

EDGE HILL UNIVERSITY

Department of Computer Science



DETECTION OF MICROCOSMS ON TWITTER

A thesis submitted in partial fulfilment of the requirements for the degree of
Doctor of Philosophy

Inuwa-Dutse, Isa

Edge Hill University, UK

2019

DEDICATION

to:

my parents -- Haj. Umma and late Malam! for everything
...Raihan, Abdullah & Muhammad for bearing the long absence
...late Mal. Buba for caring

APPRECIATION

Bismillah, all praise is to Allah, Lord of the universe and the God of those who passed away and those to come. Prayers, peace and blessings be upon the elite of His creation, the last of His prophets and messengers, our master Muhammad *sallallahu alaihim-wasallam*, his pure family and companions. It is true; a journey of a thousand miles begins with a step. It seems like yesterday when this arduous journey begins, and now it is looming into the alley of history. To quote Derek V. Ager: "*The history of any one part of the Earth, like the life of a soldier, consists of long periods of boredom and short periods of terror*"; when I first read the quote in Bill Bryson's book on *A Short History of nearly Everything*, I thought it is implicitly referring to research students! The statements aptly apply to PhD candidates as well, noting that the experience is characterised with boredom and momentarily terrifying periods – rejected papers, struggling with deadlines, setbacks in experiments, harsh reviewers, balancing to remain on the same page with the supervisory team, spontaneous eventualities, etc. In any case, as the saying goes 'what doesn't kill you makes you stronger', *Alhamdulillah*, this day I'm writing the preface for my final PhD thesis after a long period mixed with happy and not so happy moments. Indeed, the journey didn't begin with the PhD days; it is like a continuum in which you require the help and support of others along the way. At various stages, *Alhamdulillah*, I was fortunate to be in the midst of supportive and remarkable individuals. I'm terrible with verbal dexterity in expressing my gratitude. Still, I'm profoundly stunned and always grateful for the support I received. The following individuals deserve a portion in the thesis.

To begin, my sincere gratitude to my parents, Haj. Umma & *late Malam*, may Allah the beneficent and merciful grant you His Jannatul-firdaus, ameen. To Malam, *Allah Ya gafarta maka Ya kuma sa Jannatul-Firdaus ce makomar maka, ameen*. Mind and body genuinely miss you, but we have to take your absence as a destiny, and someday we will be history – down the abyss. To Mal. Buba Moyole, I owe you a debt of gratitude, may the Almighty Allah grant you the best of abodes, ameen. To my supervisors, I was fortunate to have a Greek and a British in my team; I presumed you can easily discern what the experience would be. Yannis (Dr!), you enable the space to explore my ideas after thought-provoking and intellectual arguments during the numerous meetings (and I guessed that office whiteboard will attest to that). I still remember our first official meeting (in the Ground floor office) how you describe

the research process in the easiest to understand. With your mentoring approach, my skills in academic writing and presentation of ideas have been improved – those Ms drafts and overleaf pages will be a suitable alibi! Thank you. Dr Mark, always on alert to critique and ensure that grammatically correct expressions are adhered to. Over the years, you have helped toward improving my writing skill; thank you. Prof. Franco, thank you for the support during the study period. Prof. Nik, my initial thanks for the swift reply to my first email inquiring about the eligibility for international students to apply for the then advertised GTA position. Your leadership and willingness to support us during our stay is well appreciated. Prof. Ella, thank you for the kind and humble support. To all my colleagues and students, thank you for making the voyage exciting and enjoyable. To Edge Hill University Campus and Ormskirk, thank you. To you, the less talking community; I mean our research office, please try to make it livelier. Sometimes one wonders whether there exist any remaining mortals in the room – all engross in seemingly unending tasks, give it a break and say all is well, thank you.

My sincere gratitude to my brothers: Ahmed Isah-Dutse, thank you for the support and caring, may Allah's blessing be with you and your family, ameen. My appreciation and gratitude to Ya Tijjani, Ya Abdullahi, Aliyu and Abubakar, may Allah reward you, ameen. I am equally indebted to Ya Indo, thank you and may Allah reward you abundantly, ameen. My thanks to the old grandpa (Baffa) and Kawu Magaji. To Gwaggo, Hajiya and Dada I equally thank you and may Allah reward you. To the departed souls, may Allah grant you His Jannah, ameen. I'm part of an extended family; to mention all my family members that contributed in one way or the other will consume the whole pages of this document! Nevertheless, my thanks to Ya Baffa, Adamu, Abba, Baba Alhaji, Yanun, Aminu, Balarabe, Farouk, Usman, and Abdulkadir. To my sisters, Ya Umami, Ya Hadiza, Ya Saude, Ya Maryam, Ya Sadiya, Hafsa and Bilkisu, I appreciate your concern and support. To Jamila for the dambun nama! To my younger ones, I must thank you too for your contributions. To my friends, you are all cherished, and I still remember the beautiful moment we shared. Finally, to my dear Raihan, Abdullah and Muhammad (Malam Jnr.), thank you for bearing the long absence. To crown it all, I once again thank *Allah Subhanahu Wata'ala* for the successful completion of one of the many phases, Alhamdulillah.

Isa Inuwa-Dutse

03-09-2019

Detection of Microcosms in Social Networks

Abstract

A network is a composition of many sub-networks or communities with distinct and overlapping properties. Because similarity breeds attraction and interaction, a community constitutes of sets of *nodes* and *edges* with a stronger relationship that is expressed as a function of *relatedness*. Network communities provide a crucial organising principle, which enables a better understanding of the structure and function of complex networks. Depending on the network type, communities come in various forms – from biologically- to technologically-induced communities. Of *technologically-induced communities*, social networks or social media platforms such as *Twitter* and *Facebook* support a myriad of diverse users to remain connected, leading to a highly connected and dynamic *social media ecosystem*. Within this complex ecosystem, multiple types of communications happen at various layers of granularity and intensity, leading to the formation of communities. The task of identifying embedded communities within a network has been of great interests for various reasons because a community is a functional unit of a network that captures local relationship among the network objects. Community detection paradigm involves prediction and quantification processes to identify and explain community structures in a network. Establishing the equivalence of network entities is achieved either based on (1) the equivalent units with the same connection pattern to the same neighbours and (2) the equivalent units have the same or similar connection pattern to different neighbours. Accordingly, communities are further formed around two primary modalities or sources of information: *network structure* and *features or attributes of nodes*. However, existing studies mostly focus on one aspect and the few studies based on a bi-modal source are limited in the use of a shallow set of features. In the context of Twitter, while many community detection algorithms have been proposed in the past, detection of socially cohesive communities still poses some challenges with respect to mining-related tasks. These challenges are due to (1) flexibility of interaction in *social media*, leading to a vast amount of content – relevant and irrelevant (2) a form of *logical*

social dichotomy that favours content from popular users to dominate (3) the ability to automate *users' accounts* and remain anonymous (4) the eccentricity of connection on Twitter contributes to identifying many socially unrelated users and encourage the propagation of spurious content.

Noting the challenges mentioned above, the thesis presents an effective detection method. The central themes in the research relate to the problems of *identifying genuine content* and *detection of socially cohesive groups*. The problem of *identifying genuine content* is tackled using a novel approach (*SPD strategy*) designed to filter out irrelevant content, while the problem of community detection is formulated to focus on smaller groups, which are homogeneous to many sociodemographic behavioural, and intrapersonal characteristics. Essentially, the research proposed a *multilevel clustering technique (MCT)* that leverages both *structural* and *textual* aspects to identify local communities termed *microcosms*. By recognising the harmful effect of social media spam and fake content towards undermining credible research based on analysing social media data, the thesis contributed a useful content filtering system. As a precautionary measure to avoid compromising the research outcome by irrelevant or unrepresentative data, the *SPD strategy* offers crucial insights into the sophisticatedly evolving techniques of spamming on Twitter. As a result, the detection of socially cohesive communities will be enhanced, thus providing a useful analysis tool and strengthening the validity of online content. The proposed *MCT* provides a useful, scalable framework to identify sub-groups in a network. The experimental results from the *MCT* and evaluation on benchmark models and datasets demonstrate the efficacy of the approach. Through this research work, a new dimension for the detection of cohesive communities on Twitter is contributed. The thesis contributes to the literature by offering better understanding and clarity toward describing how low-level communities of users evolve and behave on Twitter. Moreover, by identifying communities of users with strong cohesion, a well-informed recommendation that recognises *structural* and *content* similarities can be achieved.

Isa Inuwa-Dutse: _____

SUPERVISORY TEAM

Dr Yannis Korkontzelos

Dr Mark Liptrott

Prof. Francesco Rizzuto

VIVA TEAM

External Examiner: Dr Zhiwei Lin, Ulster University, UK

Internal Examiner: Prof. Ella Pereira, Edge Hill University, UK

Chair: Dr Katja Eckl, Edge Hill University, UK

CONTENT

ABSTRACT	1
PART I: INTRODUCTION	11
I INTRODUCTION	11
1.1 Introduction	11
1.2 Motivation	16
1.3 Aim and Objectives	19
1.4 Contributions	21
1.5 Research Process	23
1.5.1 Methodology	24
1.6 Thesis Structure and Summary	26
PART II: BACKGROUND	29
II SOCIAL MEDIA ECOSYSTEM	30
2.1 Introduction	30
2.2 Social Media Ecosystem	30
2.2.1 Social Media Platforms	31
2.2.2 Twitter	32
2.3 Network Communities and Sociometry	37
2.3.1 Online Interaction and Local Community	37
2.3.2 Sociometry	39
2.4 Summary	42
III COMMUNITY MEMBERSHIP MODELS	43
3.1 Introduction	43
3.1.1 Relevance of Community Detection	43

3.2	Network Models and Community Structure	44
3.2.1	Network Models	45
3.2.2	Community Structure	47
3.3	Detection Task	48
3.3.1	Detection Approach	49
3.4	Clustering and Community Detection	52
3.4.1	Clustering-related Tasks	53
3.5	Summary	57
PART III: PROBLEM FORMULATION		57
IV MICROCOSM DETECTION PROBLEM		58
4.1	Introduction	58
4.2	Community Structure in a Network	58
4.2.1	Role of a scoring function	59
4.3	The MCT Framework	61
4.3.1	Structurally-Related Nodes	62
4.3.2	Textually-Related Nodes	65
4.4	Definition and Notation	66
4.4.1	Definition	66
4.4.2	Notation	68
4.5	Summary	69
V AUTHENTICATION OF ONLINE CONTENT		71
5.1	Introduction	71
5.2	Online spam and detection methods	71
5.2.1	Spam detection methods	72
5.3	Spam-Posts Detection method	73
5.3.1	Spam detection data	73
5.3.2	SPD data annotation	74
5.3.3	Validation of SPD _{automated}	75
5.3.4	Feature extraction	78
5.4	Spam Prediction and Evaluation	80

5.4.1	Parameter tuning and prediction models	80
5.4.2	Feature importance and correlation	82
5.4.3	Performance metrics	83
5.4.4	Experimental results	84
5.4.5	Error analysis	87
5.5	Spam accounts and their features	88
5.5.1	Characterising users	88
5.6	Summary	91
	PART IV: MICROCOSMS DETECTION	91
VI	MICROCOSM: A META ANALYSIS	92
6.1	Introduction	92
6.2	Identifying Structurally-Related Nodes	92
6.2.1	Reciprocal Units	93
6.2.2	Identifying Dyads	93
6.2.3	Identifying Simmelian Ties	98
6.3	Summary	107
VII	DETECTION OF MICROCOSMS	108
7.1	Introduction	108
7.2	Structurally-Related Clusters	108
7.2.1	Network and Reciprocal Communities	109
7.2.2	Nodes Reciprocity	109
7.2.3	Spectral Clustering	113
7.2.4	Modelling Structural Communities	115
7.3	Textually-Related Clusters	118
7.3.1	Identifying Similar Content	118
7.3.2	Modelling Textual Communities	120
7.4	Microcosms Detection Algorithm	121
7.4.1	Objective Function and Optimisation	122
7.5	Summary	123
	PART IV: RESULT AND CONCLUSION	123

VIII MCT AND BASELINES	125
8.1 Introduction	125
8.2 Experimentation Setup	125
8.2.1 Datasets	125
8.3 Evaluation	127
8.3.1 Evaluation metrics	127
8.3.2 Baseline Models	129
8.3.3 Structural and textual aspects	130
8.4 Summary	133
IX CONCLUSION	134
9.1 Introduction	134
9.2 Key Findings	134
9.2.1 Content veracity	135
9.2.2 Community Detection	136
9.3 Reflection and Future Research	138
9.3.1 Spam Detection	138
9.3.2 Analyses of aspects of sociometry and clustering	139
9.3.3 Future Work	139
BIBLIOGRAPHY	157
APPENDIX	158
A SUPPLEMENTARY INFORMATION	159
1.1 Content Authentication	159
1.1.1 Tweet Object	159
1.1.2 SPD Features	160
1.2 Search Optimisation	161
1.3 Dyadic and Transitive Datasets	162
1.4 The MCT Framework	163
1.4.1 Microcosm detection dataset	163
1.4.2 MCT optimisation	164
1.4.3 Matrix decomposition	166

List of Tables

4.1	Summary of notations and descriptions	70
5.1	Summary of SPD datasets	74
5.2	Examples of collocational bigrams	76
5.3	Percentage of relevant metrics	76
5.4	Evaluation of results	88
5.5	An Example of sample tokens from misclassified tweets.	88
6.1	Dyads data statistics	94
6.2	Simmelian dataset summary	100
8.1	Microcosm detection datasets summary	126
8.2	Ground-truth data statistics	127
8.3	Clustering performance	132
A.1	Proposed SPD features	161
A.2	Search optimisation data	162
A.3	Facebook data	163

List of Figures

1.1	An illustration of multilevel clustering	16
1.2	Topological structure on Twitter	17
1.3	Iterative tasks	24
1.4	Schematic overview of the research process	25
1.5	A high-level overview of the thesis Structure.	28
2.1	Example of users on Twitter	32
2.2	An overview of Twitter	34
3.1	The notion of network communities	50
4.1	Classification of social groups	59
4.2	Cluster bands	60
4.3	MCT execution pipeline	62
4.4	Triads in a network	63
4.5	Relationship between pairs	67
5.1	SPD collection and validation	75
5.2	Activity pattern of bot account	81
5.3	Activity pattern of normal accounts	82
5.4	Relative performance of features	83
5.5	Univariate analysis	84
5.6	Performance of classification models	85
5.7	Learning curves1	86
5.8	Learning curves2	87
5.9	User types in the SPD_{manual} dataset.	90

5.10	Distribution of users in SPD data	90
6.1	Dyads proportion	95
6.2	Dyads boxplot	95
6.3	FCN pipeline	97
6.4	Dyads prediction results	98
6.5	ECDF of reciprocal ties	100
6.6	Reciprocity effect of features	101
6.7	Bayesian workflow	103
6.8	Bayesian sampling results	104
6.9	Posterior distributions	104
6.10	Existential reciprocal ties	106
8.1	<i>F-sim</i> prediction accuracy	131
8.2	Line and ECDF plot of probable tie formation	132
A.1	An overview of <i>content authentication</i> activity	160
A.2	Example of nodes degree	164
A.3	Geometric interpretation of function optimisation	165

Chapter I

INTRODUCTION

1.1 Introduction

Networks in nature are characterised by varying degrees of organisations (Lancichinetti et al. 2009), and past studies have analysed how to uncover the structure embedded within complex networks (Erdős & Rényi 1960, Scott 1988, Watts & Strogatz 1998, Albert & Barabási 2002). Complex networks were once considered to be random and the classical *random graph model* (Erdős & Rényi 1960) used to be the de facto analysis tool until the discovery of a regular pattern in various complex networks (Albert & Barabási 2002). However, with the computerisation of data acquisition methods and the development of data manipulation tools, most of the underlying design principles in networks across different domains have been well investigated (Albert & Barabási 2002). Classical network models such as the *small-world* (Watts & Strogatz 1998) and *scale-free* (Albert & Barabási 2002) are widely used and form the basis on which networks and communities are studied. As the size of a network increases, the possibility of fragmentation increases (Berelson & Steiner 1964, Shaw 1971), leading to a decrease in the homogeneity of behaviour and attitude across groups (Granovetter 1992). Because similarity breeds attraction and interaction, communities of similar members are formed, which result in the formation of strong ties among members of communities (Brass et al. 1998). A complex network is considered as a composition of many sub-networks or communities (Newman & Girvan 2004) and one of the vital tasks is to identify the community structure or network communities. Fundamentally, communities consist of a set of network objects (nodes) and a corresponding set of connections (edges) interacting with one another. The interaction is stronger within the communities and less across other communities in the network (Newman 2006). Detection of a

community structure is of interest to a myriad of researchers across various domains for many reasons:

1. it provides a means of analysing complex networks (Williams & Martinez 2000)
2. it allows for the detection of a collection of related web-pages (Flake et al. 2002, Papadopoulos et al. 2012)
3. it facilitates intelligent recommendation in social networks (Cao et al. 2015)
4. it enables detection of cliques in social networks (Newman & Park 2003)
5. it serves as a fundamental tool to discover the organisational principles in networks (Newman 2003, Yang et al. 2013), and
6. it helps in studying social behaviour of users (Arnaboldi, Guazzini & Passarella 2013).

The utility of a community structure in enabling effective analysis of complex networks makes it ideal to explore the network by identifying set of nodes and corresponding relationships. Depending on the network type, communities come in various forms: *protein–protein interaction network* (Krogan et al. 2006), *social networks* (Newman & Park 2003), *food webs* (Williams & Martinez 2000), *collaboration networks*, (Nascimento et al. 2003), *World Wide Web* (Albert & Barabási 2002).

In all the examples mentioned above, a community is a functional unit of the network that captures local relationship among the network objects, and the problem of community detection is to identify relevant partitions in the network. The underlying difference across many network communities is how the connections are defined – some connections are deterministic while some are uncertain but based on probability (Zhang & Zaïane 2018). To this end, many algorithms have been proposed in the past to detect community structures across various networks. Zhang & Zaïane (2018) highlights a simple categorisation of the algorithms based on the deterministic framework as follows: *graph-based algorithms* (Newman 2013, Kernighan & Lin 1970), *clustering-based algorithms* (Girvan & Newman 2002, Newman 2004b, Blondel et al. 2008), *genetic-based algorithms* (Pizzuti 2008), and *label-based propagation algorithm* (Raghavan et al. 2007). For a community structure in a network with edge uncertainty, approaches include the conversion of the *uncertain* network to a *certain* network by thresholding

the probability values to a binary form (Liu et al. 2012, Martin et al. 2016, Kollios et al. 2011, Zhang & Zaïane 2018).

Within the social media platforms, taking Twitter¹ as a case study, social interactions are continually evolving to support a myriad of objects to remain connected, leading to a highly connected and dynamic social media ecosystem. Within this complex ecosystem, multiple types of communications happen at different layers of granularity and intensity ranging; from *global* or *local*, *positive* or *negative*, *influential* or *less influential*, *low-level* or *high-level*, leading to the formation of communities at various levels in the network. A community detection paradigm involves prediction and quantification activities to identify a community structure and relevant details about the observed structure (Chen et al. 2009). Predicting membership and assigning items to a community or cluster is achieved using a measure of equivalence or a scoring function. A scoring function also enables *block-modelling*, based on the idea that units in a network can be grouped according to the extent to which they are equivalent using a set of experimental procedures. Depending on the procedural approach, the definition of equivalence usually leads to different partitioning of a network.

Moreover, communities are formed around two primary modalities or sources of information: *the network structure* and the *features and attributes of nodes*. However, studies mostly focus on one aspect, not both. The early work in this line of research can be found in (Balasubramanyan & Cohen 2011, Leskovec & McAuley 2012, Yang et al. 2013). A closely related approach to the thesis's method can be found in Yang et al. (2013), in which both modalities have been considered. However, the depth of the features, especially the nodes attributes, is shallow. Moreover, concerning Twitter, the structural component is not fully captured because it relies on a directed form of connections. In the context of Twitter, communities could be formed based on many factors (see Figure 1.2 for illustration) and the research is interested in revealing factors that will ensure the detection of a local community in a network.

Multilevel Clustering Technique Noting the eccentric connections pattern in Figure 1.2, which could lead to the detection of socially unrelated users and encourage the propagation of spurious content, this thesis proposed a *multi-level clustering technique (MCT)* to identify socially cohesive groups of users or local community structures on Twitter termed *microcosms*.

¹<https://twitter.com>

There are no practical reasons that will prevent the thesis’s approach to apply to other domains involving network data; however, the contrast could be true if the method is in other platforms where a reciprocal tie is the default connection between users. A directed tie is peculiar to Twitter since, in other platforms such as Facebook², an automatic reciprocal relationship is established once a friend request is accepted. Identifying the set of fully connected nodes on Twitter is challenging due to the flexible and eccentric underlying connection patterns, which enables flexible followership that results in many unidirectional links. The premise in the *MCT* is that a *community detection* or *clustering* method that recognises a set of reciprocal ties on Twitter offers a more cohesive and better representation of a community. *Dyadic and transitive ties*, primary forms of establishing a reciprocal tie, will play a central role in identifying socially cohesive groups. In social networks, these two forms of relationships are viewed from various perspectives and often with contradicting results. Previous studies (Weng et al. 2010, Kwak et al. 2010, Cha et al. 2012, Arnaboldi, Conti, Passarella & Pezzoni 2013) have examined the manifestation of dyadic ties or reciprocity for various tasks which are either based on *directed sets of nodes* or *textual content*. Transitivity defines a social preference to be friends with a *friend-of-a-friend* and has been recognised as a peculiar feature of a network (Watts & Strogatz 1998). The transitive tie is closely related to a *Simmelian tie*, a strong social relationship between three or more individuals. For network analysis, transitivity is a vital feature of a network that enables the formation of cohesive communities and enables an understanding of the structure of social ties in a network (Granovetter 1977). However, as illustrated in Figure 1.2, the prevalence of transitory connections makes it challenging to identify real transitive ties on Twitter.

Failing to recognise the particularity of connectivity on Twitter as exemplified in Figure 1.2, the approach cannot be generalised; hence, the use of Twitter as a case study (see Chapter II). The proposed *MCT framework* is based on a joint modelling of structural and intrinsic textual features (see Figure 1.1), which contributes toward a methodological paradigm for the detection of *microcosms* in a dynamic and heterogeneous social media. The *MCT framework* targets a similarity within a community of users using global and local information to measure similarity or equivalence among nodes. There are two basic approaches to the equivalence of units or nodes in a given network: (1) the equivalent units have the same connection pattern to the

²<https://www.facebook.com/>

same neighbours and (2) the equivalent units have the same or similar connection pattern to different neighbours (Doreian et al. 2005). In the context of this study, these two equivalences are related to the structural similarity of users on Twitter, which has been used to define a useful scoring function detailed in Chapter VII. Consider Figure 1.1 in which communities of users exist based on *structural* (Figure 1.1(a)) and *content* or *textual* (Figure 1.1(b)) similarities. The groups under the structural component are related based on reciprocal ties (which are rare on Twitter) and the community is more cohesive than its counterpart, which is based on *textual similarity*. A more cohesive community is the one that recognises both *structural* and *textual* similarities (Figure 1.1(c)).

The modelling of a *structural similarity* (Figure 1.1(a)) exploits the idea of *homophily*, a social science principle based on the adage *birds of a feather flock together*. A true reflection of homophily on Twitter will rely on the reciprocal relationship among users and a user with many reciprocated ties can be a resourceful representative in the quest of detecting microcosms, making it possible to analyse a group of users as a unit. For instance, many crucial aspects, such as validation or characterisation of content integrity, can be explored. It follows that a user who spreads rumours or spam content is likely to be strongly connected with similar users, hence the inclusion of structural component in the *MCT framework* adds a layer of social cohesion. For the structural component, the two basic forms establishing a reciprocal tie: *dyads* and *Simmelian tie* (see Chapter VI) are utilised. The collection of users with structural similarities are analysed using *spectral clustering*. Spectral clustering involves a series of operations ranging from the construction of adjacency or affinity matrices to grouping in a reduced dimension (Han et al. 2011). The *MCT strategy* also incorporates *textual aspect* that complements the *structural aspect* to add a layer of social cohesion in the detection task. Identifying *textually-related clusters* is a form of *document-pivot clustering* in which weights are assigned to features in the document according to a weighting scheme (Allan et al. 1998, Yang 2001, Brants et al. 2003, Fung et al. 2005). Through the *MCT strategy*, the problem of structurally unrelated users is addressed, thereby adding a layer of social cohesion to community detection approaches proposed in the past. Specifically, the proposed *MCT* advances existing techniques in the related literature (see Chapter III) through (1) *an in-depth utilisation of bi-modal sources of information for community detection* (2) *detection of network communities at various levels* (3) *a robust and*

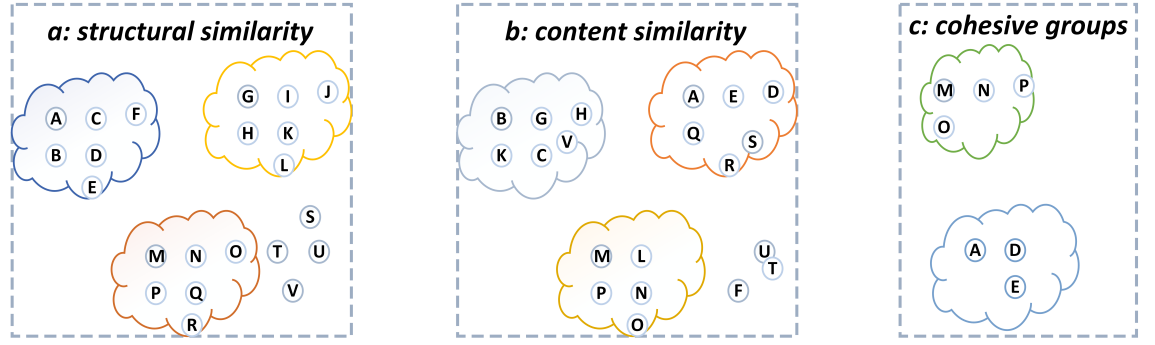


Figure 1.1: An illustration of how nodes are clustered in the multilevel approach: (a) The first stage involves the grouping of nodes according to their degrees of structural similarities. (b) The second stage is concerned with identifying collections of nodes with high degrees of similarities in content or *textual* aspect. (c) Finally, nodes with high degrees of similarities in both *structural* (a) and *content* (b) aspects are grouped accordingly; such groups constitute cohesive communities or cliques.

scalable community detection algorithm that is less affected by noise in the network data, and (4) an intuitive interpretation of the detected communities.

One of the critical dimensions in the research was motivated by the need to get rid of spurious content in the research data. Because *text corpus* is involved, there exist preprocessing requirements that need to be satisfied. The set of texts comes from Twitter, which makes it possible for participants to freely generate and consume information leading to unprecedented amounts of data. While this data is being exploited for various applications, a substantial amount of the data is being generated by spam or fake users. Without a proper data filtering mechanism, the growing amount of irrelevant *social media content* undermines the credibility of research based on analysing such data. The thesis's motivation to identify and filter out spam content in social media data culminated in the development of a novel spam detection technique, see details in Chapter V. In addition to deploying the proposed strategy in a social media data collection pipeline as an initial preprocessing task to detect and filter unwanted *tweets*, short texts posted by users on Twitter, the research complements earlier approaches (Varol et al. 2017, Lee et al. 2011, Alsaleh et al. 2015, Davis et al. 2016).

1.2 Motivation

Through clustering, a compelling summary of relationships among network objects can be found, and the detection of communities provides an effective means to understand the under-

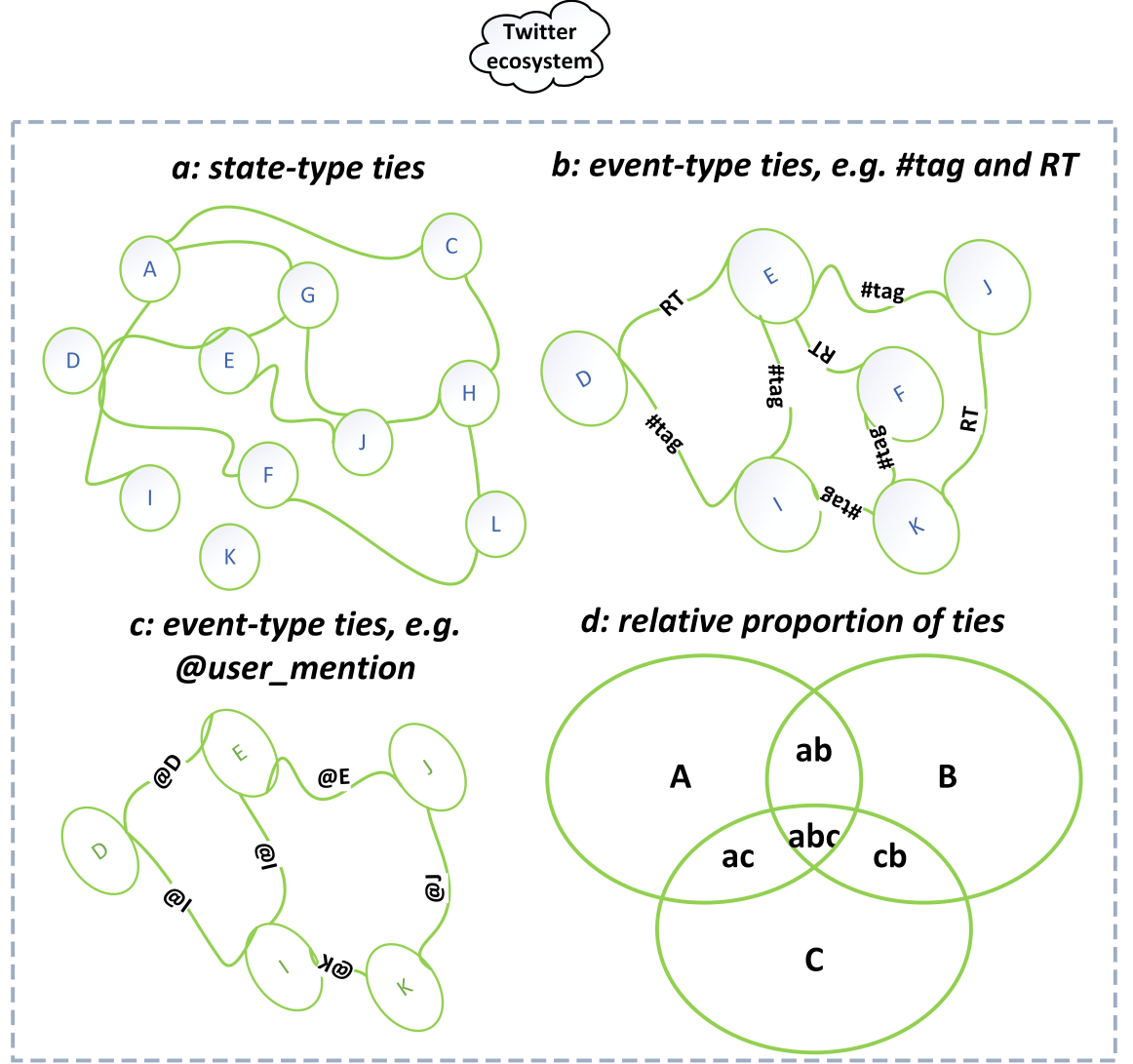


Figure 1.2: Examples of *event-type ties* ((a),(b) and (c)) on Twitter, making it possible for users to openly connect in many ways, which could be via: (a) unidirectional or directed means (e.g. a friend or a follower), bidirectional or undirected (both friend and follower) (b – c) indirect or *transitory events* such as *retweets*, *mentions* or *likes*. These flexible connections challenge cohesive community detection task and also contribute to the proliferation of spurious content on Twitter. In (d), $A = \{a, a_1, \dots, a_l\}$, $B = \{b, b_1, \dots, b_m\}$, and $C = \{c, c_1, \dots, c_n\}$ denote users and their sets of networks. It is rare to find a large scale data consisting of reciprocal ties (e.g. ac) or transitive ties (e.g. abc) on Twitter.

lying structure in a network (Watts & Strogatz 1998, Albert & Barabási 2002, Newman 2003, Doreian et al. 2005, Lancichinetti et al. 2009) and extract useful information (Zhang & Zaïane 2018). Recent advances in community detection have offered a useful framework to explore the structure of communities embedded in various kind of networks (Girvan & Newman 2002, Newman 2004b, Raghavan et al. 2007, Blondel et al. 2008, Pizzuti 2008, Newman 2013, Yang et al. 2013, Zhang & Zaïane 2018). While many community detection algorithms have been proposed in the literature (see Chapter III), detection of socially cohesive communities on Twitter is still a challenge leading to the discovery of disparate communities that are likely to be socially unrelated. Noting that social networks exhibit different properties from other networks (Newman & Park 2003), the thesis argues that the limitation of existing approaches is due to the following:

1. *methodological viewpoints*: social network theorists hold two methodological positions in investigating social relationships: *realist* and *nominalist*. While the *realist* proceeds with a preconceived notion of the existence of relationships in a network which need to be discovered, the *nominalist's* approach is based on the questions asked by the investigator (Laumann et al. 1989). Existing studies mostly adopt the *realist's* approach.
2. *connections and formation of social ties*: social ties can be based on *event-type ties* or *state-type ties*. An *event-type tie* is transitory and often results in socially distant members. With respect to Twitter, an *event-type tie* consists of subscription to *trending hashtags* or *retweeting a popular user*; see Figure 1.2. On the other hand, a *state-type tie* is based on static or structural connectivity between users, which suggests a certain degree of familiarity and trust (Borgatti & Halgin 2011). The problem of community detection on Twitter is mostly centred around a directed form of connections (*event-type ties*) based on the *realist's* approach. While this is valid in many networks, it could lead to many disparate or unrelated sets of users on Twitter. For instance, Figure 1.2 shows the various forms of establishing connections on Twitter which led the research to argue that the implication of such connections contributes to widespread spurious content and a less cohesive community of users.
3. *rapid increase in volume and complexity of online content*: large scale and transitory content (largely from influential users, see Figure 2.1) mostly dominate the space leading

to many forms of explicit communities (Kwak et al. 2010). Basing a community detection task on transitory aspects of *metadata* such as popular hashtags or trending topics does not often reflect true connectivity (Wilson et al. 2009), hence limiting the full realisation of the benefits in communities such as *cliquishness* and *local coherence*. As the size of a network increases, the possibility of fragmentation (Berelson & Steiner 1964, Shaw 1971), leading to a decrease in the homogeneity of behaviour and attitude across groups (Granovetter 1992). With an average 100m daily users contributing to 500m content³, it is becoming more challenging to keep track of socially cohesive communities on Twitter.

The aforementioned challenges affecting the detection of local communities on Twitter motivate the research; thus, addressing them will ultimately advance our knowledge concerning community detection and related problems. The thesis addresses all the identified challenges, which are further reformulated in the forthcoming section (1.3).

1.3 Aim and Objectives

While noting the breadth and depth of the issues raised in the previous sections (1.1 and 1.2), the aim of the research is to develop a useful framework for the detection of *microcosms* on Twitter. To achieve the primary goal of identifying microcosms, the following *Research Questions (RQ)* and the corresponding *research objectives* are presented as the central targets. The research questions are addressed in a sequential logical order such that the previous question informs the next one.

1. *RQ1: what algorithms exist for clustering in a dynamic Twitter environment?*

It is crucial to identify relevant clustering algorithms, which enable detection of communities on Twitter. This is required to pre-empt repeating a problem previously solved and better inform the thesis's methodological standpoint by actively engaging with the relevant literature. The following objective would answer the posed question and facilitates the achievement of the research aim: *to conduct a thorough survey of the existing literature on community detection algorithms and related methods*. Eventually, the strengths and weaknesses of the appropriate methods or algorithms relative to the research problem will be described.

³See <https://www.omnicoreagency.com/twitter-statistics/>

2. *RQ2: is data from Twitter credible and suitable for investigating community detection problems?*

Data from Twitter or *tweets* are characterised as messy and full of unnecessary and diverse jargons. Moreover, a substantial part of a collection of tweets⁴ is being generated by spam or fake users which will endanger credible outcomes if the data is utilised in its original form. The rapid growth in the volume of global spam is expected to compromise research works that use social media data, thereby questioning data credibility. Thus, researchers working with tweets expend a considerable amount of time to clean and avoid any form of pollutants in the data. Motivated by the need to identify and filter out spam content in the research data, the objective is *to investigate the credibility and suitability of tweets in addressing the research problems*. Furthermore, as the first step in clustering, it is crucial to assess the tendency of the research data for clustering (Jain et al. 1988, Han et al. 2011); hence, the need to determine whether the data is suitable and could be clustered.

3. *RQ3: how to efficiently search for relevant items on Twitter?*

Among the peculiar features of online social media platforms is the flexible ability for users to act as both producers and consumers of content. While this is empowering, it is also posing many challenges in terms of how we access relevant information efficiently. Searching for information on Twitter is particularly challenging due to both the inconsistency in writing styles and the high generation rate of spurious and duplicate content. With the growing data stream and increasing demand for instant processing where efficiency is crucial, the objective in this respect is *to improve search efficiency in a highly dynamic and stochastic environment like Twitter*. The ability to develop an efficient search strategy will ultimately help toward addressing the clustering problem.

4. *RQ4: how to detect a community of users with a strong social cohesion on Twitter?*

In sociometry, a taxonomy of social relationships is described as a function of closeness among users (Dunbar 1998), and the closer the users are, the more cohesive and trustworthy. However, in social platforms such as Twitter, where *event-type ties* are prevalent, it is difficult to identify a local community or a community of users with strong social cohe-

⁴One in every 200 social media messages and one in every 21 tweets is estimated to be spam

sion. To enable the detection of *microcosms* on Twitter, the following research objectives would facilitate the achievement of the goal:

- *to study the formation of ties on Twitter*
- *to conduct an empirical analyses of event-type ties and state-type ties on Twitter*
- *to identify social factors that enable formation of groups*
- *to formulate a useful method of predicting reciprocal relationships on Twitter*

Achieving the objectives outlined above will provide answers to the posed questions, thus satisfying the requirements needed to detect cohesive communities of users on Twitter and related application domains. The next section (1.4) presents a summary of the thesis's contributions.

1.4 Contributions

Through this research work, a new dimension of detecting cohesive communities on Twitter is contributed. The approach can be applied to study the formation of various communities at different levels of granularity. Furthermore, the research contributed to the literature by offering better understanding and clarity toward describing how low-level communities of users evolve and behave on Twitter. The following points succinctly describe the thesis's contributions.

1. **Contribution I:** *a systematic exposition of clustering and community detection algorithms*

Despite the popularity and rich research works on Twitter, there is a lack of a comprehensive body of work analysing microscopic communities with strong social cohesion on Twitter. This limitation is attributed to the dominance of *influential users*, making their content more visible on Twitter which led to many users engaging with such content via *hashtags*, *trending topics* or other *metadata* at high levels. In line with this, the study contributed a detailed analysis of relevant clustering algorithms with an exposition on their strength and weakness, suitability on clustering tweets and the associated challenges. The relevance of this contribution is not limited to the literature only, but as an initial guiding step in the research.

2. **Contribution II:** *an effective data cleaning strategy*

The ease at which tweets are produced on Twitter is also a source of the challenges concerning data quality. Tweets are generally characterised as messy, and research activity involving tweets is expected to ensure clean and avoid any form of pollution in the data. It is of utmost importance to ensure the suitability and credibility of the data to use in solving the research problem because, without effective filtering, misleading results will be inevitable. While many filtering strategies have been proposed, the growing spamming activities and sophistication of automated accounts on Twitter require an up-to-date tool to combat evolving challenges. Motivated by the need to identify and filter out spam contents in social media data, the thesis contributes a novel approach (*SPD strategy, Chapter V*) for distinguishing *spam* vs. *non-spam* social media posts. The approach in this respect offers more insight into the behaviour of spam users on Twitter. The effectiveness of the *SPD strategy* has been evaluated against numerous datasets and related baseline in the literature. The filtering mechanism can be easily deployed during data collection as a first preprocessing strategy to improve the validity of research data. All the data collected for the research have been cleaned using this strategy.

3. **Contribution III:** *an efficient search method for similar items on Twitter*

The increasing high generation rate of online content, which makes searching for relevant items difficult, is causing drawbacks to tasks such as the clustering of items which relies on a measure of relatedness. Many search approaches have been proposed in the literature; however, searching for information on Twitter is particularly challenging due to both the inconsistency in writing styles and the high proportions of spurious and duplicate content. The research contributed a useful method that enhances searching for relevant items, which will ultimately improve clustering tasks. The approach is based on state-of-the-art deep learning methods and a novel *Scalable Windowing Approach for Pairwise-similarity Search (SWAPS)* to improve search efficiency. Essentially, SWAPS optimises searches using a strategic balancing criterion to assess the trade-off between accuracy and search speed, thereby circumnavigating sequential search problems, i.e. time consumption.

The theoretical significance of SWAPS has been analysed in terms of computational complexity with a promising performance. Although using SWAPS and deep search strategy, there is no optimal way to return all related items since information about the whole net-

work structure is required, the effectiveness of the approach is adequate in this respect. This contribution also entails the use of a *deep learning strategy* to profile tweets based on their unique signature, which establishes a relationship between the status of a tweet and its longevity measured in terms of engagement lifespan since the time of initial posting. Using various benchmark datasets, the efficacy of the proposed approach was assessed using a searching criterion that ensures a high degree of true positives.

4. **Contribution IV:** *a new dimension of detecting cohesive communities of users*

The ability to follow anyone on Twitter results in many unidirectional connections between socially unrelated users, which affects effective clustering and the integrity of online content. In the proposed approach, the detection of a socially cohesive group of users starts with the identification of reciprocal ties based on the premise that a clustering method that recognises *reciprocal ties* offers a cohesive and better representation of a community. This is inspired by the idea of *homophily* and *cognitive balance theory*. However, this is a challenging task when working with a large scale dataset due to the flexible and eccentric underlying connection patterns leading to many socially unrelated users group as communities on Twitter. To counter the challenging and time-consuming task of collecting reciprocal ties on Twitter, the research proposed a prediction model that returns the likelihood of two users engaging in a pairwise relationship. As a result, the detection of socially cohesive communities is enhanced, thus providing a useful analysis tool and strengthening the validity of online content. From an application point of view, by identifying communities of users with strong cohesion, a well-informed recommendation that recognises *structural* and *textual* similarities can be achieved.

1.5 Research Process

The primary goal is to develop a method that detects *microcosm* on Twitter and the following section presents an overview of the research process. Noting the breadth and depth of the issues raised in the chapter, especially Section 1.2 and Section 1.3, a systematic research process inspired by the *gradualist's philosophy* is followed to meet the following targets:

- i satisfy the requirements needed in the domain of application, i.e. social networks with em-

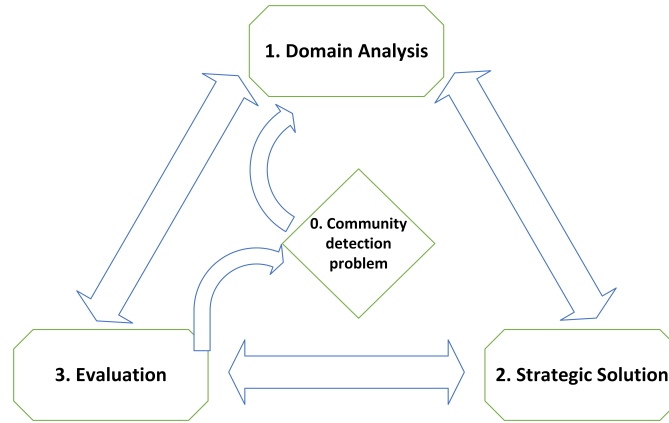


Figure 1.3: An overview of the research process showing the three iterative tasks designed to address the aim of the thesis.

phasis on Twitter

ii provide extra capabilities previously offered by related approaches in past studies

iii develop and demonstrate the applicability of the detection algorithm

Thus, to meet these goals, an iterative research process⁵ is considered as described in the following sections. Figure 1.3 depicts an overview of the iterative research process.

1.5.1 Methodology

The research methodology consists of three broad categories: *domain analysis*, *strategic solution* and *evaluation*.

Domain Analysis

To place the research in context and bring forward new knowledge, the *domain analysis* phase begins with a description of the problem and a review of relevant literature through a systematic approach. This is a crucial stage that enables understanding of the current research on clustering and community detection on Twitter. The *domain analysis* focuses on a holistic understanding of the relevant areas that could lead to achieving the thesis’s objectives. Consequently, the focus is on *social media* in general and Twitter in particular. Of interest, the research analyses how Twitter enables users’ *interactions*, *data availability* and nature of *research problems* being

⁵Following the *gradualist’s philosophy*, an iterative process is preferred to satisfy the evolving needs in the research.

addressed using tweets. This is followed by an in-depth engagement with the literature on *social networks* and *clustering problems* to identify relevant areas for contributions. Figure 1.4 shows a schematic overview of the processes involved in the *domain analysis* phase.

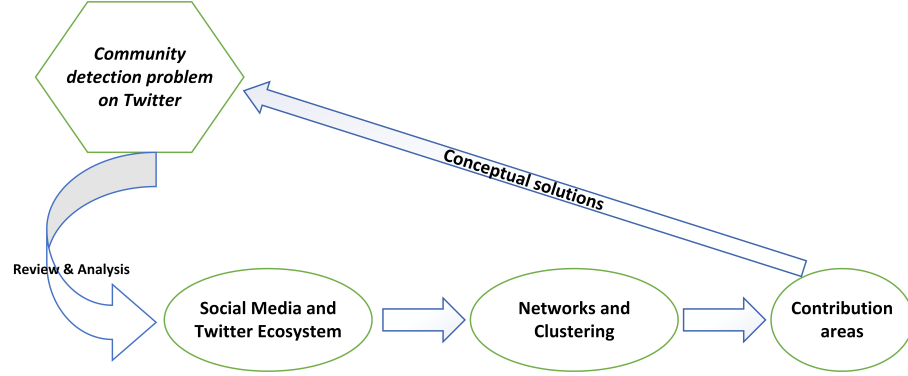


Figure 1.4: Sub-tasks in the *domain analysis* phase to analyse the state-of-the-art in clustering and community detection methods on Twitter and pinpoint contribution areas.

