Understanding the Psycho-Sociological Facets of Homophily in Social Network Communities

R Sudhesh Solomon, Srinivas P Y K L, Amitava Das
Dpt. of Computer Science and Engineering, Indian Institute of Information Technology, Sri City, INDIA
Björn Gambäck
Dpt. of Computer Science, Norwegian University of Science and Technology, Trondheim, NORWAY
Tanmoy Chakraborty
Dpt. of Computer Science and Engineering, Indraprastha Institute of Information Technology Delhi, INDIA

Abstract—"Community" in social networks is a nebulous concept. A community is generally assumed to be formed by people who possess similar attributes or characteristics, also known as "homophily". Although there has been a lot of research on community detection based on network topology, the semantic interpretation of communities is rarely studied. The present work aims to understand the behavioural similarity of users present in their personal neighbourhood communities formed by *friends*, *relatives*, or *colleagues*, and addresses two fundamental questions: (i) *Are communities formed by users who possess similar behavioural traits? If so, does this apply to all those sub-networks, i.e., friends, relatives, and colleagues?* (ii) *Does adding behavioural node-specific attributes/features to the nodes in a network lead to better community detection?* To better understand the psycho-sociological homophilic nature of personal networks, the personalities and values of Twitter users were analysed using the well-established "Big-5 personality model" and "Schwartz sociological behaviour model". Empirical results based on the psycho-sociological behaviour show that friends networks exhibit homophily, whereas relatives and colleagues networks do not exhibit such homophilic behaviour. It can also be observed that neurotic people tend to behave heterogeneously with people of various personality traits. In addition, it is shown that such empirical evidence can be used as features for the tasks of community detection and link prediction.

1 Introduction

Social networks are playing a progressively paramount role in our existence, especially in the spread of convictions and feelings among their users. The availability of huge volumes of social interaction data has opened up several new possibilities to unleash a grand avenue for research in the field of computational social science. The present paper aims to seek answers to the following questions:

(i) What are the psycho-sociological facets that govern the natural selection of societal relationships, such as friends, relatives, and colleagues?
(ii) Can we identify different psycho-sociological facets automatically?
(iii) Can psycho-sociological features be used as user-centric properties to detect community structure and to predict emerging links more accurately?

1.1 The Paradox of Homophily

Loosely, *homophily* (from Ancient Greek ὁμοῦ 'together' and φιλία 'friendship') refers to the tendency for people of having (non-negative) ties with other people who are similar to themselves in significant ways. However, there is no consensus on the exact meaning of homophily and what the definite facets of such similarities are. Here, we argue that someone's individuality (psychological) and societal upbringings define the facets of their homophily nature; Section 2.1 compares this to the spectrum of ways homophily is addressed in the literature.

Corresponding Author: Amitava Das (amitava.santu@gmail.com)

Table 1: Big-5 personality traits. The Personality model aids in understanding characteristics or a blend of characteristics at an individual level.

Personality	Behavioural Characteristics
Openness [**O**]	Imaginative, insightful, and have wide interest
Conscientiousness [**C**]	Organised, thorough, planned, and punctual
Extroversion [**E**]	Articulative, boastful, and energetic
Agreeableness [**A**]	Amiable, generous, co-operative, and altruistic
Neuroticism [**N**]	Anxious, timid, immature, and unstable in their actions.

In this paper, the theoretical point of departure is in psycholinguistic models. The psycho-sociological backgrounds of individuals play a crucial role in determining which communities they will belong to and which ones they will leave or join in the future. To understand someone's personality we have borrowed from Psychology the well-established Big-5 [1] personality model also known as the Five Factor Model (FFM) or the OCEAN model: *Openness*, *Conscientiousness*, *Extroversion*, *Agreeableness*, and *Neuroticism*, as described in Table 1. Our argument is that the Big-5 Personality model could be further considered as a personal level sentiment model, whereas the normal sentiment analysis on only text considers a ternary class of separation: *positive*, *negative*, and *neutral*. To understand the behaviour of someone in accordance to society, Schwartz' value model [2] is used. This model defines ten distinct ethical values that are illustrated in Table 2. Gavrilescu and Vizireanu [3] proposed a neural network based model that determines the Big-5 personality traits of an individual by analysing offline handwriting. Zhong et al. [4] proposed

Table 2: Schwartz' value model.

Values	Behavioural Characteristics
Achievement [**AC**]	sets goals; focused towards achieving them
Benevolence [**BE**]	help others; works on general welfare
Conformity [**CO**]	follows rules, laws and structures
Hedonism [**HE**]	seeks pleasure and enjoyment
Power [**PO**]	dominates and controls others and resources
Security [**SE**]	seeks safety, security, and social stability
Self-direction [**SD**]	free and independent in thoughts and actions
Stimulation [**ST**]	seeks excitement and thrills
Tradition [**TR**]	accepts customs and ideas provided by religion
Universalism [**UN**]	prefers peace; works toward social welfare

a personality prediction framework, consisting of outlier elimination, training dataset selection and personality prediction. Rammstedt et al. [5] comprehensively investigated the associations between both fluid and crystallised intelligence with Big-5 personality domains as well as their facets. Anglim et al. [6] examined the correlates of Schwartz' basic values with the broad and narrow traits of the HEXACO model of personality. Zhang et al. [7] assessed the possible correlations between personality traits and face images. Xu et al. [8] proposed a multi-view facial feature extraction model is proposed to evaluate the possible correlation between personality traits and face images.

The classical definition of homophily only considers that similar people can come together and become friends. Indeed, such homogeneity is perceived to a large extent in society, but there are specific nuances in human-human relationships that might get overlooked in the paradox of homophily. For example, extroverts can really handle neurotic friends whereas two highly sentimental (neurotic) people may not ideally complement to each other. To understand such relational nuances, the present research addresses issues such as:

(i) Do people seek friendship with others with similar personality traits and/or having similar societal values?
(ii) Are relatives more societal homophile than friends?
(iii) Where do colleagues fit on the personality/value spectra?

1.2 Link Prediction

Understanding the evolutionary dynamics of a social network is complicated, since network structure is not static, but varies over time. Some of the possible changes in the network may include: deletion of links, creation of new links, and addition of new nodes. Although the evolution of a network can be categorised based on various factors, the main aspect considered in this paper is *link prediction*, i.e., understand in which types of network it is possible to predict with the highest accuracy that a new link between two nodes (persons) will be established in the future.

1.3 Communities — Understanding Who's Who

The word "community" is derived from the Latin *communis*, meaning to have something in common. Communities in social networks are generally assumed to be formed by people possessing similar attributes and characteristics. Although there has been a plethora of work ([9–13]) on understanding network topology (edge density, clustering coefficients, etc.)

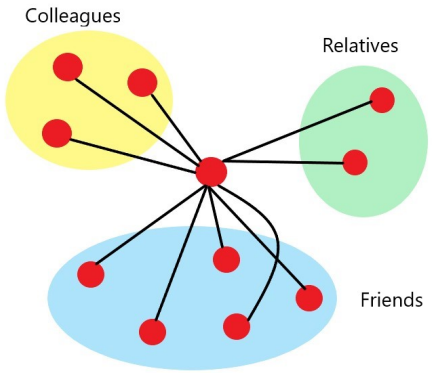

Figure 1: Personal communities of an individual in a network.

within a community, the semantic interpretation of a community has barely been contemplated. Here the questions are whether individuals in a community possess similar personalities, values, and ethical backgrounds, and whether community structure be discovered more accurately if we know psycho-sociological traits. We are interested to learn how the nature of homophily changes in terms of user-user relationships with friends, relatives, and colleagues. The annotation of such relationships was crowd-sourced, as described in detail in Section 3. Figure 1 depicts an ego-centric network of a particular user and shows their relationships with other users in different communities: *friends*, *relatives*, and *colleagues*.

