# Get graph and column name as inputs
getSiloIndex <- function(graph, colName){
  # Identify the unique column values for colName
  unique_col <- unique(get.vertex.attribute(graph, colName))
  # Initialize matrix M which is of dimension: unique_col x 2
  m=matrix(nrow=length(unique_col),ncol=2)
  # For each of the unique value, calculate Silo Index
  for (j in 1:length(unique_col)){
    # Get current unique value
    val <- unique_col[j]
    # Calculate how many nodes in the graphs have this unique
    value for colName
    tmp1 <- which(get.vertex.attribute(graph, colName) == val)
    # Calculate how many nodes in the graphs have some other value
    other than unique value for colName
    tmp2 <- which(get.vertex.attribute(graph, colName) != val)
    # Calculate how many Internal (I) links are there
    i <-length(E(graph)[tmp1 %--% tmp1])
    # Calculate how many External (E) links are there
    e <- length(E(graph)[tmp1 %--% tmp2])
    # Calculate the Silo Index for the unique value
    si <- (i-e)/(i+e)
    # Saving the unique value in the matrix at column 1
    m[j, 1]= val
    # In the same row, save the Silo Index for the unique value at
    column 2 [j, 2]= si
  }
  # Return the matrix which contains Silo Indices for the whole
  column (colName)
  m
}
```

Appendix E.2 Affinity Measure

This is the implementation of the affinity measure (Alan Mislove et al., 2007) in R.

```
# Get Affinity
# Supply an undirected graph to it
getAffinity<- function(undg){
# Get a list of vertex attributes
ls_attr<-list.vertex.attributes(undg)
# Create a matrix with number of vertex attributes x 2
m<-matrix(nrow=length(ls_attr),ncol=2)
# Count number of edges (links)
E <- ecount(undg)
# Count number of vertices (nodes)
V <- vcount(undg)
# Loop through the list of attributes
for (i in 1:length(ls_attr)){
# Initialize local variables
total_vcount <- 0
# Si is the total number of matched nodes with the same
attribute values for an attribute. Initialize it by 0
Si <- 0
# Ei which represents the expected value when attributes are
randomly assigned. Initialize it by 0
Ei <- 0
# Sort unique values for a particular attribute (say
dormitory)
unique_rel <- sort(unique(get.vertex.attribute(undg,
ls_attr[i])))
# For each unique value for a column, iterate through it
  for (j in 1:length(unique_rel)){
    # Identify the set which has the same value for both lists
indices (ls_attr[i]) == unique_rel[j])
    tmp <- which(get.vertex.attribute(undg, ls_attr[i]) ==
unique_rel[j])
    # Count how many links are there in total tmp set
ecount_rel <- length(E(undg)[tmp %--% tmp])
    # Count how many nodes are there in total tmp set
vcount_rel <- length(V(undg)[tmp])
    # Increase the total counter of nodes by adding vcount_rel
to it
    total_vcount <- total_vcount + vcount_rel
    # Add ecount_rel to local Si
Si = Si + ecount_rel
    # Add local Ei to global Ei
Ei = Ei + (vcount_rel*(vcount_rel-1))
  }
# Calculate global Si
Si <- Si/E
# Calculate global Ei
Ei <- Ei / (V * (V-1))
```

```

# Calculate Affinity A for a vertex i
A = Si/Ei
# Store the value vertex i to matrix
m[i,1] <- ls_attr[i]
# Also the calculated affinity too
m[i,2] <- A
}
# Return the overall matrix with Affinity measures of all
attributes of graph
m
}

```

Appendix E.3 Contract Graph

```

contract_graph <- function (graph, attribute){
  # Add a new vertex attribute called 'count' to the graph
  # Initialize its value by 1
  V(graph)$count <- 1
  # IGraph function which merges several vertices into one,
  # specified in the filter
  contracted_graph <- contract.vertices(graph,mapping =
  as.integer(
  as.factor(
  get.vertex.attribute(graph, attribute))),
  vertex.attr.comb <- list(count = "sum", attribute =
  toString, "ignore"))

  # Weight set 1
  E(contracted_graph)$weight <- 1
  # Remove duplicates and sum weigh
  contracted_graph <- simplify(contracted_graph)
  contracted_graph
}

```

Appendix F Annual Review Documents and Ethical Committee's letter

In this appendix we list down the documents used for the annual review, and also attach the letter from the chair of the Ethics Committee regarding our Facebook data.