Strategic Solution

In the *strategic solution*, the focus is on the description of the solutions to the problems identified during the *domain analysis* stage. The process of making tweets available for analysis by Twitter has implications for the research problem and how it is being addressed. In Figure 1.2, the description of how the topological structure on Twitter leads to the formation of high-level communities either based on *metadata*, such as *hashtags*, or popular content triggered by the few influential users is provided. Basing a community detection on those factors may not reflect true connectivity and social cohesion of traditional communities. To investigate and formulate *socially cohesive communities*, a thorough analysis of clustering and community detection algorithms is required. To this end, the strength and weaknesses of each algorithm concerning clustering tweets are analysed in conjunction with the most appropriate similarity measure to utilise. This initial approach helps toward identifying contribution areas for study and inform its methodological standpoint. Essentially, the strategic solution provides the relevant mathematical formalisation in the research problem and practical solution alongside useful insights to inform how the research questions can be addressed. This phase is also concerned with ascertaining the suitability and credibility of the research data and relevant aspects that will ensure that the research objectives have been achieved. Chapter IV and Chapter VI provide a detailed

information in this respect.

Evaluation

The *evaluation* stage assesses the utility of the proposed solution relative to appropriate baselines in the literature before asserting the usefulness of the solution approach. To ascertain the efficacy and relevance of the research output, the evaluation process involves a thorough analysis and comparison of the research results based on the proposed methods with relevant baselines in the literature. Various techniques have been proposed to evaluate community structure; for example, the survey work of Papadopoulos et al. (2012) identifies vital action areas in community detection in social media networks. Such action areas include performance methods in terms of computation complexity – memory requirements and the possibility for incremental updates of detected communities and applicability of results in the context of the real-world web. Alternate forms of evaluation involve a manual inspection to assess the performance of the detection algorithm by identifying a common property or external attribute shared by all the members of the community (Yang & Leskovec 2015). Others involve determining how visible a community is within a more extensive network which can be measured using metrics such as the *concealment metric* (Waniek et al. 2018). Accordingly, the research outcome will be based on both quantitative, experimentations on various datasets and reproduction of results from select algorithms from the literature, and an informed extrapolation. Figure 1.4 describes the evaluation phase. The next section (1.6) provides an overview of the thesis structure.

1.6 Thesis Structure and Summary

In this section, an attempt is made to make each chapter of the thesis to be as independent as possible while maintaining a logical coherency with the remaining chapters (especially the chapters before and after a given chapter in the thesis). Primarily, the remainder of the document is structured as follows (see Figure 1.6 for a visual summary).

PART I: Introduction

The *Introduction Part* consists of two segments: the Introduction chapter (I) and *Research Approach* (1.5) section. The *Introduction Chapter* mainly consists of the thesis’s motivation,

research questions and contributions. The *Research Process* section describes the life-cycle of the study and how the main targets of the thesis have been met.

PART II: Background

This part consists of chapters about the background and the context of the study. The *Background Part* consists of two chapters: *Social Media Ecosystem* and *Community Membership Models*.

PART III: Problem Formulation

The *Problem Formulation* part consists of the following chapters: *Microcosm Detection Problem*, which provides a formal formulation of the problem, and *Authentication of Online Content*, which deals the research data and filtering strategy.

PART IV: Microcosms Detection

This part consists of *Microcosm: A Meta Analysis* chapter, which focuses on a pragmatic approach to analysing various connections and formation of ties on Twitter, and *Detection of Microcosms*, provides the implementation of the *MCT framework* for the detection of *microcosms* on Twitter. The *Detection of Microcosms* chapter focuses on identifying set of *structurally-related nodes* by exploiting social homophily and equivalence, which combines with a set of *textually-related nodes* to detect local community structures known as (*microcosms*). The *MCT* is focused at in-depth utilisation of the *bi-modality* for information search for local communities detection.

PART V: Result and Conclusion

The *Result and Conclusion* part consists of *MCT and Baselines* Chapter, which provides a detailed results from the implemented *MCT framework* and how it compares with existing benchmark models in the literature; the *Conclusion* Chapter provides the concluding remarks and reflection. Finally, the document ends with an appendix section that provides supplementary information about relevant topics raised in the thesis.

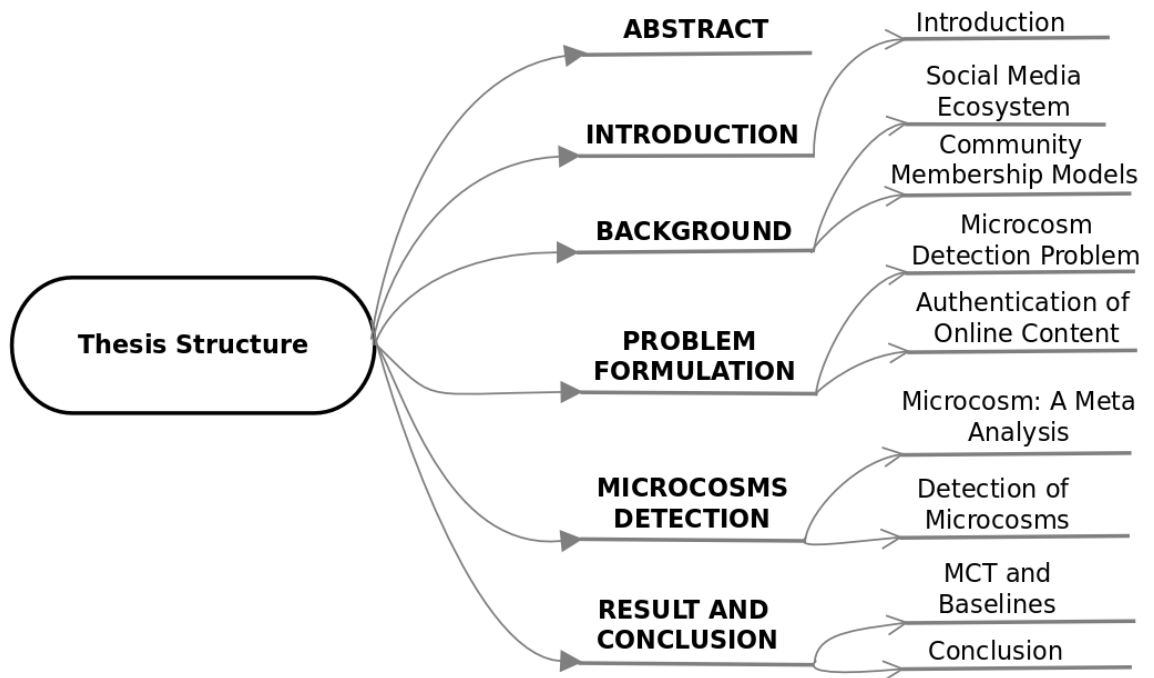


Figure 1.5: A high-level overview of the thesis Structure.

Publications List

The list of all publications resulting from the research activity is provided in this section. Majority of the thesis's content has been presented as part of conferences or journal publications. To maintain the desired logical coherency, the publications are given in chronological order – from the beginning of the research to the final research activity. Each published work is provided under the central research theme that led to its publication.

Research Proposal

Chapter I and Chapter VII are based on the original research proposal presented *In Companion of the The Web Conference 2018*.

- **Inuwa-Dutse, I.**, 2018. Modelling Formation of Online Temporal Communities. In *Companion of The Web Conference 2018 on The Web Conference 2018* (pp. 867-871). International World Wide Web Conferences Steering Committee. [(Inuwa-Dutse 2018)].

Content Authentication

- **Inuwa-Dutse, I.**, Liptrott, M. and Korkontzelos, I., 2018. Detection of spam-posting accounts on Twitter. *Neurocomputing*, 315, pp.496-511. [(Inuwa-Dutse, Liptrott &

Korkontzelos 2018)].

- **Inuwa-Dutse, I.**, Bello, B.S. and Korkontzelos, I., 2018. Lexical analysis of automated accounts on Twitter. arXiv preprint arXiv:1812.07947. **[(Inuwa-Dutse, Bello & Korkontzelos 2018)]**
- **Inuwa-Dutse, I.**, Bello, B.S. and Korkontzelos, I., 2018. The effect of engagement intensity and lexical richness in identifying bot accounts on Twitter. *International Journal on WWW/Internet*, 16(2). **[(Inuwa-Dutsea et al. 2018)]**.

Microcosms Detection

- **Inuwa-Dutse, I.**, Liptrott, M., Korkontzelos, Y. (2019) Analysis and Prediction of Dyads in Twitter. In: Métais E., Meziane F., Vadera S., Sugumaran V., Saraee M. (eds) *Natural Language Processing and Information Systems. NLDB 2019. Lecture Notes in Computer Science*, vol 11608. Springer, Cham. **[(Inuwa-Dutse et al. 2019b)]**.
- **Inuwa-Dutse, I.**, Liptrott, M., Korkontzelos, Y. (2019) Simmelian ties on Twitter: empirical analysis and prediction. *The Sixth IEEE International Conference on Social Networks Analysis, Management and Security (SNAMS-2019)*. **[(Inuwa-Dutse et al. 2019c)]**.

Search Optimisation

- **Inuwa-Dutse, I.**, Liptrott, M., Korkontzelos, Y. (2019) A Deep Semantic Search Method for Random Tweets. *Online Social Networks and Media, Elsevier*. **[(Inuwa-Dutse et al. 2019a)]**.

WIP

Work in progress – under review and about to be submitted:

- **Inuwa-Dutse, I.**, Liptrott, M., Korkontzelos, Y. (2019) Migration and refugee crisis: a critical analysis of online public perception; **[under review]**.
- **Inuwa-Dutse, I.**, Liptrott, M., Korkontzelos, Y. A multilevel clustering technique for community detection; **[about to be submitted]**

Chapter II

SOCIAL MEDIA ECOSYSTEM

2.1 Introduction

Social media ecosystem consists of platforms or networks such as Twitter and Facebook, which are very popular with the public. The *social media networks* have transformed the way sociological research is being conducted in terms of participants and size of data with profound effect; they offer useful utility in understanding modern society and how it functions. Users of the platforms can generate and consume content simultaneously leading to different kinds of information – fads, opinions, breaking news – on various social phenomena. The utility offered by modern social media makes it possible for users to socialise and serve as a news source (Sundaram et al. 2012). The quest to turn every aspect of humans' lives into computerised data for competitive value has been gaining momentum in the past decades. The *social media* is currently one of the most prestigious sources of commoditised data attracting huge attention. Because users can share information about virtually all aspects of their social life, social media platforms are ideal for studying various aspects of social events such as the detection of local communities. Thus, it is worthwhile understanding the *social media ecosystem* in great depth; hence, the purpose of this chapter.

2.2 Social Media Ecosystem

As a result of technological advancements, various aspects of social phenomena are witnessing transformative process at a faster pace. For instance, communication and interaction of people witnessed a tremendous transformation, especially with the advent of Online Social Media Networks or Platforms (OSMPs). *Social Media Network*, a fusion of *social* and *media network*

(Dijk van Jan 2006), facilitates social interactions of diverse users on a larger scale. Social interactions continuously evolve to support a myriad of objects to remain connected, leading to a highly connected and dynamic social media ecosystem that enables various forms of interactions among diverse people. Within this complex ecosystem, multiple types of communications happen at different layers of granularity and intensity. Thanks to the architecture of social media network, which is designed to enable users to both generate and consume information. This capability contrasts it with the early unidirectional *two-step* communication model in which only a handful of individuals serve as intermediaries between the mass communication and the public (Katz et al. 2017). The contemporary social media ecosystem consists of numerous platforms which support various aspects of humans' social engagements and enable users to act as both *producers* and *consumers* of information. Unlike the early *two-step* flow model, the influence network model of Watts & Dodds (2007) enables *multi-way* flow of information where users can simultaneously generate and consume information. By allowing the multi-way flow of information, the communication model of the OSMPs can be likened to the *influence network model*. The utility offered by modern social media makes it possible for users to socialise and serves as a news source.

2.2.1 Social Media Platforms

The online social media is one of the defining phenomena in this *technology-driven* era in which various platforms play an instrumental role in enabling global connectivity. Social media networks have been instrumental in globalisation and enable socio-technological research to understand modern society better. Essentially, the communication model of the *OSMPs* is a form of *multi-flow channel*, which enables both the consumption and generation of information by its users and have devised various means to boost the number of active users using various strategies such as the promotion of enticing contents that will attract wider attention. A closer examination of Twitter will reveal the dominant effect of *influential users* thereby creating a logical division: a clique of *content pushers* and *consumers* (see Figure 2.1), which resembles the *two-step flow model* (Katz et al. 2017). The division in Figure 2.1 is being strengthened due to various strategies¹ used by the OSMPs to support users to increase their network of

¹Such as the promotion of contents that will entice users to use and engage more with the platforms or recommendations to follow or add a friend.

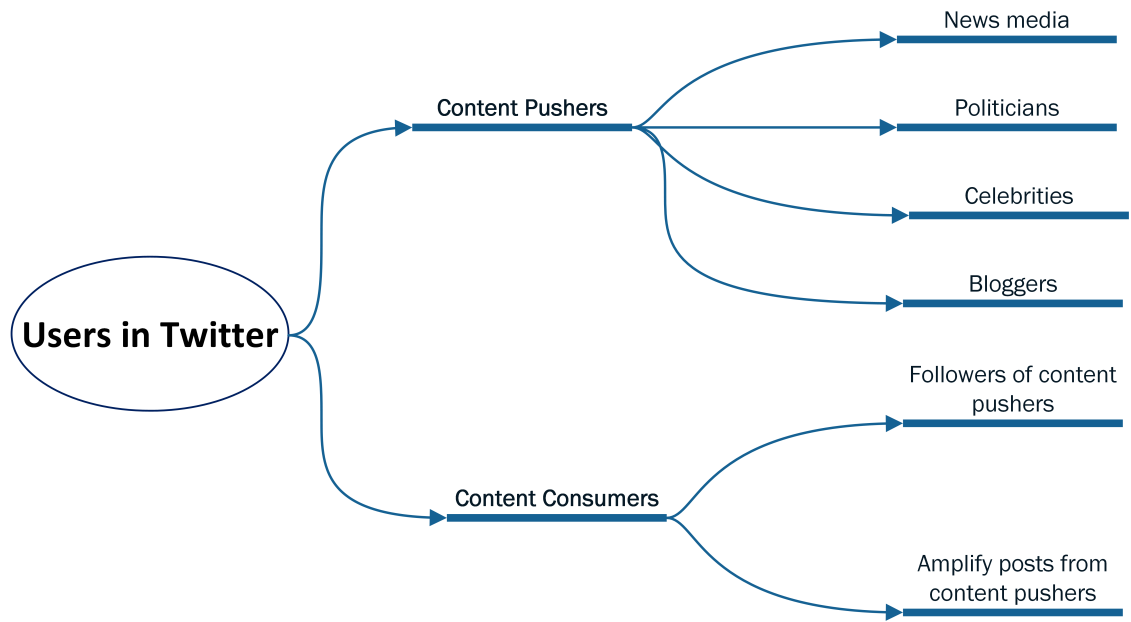


Figure 2.1: Example of users on Twitter showing *content pushers* such as news channel, celebrities, and politicians, who often generate a high proportion of tweets and attract large followers. In the other extreme, *content consumers* or *followers* further amplify posts from the *content pushers*.

friends which in turn generates more value to the platforms. The task of community detection is greatly simplified under this scenario since the communities are explicit (Kwak et al. 2010). With the apparent preference in favour of influential users, research data from Twitter is prone to bias and challenging to detect communities not based on trending topics or popular hashtags mostly triggered by prominent users. Communities with a low presence on Twitter are implicit. They require extensive exploration to understand the underlying mechanism governing their behaviour (Palla et al. 2007). This thesis posits that despite the freedom and flexibility to disseminate information in the present-day OSMPs, the flow of information is being influenced by a few users making it difficult to detect relevant segments within a network like Twitter.

2.2.2 Twitter

The growing relevance of online socialisation, facilitated by many platforms such as *Twitter* and *Facebook*, attracts much research interest and questions to be addressed. The problem of *detecting microcosms in social networks* is described from the perspective of *Twitter*, which facilitates low-level access to news in real-time distinguishing it from conventional media (Kwak et al. 2010, Cataldi et al. 2010). There are no practical reasons that will prevent the proposed

approach in this research from being applicable in other social network platforms. However, the contrast could be true if the approach is only on other platforms where the formation of a reciprocal tie is the default setting. For instance, using *Facebook*, without recognising the particularity or the eccentric topological structure on *Twitter* (see Figure 1.2) will make the approach less generalisable. Hence, focusing on *Twitter* ensures a more encompassing framework that can be mapped to other networks. In a nutshell, *Twitter* is chosen because of the following reasons:

- the peculiarity of connection patterns, e.g. the ability to follow anyone results in many unidirectional connections between socially disconnected users; this porosity of connection challenge many tasks such as *community detection* and *content authentication*.
- data from *Twitter* is easily available on a relatively larger scale² compared to other platforms.
- it is difficult to deal with *content* from *Twitter* due to the flexible posting styles.
- the requirements that satisfy *Twitter* can be easily mapped to other platforms.

Connections on Twitter

While many properties are common across various networks, social networks exhibit different properties (Newman & Park 2003). This deviation can be seen to be rooted in the methodological point of view. Social network theorists hold two methodological positions in investigating social relationships: *realist* and *nominalist*. The *realist* proceeds with a preconceived notion of the existence of relationships in a network which need to be discovered while the *nominalist's* approach is based on the questions asked by the investigator (Laumann et al. 1989). Furthermore, the formation of social ties can be based on *event-type ties* or *state-type ties*. An *event-type tie* is transitory and often results in socially distant members. With respect to *Twitter*, *event-type ties* consist of subscription to *trending hashtags* or *retweeting a popular user*; see Figure 1.2. On the other hand, a *state-type tie* is based on static or structural connectivity between users, which suggests a certain degree of familiarity and trust (Borgatti & Halgin 2011). The *Twitter* social platform facilitates global connections and interactions of diverse

²for instance, users on *Twitter* can download up to 6000 tweets per second within the allowed 1% threshold

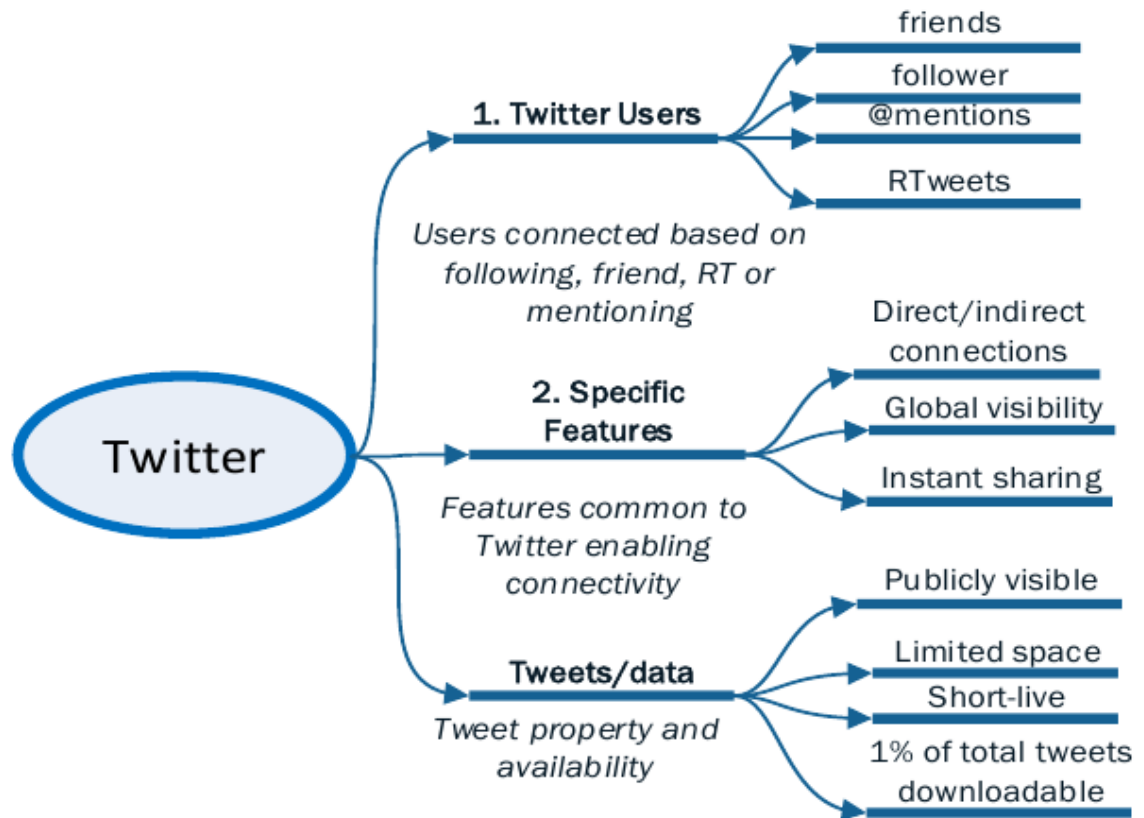


Figure 2.2: An overview of relevant categories of attributes that support global inter-connectivity among users on Twitter. The features utilised in this study are derived from these categories directly or indirectly.

users (Qazvinian et al. 2011). Figure 2.2 presents an overview of the platform and its relevant attributes that enable users to connect and interact. According to Figure 1.2, connections on *Twitter* may manifest differently, such as *sharing a link*, *re-tweeting (RT)*, using the same or similar *hashtags*, *user mention (@)* or *follower-ship*. Evidently, the connection is porous allowing a users to connect with many diverse users and limiting the chances of *reciprocal ties* or *dyads*. The thesis argues that the presence of random connection among some users on Twitter (see Figure 1.2) contributes to the limited overall cohesiveness, and the growing proportions of *fake* and *spam contents*. The importance of a small group of users with a positive relationship has been recognised as a critical feature in the structural analysis of networks (Freeman 1996). Anecdotal and cognitive evidence suggests that users are more likely to believe information from closely related individuals (Carley 1991). Moreover, according to *cognitive balance*, strong ties among users prevent misuse of the network and any form of *psychological strain*

(Newcomb 1978). This study opined that, if a user genuinely engages with other users in a bidirectional means, it will maximise the chances of detecting more cohesive communities, and help in curbing the circulation of irreverent information from unknown sources. Users with reciprocal ties are more likely to be genuine, trustworthy and will probably result in more cohesive clusters.

Social Media Data

Until recently, access to a large amount of data is exclusive to large research facilities such as weather forecast stations, astronomical stations or scientific laboratories (Dijk van Jan 2006). The advent of the *social media platforms* have transformed the way sociological research is being conducted in terms of participants and size of data with profound effect. The platforms offer useful utility in understanding modern society and how it functions (Miller et al. 2015). Within the *social media ecosystem*, it was estimated that 2.46 billion users are connected and by the year 2020, one-third of the global population will be connected³. This massive connections of diverse users contribute a significant proportion of the current data traffic. *Datafication*⁴ is a relatively new term that denotes the continuous quest to turn every aspect of humans' lives into computerised data for competitive value (Cukier & Mayer-Schoenberger 2013). The *social media* is currently one of the most prestigious sources of commoditised data attracting huge attention, and several domains have already recognised the crucial role of social media analysis in improving productivity and gaining competitive advantage. Some common use case examples where information derived from the *social media* has been utilised include health-care to support effective service delivery (Rojas et al. 2016, Yee et al. 2008), in sport to engage with fans (Davenport 2014), in the entertainment industry to complement intuition and experience in business decisions (Deloitte 2014) and in politics to track election processes, promote wider engagement with supporters (Contractor et al. 2015) and predict poll outcomes. However, alongside the benefits, the rapid increase in social media spam content threatens the credibility of research based on analysing this data. As a precautionary measure to avoid compromising the research outcome by irrelevant or unrepresentative data, the research proceeds by developing an effective method for spammers and automated accounts detection, and offers crucial

³www.statista.com/topics/1164/social-networks

⁴see <https://www.foreignaffairs.com/articles/2013-04-03/rise-big-data>

insights into the sophisticatedly evolving techniques for spamming on Twitter. See Chapter V for details.

Tweets

In this age of social media, a massive amount of data can be obtained easily. Data from platforms such as Twitter is prompt with high-propagation capacity. The posts that users share on Twitter are called *tweets*, short text snippets on Twitter and they enable longitudinal studies involving various social aspects (Würschinger et al. 2016). A *tweet object* is a complex data structure consisting of many attributes describing specific information about the *tweet* and the *account holder*. It is normally returned as a *JavaScript Object-Notation (JSON)* format, which makes it relatively easy to retrieve specific segments. As a marked-up piece of text, the different fields in the tweet object define important characteristics of the tweet; see Section 1.1.1 in Appendix A for some examples of important components. The complexity of a tweet and its unstructured nature makes it difficult to process directly into a usable form; hence, the need for a series of preprocessing before effective analysis can be conducted. The stream of tweets differs from conventional stream texts in terms of *posting rate*, *dynamism* and *flexibility*. Tweets are generated at a rapid rate and tend to be highly dynamic (Guille & Favre 2015); *tweets* are generally informal, sparse and display an inherently poor structure that results from various writing styles. These attributes make *tweets* a noisy data consisting of widespread abbreviations and personalised terms (Chakraborty et al. 2016). Moreover, *fake news* and *content* from *automated accounts* or *social bots* are prevalent, which questions the credibility of data without active filtering. On the issue of processing, social media data is particularly challenging to process because standard writing styles are not usually adhered to (Pang et al. 2008, Chakraborty et al. 2016). Thus, utilising social media data for research without extensive filtering to retain only the relevant part will lead to a spurious result. However, with proper preprocessing, social media data is capable of providing a compelling research outcome (Miller et al. 2015). In response to the concerns mentioned above, this research begins with an in-depth study that mitigates the challenges related to the *quality* and *processing* of the research data. Chapter V provides a detailed description of the data cleaning strategy. The need for a custom method in this respect is to ensure the credibility of the data and to utilise an approach that meets the requirements of the research.

2.3 Network Communities and Sociometry

Phenomena in real life are associated with numerous network structures and embedded communities. Network dynamism leads to changes in size and configuration of entities within the network (Aggarwal & Subbian 2012). This is more prominent in the *web* and *social networks* due to the continuous addition of users that results in a network with a shrinking diameter over time (Backstrom et al. 2006). Relationships and structural properties of networks have been extensively studied at different levels of granularity and sophistication, ranging from the structure of microscopic organisms to complex networks, such as the internet (Erdős & Rényi 1960, Watts & Strogatz 1998, Scott 1988, Albert & Barabási 2002). While many properties are shared across various networks, social networks are different concerning the degree of correlation and tendency for clustering; the formation of clusters is more straightforward and the correlation degree between users tend to be positive (Newman & Park 2003). Information diffusion and evolution of communities have been extensively studied from different fields, and social networks provide means of information diffusion via *shared content* and *community formation*. Users continuously engage and disengage in discussions with varying degree of interactions, leading to the formation of distinct online communities. Such communities are often formed at high-level either based on metadata, such as hashtags on Twitter, or trending content triggered by a few *influential users*. As such, these online communities often do not reflect true connectivity and lack the cohesiveness of traditional communities. From Figure 2.1, if *content pushers* are excluded from the networks, a significant layer of users which engage and form communities at a microscopic level will be uncovered. Consequently, the study investigates the detection of a socially cohesive community of users and analyse relevant aspects of *sociometry*.

2.3.1 Online Interaction and Local Community

The social media ecosystem enables various forms of interactions among diverse users at multiple levels. *Online socialisation*, facilitated by platforms such as *Twitter* and *Facebook*, attracts much research interests and poses many questions. The architecture in *OSMPs* follows a model designed to influence or support users to expand their networks, thereby resulting in a massive network of densely interrelated users dubbed *friend-of-a-friend network*. Managing social relationships has been linked to the cognitive capability of the human brain, in which a large

network size affects a user's ability to maintain a cohesive social interaction. It has been reported that humans are capable of attaching names to about 2000 faces, but have a cognitive group size of about 150 to actively maintain social relationships (Dunbar 1998). This limitation will be more pronounced in platforms such as Twitter where a user can have a vast network size making detection of socially cohesive groups a challenge. In platforms such as Twitter, the ability to follow anyone results in many unidirectional connections between socially disconnected users and ultimately affects clustering users and, in turn, the integrity of online content. The importance of a small group of users with a positive relationship has been recognised as a critical feature in the structural analysis of network (Freeman 1996). Due to strong social cohesion, a small group of users, of about five members, are more intimate with a high degree of familiarity (Dunbar 1998). A sufficient understanding of the structural properties of online platforms is considered as a crucial factor in the design of a more *human-centric* future internet (Arnaboldi, Guazzini & Passarella 2013). However, the growing complexity and heterogeneity of connections make the task of identifying social relationships at the micro-level more challenging.

A *community* can be simply defined as a collection of entities that share common space and values. Participants in the same community exhibit a high degree of similarity, while those in different communities exhibit no or low levels of similarity (Lancichinetti et al. 2009, Newman 2004a). A local community is a crucial organising principle, especially in a vast network in enabling a better understanding of the structure and function of networks (Watts & Strogatz 1998, Newman & Park 2003). What constitutes a *community* may differ, but the central concept seems to apply to all communities – *entities*, *links* and *interactions*. With a computer-mediated communication (CMC) research work, there is a growing interest in the characterisation of *virtual communities* that transcend geographic communities (Sundaram et al. 2012). On Twitter, such communities evolve by sharing common interests on a topical issue with other users across the globe (Yang & Leskovec 2012), and the formation of such communities may manifest in different ways such as *sharing a link*, *retweet (RT)*, use of similar *hashtags*, *user mention (@)* or *follower-ship*. All these aspects can lead to random communities of users as exemplified in Figure 1.2. The problem of community detection on Twitter is mostly centred around a directed form of connections, i.e. based on *event-type ties* according to the *realist's* approach. While

this is valid in many networks, such an approach could lead to many unrelated sets of users. However, by focusing on smaller groups with a high degree of *structural* and *content* similarities, it is possible to identify communities, which are homogeneous to many sociodemographic behavioural, and intrapersonal characteristics (Miller McPherson et al. 2001).

2.3.2 Sociometry

Social networks come in various forms depending on the cohesiveness and size – from the most intimate to tenuous relationships. For a long time, a *social network* has been recognised as a useful tool for linking *micro* and *macro* levels of sociological theory (Granovetter 1977). As a result, many forms of social relationships have been analysed at various levels, from the structure of microscopic organisms to complex networks, such as the internet (Scott 1988, Watts & Strogatz 1998, Albert & Barabási 2002). It can be argued that understanding social interactions today would be incomplete without taking online social relationships into account, where various forms of interactions among diverse users happen. This capability makes it possible to empirically quantify and evaluate social relationships among users on an unprecedented scale. Primarily, many *social network theories* and analytical solutions can now be tested using real social data from platforms such as Twitter. In the *social science* domain, *sociometry* is a means to measure social relationships⁵ between people (Wasserman & Faust 1994). *Sociometry* can be analysed along various dimensions and this research focuses on *communities of users*, *homophily*, *centrality metrics* – degree, closeness, betweenness – and content diffusion and veracity.

Homophily

Homophily is a social phenomenon that suggests individuals with a certain degree of similarity are more likely to interact, and it is central to many aspects of human's social interaction (McPherson et al. 2001). Homophily describes the tendency for humans to connect with people of similar characteristics. For example, *homophily* has been investigated in the context of *geolocation* and *popularity* (Kwak et al. 2010), and how users in reciprocal ties discuss similar topics (Weng et al. 2010). The notion of *homophily* is related to the idea of *equivalence*, which suggests that individuals compare themselves with one another and adopt similar attitudes and

⁵Usually via the use of a relational data that is often presented in two-way matrices or *sociomatrices*

behaviours of, those who occupy an equivalent position in a network or an organisation (Brass et al. 1998). Correspondingly, there are two forms of equivalence: *structural equivalence* and *regular equivalence*. A *structural equivalence* refers to two actors having similar interaction partners, which can be mapped to a *state-type tie*, even if they are not directly connected. On this basis, *structural similarity* according to user's attributes can be inferred (see Figure 4.4). On the other hand, a *regular equivalence* refers to similar patterns of interactions, for instance, profiles information in the context of Twitter, even though the interaction partners may be entirely different (Scott 1988). In the context of this work, profile information can be applied to study such equivalence, and the ideas are explored further in Chapter IV.

Centrality metrics

Because the taxonomy of social relationships is expressed as a function of closeness among users in a network, in which the closer the users are, the more cohesive and trustworthy, it is possible to analyse various centrality measures. Spectral clustering is instrumental in identifying relevant groups or clusters induced by users' connections. Thus, the detected clusters will enable the analysis of metrics associated with centrality such as *degree centrality*, *closeness centrality* and *betweenness*.

Personal Network and Reciprocal Ties

People's networks are homogeneous concerning many sociodemographic behavioural, and intrapersonal characteristic (Miller McPherson et al. 2001). A small group of users, consisting of about five members, promotes strong social cohesion (Dunbar 1998), and its importance has been recognised as a critical feature in the structural analysis of networks (Freeman 1996). Focusing on smaller groups will be more desirable in discovering personal network, which are homogeneous concerning many sociodemographic behavioural, and intrapersonal characteristic (Miller McPherson et al. 2001), on Twitter. This thesis considered the following *reciprocal ties* as crucial in identifying a form of *personal network*.

Dyadic ties: In social networks, the concept of a *dyad*, or *reciprocity*, has been viewed from various perspectives and often with contradicting results. Previous studies (Weng et al. 2010, Kwak et al. 2010, Cha et al. 2012, Arnaboldi, Conti, Passarella & Pezzoni 2013) have examined

reciprocity for various tasks which are either based on a *directed set of nodes* or *textual content*. Regarding how popular users prefer to follow other popular users, Kwak et al. (2010) reports a low percentage of reciprocity and a high proportion of directed or unreciprocated connections on Twitter. However, Weng et al. (2010) reports a high percentage of reciprocity by computing the ratio of *follower/following* on Twitter. The analysis in this work is based on the premise that the level of trust is stronger among users that share dyadic ties and it is highly unlikely for a user in the group to misuse the network e.g. spread fake news or spam. However, acquiring large amounts of tweets sufficient to identify such cohesive groups is challenging and time-consuming because *dyadic ties* on Twitter are rare due to the prevalence of directed ties. A directed tie is peculiar to Twitter since, in other platforms such as Facebook, an automatic reciprocal relationship is established once a friend request is accepted. The research argues that the connection topology on Twitter contributes to widespread spurious content and a less cohesive community of users due to many unreciprocated or *event-type ties*.

Transitive or *Simmelian ties*: Transitivity defines a social preference to be friends with a *friend-of-a-friend* and has been recognised as a common feature of a network (Watts & Strogatz 1998). The concept of a *transitive relationship*⁶ is similar to a *Simmelian tie* (Simmel et al. 1950), which is referred to as a strong social relationship within three-person groups or more. Transitive ties are crucial, especially by noting how the flexible connection types on Twitter makes it difficult to establish reciprocal ties that could lead to *Simmelian ties*. This form of relationship is essential toward understanding the structure of social ties in a network. However, as illustrated in Figure 1.2, the prevalence of transitory connections makes it challenging to identify transitive links in accordance with *state-type ties* on Twitter. Arguably, this also explains the limited number of studies exploring such links in clustering and content validation tasks. The position of the research is that such limitation contributes to the challenge of detecting socially cohesive communities and the proliferation of spam and fake contents on Twitter. For network analysis, transitivity is a vital feature of a network that enables the formation of cohesive communities and (Granovetter 1977). In this analysis, a set of *transitive users* is regarded as a facilitator for the detection of socially cohesive communities of users, which relies on the premise that the underlying mechanisms to predict transitivity on Twitter is understood, then

⁶Transitive and Simmelian ties are synonymous in this study

tasks such as cohesive clustering and content validation can be greatly enhanced.

2.4 Summary

The popularity of social networks has attracted huge interest in the utilisation of social media for studying various aspects of humans' lives. Several domains spanning health-care, sport, entertainment, politics, and all forms of social phenomena have already recognised the crucial role of social media analysis in improving productivity and gaining a competitive edge. This chapter presents relevant areas and challenges that relate to the thesis and how the identified challenges motivate the approach proposed in the thesis. The next chapter (Chapter III) describes in details community detection and clustering methods.

Chapter III

COMMUNITY MEMBERSHIP MODELS

3.1 Introduction

As alluded in Chapter II, the complexity and dynamism of the *social media ecosystem* results in multiple types of interactions at different layers of granularity and intensity, ranging from *global* or *local*, *positive* or *negative*, *influential* or *less influential*, *high-level* or *low-level*. Such interactions culminate in the formation of communities at various levels within the network. This chapter describes relevant existing methods of identifying community structures in a network, and what qualifies a network entity to belong to a community.

3.1.1 Relevance of Community Detection

The set of closely-related individuals constitute a close-knit, or a small-group is useful for understanding the structural property of a network (Freeman 1996). A small group of users implies a high degree of familiarity due to strong social cohesion (Dunbar 1998). According to the theory of *cognitive balance*, if strong ties exist among three users, anything short of positive relationship would lead to a *psychological strain* and will be avoided Newcomb (1978). Related insights from anecdotal and cognitive evidence suggest that users are more likely to believe information from closely-related individuals (Carley 1991); this is an indication of the usefulness of identifying communities within a large network for various purpose. For instance, an increase in a network size affects the capability of maintaining strong social relationships – an increase in a network size diminishes community cohesion (Dunbar 1998). This research posits that the limitation will be more pronounced in social platforms, such as Twitter, where a

user can have a vast network size making detection of socially cohesive groups a challenge. In response to the growing volume of online content, identifying community is a critical endeavour across various domains for many reasons. Consequently, community detection task:

- serves as a fundamental tool to discover the organisational principles in networks (Albert & Barabási 2002, Yang et al. 2013), thus enabling a means of analysing complex networks (Williams & Martinez 2000).
- provides an effective means of analysing content in a network as a unit and allows for the detection of a collection of related pages in World Wide Web (Flake et al. 2002, Papadopoulos et al. 2012).
- enables the detection of social circles or cliques in social networks (Newman & Park 2003) and facilitates an intelligent recommendation (Cao et al. 2015).
- helps in studying the social behaviour of users (Arnaboldi, Guazzini & Passarella 2013).

3.2 Network Models and Community Structure

Advances in technological innovation brought about significant changes in the way real and virtual networks are being studied. The ability to computerise data acquisition methods and the development of sophisticated data manipulation tools have paved the way for the discovery of various network models. Through statistical analyses, many interesting properties of various networks have been discovered. According to Watts & Dodds (2007), the complexity of a network can be described using the following key concepts:

- *Clustering coefficient* is a useful metric that is used to quantify the clustering tendency of a given node in relation to other nodes within a network. The metric implies the notion of *my friends will likely know each other*, as seen in a small group. To quantify the clustering coefficient for a given node (say, i^{th} node) in the network, firstly the number of edges between the first neighbours of the i^{th} node is computed using:

$$\#edges_i = \frac{k_i(k_i - 1)}{2}$$

where k_i denotes the *number of edges connecting i^{th} node* to k_i other nodes in the network. It follows that the *clustering coefficient* can be defined:

$$C_{coeff_i} = \frac{2E_i}{k_i(k_i - 1)}$$

where E_i denotes the *actual number of existing edges between k_i nodes*. The clustering coefficient for each node in the network is computed using:

$$E_i \propto \frac{k_i(k_i - 1)}{2}$$

Thus, the *clustering coefficient* C_{coeff_i} , of the whole network, turns out to be the mean of individual C_{coeff_i} 's with respect to the i 's in the network.

- *Degree distribution* quantifies the distribution of nodes in a network such that the spread of edges in the network is captured by a probability distribution function $p(k)$, in which $p(k)$ gives the probability that a random node in the network has exactly k edges. Accordingly, many large networks have been shown to exhibit *power law degree distribution* or *scale-free network* (Barabási et al. 2002).
- *Small-worldliness* is a useful feature, which has been shown to be present in numerous networks with the following properties (Watts & Strogatz 1998): *short path-length* (i.e. many structured short-range connections and few random long-range connections), *diameter of the network* is exponentially smaller than their size and bounded by a polynomial in $\log N$, where N is the number of nodes in the network.

3.2.1 Network Models

The structure and properties of various networks have been examined in the past, and various networks in nature are characterised as possessing varying degrees of organisation (Erdős & Rényi 1960, Scott 1988, Watts & Strogatz 1998, Albert & Barabási 2002, Lancichinetti et al. 2009). The underlying design principles in networks based on topological structure and other quantitative measures associated with the network are useful toward understanding complex networks across different domains (Albert & Barabási 2002). Classic network models such as the *random graph* (Erdős & Rényi 1960), *small-world* (Watts & Strogatz 1998) and *scale-free*

(Barabási et al. 2002) form the basis on which networks and local communities are studied. The following section gives a brief description of the models.

Random Networks

Previously, complex networks have been considered to be random, and one of the oldest network model proposed in Erdős & Rényi (1960) was used to identify relevant structures. In its simplistic design, the *random network model* views a complex network as consisting of N *connected nodes* such that for any pair of nodes (say, n and m), the connection probability p , leads to a graph whose edges are randomly distributed with an approximate value of $p(\frac{N(N-1)}{2})$. Networks that are modelled according to the *random graph model* exhibit skewed distributions. The skewed distributions contrast the *random model* from most networks, which have *power degree distribution* or its variants, e.g. *truncate power-law* or *exponential* or *strongly peaked distributions* (Newman 2003). Generally, the *random model* is viewed as lacking clear design principles; as such, any network with no known design principle can be modelled using the model (Albert & Barabási 2002). Also, it is limited in modelling complex phenomena since most networks in nature are not random, but based on underlying guiding principles (Lancichinetti et al. 2009). However, one of the profound effects of the *random graph model* is its utility as a kind of a *litmus test* for the existence of a regular network. The idea is to investigate how well the topology of a given complex network deviates from that of a random graph. If a significant deviation is observed, then there exists an organising structure which needs to be uncovered; otherwise, the network is random (Watts & Dodds 2007).

Small-world

The idea of a *small-world* was first empirically investigated by Travers & Milgram (1977), in which an average of five chains (culminating to the famous six degrees of separation¹) are enough to reach two random individuals via a network of acquaintances. This notion has been shown to manifest in many complex structures such as *metabolic systems*, *genes pathway*, and *technology* (Watts & Strogatz 1998). In modelling the *small-world* phenomenon, it is assumed that connection topology in a network is completely regular or random. Most networks in biology, society and technology lie somewhere in between these two extremes (i.e. completely

¹This is based on the concept of *small-world* and has been dramatised in Guare (1990)

regular or random) (Watts & Dodds 2007). Most complex networks are believed to exhibit small-world behaviour such as *short path length* and *many structured short-range connections* (Kleinberg 2000, Kwak et al. 2010, Bakhshandeh et al. 2011, Szüle et al. 2014).

Scale-free networks

A substantial number of real-world networks are dynamic and witness changes in terms of size and configuration of entities in the network (Aggarwal & Subbian 2012). For example, networks such as the *World Wide Web (WWW)* and the *social networks* are characterised as possessing a *short diameter* that shrinks over time (Backstrom et al. 2006, Aggarwal & Subbian 2012). This property makes it crucial in studying the dynamics in networks as any modification in the configuration of the network (e.g. an addition or removal of a node) often results in a change in the network's diameter. In response to the limitations of *random graph* or *small-world* models in that they lack temporal variables, *scale-free networks* (Barabási et al. 2002) are useful in modelling dynamic networks with temporal aspects such as the WWW (Newman 2006). The *scale-free model* exhibits *exponential growth*, incorporates a *preferential attachment* and displays a high *clustering coefficient* (Albert & Barabási 2002).

3.2.2 Community Structure

As the size of a network increases, the possibility of fragmentation increases (Berelson & Steiner 1964, Shaw 1971), leading to a decrease in the homogeneity of behaviour and attitude across groups (Granovetter 1992). Because similarity breeds attraction and interaction, communities of similar members are formed, which result in the formation of strong ties among members of the same communities (Brass et al. 1998). Noting that nodes in the same community share high degree of similarity and edges that run among communities are relatively low, the problem of community detection is usually formalised to identify relevant partitions in the network that satisfy specific requirements. A *community structure* in a network is characterised as possessing densely connected groups of nodes and sparser connections between other communities (Newman 2004b). An important goal of a clustering technique is to summarise the relationships between objects in a network and many forms of the network have been analysed to understand how various community structures manifest. With structured data, a *scoring function* is used to easily identify a community structure in which set of nodes or vertices are

integrated based on their relatedness (Chen et al. 2009).

Scoring Function

The goal of a scoring function is to identify a pair of nodes, which are closely related in some respect. Depending on the *scoring function* used, various similarity matrices can be constructed. The choice of an effective similarity measure is crucial due to its strong correspondence between the ability of a clustering algorithm to correctly identify groups and the *signal-to-noise-ratio* within the matrix of instances (Lawson & Falush 2012). *Block-modelling* (see Section 3.3.1) is an approach that enables the grouping of network entities according to the extent to which they are equivalent using a set of empirical procedures. Depending on the procedural approach, the definition of equivalence usually leads to a different partitioning of a network according to the following approaches (Doreian et al. 2005):

- equivalent units have the same *connection pattern* to the *same neighbours*, and
- equivalent units have the same or similar *connection pattern* to *different neighbours*.