The remainder of this paper is structured as follows. Section 2 discusses related work. Section 3 presents the data collection process. Section 4 explains the classifiers built to assign personality and values to each user in the network. Section 5 then provides a complete description of the methodology. Section 6 presents three link prediction models and a comparative study to analyse the performance of the model in each categorical network (friends, relatives, and colleagues), while Section 7 shows how communities can be detected without node features and how the addition of appropriate features to nodes can increase community detection accuracy. Finally, Section 8 concludes this paper and discusses future work.

2 Related Work

As in Section 1, the description of related work will be structured around the paper's three main themes: the psycho-sociological facets underlying the homophily that governs the selection of societal relationships, ways to automatically identify such relationships and links between individuals, and ways to detect overall community structures.

2.1 Homophily — the Facets of Commonality

The phenomenon of homophily has been studied for a long time in the fields of social science and social network analysis [14–21]. Previous work has focused on understanding the homophily phenomenon mostly using two main approaches: i) investigating real-world socio-demographic information (age, gender, education, occupation, school, workplace, home-town, etc.) or social media demographics

(interests, opinions, perspectives, etc.), or ii) complex network analysis, where homophily is considered as a similarity metric (*assortativity*), representing to what extent nodes in a network are associated with each other.

2.1.1 User Demographic and Properties

Bisgin et al. [16] hypothesised that new ties are formed between individuals who demonstrate similar characteristics, either in the real world or in social media. However, retrieving demographic information from social media is hard as most of the times it is missing, and even if it is available, it might not be trustworthy. Therefore, problems such as identifying age, gender, income, and religious beliefs [20, 22, 23] are all addressed as topical research problems. Furthermore, user personality and their psycho-sociological behaviour change with their demographic. An empirical analysis and an interactive map are available from the University of Pennsylvania.[1] Wilson et al. [24] present a cross-cultural analysis of value-behaviour relationships spanning writers from the United States and India, based mainly on language in social media. Still, all such characteristic features are explicit in nature; here we concentrate more on psychological and sociological (personality and values) characteristics, that are innate but often implicit.

For the same reason, Bisgin et al. [16] only analysed interest-based homophily through i) dyadic relations: identifying the overlap of interest among individuals (those with no common interest and those with one or more common interest), ii) community structure: clustering users into communities based on their interests; using the Fast Modularity [25] and Graculus [26] algorithms for the interest-based community identification, and iii) content: identifying user interests based on the blog posts that they share.

Anagnostopoulos et al. [14] tried to understand the relationship and influence that exist among users in social networks. They explained the importance of achieving social ties through *social influence*, the ways by which users induce their friends to show similar behaviour in tagging (in this case the usage of tags/keywords on Flickr). Kafeza et al. [27] argued that users in social media communities are more probable to make contact if they are not "biased" with respect to user personality.

Several other types of information sources have also been used to extract personalities and relations. So did Gavrilescu and Vizireanu [3] propose a neural network based model that determines the Big Five personality traits of an individual by analysing offline handwriting. Zhong et al. [4] utilised a personality prediction framework consisting of outlier elimination, training dataset selection and personality prediction. Rammstedt et al. [5] comprehensively investigated the associations between both fluid and crystallised intelligence with Big Five personality domains as well as their facets. Anglim et al. [6] examined the correlations of Schwartz' basic values with the broad and narrow traits of the HEXACO personality model. Zhang et al. [7] assessed the possible correlations between personality traits and face images. Xu et al. [8] proposed a multi-view feature extraction model to evaluate the possible correlation between personality traits and facial images.

1. https://map.wwbp.org/

2.1.2 Assortativity

A network is said to be assortative if nodes with higher degrees are connected with other nodes that have similar high degrees. Newman [28] showed that social networks tend to be assortative, while other networks such as technological (internet, world wide web) and biological (protein interactions, food web) are disassortative, i.e., nodes with higher degrees are connected with nodes with lower degrees in the network. Mulders et al. [21] proposed a method to show how assortativity could be used to enhance the demographic predictions of individuals within a social network.

2.2 Link Prediction in Online Social Networks

Studying whether a link could be established between a pair of users in a social network is called link prediction. It could be used for friend recommendation, community evolution, and for solving several other real-life problems. Studies on link prediction can be categorised into three broad genres: i) similarity-based approaches, ii) path-based approaches, and iii) learning-based approaches.

2.2.1 Similarity-Based Approach

This is a node-centric approach. The similarity between disconnected pairs of nodes in a social network is calculated. Although there are several metrics to measure similarity between nodes, the most commonly used are: Common Neighbours (CN) [29], Salton Cosine Similarity (SC), Jaccard Coefficient (JC) [30], and Adamic-Adar (AA) [31]. The similarity measures can be i) content specific [32], ii) network specific, and iii) activity specific [33].

2.2.2 Path-Based Approach

Several researchers approached this problem by exploring whether a direct link can be established between a pair of users. A variety of methods have been used to compute the possibility of a link between a pair of nodes based on the ensemble of all the indirect paths available between the nodes: Local Path (LP) [34], Katz' metric [35], PageRank [30], SimRank [36], etc.

2.2.3 Learning-Based Approach

Link prediction can also be viewed as a binary classification problem, for which several feature-based supervised machine learning classifiers such as Decision Trees and Support Vector Machines (SVM) [37] have been used. Consider a graph $G(V, E)$, with vertices V and edges E, where $x, y \in V$ represent the nodes and $l^{x,y}$ represents the label for each pair of nodes (x, y). If a link exists between a pair of nodes, it is labelled as positive, otherwise as negative. Jiang et al. [38] showed that using attributes such as age, interest, characteristic, and common friends for nodes and edges in the network can notably increase the performance of the link prediction system. Li and Chen [39] unveiled the task of link prediction in a bipartite graph, using a graph kernel-based learning method based on features such as user age, level of education, book title, and keywords.

In this paper, link prediction is performed using all the three approaches: i) similarity-based, by calculating the cosine similarity between nodes based on their psychological and sociological attributes, ii) path-based, by using the

node2vec model [40] of link prediction on the three categorical networks. iii) learning-based, by creating a node2vec link prediction system which uses the path-based approach incorporating the psychological and sociological attributes of the nodes/users.

2.3 Community Detection

Discovering hidden structure within a network, i.e., nodes that are tightly connected within themselves and loosely connected with others, is the main agenda of community detection research. Community detection within a social network has been widely studied during the last two decades. Most community detection algorithms focus only on the network information (see [41, 9] for detailed reviews). Yang et al. [42] proposed CESNA, a community detect method from edge structure and node attributes, which takes into consideration both the network structure and the features and attributes of the nodes in the network. Leskovec and Mcauley [43] referred to the categorised network as egocentric, representing how the central user is connected with other members in various networks (high school friends, college friends, and family members). Chakraborty et al. proposed a metric called "permanence" to detect disjoint [44], overlapping [45], and dynamic [46] communities. Zheng et al. [47] introduced ComEmbed, the first community embedding method, which jointly optimises the community embedding and node embedding; M. et al. [48] then proposed another novel community embedding framework. Zuo et al. [49] suggested a temporal network embedding based on Hawkes processes, which is useful for node classification, link prediction, and embedding visualisation, while Wang et al. [50] used a modularised nonnegative matrix factorisation model to incorporate the community structure into the network embedding. Furthermore, Kafeza et al. [27], argue that communities in social media, e.g. Twitter, are more probable to contact information easily if they are not "biased" with respect to user personality

Our work is motivated by Maheshwari et al. [51], who showed that adding the psycholinguistic behaviour of individuals as additional features to community detection helps in accurately detecting structure within a network. Here, we add personality and value features to the nodes and analyse if these node features increase the performance of state-of-the-art community detection algorithms (such as CESNA).

3 Data: the Relations Between Users

There has been a growing interest in the scientific community in doing automatic personality recognition based on language usage and behaviour in social media. A milestone in this area were the 2013 and 2014 workshops and shared tasks on Computational Personality Recognition (WCPR) [52, 53]. Two corpora were released for the 2013 task: a Facebook corpus consisting of about 10,000 Facebook status updates from 250 users, plus their Facebook network properties, labelled with Personality traits, and a corpus of 2,400 essays written by several participants labelled with Personality traits.