Appendix F.1 Progress Report 2009-10

Progress REPORT 2009-2010

Name of candidate: Syed Muhammad Ali Abbas
 Director of Studies: Bruce Edmonds
 Supervisors: Bruce Edmonds, Emma Norling

Date of review meeting: June 2010
 Date of original registration: 01/07/09
 Date of MRes completion: n/a
 Date of PhD transfer: 01/07/09
 Expected PhD submission date: 30/06/12
 Current submission date: 30/06/12

Reviewer: Paul Brook

Area of study or thesis title:

Segregation and Diversity in Social Networking System (SNS)

How do you feel about your progress since the last review? (Please tick appropriate box) Disappointed / _____ / _____ / _____ / _____ / Pleased

n/a

Briefly Comment:

n/a

MRes Results

Final
mark

Literature Review
 Philosophy of Knowledge
 Qualitative Methods
 Quantitative Methods
 Subject disciplines
 Dissertation

Written research programme submitted to supervisor(s)

Discussed Agreed

literature review	✓	✓
research objectives	✓	✓
methodological issues	✓	✓
data sources	✓	✓
sample	✓	✓
data acquisition/collection	✓	✓
data analysis methods	✓	✓

results development
 results evaluation
 possible contributions

OUTPUT

Working Papers, thesis chapters or papers submitted in this academic year relevant to your PhD research.

Titles: 1. Syed Muhammad Ali Abbas (2010) A segregation model of Facebook, in 6th UK Social Network Conference (UK-SNA 2010), Manchester, UK.

PLEASE TICK APPROPRIATE BOX TO INDICATE COMPLETION OF TASKS

Literature Review – best estimate for completion of critical review to date 0% ___/___/___/___ ✓ /___100%

MAIN PHASES OF RESEARCH

You may like to consider how much you have done so far
 (Not all will apply so just indicate those that do by ticking the appropriate box)

Data Acquisition 0% 100%
 Data collection instrument ___/___/___/___ ✓ /___
 Access to sources assured ___/___/___/___ ✓ /___
 Acquisition of data completed ___/___/___ ✓ /___/___

Data Analysis 0% 100%
 Analysis of data completed ___/___/___ ✓
 Implications/interpretation/meaning of results thought through ___/___/___/___ ✓ /___
 Model development (if appropriate) ___/___/___ ✓ /___/___
 Conceptual development ___/___/___ ✓ /___/___

Thesis Chapters Drafted

Chapter T	Title	First Draft	Ready for thesis
1			
2		___/___/___	___/___/___
3		___/___/___	___/___/___
4		___/___/___	___/___/___
5		___/___/___	___/___/___
6		___/___/___	___/___/___
7		___/___/___	___/___/___
8		___/___/___	___/___/___
9		___/___/___	___/___/___

RESEARCH FORECAST

Future Timetable
 Work targets for next half year
 Student segregation model: Based on the Facebook data of US universities, I have developed a model of school

Is transfer to PhD requested at this Review?
 Yes No
 Already transferred
 Critical Problems

segregation. The analysis and the model to data matching have to be done. In the end, a paper will be developed out of this work – it is about to finish.

In the next half year

I have crawled Facebook and collected a dataset of a subset of it. This data not only has people's SN, but also contains their name and profile pictures. I need to identify racial background from these two information; and then eventually make a segregation model of a bigger dataset which is quite diverse. This work will be followed after the target 1

Further ahead

Getting acquainted with immigration and diversity research and theories – SCID project.

Outline timetable until thesis completed:

In the next 6 months, I will achieve all the three targets mentioned above. Along with that, based on the data I have collected so far, and the new data from the SCID project will help me identify the key parameters and factors involved in the segregation of a society – both in online and real world. Missing information by online questionnaires will be asked to get filled. Based on the outcomes of first set of models, and the missing information, fine course grained models will be established. After that, till the end of year 2, a detailed set of interviews, from a subset of the target network, will be carried out to grasp the micro level information from users - which couldn't be fetched earlier. A new set of model(s) will be developed, which would shed light on the social processes behind segregation and diversity. The analysis of the model(s) and thesis writing will take place in the last year – 3rd year.