In social network research, equivalence suggests that individuals compare themselves with one another and adopt similar attitudes and behaviours of others who occupies an equivalent position in the organisation (Brass et al. 1998). There are two forms of equivalence: *structural equivalence* and *regular equivalence*. The *structural equivalence* refers to two actors having similar interaction partners, which is mapped to the idea of *state-type ties* in this research to infer *structural similarities* in terms of users' attributes (see Section 7.2). The *regular equivalence* denotes similar patterns of interactions, even though the interaction partners may be entirely different (Scott 1988). In the context of this study, these two equivalences have been considered in defining a useful scoring function that is targeted at identifying a high degree of similarities among users using global and local information. In line with this, the thesis proposed a composite scoring function based on *structural similarity* and *textual similarity*. A detailed description is given in Section 4.3.

3.3 Detection Task

A community is a functional unit of a network that captures local relationship among the network objects. Because a complex network is considered as a composition of many sub-networks

or communities, and one of the vital tasks is to identify the community structures or network communities (Newman & Girvan 2004). The detection of a community structure is of interest to myriad researchers across various domains, hence attracting many interests and research perspectives, notably from *Social Science* and *Computer Science* domains. To this end, many algorithms have been proposed in the past to detect community structures across various networks. Zhang & Zaïane (2018) highlights a useful categorisation of detection algorithms along the following *deterministic* or *non-deterministic* dimensions: graph-based algorithms (Newman 2013, Kernighan & Lin 1970), clustering-based algorithms (Girvan & Newman 2002, Newman 2004b, Blondel et al. 2008), genetic-based algorithms (Pizzuti 2008), and label-based propagation algorithm (Raghavan et al. 2007). The underlying difference across many network communities is how the connections are defined – some links are deterministic or certain while some are non-deterministic or uncertain, but based on probability (Zhang & Zaïane 2018). Figure 3.1 shows a summary of relevant methods in the *detection task*.

3.3.1 Detection Approach

A detection task entails prediction and quantification activities to identify sets of related nodes and relevant details about the network data (Chen et al. 2009). As described in Section 3.2.2, the act of predicting membership and assigning items to a community or cluster is achieved using a measure of equivalence or a scoring function. The *scoring function* also enables *block-modelling*, which is based on the idea that units in a network can be grouped according to the extent to which they are equivalent using a set of experimental procedures. Depending on the procedural approach, the choice of equivalence usually leads to detecting different partitions within a network. One of the reasons is related to the fact that communities are formed around two primary modalities or sources of information: (1) the *network structure* and (2) the *features and attributes of nodes*. However, studies mostly focused on one aspect, not both. Few studies in the past have applied both modalities in detection tasks (Balasubramanyan & Cohen 2011, Leskovec & Mcauley 2012, Yang et al. 2013). Of these studies, a closely-related work to the approach in this research can be found in Yang et al. (2013), in which a generative model for networks with node attributes is proposed. However, the approach differs from the *MCT approach* as follows. The node attribute considered (*hashtag* is used) is insufficient in analysing the depth of similarity between network entities in a complex environment like Twitters. There

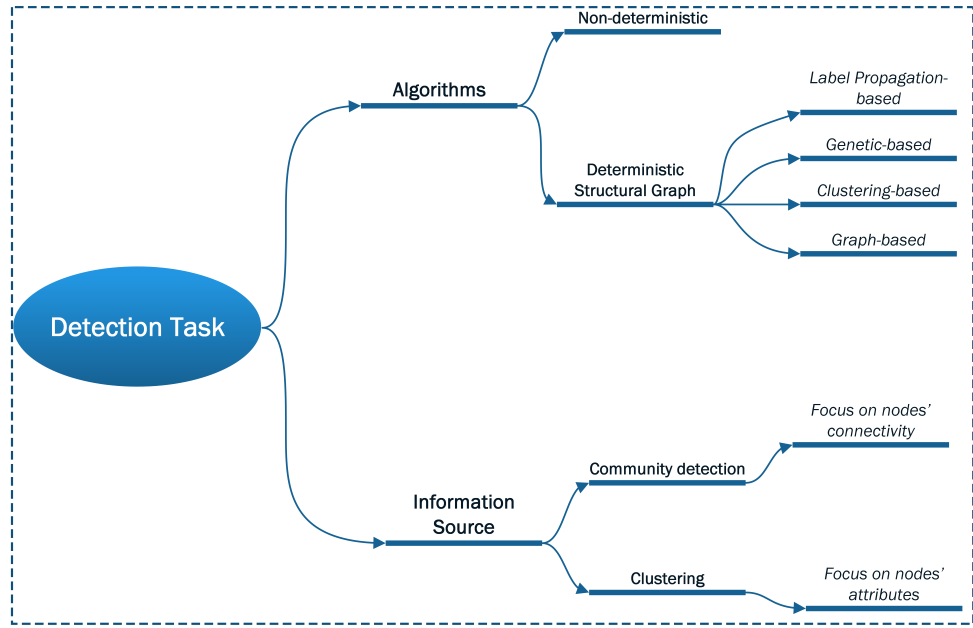


Figure 3.1: A summary of key aspects – *algorithm* or *method* and *information source* – in community detection tasks. Depending on the problem, the detection method further sub-divides into deterministic or otherwise. Majority of studies proceed with a preconceived notion of a community structure in a network; hence, the deterministic approach is often the preferred one.

is a need to include aspects of the text and other profile information for in-depth structural insight. In many cases, structural and textual information evolve simultaneously, and each is used as a basis for the formation of communities. Hence, identifying communities of nodes with shared information in both components will ensure the detection of more cohesive communities.

Identifying Partitions in a Network

The task of partitioning a network into distinct groups has engaged many experts from various domains leading to various perspectives. There are basically two principal lines of research in this direction: *graph partitioning* and *block* or *hierarchical modelling* (Newman 2004a). An alternate view presents a categorisation along the following dimensions: *method-based* and *dimensionality reduction* and *graph partition* and *hierarchical* methods (Aggarwal & Subbian 2014). According to Manning et al. (1999), algorithms for partitioning a network into clusters fall under *hierarchical* or *non-hierarchical*. This section focuses on the classification given in Newman (2004a) because it reflects the approach taken in the thesis.

- *Graph partitioning*² entails dividing a network into predefined groups of nodes. It is suitable for applications in which the required number and size of the communities are known³. The *graph partitioning* can be achieved using an *iterative bisection*, which begins with the division of the network into the best two-parts, which are further subdivided until the required division is identified. The ultimate goal in the approach is to reach the best split point (Newman 2006). One of the drawbacks of the bisection approach is the need to preset the maximum number of bisections, which may affect performance. With respect to text, a *graph-based clustering* involves a series of activities, such as transformation of a corpus into a *node-node similarity graph* or a *node-word occurrence graph*, followed by an application of a decomposition technique such as *non-negative matrix factorisation* (Aggarwal & Subbian 2014). Metrics such as *betweenness* or *shortest loop edges*, are central to the operation of algorithms that process graphs to detect relevant groups (Pothen et al. 1990).
- *Block or hierarchical modelling* tends to employ a different technical approach from the *graph-based approach*. For instance, from sociological point of view, the analyses and interpretation of communities in a network are based on the *hierarchical clustering*, which used a similarity measure⁴ to distinguish between pairs in the network (Scott 1988). The *hierarchical clustering* proceeds by assuming that the network splits into subgroups naturally, and the goal is to discover such division using techniques that rely on a *similarity function*. For example, using a function that computes the distance between points is used to iteratively assign nodes to clusters or partitions in a deterministic way. Moreover, it is common in hierarchical clustering to iteratively aggregate similar clusters into larger ones (Aggarwal & Subbian 2014).
- Furthermore, categorisation based on *generative* or *model-based* and *discriminative* or *similarity-based* is used in the literature (Berkhin 2006). *Model-based clustering algorithms* are a form of *Expectation Maximisation (EM)* that aim to learn a generative model from each document or data segment where each model represents a cluster. Models

²This is widely used within the computing domain

³e.g. in parallelisation of computing processors

⁴Examples include *Euclidean distance*, *Pearson correlation*, or based on *vertex or edge count* – independent paths between vertices

based on the *EM* estimate the maximum likelihood of data-points to belong to a cluster and is suitable where there is incomplete data. For instance, the *Latent Dirichlet Allocation (LDA)* is a form of *generative model* (Blei et al. 2003, Balasubramanyan & Cohen 2011, Yan et al. 2013). On the other hand, *similarity-based clustering algorithms* are based on optimising a *scoring function* in which the pairwise similarity between data-points is computed and optimised for clustering. This form of clustering follows the hierarchical agglomerative clustering (or the *block modelling*).

The distinguishing factor between the *group partitioning method* and the *detection method* is that the former has a predefined *number* of and *size* of the communities, and the latter is being decided by the network, not by the experimenter (Newman 2006). Also, there is no hard and fast rule to say that the network must have the best split; that depends on the network topology, which resembles more of a real-world network. A core requirement for a community detection algorithm is the ability to naturally detect divisions among vertices without external influence or imposing restrictions on the divisions (Newman & Girvan 2004). The problem being addressed in this thesis is the detection of local communities, which incorporates aspects of *graph* and *block* modelling in a bi-modal manner to achieve the goal of identifying relevant divisions with less or no external interference (see details in Chapter VII).

3.4 Clustering and Community Detection

A network is an interconnection of heterogeneous vertices or nodes in which a connection is established via edges or links, making it possible to connect various nodes. Each node in a network tends to incline or display an affinity towards a given group or community. A community evolves when interactions among subsets of vertices within the network are dense and infrequent with other vertices (Newman & Girvan 2004). Naturally, humans are endowed with the capability of abstracting many complex phenomena; we perform various forms of clustering effortlessly without paying much attention. However, to automate the process at an appreciable level of efficacy is a daunting task. One of the reasons is that in most applications, the clustering data is multidimensional, hence patterns are not easily recognisable (Bishop 2006). As a result, various clustering algorithms are being developed to handle multidimensional data on a large scale across various domains. The following section reviews relevant studies that employ

clustering and *community detection* techniques, which are often used interchangeably in various contexts. The literature is rather vague in distinguishing between *clustering* and *community detection*. A *clustering task* tends to focus on a single modality, e.g. using the attributes of nodes as the basis for grouping network objects while a *community detection task* focuses on detecting communities based on the network structure as a function of connectivity strength.

3.4.1 Clustering-related Tasks

The objective in this section is to analyse relevant methods in the literature, which are geared toward clustering and the detection of *socially cohesive communities* on Twitter. Essentially, the survey focuses on the following related areas: *clustering*, *topic/event detection* and *information diffusion*. The relevant algorithms are thoroughly described, and their relevance in tackling the thesis's problem is discussed. Thus, the strengths and weaknesses of each algorithm concerning the problem are described.

Clustering tasks: *Clustering tasks* typically proceed without prior knowledge of the structure in the data by exploring many options to uncover meaningful patterns inherent in the data. The activity can be modelled as *graph-based* or *similarity-based*. As a form of dimensionality reduction technique, clustering entails the partitioning of an extensive collection of objects into groups of related objects in an unsupervised learning paradigm. Under this paradigm, the grouping of similar items is achieved using a *domain-specific similarity function*, in which the goal is to *maximise in-group similarity*. Graph-based clustering methods often proceed with a preconceived notion of the existence of a community structure in the data. The approach may involve *hierarchical agglomerative* clustering of objects (Pons & Latapy 2006) according to a *random walk model* (Erdős & Rényi 1960) or based on *modularity* (Newman 2004b) optimisation, such as in the *Louvain* detection algorithm (Blondel et al. 2008) or its variants (Kim et al. 2013). The graph-based approach involves detecting subsets in a graph exhibiting *dense intra-cluster* and *sparse inter-cluster* connectivity (Newman 2002). Metrics such as *betweenness* or *shortest loop edges*, are central to the operation of algorithms that process graphs to detect relevant groups (Pothén et al. 1990).

Modularity is a popular method to detect groups in a network in which the true community structure corresponds to an interesting statistical arrangement of edges which can be quantified

using the *modularity score*. *Modularity* is defined as the number of edges falling within groups (up to a multiplicative constant) minus the expected number in an equivalent network with edges placed at random (Newman 2004b). A value of modularity $Q > 0$ signifies the possible presence of community structure in a network, and this value enables the experimenter to search for relevant communities with $Q > 0$ or higher positive values denoting positive modularity (Newman 2006). Pons & Latapy (2006) developed an approach for hierarchical agglomerative clustering of objects based on a *random walk model* (Erdős & Rényi 1960), while the *Louvain* detection algorithm (Blondel et al. 2008) is based on optimising *modularity score*; both the methods assume the existence of community structures in the data. A variant of the *Louvain algorithm* was utilised to cluster a static collection of tweets in Kim et al. (2013) and Tsur et al. (2013) proposed a clustering technique that relies on message tags, such as hashtags, on Twitter.

Multiview and bi-modal clustering: *Multiview clustering* is suitable where attributes (of nodes in the network) can be split into two independent groups, and each can be used independently for learning. The possibility of using multiple independent sources of data motivates *multiview clustering*, which relies on data capable of being split into two independent sub-features or attributes. For instance, a *web page* can be described by the words that appear on the page and the underlying links or *urls* directed at the page from other external sources or pages (Bickel & Scheffer 2004, Chao et al. 2017). In line with this, many approaches dealing with various forms of data clustering improvement have been proposed in the past. The work of Bickel & Scheffer (2004) examined how the utilisation of multiview clustering outperforms its single view counterpart using algorithms based on *K-means* and *Expectation Maximisation*. The *multilevel clustering technique (MCT)* proposed in this study is similar to *multiview clustering* in which different features are utilised to improve clustering problems (Chaudhuri et al. 2009, Liu et al. 2013). Noting that communities on Twitter could be formed based on many factors as described in Figure 1.2, the *MCT framework* is a two-stage clustering technique that recognises different modalities at various levels – *structural* and *textual* – using various independent features as information sources for clustering.

In *bi-modal clustering*, network structure and nodes' attributes are the two primary modalities or sources of information upon which communities of nodes are formed. Until recently, studies mostly focus on a single modality. Early work in this line of research can be found

in (Ester et al. 2006, Zhou et al. 2009, Balasubramanyan & Cohen 2011, Leskovec & McAuley 2012, Yang et al. 2013). The work of Ester et al. (2006) proposed a *connected k-centre* approach that employs both structural and attributes information of a given partition in the network. The approach was shown to be NP-hard problem, leading to many heuristics, hence affecting performance. Similarly, the work of Zhou et al. (2009) proposed an approach for community detection (*SA-cluster*), that combines structural and attributes' similarities, which is based on partitioning a network into cohesive k-clusters according to the structural and attribute information. Structural and nodes' attributes information is used to compute a distance metric, which estimates the pairwise similarity or closeness between a pair of vertices. This is followed by applying a random walk model to explore the node's neighbourhood to identify relevant or similar nodes for clustering. A closely related study to the thesis's approach can be found in (Yang et al. 2013), in which a generative model for networks with node attributes is proposed. However, the depth of the features, especially the nodes attributes, is shallow and the node attribute (*hashtag*) is insufficient in analysing the depth of similarity between network entities in a complex environment like Twitter. With reference to Twitter, the structural component is not fully captured because it relies on a directed form of connections and is devoid of the necessary attributes to afford an in-depth structural insight.

In addition to using a *bi-modal information source*, there is a growing interest in the detection of communities in networks with edge uncertainty or incomplete information. Conventional methods such as *normalised cut* (Shi & Malik 2000) and modularity (Newman 2006) are based on the topological structure of the network. However, many networks, such as *terrorist network* or a *food web*, come with incomplete information (Lin et al. 2012). Thus, inferring link information in incomplete networks is challenging because the information is usually localised within a small group with linkage information. To have a broader picture of the network by accounting for the missing links, the work of Lin et al. (2012) leverages the *complete information* available in the data to learn a generalisable distance metric that will help to estimate the missing information in the network. The problem with the approach in Lin et al. (2012) is that it is centred around the structural aspect and does not account for the breadth required in *textual* aspect in networks such as Twitter.

Topic Detection

In Becker et al. (2011), tweets are clustered by distinguishing posts related to real-world events from posts related to non-events. A method for automatic clustering and classification of tweets into sub-categories based on discussion hashtags was proposed in Rosa et al. (2011). Vicient & Moreno (2014) observed that those approaches detect high-level communities of users, which exhibit limited conformity between hashtags and associated sets of tweets. Although hashtags can be used as the representative of the topics being discussed on Twitter (Dann 2010), they are quite limited in exposing them fully; this is because users may use keywords relevant to the trending discussions, but the content may not reflect the keyword and ultimately lead to communities of unrelated users. The sparsification technique proposed in Karthick et al. (2014) relies on influential users as the basis for community formation, which leads to identifying followers as against cohesive communities.

Topics on Twitter are associated with varying degrees of extent and intensity, with exciting topics exhibiting high activity burstiness (Kleinberg 2002). Ruchi & Kamalakar (2013) proposed an event detection method that relies on keywords used in defining events on Twitter. A study by Gadek et al. (2017) focuses on understanding the relationship between users and the topic of discussion to detect communities on Twitter. The study, however, focuses on a single topic to be discussed by each group at high-level. This is usually not the case since users participate in several discussions on Twitter. Similarly, the work of Feng et al. (2015) relies on *hashtagged* messages for clustering activity; because tweets not associated with the hashtag are ignored, the method will ultimately lead to less cohesive communities since contextual properties are not fully represented. The work of Cataldi et al. (2010), Lu & Lee (2015) and Feng et al. (2015) focus on identifying the role of a temporal quantity in clustering dynamic network data. Similar to community detection based on trending topics, Cataldi et al. (2010) focuses on the popularity of terms over time, which makes the approach limited in capturing the full spectrum of communities, especially when terms are becoming less popular. During its life-cycle on Twitter, the popularity of an average hashtag is uneven – exhibits peaks and drops at various points in time; this property motivated the development of *Topic-over-Time Mixed Membership Model (TOT-MMM)*, a system that captures the temporal clustering effect of latent topics in tweets (Lu & Lee 2015). Lagnier et al. (2013) investigate how information diffuse within communities by

leveraging interaction dynamism, the generated contents and profile information.

The surveyed studies primarily focus on either high-level meta-data (such as hashtag) or static trending content with a large number of users with less connectivity. Such an approach will conceal crucial segments in the network with no subscription to trending topics (Guille & Favre 2015). With an average of 6000 posts per second on Twitter, it is highly likely that low trending topics or events will be buried, and communities of such users will go undetected. The goal here is to avoid these sorts of limitations and aspire for low-level clustering of users.

3.5 Summary

Almost all networks in nature are characterised by a certain level or degrees of organisations in which groups of nodes form tightly connected units called communities. *Sub-networks* or *communities* represent functional entities which reflect the topological relationships between elements of the underlying system or network (Newman 2006). Communities can range from the structure of microscopic organisms to complex networks, such as the internet (Erdős & Rényi 1960, Watts & Strogatz 1998, Scott 1988, Albert & Barabási 2002). Because different attributes and features define how a given dataset may be clustered, various clustering and community detection algorithms have been proposed. The formalisation of a community detection task requires a scoring function that quantifies how much the connectivity pattern of a given set of nodes resembles the connectivity structure of a network community (Yang & Leskovec 2015). The content in this chapter is posed to put the research in perspectives, and the next chapter (Chapter IV) gives a formal description of the thesis problem.

Chapter IV

MICROCOSM DETECTION PROBLEM

4.1 Introduction

In the context of Twitter, communities could be formed based on many factors as described in Figure 1.2 and the research is concerned with identifying sub-networks with strong social cohesion. Thus, the problem is formulated to focus on smaller groups with high degrees of *structural* and *textual* similarities. This chapter discusses and formulates the problem of identifying such communities, relevant definitions and notations.

4.2 Community Structure in a Network

A *community structure* in a network is defined as the existence of densely connected groups of nodes who have sparser connections with other communities in the network (Newman 2004b). Such densely connected nodes constitute a *small group*, which results in a homogeneous community displaying sociodemographic behavioural, and intrapersonal similarities (Miller McPherson et al. 2001). In Section 2.3.2 (Chapter II), *sociometry* is defined as a means to measure social relationships between people (Wasserman & Faust 1994), and Figure 4.1 shows a classification of *social groups* and the corresponding degrees of cohesiveness, in which the smaller the size, the more cohesive. Also, according to *Block-modelling* (see Section 3.3.1), nodes can be grouped based on the extent to which they are equivalent using a set of empirical procedures. Putting these insights in the context of Twitter, a comparison of *social groups*, which defines *social cohesion* as a function of *group size*, the notion of *socially cohesive groups* seems far-fetched. This is ascribed to the different forms of connections on Twitter (Figure 1.2) where most of the relationships tend to be outside the *active zone* in Figure 4.1. To achieve the objec-

tive of identifying *microcosms*, the problem of community detection is formulated to focus on smaller groups with a high degree of *structural* and *textual* similarities (see Figure 4.3).

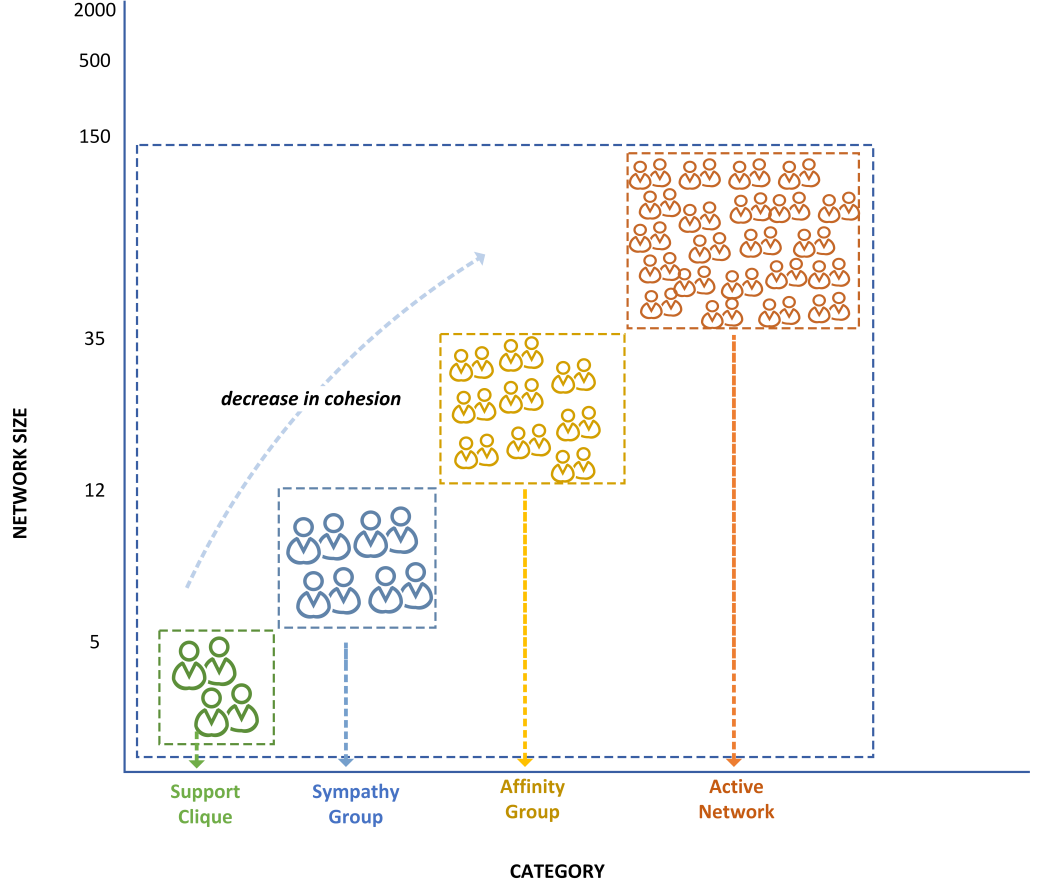


Figure 4.1: Classification of social groups and corresponding degrees of cohesiveness as a function of group size – the smaller the size, the stronger the cohesion. In the context of Twitter, most of the connections can be regarded as outside the active zone, hence leading to an increase in irrelevant content and less cohesive groups.

4.2.1 Role of a scoring function

The partitioning of a network into groups of related objects is conducted according to an unsupervised learning paradigm using a *domain-specific scoring function*. The role of a scoring function is to identify pairs of objects which are closely related in some respect, hence the choice of an effective similarity measure is crucial (see Section 3.2.2 for details). In the research context, a *joint similarity* comprising of both *structural* and *textual* modalities is used to account for *global* and *local* information. As illustrated earlier in Figure 1.1, both *structural* and *textual* components of a node are used to define a *joint similarity* in the *MCT framework*

(Section 4.3). The *structural similarity* involves predicting the likelihood of a reciprocity between pairs of nodes (a form of a *state-type tie*) and the *textual similarity* is based on computing similarity of sets of texts between pairs of nodes (a form of an *event-type tie*). According to the degrees of similarities in each component, any pairs of network entities can be placed in the relevant communities; see the illustrative example of Figure 4.2.

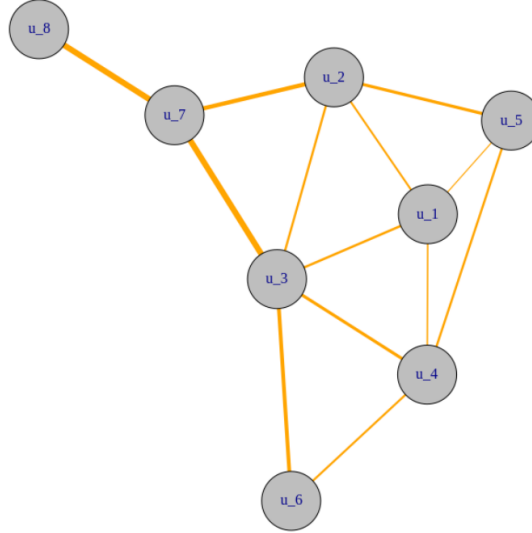


Figure 4.2: A hypothetical example of possible *cluster bands*, which are expressed as a function of degree of similarities. For instance, if the maximum similarity is 1.0, a value in the range of: (a) 0 – 0.2 denotes *no relation* (b) 0.25 – 0.4 signifies a similarity in textual aspect or *C-band* (c) 0.45 – 0.65 signifies a similarity in structural aspect or *B-band* and (d) 0.7 – 1.0 denotes similarity in both textual and structural aspects or *A-band*. The edges in the figure signifies the magnitude of similarity between nodes in which the thicker the edge, the more cohesive.

In accordance with the degrees of similarities in both the *structural* and *content* aspects, any pair of nodes can be placed in different sets or *bands of communities* classified as *A-band*, *B-band* and *C-band* (see Figure 4.2). Under this process, the degree of cohesiveness varies across the communities, where *A-band cluster* is the most cohesive and *C-band cluster* is the least cohesive. The members of *A-band cluster* exhibit the highest degrees of similarities, which exceed a pre-defined threshold in both the structural and textual components. The members in *B-band cluster* exhibit a higher degree of *structural similarity* (a form of *state-type tie*), but a low or no similarity in the *textual* component. Finally, the members in *C-band cluster* exhibit high similarities in the *textual* component but low or no structural similarity. Figure 1.1 provides a visual illustration of the process and Figure 4.2 shows memberships of various clusters based

on the strength of the two measures among nodes.

4.3 The MCT Framework

Noting that communities on Twitter could be formed based on many factors as described in Figure 1.2, the *MCT framework* is a two-stage clustering technique that recognises different modalities as information sources for clustering. Fundamentally, the framework (see Figure 4.3) consists of the following:

- i The first stage in the framework deals with the *structural aspect* and begins with a description of factors that influence reciprocal ties, which could ultimately result in a local community. The *structural similarity* involves predicting the likelihood of a reciprocal tie between pairs of nodes. Inspired by social homophily, which promotes group formation, the goal is to identify as much as possible a set of nodes with structural similarity. On this basis, a prediction model that returns the likelihood of similarity between nodes, is proposed; see Section 7.2 and Section 7.2.2 for more details.
- ii The second stage is concerned with *textual analysis* of content from *structurally-related nodes* in the network. In this stage, the *textual similarity* captures relatedness in the *texts* produced by *structurally-related nodes*. In response to the limitations of a single tweet in conveying sufficient information about a topic, for each node in the collection of structurally-related nodes, a finite number of texts, making a corpus, is used to compare the similarities between the topics of discussions of the nodes within a given time-frame. Ultimately, the goal is to understand the specific topics being discussed and identify relevant clusters.

In Figure 1.1, nodes in communities under the structural component (a form of a *state-type tie*) are related based on reciprocal ties (which are rare on Twitter) and the community is more cohesive than the community of users based on *textual similarity* (a form of an *event-type tie*). A more cohesive community is the one that recognises both *structural* and *textual* similarities (Figure 1.1(c)). These two aspects are being explored to enable detection of *microcosms*. Figure 4.3 depicts a high-level illustration of the basic steps in the *MCT framework*.

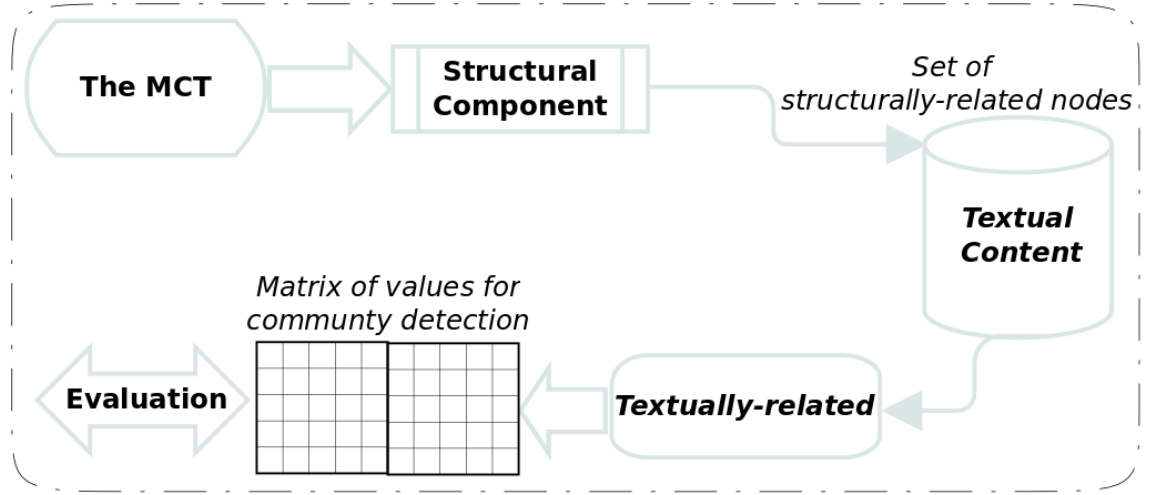


Figure 4.3: A high-level depiction of the execution pipeline in the MCT framework. As the first activity in the pipeline, the *structural component* is based on a collection of nodes with structural similarities, which promotes group formation among nodes. The second stage is concerned with *textual analysis* of content from *structurally-related nodes*, to identify groups of nodes according to discussion topics. The idea is to enable the combination of both *state-type* and *event-type* ties to improve clustering.

4.3.1 Structurally-Related Nodes

Research findings discussing *social networks* suggest that individuals compare themselves with one another and adopt similar attitudes and behaviours of, those who occupy an equivalent position (Brass et al. 1998). The meaning of *structural equivalence*, two actors having similar interaction partners, can be mapped to a *state-type tie* to infer structural similarity using the node’s attributes. The *structural similarity* component in Figure 4.3 is inspired by the idea of *homophily* and the need to simplify the challenging task of identifying reciprocal nodes on Twitter. Figure 4.4 shows relevant features to leverage toward analysing *homophily* to identify *structurally-related nodes* \mathcal{S}_r . For each node v_i in \mathcal{S}_r , it is possible to identify at least a node v_j , which is structurally similar to v_i . It follows that $\forall v_i \in \mathcal{S}_r \exists v_j : p(R_{v_i, v_j}) \geq \tau$, i.e. \mathcal{S}_r is a collection of nodes with high degrees of *structural similarity*.

Spectral Clustering

Spectral clustering involves a series of operations, ranging from the construction of adjacency or affinity matrices to clustering in a reduced dimension (Han et al. 2011). Because the structural similarities can be easily transformed into a graph structure based on the likelihood of reci-

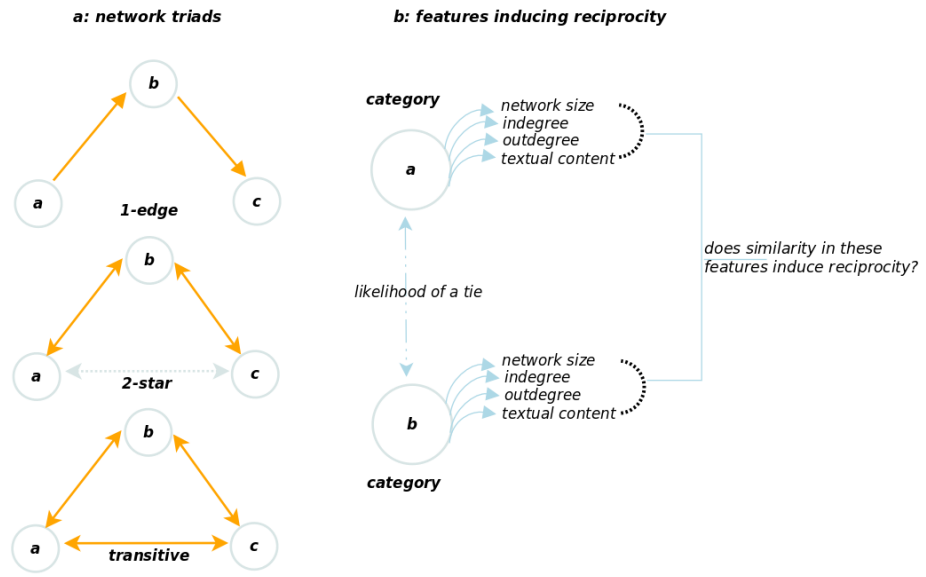


Figure 4.4: (a) Possible social ties in a network triad. The network composition of each node in the example consists of a set of nodes with a directed or reciprocal ties to the node. In *sub-figure (b)*, a and b denote nodes and their corresponding network compositions given by N_a and N_b , e.g. $a_1, a_2, a_3, \dots, a_n \in N_a$. Thus, a reciprocal tie is established if $\exists a_i \in N_a : a \in N_{a_i}$, where N_{a_i} denotes the network composition of node a_i . (b) An example of a dyad and the corresponding features that are responsible for tie formation between nodes on Twitter. For each of these attributes, a probability score is assigned to discover the inter-dependencies between the features in enabling reciprocal ties.

procuity, spectral clustering can be applied to identify relevant clusters resulting from *structural equivalence*. The detected clusters will enable the analysis of various metrics for sociometry (see Section 2.3.2). The essential steps in the *spectral clustering* include the following (see Section 7.2.3 for details).

- i *Adjacency and Degree Matrices*: The *adjacency matrix* M_A is constructed from the structural similarities among pairs of users (represented by the *row* and *column* indices) and the entries in the matrix represents the presence or absence of reciprocity (see Eq. 7.5). The *degree matrix* M_D , is a diagonal matrix obtained by summing the entries in the adjacency matrix M_A across the rows, in which the entry i, i denotes the degree of each node. In essence, the degree matrix represents the number of edges each node is connected to and is a crucial requirement in the *spectral clustering*. Each value in the diagonal of the matrix is given by Eq. 7.6.
- ii *Graph Laplacian and Clustering*: This step involves the construction of a *Laplacian matrix* and the associated *eigenvectors* and *eigenvalues*. The *Laplacian matrix* M_L , is obtained by subtracting the *adjacency matrix* M_A , from the *degree matrix* M_D , (see Eq. 7.7). Given the matrix M_L , if there exists a vector x and a scalar quantity λ , such that $M_L x = \lambda x$, then x is an *eigenvector* and λ the *eigenvalue* of M_L . Because of the special property of *eigenvectors*, i.e. they remain unchanged after undergoing matrix transformation, they have a wide range of applications. In the spectral clustering, the *eigenvectors* of the *Laplacian matrix* represent the defining features for vertices in the data and are employed for clustering. The clustering can be done using classic clustering algorithms, e.g. *k-means*, or a custom algorithm to detect local structures in the network.

The next phase after identifying *structurally-related nodes* in the *MCT framework* is to retrieve sets of relevant textual content from each node. In addition to detecting communities, the *textual content* will enable a means of studying *constructionism*, a sociological premise that people who share knowledge are more likely to interact (Carley 1991), hence facilitate the formation of social ties on Twitter.

4.3.2 Textually-Related Nodes

Textual similarity captures relatedness in the *texts* produced by *structurally-related nodes* \mathcal{S}_r . The task of identifying *textually-related nodes* \mathcal{T}_r starts by aggregating a finite collection of *textual content* \mathcal{T} , from each node v_i described by the following set of features, $\{set\ of\ texts, network\ features, auxiliary\ features\}$, to compute similarity. Given a stream of texts $t_1, t_2, t_3, \dots, t_k \in \mathcal{T}$, each $t_i \in \mathcal{T}$ consists of *n-gram features*¹ given by:

$$f_{i1}, f_{i2}, f_{i3}, \dots, f_{im} \in t_i \in \mathcal{T}$$

Then, for each node v_i in the structurally-related nodes \mathcal{S}_r , a finite number k , of texts making a corpus, is considered:

$$\mathcal{T}_{v_i} = \{t_{i,1}, t_{i,2}, t_{i,3}, \dots, t_{i,k}\}$$

Each tweet is preprocessed to extract shingles² for transformation based on *term-frequency-inverse document frequency (tf-idf)* scheme. Each of the *tf-idf* vector \mathbf{v}_i , can be left unnormalised or normalised. This thesis utilises the $L_2 - norm$ normalisation given by:

$$\mathbf{v}_i = \frac{1}{n} \sum_{i=1}^n \frac{\mathbf{v}_i}{\|\mathbf{v}_i\|_2}$$

The shingle, used as the unit of the analysis for the text, can also be represented by an embedding vector, which is useful in capturing a greater semantic in the texts. The aggregated texts are analysed for clustering and assessment of other relevant metrics. Through aggregating *textual content*, each node has a unique fingerprint, which can be used for making an in-depth comparison. To achieve this aim, the *similarity* between pair of corpora $\mathcal{T}_{v_i}, \mathcal{T}_{v_j}$, is based on training a *topic model* on various corpora such that each corpus or document (the set of k tweets from each node) will have finite distributions (over all topics, and all topics will have distribution over all words). Each document (or *anchor document*) is compared with other documents in the corpus and the most similar documents to the *anchor document* are returned. Algorithm 3 describes how to obtain the *textually-related clusters*.

¹ n could be any positive integer, e.g 1, 2, 3 for *unigram*, *bigram* and *trigram* respectively.

²*shingles* are obtained by removing stopwords and other non-content bearing terms in a text

4.4 Definition and Notation

The following section presents the formal definitions of relevant concepts and terms that are central to the thesis. More definitions and notations will be introduced at the appropriate sections of the thesis. Table 4.1 provides a summary of all relevant notations used in the study.

4.4.1 Definition

Network data: A network data \mathcal{D} , consists of a set of vertices or nodes $v_1, v_2, \dots, v_m \in \mathcal{V}$, and a set of relations or edges $e_1, e_2, \dots, e_n \in \mathcal{E}$. Each node in the data is described according to its structural and textual features.

Indegree, Outdegree, Category: These are the attributes utilised in identifying structurally-related nodes. The *indegree* (*ind*) corresponds to the number of *followers* of a user; the *outdegree* (*out*) corresponds to the number *friends* or *followings* of a user; and the *category* (*cat*) denotes whether the account holder is *verified* or *unverified*. Account verification is done by Twitter to ascertain who the account holder claims to be.

Relational pairs: A relation \succ , between pair of nodes v_i, v_j , over a set of network data \mathcal{D} , is denoted by:

$$v_i \succ v_j \quad \forall v_i, v_j \in \mathcal{D}$$

Dyadic tie: A relation \succ between a pair of nodes $v_i, v_j \in \mathcal{D}$ is *dyadic*³ if v_i follows v_j and vice versa:

$$\text{iff} \quad v_i \succ v_j \text{ is true,} \quad \forall v_i, v_j \in \mathcal{D}$$

In the context of this thesis, v_i follows v_j is a directed relationship; if v_j follows v_i back, then it is undirected, which is referred to as a *dyad* (see Figure 4.5).

Transitive or Simmelian tie: Transitivity, a social preference to be friends with a *friend-of-a-friend*, is a common feature of a network (Watts & Strogatz 1998). The concept of a *transitive tie*⁴ is similar to a *Simmelian tie* (Simmel et al. 1950), which is referred to as a strong social

³Dyadic tie, pairwise or binary relations are used interchangeable in the research.

⁴transitive and Simmelian ties are synonymous in this study

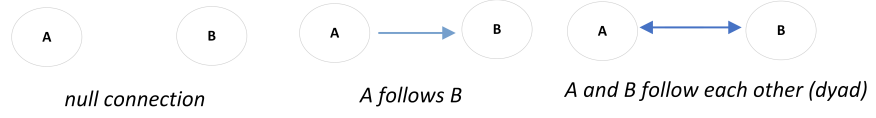


Figure 4.5: Examples of relationships between a pair of nodes A and B on Twitter showing: no relationship (*null connection*), directed relationship ($A \longrightarrow B$) and dyadic or pairwise relationship ($A \longleftrightarrow B$).

relationship within *three-person groups* or more. A binary relation \succ , over a set \mathcal{D} is *transitive*:

$$\text{iff} \quad v_i \succ v_j \text{ and } v_j \succ v_k \text{ then } v_i \succ v_k \quad \forall v_i, v_j, v_k \in \mathcal{D}$$

Structural similarity: Any pairs of nodes v_i and v_j are *structurally-similar* or *-related*, if their degree of reciprocity $p(R_{v_i, v_j})$, is greater than a predefined threshold τ . It follows that:

$$\mathcal{S}_r : \forall v_i \in \mathcal{S}_r \exists v_j \text{ such that } p(R_{v_i, v_j}) \geq \tau \quad \mathcal{S}_r \subset \mathcal{D}$$

See Figure 4.4 for examples of various forms of social ties.

Textual similarity: Any pairs of texts or textual documents t_i and t_j are *textually-similar* or *-related* \mathcal{T}_r , if their degree of similarity ϕ , is greater than a predefined threshold τ . It follows that:

$$\mathcal{T}_r : \forall t_i \in \mathcal{T}_r \exists t_j \text{ such that } \phi(t_i, t_j) \geq \tau \quad \mathcal{T}_r \in \mathcal{S}_r$$

A finite collection of texts is extracted from each node in the structurally-related nodes \mathcal{S}_r , thus, each $t_i \in \mathcal{T}_r$ consists of the node and the corresponding set of texts.

Cohesive community and microcosm: A *cohesive community* is a collection of nodes \mathcal{V} with high degrees of similarities in both *structural* and *textual* aspects (see Figure 1.1). Thus, the problem of *microcosm detection* can be formally expressed as:

given a collection of network data \mathcal{D} , define by a set of nodes \mathcal{V} and a set of edges \mathcal{E} , for each node $v_i \in \mathcal{V}$ consisting of sets of structural and textual features, the goal is to identify a collection of sub-networks \mathcal{P} , with high similarities given by:

$$\mathcal{P} : \forall v_i \in \mathcal{S}_r \exists v_j \text{ such that } p(R_{v_i, v_j}) \geq \tau \text{ and } \forall t_i \in \mathcal{T}_r \exists t_j \text{ such that } \phi(t_i, t_j) \geq \tau \quad \mathcal{P} \subset \mathcal{D}$$

The above formulation means that for all pair of nodes in the partition \mathcal{P} , both the *structural* and *textual* similarities are greater than their respective threshold τ . For all relevant experiments in

the thesis, $\tau \geq 0.5$, i.e. the pairs are considered similar (1) if $\tau \geq 0.5$, otherwise dissimilar (0). Once an appropriate clustering criterion has been identified, an optimisation procedure, which aims to find a function ψ that assigns nodes to relevant clusters, is defined. See Chapter VII for details.

4.4.2 Notation

This section presents some relevant notations that will keep reoccurring throughout the thesis. Table 4.1 provides a list of the notations and their corresponding descriptions. Uppercase letters denote random variables and lowercase letters are used for scalar values or functions of random values. Matrices are denoted by bold capital letters and vectors by bold small letters.

Training features: The *structural similarity* is based on predicting the likelihood of reciprocity between pair of nodes using a set of training features. The set of all possible features \mathcal{A}_f , that can be extracted from a node's profile is exemplified in Figure 4.4. A subset of the features $\mathcal{X}_f \subset \mathcal{A}_f$, for making the comparison consists of easily accessible features, see Figure 4.4, that enables a user to make a quick decision about reciprocity is given by:

$$\mathcal{X}_f = \{ind, out, cat\} \subset \mathcal{A}_f$$

Thus, for a given pair of nodes v_i, v_j , their corresponding features are denoted by:

$$\mathcal{X}_{f_{v_i}} = \{ind_{v_i}, out_{v_i}, cat_{v_i}\}; \quad \mathcal{X}_{f_{v_j}} = \{ind_{v_j}, out_{v_j}, cat_{v_j}\}$$

When using a deep learning classifier, the training examples are denoted by \mathcal{X} and target labels are denoted by γ . The input \mathcal{X} , consists of both *structural* and *textual* features and a training instance is given by:

$$\{x_i, y_i\}_{i=0}^n \in \mathbb{R}$$

A potential reciprocal relationship is extracted from $R \subseteq \mathcal{V} \times \mathcal{V}$, such that $\mathcal{D} = (\mathcal{V}, R)$, defines a network of nodes with implicit *structural-similarity* where relevant *communities* or *partitions* \mathcal{P} are expected to exist.