For this research, the micro-blog Twitter was used as the source for the data collection. The dataset is comprised of information about users and their relationships: friends, relatives, and colleagues, with their followers and followings.

Table 3: Dataset statistics.

Number of users who took part in the survey	559
Total number of followers/followings	14,710
Total number of tweets obtained	79,129,245
Minimum number of tweets per user	100
Maximum number of tweets per user	10,000
Average number of tweets per user	5,379

A web interface was created to collect the data. In it, the users were first asked to enter their Twitter handles, and were then redirected to pages that respectively displayed their **Followers** (people who follow them) and **Followings** (people who they are following). The followers and followings were populated dynamically with the help of the Twitter API Twitter4j.[2] The users needed to categorise their followers/followings into four pre-defined categories: friends, relatives, colleagues, and other. The first three categories are direct, but selecting the fourth category opened up an additional option, where the user was asked to explicitly describe the "other" type of relationship, e.g., a relationship between a professor and a student.

The data was collected using the crowdsourcing services Amazon Mechanical Turk (AMT)[3] and Rapid Workers[4] [54]. Nearly 600 users participated in the data collection process, and each of them was asked to select at least 10 followers and 10 followings along with their relationship (friends, relatives, colleagues, other). However, not all the responses were qualitative as some of the users did not complete the survey, and a few of them had their Twitter account as private. Furthermore, users who had posted less than 100 tweets were discarded, since they would be not useful for the personality and values classification. Table 3 gives the details of the remaining data.

In addition, a Twitter corpus with personality gold labels developed by Maheshwari et al. [55] was used. It contains 367 unique users with on average 1,608 tweets per user.

4 Psycho-Sociological Classifiers

Inspired by the work published at the Workshop and Shared Task on Computational Personality Recognition [52, 53], four classifiers were built to classify the users' personality, values, age, and gender. The classifiers were trained using several machine learning algorithms: SVM, Multinomial Naïve Bayes (MNB), Simple Logistic Regression (LR), and Random Forests (RF). A comprehensive set of linguistic and non-linguistic features were used in these models. Those features are described below, before going into details of the classifiers that were built.

4.1 Features

Varying lengths of word n-grams (unigrams, bigrams, trigrams) were extracted, since they have proven to be effective for various text classification works [56]. In addition, categorical features such as Part of Speech (POS) were

2. http://twitter4j.org
3. https://www.mturk.com
4. http://www.rapidworkers.com

Table 4: Best LIWC feature selection (accuracy) for each Schwartz value type. "Before Ablation" is based on the full (69) feature set, while "After Ablation" used only the number of features (last row) selected through Pearson correlation analysis.

Feature Ablation	AC	BE	CO	HE	PO	SE	SD	ST	TR	UN
Before Ablation	65.84	56.06	64.02	58.02	58.80	53.06	60.89	56.58	64.28	65.58
After Ablation	65.84	58.54	64.80	58.93	59.58	55.80	61.53	56.84	65.06	66.10
Number of Features	52	37	65	38	54	47	65	53	39	48

Table 5: Speech-Act class distribution in the Facebook and Quora corpus along with performance of the Speech-Act classifier.

Speech-Act	SNO	Wh	YN	SO	AD	YA	T	AP	RA	A	O	Avg.
Class Distribution (%)	33.37	11.45	15.45	5.16	6.88	15.08	0.41	3.26	0.71	0.07	14.59	
F_1-score	0.45	0.88	0.88	0.72	0.45	0.60	0.72	0.60	0.12	0.77	0.12	0.69

used, together with a few other word-level features like capital letters and repeated words. Apart from these, features extracted from several psycholinguistic and sensorial lexica were utilised, as well as speech-act features and non-linguistic features.

4.1.1 Linguistic Inquiry and Word Count (LIWC)

LIWC [57] is a well-developed handcrafted lexicon consisting of 6,000 distinct words and 69 broad categories (emotions, psychology, affection, social processes, etc.). In order to select the words of interest from the lexicon, three corpora were collected—from user essays, Facebook, and Twitter—and analysed in terms of Big 5 personality traits and Schwartz' ten value classes. To understand the contribution of each LIWC feature, feature ablation was performed and the Pearson correlation calculated for each feature over all personality and value traits.

Table 4 delineates the accuracy for the ten Schwartz value traits when using either all 69 LIWC features or when only using a smaller set of features selected by Pearson correlation (given in the "after ablation" row). The final values classifier was trained only with features that contributed to a particular value trait, giving a performance boost and a significant reduction in time complexity (both training and testing time for the model). Table 4 shows that the Achievement (AC) class has the same accuracy (65.84%) before and after feature ablation, whereas the lowest obtained accuracy for the Security (SE) class (53.06%) is increased to 55.80% when considering only 47 features (after ablation).

4.1.2 Psycholinguistic Lexica

In addition to the basic LIWC features, information was extracted from two other psycholinguistic lexica: Harvard General Inquirer and the MRC psycholinguistic database.

The Harvard General Inquirer lexicon contains 182 categories including two broad categories, positive and negative. Apart from these, there are also other psycho-linguistic categories such as words that indicate pleasure, pain, virtue, etc., and words that indicate overstatement and understatement. These categories have been used in content analysis research applications in social science to assess particular situations, emotion-laden words, and cognitive orientation.

The MRC lexicon contains 1,508,837 words with up to 26 linguistic and psycholinguistic attributes. Here, 14 features from the MRC lexicon were used to create the model, including the number of phonemes and syllables, Kucera-Francis number of categories and number of samples, Kucera-Francis frequency, Thorndike-Lorge frequency, Brown verbal frequency, ratings of familiarity, concreteness, imageability, and age of acquisition. All MRC psycholinguistic features were obtained using an API.[5]

4.1.3 Sensicon

Sensicon [58] is a sensorial lexicon, which is comprised of words with sense association scores pertaining to the five basic senses: sight, hearing, smell, touch, and taste. For example, when the human mind comes across the word "apple", it will automatically visualise the appearance of an apple, stimulating the eye-sight, feel the smell of the apple in the nose and the taste on the tongue. Sensicon provides probabilistic mappings indicating the level to which each of the five senses is used to understand a particular concept.

4.1.4 Speech-Act Features

Speech-acts define the various ways we make conversation by uttering different sentence semantics such as opinion, question, acknowledgement, etc. For this research, speech-acts were classified into eleven types: statement non-opinion (SNO), wh-question (Wh), yes-no question (YN), statement opinion (SO), action directive (AD), yes-answers (YA), thanking (T), appreciation (AP), response acknowledgement (RA), apology (A), and others (O). A corpus comprised of 7,000 utterances was gathered from Facebook and Quora pages, and annotated manually. A speech-act classifier was built using an SVM trained on the following features: bag-of-words (top 20% bigrams), presence of "wh" words, presence of question marks, occurrence of "thanks/thanking" words, POS tag distributions, and sentiment lexica such as the NRC Sentiment and Emotion lexicon,[6] SentiWordNet [59], and WordNet Affect [60]. The classifier obtained an F_1-score of 0.69 after 10-fold cross-validation. The corpus distribution and classifier performance are shown in Table 5.

4.1.5 Non-Linguistic Features

Apart from the linguistic features, social networks topological features were used to build the classifiers. For the Facebook corpus, network properties such as network size, betweenness centrality, density, and transitivity were used. For the Twitter values corpus, properties such as the total

5. http://ota.oucs.ox.ac.uk/headers/1054.xml

6. https://www.nrc-cnrc.gc.ca/eng/solutions/advisory/emotion_lexicons.html

Table 6: Performance of the GloVE + CNN Values classifier.