Appendix F.2 RDF Form

Segregation in online Social Networking Systems (SNS)

Rationale

Humans are social animals: they can only achieve their goals collectively, they tend to develop and manage large and complex social groups, they have a fundamentally affiliation motive [1], and they need to share and gather information about the people and environment around us [2]. This motivates us to form Social groups, which forms the cornerstone of an individual's Social Capital [3].

The recent trend in the cyber world is the emergence of Social Networking Systems (SNS) and the subsequent transformation of Web 2.0. Very trivially, social networking could be described as the grouping of individuals into specific communities, like small rural communities or neighbourhood subdivisions. However, and this is an important point, while social networking exists and has always existed in these domains, the online world opens up new vistas for the formation of groups of magnitude and diversity, which would have previously been deemed unthinkable.

The first major SNS that attracted worldwide attention and proved itself a trailblazer, was Friendster. It popularized the features that define contemporary social network sites, such as profiles, public testimonials or comments, and public lists of friends [4]. Facebook being the most popular SNS has over 350 Million users to its credit [5]. In short, the impact of online SNS has been tremendous.

When the internet came to its existence, the idea of online communication between people arose. Popular views of the impact of the Internet on race dynamics, taking their lead from technological visionaries, conceived of the online world as a disembodied, utopian space where anything can happen [6]. In a space such as this, no colour, gender or ethnicity would be considered. Aspects of visual identities that lead to discrimination in the offline world would matter little in the fluid identities expressed in the democratic, disembodied world of cyberspace [7].

However, the reality does not depict such a utopian environment. A recent research work shows that people do not act independent of their offline identity when they interact online; their class, ethnicity as well as their parents' level of education matter [8]. For instance, minorities represent themselves differently from the majority, by having well-articulated profile about their presence in SNS [9]. They try to emphasize what their background and ethnicity are in an elaborated manner. The need of the hour is to make diverse people connect so that the differences could be reduced. According to the contact theory, diversity fosters interethnic tolerance and social solidarity.

In network analysis, Barabasi and Albert[10] have shown that, if nodes prefer to connect with high degree nodes, or in this case, people preferring to connect to highly connected profile of others (the one with high degree of links), in a so-called preferential attachment, the network will become scale free. Huge clusters/communities will be developed, leaving the least unpopular ones on their own as segregated. In short, it says that if you are not accepted to the main stream, there is a high chance that you would be left alone.

A recent study conducted by Ofcom [11] shows that people on Social Network Sites (SNS) predominantly bring their offline social network online, and use SNS to revitalize old relations. This is further endorsed by a study done by Mayer and Puller [12] on one of the SNS - Facebook, which shows that only a small percentage – 0.4 to be specific, of the friendships ties reflects “merely online interactions”. If this is true, do we carry the same notions/attitudes of our offline life onto the online one? The marginalization of minorities based on ethnicity, religion and social status is still being done?

According to [13], US schools are racially segregated. They carried out their research on 10 public and private universities of US, and came to this conclusion. This study, however, is limited and needs to be further investigated for a broader target.

Research Aims

- To gather and organize data related to social ties of online social networking systems (SNS) in order to identify individual friendship-choice behaviours and aggregate network patterns.
- To analyze the data to assess the extent to which SNS communities based on ethnicity, religion or race, are segregated.
- To discover some of the differences between the online ties with the real world ties (online vs. offline social network) in terms of friendship-making behaviour and network patterns.
- To build a series of evidence-driven agent-based models to capture the social processes involved in SNS network development.

Research Questions

- Does an SNS represent segregated network; and on what factor(s) does/do it segregate?
- To what extent are processes involved local – that is based only on individual's own preference?
- What characterises those with more diverse the online ties?