Similarity matrices: \mathcal{S}_f and \mathcal{T}_f denote sets of features of *structurally-related* and *textually-related* nodes, respectively. Accordingly, $\mathcal{M}_{\mathcal{S}_f}^{m \times m}$ defines an *adjacency matrix* among pairs of

nodes based on the *structural similarities* and $\mathcal{M}_{T_f}^{n \times n}$ defines an *affinity matrix* based on the *textual similarities* among nodes. Thus, for each matrix constructed based on $\mathcal{S}_f, \mathcal{T}_f$, there exist sets of communities $(\mathcal{K}, \mathcal{Q})$, such that $k_1, k_2, k_3, \dots, k_m \in \mathcal{K}$ consists of possible communities that can be identified in \mathcal{M}_{S_f} . Similarly, $q_1, q_2, q_3, \dots, q_n \in \mathcal{Q}$ consists of possible communities that can be identified in \mathcal{M}_{T_f} . A concise representation is given by the following tuple $(\mathcal{V}, \mathcal{S}_f, \mathcal{T}_f \in \mathcal{D})$, consisting of a set of nodes and associated features for clustering. It follows that $\mathcal{S}_f \in \mathbb{R}^{m \times m}, \mathcal{K} \in \mathbb{R}^{m \times k}$ and $\mathcal{T}_f \in \mathbb{R}^{n \times n}, \mathcal{Q} \in \mathbb{R}^{n \times q}$.

Community of related nodes: The members of a community are similar in some respect, i.e. based on the clustering *scoring function* (Section 3.2.2). For instance, a pair of nodes is similar:

$$v_i \sim v_j \iff \exists c_i \in \mathcal{C} : v_i, v_j \in c_i$$

where \mathcal{C} denotes a set of communities. Given the matrix of reciprocal relationships (adjacency matrix) $R \subseteq \mathcal{V} \times \mathcal{V}$ and the network data $\mathcal{D} (\mathcal{V}, R \in \mathcal{D})$, to find communities $c_1, c_2, \dots, c_k \in \mathcal{C}$, in which:

$$\emptyset \subset c_i \subseteq \mathcal{V}$$

A more socially cohesive community of nodes (*microcosm*) can be formed by identifying overlapping nodes in both \mathcal{K} and \mathcal{Q} . See Figure 1.1 for a visual illustration.

4.5 Summary

This chapter offered a formal description of the thesis's problem, including notations and definitions of terms that will be used throughout the work. For ease of referencing, Table 4.1 summarises the notations used in the study. The next chapter (Chapter V) focuses on the research data and its suitability in answering the research questions.

Table 4.1: Relevant notations and their corresponding descriptions.

Notation	Description
D and \hat{D}	observed and synthetic data
θ	vector of unknown parameters, e.g. μ and σ
$p(\theta)$	prior distribution
$p(D \theta)$	likelihood function
$p(\theta D)$	posterior distribution
β_{ui}	mean reciprocity among users
γc_{ui}	mean reciprocity between users' categories
ϵ_{ui}	error term in the proposed model
χ	set of features inducing reciprocity
$a \succ b$	a binary relation between a and b
κ	set of reciprocal ties
$v_i \sim v_j$	a pair of similar nodes v_i and v_j
\approx	approximately equal
$P(X = x)$	probability that a random variable X takes on the value x
$X \sim p(\cdot)$	random variable X selected from distribution $p(x) = p(X = x)$
$E[X]$	Expectation of a random variable X s.t. $E[X] = \sum_x p(x)x$
$\operatorname{argmax}_a f(a)$	value f a at which $f(a)$ takes its maximum value
$\exp x$ or e^x	base of a natural logarithm, \ln , $\exp(\ln x) = x$; $e \approx 2.71828$
\mathbb{Z}^+	set of positive integers
\mathcal{R}	set of real numbers or values
d_i, p_i, q_i, u_i	respective i th component of vectors, $\mathbf{D}, \mathbf{P}, \mathbf{Q}, \mathbf{U}$
χ and γ	sets of training examples and target labels respectively
$\{x_i, y_i\}_{i=0}^n \in \mathbb{R}$	a training instance
ϕ	similarity function
τ	a predefined threshold for making comparison, e.g. $\phi \geq 0.5$

Chapter V

AUTHENTICATION OF ONLINE CONTENT

5.1 Introduction

The utility offered by platforms such as Facebook and Twitter is instrumental in enabling global connectivity. There are an estimated 2.46 billion connected users, and one-third of the global population will be connected by 2020¹. The users of these platforms freely generate and consume information leading to unprecedented amounts of data. While such data is being exploited for various applications, a substantial amount of it is contributed by spam or fake users. The growing amount of irrelevant *social media content* threatens the credibility of research based on analysing such data. This chapter is motivated by the need to identify and filter out spam content in social media data.

5.2 Online spam and detection methods

The growing volume of spam posts and the use of autonomous accounts (or social bots) to generate posts raises many concerns about the credibility and representativeness of the data for research. It was estimated that on average, one spam post occurs in every 200 social media posts (NexGate 2013), and approximately 15% of active accounts on *Twitter* are automated (Varol et al. 2017). Online spamming activities come in different forms such as malware dissemination, posting of commercial URLs, fake news or abusive contents, automated generation of large volume of contents (Varol et al. 2017) and following or mentioning random users (Lee et al.

¹See www.statista.com/topics/1164/social-networks

2011). Another form of online spamming is the growing use of machine learning models to generate fake reviews on products and services (Yao et al. 2017), and the use of social bots to influence the opinion of users (Subrahmanian et al. 2016). The research posits that one of the reasons for the proliferation of *spam content* in social media is attributed to the lack of physical contact between the communicating parties, which makes it difficult to ascertain the true identity of users. The following section presents a review of relevant methods that are used for spam detection.

5.2.1 Spam detection methods

Approaches for spam detection can usually be classified under the following categories: social graph analysis (Yang et al. 2012, Yu et al. 2008, Danezis & Mittal 2009), text analysis and activity patterns (Gao et al. 2010), analysis of user profile meta-data, URL usage and the effect of URL obfuscation (Thomas et al. 2011, Lee & Kim 2012, Benevenuto et al. 2010), analysis of interaction behaviour (Howard & Kollanyi 2016, Varol et al. 2017, Lee et al. 2011), and URL blacklisting and its effect (Grier et al. 2010).

Until November 2017, users on Twitter are limited to the use of a maximum of 140 characters² to compose a tweet, the use of *Uniform Resource Locators (URLs)* and corresponding shortening services were widespread. Thomas et al. (2011) and Lee & Kim (2012) analysed stream of *URLs* used by spam users and studied how spammers exploit *URLs obfuscation* to redirect users to malicious sites. Grier et al. (2010) analysed a large number of distinct *URLs* pointing to blacklisted sites due to their involvement in scam, phishing and malware activities. Although the approach is effective, it is often slow and fails to detect *URLs* that point to malicious sites not previously blacklisted. Gao et al. (2010) also studied *URL* usage on Facebook to detect spamming activity, and observed how it is limited to compromised accounts rather than accounts created solely for spam activity. Benevenuto et al. (2010) studied the statistical properties of various user accounts and how *URL* shortening services affect spam detection mechanisms. However, the universal use of *URL shortening* by the majority of Twitter users makes it difficult to identify potentially nefarious *URLs* directly on a large scale. In general, the use of *URLs* relies on historical information, limiting the possibilities for real-time detection. Danezis & Mittal (2009) utilised a social network model to infer legitimate user accounts that

²See www.blog.twitter.com/official/

are being controlled by an adversary, and Lee et al. (2011) created *social honeypot accounts* mimicking naive Twitter users to entice spam posting users. Users who fall prey by engaging with the *honeypot accounts* are assumed to violate usage policy and are harvested for further analysis. The harvested users were analysed to distinguish different user types focusing on *link payloads* and features that can capture the dynamics of *follower-following* networks of users. Varol et al. (2017) employed many features related to users, content and the network to develop a system of detecting automated accounts. The use of a large number of features introduces extra overheads to the detection system and some of which may be unavailable for real-time use. Subrahmanian et al. (2016) offered insights into the techniques utilised in identifying *influence bots*, i.e. autonomous entities that are determined to influence discussions on Twitter. Influence bots comprise a category of social bot accounts that seek to assert influence on topical or new discussions, thereby generating unrepresentative or fake data.

The surveyed studies on spam detection mainly rely on historical tweets of a user to extract features, which contribute to an extra overhead in the detection system (Wang et al. 2015). Moreover, spammers evolve rapidly to evade detection systems, hence, rendering some approaches obsolete and ineffective in responding to the new tricks introduced by the spammers. The proposed approach (Section 5.3) in this part of the research relies on readily available features in real-time for better performance and broader applicability.

5.3 Spam-Posts Detection method

The following section presents the proposed method of detecting spam accounts on Twitter, termed *Spam-Posts Detection (SPD)*. Firstly, the *dataset* and corresponding features utilised in the experiment are described. This is followed by a series of experiments and the evaluation process.

5.3.1 Spam detection data

The following datasets have been used in the *spam detection* experiments: *Honeypot dataset* (Lee et al. 2011), automatically annotated SPD dataset (*SPD_{automated}*) and manually annotated dataset (*SPD_{manual}*). The *Honeypot dataset* is available to the public, and is useful for studying spam activity on Twitter. The research leveraged the *honeypot dataset* both as a dataset per se

Table 5.1: Summary of the three different datasets utilised in the *content authentication* research. Each of the data category consists of two classes – legitimate and polluter or spam. The *size of original data* refers to data collected before carrying out relevant preliminary preprocessing such as discarding non-English tweets and duplicates.

Dataset	Size of Original Data	Size of Preprocessed Datasets	Class	Collection	Verified?
<i>Honeypot</i>	19,297	19,276	Legitimate	Automated	No
<i>Honeypot</i>	23,869	22,223	Polluter	Automated	No
<i>SPD_{automated}</i>	10,318	8,515	Legitimate	Automated	Yes
<i>SPD_{automated}</i>	25,568	9,831	Spam	Automated	No
<i>SPD_{manual}</i>	2,000	1,300	Legitimate	Manual	Yes
<i>SPD_{manual}</i>	2,000	700	Spam	Manual	No

and for collecting the *SPD* datasets using a set of keywords. Because keywords play a crucial role in retrieving specific documents from a large corpora (Lott 2012), the study speculates that keywords extracted from the *Honeypot dataset* can be used in retrieving a large quantities of similar data. The last two datasets (*SPD_{automated}* and *SPD_{manual}*) have been collected for the study, and is made publicly available³. Table 5.1 presents relevant statistics about the *SPD* datasets.

5.3.2 SPD data annotation

In Table 5.1, *legitimate* refers to data from genuine users whose accounts have been verified by Twitter. A *verified account* can be ascertained according to the information available from the *meta-section* of a *tweet object* (see Section 2.2.2 in Appendix A for some examples). In contrast to the randomised approaches utilised in Lee et al. (2011) to validate legitimacy of users on Twitter, the *SPD strategy* relies on accounts verified by Twitter in building the set of legitimate users to avoid the potential risk of a high false positive rate. As the name suggests, the *SPD_{manual}* is a manually annotated dataset created to supplement evaluation. It contains tweets randomly selected from the full set of tweets that have been downloaded between *February, 2017* and *June, 2017* via the traditional Twitter *Application Programming Interface (API)*⁴ using relevant keywords as query terms.

The *SPD_{manual}* consists of 1,700 tweets of *legitimate users* and 300 tweets of *spam users*. Due to the imbalance class proportion in the data, the defect is alleviated using a re-sampling technique. An upsampling of the minority class is achieved using the SMOTE technique

³See <https://github.com/ijdutse/spd>

⁴This is a dedicated channel provided by Twitter to enable access to its public datasets.

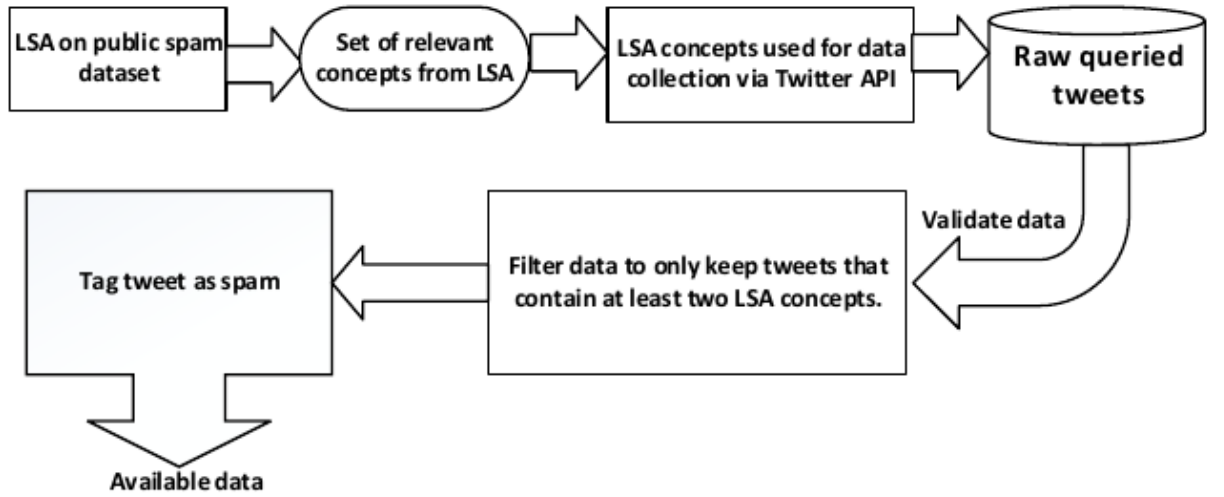


Figure 5.1: A high-level overview of the collection and validation of the spam part of the $SPD_{automated}$ dataset.

(Chawla et al. 2002). Similarly, the $SPD_{automated}$ contains tweets that have been collected between February, 2017 and June, 2017, and have been automatically annotated as *legitimate* or *spam* according to preset criteria. Firstly, tweets posted by users whose accounts have been verified by Twitter were marked as legitimate. The annotation of the *spam* content is more challenging and is based on the following. The initial set of keywords used in retrieving the research data is based on the *honeypot dataset*. Keywords, both for querying Twitter and validating spam, were obtained by applying Latent Semantic Analysis (LSA) on the *Honeypot* dataset (Halko et al. 2011). The *LSA* is a topic modelling method that is useful in capturing the semantics and relevance of terms to a document (Wiemer-Hastings et al. 2004). Prevalent keywords from the *LSA concepts* include *free*, *new*, *lots*, *win*, *follow*, *trade*, *good*, *great*, *make*, *create*, *twitter*, *followers*, *check*, *gain*, *buy*, *account*, *get*, *making*, *online*, *want*. Tweets that contained at least two of the most representative keywords in the *Polluter* part of the *Honeypot* dataset were considered as spam. Table 5.2 shows some examples of tweets that satisfy this criterion. A block diagram of the collection and validation process is shown in Figure 5.1.

5.3.3 Validation of $SPD_{automated}$

The act of labelling data-points in the $SPD_{automated}$ as spam is based on the hypothesis that spam users are more likely to use at least two of the terms obtained via the *LSA concept terms* on the part of the *Honeypot* dataset that is known to be spam (Lee et al. 2011). To validate this,

Table 5.2: Examples of collocational bigrams from the spam part of $SPD_{automated}$. Keywords returned by LSA on the *Honeypot* dataset are shown in bold face. Actual users mention were replaced with the ‘user’ placeholder to preserve anonymity.

Id	Tweet
T1	RT @user: Retweet to win up to 121+ followers must be following me 🙏
T2	Retweet this for 81+ free follows 🌸🌿
T3	Retweet for 125 free follows 🌈 '\n' Retweet and Fav 🍷 this if you have my post notifications on! 🍷 For 125 free followers 🍷🍷
T4	Watch and like this video for free 80 followers 🍷 url
T5	Retweet to win up to 130+ free followers 🍷 ' ' @user
T6	RT @user: Retweet this to gain followers faster 🍷🍷🍷
T7	Follow everyone who FAV this 🍷
T8	@user @user @user @user @user follow everyone who likes this 🍷 #SolarEclipse2017 \

Table 5.3: Percentage distributions of relevant metrics computed in the two parts of $SPD_{automated}$, i.e. legitimate and spam.

Data	% Name Similarity	% digits in names	% containing spam bigrams	% LexRich unfiltered	% LexRich filtered
Legitimate	82.59	14.07	1.05	97.43	86.74
Spam	26.27	88.84	89.51	90.94	49.46

we compute and compare in the *legitimate* and the *spam* part of $SPD_{automated}$ according to: (1) the distribution of the *co-occurring keywords* (2) *lexical richness* and *lexical density* and (3) the distributions of *user mentions* and *URLs*. Table 5.3 shows the results.

Distribution of co-occurring keywords

In Table 5.3, it can be observed that only 1.05% of the tweets in the legitimate part of $SPD_{automated}$ contain two or more keywords, extracted using LSA from the *polluter* part of the *Honeypot dataset*. In contrast, more than 89.5% of the tweets in the spam part of $SPD_{automated}$ contain keyword pairs. The observed distributions reinforced the intuition applied in labelling the data instance as spam or otherwise, and also minimises the risk of labelling legitimate users as spammers. Table 5.2 shows examples of frequent co-occurring keywords sampled from the $SPD_{automated}$.

Lexical richness and density

Type-token ratio (TTR), a measure of the richness of lexicons in a document (Biber & and Geoffrey Leech 2002), is applied to understand how distinct words are utilised in the legitimate

and the spam class of the $SPD_{automated}$. For a dataset \mathcal{D} , the TTR is computed as follows:

$$TTR = \frac{\text{unique tokens in } \mathcal{D}}{\text{tokens in } \mathcal{D}} \quad (5.1)$$

Intuitively, legitimate users are expected to use rich and diverse lexicons in response to various discussion topics on Twitter. In contrast, spam users may focus on specific targets such as promoting a certain product or marketing to increase the number of their followers; users engaging in this behaviour are highly likely to recycle specific sets of similar words quite often. A more advanced version of the TTR (Eq. 5.1) is *lexical density* (LD), which recognises *content-bearing terms* only (Biber & and Geoffrey Leech 2002). The *lexical density* is expressed as follows:

$$LD = \frac{\text{words in } \mathcal{D} \text{ excluding stopwords}}{\text{tokens in } \mathcal{D}} \quad (5.2)$$

Table 5.3 shows the result of computing the TTR (Eq. 5.1) and LD (Eq. 5.2) metrics in both datasets.

User mention

A random mentioning of users is a common tactic employed by spammers to expand the visibility or their network of followers (Lee et al. 2011, Howard & Kollanyi 2016). In Table 5.3, the lexical richness, i.e. *%LexRich (unfiltered)*, in the spam set is marginally higher than expected. Noting the high proportion of user mentions in spam data, lexical richness (*% LexRich (filtered)*) or *lexical density* is computed without considering the *user mentions* and *URLs* in both datasets. The computation in the spam dataset led to a shallow score suggesting that the large number of *user mentions* and *URLs* are responsible for the relatively high TTR score observed earlier in the spam dataset. However, the TTR score in the legitimate dataset is not affected by filtering out *user mentions* and *URLs*, which is indicative of the richness and diversity of the lexicons used by genuine users. Similarly, the low TTR score in the spam dataset indicates that the same words are being used repetitively usually not matching the discussion topics. Table 5.3 also shows metrics related to naming conventions by computing the degree of similarity between the *username* and the *screenname* of each user, and the proportion of digits in their names. This topic is discussed further in Section 5.3.4.

5.3.4 Feature extraction

The *Twitter* platform facilitates global connections and interactions of diverse users (Qazvinian et al. 2011) using myriads of features that could be used for various analyses. Figure 2.2 presents an overview of the platform depicting relevant attributes that enable users to connect, and they form the basis of the features utilised in the research.

Accessibility, dynamism and categorisations of features

While a collection of historical tweets can be retrieved, visibility of tweets on Twitter can be viewed as a *time bound*, because they remain available for a short time (approximately a week) before being replaced by more recent ones. Real-time spam detection systems that rely on historical features from past tweets are affected by this constraint and may be practically less effective. Readily available and dynamic features offer an enhanced opportunity to distinguish *spam* from *non-spam* tweets in real-time. To leverage this potential, features are categorised as follows:

- *User Profile Features (UPF)* include information about the user, such as their *user name*, *screen name*, *location* and *description*
- *Account Information Features (AIF)* consist of information such as *account creation time* (*account age*) and *account verification flag* (*verified or not verified*)
- *Pairwise engagement features* are sub-categorised into:
 - *Engage-with Features (EwF)* include features that describe users' activities on Twitter and the users can influence or choose how to alter their values. Features under this group include *friends count*, *statuses count*, *tweet type*, *tweet creation time*, *tweet creation frequency*, etc.
 - *Engaged-by Features (EbF)* are similar to features in the EwF group. The main difference is that features under this group cannot be influenced by the users directly. For instance, a user relies on other users to increase their *favourites count* or to attract more *followers*. Features in this group include *followers count*, *favourites count*, *number of retweet (RT)*, etc.

Furthermore, features can be classified as *basic features* or *derived features*. The features mentioned above, i.e. under *UPF*, *AIF*, *EwF* and *EbF*, are basic features. In contrast, derived features are computed using two or more basic features or are based on further analysis, e.g. sentiment analysis or entropy computation on textual data. Features can also be characterised as *static* or *dynamic*. Static features cannot be changed once the account is created, e.g. *user ID* and *account creation time*, whereas dynamic features keep changing depending on the user's level of engagements on Twitter, e.g. *statuses count*. A detailed description of all the features and their properties are provided in Table A.1 of Appendix A.

Feature selection

The choice of features for the experimentation is informed by insights gained from a series of exploratory analyses to understand the distribution of textual features, the composition of data, and the dynamism of features, such as *statuses count*, *friends count*, *followers count*, *favourites count*, *naming conventions* and *tweeting patterns*.

Account age & naming convention: According to the exploratory analyses results, accounts with very high statuses and friends count, but low favourites count and followers count at young age are likely to be automated spam posting accounts. For example, Figure 5.2 shows huge amounts of content generated within a short period. These observations are noted in deriving features, such as *Activeness*, *Interestingness* and *Followership* (see Table A.1 of Appendix A). In terms of naming conventions, *username* and *screenname* of a Twitter user usually exhibit a high degree of similarity. Typically, the *screennames* of legitimate users contain segments of the *username*, are not very lengthy and rarely begin with a digit. In some cases, *usernames* of legitimate users contain a reasonably small number of digits in the middle or at the end. In spam accounts, the mix of letters, digits, special characters and unusual symbols is much more widespread. Often, names begin with digits or email addresses, and as shown in Table 5.3, there is high discrepancy between *usernames* and the corresponding *screenname*. Features, such as *NameSim* and *NamesRatio* in Table A.1, are inspired by this analysis. Other static features in the metadata of a user account on Twitter, such as the *Language* and *Location* fields, may be useful to some extent for identifying spam accounts, due to the fact that most of these fields are either vacant or populated with meaningless content for spam users. Genuine users often report

a real location name. However, spam posting accounts often return irrelevant content or lengthy and unintelligible sequences of characters or just email addresses.

Tweeting activity and posting behaviour: In an earlier study, spam posting users have been observed to post four tweets per day on average (Lee et al. 2011); this research observes an automated *spam-posting account* posts on average at least 12 tweets per day at distinct periods. In some cases, the activity levels remain constant within approximately four long-lasting periods. Figures 5.2 and 5.3 show examples of spam and legitimate user activity patterns, respectively. In contrast to automated spam-posting users, a legitimate user of Twitter often follows random usage patterns and takes long breaks of inactivity. Figure 5.3 represents the activity patterns of two different users with relatively low traffic generation within the same period as the users in Figure 5.2. Table A.1 shows the features proposed for prediction model training, the corresponding feature groups and definitions.

5.4 Spam Prediction and Evaluation

This section discusses the experimental procedure and the results obtained. All experiments were conducted using the *Scikit-learn toolkit* (Pedregosa et al. 2011).

5.4.1 Parameter tuning and prediction models

An effective classifier should be able to correctly classify previously unseen data by leveraging the experience gained from training on some labelled samples, i.e. data instances and the corresponding class. The target of the classification task at hand is to predict spam-posting users or normal legitimate users correctly, by accessing one of their tweets and the associated account *meta-data*. Noting that traditional *machine learning* requires an effective hyper-parameter tuning, which is vital for significantly improving the performance of prediction models (Olson et al. 2018), relevant hyper-parameters in the model are tuned via a grid search technique on a standard *10-fold cross-validation*.

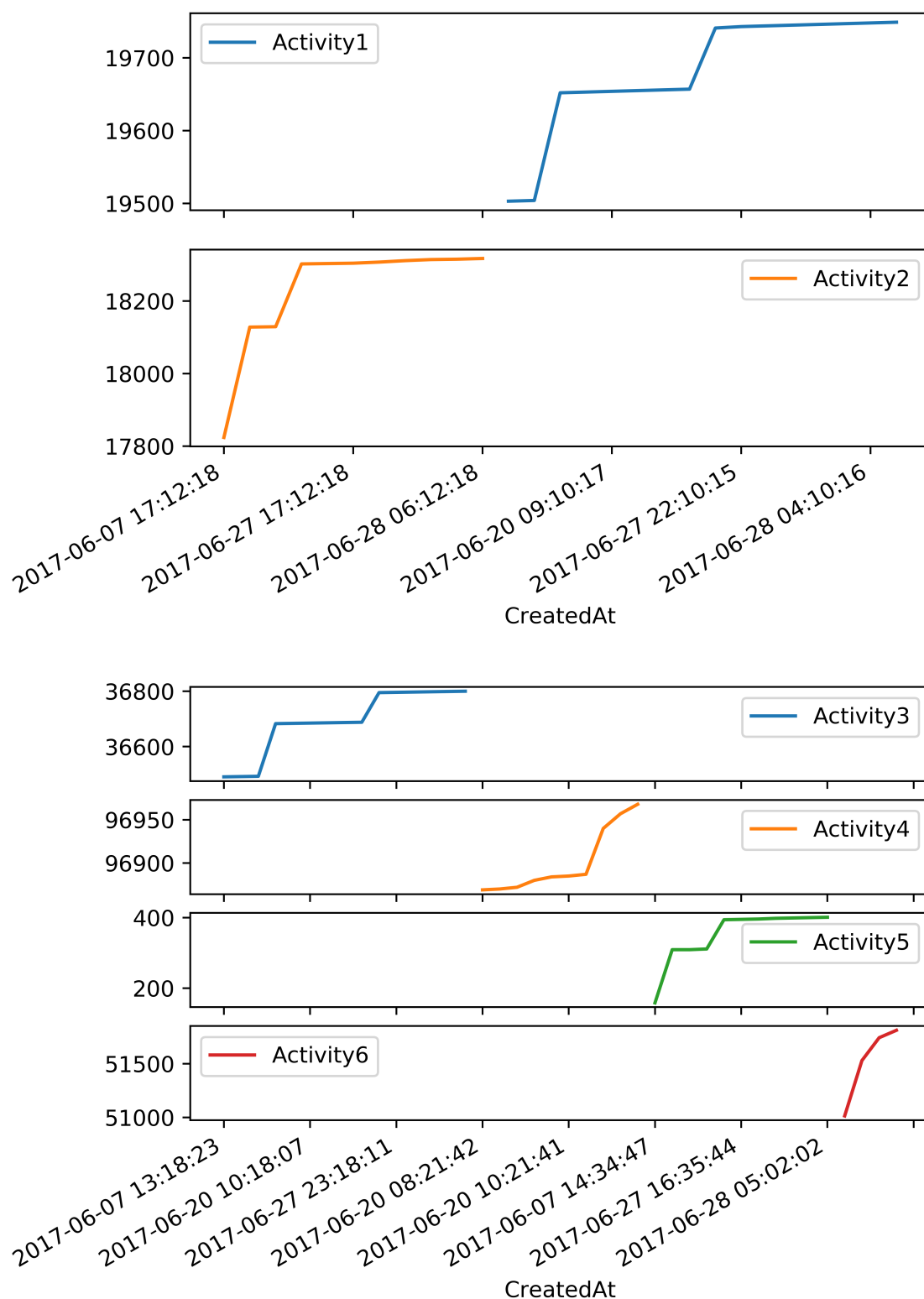


Figure 5.2: Example of activity patterns of spam-posting social bot accounts. All sub-figures depict hyperactive automated users that generated very high traffic within a short period. The activity distribution over time for most users resembles the *staircase* function. Some users generate much higher traffic than others, e.g. *Activity4* and *Activity6* represent many times more tweets than *Activity5*.

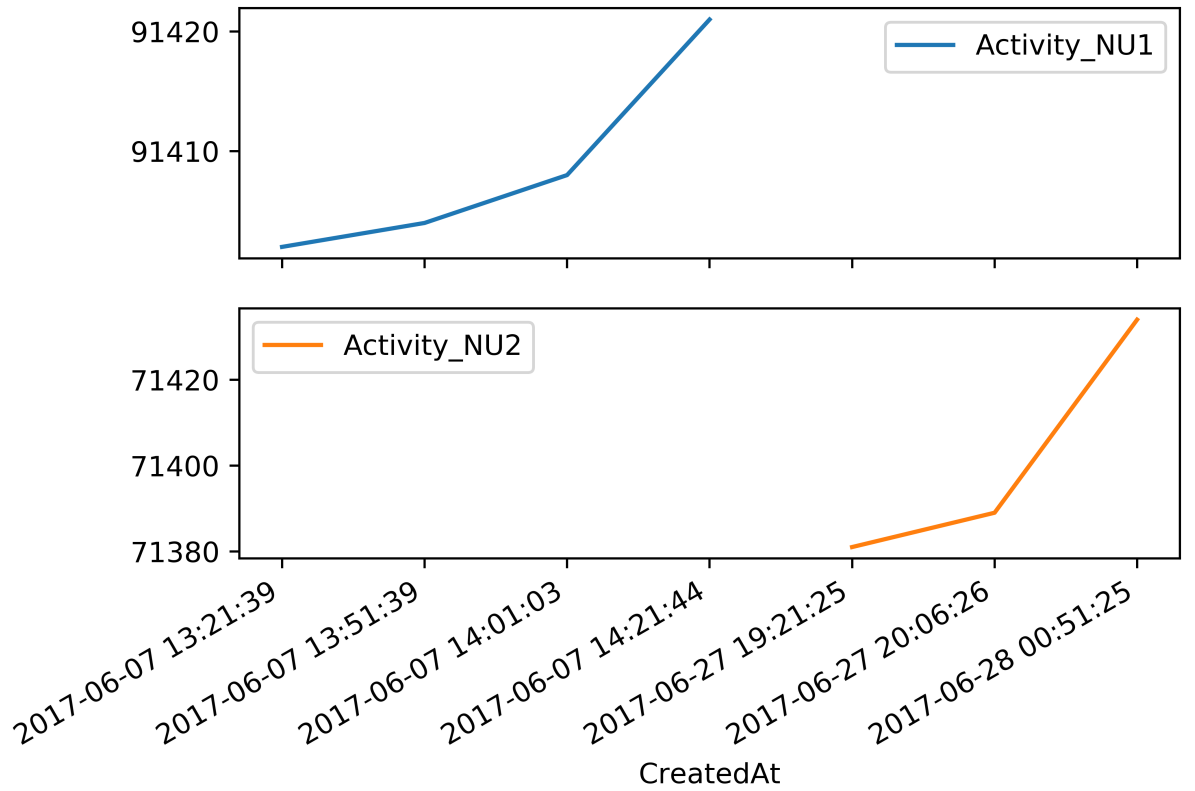


Figure 5.3: Example of activity patterns of two legitimate users. The figure shows a relatively low traffic generation within the same period as the users in Figure 5.2.

5.4.2 Feature importance and correlation

During an initial analysis stage, a large number of features have been used for training, and some features were discarded due to their relatively low contribution to the overall performance. Essentially, a recursive feature elimination approach was adopted to measure the contribution of each feature to the overall performance. Correlation analysis plays a crucial role in achieving optimum performance because features that correlate perfectly introduce redundancy, and do not add extra information into the classification models (Guyon & Elisseeff 2003). To understand the relevance of each feature in predicting the target class, the thesis conducted a *univariate feature analysis*. The results are shown in Figure 5.5, formatted as a heat-map in which the colour intensity corresponds to the correlation degree of each feature with the target class, i.e. *AccountClass*, and with other features. The main diagonal of the *heatmap matrix* represents perfect correlation because each feature is correlated with itself, and the column of the target (*AccountClass*) shows the intensity of the correlation of each feature with the target. With the

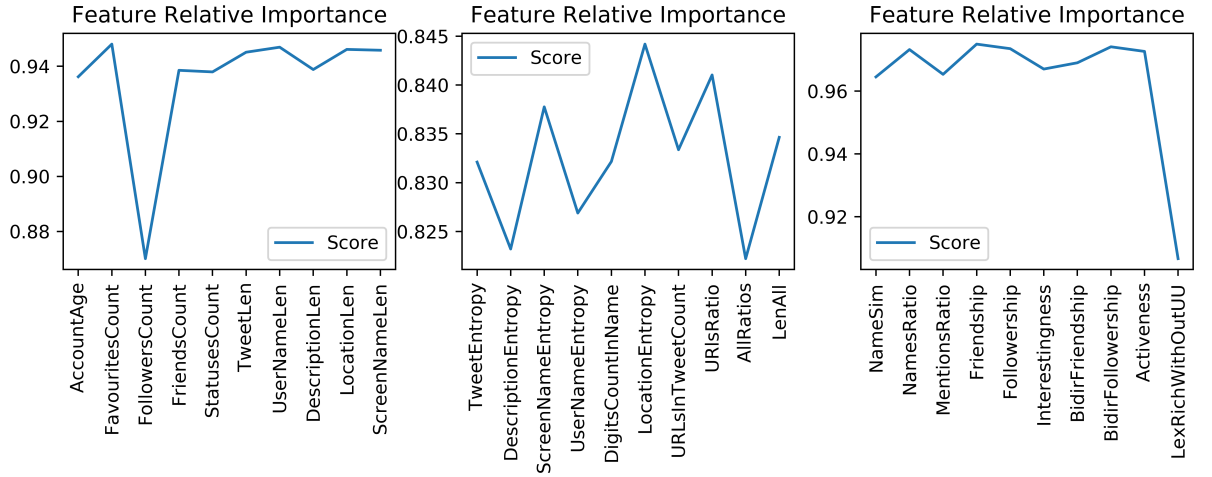


Figure 5.4: The figure shows the performance of features measured using recursive feature elimination method. The most informative feature is the lexical richness of tweets including user mentions and URLs (*LexRichWithUU*), which contributed significantly to the overall performance, as evidenced from the sharp drop in the figure.

exception of *lexical richness*, *LexRichWithUU*, and *lexical density*, *LexRichWithOutUU*, which were derived from the same root, there is no other pair of features with perfect correlation. Thus, the set of features shown in Figure 5.5 comprises of feature set for all experiments in this part of the research. In the preliminary stages of this study, many features, mainly derived as combinations of the features in Figure 5.5, have been used for experimentations. Most of these features were discarded due to correlating almost perfectly with others, hence, not contributing to the accuracy of the model.

5.4.3 Performance metrics

For evaluation, different metrics are utilised in order to avoid any bias towards the majority class, especially when the dataset is imbalanced (Fawcett 2006). In particular, the following metrics have been used to summarise the experimental results: *F-score*, *Precision*, *Recall*, *Accuracy*, the *Receiver Operating Characteristics (ROC) curve* and the *area under the ROC curve (AUC)*. The *F-score*, the geometric mean of *Precision* and *Recall*, captures a model’s prediction quality especially in sensitive areas, by requiring both *Precision* and *Recall* to be high. The *AUC* offers a more encompassing metric, insensitive to the imbalance between classes that sometimes provides better evaluation than accuracy (Japkowicz 2000). Specifically, the higher the AUC score, the larger the area under the curve, i.e remains well above the diagonal, see

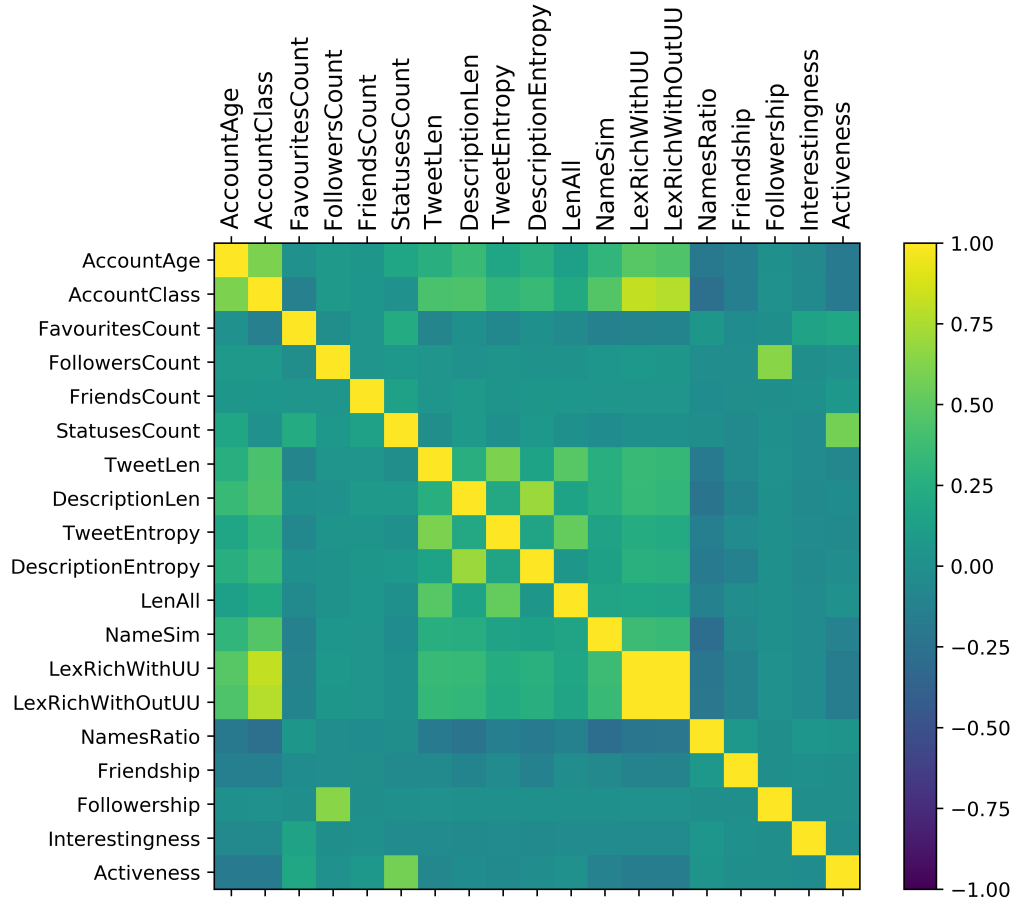


Figure 5.5: Visual representation of the univariate analysis of correlation of each feature with the target, i.e. *AccountClass* and other features. Correlation magnitudes range from 1 to -1, with 1 denoting perfect positive correlation, 0 no correlation and -1 perfect negative correlation. Features highly correlated with the target constitute the optimum features set.

Figure 5.6.

5.4.4 Experimental results

Series of experiments were conducted with different classification models and were assessed using the metrics previously described in Section 5.4.3. The first experiment aimed to investigate the effectiveness of the proposed *SPD features*, and compared the suitability of different classification models for the task at hand. In total, six different classification models were used: Maximum-Entropy (MaxEnt), Random Forest, Extremely Randomised Trees (ExtraTrees), C-Support Vector Classification (SVC), Gradient Boosting and Multi-layer Perceptron (MLP). Because it is possible interrupt the MLP classifier and extract the information learned from

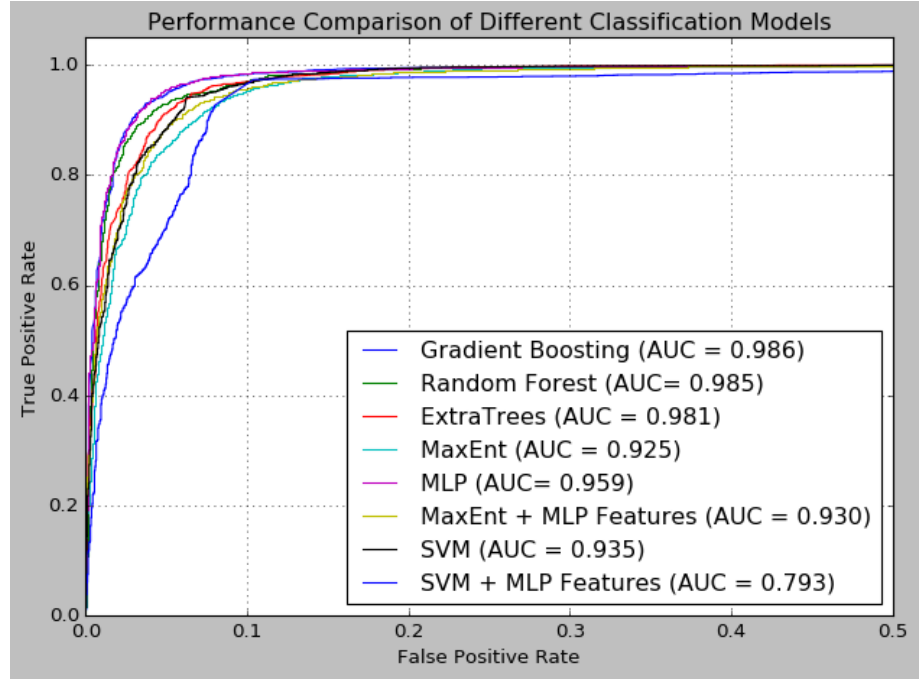


Figure 5.6: Performance of different classification models evaluated on the $SPD_{automated}$ dataset using 10-fold cross-validation.

the training data, the following models are used: *MaxEnt* + *MLP* and *SVM* + *MLP*. The two models are a form of a hybrid model and have been used to improve the prediction task by harnessing the power of the composite models. Instead of using the raw features, the former (*MaxEnt* + *MLP*) relies on the features learned during the training of the MLP model to train the MaxEnt classifier. In contrast, the latter uses them to train an SVM classifier. Figure 5.6 shows the learning curves and corresponding AUC scores achieved by each model on the best hyperparameter values, as explained in Section 5.4.1. All the models were trained and evaluated on the $SPD_{automated}$ dataset using 10-fold cross-validation. The chart shows a relative consistency in terms of performance across the different classification models, which can be attributed to the effectiveness of the proposed *SPD* features. Henceforth, *Gradient Boosting* is chosen for the remaining experiments due to its relatively higher performance.

The second experiment compared the features proposed in this study, the *SPD* features, with the *Honeypot* features, proposed in Lee et al. (2011). Since the study of Lee et al. (2011) is the main baseline, a comparison of the two feature sets is conducted on the *Honeypot* dataset and the $SPD_{automated}$ dataset, using a 10-fold cross validation. The associated learning curves are shown in Figures 5.7 and 5.8, respectively. The figures show that *SPD* features perform

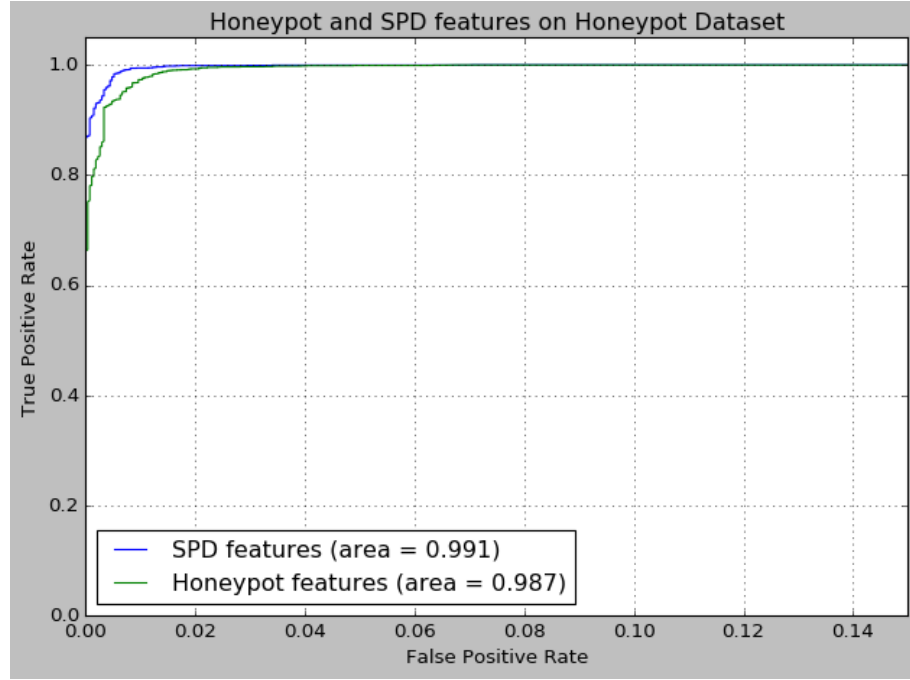


Figure 5.7: Learning curves of the *SPD* features and the *Honeypot* features on the *Honeypot* dataset Lee et al. (2011). The *SPD* features achieve a slight improvement in performance.

better than the *Honeypot* features for both datasets, in which the improvement is small for the *Honeypot* dataset, whereas it is significant for the *SPD_{automated}* dataset. It should be noted that the *Honeypot* dataset does not provide enough information for computing all *SPD* features. As a result, the *SPD* features line in Figure 5.7 is based on some *SPD* features only. Features such as *Interestingness*, *Activeness*, *NameSim* and *Lexical Richness* are not used in this experiment. The lack of these features explains why the improvement in performance is minimal.

Finally, Table 5.4 presents experimental results for all combinations of the datasets and the features sets in the current study and that of Lee et al. (2011). To address the imbalance in the *SPD_{manual}* dataset, the *SMOTE* technique (Chawla et al. 2002) is used to up-sample the minority class during training the classifier. The set of features proposed in this work, *SPD*, performs better than the *Honeypot* (Lee et al. 2011) on all datasets. The lightweight version of *SPD* features, as computed by the feature selection process in Section 5.4.2, perform better than the *Honeypot* features set when applied on *SPD_{automated}* but worse than the *Honeypot* feature set when applied on *Honeypot* and *SPD_{manual}*. The lightweight version of *SPD* features consistently performs worse than the full *SPD* features set, as expected.