Class	Precision	Recall	F_1-score
Achievement	0.85	0.83	0.86
Benevolence	0.78	0.74	0.75
Hedonism	0.86	0.78	0.81
Conformity	0.89	0.75	0.81
Security	0.86	0.70	0.77
Power	0.88	0.80	0.83
Stimulation	0.87	0.76	0.81
Traditional	0.85	0.83	0.86
Self-Direction	0.81	0.75	0.77
Universalism	0.84	0.70	0.76

number of tweets/messages of one user, number of likes, average time difference between two tweets, number of favourites and re-tweets of all the messages by one user, and their in-degree and out-degree centrality scores on the networks of friends and followers were used as features.

4.2 The Personality and Values Classifiers

Several psycholinguistic features were tested for categorising user personality and values. The classifiers as described by Maheshwari et al. [55] were recreated using features including network properties (network size, betweenness centrality, density, transitivity), linguistic features (such as LIWC, MRC, Harvard General Inquirer, and Sensicon), and speech-act classes. Three machine learning algorithms (SVM, LR, and RF) were used to perform the classification. All results obtained are based on 10-fold cross-validation. The SVM-based model outperformed the state-of-the-art system [61], achieving an average F_1-score of 0.80 for personality classification and 0.81 for values classification.

For comparison, a classifier using GloVE word embeddings [62] and a convolutional neural network (CNN) was also created for the personality and values classification.[7] This classifier performed slightly worse than the SVM, producing an F_1-score of 0.78 for personality and 0.80 for values classification (see Table 6 for detailed results).

Furthermore, the SVM personality classifier was compared to the deep learner proposed by Majumder et al. [63]. Their model achieved an F_1-score of 0.73 on our data, which is 9.4% lower than that of the SVM model. In short, the SVM model beats three compatible systems. Readers are referred to the supplementary material for detailed feature analysis.

4.3 The Age Classifier

The dataset released for the PAN 2016 Author Profiling task [64] was used to develop an age/gender classification model. In this classifier, the same set of linguistic features was used as in the personality and values classifier along with some additional features: word *n*-grams, POS tags, sentiment amplifiers (exclamation marks, quotes, interjections, emoticons, etc.), and misspelt words (words used in text messages and while chatting, typographical errors, etc.). A Random Forest classifier was trained to predict the age of the users according to the following categories: between 10 and 20, between 20 and 30, and older than 30 years. The model achieved an F_1-score of 0.56, which is very close to the state-of-the-art system by Rangel et al. [64], which had an F_1-score of 0.59.

7. See the supplementary material for detailed network architectures.

4.4 The Gender Classifier

For the gender classifier, an additional feature based directly on Indian usernames was utilised (checking for possible occurrences of "-abu" and "-um/min", that represent common suffixes for men and women, respectively) along with all the features used for the age classifier: word *n*-grams, POS tags, sentiment amplifiers, and misspelt words. A Support Vector Machine classifier was trained to predict user's gender based on their tweets. The model obtained an F_1-score of 0.76, which is comparative to the highest score achieved by the participants in the author profiling task [64].

5 Psycho-Sociological Factors

To understand how psycho-sociological factors impact on the categorised networks, three networks of friends, relatives, and colleagues were created. The networks were created in such a way that source nodes represent the users and target nodes represent their friends, in the friends network, with similar relationships represented in the relatives and colleagues networks. An automatic model (discussed in Section 4) was used to categorise people into corresponding personality and value types by analysing their language usage in social media and their social network behaviour. Each member of the network was assigned personality and values based on the scores obtained for each trait as a result of the classification task, with the personality and values having the highest score being assigned as the most dominant characteristics of the particular user.

For each of the three networks, the number of Big-5 personality-personality pairs (Openness-Openness, Openness-Conscientiousness, and so on) was calculated, resulting in a 5×5 matrix. Then the number of Schwartz value pairs (Achievement-Achievement, Achievement-Benevolence, etc.) that existed in each of the three networks was calculated; resulting in a 10×10 matrix.

These results represented a count of the personality/values pairs and differed greatly. Hence the results were scaled using max-min normalisation (i.e., $(x - x_{min})/(x_{max} - x_{min})$) so that they fall in the $[0, 1]$ range. The scaled personality and values scores for each network were then analysed.

5.1 Friends Are Homophilic

The scaled personality scores for the friends network are reported in Table 7(a). Users with the same personality traits show higher connectivity, i.e., prefer their friends to be of similar personality, thus the network exhibits the homophily phenomenon. Furthermore, people who are agreeable (A) tend to be connected strongly to almost all the other personality traits as they are friendly and amiable by nature. Some other characteristics of these people include geniality, generosity, kindness, and altruism. The extroverts (E), who spirited, talkative, sociable, and energetic, are also highly connected with people of other personality traits. The neurotic people (N), who by nature are docile, passive, and unstable in their personalities, are in the friends network strongly connected with conscient (C) users, who are dependable, reliable, and mature. In this network, neurotic people are not complementary to each other, as they are

Table 7: Friends are homophilic. The diagonals (left to right) show highest connectivity between similar personality and value pairs.

		Friend				
		O	C	E	A	N
User	O	**1.00**	0.22	0.04	0.19	0.00
	C	0.00	0.89	0.68	0.32	**1.00**
	E	0.00	0.26	**1.00**	0.92	0.74
	A	0.00	0.74	0.54	**1.00**	0.71
	N	0.11	**1.00**	0.29	0.00	0.41

(a) Friends' personalities. The highest scores are for matching personalities (e.g., openness-openness) except for the neurotic-neurotic and conscientiousness-conscientiousness pairs.

		Friend									
		AC	BE	CO	HE	PO	SE	SD	ST	TR	UN
User	AC	**1.00**	0.07	0.31	0.36	0.42	0.00	0.48	0.40	0.28	0.55
	BE	0.27	1.00	0.00	0.37	0.13	0.22	0.41	0.41	0.10	0.11
	CO	1.00	0.00	0.86	0.37	0.40	0.31	0.49	0.56	0.28	0.08
	HE	0.16	0.18	0.33	**1.00**	0.42	0.00	0.17	0.29	0.13	0.30
	PO	0.24	0.09	0.00	0.15	1.00	0.77	0.15	0.17	0.99	0.21
	SE	0.40	0.74	0.70	0.19	0.39	0.51	0.31	0.00	0.58	**1.00**
	SD	0.02	0.82	0.26	0.13	0.01	0.61	**1.00**	0.00	0.11	0.59
	ST	0.00	0.26	0.18	0.29	0.25	0.15	0.20	**1.00**	0.22	0.10
	TR	0.16	**1.00**	0.73	0.52	0.21	0.89	0.23	0.24	0.33	0.00
	UN	0.38	0.06	**1.00**	0.00	0.04	0.07	0.11	0.23	0.07	0.92

(b) Friends' values. Eight of the ten user pairs with the same value traits (achievement-achievement, benevolence-benevolence, etc.) have high connectivity in this network.

mostly tense, anxious, and moody, which does not fit into the characteristics of friends.

The relationships between people of different value traits are tabulated in Table 7(b). Similar to the personality classification, people of the same value trait are strongly connected among themselves, thus exhibiting the phenomenon of homophily. The security-oriented people (SE) maintain good connectivity with people of most other value traits, which is an astounding inference. These people place more emphasis on safety, harmony, the stability of society, social relationships, and of self. The power-oriented people (PO), on the other hand, are by nature dominant over other people and resources. Hence they are not connected to people of other values (except with SE and TR), substantiating that friends like to have similar intentions and do not want to be dominated by others.

5.2 Relatives Are Cultural Homophilic

The relationships between users of various personality and values traits in the relatives network are delineated in Tables 8(a) and 8(b), respectively. Unlike the relationship in the friends network, the homophily phenomenon does not appear in this network, that is, most of the relationships are not oriented towards people of the same personality/value trait. This is since the users may choose friends based on their personality/value traits, but cannot do so in the case of relatives. In this network, surprisingly the extroverts (E) are highly connected with the neurotic (N) people, which is not a common phenomenon in the friends network. This shows that in this network the extroverts can really handle the neurotic people because of the conservative behaviour of people in the network. People who are open (O) are connected highly among themselves as these people are

Table 8: Relatives are conservative.