Methodology

The basic motivation of this PhD thesis comes from Schelling's model of residential segregation [14], which gave the concept of Agent-Based Modeling. He showed, if a person has a colour preference for her neighbours, that could lead to total residential segregation. For this thesis, similar methodology of ABM through social simulation will be used, which tries to bring formal and descriptive approaches of social sciences toward the study of processes/mechanisms and behaviours that constitute the society. A very detailed representation of this method was used by Joshua M. Epstein and Robert Axtell [15] who developed the first large scale agent model, called "Sugarscape", to simulate and explore the role of social phenomena such as seasonal migrations, pollution, sexual reproduction, combat, transmission of disease, and even culture.

Data collection itself will be done at a number of different levels. Three modes of information gathering will be used. Firstly I will collect data out of the publicly accessible information of people's profiles in an SNS, such as Facebook. Information like one's ethnic background, along with her friendship list will be fetched. Obviously, not the whole SNS could be explored. A subset of it which would be accessible and feasible too, shall be examined. Let us call it as "Target Network". To determine the social ties of the target network, a graph traversal method of Breadth First Search (BFS) will be used. This search is almost finished. Secondly, after careful analysis of this data, if there is any missing information, such as if friends' list is inaccessible due to privacy settings of any user, an online questionnaire will be asked by the users to fill the missing information - this is done in second mode. Lastly, in order to gain some fine grain information, either electronic (or where possible face-to-face) interviews will be conducted on a subset of the target network. Thus three complementary data sets will be collected, the smaller and richer samples allowing qualitative insight into the larger network data.

Based on the data of SNS, an evidence-driven agent-based model [16] will be developed. It is going to be an iterative stage, where the data collection and analyses will be done first, and then a detailed agent-based simulation model will be developed. But before actually designing the model, it will be required to identify different categories of agents and then filtering the required ones, in order to have a coarse understanding. Mechanisms will be built into the framework for testing the sensitivities of simulations to random factors and changes in model configuration and parameters.

The main factor to be investigated here would be to determine if the online representation of social ties do represent segregated network or not. If it does exist, what are the main reasons for such an emergence? Can this segmentation be minimized? And finally, how the network will look if people's behaviours don't change.

Time line

In this first six months, literature review will be done, along with the data collection by mode 1, i.e., acquisition of data through publicly available profiles and a trial model explored (to test the technique). This has been completed. In the next six months (end of first year), analysis of this data and key questions are going to be explored and the first phase of evidence-driven models will be developed. Key parameters and factors are going to be analyzed and feed into these set of models. Missing information by mode 2, online questionnaire, will be asked to get filled. Based on the outcomes of first set of models, and the missing information, fine course grained models will be established (next 6 months). After that, till the end of year 2, a detailed set of interviews, from a subset of the target network, will be carried out to grasp the micro level information from users - which couldn't be fetched earlier. A new set of model(s) will be developed, which would shed light on the social processes behind online segregation.

References

1. H. Murray. Explorations in Personality. Oxford University Press, USA, 2007.
2. L. Festinger. Laboratory Experiments: The Role of Group Belongingness. Experiments in Social Process, 1950.
3. Coleman, J. S. Social capital in the creation of human capital. American Journal of Sociology, 94 (Supplement), S95-S120. 1998.
4. D. Boyd, "Why youth (heart) social network sites: The role of networked publics in teenage social life," Identity. MacArthur Foundation [http://www.danah.org/papers/\[March 15, 2007\], 2007](http://www.danah.org/papers/[March 15, 2007], 2007).
5. <http://www.facebook.com/press/info.php?statistics>
6. Negroponte, N. Being digital. New York: Alfred Knopf. 1995.
7. Turkle, S. Life on the screen: Identity in the age of the Internet. New York: Simon & Schuster, 1995.
8. Hargittai, E. Whose Space? Differences Among Users and Non-Users of Social Network Sites Journal of Computer Mediated Communication, 13 (1), 2007.
9. Grasmuck, S. and Martin, J. and Zhao, S. Ethno-Racial Identity Displays on Facebook. Journal of Computer-Mediated Communication, 15, 1, 158--188, 2009.
10. Barabási, A. L. and Albert, R., Emergence of scaling in random networks, Science 286, 509-512, 1999.
11. Ofcom. Social networking: A quantitative and qualitative research report into attitudes, behaviours and use. London: Ofcom/Office of Communications, 2008.