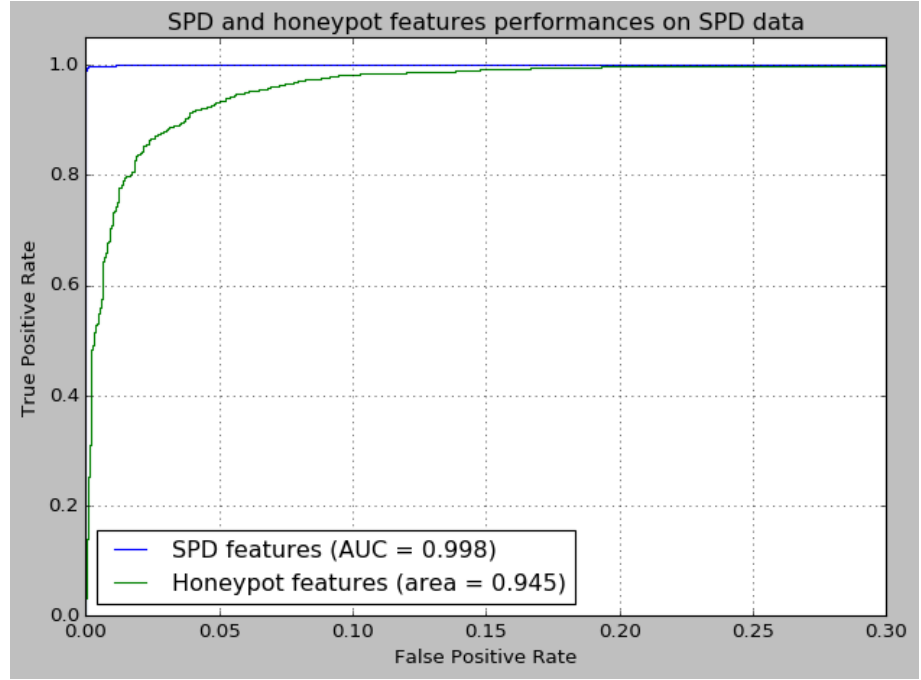


Figure 5.8: Learning curves of the *SPD* features and the Honeypot features Lee et al. (2011) on the *SPD_{automated}* dataset. The *SPD* improve performance significantly.

5.4.5 Error analysis

To understand the reasons that may have led to misclassification of some of the representative samples, shown in Figure 5.5, an *error analysis* is carried-out to investigate cases that were not classified correctly by the classification model. Because the collection of the *SPD* dataset consists of tweets in English, some tweets in the dataset, such as tweet #1 in Table 5.5, were not in English and were misclassified. This can be attributed to the fact that although the first language field in the meta-section of some user profiles was set as English, the actual interaction language in the tweet is not English. Moreover, as shown in Figure 5.4, lexical richness and density are essential classification features. However, the occurrence of irrelevant tokens in a tweet, which were regarded as unique, leads to a more sumptuous lexicon, which in turn increases the chance of classifying the tweet as legitimate. Tweets #2-#6 in Table 5.5 contain some irrelevant symbols, which were counted as unique, increased the corresponding lexical score and misled the classifier. Emoticons are also a source of confusion for the classifier, especially when computing the lexicon of unique tokens for a tweet and its similarity to lexicons of other tweets.

Table 5.4: Evaluation results for various combinations of datasets and feature sets. ‘(0, 1)’ denotes performance on the spam part and the legitimate part of each dataset, respectively.

Dataset	Features	Accuracy %	AUC %	Precision % (0, 1)	Recall % (0, 1)	F-score % (0, 1)
<i>Honeypot</i> Lee et al. (2011)	<i>Honeypot</i> Lee et al. (2011)	98.53	98.55	(99, 98)	(98, 99)	(99, 98)
	<i>SPD selected</i>	93.26	93.29	(95, 92)	(92, 95)	(93, 93)
	<i>SPD</i>	98.93	98.94	(99, 99)	(99, 99)	(99, 99)
<i>SPD_{automated}</i>	<i>Honeypot</i> Lee et al. (2011)	94.93	94.94	(96, 94)	(95, 95)	(95, 95)
	<i>SPD selected</i>	95.12	95.16	(96, 94)	(95, 96)	(95, 95)
	<i>SPD</i>	99.75	99.75	(99, 99)	(99, 99)	(99, 99)
<i>SPD_{manual}</i>	<i>Honeypot</i> Lee et al. (2011)	89.38	59.19	(35, 93)	(23, 96)	(27, 94)
	<i>SPD selected</i>	88.50	60.99	(39, 92)	(27, 95)	(32, 94)
	<i>SPD</i>	98.27	98.28	(97, 100)	(100, 97)	(98, 98)

Table 5.5: An Example of sample tokens from misclassified tweets.

Id	Tweet
1	gain followers 🍌@ 🍌🍌🍌 アメリカとカナダでまっています。🍌地域社会に溶けめずにいた民が、ディナ会での出會。
2	retweet this 🍌🍌
3	like this 🍌🍌
4	follow like & retweet 🍌🍌
5	follow back follow you 🍌🍌
6	gain followers 🍌...’, 68), gain followers 🍌...’, this 🍌🍌; retweet this 🍌🍌

5.5 Spam accounts and their features

This section presents an additional analysis of the manual annotations in the *SPD_{manual}* dataset, a description of the different user groups and discusses the distribution of relevant features in the dataset.

5.5.1 Characterising users

A thorough inspection of the tweets in the spam and legitimate parts of the *SPD_{manual}* dataset suggests that there are two kinds of users on Twitter: *human* users and *social bot (autonomous entity)* users. Each user type consists of a legitimate (non-spam) and a spam part, as depicted in Figure 5.9, with the following characteristics.

Legitimate users

Legitimate users interact with moderate frequency, within the reasonable and acceptable Twitter usage policy, and this user group also contains *genuine multiple users*, i.e. accounts managed by organisations or *useful social bots*. The users in this group tend to exhibit a proportionate

interaction level and (*activeness*), i.e. their statuses count matches their account age and the tweets they post are of interest to followers, hence exhibit high *interestingness*. Followers of users in this group often outnumber friends, sometimes even by twice as much. This is expected since most users subscribe or follow an account due to their interest in it. In terms of *naming convention*, the *username* and *screenname* of useful social bot accounts often contain the word ‘bot’ as part of name, e.g. *AIBigDataCloudIoTBot* and *Troll Bot*. In some cases, groups of *screennames* share the same suffix separated by the underscore character from a description of the account. Accounts in this group achieve relatively high *interestingness* levels and an almost equal proportion of friends and followers. They also show a moderate similarity between their *username* and *screenname*, and use a wide variety of words and expressions, i.e. diverse lexicons.

Spam-posting users

Spam-posting users are hyperactive and generate irrelevant content, potentially offensive to other users and in violation of Twitter’s terms of use⁵. Accounts in this group exhibit very low *interestingness*, and disproportionate *activeness* levels i.e. the statuses count does not match the account age, indicating that they employ flooding techniques. Friends of users in this group usually outnumber followers. The interaction patterns of spam-posting social bot accounts are often randomised rather than well-defined, as shown in Figure 5.2. There is also a high level of inconsistency in naming conventions and a high dissimilarity between *usernames* and corresponding *screennames*. The *screenname* of spam-posting social bot accounts is often unintelligible, mostly containing digits and special characters. Spam-posting users also exhibit low lexicon richness due to the high proportion of *URLs*, *retweets*, and *user mentions*. Spam users generally engage in subscribing to different conversations on Twitter (based on hashtags) and generate tweets not related to the topic of discussion. Figure 5.10 shows a summary of user groups on Twitter, *human* and *social bot*, *legitimate* and *spam-posting*.

⁵See www.help.twitter.com/en/rules-and-policies/twitter-rules/

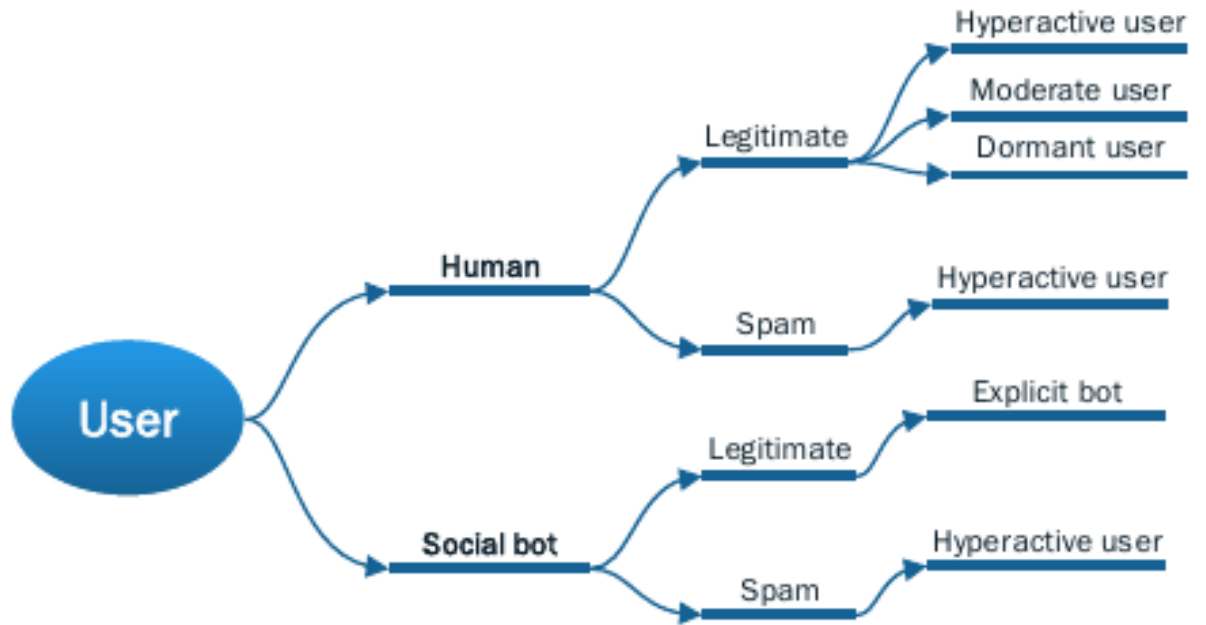


Figure 5.9: User types in the SPD_{manual} dataset.

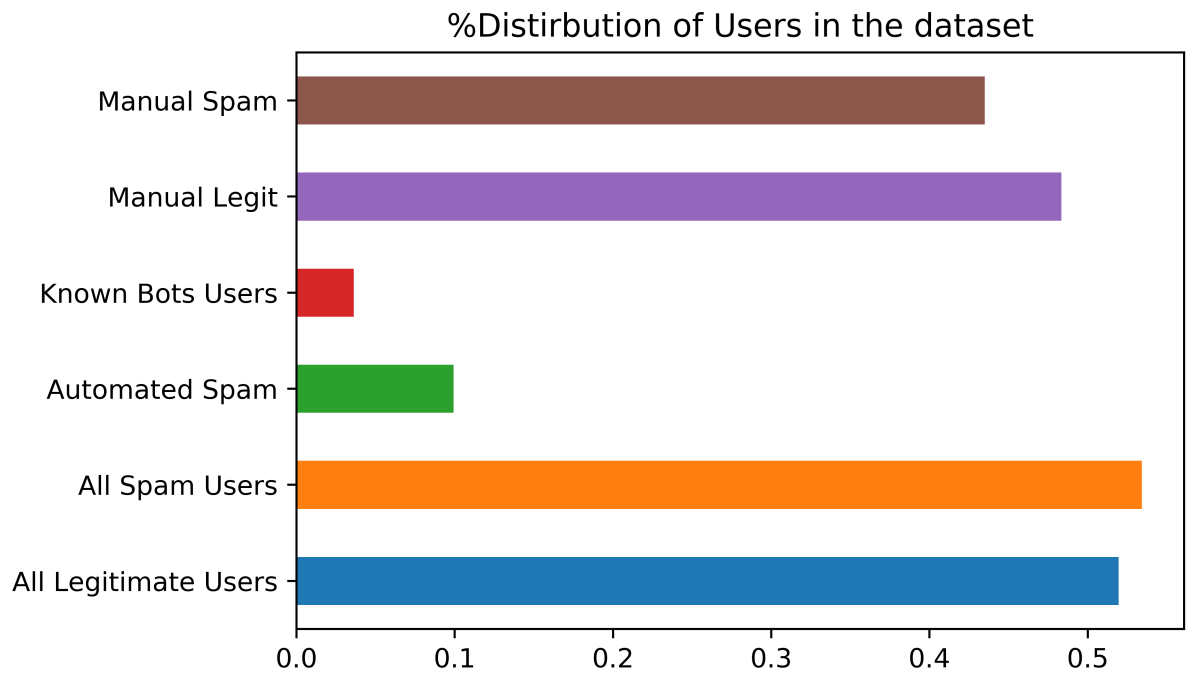


Figure 5.10: Distribution of different users in the SPD_{manual} dataset. The y -axis represents account types and the x -axis shows the percentage proportion (%) of account types. Known bots are accounts that mention the word ‘bot’ explicitly as part of their name and share some basic features similarities with normal users such as the level of name similarity ($NameSim$). Known bots in the dataset account for less than 10% of all users.

5.6 Summary

The growth rate of spam volume within the *social media ecosystem* can be attributed to the lack of physical contact between the communicating parties. This makes it difficult to ascertain the actual identity of the user and the legitimacy of the content being posted. As a result, much irrelevant content is encountered. Utilising data directly from social media platforms without active filtering may mislead and lead to wrong conclusions due to unrepresentative data. This chapter offers an effective method for spam detection and new insights into the sophisticatedly evolving techniques for spamming on Twitter. The proposed spam detection method utilised an optimised set of readily available features. Being independent of historical tweets, which are often unavailable on Twitter, makes them suitable for real-time spam detection. The efficacy and robustness of the proposed features set are shown by testing several machine learning models and various datasets. To ensure the quality and representativeness of the research data, the proposed *SPD strategy* is applied in all the data collection pipelines used in the research in filtering out irrelevant content as a form of an early pre-processing stage.

Chapter VI

MICROCOSM: A META ANALYSIS

6.1 Introduction

The rationale behind this chapter is to present a set of analyses that will inform the final detection framework described in Chapter IV. The chapter focuses on a pragmatic approach to identify nodes' attributes (Figure 4.4(b)) exhibiting strong correlations, which will eventually lead to improving mining tasks involving Twitter. The approach involves statistical analysis and experimentations on various datasets.

6.2 Identifying Structurally-Related Nodes

The thesis posits that the connection topology on Twitter contributes to widespread spurious content and a less cohesive community of users due to many unreciprocated or *event-type ties*. As a precursor for identifying *microcosms* or cohesive communities on Twitter, it is vital to account for both *structural* and *content* similarities. This is in response to the growing complexity and heterogeneity of connections, which challenge many mining-related tasks such as *communities detection* and *authentication of online content*. Concerning the *structural similarity*, the research examines the basic units of structural relationships – *dyadic* and *transitive* ties (Section 4.4) – under the premise that the level of trust is stronger among users that share reciprocal ties, and it is highly unlikely for a user in the group to misuse the network e.g. spread fake news or spam. By focusing on smaller groups, it enables the discovery of a form of a personal network on Twitter, which is homogeneous to many sociodemographic behavioural, and intrapersonal characteristics (Miller McPherson et al. 2001). Such groups are more intimate, with a high degree of familiarity due to strong social cohesion (Dunbar 1998); and play a vital

role in the structural analysis of a network (Freeman 1996). However, acquiring large amounts of tweets sufficient to identify such cohesive groups is challenging and time-consuming; the forthcoming sections expound on the challenges and the proposed mitigation measures.

6.2.1 Reciprocal Units

In social science, a taxonomy of social relationships is described as a function of closeness among users. The closer the users are, the more cohesive and trustworthy. Social networks are useful for linking *micro* and *macro* levels of sociological theory by enabling the analysis of various forms of relationships.

Dyadic and Simmelian ties constitute the basic unit of analysing structurally-related nodes. The concept of *dyad* has been viewed from various perspectives and often with contradicting results. Previous studies (Weng et al. 2010, Kwak et al. 2010, Cha et al. 2012, Arnaboldi, Conti, Passarella & Pezzoni 2013) have examined reciprocity for various tasks, which are either based on *directed sets of nodes* or *textual content*. A directed tie is peculiar to Twitter since, in other platforms such as Facebook, an automatic reciprocal relationship is established once a friend request is accepted. *Transitivity* defines a social preference to be friends with a *friend-of-a-friend* and has been recognised as a peculiar feature of a network (Watts & Strogatz 1998). For network analysis, transitivity is a vital feature of a network that enables the formation of cohesive communities and enables understanding of the structure of social ties in a network (Granovetter 1977). However, as illustrated in Figure 1.2, the prevalence of transitory connections makes it challenging to identify both the dyadic and transitive ties based on *state-type ties* on Twitter. Identifying the set of fully connected nodes on Twitter is challenging due to the eccentric underlying connection patterns, which enables flexible followership that results in many unidirectional links.

6.2.2 Identifying Dyads

The process of identifying dyads on Twitter is captured by *Algorithm search-dyads* (Algorithm 1), which searches and profiles users as *directed or 1-edge* or *undirected or dyads* for further analysis. The goal of the algorithm is to formalise and simplify the task of searching for nodes with actual reciprocal ties (dyads) on Twitter for evaluation.

Algorithm 1 *Algorithm search-dyads* profiles users with *directed/1-edge* and *undirected/dyads* ties on Twitter

```

1: Initialisation:  $1 - edge \rightarrow \{\}, dyads \rightarrow \{\}$ 
2: Input: begin with an arbitrary set of seed users, say  $k$ 
3: while  $k \neq \emptyset$  do
4:    $\forall v_i \in k$ , get sets of friends  $fr_{v_i}$ , followers  $fl_{v_i}$ ,  $fr_{v_i}, fl_{v_i} \in m_{v_i}$ ;  $m_{v_i}$  denotes  $v_i$  network
5:    $\forall v_j \in fr_{v_i}$ , retrieve the sets  $fr_{v_j}$  and  $fl_{v_j}$ ,  $fr_{v_j}, fl_{v_j} \in m'_{v_j}$ ;  $m'_{v_j}$  denotes  $v_j$  network
6:   if  $v_i \in fr_{v_j}$  then
7:      $v_i \sim v_j$   $\triangleright$  both follows one another
8:     update dyads
9:   else
10:     $v_i$  follows  $v_j$ 
11:    update 1-edge
12:   end if
13: end while

```

Dyadic Analysis The dataset for *dyads analysis* was collected using the Twitter API according to *Algorithm search-dyads* (Algorithm 1). The process begins with 4022 seed users¹ from *verified* and *unverified* accounts. The network profile of each seed user² or network composition m_{v_i} , consisting of a list of *friends* fr_{v_i} and *followers* fl_{v_i} , was searched to determine pairs of users that follow each other. Table 6.1 shows the basic statistics of users visited by the collection crawler. In particular, it shows the counts of directed ($1 - edge$) connections and dyadic ties. Similarly, Figure 6.1 summarises dyads in the *verified* and *unverified* user category. Unlike previous studies (Leskovec & Mcauley 2012, Yoshida 2013, Yang & Leskovec 2015) in which datasets from various social networks were collected, this study focuses on nodes with real reciprocity not directed ties.

Table 6.1: A summary of the dyadic data. Many *unverified* users had to be visited due to the large number of $1 - edge$ or directed connections, occurring when followers are not being followed back.

Category	Seed Size	Visited Users	Retrieved	Remark
Unverified dyads	2,023	13,409,661	8,715	utilised for prediction
Verified dyads	1,999	3,893,075	–	not used for prediction
1-edge and null tie	1,700	–	7,014	utilised for prediction

Network topology: The datasets in Table 6.1 is used to explore the potential contribution of dyadic ties to the detection of local communities. In the collected data³, 55% and 21% of

¹These are genuine users devoid of spammers or social bots collected based on the SPD filtering technique (Inuwa-Dutse, Liptrott & Korkontzelos 2018).

²A 'list' on Twitter allows a user to store a set of preferred users and can be used to obtain relevant information. However, not all users maintain it, and where it exists, it may likely contain a curated news source for the user.

³Data available at https://github.com/ijdutse/dyads_in_Twitter

unverified and *verified* profiles, respectively, are involved in dyadic ties (see Figure 6.1). This is a useful observation because, despite the large proportion of dyads, a random collection of tweets corresponds to far fewer users in dyadic ties. In Figure 6.1, the *Verified users* have more

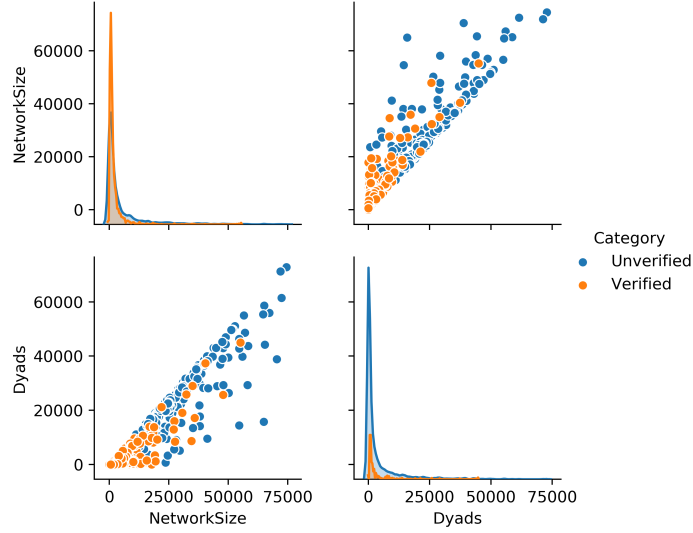


Figure 6.1: The proportion of dyadic ties and network size in the data. The *verified* category exhibits larger network sizes but fewer dyads in comparison to the *unverified* category.

network neighbours than their *unverified counterparts*, but there is a higher proportion of dyadic ties in the *unverified category*. With reference to Figure 6.1 and Table 6.1, the huge numbers of *visited users* for the number of *dyads* reveal a high proportion of *null connections* and *1-edge connections*. In view of this, subsequent analysis will focus on ordinary *unverified users*. The

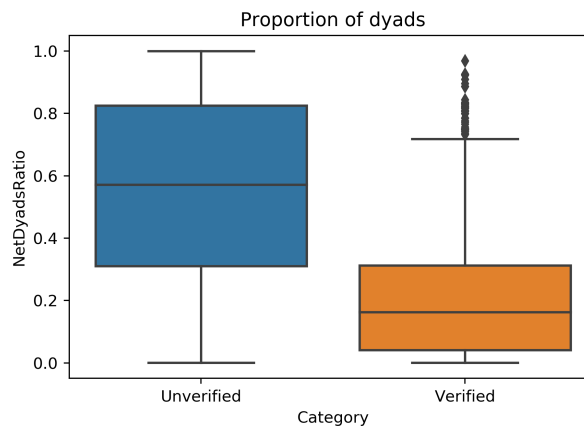


Figure 6.2: Dyads proportions in *verified* and *unverified* profiles

analysis of *network topology of dyads* shows that, given a pair of users (a, b) , b is likely to follow a back:

- if a and b are both in the *unverified user's category*

- if both a and b have a low or relatively large number of followers or network size, i.e. based on the average of those metrics in the users' categories
- if a has more followers than b or if a is a verified user.

The opposite of the above statements holds for *verified users*.

Dyads Prediction

Noting the flexibility of connections on Twitter and the rarity of reciprocal links, identifying dyadic ties at scale is difficult and time-consuming. A manual process will be prone to a *curse of dimensionality*. Consequently, the goal is to efficiently predict a large collection of users with a high likelihood of reciprocal ties for clustering purpose. The probability of a user reciprocating a relationship, i.e. by following back, and how users of varying influence on Twitter reciprocate their followers have been analysed in (Cha et al. 2012); this motivates the tie prediction aspect. A *user-centric* approach is applied to predict the formation of a tie between any pair of users; that is modelled as a function of easily accessible features (see Figure 4.4), which enable a user to decide about reciprocating a *following request* on Twitter.

Features for the prediction model consist of a rich set of *meta-data information* describing users based on their behaviour and the *textual* part of their account description. This is essential because if the users comprising a potential dyad have conflicting ideologies expressed in the profile descriptions, the likelihood of a dyadic tie is minimal. Thus, the set of network and text features is given by $\chi = \{f_n, f_t\}$, where f_n is the *network features* consisting of *followers*, *friends*, *account category* and f_t is the textual features consisting of *account description* for training the prediction model. Among other intrinsic factors, these are the likely features a user can easily access in deciding to follow back a request or not. Other latent factors that could induce reciprocity have been ignored, and the assumption is that reciprocity is based on the available attributes identified in Figure 4.4.

Prediction Features & Results Presumably, the decision to reciprocate is correlated with the idea of *homophily*, hence the hypothesis that a high number of similarities between users could be a strong indicator of reciprocal ties. As in Ahn et al. (2010), where attributes' similarities were used for communities detection task, the probability of reciprocity between pairs $p(R_{v_i, v_j})$,

resulting from a feature, $f \in \chi$, is based on the *attributes similarities* in Figure 4.4(b). The problem of predicting the likelihood of a tie between any pair of users (i.e. *the possibility that user v_i who follows user v_j will be followed back*) is modelled as binary classification. This was achieved by building a deep learning classifier that predicts the probability of a dyadic tie between two users on Twitter, and then compares the results with actual dyads collected for evaluation. The predictor is based on a deep learning method that returns the likelihood of two users engaging in a pairwise relationship. In other words, given two users a and b or v_i and v_j connected with one edge connection, the goal is to predict whether a pairwise relationship will be established. The following vector of reciprocal relationships represents a user \mathbf{v}_i :

$$\mathbf{v}_r^i = [v_{i,j}, v_{i,k}, \dots, v_{i,n}]$$

where users $j, k \dots n$ have dyadic or reciprocal ties with the user \mathbf{v}_i . Features from the *account description text* are learned by applying a *Convolutional Neural Network (CNN)* or *convnet* (Kim 2014) on the *n-dimensional* embeddings of the tokens⁴ in the text. The *convnet* has been applied to various domains and many successful studies in *Natural Language Processing (NLP)* have used them (Kim 2014, Zhang et al. 2015, Wang 2017). The final output from the *convnet* is encoded using *Long Short-Term Memory (LSTM)* (Hochreiter & Schmidhuber 1997). See Figure 6.3 for a summarised process workflow. Figure 6.4 shows some results from the prediction

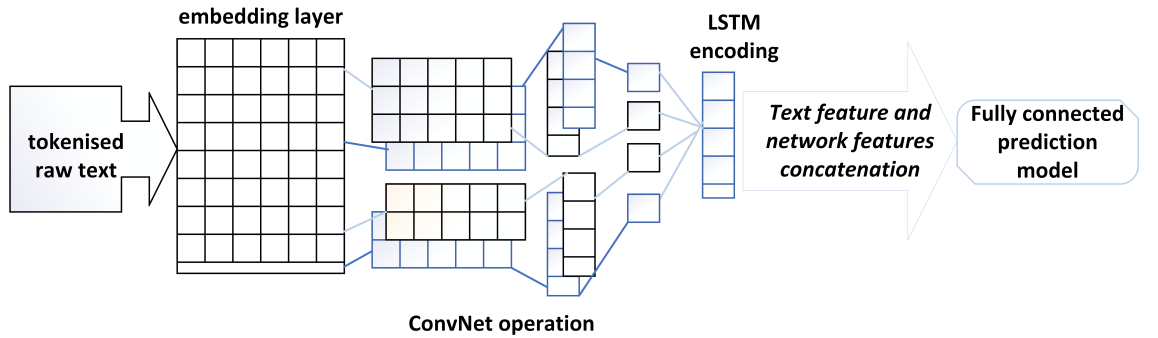


Figure 6.3: A simplified workflow in the dyad predictor. The *embedding layer* accepts tokenised text and encodes each token in a dense 100-dimensional vector to be used by the *ConvNet* part. The *LSTM* layer transforms the output to a lightweight vector that is merged with the *network features* for training.

model. Although the performance in *sub-figures a and b* is good, it is unstable and seems to be prone to overfitting, noting the proportional relationship between the training accuracy and the

⁴This thesis utilised Glove word embeddings (Pennington et al. 2014), pre-trained on tweet collections

validation loss, i.e. both are increasing. More training was conducted by increasing the training epochs to 200 and adding more layers to the network for stability (see *sub-figure c*). The funda-

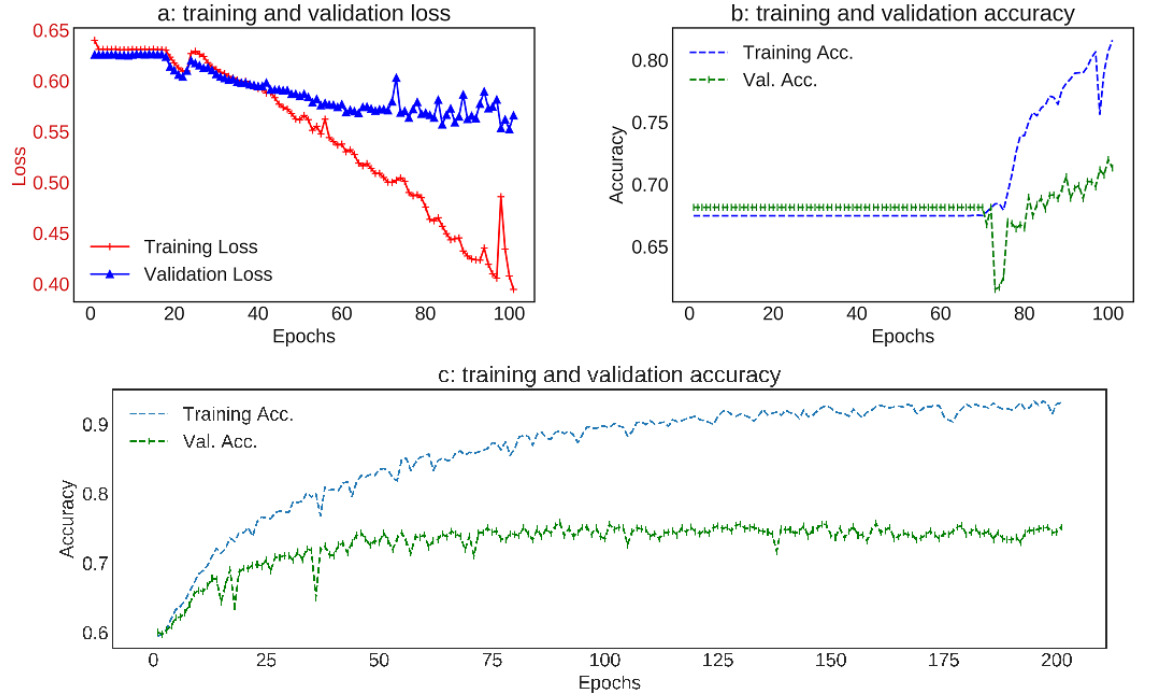


Figure 6.4: Performance of the proposed model on the training and the validation set. The performance remains stable after the first 100 epochs.

mental conclusion is that dyadic ties can be accurately predicted (if the pair of users are socially active) enabling the identification of low-level communities which offer a better reflection of true connectivity with strong social cohesion. The next section (Section 6.2.3) focuses on a scaled version of *dyadic ties* i.e. *transitive ties*.

6.2.3 Identifying Simmelian Ties

A *Simmelian tie* is a small group of interconnected users, which is synonymous to a *transitive tie*, in a network. A set of *transitive nodes* is regarded as a facilitator for the detection of socially cohesive communities and is based on the premise if the underlying mechanisms to predict transitivity on Twitter can be understood, tasks such as cohesive clustering and content validation could be greatly enhanced. Consequently, the following part of the *meta-analysis* focuses on the investigation of:

- i how to identify and quantify the proportion of *Simmelian ties* in a given collection?

- ii how to infer the latent variables making it possible for two or more users to establish reciprocal ties (*reciprocity effect*)?
- iii how to develop a *Simmelian tie* prediction model and quantify the uncertainties surrounding the prediction?

To address these questions, which could ultimately lead to an improved detection task, an empirical analysis of datasets consisting of *Simmelian ties*⁵ retrieved from over 30m Twitter accounts (see Table 6.2) is utilised in developing a learning model for the prediction.

Transitive Dataset Similar to the *dyadic dataset* collection criteria, *Algorithm search-dyads* (Algorithm 1) is used to retrieve empirical data from Twitter using an initial set of (*seed users*) from *verified* and *unverified* account categories. If m_{v_i} denotes the network of a user v_i , consisting of sets of followers fl_{v_i} and friends fr_{v_i} , a reciprocal tie exists between v_i and v_j if $p(R_{v_i, v_j}) \geq \tau$, otherwise directed tie. The set of reciprocal pairs is denoted by κ where $\kappa \in m_{v_i}$. According to Figure 1.2(d), if commonality exists between the nodes in the sets of networks of users (v_i, v_j, v_k) , then the search stops and transitive users spanning three generations – *parent* \mapsto *children* \mapsto *grandchildren* are found. Table 6.2 shows a summary of the *transitive dataset* consisting of information about *reciprocated ties* and *unreciprocated ties*. To aid evaluation, an additional set of publicly available data⁶ is used as a benchmark (termed ego-Twitter in Table 6.2). The dataset is the closest available to the research, and consists of directed ties only; hence, its applicability in this study is limited since it was utilised for prediction and comparison only.

Network Topology Having obtained the ground-truth data of transitive ties, a pragmatic approach is applied to examine the distribution of ties and compute relevant metrics. Figure 6.5 shows the *empirical cumulative distribution function (ECDF)* of relevant metrics across user categories in the dataset. In Figure 6.5, there is a higher proportion of reciprocal ties in the *unverified* users category, and a plausible reason for the low percentage of reciprocal relationships among the *verified* users can be likened to the reasons given in Cha et al. (2012) that such users are authorities or institutions, who rely on independent sources of information that are

⁵consisting of many *pairwise ties* and *transitive ties*

⁶obtained from the Stanford public data repository (Leskovec & Krevl 2014)

Table 6.2: A summary of the transitive datasets. C: Category; S: Seed Size; V: Visited users; P: Pairwise ties; T: Transitive ties; D: Search duration.

C	S	V	P	T	D(min.)
1:verified	1000	1832630	708	–	1122
2:verified	1990	3893075	2155	–	2247
3:verified	6803	14413641	1317	541	7965
1:unverified	1000	1793806	640	–	2162
2:unverified	2023	13409661	1834	–	4084
3:unverified	7121	32065133	2150	347	13071
ego-Twitter	81,306	–	–	–	–

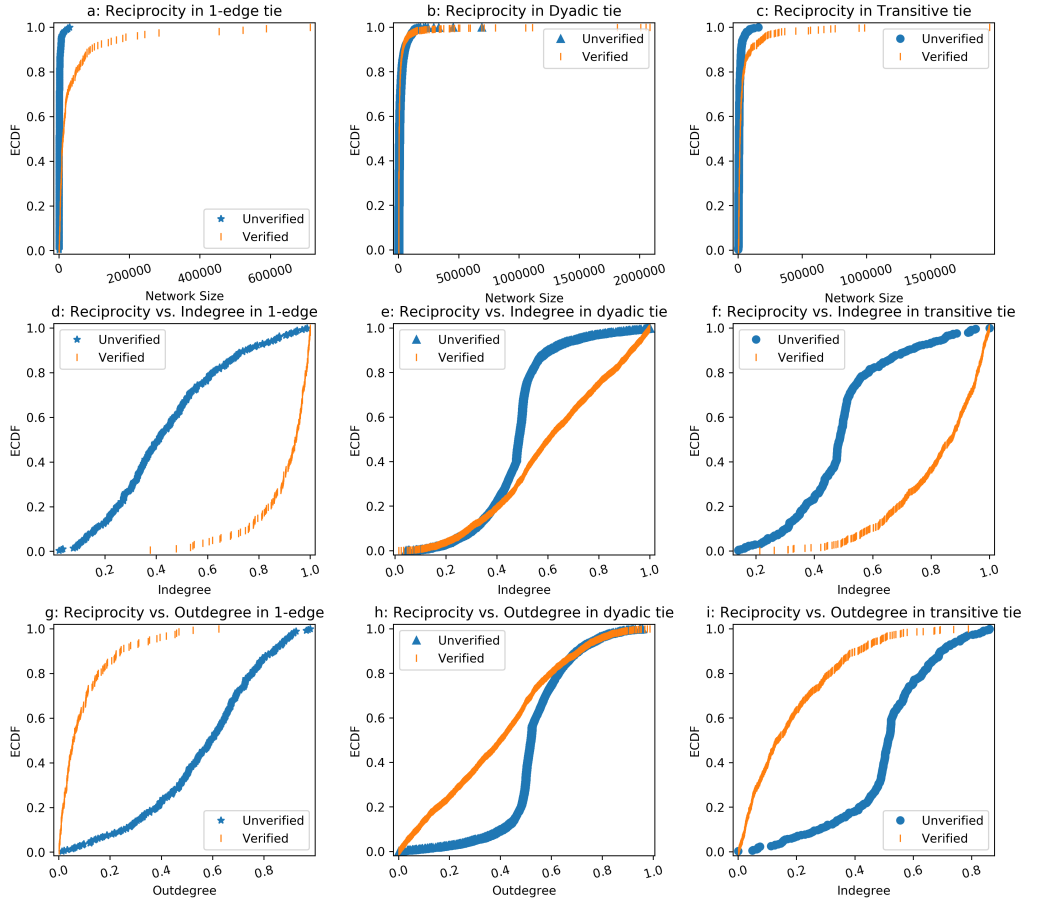


Figure 6.5: The *ECDF* of various types of connections and network sizes in the data. The network neighbours of *verified* users are higher, but the *unverified* counterparts show a higher proportion of reciprocal ties. The relatively high proportion of 1-edge in the network can be explained by many followers not being followed back on Twitter.

outside the network. In Figure 6.6, the percentage of reciprocity is higher among *unverified*. A manual check on some random samples from the *verified* category reveals a high proportion of reciprocity with other *verified users*. This can be attributed to trust, i.e. the identity of the users is known. Based on this observation, it can be assumed that the network size of a user is highly

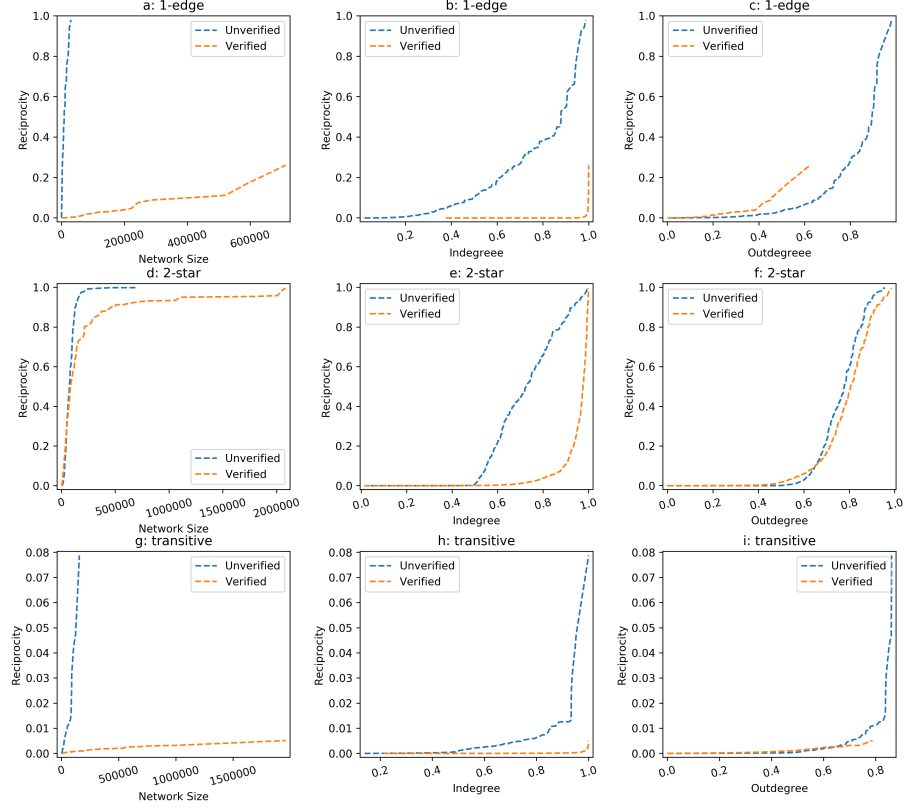


Figure 6.6: The effect of user’s attributes in enabling reciprocal ties. Both *indegree* and *outdegree* appear to be instrumental in enabling a high degree of reciprocity; this information provides useful insights about the effects of the user’s attribute in influencing reciprocity which can be used to inform the prediction model.

likely to grow if the user in the verified category and has a large number of followers; however, there will be a decreased likelihood of reciprocity. Conversely, there is an increased likelihood of reciprocity if the user is unverified and has a relatively large network size. Similarly, users in the *unverified category* are more likely to reciprocate a followership request and users with large network size (usually greater than 20k) have a low proportion of reciprocated ties. The majority of reciprocated links have network sizes below 20k.

Generative Process The ground-truth data is relatively small in size (see Table 6.2), which could hamper comprehensive analysis. However, the likelihood of establishing reciprocal ties among users can be defined as a generative process. In view of this, an approach based on Bayesian inference is employed to simulate the real data on a scale in a controlled setting. Accordingly, the focus is on (1) how possible is it to understand why some users have reciprocated ties, and some do not? (2) how to predict the likelihood of reciprocal ties among users and the associated uncertainties?

To answer the above-mentioned questions, the experiment begins by building the *Bayesian network* and applying a *variational inference* to assess the predicted posterior and associated uncertainties. The *variational inference* is a Probabilistic Programming paradigm that is designed to optimise a posterior distribution function. The implementation in this thesis is based on ADVI (Kucukelbir et al. 2015), which is available in *PYMC3*⁷. Moreover, a *log-linear model*, which relies on the relevant attributes of a user, is applied. A *log-linear model* is commonly used in problems involving probabilistic prediction (Karlis & Ntzoufras 2003, Baio & Blangiardo 2010) and the following *log-linear model* (Eq. 6.1), which is expressed as a linear combination of the user’s attributes is used to study the propensity of reciprocity.

$$y_{v_i} = \beta_{v_i} + \gamma_{c_{v_i}} + \epsilon_{v_i} \quad (6.1)$$

In Eq. 6.1, β_{v_i} , $\gamma_{c_{v_i}}$ and ϵ_{v_i} denote the mean reciprocity among users, mean reciprocity between users’ categories and error term respectively. With the setup in Eq. 6.1, it is possible to apply a conventional classification algorithm, as applied in Section 6.2.2, however, the goal is to incorporate prior knowledge about the problem. Furthermore, the dataset is relatively small, hence the need for a Bayesian approach.

The Inference Workflow and Parameters The parameters in Eq. 6.1 are treated as random variables specified by probability distribution functions $p(\cdot)$ consisting of a range of values making it possible to define relevant statistical quantities such as mean μ values of the parameters. Refer to Table 4.1 for a summary of the parameters and their respective distributions. Figure 6.7 shows the basic execution pipeline in the *Bayesian Inference* in which the final hypothesis (d) i.e. inference is the estimated underlying probability correct for finite trials, say n ,

⁷A toolkit of probabilistic programming in Python (Salvatier et al. 2016)

in experiment j , e.g. θ_j . Informed by previous knowledge about the data, the *prior* θ , and the *likelihood* $f(y|\theta, x)$, represent sets of variables that are likely to characterise the observed data. The assumption is that $\theta_v i$ comes from a probability distribution that describes the individual difference among users. The *posterior*, denoting the evidence in the data $p(\theta|\mathcal{D})$ or $p(\theta|y, x)$, is expressed as a function of the *likelihood* and the *prior* and is obtained based on *Bayes' rule* which entails updating beliefs about θ given the observed data \mathcal{D} .

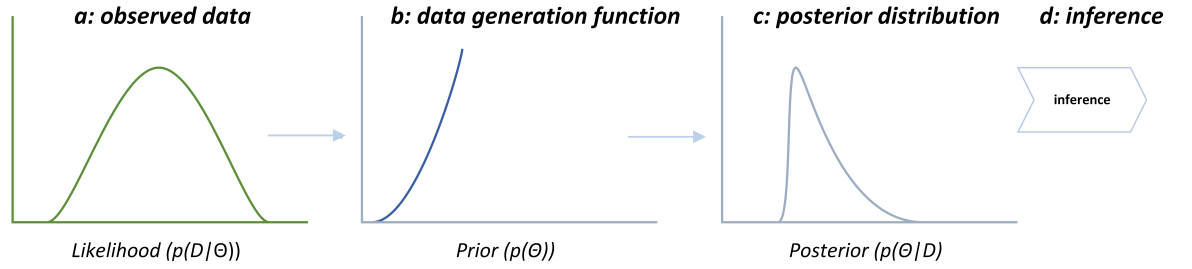


Figure 6.7: A simple workflow in a hierarchical model for Bayesian inference. Essentially, (a) the collected data/observations \mathcal{D} (b) is assumed to be generated by a function of a set of unknown variables denoted by θ to (c) compute posterior distribution for (d) making an inference.

The essence of the *linear model* (Eq. 6.1) is to enable the simulation of the observed data \mathcal{D} and generate a synthetic version $\hat{\mathcal{D}}$ indistinguishable from the observed data \mathcal{D} . In Figure 6.7, the data generation proceeds in the forward direction and the inference in the backward direction using the *linear model*. Finally, the inference (Figure 6.7(d)) involves backtracking to determine the parameter that produced the observed data points. Many algorithms for inference, such as the *maximum likelihood estimation* (Myung 2003) are used to estimate the parameter values that maximise the likelihood (given the observed data). The *PYMC3 toolkit* incorporates all the required dependencies for the analysis and is utilised for the implementation. Figure 6.8 and Figure 6.9 show data sampling and posterior distribution respectively.