		Relative				
		O	C	E	A	N
User	O	**1.00**	0.37	0.31	0.63	0.00
	C	**1.00**	0.52	0.00	0.07	0.21
	E	0.00	0.10	0.36	0.34	**1.00**
	A	0.03	0.00	0.63	**1.00**	0.42
	N	0.00	**1.00**	0.77	0.05	0.17

(a) Relatives' personalities. All other personality types are highly connected to the conscient people (second column), as they are dependable, decisive, and conventional.

		Relative									
		AC	BE	CO	HE	PO	SE	SD	ST	TR	UN
User	AC	0.06	0.01	0.24	**1.00**	0.01	0.34	0.18	0.00	0.07	0.11
	BE	0.84	**1.00**	0.25	0.35	0.22	0.73	0.66	0.63	0.29	0.00
	CO	0.56	0.46	0.17	**1.00**	0.43	0.00	0.62	0.39	0.30	0.37
	HE	0.25	0.19	**1.00**	0.14	0.07	0.49	0.00	0.09	0.21	0.70
	PO	0.07	0.14	0.00	0.09	**1.00**	0.04	0.09	0.01	0.74	0.32
	SE	0.00	0.10	0.30	0.03	0.43	0.39	0.16	0.13	0.14	**1.00**
	SD	0.83	0.61	0.11	0.01	0.15	0.51	**1.00**	0.65	0.00	0.10
	ST	0.21	0.23	0.55	0.24	0.24	0.00	0.23	**1.00**	0.11	0.30
	TR	0.89	0.83	0.27	**1.00**	0.68	0.00	0.73	0.61	0.47	0.58
	UN	0.08	0.12	0.22	0.11	0.27	0.26	0.03	0.00	**1.00**	0.05

(b) Relatives' values. People who are traditional (TR) maintain good connectivity to most other value traits, demonstrating that people in this network are conservative and cultural.

intellectual, thoughtful, simple, and lovers of art (poetry, literature, music, etc.) and hence end up exchanging creative ideas. On the other hand, similar to the friends network, there is a strong connectivity between people who are neurotic (N) and conscient (C). People who are conscient are dependable as they are reliable, mature, and discreet; thus the neurotic people, who tend to be unstable and dependent, have strong connections to them.

Table 8(b) summarises the relationship between people of various Schwartz type values in the relatives network. There is a strong relationship between people who are traditional (TR) and other values traits, as these people respect and accept the customs and ideas that the culture and religion of another person provide, thus maintaining mature relationships with others. Conformity-oriented (CO) persons are well connected with hedonistic and self-directed people (likewise for benevolent persons). These persons are restrained in their actions and less likely to harm others and violate social expectations. From these observations it is clear that people in the relatives network are more conservative, i.e., they are more conventional and orthodox.

5.3 Colleagues Are Not Necessarily Homophilic

Empirical results also provide interesting observations in the colleagues network represented in Table 9(a) for personality and Table 9(b) for values. The relationships in the colleagues network are not necessarily considered to be homophilic in nature. It is inferred that there is a strong relationship between users of various personality traits and those who are open (O) (known for their wisdom, intellectuality, innovativeness, and clever thinking).

The users who are agreeable (A) tend to be cooperative, agreeable, and altruistic. The relationship between the openness (O) and agreeableness (A) characteristics shows the need for growth/enhancement of skills among the users

Table 9: Colleagues focus on growth/enhancement.

		Colleague				
		O	C	E	A	N
User	O	**1.00**	0.87	0.70	0.79	0.00
	C	0.00	0.70	0.16	0.11	**1.00**
	E	**1.00**	0.20	0.07	0.00	0.89
	A	0.15	0.00	**1.00**	0.93	0.26
	N	0.17	**1.00**	0.27	0.43	0.00

(a) Colleagues' personalities. People possessing the openness personality trait tend to be intelligent, wise, pensive are highly connected with all the other personality traits.

		Colleague									
		AC	BE	CO	HE	PO	SE	SD	ST	TR	UN
User	AC	0.69	0.00	**1.00**	0.45	0.95	0.71	0.67	0.36	0.20	0.95
	BE	0.00	0.58	**1.00**	0.06	0.20	0.19	0.70	0.38	0.57	0.09
	CO	0.52	0.00	0.24	0.55	0.69	0.70	0.51	**1.00**	0.21	0.76
	HE	0.54	0.11	0.71	**1.00**	0.12	0.13	0.03	0.11	0.00	0.24
	PO	0.18	0.25	0.00	0.36	**1.00**	0.13	0.29	0.01	0.16	0.35
	SE	0.32	0.06	0.17	0.65	0.18	0.36	0.31	0.10	0.00	**1.00**
	SD	0.94	0.80	0.29	0.03	0.15	**1.00**	0.80	0.94	0.00	0.22
	ST	0.11	0.51	0.08	0.06	0.20	0.11	0.00	0.55	**1.00**	0.04
	TR	0.55	0.00	0.29	0.41	0.70	0.56	0.19	0.35	**1.00**	0.34
	UN	0.00	0.43	0.19	0.01	0.07	0.21	0.15	0.19	**1.00**	0.03

(b) Colleagues' values. Achievement-oriented people appear to have strong relationships with persons possessing most other value traits in the colleagues network.

in the network. This relationship can be justified by the fact that when some new initiative or idea is posted by a person who is open, the agreeable users tend to be more supportive than controverting. On the other hand, people who are neurotic show heterogeneity in their relationships as they are mostly tense, anxious, and moody, which can be interpreted as instability in their characteristics.

Table 9(b) represents the relationships between people of various values in the colleagues network. The achievement-oriented people, who are goal-focused, are highly connected people of power, self-enhancement, and universality, and with other achievement oriented people in their neighbourhood (likewise for CO people), as their primary objective is to achieve their targets. In this network, the self-directed people (SD) are highly connected to the stimulant (ST) people, showing their intrinsic interest in novelty and mastery, similar to the behaviour of people who are open. There is a strong relationship between people who are benevolent (BE) and conformity-oriented (CO) people.

5.4 Take-Away Points

The overall observation from this analysis on personality and values in three networks can be summarised as follows:

- People who are agreeable (A) are connected strongly with users of other personality traits in all three networks.
- People who are neurotic (N) show inconsistent behaviour in all the networks because of the unstable character pertaining to their personality. For friendship and in professional settings they rely mostly on conscient people, whereas among relatives they prefer to mingle with extroverts.
- In none of the networks do people connect much with those who are power-oriented (PO), because these people try to impose their dominance in relationships.
- The friends network exhibits the homophily phenomenon, where the relationships among people of the same personality traits are strong, whereas the other networks do not produce significant results in terms of homophily.
- The relatives network is conservative, i.e., traditional, orthodox, and conventional. Traditional people show a strong connection with others in the network, thus making this network culturally homophilic in nature.
- The colleagues network showed diversity in the relationships between personalities and values, but a significant observation is that people in this network are focused on achieving goals and emphasise growth/enhancement.

6 Link Prediction

An abstract definition of link prediction is: given a snapshot of a network, is it possible to predict the links that are likely to be formed in the network? This section first defines the link prediction problem, then describes the experimental setup and the methodology used to develop a link prediction system.

Definition 1 (The Link Prediction Problem). Given a network $G(V, E)$ where E represents the edge set and V the vertex set in G at time t, is it possible to predict the links that could be established in G at a future time t'?