12. Mayer, A., Puller, S.L.. The old boy (and girl) network: social network formation on university campuses. Journal of Public Economics 92, 329–347, 2008.
13. Mayer, A.. Online social networks in economics, Decision Support Systems, 47(3),169-184, 2009.
14. Schelling, Thomas C. Dynamic Models of Segregation. Journal of Mathematical Sociology 1:143-186, 1971.
15. Epstein J M & Axtell R L Growing Artificial Societies: Social Science from the Bottom Up. The MIT Press, 1996.
16. Geller, A and MOSS, S.. Growing Qawm: An Evidence-Driven Declarative Model of Afghan Power Structures. Advances in Complex Systems, 11(2), pp. 321-335, 2008.

Appendix F.3 Letter from the chair of the Ethics Committee

From: **Stephen Whittle** <S.T.Whittle@mmu.ac.uk>

Date: 9 January 2017 at 18:54

Subject: RE: Face book scraping

To: Bruce Edmonds <bruce@edmonds.name>

Dear Bruce

Re. PhD student is Ali Abbas Ethical Clearance; Retrospective Chairs Action.

1. The Research Data.

The Data Protection Act 1998 requires that research data must not be kept beyond the period for which it is required. This applies only to personal data, i.e. data that could be used to identify a living individual.

In this case, once the student has been examined, I don't believe that there will be any need to further retain this data, and consideration should be given to its destruction. As soon as the data no longer needs to be retained, the student needs to ensure all copies are securely disposed of. This may require the physical destruction of discs and drives. Guidance could be sought from Computer Sciences.

During the period of time in which the data still needs to be retained, the data must be kept in a secure storage system.

Bearing in mind the potential sensitivity of Facebook pages, linking the Facebook scrapings and the name data should only be possible with separate encryption keys, which are to be stored separately from both sets of data.

- a. The Facebook scrapings must be anonymised i.e. have the subject names removed and replaced by a random coding.
- b. The Facebook scrapings data should then be encrypted with an original encryption key.
- c. The name data and the linking code should then be encrypted with a separate original encryption key.
- d. The encryption keys required to connect the data should then be further encrypted with a new original encryption key.
- e. The encrypted sets of data and encryption keys must be kept within a high security system, requiring password access.
 - i. This may be online, but extra care must be taken in those circumstances to avoid any risk of hacking. Alternatively,
 - ii. this information may be kept on discs or drives, in which case the disc or drive containing the encryption keys must be kept in a separate space from the data e.g. by the student's supervisor, until such time as the data can be destroyed.

2. Protection of Research Subjects

Just to clarify; I gather that very few of these subjects were contacted, and those who were, it was then only to clarify their ethnicity.

At the time of the data collection, I understand that even these subjects were not aware this information was being scraped. I also gather that the information retained did not concern personal or sensitive matters that might cause distress or harm if discussed with the subject. Furthermore, no such discussions took place anyway.

If that is the case, there are no ethical issues in relation to potential harm to the subjects of the research in the data collection process.

I am not sure what the ethics are of Facebook scraping these days, but so long as the student abided with the rules of Facebook publication at the time, I can see no ethical concerns there.

3. Conclusion

The supervisor should note that mass accessing of personal self-published online information for research purposes is fraught with potential ethical issues. It would not now be possible to afford ethical clearance for most future research that uses an online trawl of personal data.

The only way in which ethical approval could be given would be if the research subjects are fully informed of the nature of the trawl; the type of information likely to be captured, and their right and ability to withdraw themselves from being a subject of the research.

However, in this case, so long as all of the above has been, or is, abided by, I can see no problem with providing retrospective ethical clearance being provided

Stephen Whittle, 09/01/2017

Acting Chair of the Faculty of Business and Law Research Ethics and Governance Committee

All the best

Stephen

Stephen Whittle

Professor of Equalities Law,

Acting Chair of the Faculty of Business and Law Research Ethics and Governance Committee

Telephone : +44 (0)161 442 4772 (Monday, Thursday and Friday)

Mobile: +44 (0)7809 621395 (Tuesday and Wednesday)

Direct email: s.t.whittle@mmu.ac.uk Office email: law@mmu.ac.uk