In Figure 6.9, some of the users are measurable below the mean throughout the experiments, the *indegree*, for instance, is below the mean shown in the *meta-analysis* section. Although the mean is below the observed mean, it suggests that it is greater than chance, which will be useful in making credible assumptions about the data. For instance, it is possible to quantify the uncertainties in the data when making predictions, hence, enabling a well-informed decision.

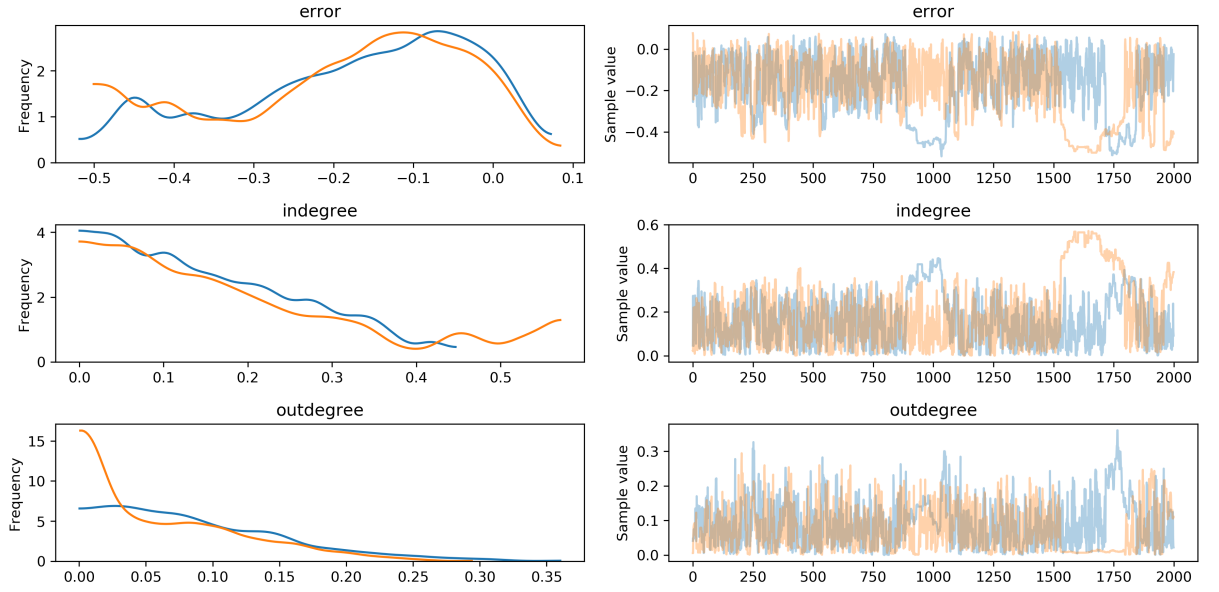


Figure 6.8: Due to the availability of a ground-truth, the values of *indegree* and *outdegree* can be estimated. Sampling results showing the *error term*, *indegree* and *outdegree* after two independent runs, hence, the two different lines in the *traceplots*. A total sample size of 2000 is used, and the probability is computed. Based on the sampled data, the *indegree* in reciprocal ties ranges in the region of 0 – 0.5 and *outdegree* ranges from 0 – 0.35; each with different frequencies. This means that in the evidenced data, high *indegree* is crucial in enabling reciprocity. The figure in the right serves as a diagnostic tool to inspect the samples drawn and the degree of correlation. Some of the samples are unstable as evidenced by the perturbations in the results in the second column.

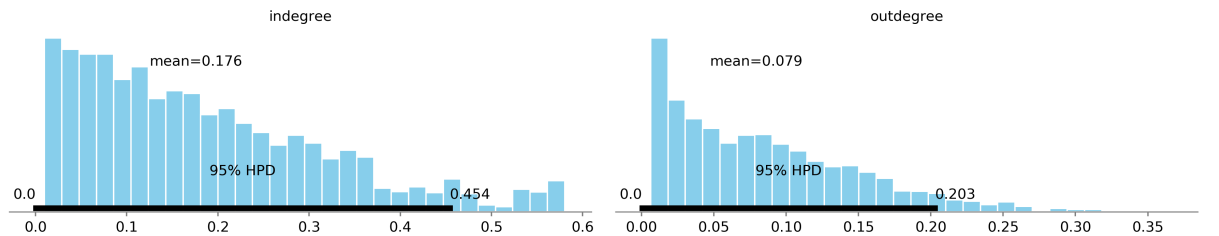


Figure 6.9: Some samples from the posterior distributions showing various mean values of the *indegree* and *outdegree*. Some of the values are measurable below the mean observed in the *meta-analysis* section, but it suggests that it is greater than chance, which will be useful in making a well-informed decision.

Utility of Structurally-Related Nodes

A characteristic feature of a network is the presence of a core-periphery pattern with a central group of closely related users. Usually, these users, acting as social bridges, are less connected to the core network and one another but play a significant role in connecting disparate parts of a community (Brass 1985). In this thesis, those users with a high proportion of reciprocated ties are referred to as *hop-skipppers*⁸, see Figure 6.10. By virtue of their centrality, the *hop-skipppers* provide the means, i.e. the local information required in detecting local communities or cohesive communities in a network. The *hop-skipppers* could be used to initiate the process of local community detection and enable a modular analysis of the network. For instance, based on the adage, *birds of a feather flock together*, users who spread rumours or spam are likely to be structurally-connected.

Moreover, a user with many reciprocal ties would be a resourceful representation of users with strong social cohesion. From the perspective of content integrity, a small group of users with reciprocal ties provides a useful means for analysing user groups with common online traits. According to Granovetter (1977), the hypothesis that allows a strong reciprocal tie ($a \iff b$) is given by: *the stronger the tie between a and b, the larger the proportion of entities to whom they will both be tied*, i.e. connected by a weak tie or strong tie. Referring back to Figure 4.4(b), there is less overlap in friendship circles if the tie between a and b is *non-existent*; *intermediate* if it is weak and *most* when it is strong. Similarly, if strong ties connect $a \leftrightarrow b$ and $a \leftrightarrow c$, both c and b , are similar to a , hence the likelihood of friendship increases once they meet. It has been suggested that if strong ties exist among three users, anything other than a positive tie will lead to a *psychological strain* (Newcomb 1978), and increase the likelihood of losing a third-party relationship (Brass 1985). Harnessing these insights will play a crucial role in restraining users in a small group to engage in demeaning behaviour or uncivilised attitudes, hence improving the quality of the content posted by members. The experiment using a collection of *Simmelian ties* exhibit useful behaviours such as connecting large groups of users or acting as network bridges on Twitter; this observation has potential relevance in tasks such as *clustering*, *content veracity* and *information diffusion*.

⁸The term *hop-skipppers* (à la Jacobs (1992)) is used to denote users with a large number of reciprocal ties on Twitter

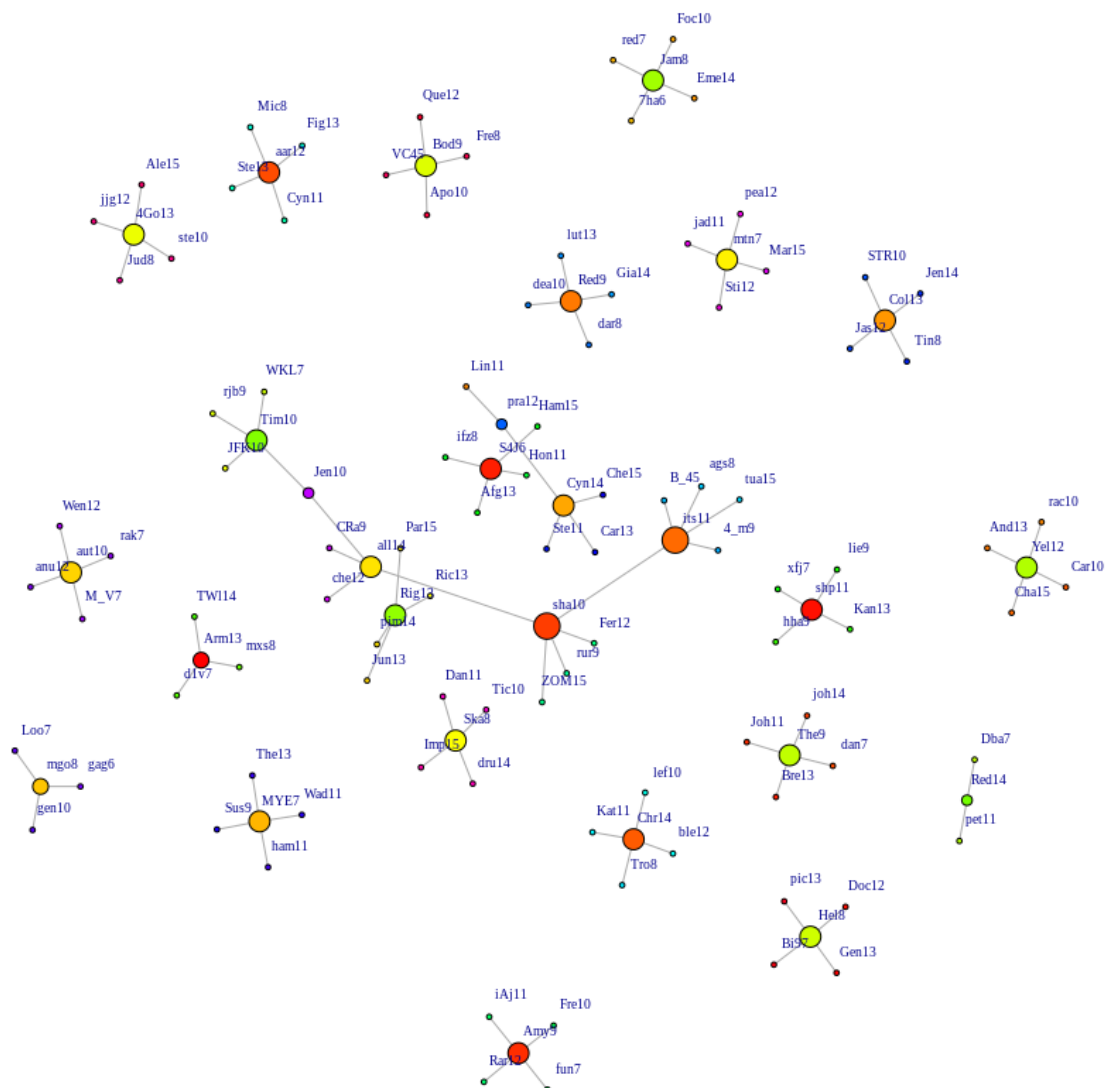


Figure 6.10: Example of users with reciprocated ties. For brevity, the actual network size of each user has been truncated; however, the size of each node in the figure reflects the user's network size. The names of the users are anonymised by retaining only the first three letters of each screen-name and attaching its length as a postfix in order to preserve identity.

6.3 Summary

The importance of a small group of users with a positive relationship has been recognised as a critical feature in the structural analysis of network (Freeman 1996) and is more intimate, with a high degree of familiarity due to strong social cohesion (Dunbar 1998). Moreover, anecdotal and cognitive evidence suggests that people are more likely to believe information from closely related individuals Carley (1991). A *Simmelian triad* consists of a small cohesive group that reflects a personal network on Twitter, which is homogeneous concerning many sociodemographic behavioural, and intrapersonal characteristics (Miller McPherson et al. 2001). Through the experiments and analyses presented in this chapter, the following have been recognised and are useful in the final detection framework (see Chapter VII).

- i availability of a large scale dataset consisting of reciprocal and transitive ties for various evaluations, i.e. a reliable, independent sets of ground-truth structural data;
- ii an effective prediction model that circumnavigates the difficult and time-consuming approach to finding *state-type ties*. This ensures efficiency and mitigates the *curse of dimensionality* that could result from manual profiling of reciprocal ties on Twitter;
- iii establishing the relevance and applicability of *Simmelian ties* in enhancing clustering, information diffusion and as an effective means to reach out to socially cohesive groups of users.

To fully harness the insights gained in this chapter, the final prediction framework in the subsequent chapter (Chapter VII) provides a detailed implementation.

Chapter VII

DETECTION OF MICROCOSMS

7.1 Introduction

The illustration in Figure 1.2 suggests that communities on Twitter could be formed based on many factors, which affect the detection of *microcosms*. In a clustering task, the role of a scoring function (Section 3.2.2) is to identify pairs of objects which are, in some respect, closely related. In the *MCT framework* (Section 4.3), a *joint scoring function*, based on *structural* and *textual* similarities, is used to assign nodes to relevant clusters as exemplified in Figure 4.2. With this background insight, this chapter builds on the previous one (Chapter VI) to develop the framework for *microcosms* detection on Twitter.

7.2 Structurally-Related Clusters

Social equivalence, which could be *structural* or *regular*, reveals that individuals compare themselves with one another, and adopt similar attitudes and behaviours of those who occupy an equivalent position in an organisation (Brass et al. 1998). The *structural equivalence* refers to two actors having similar interaction partners, which can be mapped to a form of *state-type tie* to infer structurally-related nodes based on relevant attributes. In the same vein, a *regular equivalence* refers to similar patterns of interactions even though the interaction partners may be entirely different (Scott 1988). In the context of this research, *regular equivalence* is analysed based on *retweets* and other forms of interactions on Twitter as depicted in Figure 1.2. Both of these aspects are explored in the research to enable detection of *microcosms*. Accordingly, the *structural aspect* in the *MCT* exploits the idea of *homophily* and *structural equivalence*.

7.2.1 Network and Reciprocal Communities

One of the goals in the *MCT strategy* is to offer a compact way to represent and find a co-occurring group of users, which will make it possible to explore both the local and global clustering requirements. The insights gained from Chapter VI, specifically Figure 6.5, reveal that network entities can be grouped on the basis of *network size*, defined by the proportion of *indegree*; this is to introduce a high-level structure that will enable more granular detection.

Network-Communities

For a given network data \mathcal{D} , there exists sets of *high-level communities of nodes* prompted by a measure of relatedness, which can be explained using *social homophily*. For instance, the status of a user's profile vis-à-vis *followers count* can be used as a proxy for social status, such that each user can be categorised accordingly. It follows that there exists a set of a *high-level communities of nodes* induced by the *network size*, such that the status of a user's profile makes it possible to categorise users. In this thesis, users with network size of up to 500 are grouped together, then up to 1000 and the grouping continues at an incremental pace of a thousand network size until the highest size is reached. Thus, a high-level *network-communities* \mathcal{C}_{nc} , are formed given by:

$$\mathcal{C}_{nc} = \mathcal{C}_{nc_1}, \mathcal{C}_{nc_2}, \dots, \mathcal{C}_{nc_h} \subset \mathcal{D}$$

and are represented by an $n \times p$ matrix ($\mathcal{M}_{\mathcal{C}_{nc}}^{n \times p}$) where n and p denote the number of *users* and number of *network-communities* respectively.

7.2.2 Nodes Reciprocity

In addition to the network-communities \mathcal{C}_{nc} , which are part of the structurally-related communities, the following *reciprocal-communities* \mathcal{C}_{rc} (or $\mathcal{C}_{rc} = \mathcal{C}_{rc_1}, \mathcal{C}_{rc_2}, \dots, \mathcal{C}_{rc_j} \subset \mathcal{D}$), which are based on the prediction of reciprocal ties between pairs of nodes, are defined. The reciprocal-communities are based on the following. Given any pair of nodes $v_i, v_j \in \mathcal{V}$, the goal is to find *the likelihood of establishing a reciprocal tie* using relevant information about the nodes. Inspired by the work of Ahn et al. (2010), where attributes' similarities were used for community tasks, the probability of reciprocity resulting from corresponding features of nodes is based on a high degree of *attributes' similarities*. Referring back to the illustration in Figure 4.4(b), which

shows possible attributes that induce friendships, other latent factors also play a role; however, the formulation in this thesis is based on the assumption that reciprocity is only based on the available attributes identified in the figure. For the prediction, consider the sets of nodes \mathcal{V} and edges \mathcal{E} , the likelihood of *reciprocity* that could lead to identifying reciprocal units, *dyads* or *transitive* (see Section 6.2.1), $p(R_{v_i, v_j}) \forall v_i, v_j \in \mathcal{V}$ is described by Eq. 7.1 through Eq. 7.3. In Chapter IV, the set of all possible attributes or features \mathcal{A}_f , can be extracted from the node's profile and compare with other nodes in the network. The subset of the features¹ $\mathcal{X}_f \subset \mathcal{A}_f$, for making the comparison consists of easily accessible attributes that enable a quick decision about reciprocity, given by $\{ind, out, cat\} \subset \mathcal{X}_f \subset \mathcal{A}_f$. It follows that, for a pair of nodes v_i, v_j , their corresponding features are given by:

$$\mathcal{X}_{f_{v_i}} = \{ind_{v_i}, out_{v_i}, cat_{v_i}\}, \quad \mathcal{X}_{f_{v_j}} = \{ind_{v_j}, out_{v_j}, cat_{v_j}\}$$

The ratio of the corresponding attributes, e.g. *ind* or *out*, between pairs is a real value quantity expressed as:

$$\frac{ind_{v_i}}{ind_{v_j}} \in \mathbb{R} \quad \forall f \in \mathcal{X}_{f_{v_i, v_j}}$$

If the above computation evaluates to a value within $[0.75, 1.25]$, the pairs are assumed to have similar attributes (1), or dissimilar attributes (0); this interval is to allow extra freedom for minor discrepancies between the corresponding features. For instance, if the ratio equals 1.0, the pairs have precisely similar attributes which is useful in analysing aspects of *homophily* and *social equivalence*. The binary values from the comparison of corresponding attributes or features are used to compute the overall similarity between pairs using *Jaccard Similarity Coefficient*, J (Eq. 7.1) given by:

$$J(\mathcal{X}_{f_{v_i}}, \mathcal{X}_{f_{v_j}}) = \frac{|\mathcal{X}_{f_{v_i}} \cap \mathcal{X}_{f_{v_j}}|}{|\mathcal{X}_{f_{v_i}} \cup \mathcal{X}_{f_{v_j}}|} \quad (7.1)$$

The formal process is given by *Algorithm f-sim* (Algorithm 2).

Reciprocity & Constant Error Term

The formulation of the reciprocity prediction is associated with the following observation. The response to a *friendship* request is either *yes* (reciprocate) or *no* (do not reciprocate), and is as-

¹For brevity, the features *indegree*, *outdegree*, *category* are trimmed to *ind*, *out*, *cat* respectively. See Section 4.4.1 for details.

sociated with a decision error, which is modelled using a *probabilistic preference model* or *response probability*. The response probability model aims to capture various scenarios in which an actor is offered a set of features, and the decision process is associated with a constant probability of making an error in the choice (Marley & Regenwetter 2016). The model enables the mapping of each possible *response* into a probabilistic space, and utilises the *constant* or the *trembling hand* error, which assigns a constant value to a choice probability. The *error term* ζ , associated with each likelihood of reciprocity, is based on the assumption that there is a 50 – 50 chance of reciprocity or otherwise, between any pairs in the network. The error is associated with the weight or contribution of each attribute shown in Figure 4.4(b) towards influencing the overall response decision. The influence or reciprocity effect of each feature R_f is mapped to a value in the interval $[0 - 1]$ given by $R_f \in [0, 1]^{|\mathcal{X}_f|}$, where $f \in \mathcal{X}_f$ and $|\mathcal{X}_f|$ is the number or length of a set of features \mathcal{X}_f . For instance, $ind \mapsto 0.1$ and $out \mapsto 0.02$ denote two features and corresponding influences. Figure 6.6 and the previous *Bayesian analyses* (Section 6.2.3) offer useful insights in this respect. Through the degree of similarities between the corresponding features (computed using Eq. 7.1), it is possible to improve the prediction by expressing the error term as a function of the similarity index $J(v_i, v_j)$, between pairs. Consequently, the prediction error ϵ_{v_i, v_j} (Eq. 7.2), and the similarity index (Eq. 7.1), are expressed in such a way that the prediction is within a practical significance range that closely match a realistic prediction using the following relation:

$$\epsilon_{v_i, v_j} = \frac{1}{\zeta \times (1 + \log(J(v_i, v_j) + \zeta))} \quad (7.2)$$

A constant value of $1/3$ is assigned to ζ , which corresponds to the *constant error term*, thus the final relation is given by:

$$p(R_{v_i, v_j}) = \frac{1}{1 + \exp \varphi} \quad (7.3)$$

where

$$\varphi = -\log(\epsilon_{v_i, v_j} + J(v_i, v_j)) \times (\epsilon_{v_i, v_j} + J(v_i, v_j))$$

With Eq. 7.3, it is possible to compute the probability of reciprocity between any pairs of nodes given their corresponding features (as described in the forthcoming section). The prediction of reciprocity makes it possible to identify as many sets of nodes as possible with a high likelihood of establishing reciprocal ties, thus adding a layer of social cohesion to the *MCT framework*.

It follows that the likelihood of a reciprocal tie between any pair of users can be expressed as follows:

$$L(R_{v_i, v_j}) = 1 - \prod_{f \in \chi_f} (1 - p(R_{v_i, v_j})) \quad (7.4)$$

The relation $L(R_{v_i, v_j})$ (7.4) can be viewed as a generative process where $p(R_{v_i, v_j})$ predicts the reciprocity between pairs of nodes using the marginal reciprocity effect of each feature $f \in \chi_f$. Algorithm 2 formalises this implementation. It follows that \mathcal{S}_r is a collection of nodes with a

Algorithm 2 : *Algorithm f-sim* returns the likelihood of reciprocity between pairs

```

1: Initialisation:  $\{\}$   $\leftarrow \mathcal{S}_r$ ;  $\{\}$   $\leftarrow \mathcal{S}_u$ 
2: Input: a finite collection of network data  $\mathcal{D}$ 
3: while  $\mathcal{D} \neq \emptyset$  do
4:    $\forall v_i, v_j \in \mathcal{D}$ , compute  $p(R_{v_i, v_j})$  using Eq. 7.3  $\triangleright v_i \neq v_j$ 
5:   if  $p(R_{v_i, v_j}) \geq \tau$  then  $\triangleright \tau$ , a predefined threshold
6:      $\mathcal{S}_r \leftarrow (v_i, v_j)$   $\triangleright$  structurally-related
7:   else
8:      $\mathcal{S}_u \leftarrow (v_i, v_j)$   $\triangleright$  structurally-unrelated
9:   end if
10: end while
11: Output:
12:  $\mathcal{S}_r, \mathcal{S}_u, \mathcal{M}_{A_{i,j}}^{n \times n}$   $\triangleright$  adjacency matrix  $\mathcal{M}_{A_{i,j}}^{n \times n}$ 

```

high *structural similarity*, i.e $\forall v_i \in \mathcal{S}_r \exists v_j : p(R_{v_i, v_j}) \geq \tau$.

What does it mean for nodes to be structurally-related? The posed question is answered using the following example. Consider the set \mathcal{V} consisting of 13 nodes v_1 through v_{13} , i.e $\{v_1, v_2, v_3, \dots, v_{13}\} \in \mathcal{V}$. After executing *Algorithm f-sim* (Algorithm 2), the following 7 pairs of nodes are *structurally-similar* or *related*²:

$$v_1 \sim v_2, v_1 \sim v_3, v_1 \sim v_5, v_2 \sim v_4, v_2 \sim v_5, v_2 \sim v_9, v_3 \sim v_{11}$$

Therefore, three different communities, also known as the *reciprocal communities* \mathcal{C}_{rc} (see Section 7.2.2), can be identified as follows:

$$c_{rc1} = \{v_1, v_2, v_3, v_5\}, c_{rc2} = \{v_2, v_4, v_5, v_9\}, c_{rc3} = \{v_3, v_9\}$$

²The symbol ' \sim ' is used to denote structural similarity between pairs

These communities are not discrete, but consist of overlapping entities. A cursory glance may present a problem by suggesting that any overlapping nodes should be in the same cluster since they have similar structure. However, noting that the degree of freedom in deciding similarity or otherwise spans a wide range of values within 0.75 to 1.25, the overlapping nodes could be due to having values in the extremum of the range, e.g. $v_1 \sim v_2$ evaluates to 0.75 and $v_2 \sim v_4$ evaluates to 1.25, hence belonging to different communities. A similar line of reasoning applies to the *textually-related nodes* (Section 7.3).

7.2.3 Spectral Clustering

Because the collection of the *structurally-related nodes* (\mathcal{S}_r), can be easily transformed into a graph structure according to the affiliation of nodes to *network-communities* or *reciprocal-communities*, a *spectral clustering* is applied to identify local structures or clusters. By leveraging the detected local structures, it is possible to analyse many aspects of *sociometry* such as *centrality measures* (Section 2.3.2). Referring to Figure 6.10, in which nodes with many reciprocal ties are shown, the use of spectral clustering should be able to avail information about the network structure in the data since it is practically infeasible to visualise the network as the number of nodes increases. Hence, the use of spectral clustering, which compares the clusters discovered based on the sets of *empirical*, *predicted* and *random data* (see Table 6.2). The description given in Section 4.3 shows that the spectral clustering involves a series of operations ranging from the construction of an adjacency or affinity matrix to clustering in a reduced dimension. The essential stages in the spectral clustering include:

- i The *preprocessing step* involves the construction of the *adjacency matrix*, *similarity graph* and the *degree matrix*. The *similarity graph* can be defined either using ϵ – *neighbourhood* or k –*nearest neighbour*. In the ϵ – *neighbourhood*, each vertex in the network is connected to vertices that are within a predefined threshold value ϵ^3 . In the k –*nearest neighbour*, each vertex is connected to the nearest vertices (k –*nearest neighbours*). Both the ϵ and k are hyper-parameters that need to be tuned according to the problem to help in capturing the local connectivity among vertices. The approach in this thesis is similar to the k –*nearest neighbour*, since the similarity score, inspired by *homophily*, is based on nodes' attributes

³For instance, using the radius of a sized-ball in which vertices within the radius are considered to be related

(see Figure 4.4) to predict reciprocity. Accordingly, the *adjacency matrix* \mathcal{M}_A (Eq. 7.5), is created such that the entries consist of the predicted connectivity between pairs of nodes. Each entry is expressed as a binary response to represent whether a given node is connected to another node (1) or not connected (0) (represented by the *row* and *column* indices).

$$\mathcal{M}_{A_{i,j}} = \begin{cases} 1 & \text{if } p(R_{v_i, v_j}) \geq \tau \\ 0 & \text{if } otherwise \end{cases} \quad (7.5)$$

The *degree matrix* \mathcal{M}_D (Eq. 7.6), is a diagonal matrix obtained by summing the entries in the adjacency matrix (Eq. 7.5) across the rows, in which the entry i, i denotes the degree of each node⁴. Thus, each entry in the diagonal d_i , of matrix \mathcal{M}_D is defined by:

$$d_i = \sum_{\{j | (i,j) \in \mathcal{E}\}} p(R_{v_i, v_j}) \geq \tau \quad (7.6)$$

- ii The *decomposition and clustering stages* entail the construction of a *Laplacian matrix* and associated *eigenvectors and eigenvalues*. The *Laplacian matrix* \mathcal{M}_L , is obtained by subtracting the *adjacency matrix* \mathcal{M}_A , from the *degree matrix* \mathcal{M}_D :

$$\mathcal{M}_L = \mathcal{M}_D - \mathcal{M}_A$$

Because of the special property of *eigenvectors*, by remaining unchanged after undergoing matrix transformation, they have a wide range of applications. Given the matrix \mathcal{M}_L , if there exists a vector x and a scalar quantity λ , if the transformation $\mathcal{M}_L x = \lambda x$ holds, then x is an *eigenvector* and λ the *eigenvalue* of \mathcal{M}_L . In the spectral clustering, the *eigenvectors* of the *Laplacian matrix* \mathcal{M}_L , represents the defining features for vertices in the data and are utilised for clustering. The entries in \mathcal{M}_L are defined by the following (Eq. 7.7):

$$\mathcal{M}_{L_{i,j}} = \begin{cases} d_i & \text{if } i = j \\ -p(R_{v_i, v_j}) & \text{if } p(R_{v_i, v_j}) \geq \tau \quad \triangleright \text{edge}(i, j) \text{ exists} \\ 0 & \text{if } otherwise \end{cases} \quad (7.7)$$

In Eq. 7.7, a negative entry ($-p(R_{v_i, v_j})$) signifies an edge between pair of nodes. The entries

⁴or the number of edges incident on the node or connected to the node

in the main diagonal of the matrix denote the degree of nodes d_i (defined by Eq. 7.6) and other entries remain zero as long as no connection or edge exists between them.

7.2.4 Modelling Structural Communities

The high-level grouping of nodes into *network-communities* and *reciprocal-communities* (Section 7.2.1) makes it possible to study relevant phenomena at local level. For example, information about the reciprocal-communities is expressed as a matrix of *users* vs *communities*, in which the entries in the matrix correspond to the strength in reciprocity between pairs of users. Thus, a local community structure based on the similarity values in the range of 0.75 to 1.25 are formed; hence, each node is associated with a local community, and each high-level network community is associated with local communities. Building on this high-level grouping of nodes, sets of hidden local communities exists, which can be obtained by a matrix decomposition technique. In this thesis, the *network-communities* ($\mathcal{M}_{C_{nc}}^{n \times p}$) matrix is decomposed into its approximate constituents (Eq. 7.8):

$$\mathcal{M}_{C_{nc}} \approx \mathcal{M}_{C_{rc}} \mathcal{M}_{C_{nr}}^T \quad (7.8)$$

The following matrices denote interactions information and the corresponding dimensions in Eq. 7.8:

- $\mathcal{M}_{c_{nc}} \mapsto n \times p$: a collection of nodes according to network sizes
- $\mathcal{M}_{c_{rc}} \mapsto n \times k$: a collection of nodes according to reciprocal-communities
- $\mathcal{M}_{c_{nr}} \mapsto p \times k$: a matrix of high-level communities and local communities

Due to the assumption that hidden communities exist in the data, Eq. 7.8 is solved by applying a decomposition technique in iterative fashion.

Optimisation of Structural Communities

Because interpretability is a desired requirement in the *MCT framework*, the *modelling of structural communities* follows a *non-negative matrix factorisation (NMF)* scheme (Lee & Seung 1999). The *NMF* is a highly interpretable type of matrix factorisation in which *non-negative*

constraints are imposed on the optimisation parameters (Aggarwal 2018). Essentially, the goal is to optimise Eq. 7.9.

$$\min_{\mathcal{M}_{c_{rc}}, \mathcal{M}_{c_{nr}}} \|\mathcal{M}_{c_{nc}} - \mathcal{M}_{c_{rc}} \mathcal{M}_{c_{nr}}^T\|_F^2 \quad \text{subject to } \mathcal{M}_{c_{rc}}, \mathcal{M}_{c_{nr}} \geq 0 \quad (7.9)$$

For simplicity, the following conventions are used to represent the parameters in Eq. 7.9:

$$\mathcal{M}_{c_{nc}} \mapsto D; \quad \mathcal{M}_{c_{rc}} \mapsto P \equiv [p_{is}]; \quad \mathcal{M}_{c_{nr}} \mapsto Q \equiv [q_{js}]$$

The process of obtaining a satisfactory approximation of the factored matrix ($\mathcal{M}_{c_{nc}}$ or D) requires training based on an iterative update, in which the object is to minimise the error margin between the original matrix (D) and its constituents (P, Q). A successful iterative update ensures that the underlying matrices exhibit strong correlations among their different entries. The formulation of the problem in Eq. 7.9 makes it a suitable candidate for optimisation under constraints, which is solved in a *Lagrangian* manner to optimise the squared *Frobenius norm* ($\|\cdot\|_F^2$) of the matrix. This thesis utilises a more liberal method based on the *Lagrangian relaxation* to optimise the parameters in Eq. 7.9. Consequently, the non-negative constraints in the NMF scheme are relaxed by introducing a new set of parameters (α and β), known as the *Lagrangian multipliers* with values ≤ 0 , to the corresponding entries of the optimisation parameters (P, Q). In response to the effect of the additional parameters (α, β) induced by the *Lagrangian relaxation*, the objective function M_{sr} is given by:

$$M_{sr} = \|D - PQ^T\|_F^2 + \sum_{i=1}^n \sum_{s=1}^k p_{is} \alpha_{is} + \sum_{j=1}^p \sum_{s=1}^k q_{js} \beta_{js} \quad (7.10)$$

As a form of a *minmax problem*, Eq. 7.10 requires a simultaneous *minimisation* of M_{sr} over P, Q and the Lagrangian multipliers (α and β). Similarly, a *maximisation* over all applicable values of the α and β is required⁵. To solve the optimisation problem, the process begins with computing the gradient of the Lagrangian relaxation with respect to the first aspect of the *min-max* (i.e. minimisation) optimisation variables. Although the introduction of α and β offers a degree of flexibility (which comes with a cost), achieving the optimal solution requires the optimisation condition to be based on P, Q only. To eliminate the introduced *Lagrangian multipliers*, a handy approach based on *Karush-Kuhn-Tucker (KKT) optimality condition* (Bertsekas

⁵The maximisation is needed because they are initialised with negative values

1997), which suggests that $p_{is}\alpha_{is} = 0$ and $q_{js}\beta_{js} = 0$, is applied. After simplification of the equations, the iterative update is initiated by assigning non-negative random values in $(0, 1]$ for the optimisation parameters (P, Q) according to the following update rule:

$$p_{is} \leftarrow \frac{(DQ)_{is}p_{is}}{(PQ^TQ)_{is}} \quad q_{js} \leftarrow \frac{D^T(P)_{js}q_{js}}{(D^TP)_{js}} \quad \forall i, \in 1, \dots, n, j, \in 1, \dots, n, s \in 1, \dots, k$$

The process of updating P, Q involves comparing their values to the original matrix D , and the goal is to minimise the difference or *error*. The iterative update of the parameters (p_{is} and q_{js}) continues until convergence. See Section 1.4.2 in Appendix A for relevant details.

Reciprocal Communities

The approach taken in Eq. 7.10 is based on a matrix factorisation, which poses a challenges with respect to exact or one-one mapping to the *textually-related clusters* (\mathcal{T}_r), i.e. ensuring, where applicable, that nodes in the same cluster share both \mathcal{S}_r and \mathcal{T}_r clusters as illustrated in Figure 1.1. Establishing the local mapping for each clusters will further improve the cohesion magnitude in the detected communities. What is known is that the two are related at a higher level since $\mathcal{T}_r \subset \mathcal{S}_r$, but the details about the shared clusters is not fully established. To address this challenge, the following approach based on the similarity of nodes is used. The similarity of nodes are represented in an affinity matrix of nodes ($\mathcal{M}_{va}^{n \times n}$) in which the magnitude of similarity between pairs represent entries in the matrix. In view of this, given a collection of nodes, the following *nodes-reciprocal communities* (\mathcal{C}_{vr}) are used according to the similarities:

$$\mathcal{C}_{vr} = c_{vr_1}, c_{vr_2}, \dots, c_{vr_k} \subset \mathcal{D}$$

and are represented by an $n \times k$ matrix ($\mathcal{M}_{C_{vr}}^{n \times k}$) where n and k denote the number of *users* and number of *reciprocal-communities* respectively. Membership of any of the clusters in \mathcal{C}_{vr} is qualified by Eq. 7.1 through Eq. 7.3, and because each $c_{vr_i} \in \mathcal{C}_{vr}$ is associated with parameters that determine membership, nodes in the network will eventually be clustered accordingly. It follows that the probability of forming a tie is higher among nodes in the same clusters.

Once a sufficient number of nodes with a high likelihood of reciprocity have been identified, the *structural aspects* exploits the *reciprocity potentials* to identify local community structure induced by network size. The next activity in the detection pipeline is to extract relevant textual

data from those nodes. Section 7.3 describes the process of identifying *textually-related nodes* and their mapping to the corresponding *structurally-related nodes* where applicable.

7.3 Textually-Related Clusters

Due to the relatively high generation rate of content in dynamic networks such as Twitter, where a user may generate a high amount of content within a short span, a fixed number of texts is extracted from each node for comparison. This is in response to the limitations of a single tweet in conveying sufficient information about a topic. Moreover, it is a challenge to compare the context of discussions within tweets owing to their high numbers and relatively small sizes. The goal in this respect is to understand the specific topics being discussed and the degree of similarities among documents. Identifying *textually-related clusters* is a form of *document-pivot clustering* in which weights are assigned to features in the document according to a weighting scheme (Allan et al. 1998, Yang 2001, Brants et al. 2003, Fung et al. 2005).

7.3.1 Identifying Similar Content

The process of identifying textually similar content or textually-related nodes \mathcal{T}_r begins by aggregating a finite collection of *textual content* \mathcal{T} , from each node v_i described by the following set of features $\{set\ of\ texts, network\ features, auxiliary\ features\}$, to compute similarity. Given a stream of texts $t_1, t_2, t_3, \dots, t_n \in \mathcal{T}$, each $t_i \in \mathcal{T}$ consists of *n-gram features*⁶ given by:

$$f_{i1}, f_{i2}, f_{i3}, \dots, f_{im} \in t_i \in \mathcal{T}$$

Then, for each node v_i in the *structurally-related nodes* \mathcal{S}_r , a finite number k , of texts making a corpus, is considered:

$$\mathcal{T}_{v_i} = \{t_{i,1}, t_{i,2}, t_{i,3}, \dots, t_{i,k}\}$$

The aggregated texts are analysed for clustering and assessing other relevant metrics. Identifying *textually-related nodes* \mathcal{T}_r is based on applying a *topic modelling technique* to compare the similarities between the topics of discussions of the nodes within a given time-frame. The *topic model* utilised is *Latent Dirichlet Allocation (LDA)*, which has been applied for various tasks (Airoldi et al. 2008, Yan et al. 2013, Yali et al. 2014). The *LDA* is a useful probability model

⁶ n could be any positive integer, e.g 1, 2, 3 for *unigram*, *bigram* and *trigram* respectively.

of a corpus that assigns a high probability to members of a corpus and other similar documents, which makes its use appropriate. In the *LDA*, documents are represented as random mixtures over latent topics in which each topic is characterised by a distribution over words (Blei et al. 2003). As a probabilistic generative model, the *LDA* assigns the *word distributions* to *topics* and *topic distributions* to *documents* in a corpus. In this thesis, the sets of texts collected from each node v_i , define a corpus \mathcal{T}_{v_i} , in which the overall theme in the corpus is analysed for comparison with other nodes in the network. Through aggregating *textual content*, each node has a unique fingerprint which can be used for making an in-depth comparison.

Similar Documents

The collections of aggregated texts from any pair of nodes $\mathcal{T}_{v_i}, \mathcal{T}_{v_j} \in \mathcal{T}_r \subset \mathcal{S}_r$, are used to extract relevant *features* or *shingles* and transformation to numeric according to the *tfidf* weighting scheme (Salton & Buckley 1988). The corpus from each node is trained such that each document in the corpus has a finite distribution over all topics, and all topics have distributions over all words. It is the distribution of each document that is used to compare with other documents for similarity. For each corpus or set of documents generated by each node, the *LDA* model produces the corresponding topics which can be compared with other documents. The distribution of each document or *anchor document* is compared with other documents in the corpus to return the most similar documents to it. Geometrically, the *LDA* space is a *simplex* in which each document is positioned according to relatedness to topics – the closer a document is to the corner of the topic, the more similar.

Given that *LDA-based* comparison relies on the probability distributions of documents, the *Jensen-Shannon Divergence (JSD)* is used to measure the distance between topical themes. The *Jensen-Shannon* is a suitable statistical metric to measure the similarity between the documents based on their distributions in which divergence in the distributions is used to assess similarity. Unlike the asymmetric *Kullback-Leibler Divergence*, the *Jensen-Shannon* is symmetrical, which is crucial since the task of comparing two documents should be the same irrespective of the order (e.g., $a \mapsto b$ or $b \mapsto a$ should remain the same). Given two discrete distributions X and Y , the *JSD* is defined by:

$$JSD(X||Y) = \frac{1}{2}D(X||\mu) + \frac{1}{2}D(Y||\mu) \quad (7.11)$$

where $\mu = \frac{1}{2}(X + Y)$ denotes the mean of the distributions and D denotes the *Kullback-Leibler Divergence (KLD)* given by:

$$D(X||Y) = \sum_i X(i) \log \frac{X(i)}{Y(i)}$$

By substituting the *KLD* in Eq. 7.11, the complete *JSD* is given by:

$$JSD(X||Y) = \frac{1}{2} \sum_i \left[X(i) \log \left(\frac{X(i)}{\frac{1}{2}(X(i) + Y(i))} \right) + Y(i) \log \left(\frac{Y(i)}{\frac{1}{2}(X(i) + Y(i))} \right) \right]$$

Thus, the distance measure of *JSD* is obtained by squaring its *divergence relation* given by:

$$JS_{dist} = \sqrt{(JSD(X||Y))} \quad (7.12)$$

Both Eq. 7.11 and Eq. 7.12 have been computed using the *Scipy implementation*⁷. For an efficient implementation, the outputs of the *LDA* (the learned topics) are represented in a dense matrix $\mathcal{M}_{lda}^{m \times k}$ of size $m \times k$ consisting of m number of documents (texts from nodes) and their k corresponding number of *topics*.

Algorithm 3 : *Algorithm text-sim identifies textually-related clusters*

- | | |
|--|--|
| 1: Initialisation: $\{\} \leftarrow \mathcal{T}_r; \{\} \leftarrow \mathcal{T}_u$ | |
| 2: Input: collection of <i>structurally-related nodes</i> \mathcal{S}_r | |
| 3: $\forall v_i \in \mathcal{S}_r$, get k texts $g(\mathcal{T}_{v_i})$ | $\triangleright g(\mathcal{T}_{v_i})$ set of k texts of v_i |
| 4: $\mathbf{x}_i \leftarrow t_i \in g(\mathcal{T}_{v_i})$ | \triangleright get texts vectors \mathbf{x}_i |
| 5: truncate \mathbf{x}_i | \triangleright retain b top terms in vector \mathbf{x}_i |
| 6: $m(\mathcal{T}_{v_i}) = \frac{1}{n} \sum_{i=1}^n \frac{\mathbf{x}_i}{\ \mathbf{x}_i\ _2}$ | \triangleright mean of L_2 normalised \mathbf{x}_i |
| 7: $LDA(m(\mathcal{T}_{v_i}))$ | \triangleright invoke the <i>LDA</i> on $m(\mathcal{T}_{v_i})$ |
| 8: $\mathcal{T}_{sim}(\mathcal{T}_{v_i}, \mathcal{T}_{v_j}) = JS_{dist}(\mathcal{T}_{v_i} \mathcal{T}_{v_j})$ | \triangleright get similar texts using Eq. 7.12 |
| 9: if $\mathcal{T}_{sim}(\mathcal{T}_{v_i}, \mathcal{T}_{v_j}) \geq \tau$ then | |
| 10: update \mathcal{T}_r | \triangleright textually-related |
| 11: else | |
| 12: update \mathcal{T}_u | \triangleright textually-unrelated |
| 13: end if | |
| 14: Output: | |
| 15: $\mathcal{T}_r, \mathcal{T}_u, \mathcal{M}_{ta}^{m \times m}$ | \triangleright affinity matrix $\mathcal{M}_{ta}^{m \times m}$ |
-

7.3.2 Modelling Textual Communities

Each of the nodes belongs to a *topically-induced community* according to the similarity in the *LDA* distribution computed using the Eq. 7.12. Based on the top topics or most relevant topics,

⁷See <https://www.scipy.org/>

the following matrices and corresponding dimensions are obtained:

- $\mathcal{M}_{vt} \mapsto m \times k$: *matrix of nodes and top k topics*
- $\mathcal{M}_{va} \mapsto m \times m$: *affinity matrix of nodes according to similarity in topics*

Accordingly, communities of nodes are formed around similar topics of discussions, which need to be identified. The goal of the following model (Eq. 7.13) is to identify clusters of nodes or users according to topical similarities.

$$tr(\mathcal{M}_{vt}^T \mathcal{M}_{va} \mathcal{M}_{vt}) \quad (7.13)$$

The goal is to identify clusters of nodes according to the topical similarities using Algorithm 3. Using the affinity matrix based on \mathcal{S}_r or \mathcal{T}_r , various community detection algorithms can be used to identify relevant partitions. Thus, community detection is based on optimising the joint similarities of \mathcal{S}_r and \mathcal{T}_r : $\psi(\mathcal{S}_r, \mathcal{T}_r) = (\phi(\mathcal{S}_r) + \phi(\mathcal{T}_r))$. Chapter VIII gives details about the algorithms used in the thesis.

7.4 Microcosms Detection Algorithm

Fundamentally, the problem of discovering community structures within a network is modelled as a clustering task, in which nodes are grouped according to a scoring function. The following section describes the *MCT* implementation.

Algorithm 4 : *Algorithm MCT identifies local communities known as microcosms in a network.*

- | | |
|--|--|
| 1: Initialisation: $\{\} \leftarrow \mathcal{S}_r, \{\} \leftarrow \mathcal{S}_u, \{\} \leftarrow \mathcal{T}_r, \{\} \leftarrow \mathcal{T}_u$ | |
| 2: Input: a collection of network data \mathcal{D} | |
| 3: structural-component: | ▷ <i>invoke f-sim (alg. 2)</i> |
| 4: f-sim(\mathcal{D}) $\mapsto \{\mathcal{S}_r, \mathcal{S}_u\}, \mathcal{M}_{A_{i,j}}^{n \times n}$ | ▷ <i>alg. 2 output</i> |
| 5: textual-component: | ▷ <i>invoke text-sim (alg. 3)</i> |
| 6: $\forall v_i \in \mathcal{S}_r$ get k tweets | ▷ <i>set of texts \mathcal{T}_{v_i}</i> |
| 7: text-sim(\mathcal{S}_r) $\mapsto \{\mathcal{T}_r, \mathcal{T}_u\}, \mathcal{M}_{ta}^{m \times m}$ | ▷ <i>alg. 3 output</i> |
| 8: compare all topics($\mathcal{T}_{v_i} \in \mathcal{S}_r$) using Eq. 7.11 | ▷ <i>affinity matrix</i> |
| 9: local clusters: | |
| 10: $\psi(\mathcal{S}_r, \mathcal{T}_r)$ | ▷ $\mathcal{S}_r, \mathcal{T}_r \geq \tau$ |
| 11: Output: | |
| 12: $\mathcal{C}_{c_i}^{n \times p}$ | ▷ <i>local communities</i> |
-

7.4.1 Objective Function and Optimisation

The *MCT* approach is presented according to the following broad categories: *optimisation of matrices of values* and *optimisation of intra-cluster similarity*.