6.1 Experimental Setup

A comparative study on the link prediction system was done on the three networks (friends, relatives, and colleagues). Each network varies diversely with the number of nodes present, with the friends network being comprised of approximately 1,200 nodes, the relatives network of 370 nodes, and the colleagues network of 440 nodes. Once a user's personality, values, age, and gender have been classified, a model can be built to predict links based on nodes with similar attributes. 60% of the dataset was used for training, 10% for validation, and 30% for testing. In a social network, the attribute information of nodes plays a decisive role in link prediction [38]. A comprehensive set of attributes was incorporated as node features (for models 1 and 3 described below): *personality* (openness, conscientiousness, extroversion, agreeableness, neuroticism), *values* (achievement, benevolence, conformity, hedonism, power, security, self-direction, stimulation, traditional, universalism), *gender* (male, female), and *age* (< 20, $20 - 30$, > 30).

6.2 The Link Prediction Systems

Three models were used to perform link prediction. The three models follow node2vec (model 1), cosine similarity between the node attributes (model 2), and node2vec with additional node attributes (model 3).

6.2.1 Model 1

The first model draws inspiration from the work by Grover and Leskovec [40] on node2vec. The model utilises random walks with SkipGrams and can be viewed as a speculation of DeepWalk [65]. The contrast between the two techniques

Table 10: Performance Evaluation of the Link Prediction model in terms of AUC (area under curve), precision, recall, and F_1-score. The improvements of Model 3 w.r.t. Model 1 and Model 2 are statistically significant with $p < 0.05$ and $p < 0.01$, respectively.

Metric	Model 1	Model 2	Model 3
AUC	0.97	0.64	0.98
Precision	0.98	0.64	0.98
Recall	0.96	0.66	0.98
F_1-score	0.97	0.65	0.94

(a) Friends Network

Metric	Model 1	Model 2	Model 3
AUC	0.89	0.43	0.92
Precision	0.84	0.45	0.89
Recall	0.80	0.41	0.94
F_1-score	0.82	0.43	0.91

(b) Relatives Network

Metric	Model 1	Model 2	Model 3
AUC	0.92	0.41	0.94
Precision	0.95	0.49	0.97
Recall	0.91	0.46	0.95
F_1-score	0.93	0.47	0.96

(c) Colleagues Network

is that node2vec's walks are arbitrary, however, one-sided by two pre-relegated parameters p and q (considered to be the hyper-parameters). When $p = q = 1$, this is equivalent to DeepWalk. In the production of the walks, the parameters are utilised to elevate the chances of the walk coming back to the parent node or going further. The technique makes use of a semi-directed approach, requiring a few models to be created and some nodes labelled, so that the best values for the parameters p and q can be chosen. (Note that this is the state-of-the-art model, so can be considered as a baseline.)

6.2.2 Model 2

The second model calculates the similarity between features (including the personality and values) of each user. Consider that there are m attributes for n nodes in a network. A vector F_i can be created for each node v_i, such that $F_i = (f_{i1}, f_{i2}, f_{i3}, \cdots, f_{im})$ represents the features of the node (user), with i ranging from 1 to n. The similarity between the attributes of each node pair v_i and v_j is obtained by calculating the cosine similarity between the vectors F_i and F_j. Links are predicted if the similarity score is greater than a threshold.

6.2.3 Model 3

Utilising a combination of the former two models, i.e., using the node2vec model (Model 1) and incorporating the personality and values features (Model 2). In this system, additional node attributes (age, gender) are also used.

6.3 Evaluation of the Link Prediction Models

Table 10 reports the performance of link prediction based on Area Under Curve (AUC), precision, recall, and F_1-score, respectively, for the three categorical networks. Empirical results suggest that all the link predictors perform better on the friends network than the other two networks, justifying the phenomenon of homophily in the friends network.

7 Psycho-Sociological Community Detection

It could also be inferred that adding node attributes increases the performance of node2vec to a significant extent, corroborating the results by Jiang et al. [38]. Specifically, experiments were carried out to investigate whether community detection accuracy improves if appropriate features are added to the nodes [66]. The state-of-the-art algorithm CESNA [42] was used for community detection, as it considers both the network structure and the node features (personality and values).

The community detection algorithm was evaluated with various performance measures: Normalised Mutual Information (NMI),[8] Adjusted Random Index (ARI),[9] and Community F-score (C_F) [41] for different feature sets: Mutual Information (I) is the shared information between two distributions [67]:

$$I(X;Y) = \sum_{y \epsilon Y} \sum_{x \epsilon X} p(x,y) \log \frac{p(x,y)}{p_1(x) p_2(y)}. \quad (1)$$

NMI is the normalisation of mutual information. The results are scaled (given in Equation 2) such that they lie between 0 (no mutual information) and 1 (perfect mutual information) [67, 68].

$$\mathrm{NMI}(X;Y) = \frac{I(X;Y)}{\sqrt{H(X)H(Y)}}. \quad (2)$$

The ARI compares two partitions (a scalar). In the case of random partition this index has a 0 expected value, and in the case of a perfect agreement between the partitions this index has the value of 1 [69].

As suggested by Chakraborty et al. [41], C_F is computed as follows: Given the ground-truth community structure $\bar{C}$ and a detected community structure C, decide which $c_i \in C$ corresponds to $\bar{c_j} \in \bar{C}$. Then C_F is defined to be the average of the F_1-scores of the best matching c_i to $\bar{c_j}$ and vice-versa (since the relationship is asymmetric):

$$C_F = \frac{1}{2} \left(\frac{1}{|C|} \sum_{c_i \in C} F_1(c_i, \bar{c_j}) + \frac{1}{|\bar{C}|} \sum_{\bar{c_i} \in \bar{C}} F_1(c_{j'}, \bar{c_i}) \right), \quad (3)$$

where the best matching j and j' are defined as $j = \mathrm{argmax}_j F_1(c_i, \bar{c_j})$ and $j' = \mathrm{argmax}_{j'} F_1(c_j, \bar{c_j})$, with $F_1(.,.)$ being the harmonic mean of precision and recall.

Table 11 represents the NMI, ARI and C_F scores for CESNA over different feature sets. The comparison is made based on community detection using CESNA in four stages where initially community detection is performed only with network information. Second, community detection is performed with network information and values features added to the node attributes. Third, community detection is performed with network information and personality

8. http://scikit-learn.org/stable/modules/generated/sklearn.metrics.normalized_mutual_info_score.html

9. https://www.rdocumentation.org/packages/mclust/versions/5.4/topics/adjustedRandIndex

Table 11: Performance of CESNA with different feature sets. Statistical significance testing is performed to show that the improvement of using the network information together with personality and values is significant ($p < 0.05$) w.r.t. using only the network information.

Features	NMI	ARI	C_F
Network Information	0.82	0.51	0.49
Network + Values	0.82	0.52	0.50
Network + Personality	0.83	0.52	0.51
Network + Personality & Values	**0.85**	**0.53**	**0.52**
p-value	0.035	0.037	0.029

features added to the node attributes. Finally, community detection is performed with network information and both personality and values features added to the node attributes.

CESNA performs better when personality and values features are incorporated as node attributes. This result signifies that the appropriate additional information related to the nodes notably improves the performance of the community detection algorithm [42]. To show that the accuracy improvement with personality and values taken together with the network features is statistically significant with respect to only considering the network features, the p-value for community detection was measured on the categorical networks (last row of Table 11). The p-value is lower than 0.05, signifying strong evidence against the null hypothesis.

8 Conclusion and Future Work

The contributions of this paper are four-fold: First, the relationships present between people of various communities (friends, relatives, and colleagues) were analysed by considering their psychological (personality) traits. Second, the behaviour of people in these communities was further analysed by considering their sociological (value) traits. Third, it was shown that the detected personality and values of individuals can be used as additional node attributes to detect better community structure. Finally, a link prediction system was developed to show that because the friends network is homophilic in nature, the performance of link prediction is significantly higher when compared to the performance of link prediction for the other two networks.

Future work would aim to look at closeness and reciprocity in categorical networks to obtain some insight on the behaviour of users with respect to various personality/values in each of the networks. We believe that these kinds of models may become extremely useful in the future for various purposes such as social media advertising, recommendation systems, computational psychology, and psycho-sociological analysis of users in social media.