Optimising matrices of values

In many applications, it is often the case that the dynamism of a system vis-à-vis some variables is analysed. To this end, it is useful to compute the greatest rate of change – increase or decrease – of a given function at a specified point, wherefore the direction of change defines the point of interest or direction of the steepest ascent or descent (Strauss et al. 2002). Because the final function in *MCT* gives the maximum achievable values in both $\mathcal{S}_r, \mathcal{T}_r$, the aim is to determine the extremum to achieve the goal of detection. Consequently, given a collection of network data \mathcal{D} , defined by a set of nodes \mathcal{V} and a set of edges \mathcal{E} , for each node v_i , there exists sets of *structural* and *textual* features, the *MCT algorithm* identifies subset of nodes \mathcal{P} with a high degree of similarity in both *structural* and *textual* aspects:

$$\mathcal{P} \subset \mathcal{D} : \forall v_i, v_j \in \mathcal{P} \quad p(R_{v_i, v_j}), \phi(t_i, t_j) \geq \tau \quad (7.14)$$

The *MCT framework* is considered as a form of *multivariate function* comprising of *structural* and *textual* components; this makes it suitable to define an objective function that maximises the overall joint similarity. Eq. 7.14 is the clustering criteria to satisfy in achieving the goal. The collection of *textually-related nodes*, represented in the matrix \mathcal{M}_{vt} , is based on the *structurally-related nodes* (or the subset of $\mathcal{S}_r : \mathcal{T}_r \subseteq \mathcal{S}_r$, at the most equal). In line with the matrix representation $\mathcal{M}_{vt} \subseteq \mathcal{M}_{c_{vr}}$, and since the goal of the *optimisation* is to maximise \mathcal{M}_{vt} , the two are considered to be equal. Thus, the constrained function is expressed as follows:

$$\mathcal{M}_{vt} = \mathcal{M}_{c_{vr}} \text{ thus } \mathcal{M}_{vt} - \mathcal{M}_{c_{vr}} = 0 \quad (7.15)$$

Because \mathcal{M}_{vt} is equated with $\mathcal{M}_{c_{vr}}$, the simplified representation used in Eq. 7.10 also applies to \mathcal{M}_{vt} , hence $\mathcal{M}_{vt} = \mathcal{M}_{c_{vr}} = P$. Therefore, the goal is to maximise the joint models (Eq. 7.10 and Eq. 7.13) under the constrained function (Eq. 7.15) given by Eq. 7.16:

$$\psi(\mathcal{S}_r, \mathcal{T}_r) = \|D - PQ^T\|_F^2 + \sum_{i=1}^n \sum_{r=1}^k p_{ir} \alpha_{ir} + \sum_{j=1}^p \sum_{r=1}^k q_{jr} \beta_{jr} - \lambda \text{tr}(\mathcal{M}_{vt}^T \mathcal{M}_{va} \mathcal{M}_{vt}) \quad (7.16)$$

In Eq. 7.16, λ , the *Lagrange multiplier*, serves as the proportionality constant; see relevant details in Section 1.4.2 of Appendix A.

Optimising intra-cluster similarity

This resembles the approach in Section 7.4.1 in the use of matrices of values, but differs by using a different objective function. The objective function is based on maximising the joint similarity between \mathcal{S}_r and \mathcal{T}_r using aggregation criteria inspired by Aggarwal & Subbian (2012). Accordingly, the goal is to maximise:

$$\psi_{st}(v_i, v_j) = (\lambda) \cdot \mathcal{S}_r(v_i, v_j) + (1 - \lambda) \cdot \mathcal{T}_r(v_i, v_j) \quad (7.17)$$

In Eq. 7.17, the similarity between pairs is computed, and a balancing parameter λ^8 , with values in $(0, 1)$, is specified by the user. Using Algorithm 5, the process continues by assigning subsequent objects to existing or new clusters until an arbitrary upper bound M is reached. When the maximum number of clusters M , which is a user-defined integer signifying the number of clusters, is reached, a new batch of communities will be initiated.

7.5 Summary

Local communities naturally evolve as the size of a network increases, which lead to diverse groups within the network (Berelson & Steiner 1964, Shaw 1971, Granovetter 1992). Because similarity breeds attraction and interaction, communities of similar members are formed, which result in the formation of strong ties among members of communities (Brass et al. 1998). This chapter identified a set of *structurally-related nodes* by exploiting social homophily and equivalence, which combines with a set of *textually-related nodes* under the *MCT framework* to detect local community structures known as (*microcosms*). The *MCT* is focused at an in-depth utilisation of the *bi-modality* for information search for local communities detection. Chapter VIII provides a detailed results from the implemented *MCT framework* and how it compares with existing benchmark models in the literature.

⁸Note that this is different from the one used in optimisation based on matrices of values

Algorithm 5 : *Algorithm MCT-2* identifies local communities known as *microcosms* in a network.

```

1: Input: a collection of network data  $\mathcal{D}$ 
2: structural-component:
3:    $\text{f-sim}(\mathcal{D}) \mapsto \{\mathcal{S}_r, \mathcal{S}_u\}, \mathcal{M}_{A_{i,j}}^{n \times n}$ 
4:   textual-component: ▷
5:      $\forall v_i \in \mathcal{S}_r$  get k tweets
6:      $\text{text-sim}(\mathcal{S}_r) \mapsto \{\mathcal{T}_r, \mathcal{T}_u\}, \mathcal{M}_{ta}^{m \times m}$ 
7:     compare all topics( $\mathcal{T}_{v_i}; v_i \in \mathcal{S}_r$ ) using Eq. 7.11
8: Clusters initialisation:
9:   select four random seed nodes:  $v_i, v_j, v_k, v_l \in \mathcal{S}_r$  and  $v_i, v_j, v_k, v_l \in \mathcal{T}_r$ 
10:  compute pairwise similarities among  $v_i, v_j, v_k, v_l$ 
11: if  $\mathcal{T}_{sim}(\mathcal{T}_{v_i}, \mathcal{T}_{v_j}) \geq \tau$  then
12:   create single cluster  $\mathcal{C}_{ij}$ 
13: else
14:   create two clusters  $\mathcal{C}_i, \mathcal{C}_j$ 
15: end if
16: repeat 9 – 15 until  $|\mathcal{C}_{ik}|_{k=1}^M = M$  ▷ user-defined maximum clusters  $M$ 
17: Assign nodes to clusters:
18:    $\forall v_i \in \mathcal{S}_r$  compute similarity with cluster's mean
19:    $\max_{\phi(v_i, \mu_{\mathcal{C}_i})}$  ▷ assign  $v_i$  to the most similar  $\mu_{\mathcal{C}_i}$ 
20: update cluster's mean:  $\mu_{\mathcal{C}_i} \leftarrow \mu_{\mathcal{C}_i}$ 
21: Output:
22:    $\mathcal{C}_{c_i}^{n \times p}$  ▷ local communities

```

Chapter VIII

MCT AND BASELINES

8.1 Introduction

Phenomena in real life are associated with numerous network structures at various levels – from microscopic to large networks. Because a network is a collection of nodes and associated connections or edges, the arrangement of such entities is not random, but follows certain pattern which could be explicit or implicit. The social media ecosystem enables various forms of interactions among diverse users at various levels. In the context of this research, the formation of groups among nodes in a network is create along the following dimensions: *structural* or *textual*. The following chapter presents various results from the experiments conducted in the thesis, wherefore the central focus is on the evaluation of the proposed detection method and comparison with baseline models. Both the *MCT* and *baseline models* have been evaluated on the same set of datasets as described in the forthcoming sections.

8.2 Experimentation Setup

The experimentation process begins with a description of the datasets, baseline models and evaluation metrics.

8.2.1 Datasets

The datasets utilised for the research consist of both *empirical* or *ground-truth*, *predicted* and *ego-networks datasets* from public repositories. The ground-truth data is crawled from Twitter via the provided API to return a collection of *tweets objects*. A *tweet object* is a complex data

object composed of various descriptive fields (see Section 2.2.2 in Appendix A), which enables the extraction of both the *structural* and *textual* components.

Table 8.1: A summary of the datasets utilised in the detection of microcosms. V and E denote *nodes size* and *edges size* respectively. The *ground-truth (G-)* is based on users' categorisation – *verified* or *unverified*.

Dataset	Size	V	E	Description
G-verified	5300	5300	1832630	ground-truth collected for the research
G-unverified	7100	7100	3893075	collected for the research
ego-Facebook	4039	4039	88234	obtained from Leskovec & Krevl (2014)
ego-Twitter	81,306	81,306	1768149	obtained from Leskovec & Krevl (2014)
Predicted	26350	26350	90,732	predicted reciprocity according to 2

Ego-Networks

Ego-Networks datasets are have been obtained from online social networks, consisting of edges representing interactions between people and set of nodes. This category of the data is publicly available at the Stanford data repository (Leskovec & Krevl 2014). Of interest to the research is the *Facebook* and *ego-Twitter* datasets. The data from Facebook consists of anonymised *circles* or *friends lists*, *node features (profiles)*, which were obtained from survey participants. Each node in the Facebook data consists of nodes' ids, set of connections or edges, and anonymised features containing information about the node's *circle*. Some features in the data include *education-classes-id-anonymised feature #*, see full details in Section 1.4.1 of Appendix A. For preprocessing purpose, for each node, if the feature mentioned above is available, a value of 1 is used, otherwise 0. The rationale of using this dataset is to explore communities using the network circle of each user in terms of the size of the circle and the diversity of membership to the circles. For instance, the question can be asked, are the users in the same circle sharing similar information features? Some of the features are anonymised because they are sensitive. However, the annonymised features suggest that the users who share sensitive information could have something in common. Table 8.1 gives a summary of the datasets. Table 8.2 shows relevant information, such as *closeness centrality*, *betweenness* and *clustering coefficient*, about the empirical data.

Table 8.2: Relevant statistics about the ground-truth data.

$\mu_{degree-dist}$	$\mu_{reciprocity}$	$\mu_{closeness}$	$\mu_{betweenness}$	$\mu_{edge-density}$
3.079×10^{-4}	1.12×10^{-3}	1.47×10^{-4}	1.95×10^{-3}	1.54×10^{-4}

8.3 Evaluation

To ascertain the efficacy and relevance of the research output, the evaluation process involves a thorough analysis and comparison with relevant baselines drawn from the literature (Section 8.3.2). The evaluation activity involves quantitative analyses from experimentations on various datasets using the algorithms. Other forms of evaluation are specific to the different levels – *structural* and *textual* – outlined in Chapter VII. In summary, the evaluation process entails the following:

- investigates the effect of using *structurally-related nodes* in identifying local communities in social networks
- examines the relationship between *structurally-related clusters* and *textually-related clusters*
- evaluates the performance of the proposed *MCT* and compare with baseline models

8.3.1 Evaluation metrics

The following metrics have been used in the research for relevant evaluations.

- *Clustering coefficient* (C_{coeff}) was presented earlier in Section 3.2; it is used to quantify clustering tendency of a given node in relation to other nodes within a network (Watts & Dodds 2007). Computing the C_{coeff} requires the following: $edges = \frac{k_i(k_i-1)}{2}$ and $C_{coeff_i} = \frac{2E_i}{k_i(k_i-1)}$ where i, k_i, E_i denote a *network node*, *number of edges connecting i to k_i other nodes in the network*, and *actual number of existing edges between k_i nodes*. The ratio $E_i \propto \frac{k_i(k_i-1)}{2}$ defines the clustering coefficient of a node.
- *Community Cohesion* demonstrates the level of connectivity within a community and is captured by measuring the degree of *cohesiveness*. A well-connected community is intuitively difficult to divide into sub-communities due to the presence of strong connectivity

among the nodes (Leskovec et al. 2010). Any useful metric that reveals the degree of cohesion can be used to evaluate cohesiveness if it satisfies the requirement of community cohesion, i.e. to be well-connected and difficult to partition further. In this thesis, *cohesiveness* is measured according to the degrees of similarities in \mathcal{S}_r and \mathcal{T}_r .

- *Average degree* (μ_{degree}) is defined as the average or mean number of node's degree to other member nodes (Radicchi et al. 2004). This measure is not so evident if the network is considered as a whole because there are many disparities across the nodes (see Figure A.2 in Appendix A). However, upon inspection of the various sub-networks or network bands, the value seems to be uniform across the different bands. The average degree is computed by isolating all nodes in the respective networks.
- *Modularity measure* (Q) is a useful metric that measures the strength of communities, which is defined as the number of edges falling within groups minus the expected number in an equivalent network (Newman 2004b). It is also used to detect groups in a network according to the notion that a community structure in a network corresponds to an interesting statistical arrangement of edges which can be quantified using the *modularity score*. The operational intuition behind the modularity is summarised as follows: assuming that a network \mathcal{D} is partitioned into k sub-networks, it is possible to define a symmetric matrix $e = k \times k$ such that e_{ij} denotes the fraction of all edges linking the nodes in community i to nodes in community j . It follows that the *trace* of e , $tr(e) = \sum_{i=1} e_{i,i}$ gives the fraction of all edges in the network that connect nodes in the same community. For instance, a value of $Q > 0$ signifies the possible presence of community structures, wherefore nodes within the same community should be more tightly connected than if it were by chance (Newman 2006). For example, the modularity value of real networks ranges from .3 to .7, in which higher values signify better community structure (Newman 2004a).
- *Normalised Mutual Information* (NMI) is a useful statistical tool to evaluate the quality of clusters in a network (Danon et al. 2005). Basically, the NMI is an information-theoretic method to evaluate the degree of agreement between partitions in a network (Fred & Jain 2002). Computing the NMI is based on the assumption that the network is partitioned

into communities in which each node $v_i \in \mathcal{V}$ is associated with both *true community* and *predicted community* such that $l_{v,p} = i$ defines the predicted community i of a node. If k_t denotes the number of true communities, then the frequency count gives its entropy T , which is used to compute the entropy of the predicted communities P as follows; thus the entropy of T is expressed as:

$$H(T) = - \sum_{i=1}^{k_t} \frac{n_i^t}{n} \log\left(\frac{n_i^t}{n}\right) \quad (8.1)$$

where n_i^t denotes the number of nodes in community i . Therefore, the mutual information (MI) between T and P is given by:

$$MI(T, P) = \sum_{i=1}^{k_t} \sum_{j=1}^{k_p} \frac{n_{i,j}^{tp}}{n} \log\left(\frac{n_{i,j}^{tp}}{n} / \left(\frac{n_i^t}{n} \cdot \frac{n_j^p}{n}\right)\right) \quad (8.2)$$

Eq. 8.2 is normalised by the maximum value $\frac{H(T)+H(P)}{2}$ to obtain the *NMI* as follows:

$$NMI = \frac{-2 \sum_{i=1}^{k_t} \sum_{j=1}^{k_p} n_{i,j}^{tp} \log\left(\frac{n_{i,j}^{tp}}{n_i^t \cdot n_j^p}\right)}{\sum_{i=1}^{k_t} n_i^t \log\left(\frac{n_i^t}{n}\right) + \sum_{j=1}^{k_p} n_j^p \log\left(\frac{n_j^p}{n}\right)} \quad (8.3)$$

8.3.2 Baseline Models

One of the major goals of a clustering algorithm is to maximise *intra-cluster similarity*, which ensures a distinctive property for different clusters. The detection of relevant clusters involves repeatedly optimising partitions to achieve the goals. For the evaluation task, the *MCT* is applied alongside with three different detection algorithms with different mode operations on the datasets described Table 8.1 to identify cohesive groups or local community structures.

Girvan-Neuman An essential requirement for a community detection algorithm is the ability to naturally detect divisions among vertices without external influences or imposing restrictions on the divisions (Newman & Girvan 2004). Accordingly, Girvan & Newman (2002) proposed an iterative approach (*G-N algorithm*) based on a progressive removal of edges in the network according to *betweenness score*, a metric to quantify traffic flow among nodes in a network. The magnitude of each node's *betweenness score* dictates which edge to remove in the network. Intuitively, the most critical nodes in the network may experience a high *traffic flow* and are

liable to create a bottleneck within the network. Ultimately, the *G-N algorithm* traces and discards those nodes; hence getting rid of the bottleneck in the network and achieving a natural division of the network into communities.

Label propagation *Label Propagation (LP)* algorithm is an iterative clustering method suitable for use with unlabelled data. The *LP* algorithm converts unlabelled data to labelled data using an initial set of labelled data. The process of assigning the labels involves a repetitive random reshuffling and tagging of nodes with the most frequent labels to its neighbours until convergence (Zhu & Ghahramani 2002). Information about the labelled data is then propagated across the whole network data.

8.3.3 Structural and textual aspects

This section reports on the effect of using structurally-related nodes in identifying local communities on social networks. Because of the availability of empirical data, the effectiveness of the model is quantified with respect to the degree of conformity with the data. *Algorithm f-sim* (Algorithm 2) computes the similarity between the corresponding features of any pairs of nodes; however, it is crucial to evaluate its efficacy in the prediction task. This is vital because the *tie prediction* segment will be of no relevance if it does not add value to the overall detection framework. Accordingly, the evaluation process entails the following sequence: (1) retrieve a set of nodes with actual reciprocal ties on Twitter (2) compute the proportion of similarity between pairs in the set (3) compare the similarity with the reciprocal nodes predicted using *Algorithm f-sim* (4) perform clustering.

To assess *Algorithm f-sim*, the data in Table 6.2 is utilised. The dataset consists of 1986 instances, in which 1023 are *unverified* and 963 are *verified users*. The *accuracy* of the prediction is obtained by computing the ratio of predicted dyads to actual dyads collected from Twitter based on *Algorithm search-dyads* (Algorithm 1). The best achievable result is 0.608 accuracy, and depending on the threshold τ , the accuracy may be lower or higher. Figure 8.1 shows the possible values τ can be and the corresponding accuracy values. Concerning *homophily* and *structural equivalence*, nodes with similar profiles or social status are more likely to interact and establish a small community. Figure 8.2 shows *homophily* as a form of *structural equivalence* based on *network size* and *indegree* to examine the *probability of an edge formation*.

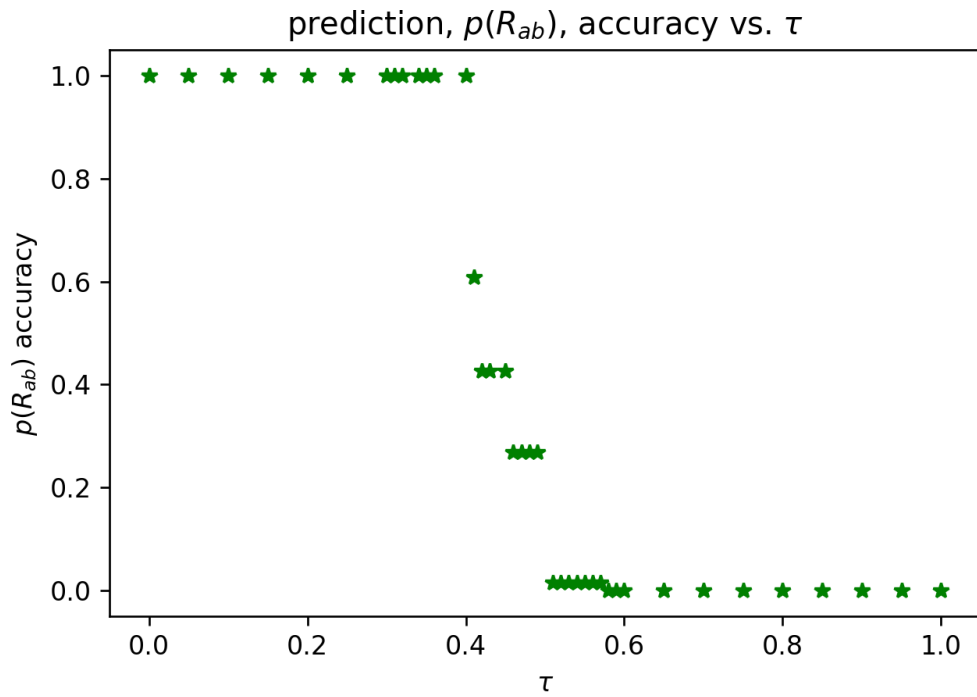


Figure 8.1: The prediction accuracy versus the threshold value: the prediction accuracy is almost 100% when the value of the threshold is low; conversely, the accuracy is almost 0% when the threshold value is very high. A *switch-point* can be observed toward the midpoint in which the accuracy is just above 60% at a threshold value of about 0.41. With additional features, the prediction can be improved. For instance, the inclusion of a description feature led to a significant improvement; however, it requires training on a large corpus to obtain the embedding of terms in the text. The use of features with ease of accessibility is more efficient.

Table 8.3: The result of experiments on three different datasets for community detection using two different algorithms. G–N and LP: Girvan–Neuman and Label Propagation respectively, MCT: Multilevel Clustering Technique; #DC: Number of Detected Communities

Dataset	G–N			LP			MCT		
	Metric		#DC	Metric		#DC	Metric		#DC
	Q	NMI		Q	NMI		Q	NMI	
Ground-truth	.908	.794	308	.77	.602	1319	.915	.791	263
ego-Twitter	.334	.197	1431	.215	.131	2131	.307	.230	1131
ego-Facebook	.522	.590	1037	.421	.304	1780	.503	.372	1845
Predicted	.473	.311	1107	.360	.267	2071	.601	.472	985

Accordingly, the figure (8.2), below, depicts a behaviour that resembles an inverse relationship: increase in *network size* results in decrease in reciprocity.

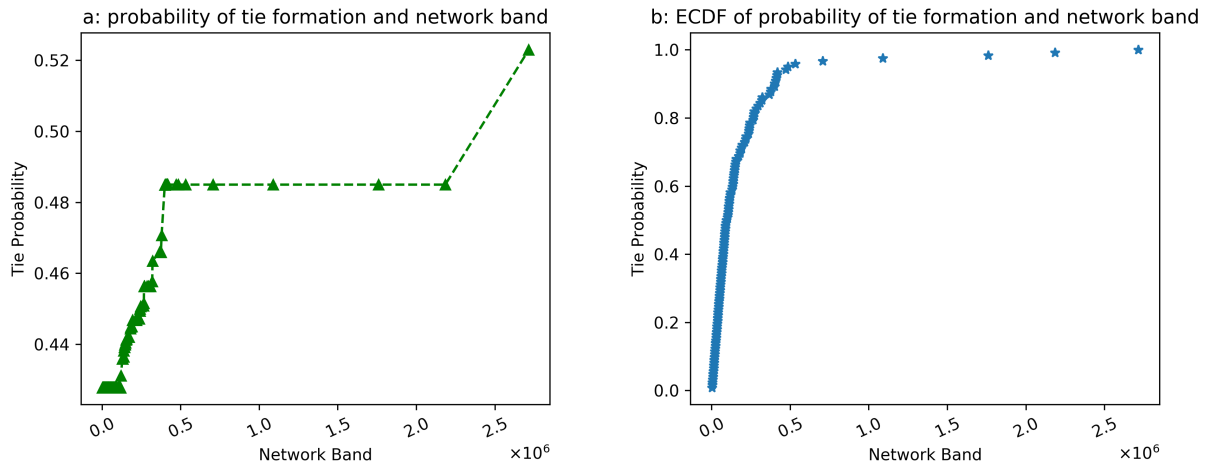


Figure 8.2: The the probability of a tie formation as a function of *network size*. There is a high chance of reciprocating a tie among users in a network band of 0.5×10^6 .

Community Structure

It is crucial to determine how the use of a collection of nodes with reciprocal ties affects the detection of communities in a network. To assess the utility of using a set of nodes consisting of *reciprocal units* (predicted by Algorithm 2) in clustering, the *MCT* (Chapter VII), *G-N* (Girvan & Newman 2002) and *LP* (Zhu & Ghahramani 2002) community detection algorithms are used to examine how the use of a collection of *structurally-related nodes* affects the detection of communities in a network and *compare algorithms' performance*.

The evaluation metrics in this respect consist of the *Modularity measure* and *Normalised Mutual Information (NMI)*, and Table 8.3 shows the results from applying the community de-

tection algorithms on the data. Although all the algorithms detected community structures, there are quantitative variations among the outcomes which are described in details in the following section. With respect to the datasets, as expected, the performance using the *ground-truth* is good across the algorithms. This is followed by the *ego-Facebook* then the *predicted*, and lastly the *ego-Twitter*. The results in Table 8.3 also demonstrate the performance of each model. The results from the *MCT* indicate a more localised structure noting the magnitude of Q , NMI and the number of detected communities ($\#DC$) in comparison with the *ground-truth*. The improvement is attributed to the use of in-depth *structural features* which introduce a layer of connectivity. The performance improvement is based on a multilevel clustering that detects cohesive groups or local community structures at the primary or local level by using an initial high-level grouping of nodes into communities according to the network size and the recognition of *bi-modal information sources* for clustering. The *ego-Facebook* data consists of nodes with reciprocal ties, but the set of the textual features (as described in Sections 8.2.1 and 1.4.1 of Appendix A) is small, making it less complex in comparison with the other datasets.

8.4 Summary

The eccentric connections pattern affects mining tasks involving Twitter. In view of this, the thesis proposed a *multi-level clustering technique (MCT)* to identify a socially cohesive group of users on Twitter termed a *microcosm*. The research demonstrated the experimental results from the *MCT* and evaluation benchmark models, illustrating the efficacy of the approach. This is important because until recently, community detection algorithms focused on single modality, e.g. using *nodes' attributes* or *nodes' connectivity*. This chapter focuses on identifying sets of *structurally-related nodes* by exploiting the idea of *homophily* under the assumption that preference over nodes' attributes induces many reciprocal units, and combine to *textually-related nodes* to detect local community structures (*microcosms* using the *MCT framework*). The *MCT* offers a useful, scalable community detection algorithm that advances existing knowledge in the detection of a community structure in a network. The chapter contributes to the following innovative approach (1) an in-depth utilisation of the *bi-modality* for information search from community detection, and (2) detection of a community of nodes at various levels. The next chapter (Chapter IX) presents the thesis's concluding remarks.

Chapter IX

CONCLUSION

9.1 Introduction

The present technological advancements culminated in the computerisation and automation of various tasks, which leads to a complex ecosystem of information exchange. Within the social network's ecosystem, social interactions are continually evolving to support a myriad of objects to remain connected. Because many forms of social phenomena can be described in the language of networks, it is not surprising that various forms of organising principles have been uncovered in networks. One of the crucial features of a network is the existence or notion of *community structure*, which makes it possible for numerous analyses. The detection of communities provides an effective means to understand the underlying network structure and extract useful information. The understanding and usefulness of uncovering the structure embedded in networks have been attracting huge interest from a wide range of researchers with diverse backgrounds (Erdős & Rényi 1960, Scott 1988, Watts & Strogatz 1998, Albert & Barabási 2002, Newman & Girvan 2004, Newman 2006, Lancichinetti et al. 2009). Understanding the underlying mechanism governing the behaviour and detection of low-level communities on Twitter, which are often implicit, requires extensive exploration (Palla et al. 2007). The following sections provide a summary of the key findings of the research, reflection and future work.

9.2 Key Findings

While many community detection algorithms have been proposed in the literature (as detailed in Chapter III), detection of a socially cohesive community on Twitter is still a challenge, wherefore many disparate communities that are likely to be socially unrelated are detected. In sum-

mary, the contributions of the research are centred around the following fundamental problems:

- Noting the prevalence of irrelevant content on Twitter, *how best to assess the credibility of social media data for the research?*
- Identify relevant methods for *clustering* and *community detection* in a dynamic Twitter environment
- How to develop an effective means of detecting cohesive communities

These aspects are further elaborated in the forthcoming sections.

9.2.1 Content veracity

One of the peculiar features of online social media is enabling diverse users to interact at rapid rate¹. Undoubtedly, the *social media networks* have transformed the way sociological research is being conducted in terms of participants and size of data with profound effect. They offer a useful utility in understanding modern society and how it functions, leading to the new concept of *datafication*, the continuous quest to turn every aspect of humans' lives into computerised data for a competitive value. Despite the relevance of social media, the menace of fake and spam content is still a challenge. One of the implications is undermining the credibility of research based on analysing social media data without extensive filtering. As a precautionary measure to avoid compromising the research outcome by irrelevant or unrepresentative data, the research contributed a useful method for spammers and automated accounts detection. Moreover, the study offers crucial insights into the sophisticatedly evolving techniques used by spammers on Twitter.

Rapid increase in data and prevalence of transitory content affect mining tasks: As the size of a network increases, the possibility of fragmentation (Berelson & Steiner 1964, Shaw 1971), leading to a decrease in the homogeneity of behaviour and attitude across groups (Granovetter 1992). Fundamentally, a community is a functional unit of the network that captures local relationship among the network objects, and the problem of community detection is to identify relevant partitions in the network. The rapid increase in volume and complexity of

¹On average, 100m daily users contribute to 500m, see <https://www.omnicoreagency.com/twitter-statistics/>

online content, of which large scale transitory content mostly from influential users dominate, are posing a challenge to the detection of socially cohesive groups and ascertaining the veracity of content on Twitter. Moreover, the dominance of *influential users' content* results in many forms of communities of users with no structural connection, which hinders the full realisation of the benefits in communities such as *cliquishness* and *local coherence*. To overcome these challenges in mining tasks and ensure the detection of groups with strong social cohesion, transitory components, such as popular hashtags or trending topics, should be complemented with static components, such as connections based on reciprocal ties among nodes. This will ensure the detection of groups with strong social cohesion.

9.2.2 Community Detection

Similarity breeds attraction and interaction, leading to the formation of communities (Brass et al. 1998). The detection of communities provides an effective means to understand the underlying network structure and extract useful information. Through clustering, a compelling summary of relationships among network objects can be found. To achieve a better outcome, it is vital to take cognisance of the following in mining-related tasks such as *community detection*.

The methodological approach affects community detection: In Section 1.2, there are two approaches to investigating social relationships – *realist* and *nominalist*. Because the assumption of the existence of communities in a network, as held by the *realist* viewpoint, may results in many unrelated groups of users, adopting the *nominalist* approach, which is based on the questions asked by the investigator (Laumann et al. 1989), offers a better chance of enabling identification of cohesive communities on Twitter.

Connection types affect the formation of communities: The formation of a social tie can be achieved either based on *event-type ties*, mostly transitory in nature, e.g. using similar *hashtag*, or *state-type ties*, which are mostly static and are based on structural similarity. Although both are useful for data mining tasks, connections based on *state-type ties* provide better results since *event-type ties* are transitory and prone to detecting socially distinct members. *Dyadic and transitive ties*, primary forms of establishing a reciprocal tie, play crucial roles in identifying socially cohesive groups.

Impact of bi-modal features: Identifying the set of fully connected nodes on Twitter is challenging due to the flexible and eccentric underlying connection patterns, which enables flexible followership that results in many unidirectional links. The presence of many unreciprocated connections affects many mining-related tasks involving Twitter, such as identifying communities with less cohesion and the promotion or proliferation of spam content. In platforms with the particularity of Twitter, the detection of communities will be more effective if both the *network structure* and *nodes' features or attributes of nodes* are considered. One of the premises in the research is that the recognition of a set of reciprocal units for *community detection* on Twitter offers a more cohesive and better representation of a community. Reciprocal relationships have been interpreted differently, which are either based on *directed sets of nodes* or *textual content* (Weng et al. 2010, Kwak et al. 2010, Cha et al. 2012, Arnaboldi, Conti, Passarella & Pezzoni 2013). The widespread availability of transitory connections makes it challenging to identify reciprocal ties based on *state-type ties* on Twitter. This research utilised two primary forms of establishing a reciprocal tie and community formation – *dyads* and *Simmelian tie*.

Multilevel Clustering Technique Noting the eccentric connections pattern in Figure 1.2, which could lead to the detection of socially unrelated users and encourages the propagation of spurious content, the thesis proposed a *multi-level clustering technique (MCT)* to identify a socially cohesive group of users on Twitter termed a *microcosm*. Recent approach to identifying communities involves the use of *bi-modal information source* – *the network structure* and *the features and attributes of nodes*. Through the *MCT strategy*, the problem of structurally unrelated users is addressed, thereby adding a layer of social cohesion to community detection approaches proposed in the past. In summary, the proposed *MCT* advances existing techniques in the related literature (see Chapter III) through:

- a systematic exposition of community detection or clustering algorithms
- an in-depth utilisation of the bi-modality for community detection
- detection of network communities at various levels
- an intuitive interpretation of the detected communities

The proposed method contributed to a methodological paradigm for the detection of *microcosms* in a dynamic and heterogeneous social media like Twitter.

A bi-modal approach adds a layer of social cohesion in the detection task: This is important because until recently, community detection algorithms focused on single modality, e.g. using *nodes' attributes* or *nodes' connectivity*. Using a detection framework that recognises various aspects of the network contributes towards a useful analysis of complex network to identify sets of nodes and corresponding relationships through the following: *structural dimensions* or *state-type ties* are more cohesive and more reliable. Failure to combine both information modalities limits the capability of capturing the nuances and relevant connections, which enables community detection.

9.3 Reflection and Future Research

The following section provides relevant reflection about the research process, challenges and insight about future research direction.

9.3.1 Spam Detection

The connection topology on Twitter contributes to widespread spurious content and a less cohesive community of users due to many unreciprocated or *event-type ties*. The stream of tweets differs from conventional stream of texts in terms of *posting rate*, *dynamism* and *flexibility*; these attributes make *tweets* noisy data and difficult to work with. Moreover, the proliferation of *fake news* and *content* from *automated accounts* or *social bots* threatens the credibility of data without active filtering. The social media ecosystem is continuously faced with new sets of threats determined to undermine civilised interactions. To further improve on the spam strategy proposed in this thesis, a new dimension can be explored, such as through a detailed investigation into how homophily will impact spam detection tasks. For instance, many crucial aspects, such as validation or characterisation of content integrity, can be explored since a user who spreads rumours or spam content is likely to be strongly connected with similar users, hence remain connected. Exploiting data from users with reciprocal connections on networks with high porosity such as Twitter will help in curtailing some of the challenges in data mining and

spam detection. One of the challenges in this respect is acquiring large scale ground-truth data for the analysis. Motivated by the utility of *structural component* (analysed in Section 7.2), a deeper understanding can be obtained.

9.3.2 Analyses of aspects of sociometry and clustering

In social science, a taxonomy of social relationships is described as a function of closeness among users. The closer the users are, the more cohesive and trustworthy. *Dyadic* and *Simmelian* ties constitute the basic unit for analysing structurally-related nodes; *how best to utilise these insights for more robust analysis of other aspects of sociometry such as centrality in a network?* This will be an essential dimension to explore further. In terms of conversation and *mentioning users*, it is often the case that those users are engaged in reciprocal ties, which signifies a strong social cohesion and is a crucial insight noting how social networks come in various forms depending on the cohesiveness and size – from the most intimate to tenuous relationships. This is more amplified in dynamic networks such as Twitter. Drawing on the idea of social *homophily*, users with many reciprocated ties play a central role in enabling the analysis of groups with strong social cohesion.

Harnessing relevant theories in social science to improve mining tasks: Useful theories in social science, such as *homophily* and *social equivalence*, support the detection of relevant groups with structural similarities. Users with many reciprocal ties offer a useful representation of or resourceful representative in the quest of detecting microcosms, making it possible to analyse a group of users as a unit. Motivated by this insight, the research examined *reciprocal ties* as the basic units of the relationship on Twitter.

9.3.3 Future Work

With the proliferation of deep learning models and the vast availability of social media data, many exciting research problems are open for exploration. By leveraging effective computational frameworks, many relevant theories in social networks can now be validated experimentally or empirically. The followings are some of the future works.

To explore the intersection between advances in other areas of deep learning and social network analysis: Worthy of mention include: (1) to explore how Generative Adversarial

Networks (GANs) can be harnessed to generate users with seemingly genuine relationships (which are rare on Twitter) (2) to investigate new developments in community detection algorithms especially with respect to evaluations. Of interest, is to assess *how active or visible are the communities built based on dyadic or Simmelian ties*? This will make it possible to evaluate the degree of visibility of such communities within the broader ecosystem that is mostly dominated by influential users and popular content. With respect to SWAPS, which balances the trade-off between speed and accuracy, to minimise omitting relevant items and maximise the algorithm’s functionality (Inuwa-Dutse et al. 2019a), future work will focus on Deep Reinforcement Learning (DRL) to utilise the algorithm in crafting a policy that will guide the operation of the DRL’s agent.

Computational Sociometry: With the current data deluge, it will be worthwhile to analyse various forms of relationships. For example, casual acquaintances, defined by a weak tie that can be measured on the basis of frequency of communication (Brass et al. 1998), can be analysed using features such as *reciprocity*, *replies*, *mentions*, *the similarity in profiles*, *Dunbar’s number (minimum reciprocal ties)* and *structural holes*. In the same vein, *multiplexity* of relationships, defined as the degree to which two actors are linked by more than one types of relationship, e.g. friend, business associate, school, neighbour (Burt & Minor 1983, Brass et al. 1998), can be analysed more effectively using social media data. Noting that *multiplexity of relationships* adds a constraint on unethical behaviour (Brass et al. 1998), a reflection of it can be extended to online content veracity and community detection tasks. Multiplexity of relationships manifests in a dense network and network with a *structural hole*. Network density is defined as the proportion of network ties compared to the total number of possible ties in a network (Scott 1988). The more interconnected the network, the higher the density. A structural hole, first used in Burt & Minor (1983), refers to the absence of a link between two actors in a network; consider the following example involving three nodes a , b , and c : $a \rightarrow b$ and $a \rightarrow c$; because there is no direct link between b and c (i.e. $b \rightarrow c$ does not exist), there is a structural hole since a controls flow of information between b and c . The existence of structural holes promotes unethical behaviour in organisations because it does not affect surveillance and refutation much (Brass et al. 1998). Those are some of the crucial areas to explore further.

BIBLIOGRAPHY

Aggarwal, C. C. (2018), *Machine learning for text*, Springer.

Aggarwal, C. C. & Subbian, K. (2012), Event detection in social streams, *in* ‘Proceedings of the 2012 SIAM international conference on data mining’, SIAM, pp. 624–635.

Aggarwal, C. & Subbian, K. (2014), ‘Evolutionary network analysis: A survey’, *ACM Computing Surveys (CSUR)* **47**(1), 10.

Ahn, Y.-Y., Bagrow, J. P. & Lehmann, S. (2010), ‘Link communities reveal multiscale complexity in networks’, *nature* **466**(7307), 761.

Airoldi, E. M., Blei, D. M., Fienberg, S. E. & Xing, E. P. (2008), ‘Mixed membership stochastic blockmodels’, *Journal of machine learning research* **9**(Sep), 1981–2014.

Albert, R. & Barabási, A.-L. (2002), ‘Statistical mechanics of complex networks’, *Reviews of modern physics* **74**(1), 47.

Allan, J., Papka, R. & Lavrenko, V. (1998), On-line new event detection and tracking., *in* ‘Sigir’, Vol. 98, Citeseer, pp. 37–45.

Alsaleh, M., Alarifi, A., Al-Quayed, F. & Al-Salman, A. S. (2015), Combating Comment Spam with Machine Learning Approaches, *in* ‘14th IEEE International Conference on Machine Learning and Applications (ICMLA)’, Miami, FL, USA, pp. 295–300.

Analytics, P. (2009), ‘Twitter study’.

Arnaboldi, V., Conti, M., Passarella, A. & Pezzoni, F. (2013), Ego networks in twitter: an experimental analysis, *in* ‘2013 Proceedings IEEE INFOCOM’, IEEE, pp. 3459–3464.

- Arnaboldi, V., Guazzini, A. & Passarella, A. (2013), 'Egocentric online social networks: Analysis of key features and prediction of tie strength in facebook', *Computer Communications* **36**(10-11), 1130–1144.
- Backstrom, L., Huttenlocher, D., Kleinberg, J. & Lan, X. (2006), Group formation in large social networks: membership, growth, and evolution, in 'Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining', ACM, pp. 44–54.
- Baio, G. & Blangiardo, M. (2010), 'Bayesian hierarchical model for the prediction of football results', *Journal of Applied Statistics* **37**(2), 253–264.
- Bakhshandeh, R., Samadi, M., Azimifar, Z. & Schaeffer, J. (2011), Degrees of separation in social networks, in 'Fourth Annual Symposium on Combinatorial Search'.
- Balasubramanyan, R. & Cohen, W. W. (2011), Block-lda: Jointly modeling entity-annotated text and entity-entity links, in 'Proceedings of the 2011 SIAM International Conference on Data Mining', SIAM, pp. 450–461.
- Barabási, A.-L., Jeong, H., Néda, Z., Ravasz, E., Schubert, A. & Vicsek, T. (2002), 'Evolution of the social network of scientific collaborations', *Physica A: Statistical mechanics and its applications* **311**(3-4), 590–614.
- Becker, H., Naaman, M. & Gravano, L. (2011), Beyond trending topics: Real-world event identification on twitter, in 'Fifth international AAAI conference on weblogs and social media'.
- Benevenuto, F., Magno, G., Rodrigues, T. & Almeida, V. (2010), Detecting Spammers on Twitter, in 'In Collaboration, Electronic messaging, Anti-Abuse and Spam Conference (CEAS', Vol. 6.
- Berelson, B. & Steiner, G. A. (1964), 'Human behavior: An inventory of scientific findings.'
- Berkhin, P. (2006), A survey of clustering data mining techniques, in 'Grouping multidimensional data', Springer, pp. 25–71.
- Bertsekas, D. P. (1997), 'Nonlinear programming', *Journal of the Operational Research Society* **48**(3), 334–334.

- Biber, D. & Geoffrey Leech, S. C. (2002), *The Longman Student Grammar of Spoken and Written English*, Longman.
- Bickel, S. & Scheffer, T. (2004), Multi-view clustering., in 'ICDM', Vol. 4, pp. 19–26.
- Bishop, C. M. (2006), *Pattern recognition and machine learning*, springer.
- Blei, D. M., Ng, A. Y. & Jordan, M. I. (2003), 'Latent dirichlet allocation', *Journal of machine Learning research* **3**(Jan), 993–1022.
- Blondel, V. D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. (2008), 'Fast unfolding of communities in large networks', *Journal of statistical mechanics: theory and experiment* **2008**(10), P10008.
- Borgatti, S. P. & Halgin, D. S. (2011), 'On network theory', *Organization science* **22**(5), 1168–1181.
- Brants, T., Chen, F. & Farahat, A. (2003), A system for new event detection, in 'Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval', ACM, pp. 330–337.
- Brass, D. J. (1985), 'Men's and women's networks: A study of interaction patterns and influence in an organization', *Academy of Management journal* **28**(2), 327–343.
- Brass, D. J., Butterfield, K. D. & Skaggs, B. C. (1998), 'Relationships and unethical behavior: A social network perspective', *Academy of management review* **23**(1), 14–31.
- Burt, R. S. & Minor, M. J. (1983), *Applied network analysis: A methodological introduction*, Sage Publications, Inc.
- Calheiros, A. C., Moro, S. & Rita, P. (2017), 'Sentiment classification of consumer-generated online reviews using topic modeling', *Journal of Hospitality Marketing & Management* **26**(7), 675–693.
- Cao, C., Ni, Q. & Zhai, Y. (2015), An improved collaborative filtering recommendation algorithm based on community detection in social networks, in 'Proceedings of the 2015 annual conference on genetic and evolutionary computation', ACM, pp. 1–8.

- Carley, K. (1991), 'A theory of group stability', *American sociological review* pp. 331–354.
- Cataldi, M., Di Caro, L. & Schifanella, C. (2010), Emerging topic detection on twitter based on temporal and social terms evaluation, in 'Proceedings of the tenth international workshop on multimedia data mining', ACM, p. 4.
- Cha, M., Benevenuto, F., Haddadi, H. & Gummadi, K. (2012), 'The world of connections and information flow in twitter', *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans* **42**(4), 991–998.
- Chakraborty, R., Kundu, S. & Agarwal, P. (2016), Fashioning data-a social media perspective on fast fashion brands, in 'Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis', pp. 26–35.
- Chao, G., Sun, S. & Bi, J. (2017), 'A survey on multi-view clustering', *arXiv preprint arXiv:1712.06246*.
- Chaudhuri, K., Kakade, S. M., Livescu, K. & Sridharan, K. (2009), Multi-view clustering via canonical correlation analysis, in 'Proceedings of the 26th annual international conference on machine learning', ACM, pp. 129–136.
- Chawla, N. V., Bowyer, K. W., Hall, L. O. & Kegelmeyer, W. P. (2002), 'SMOTE: Synthetic Minority Over-sampling Technique', *Journal of Artificial Intelligence Research* **16**, 321–357.
- Chen, J., Zaiane, O. R. & Goebel, R. (2009), Detecting communities in large networks by iterative local expansion, in '2009 International Conference on Computational Aspects of Social Networks', IEEE, pp. 105–112.
- Contractor, D., Chawda, B., Mehta, S., Subramaniam, L. & Faruque, T. A. (2015), Tracking Political Elections on Social Media: Applications and Experience, in 'Proceedings of the 24th International Conference on Artificial Intelligence', AAAI Press, pp. 2320—2326.
- Cukier, K. & Mayer-Schoenberger, V. (2013), 'The rise of big data: How it's changing the way we think about the world', *Foreign Aff.* **92**, 28.

- Danezis, G. & Mittal, P. (2009), SybilInfer: Detecting Sybil Nodes using Social Networks, in 'Proceedings of the Network and Distributed System Security Symposium (NDSS)', San Diego, California, USA.
- Dann, S. (2010), 'Twitter content classification', *First Monday* **15**(12).
- Danon, L., Diaz-Guilera, A., Duch, J. & Arenas, A. (2005), 'Comparing community structure identification', *Journal of Statistical Mechanics: Theory and Experiment* **2005**(09), P09008.
- Davenport, T. (2014), Analytics in Sports: The New Science of Winning, Technical report, International Institute for Analytics. White paper.
- Davis, C. A., Varol, O., Ferrara, E., Flammini, A. & Menczer, F. (2016), BotOrNot: A System to Evaluate Social Bots, in 'Proceedings of the 25th International Conference on World Wide Web (WWW), Companion Volume', Montreal, Canada, pp. 273–274.
- Deloitte (2014), Social Analytics in Media Entertainment the Three-minute Guide, Technical report, Deloitte Development LLC.
- Dijk van Jan, A. (2006), 'The network society. social aspects of new media'.
- Doreian, P., Batagelj, V. & Ferligoj, A. (2005), 'Positional analyses of sociometric data', *Models and methods in social network analysis* **77**, 77–96.
- Dunbar, R. I. (1998), 'The social brain hypothesis', *Evolutionary Anthropology: Issues, News, and Reviews: Issues, News, and Reviews* **6**(5), 178–190.
- Erdős, P. & Rényi, A. (1960), 'On the evolution of random graphs', *Publ. Math. Inst. Hung. Acad. Sci* **5**(1), 17–60.
- Ester, M., Ge, R., Gao, B. J., Hu, Z. & Ben-Moshe, B. (2006), Joint cluster analysis of attribute data and relationship data: the connected k-center problem, in 'Proceedings of the 2006 SIAM International Conference on Data Mining', SIAM, pp. 246–257.
- Fawcett, T. (2006), 'An introduction to ROC analysis', *Pattern Recognition Letters* **27**(8), 861–874. ROC Analysis in Pattern Recognition.

- Feng, W., Zhang, C., Zhang, W., Han, J., Wang, J., Aggarwal, C. & Huang, J. (2015), Stream-cube: hierarchical spatio-temporal hashtag clustering for event exploration over the twitter stream, in '2015 IEEE 31st International Conference on Data Engineering', IEEE, pp. 1561–1572.
- Flake, G. W., Lawrence, S., Giles, C. L. & Coetzee, F. M. (2002), 'Self-organization and identification of web communities', *Computer* pp. 66–71.
- Forman, G. & Scholz, M. (2010), 'Apples-to-apples in Cross-validation Studies: Pitfalls in Classifier Performance Measurement', *ACM SIGKDD Explorations Newsletter* **12**(1), 49–57.
- Fred, A. L. & Jain, A. K. (2002), Data clustering using evidence accumulation, in 'Object recognition supported by user interaction for service robots', Vol. 4, IEEE, pp. 276–280.
- Freeman, L. C. (1996), 'Some antecedents of social network analysis', *Connections* **19**(1), 39–42.
- Fung, G. P. C., Yu, J. X., Yu, P. S. & Lu, H. (2005), Parameter free bursty events detection in text streams, in 'Proceedings of the 31st international conference on Very large data bases', VLDB Endowment, pp. 181–192.
- Gadek, G., Pauchet, A., Malandain, N., Khelif, K., Vercouter, L. & Brunessaux, S. (2017), 'Topical cohesion of communities on twitter', *Procedia Computer Science* **112**, 584–593.
- Gao, H., Hu, J., Wilson, C., Li, Z., Chen, Y. & Zhao, B. Y. (2010), Detecting and characterizing social spam campaigns, in 'Proceedings of the 10th ACM SIGCOMM Internet Measurement Conference (IMC)', Melbourne, Australia, pp. 35–47.
- Girvan, M. & Newman, M. E. (2002), 'Community structure in social and biological networks', *Proceedings of the national academy of sciences* **99**(12), 7821–7826.
- Granovetter, M. (1992), 'Problems of explanation in economic sociology', *Networks and organizations: Structure, form, and action* pp. 25–56.
- Granovetter, M. S. (1977), The strength of weak ties, in 'Social networks', Elsevier, pp. 347–367.

- Gräßer, F., Kallumadi, S., Malberg, H. & Zaunseder, S. (2018), Aspect-based sentiment analysis of drug reviews applying cross-domain and cross-data learning, *in* ‘Proceedings of the 2018 International Conference on Digital Health’, ACM, pp. 121–125.
- Grier, C., Thomas, K., Paxson, V. & Zhang, C. M. (2010), @spam: The Underground on 140 Characters or Less, *in* ‘Proceedings of the 17th ACM Conference on Computer and Communications Security (CCS)’, Chicago, Illinois, USA, pp. 27–37.
- Guare, J. (1990), *Six degrees of separation: A play*, Vintage.
- Guille, A. & Favre, C. (2015), ‘Event detection, tracking, and visualization in twitter: a mention-anomaly-based approach’, *Social Network Analysis and Mining* **5**(1), 18.
- Guyon, I. & Elisseeff, A. (2003), ‘An introduction to variable and feature selection’, *Journal of Machine Learning Research* **3**, 1157–1182.
- Halko, N. P., Martinsson, P.-G. & Tropp, J. A. (2011), ‘Finding Structure with Randomness: Probabilistic Algorithms for Constructing Approximate Matrix Decompositions’, *Society for Industrial and Applied Mathematics (SIAM) Review* **53**(2), 217–288.
- Han, J., Pei, J. & Kamber, M. (2011), *Data mining: concepts and techniques*, Elsevier.
- Hochreiter, S. & Schmidhuber, J. (1997), ‘Long short-term memory’, *Neural computation* **9**(8), 1735–1780.
- Howard, P. N. & Kollanyi, B. (2016), ‘Bots, #StrongerIn, and #Brexit: Computational Propaganda during the UK-EU Referendum’, *Social Science Research Network (SSRN)* .
- Inuwa-Dutse, I. (2018), Modelling formation of online temporal communities, *in* ‘Companion of the The Web Conference 2018 on The Web Conference 2018’, International World Wide Web Conferences Steering Committee, pp. 867–871.
- Inuwa-Dutse, I., Bello, B. S. & Korkontzelos, I. (2018), ‘Lexical analysis of automated accounts on twitter’, *arXiv preprint arXiv:1812.07947* .
- Inuwa-Dutse, I., Liptrott, M. & Korkontzelos, I. (2018), ‘Detection of spam-posting accounts on twitter’, *Neurocomputing* **315**, 496–511.

- Inuwa-Dutse, I., Liptrott, M. & Korkontzelos, I. (2019a), 'A deep semantic search method for random tweets', *Online Social Networks and Media* **13**, 100046.
- Inuwa-Dutse, I., Liptrott, M. & Korkontzelos, Y. (2019b), Analysis and prediction of dyads in twitter, *in* 'International Conference on Applications of Natural Language to Information Systems', Springer, pp. 303–311.
- Inuwa-Dutse, I., Liptrott, M. & Korkontzelos, Y. (2019c), Simmelian ties on twitter: Empirical analysis and prediction, *in* '2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS)', IEEE, pp. 118–125.
- Inuwa-Dutsea, I., Bello, B. S. & Korkontzelos, I. (2018), 'The effect of engagement intensity and lexical richness in identifying bot accounts on twitter.', *IADIS International Journal on WWW/Internet* **16**(2).
- Jacobs, J. (1992), 'The death and life of great american cities. 1961', *New York: Vintage* .
- Jain, A. K., Dubes, R. C. et al. (1988), *Algorithms for clustering data*, Vol. 6, Prentice hall Englewood Cliffs.
- Japkowicz, N. (2000), The class imbalance problem: Significance and strategies, *in* 'In Proceedings of the 2000 International Conference on Artificial Intelligence (ICAI)', pp. 111–117.
- Karami, A., Gangopadhyay, A., Zhou, B. & Kharrazi, H. (2018), 'Fuzzy approach topic discovery in health and medical corpora', *International Journal of Fuzzy Systems* **20**(4), 1334–1345.
- Karlis, D. & Ntzoufras, I. (2003), 'Analysis of sports data by using bivariate poisson models', *Journal of the Royal Statistical Society: Series D (The Statistician)* **52**(3), 381–393.
- Karthick, S., Shalinie, S. M., Kollengode, C. & Priya, S. M. (2014), A sparsification technique for faster hierarchical community detection in social networks, *in* '2014 3rd International Conference on Eco-friendly Computing and Communication Systems', IEEE, pp. 29–34.
- Katz, E., Lazarsfeld, P. F. & Roper, E. (2017), *Personal influence: The part played by people in the flow of mass communications*, Routledge.

- Kernighan, B. W. & Lin, S. (1970), ‘An efficient heuristic procedure for partitioning graphs’, *Bell system technical journal* **49**(2), 291–307.
- Kim, Y. (2014), ‘Convolutional neural networks for sentence classification’, *arXiv preprint arXiv:1408.5882*.
- Kim, Y.-H., Seo, S., Ha, Y.-H., Lim, S. & Yoon, Y. (2013), ‘Two applications of clustering techniques to twitter: Community detection and issue extraction’, *Discrete dynamics in nature and society* **2013**.
- Kleinberg, J. (2002), Bursty and hierarchical structure in streams, data mining and knowledge discovery, in ‘elected Papers from the 8th ACM SIGKDD Int. Conf. on Knowledge I Discovery and Data Mining? Part’, pp. 372–397.
- Kleinberg, J. M. (2000), ‘Navigation in a small world’, *Nature* **406**(6798), 845.
- Kollios, G., Potamias, M. & Terzi, E. (2011), ‘Clustering large probabilistic graphs’, *IEEE Transactions on Knowledge and Data Engineering* **25**(2), 325–336.
- Krogan, N. J., Cagney, G., Yu, H., Zhong, G., Guo, X., Ignatchenko, A., Li, J., Pu, S., Datta, N., Tikuisis, A. P. et al. (2006), ‘Global landscape of protein complexes in the yeast *saccharomyces cerevisiae*’, *Nature* **440**(7084), 637.
- Kucukelbir, A., Ranganath, R., Gelman, A. & Blei, D. (2015), Automatic variational inference in stan, in ‘Advances in neural information processing systems’, pp. 568–576.
- Kwak, H., Lee, C., Park, H. & Moon, S. (2010), What is twitter, a social network or a news media?, in ‘Proceedings of the 19th international conference on World wide web’, AcM, pp. 591–600.
- Lagnier, C., Denoyer, L., Gaussier, E. & Gallinari, P. (2013), Predicting information diffusion in social networks using content and user’s profiles, in ‘European conference on information retrieval’, Springer, pp. 74–85.
- Lancichinetti, A., Fortunato, S. & Kertesz, J. (2009), ‘Detecting the overlapping and hierarchical community structure in complex networks’, *New journal of physics* **11**(3), 033015.

- Laumann, E. O., Marsden, P. V. & Prensky, D. (1989), ‘The boundary specification problem in network analysis’, *Research methods in social network analysis* **61**, 87.
- Lawson, D. J. & Falush, D. (2012), ‘Population identification using genetic data’, *Annual review of genomics and human genetics* **13**, 337–361.
- Lee, D. D. & Seung, H. S. (1999), ‘Learning the parts of objects by non-negative matrix factorization’, *Nature* **401**(6755), 788.
- Lee, K., Eoff, B. D. & Caverlee, J. (2011), Seven Months with the Devils: A Long-Term Study of Content Polluters on Twitter, in ‘Proceedings of the Fifth International Conference on Weblogs and Social Media’, Barcelona, Catalonia, Spain, pp. 185–192.
- Lee, S. & Kim, J. (2012), WarningBird: Detecting Suspicious URLs in Twitter Stream, in ‘19th Annual Network and Distributed System Security Symposium (NDSS)’, San Diego, California, USA, pp. 183–195.
- Leskovec, J. & Krevl, A. (2014), ‘SNAP Datasets: Stanford large network dataset collection’, <http://snap.stanford.edu/data>.
- Leskovec, J., Lang, K. J. & Mahoney, M. (2010), Empirical comparison of algorithms for network community detection, in ‘Proceedings of the 19th international conference on World wide web’, ACM, pp. 631–640.
- Leskovec, J. & McAuley, J. J. (2012), Learning to discover social circles in ego networks, in ‘Proceedings of NIPS’, pp. 539–547.
- Lin, W., Kong, X., Yu, P. S., Wu, Q., Jia, Y. & Li, C. (2012), Community detection in incomplete information networks, in ‘Proceedings of the 21st international conference on World Wide Web’, ACM, pp. 341–350.
- Liu, J., Wang, C., Gao, J. & Han, J. (2013), Multi-view clustering via joint nonnegative matrix factorization, in ‘Proceedings of the 2013 SIAM International Conference on Data Mining’, SIAM, pp. 252–260.
- Liu, L., Jin, R., Aggarwal, C. & Shen, Y. (2012), Reliable clustering on uncertain graphs, in ‘2012 IEEE 12th International Conference on Data Mining’, IEEE, pp. 459–468.

- Lott, B. (2012), Survey of Keyword Extraction Techniques, Technical report, UNM Education.
- Lu, H.-M. & Lee, C.-H. (2015), ‘A twitter hashtag recommendation model that accommodates for temporal clustering effects’, *IEEE Intelligent Systems* **30**(3), 18–25.
- Manning, C. D., Manning, C. D. & Schütze, H. (1999), *Foundations of statistical natural language processing*, MIT press.
- Marley, A. & Regenwetter, M. (2016), ‘Choice, preference, and utility: Probabilistic and deterministic representations’, *New handbook of mathematical psychology* **1**, 374–453.
- Martin, T., Ball, B. & Newman, M. E. (2016), ‘Structural inference for uncertain networks’, *Physical Review E* **93**(1), 012306.
- 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.
- Mencia, E. L. & Fürnkranz, J. (2008), Efficient pairwise multilabel classification for large-scale problems in the legal domain, *in* ‘Joint European Conference on Machine Learning and Knowledge Discovery in Databases’, Springer, pp. 50–65.
- Miller, C., Ginnis, S., Stobart, R., Krasodonski-Jones, A. & Clemence, M. (2015), ‘The road to representivity, a demos and ipsos mori report on sociological research using twitter’, *London: Demos. Available at: http://www.demos.co.uk/files/Road_to_representivity_final.pdf* **1441811336**.
- Miller McPherson, J., Smith-Lovin, L. & Cook, J. M. (2001), ‘Birds of a feather: Homophily in social networks’, *Annual Review of Sociology* **27**(1), 415–444.
- Myung, I. J. (2003), ‘Tutorial on maximum likelihood estimation’, *Journal of mathematical Psychology* **47**(1), 90–100.
- Nascimento, M. A., Sander, J. & Pound, J. (2003), ‘Analysis of sigmod’s co-authorship graph’, *ACM Sigmod record* **32**(3), 8–10.
- Newcomb, T. M. (1978), ‘The acquaintance process: Looking mainly backward.’, *Journal of Personality and Social Psychology* **36**(10), 1075.

- Newman, M. E. (2002), 'The structure and function of networks', *Computer Physics Communications* **147**(1-2), 40–45.
- Newman, M. E. (2003), 'Properties of highly clustered networks', *Physical Review E* **68**(2), 026121.
- Newman, M. E. (2004a), 'Detecting community structure in networks', *The European Physical Journal B* **38**(2), 321–330.
- Newman, M. E. (2004b), 'Fast algorithm for detecting community structure in networks', *Physical review E* **69**(6), 066133.
- Newman, M. E. (2006), 'Modularity and community structure in networks', *Proceedings of the national academy of sciences* **103**(23), 8577–8582.
- Newman, M. E. (2013), 'Spectral methods for community detection and graph partitioning', *Physical Review E* **88**(4), 042822.
- Newman, M. E. & Girvan, M. (2004), 'Finding and evaluating community structure in networks', *Physical review E* **69**(2), 026113.
- Newman, M. E. & Park, J. (2003), 'Why social networks are different from other types of networks', *Physical review E* **68**(3), 036122.
- NexGate (2013), 'State of Social Media Spam Research report', Online. Accessed: 18-02-2018.
- Olson, R. S., Cava, W. L., Mustahsan, Z., Varik, A. & Moore, J. H. (2018), 'Data-driven Advice for Applying Machine Learning to Bioinformatics Problems', *Pacific Symposium on Biocomputing (PSB) 2018 Online Proceedings* **23**, 192–203.
- Palla, G., Barabási, A.-L. & Vicsek, T. (2007), 'Quantifying social group evolution', *Nature* **446**(7136), 664.
- Pang, B., Lee, L. et al. (2008), 'Opinion mining and sentiment analysis', *Foundations and Trends® in Information Retrieval* **2**(1–2), 1–135.
- Papadopoulos, S., Kompatsiaris, Y., Vakali, A. & Spyridonos, P. (2012), 'Community detection in social media', *Data Mining and Knowledge Discovery* **24**(3), 515–554.

- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. & Duchesnay, E. (2011), ‘Scikit-learn: Machine learning in python’, *Journal of Machine Learning Research* **12**, 2825–2830.
- Pennington, J., Socher, R. & Manning, C. (2014), Glove: Global vectors for word representation, in ‘Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)’, pp. 1532–1543.
- Pizzuti, C. (2008), Ga-net: A genetic algorithm for community detection in social networks, in ‘International conference on parallel problem solving from nature’, Springer, pp. 1081–1090.
- Pons, P. & Latapy, M. (2006), ‘Computing communities in large networks using random walks.’, *J. Graph Algorithms Appl.* **10**(2), 191–218.
- Pothen, A., Simon, H. D. & Liou, K.-P. (1990), ‘Partitioning sparse matrices with eigenvectors of graphs’, *SIAM journal on matrix analysis and applications* **11**(3), 430–452.
- Qazvinian, V., Rosengren, E., Radev, D. R. & Mei, Q. (2011), Rumor has it: Identifying Misinformation in Microblogs, in ‘Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, (EMNLP), A meeting of SIGDAT, a Special Interest Group of the ACL’, Edinburgh, UK, pp. 1589–1599.
- Radicchi, F., Castellano, C., Cecconi, F., Loreto, V. & Parisi, D. (2004), ‘Defining and identifying communities in networks’, *Proceedings of the national academy of sciences* **101**(9), 2658–2663.
- Raghavan, U. N., Albert, R. & Kumara, S. (2007), ‘Near linear time algorithm to detect community structures in large-scale networks’, *Physical review E* **76**(3), 036106.
- Rojas, E., Munoz-Gama, J., Sepúlveda, M. & Capurro, D. (2016), ‘Process mining in health-care: A literature review’, *Journal of Biomedical Informatics* **61**, 224–236.
- Rosa, K. D., Shah, R., Lin, B., Gershman, A. & Frederking, R. (2011), ‘Topical clustering of tweets’, *Proceedings of the ACM SIGIR: SWSM* .

- Ruchi, P. & Kamalakar, K. (2013), Et: Events from tweets, *in* 'Proc. 22nd Int. Conf. World Wide Web Comput., Rio de Janeiro, Brazil', pp. 613–620.
- Salton, G. & Buckley, C. (1988), 'Term-weighting approaches in automatic text retrieval', *Information processing & management* **24**(5), 513–523.
- Salvatier, J., Wiecki, T. V. & Fonnesbeck, C. (2016), 'Probabilistic programming in python using pymc3', *PeerJ Computer Science* **2**, e55.
- Scott, J. (1988), 'Social network analysis', *Sociology* **22**(1), 109–127.
- Shaw, M. E. (1971), 'Group dynamics: The psychology of small group behavior'.
- Shi, J. & Malik, J. (2000), 'Normalized cuts and image segmentation', *Departmental Papers (CIS)* p. 107.
- Simmel, G. et al. (1950), 'The stranger', *The Sociology of Georg Simmel* **402**, 408.
- Strauss, M. J., Bradley, G. L. & Smith, K. J. (2002), *Multivariable Calculus, 3rd Edition*, Prentice Hall.
- Subrahmanian, V. S., Azaria, A., Durst, S., Kagan, V., Galstyan, A., Lerman, K., Zhu, L., Ferrara, E., Flammini, A. & Menczer, F. (2016), 'The DARPA Twitter Bot Challenge', *IEEE Computer* **49**(6), 38–46.
- Sundaram, H., Lin, Y.-R., De Choudhury, M. & Kelliher, A. (2012), 'Understanding community dynamics in online social networks: a multidisciplinary review', *IEEE Signal Processing Magazine* **29**(2), 33–40.
- Szüle, J., Kondor, D., Dobos, L., Csabai, I. & Vattay, G. (2014), 'Lost in the city: Revisiting milgram's experiment in the age of social networks', *PloS one* **9**(11), e111973.
- Thomas, K., Grier, C., Ma, J., Paxson, V. & Song, D. (2011), Design and Evaluation of a Real-Time URL Spam Filtering Service, *in* '32nd IEEE Symposium on Security and Privacy (S&P)', Berkeley, California, USA, pp. 447–462.
- Travers, J. & Milgram, S. (1977), An experimental study of the small world problem, *in* 'Social Networks', Elsevier, pp. 179–197.

- Tromble, R., Storz, A. & Stockmann, D. (2017), We don't know what we don't know: When and how the use of twitter's public apis biases scientific inference, *in* 'SSRN'.
- Tsur, O., Littman, A. & Rappoport, A. (2013), Efficient clustering of short messages into general domains, *in* 'Seventh International AAAI Conference on Weblogs and Social Media'.
- Varol, O., Ferrara, E., Davis, C. A., Menczer, F. & Flammini, A. (2017), Online Human-Bot Interactions: Detection, Estimation, and Characterization, *in* 'International AAAI Conference on Web and Social Media', AAAI Press, pp. 280–289.
- Vicent, C. & Moreno, A. (2014), Unsupervised semantic clustering of twitter hashtags., *in* 'ECAI', Citeseer, pp. 1119–1120.
- Wang, B., Zubiaga, A., Liakata, M. & Procter, R. (2015), Making the Most of Tweet-Inherent Features for Social Spam Detection on Twitter, *in* 'Proceedings of the the 5th Workshop on Making Sense of Microposts, co-located with the 24th International World Wide Web Conference (WWW)', Florence, Italy, pp. 10–16.
- Wang, W. Y. (2017), '" liar, liar pants on fire": A new benchmark dataset for fake news detection', *arXiv preprint arXiv:1705.00648* .
- Waniek, M., Michalak, T. P., Wooldridge, M. J. & Rahwan, T. (2018), 'Hiding individuals and communities in a social network', *Nature Human Behaviour* **2**(2), 139.
- Wasserman, S. & Faust, K. (1994), *Social network analysis: Methods and applications*, Vol. 8, Cambridge university press.
- Watts, D. J. & Dodds, P. S. (2007), 'Influentials, networks, and public opinion formation', *Journal of consumer research* **34**(4), 441–458.
- Watts, D. J. & Strogatz, S. H. (1998), 'Collective dynamics of 'small-world' networks', *nature* **393**(6684), 440.
- Weng, J., Lim, E.-P., Jiang, J. & He, Q. (2010), Twitterrank: finding topic-sensitive influential twitterers, *in* 'Proceedings of the third ACM international conference on Web search and data mining', ACM, pp. 261–270.

- Wiemer-Hastings, P., Wiemer-Hastings, K. & Graesser, A. C. (2004), Latent Semantic Analysis, in 'Proceedings of the 16th international joint conference on Artificial intelligence', pp. 1–14.
- Williams, R. J. & Martinez, N. D. (2000), 'Simple rules yield complex food webs', *Nature* **404**(6774), 180.
- Wilson, C., Boe, B., Sala, A., Puttaswamy, K. P. & Zhao, B. Y. (2009), User interactions in social networks and their implications, in 'Proceedings of the 4th ACM European conference on Computer systems', Acm, pp. 205–218.
- Würschinger, Q., Elahi, M. F., Zhekova, D. & Schmid, H.-J. (2016), Using the web and social media as corpora for monitoring the spread of neologisms. the case of 'rapefugee', 'rapeugee', and 'rapugee', in 'Proceedings of the 10th Web as Corpus Workshop', pp. 35–43.
- Yali, P., Jian, Y., Shaopeng, L. & Jing, L. (2014), A biterm-based dirichlet process topic model for short texts, in '3rd International Conference on Computer Science and Service System', Atlantis Press.
- Yan, X., Guo, J., Lan, Y. & Cheng, X. (2013), A biterm topic model for short texts, in 'Proceedings of the 22nd international conference on World Wide Web', ACM, pp. 1445–1456.
- Yang, C., Harkreader, R., Zhang, J., Shin, S. & Gu, G. (2012), Analyzing Spammers' Social Networks for Fun and Profit: A Case Study of Cyber Criminal Ecosystem on Twitter, in 'Proceedings of the 21st International Conference on World Wide Web', WWW '12, ACM, New York, NY, USA, pp. 71–80.
- Yang, J. & Leskovec, J. (2012), Community-affiliation graph model for overlapping network community detection, in '2012 IEEE 12th International Conference on Data Mining', IEEE, pp. 1170–1175.
- Yang, J. & Leskovec, J. (2015), 'Defining and evaluating network communities based on ground-truth', *Knowledge and Information Systems* **42**(1), 181–213.
- Yang, J., McAuley, J. & Leskovec, J. (2013), Community detection in networks with node attributes, in '2013 IEEE 13th International Conference on Data Mining', IEEE, pp. 1151–1156.

- Yang, Y. (2001), A study of thresholding strategies for text categorization, in 'Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval', ACM, pp. 137–145.
- Yao, Y., Viswanath, B., Cryan, J., Zheng, H. & Zhao, B. Y. (2017), Automated Crowdturfing Attacks and Defenses in Online Review Systems, in 'Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security (CCS)', Dallas, TX, USA, pp. 1143–1158.
- Yee, K. C., Miils, E. & Airey, C. (2008), 'Perfect Match? Generation Y as Change Agents for Information Communication Technology Implementation in Healthcare', *Studies in Health Technology and Informatics* **136**, 496–501.
- Yoshida, T. (2013), 'Toward finding hidden communities based on user profile', *Journal of Intelligent Information Systems* **40**(2), 189–209.
- Yu, H., Kaminsky, M., Gibbons, P. B. & Flaxman, A. D. (2008), 'SybilGuard: Defending Against Sybil Attacks via Social Networks', *IEEE/ACM Transactions on Networking* **16**(3), 576–589.
- Zhang, C. & Zaïane, O. R. (2018), Detecting local communities in networks with edge uncertainty, in '2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)', IEEE, pp. 9–16.
- Zhang, X., Zhao, J. & LeCun, Y. (2015), Character-level convolutional networks for text classification, in 'Advances in neural information processing systems', pp. 649–657.
- Zhou, Y., Cheng, H. & Yu, J. X. (2009), 'Graph clustering based on structural/attribute similarities', *Proceedings of the VLDB Endowment* **2**(1), 718–729.
- Zhu, X. & Ghahramani, Z. (2002), Learning from labeled and unlabeled data with label propagation, Technical report, Citeseer.

Appendix

Appendix A

SUPPLEMENTARY INFORMATION

1.1 Content Authentication

The following section provides additional information about the dataset utilised in Chapter V. The dataset consists of a collection of tweets, collected via an application programming interface (API) provided by the platform.

1.1.1 Tweet Object

Managing a tweet is quite a daunting task because each *tweet object* has over 100 *metadata*, and each describing a different aspect or property of the tweet object. This also makes it easy to grow in size. It is common for handfults of complete tweets can consume a considerable amount of space and offset the performance of the processor. Careful filtering is required to extract relevant features. Some relevant fields in a *tweet object* are shown below:

```
{"created_at": "Wed Sep 20 16:51:22 +0000 2017", "id": "###", "id_str": "###", "text": "RT @kpnewschannel: Caught in the ...", "source": "href=\\\"http:url\\\"", "truncated": false, "in_reply_to_status_id": null, "in_reply_to_status_id_str": null, "in_reply_to_user_id": null, "in_reply_to_user_id_str": null, "in_reply_to_screenname": null, "user": {"id": "###", "id_str": "###", "name": "###", "screen_name": "###", "location": " u092d\\u093", "url": null, "description": ".. !!", "translator_type": "none", "protected": false, "verified": false, "followers_count": 341, "friends_count": 364, "listed_count": 9, "favourites_count": 10919, "statuses_count": 24122, "created_at": "Sat Nov 01 10:34:43 +0000 2014", "utc_offset": null, "time_zone": null, "geo_enabled": true, "lang": "en", "contributors_enabled": false, "is_translator": false, "profile_background_color": "CODEED", "retweeted": false, "favorited": false, "filter_level": "low", "lang": "en", "timestamp_ms": "1505926282748"}
```

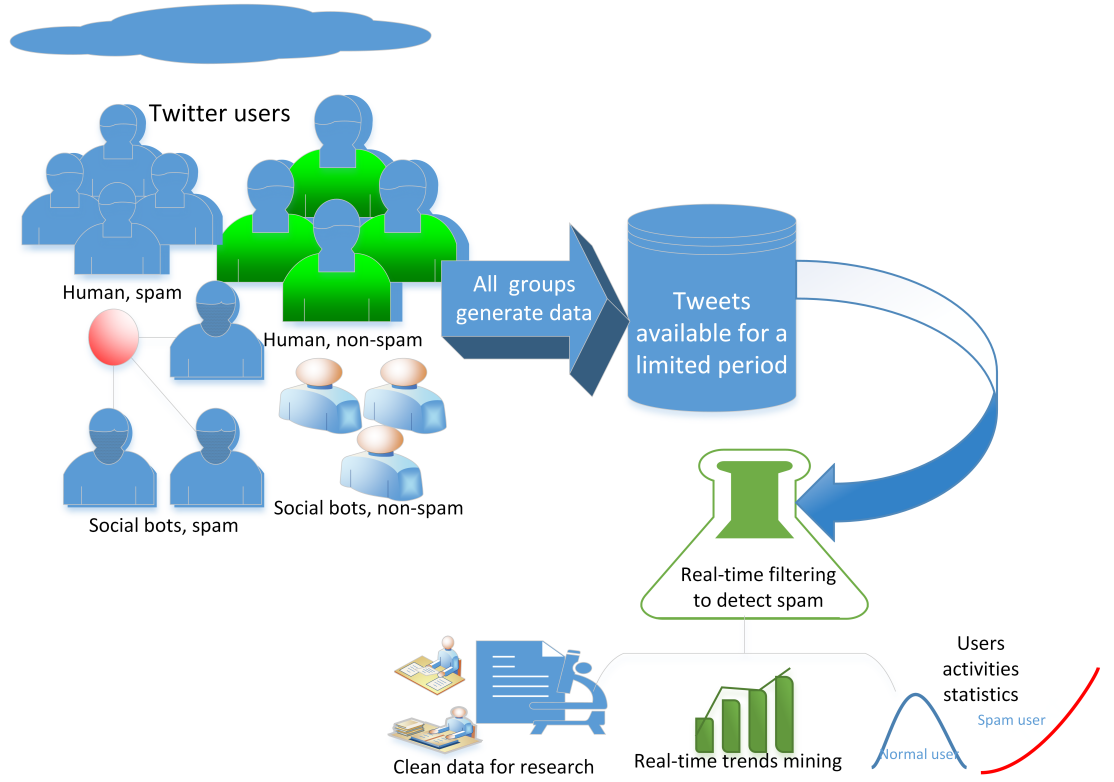


Figure A.1: An overview of *content authentication* activity

Figure A.1 shows a visual depiction of the research processes involve in validating the research data.

1.1.2 SPD Features

As detailed in Chapter V, the *SPD datasets* consist of the following: *Honeypot dataset*, a public dataset from (Lee et al. 2011), ($SPD_{automated}$), and (SPD_{manual}). These datasets have been fully described in Chapter V and Table 5.1. Table A.1 shows the features proposed for prediction model training, the corresponding feature groups and definitions. The *VerifiedAccount* feature, labelled as f_{22} , takes on binary values, ‘1’ for verified accounts or ‘0’ otherwise. These values reflect the target labels in the user profile meta-data. The feature was used in the feature set for training classification models during our early experiments. The resulting model overfitted the training data and, for this reason, the feature was later removed due to its role in leaking the correct prediction into the test data (Forman & Scholz 2010).

Table A.1: The complete set of the proposed features used in the spam detection study; each feature is associated with a group and relevant definition. The *VerifiedAccount* feature, f_{22} , was excluded from the final feature set, because in preliminary experiments it was shown to cause overfitting.

Id	Features	Groups	Status	Description/Definition
f_1	<i>AccountAge</i>	AIF	static	days since account creation to date of collection
f_2	<i>FollowersCount</i>	EbF	dynamic	in user profile meta-data
f_3	<i>FriendsCount</i>	EwF	dynamic	in user profile meta-data
f_4	<i>StatusesCount</i>	EwF	dynamic	in user profile meta-data
f_5	<i>DigitsCountInName</i>	UPF	static	number of digits in screen name
f_6	<i>TweetLen</i>	EwF	dynamic	number of characters in tweet
f_7	<i>UserNameLen</i>	UPF	static	number of characters in user name
f_8	<i>ScreenNameLen</i>	UPF	static	number of characters in screen name
$f_{9,10,11,12}$	<i>Metric entropy</i> for all textual features: tweet, user profile description, user name and screen name, respectively	UPF	dynamic	to measure randomness in text. $\frac{H(x)}{ x }$: where $ x $ is the length of a string, x , and $H(x)$ is the Shannon entropy of text: $\sum_{i=1..k} p_i \log_2 p_i$
f_{13}	<i>URIsRatio</i>	EwF	dynamic	$\frac{ \text{characters in URLs} }{ \text{tweet length} }$
f_{14}	<i>MentionsRatio</i>	EwF	dynamic	$\frac{ \text{characters in user mentions} }{ \text{tweet length} }$
f_{15}	<i>NameSim</i>	UPF	static	% proportion of similarity in User name and Screen name
f_{16}	<i>LexRichWithUU</i>	EwF	dynamic	TTR in tweets: $\frac{ \text{token types} }{ \text{total tokens} } * 100$
f_{17}	<i>Friendship</i>	EwF	dynamic	$\frac{FriendsCount}{FollowersCount}$
f_{18}	<i>Followership</i>	EbF	dynamic	$\frac{FollowersCount}{FriendsCount}$
f_{19}	<i>Interestingness</i>	EbF	dynamic	$\frac{FavouritesCount}{StatusesCount}$
f_{20}	<i>Activeness</i>	EwF	dynamic	$\frac{StatusesCount}{AccountAge}$
f_{21}	<i>LexRichWithOutUU</i>	EwF	dynamic	$\frac{ \text{lexical worlds} }{ \text{total number of words} } * 100$
f_{22}	<i>VerifiedAccount*</i>	AIF	static	in tweet metadata
f_{23}	<i>FavouritesCount</i>	EwF	dynamic	in user profile meta-data
f_{24}	<i>NamesRatio</i>	UPF	static	$\frac{ \text{screenname length} }{ \text{username length} }$

1.2 Search Optimisation

The following section describes the data collected for investigating search enhancement and various benchmark datasets during the early stage of the research. Because the research direction has shifted to a different approach, a summary of the experimentation follows. For training and evaluation purposes, two sets of data (see Table A.2) have been used: *subject-based tweets (SBT)* and *diverse tweets (DVT)* and were both collected using the *adhoc retrieval* method, which involves the use of descriptive keywords to search for relevant documents (Manning et al. 1999). In addition to the *SBT* and *DVT*, two sets of publicly available data have been used

Table A.2: Datasets and features (main and meta features comprising the tweet signature)

	Group	Pairwise Size	Unique	Description
Datasets	Diverse tweets (DVT)	35m	300000K	consists of random tweets collected using diverse keywords covering many domains (based on Analytics (2009)) to introduce a high level of randomness and improve the universality of the dataset.
	Subject-based tweets (SBT)	45m	300000K	consists of tweets collected in 2016/2017 related to EU refugee crisis
	Drug review Gräßer et al. (2018)	3107	602400	data about patients' reviews on specific drugs and related conditions
	Hotel review Calheiros et al. (2017)	400	9336	online and offline collections of customers review from on hotel service
	Health tweets Karami et al. (2018)	17413	334140	contains tweets about health news from major health news agencies
	Eur-Lex Mencia & Fürnkranz (2008)	62311	12353646	data about EU legal documents

for evaluation. Using data other than tweets counters the effect of their stochastic nature and the black box sampling strategy that Twitter uses to make them available (Tromble et al. 2017). For more details, see Inuwa-Dutse et al. (2019a).

1.3 Dyadic and Transitive Datasets

Dyadic dataset collection begins with 4022 *seed users* from *verified* and *unverified* accounts. The *seed users* are genuine users devoid of spammers or social bots collected based on the SPD filtering technique Inuwa-Dutse, Liptrott & Korkontzelos (2018). A collection crawler is designed to search the profile of each user's network (consisting of lists of friends and followers) to determine the set of users with a reciprocal tie with the seed user. Table 6.1 of Chapter VI shows basic statistics of users visited by the collection crawler.

Transitive dataset is collected in a similar manner to the *dyadic* counterpart. Essentially, the *transive collection* is a scaled version of the *dyadic data*. Table 6.2 of Chapter VI shows a summary of the *transitive dataset*.

1.4 The MCT Framework

Information about the *MCT framework* and datasets used in the detection task. Figure 4.3 shows a summary of the stages in the *MCT framework*. Figure 4.3 shows a generic version of the MCT, which recognises both related and unrelated aspects of structural and textual components.

1.4.1 Microcosm detection dataset

The following section describes the data¹ used in the detection task. Due to the availability of empirical data, the effectiveness of the model is quantified with respect to the degree of conformity with the data. The ground-truth was used earlier in Chapter VI, and supplementary information about the data is given in Figure A.2 showing relevant *nodes' degree distribution*. For the *degree distribution*, it shows less reciprocity, about 70% – 80% of the nodes have no more than two orders of degree. Moreover, there exist many nodes with few connections; this observation was echoed earlier in Chapter VI.

Facebook Data

The data from Facebook consists of the following high-level features. Table A.3 depicts how the data is structured². For each node, if the above-mentioned feature is available, a value of

Table A.3: Description of the Facebook data showing examples of high-level features

Feature	Feature
birthday-anonymised feature #	education-classes-id-anonymised feature #
education-concentration-id-anonymised feature #	education-degree-id-anonymised feature #
education-school-id-anonymised feature #	first name-anonymised feature #
work-position; location-id-anonymised feature #	last name-anonymised feature #
work-employer-id-anonymised feature #	work-end date-anonymised feature #
work-start date-anonymised feature #	home town-id-anonymised feature #
education-year-id-anonymised feature #	gender-anonymised feature #
education-degree-id-anonymised feature #	—

1 is used, otherwise 0. The *anonymised feature #* is not clear what it is, but it is sensitive to share because it could lead to identification. These are the features and have been used to suit the research implementation. All nodes with similar groups/circles are compared according to

¹For more details, see https://github.com/ijdutse/mct/blob/master/MCT-datasets_and_meta-analysis.ipynb

²More detailed information about the data and preprocessing is available at https://github.com/ijdutse/mct/blob/master/MCT-datasets_and_meta-analysis.ipynb

the features in the circle (by focusing on which features overlap and for which node in which group).

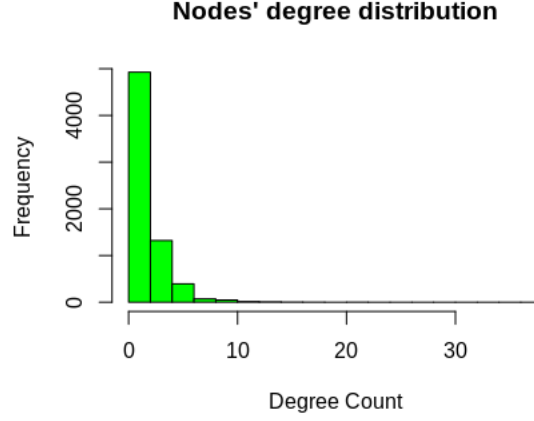


Figure A.2: Example of nodes with very high and low degree frequency.

1.4.2 MCT optimisation

For a more intuitive description, the *MCT framework* can be considered as a form of *multivariate function* comprising of *structural* and *textual* variables. This specification makes it suitable to apply an *optimisation function* that maximises the overall joint similarity. Accordingly, the *optimisation* problem can be analysed or interpreted geometrically. The *joint similarity functions* can be bounded by $(0, 1]$, meaning they cannot exceed 1, then the constrained function can be viewed as the equation of a unit circle in an n -dimensional space (see Figure A.3). The constrained function (depicted by a circle) and the optimisation function (the joint similarity function) are represented by the contour lines denoting possible values from the joint similarity of both $\phi(S_p)$ and $\phi(T_p)$.

Gradient of the functions: A gradient field ∇ , at any point in the space is a vector that can be observed in the surface of the functions and has a useful property of being perpendicular (\perp) to both the functions. Essentially, the *gradient* is considered as an *operator* on a function that produces a vector (Strauss et al. 2002). Thus, the gradient of the *optimisation function* $\nabla \mathcal{F}$ evaluated under a constraint is given by:

$$\nabla \mathcal{F}(\phi(\mathcal{S}_r), \phi(\mathcal{T}_r)) = \lambda \nabla (\mathcal{M}_{vt} - \mathcal{M}_{cvt}) \quad (1.1)$$

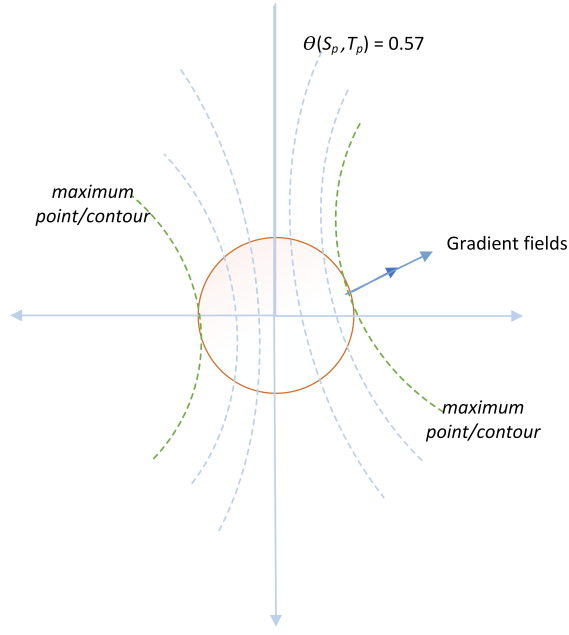


Figure A.3: A geometric interpretation of the optimisation problem constrained by a unit circle equation. The goal is to be or contain lines (signifying values of the optimisation function) to be as close as possible to the points in asterisk (*) signifying the points in the function (but cannot be outside the circle). The points close to the boundary of the constrained function signify high similarity and the constrained function can be seen as a subset of the function, but with a different property imposed by a limit (cannot be > 1). Each contour line represents a specific value, e.g. $\phi(S_p, T_p) = 0.67$ and it does not change along the contour.

where λ is the *Lagrange multiplier*, serving as the proportionality constant. The *objective function* (Eq. 7.16) and the *constrained function* (Eq. 7.15) can be described geometrically in which the maximum value is obtained when the *objective function* coincides with the highest peak in the *constrained function*. For illustration, the problem corresponds to the tangent of the *functions* which can be solved by leveraging the idea of *gradient fields* in the surface of the functions, c.f. Figure A.3. In Figure A.3, the maximum value is the point where the *contour lines* and the boundary of the *constrained function* $g(\phi(S_p), \phi(T_p)) = 1^3$, coincide and corresponds to the tangent of the function in which *gradient fields* (in the surface of the function) can be applied. The *gradient fields* have the property of being perpendicular (\perp) to both the *contours* and the *constraint*. A switch between contour lines is possible, which can be observed since moving along the contour does not change or alter the value of a function, but across the contour does. It follows that all values along the contours are constant because each contour represents a single value. The gradient field ∇ , at any point in the space, is a vector that can be observed in the

³This is the same with $g(\phi(S_r), \phi(T_r)) = 1$

surface of the functions; hence, the two functions can be equated.

1.4.3 Matrix decomposition

The *reciprocity matrix* of the nodes is decomposed into its approximate constituents according to a generic framework for *nonnegative matrix transformation*. In modelling the structural clusters, the following matrices of interactions and corresponding dimensions are used:

- $\mathcal{M}_{scva} \mapsto n \times n$: *adjacency matrix of structurally-related nodes*
- $\mathcal{M}_{cvd} \mapsto n \times n$: *nodes diagonal matrix*
- $\mathcal{M}_{cvi} \mapsto n \times n$: *nodes Laplacian matrix*
- $\mathcal{M}_{cvt} \mapsto n \times k$: *a collection of nodes according to reciprocal-communities*
- $\mathcal{M}_{cra} \mapsto k \times k$: *reciprocal-communities adjacency matrix*
- $\mathcal{M}_{crd} \mapsto k \times k$: *reciprocal-communities diagonal matrix*

Frobenius Norm ($\|\cdot\|_F^2$) and Residual Matrix In line with the tenets of optimisation models with constraints, the goal is to maximise similarity between entries of D and P, Q^T subject to nonnegative entries in P, Q . The optimisation of Eq. 7.9 is based on the squared *Frobenius norm* ($\|\cdot\|_F^2$) of the model. The squared *Frobenius norm*⁴, returns the sum of the squares of the entries in the *residual matrix*⁵. After each iteration, the error magnitude is compared, and the goal is to minimise the margin between the actual and the predicted entries. The iterative update of the parameters (p_{ij} and q_{ij}) continues until convergence.

Data repositories

Some relevant files available at the *Github repository*⁶:

- *SPD*: <https://github.com/ijdutse/spd>

⁴also referred to as the *energy* due to summation over the second moments (i.e. the variance) over all data points about the origin

⁵containing the residue $D - PQP^T$ of the original matrix, i.e. the residual errors obtained from a low-rank factorisation of the original matrix D

⁶<https://github.com/>

- *Dyads*: https://github.com/ijdutse/dyads_in_Twitter
- *Simmelian ties*: https://github.com/ijdutse/simmelian_ties_on_Twitter
- *MCT*: <https://github.com/ijdutse/mct>