The code and the datasets used in this study are publicly available at https://github.com/Hilt-Reseach/CIM.

Acknowledgments

The authors thank the anonymous reviewers and the editor for providing valuable feedback and suggestions. Tanmoy Chakraborty acknowledges the support of the Ramanujan Faculty Fellowship and the Infosys Centre of AI, IIIT Delhi.

References

[1] L. R. Goldberg, "An alternative "description of personality": The big-five factor structure," *Journal of Personality and Social Psychology*, vol. 59, no. 6, p. 1216, Dec. 1990.

[2] S. H. Schwartz, "Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries," *Advances in Experimental Social Psychology*, vol. 25, pp. 1–65, Dec. 1992.

[3] M. Gavrilescu and N. Vizireanu, "Predicting the Big Five personality traits from handwriting," *EURASIP Journal on Image and Video Processing*, vol. 2018, no. 1, p. 57, Jul. 2018.

[4] X.-F. Zhong, S.-Z. Guo, L. Gao, H. Shan, and D. Xue, "A general personality prediction framework based on facebook profiles," in *Proceedings of the 10th International Conference on Machine Learning and Computing*. Macau, China: ACM, 2018, pp. 269–275.

[5] B. Rammstedt, C. M. Lechner, and D. Danner, "Relationships between personality and cognitive ability: A facet-level analysis," *Journal of Intelligence*, vol. 6, no. 2, p. 28, 2018.

[6] J. Anglim, E. R. Knowles, P. D. Dunlop, and A. Marty, "HEXACO personality and Schwartz's personal values: A facet-level analysis," *Journal of Research in Personality*, vol. 68, pp. 23–31, June 2017.

[7] T. Zhang, R.-Z. Qin, Q.-L. Dong, W. Gao, H.-R. Xu, and Z.-Y. Hu, "Physiognomy: Personality traits prediction by learning," *International Journal of Automation and Computing*, vol. 14, no. 4, pp. 386–395, Jun. 2017.

[8] J. Xu, W. Tian, Y. Fan, Y. Lin, and C. Zhang, "Personality trait prediction based on 2.5 D face feature model," in *International Conference on Cloud Computing and Security*. Springer, 2018, pp. 611–623.

[9] S. Fortunato, "Community detection in graphs," *Physics Reports*, vol. 486, no. 3, pp. 75–174, Nov. 2010.

[10] I. X. Leung, P. Hui, P. Liò, and J. Crowcroft, "Towards real-time community detection in large networks," *Physical Review E*, vol. 79, no. 6, p. 066107, Jun. 2009.

[11] M. E. Newman, "Fast algorithm for detecting community structure in networks," *Physical Review E*, vol. 69, no. 6, p. 066133, Jun. 2004.

[12] F. Radicchi, C. Castellano, F. Cecconi, V. Loreto, and D. Parisi, "Defining and identifying communities in networks," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 9, pp. 2658–2663, Mar. 2004.

[13] J. Yang and J. Leskovec, "Defining and evaluating network communities based on ground-truth," *Knowledge and Information Systems*, vol. 42, no. 1, pp. 181–213, Jan. 2015.

[14] A. Anagnostopoulos, R. Kumar, and M. Mahdian, "Influence and correlation in social networks," in *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2008, pp. 7–15.

[15] S. Aral and D. Walker, "Identifying influential and susceptible members of social networks," *Science*, vol. 337, no. 6092, pp. 337–341, July 2012.

[16] H. Bisgin, N. Agarwal, and X. Xu, "A study of homophily on social media," *World Wide Web*, vol. 15, no. 2, pp. 213–232, March 2012.

[17] J. J. Brown and P. H. Reingen, "Social ties and word-of-mouth referral behavior," *Journal of Consumer Research*, vol. 14, no. 3, pp. 350–362, Dec. 1987.

[18] P. N. Krivitsky, M. S. Handcock, A. E. Raftery, and P. D. Hoff, "Representing degree distributions, clustering, and homophily in social networks with latent cluster random effects models," *Social Networks*, vol. 31, no. 3, pp. 204–213, Jul. 2009.

[19] H. Kwak, C. Lee, H. Park, and S. Moon, "What is Twitter, a social network or a news media?" in *Proceedings of the 19th International Conference on World Wide Web*. ACM, Apr. 2010, pp. 591–600.

[20] M. McPherson, L. Smith-Lovin, and J. M. Cook, "Birds of a feather: Homophily in social networks," *Annual Review of Sociology*, vol. 27, no. 1, pp. 415–444, Aug. 2001.

[21] D. Mulders, C. De Bodt, J. Bjelland, A. S. Pentland, M. Verleysen, and Y.-A. de Montjoye, "Improving individual predictions using social networks assortativity," in *Proceeding of the 12th International Workshop on Self-Organizing Maps and Learning Vector Quantization, Clustering and Data Visualization*. Nancy, France: IEEE, Jun. 2017, pp. 169–178.

[22] M. Q. Pasta, F. Zaidi, and C. Rozenblat, "Generating online social networks based on socio-demographic attributes," *Journal of Complex Networks*, vol. 2, no. 4, pp. 475–494, Dec. 2014.

[23] R. Reagans, "Demographic diversity as network connections: Homophily and the diversity performance debate," in *The Oxford Handbook of Diversity and Work*, Q. M. Roberson, Ed. Oxford University Press, Oct. 2013, pp. 192–206.

[24] S. R. Wilson, R. Mihalcea, R. L. Boyd, and J. W. Pennebaker, "Disentangling topic models: A cross-cultural analysis of personal values through words," in *Proceedings of the First Workshop on NLP and Computational Social Science*. Austin, Texas: ACL, 2016, pp. 143–152.

[25] A. Clauset, M. E. Newman, and C. Moore, "Finding community structure in very large networks," *Physical Review E*, vol. 70, no. 6, p. 066111, Dec. 2004.

[26] I. S. Dhillon, Y. Guan, and B. Kulis, "Weighted graph cuts without eigenvectors a multilevel approach," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 11, pp. 1944–1957, Nov. 2007.

[27] E. Kafeza, A. Kanavos, C. Makris, and D. Chiu, "Identifying personality-based communities in social networks," in *Advances in Conceptual Modeling*, J. Parsons and D. Chiu, Eds. Basel, Switzerland: Springer, 2014, pp. 7–13.

[28] M. E. Newman, "Assortative mixing in networks," *Physical Review Letters*, vol. 89, no. 20, p. 208701, Oct. 2002.

[29] ——, "Clustering and preferential attachment in growing networks," *Physical Review E*, vol. 64, no. 2, p. 025102, Jan. 2001.

[30] D. Liben-Nowell and J. Kleinberg, "The link-prediction problem for social networks," *Journal of the Association for Information Science and Technology*, vol. 58, no. 7, pp. 1019–1031, Nov. 2007.

[31] L. A. Adamic and E. Adar, "Friends and neighbors on the web," *Social Networks*, vol. 25, no. 3, pp. 211–230, July 2003.

[32] P. Bhattacharyya, A. Garg, and S. F. Wu, "Analysis of user keyword similarity in online social networks," *Social Network Analysis and Mining*, vol. 1, no. 3, pp. 143–158, July 2011.

[33] A. Anderson, D. Huttenlocher, J. Kleinberg, and J. Leskovec, "Effects of user similarity in social media," in *Proceedings of the 5th ACM International Conference on Web Search and Data Mining*. Seattle, WA, USA: ACM, Feb. 2012, pp. 703–712.

[34] L. Lü, C.-H. Jin, and T. Zhou, "Similarity index based on local paths for link prediction of complex networks," *Physical Review E*, vol. 80, no. 4, p. 046122, Oct. 2009.

[35] L. Katz, "A new status index derived from sociometric analysis," *Psychometrika*, vol. 18, no. 1, pp. 39–43, Mar. 1953.

[36] G. Jeh and J. Widom, "SimRank: A measure of structural-context similarity," in *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, Jul. 2002, pp. 538–543.

[37] M. A. Hearst, "Support vector machines," *IEEE Intelligent Systems*, vol. 13, no. 4, pp. 18–28, Jul. 1998.

[38] M. Jiang, Y. Chen, and L. Chen, "Link prediction in networks with nodes attributes by similarity propagation," *arXiv preprint arXiv:1502.04380*, 2015.

[39] X. Li and H. Chen, "Recommendation as link prediction in bipartite graphs: A graph kernel-based machine learning approach," *Decision Support Systems*, vol. 54, no. 2, pp. 880–890, Jan. 2013.

[40] A. Grover and J. Leskovec, "node2vec: Scalable feature learning for networks," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, Aug. 2016, pp. 855–864.

[41] T. Chakraborty, A. Dalmia, A. Mukherjee, and N. Ganguly, "Metrics for community analysis: A survey," *ACM Computing Surveys*, vol. 50, no. 4, pp. 54:1–54:37, November 2017.

[42] J. Yang, J. McAuley, and J. Leskovec, "Community detection in networks with node attributes," in *Proceeding of the 13th IEEE International Conference on Data Mining*. Dallas, Texas: IEEE, 2013, pp. 1151–1156.

[43] J. Leskovec and J. J. Mcauley, "Learning to discover social circles in ego networks," in *Advances in Neural Information Processing Systems 25*. Curran Associates, 2012, pp. 539–547.

[44] T. Chakraborty, S. Srinivasan, N. Ganguly, A. Mukherjee, and S. Bhowmick, "On the permanence of vertices in network communities," in *Proceeding of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, USA, Aug. 2014, pp. 1396–1405.

[45] T. Chakraborty, S. Kumar, N. Ganguly, A. Mukherjee, and S. Bhowmick, "GenPerm: A unified method for detecting non-overlapping and overlapping communities," *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 8, pp. 2101–2114, Aug. 2016.

[46] P. Agarwal, R. Verma, A. Agarwal, and T. Chakraborty, "DyPerm: Maximizing permanence for dynamic community detection," in *Proceedings of the 22nd Pacific-Asia Conference on Knowledge Discovery and Data Mining*, vol. 1, Melbourne, Australia, Jun. 2018, pp. 437–449.

[47] V. W. Zheng, S. Cavallari, H. Cai, K. C.-C. Chang, and E. Cambria, "From node embedding to community embedding," *arXiv preprint*, no. 1610.09950, 2016.

[48] C. M., V. W. Zheng, H. Cai, K. C.-C. Chang, and E. Cambria, "Learning community embedding with community detection and node embedding on graphs," in *Proceedings of the 2017 ACM Conference on Information and Knowledge Management*. Singapore: ACM, 2017, pp. 377–386.

[49] Y. Zuo, G. Liu, H. Lin, J. Guo, X. Hu, and J. Wu, "Embedding temporal network via neighborhood formation," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. London, United Kingdom: ACM, 2018, pp. 2857–2866.

[50] X. Wang, P. Cui, J. Wang, J. Pei, W. Zhu, and S. Yang, "Community Preserving network embedding." in *Proceedings of the 31st AAAI Conference on Artificial Intelligence and the 29th Innovative Applications of Artificial Intelligence Conference*. San Francisco, California: AAAI Press, Feb. 2017, pp. 203–209.

[51] T. Maheshwari, A. N. Reganti, U. Kumar, T. Chakraborty, and A. Das, "Revealing psycholinguistic dimensions of communities in social networks," *IEEE Intelligent Systems*, vol. 33, no. 4, pp. 36–48, Jul 2018.

[52] F. Celli, F. Pianesi, D. Stillwell, and M. Kosinski, "Workshop on computational personality recognition: Shared task," in *Proceedings of the Workshop on Computational Personality Recognition*. Boston, MA, USA: AAAI Press, Jul. 2013, pp. 2–5.

[53] F. Celli, B. Lepri, J.-I. Biel, D. Gatica-Perez, G. Riccardi, and F. Pianesi, "The workshop on computational personality recognition 2014," in *Proceedings of the 22nd ACM International Conference on Multimedia*. Orlando, FL, USA: ACM, Nov. 2014, pp. 1245–1246.

[54] U. Kumar, A. N. Reganti, T. Maheshwari, T. Chakraborty, B. Gambäck, and A. Das, "Inducing personalities and values from language use in social network communities," *Information Systems Frontiers*, vol. 20, pp. 1219–1240, December 2018.

[55] T. Maheshwari, A. N. Reganti, U. Kumar, T. Chakraborty, and A. Das, "Semantic interpretation of social network communities," in *Proceedings of the 31st AAAI Conference on Artificial Intelligence*. San Francisco, CA, USA: AAAI Press, Feb. 2017, pp. 4967–4968.

[56] T. Maheshwari, A. N. Reganti, S. Gupta, A. Jamatia, U. Kumar, B. Gambäck, and A. Das, "A societal sentiment analysis: Predicting the values and ethics of individuals by analysing social media content," in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, Valencia, Spain, Apr. 2017, pp. 731–741.

[57] J. W. Pennebaker, R. J. Booth, R. L. Boyd, and M. E. Francis, *Linguistic Inquiry and Word Count: LIWC2015*, Austin, TX, USA, 2015.

[58] S. S. Tekiroğlu, G. Özbal, and C. Strapparava, "Sensicon: An automatically constructed sensorial lexicon," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*. Doha, Qatar: ACL, Oct. 2014, pp. 1511–1521.

[59] S. Baccianella, A. Esuli, and F. Sebastiani, "SentiWordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining," in *Proceedings of the 7th International Conference on Language Resources and Evaluation*. Valletta, Malta: European Language Resources Association, May 2010, pp. 2200–2204.

[60] C. Strapparava and A. Valitutti, "WordNet-Affect: An affective extension of WordNet," in *Proceedings of the 4th International Conference on Language Resources and Evaluation*. Lisbon, Portugal: European Language Resources Association, May 2004, pp. 1083–1086.

[61] B. Verhoeven, W. Daelemans, and T. De Smedt, "Ensemble methods for personality recognition," in *Proceedings of the Workshop on Computational Personality Recognition*. Boston, MA, USA: AAAI Press, Jul. 2013, pp. 35–38.

[62] J. Pennington, R. Socher, and C. Manning, "GloVe: Global vectors for word representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*. Doha, Qatar: ACL, Oct. 2014, pp. 1532–1543.

[63] N. Majumder, S. Poria, A. Gelbukh, and E. Cambria, "Deep learning-based document modeling for personality detection from text," *IEEE Intelligent Systems*, vol. 32, no. 2, pp. 74–79, Mar. 2017.

[64] F. Rangel, P. Rosso, B. Verhoeven, W. Daelemans, M. Potthast, and B. Stein, "Overview of the 4th author profiling task at PAN 2016: Cross-genre evaluations," in *Working Notes Papers of the CLEF 2016 Evaluation Labs. CEUR Workshop Proceedings*, K. Balog, Ed., Sep. 2016, pp. 750–784.

[65] B. Perozzi, R. Al-Rfou, and S. Skiena, "DeepWalk: Online learning of social representations," in *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2014, pp. 701–710.

[66] T. Maheshwari, A. N. Reganti, T. Chakraborty, and A. Das, "Socio-ethnic ingredients of social network communities," in *Companion of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. Portland, OR, USA: ACM, Feb. 2017, pp. 235–238.

[67] L. Tang and H. Liu, *Community Detection and Mining in Social Media*. Morgan & Claypool, 2010.

[68] L. Martin, G. B. Gloor, S. Dunn, and L. M. Wahl, "Using information theory to search for co-evolving residues in proteins," *Bioinformatics*, vol. 21, no. 22, pp. 4116–4124, Nov. 2005.

[69] L. Hubert and P. Arabie, "Comparing partitions," *Journal of Classification*, vol. 2, no. 1, pp. 193–218, Dec. 1985.