Multimedia Tools and Applications (2023) 82:8811–8854
https://doi.org/10.1007/s11042-021-11857-1

1209: RECENT ADVANCES ON SOCIAL MEDIA ANALYTICS AND MULTIMEDIA SYSTEMS: ISSUES AND CHALLENGES

The homophily principle in social network analysis: A survey

Kazi Zainab Khanam[1] · **Gautam Srivastava**[1,2,3] · **Vijay Mago**[1]

Received: 20 November 2020 / Revised: 21 September 2021 / Accepted: 23 December 2021 /
Published online: 18 January 2022

Abstract
In recent years, social media has become a ubiquitous and integral part of social discourse. Homophily is a fundamental topic in network science and can provide insights into the flow of information and behaviours within society. Homophily mainly refers to the tendency of similar-minded people to interact with one another in social groups than with dissimilar-minded people. The study of homophily has been very useful in analyzing the formations of online communities. In this paper, we review and survey the effects of homophily in social networks and summarize the state-of-art methods that have been proposed in the past recent years to identify and measure those effects in multiple types of social networks. We conclude with a critical discussion of open challenges and directions for future research.

Keywords Homophily · Social network analysis · Natural language processing · Machine learning

1 Introduction

Homophily is defined as the tendency for people to seek out or be drawn to others who are similar to themselves. The people's networks tend to be more homogeneous than

✉ Gautam Srivastava
srivastavag@brandonu.ca

Kazi Zainab Khanam
kkhanam@lakeheadu.ca

Vijay Mago
vmago@lakeheadu.ca

[1] Department of Computer Science, Lakehead University, Thunder Bay, Canada

[2] Department of Mathematics and Computer Science, Brandon University, Brandon R7A 6A9, Canada

[3] Research Center for Interneural Computing, China Medical University, Taichung 40402, Taiwan, Republic of China

heterogeneous such that the communication between similar people occurs more frequently than with dissimilar people [70]. It is a well-established phenomenon that has been observed to occur frequently in social networks, where users with similar contexts tend to connect constantly. This principle is also a meticulously researched field in the domain of social sciences [33, 39, 40, 66, 86, 125]. The main driving forces for initiating these networks are social influence and homophily. In other words, the importance of establishing connections between people does not rely upon 'what you know' but 'who you know.' To study this phenomenon, various studies have been conducted by sociologists on multiple sociodemographic dimensions of race, age, social class, culture, and ethnicity. For example, friends, colleagues, spouses, and other associates are inclined to mix with those who are similar to them than with randomly selected members of the same population [70, 116].

Homophily research has traditionally been carried out by surveying a group of human participants who, in most cases, belonged to a certain geographical region [1, 15, 29, 114]. According to one study, American high school students are more likely to befriend peers of the same race and gender [75]. Homophily was first divided into two types: *status* homophily and *value* homophily [56, 60, 110]. The concept of status homophily is centered on an individual's social standing, implying that people in similar social situations are more likely to mix. Status homophily also includes the major sociodemographic dimensions such as race, ethnicity, sex, or age, and procured characteristics like religion, education, occupation, or behaviour patterns [4]. Value homophily, on the other hand, is based on the similarity of people's thinking, leading to the assumption that people with similar thoughts are more likely to interact with others, even though differences may lie in their social positions. It includes the internal states that are thought to shape an individual's orientation towards the future [25, 27, 79]. Homophily is a broad field with also some strong connections to social cognitive mechanisms [30]. Albert Bandura proposed the Social Cognitive Theory, which stresses the dynamic interplay between individuals (personal variables), their behaviour, and their settings [3]. The behaviour between the individuals includes the heartfelt emotions (empathy) that the individuals feel for one another.

Although researchers have successfully conducted experiments with human beings, the results were often based upon real-world scenarios of only small groups [107]. To fill the gap in the analysis, social media platforms come in handy as social networking sites such as Twitter and Facebook have become extremely widespread, with over 126 million daily Twitter users [51, 91, 106] and Facebook had approximately 1.2 billion daily users [32, 109]. Reactive interfaces like those available through social networks provide users with the opportunity to be more open about their opinions, perspectives, thoughts, likes, and dislikes [18, 53, 62]. As a result, social media platforms are becoming more and more popular among users [6, 77]. These platforms are known to help users feel more involved. Users feel that they can participate in events that are happening around the world. Furthermore, such platforms help users in raising their voice against unjust acts or issues [34]. Therefore, both status and value homophily have been analyzed recently in social networks to evaluate whether these types of homophily phenomenon exists in these types of networks [7, 80, 111]. Moreover, the effect of homophily has been vastly studied in different types of social media data [86]. However, no detailed survey has been conducted to date based on the works of social media networks related to the homophily principle. Therefore, the main aim of this paper is to focus on providing a thorough review of the related works conducted on social media networks based on the homophily principle. The rest of the paper is organized as follows. Section 2 presents the methodology that has been used to extract high-quality articles to conduct the survey. Section 3 discusses the role of homophily in the various ways in which the homophily effect has been analyzed in multiple domains of social media data.

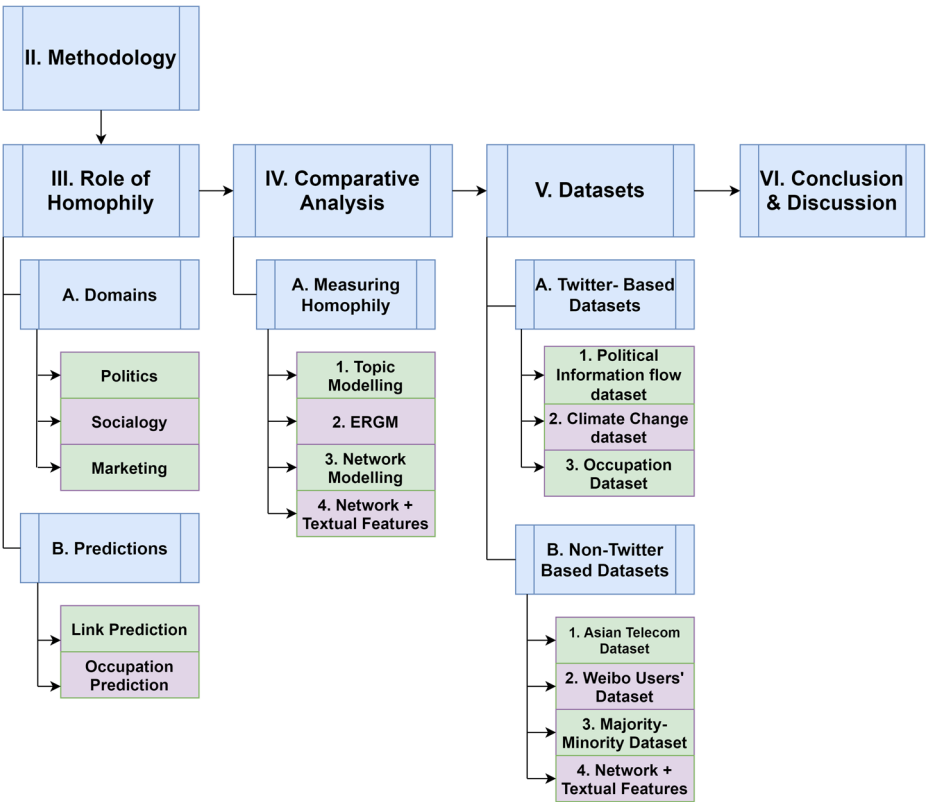

Fig. 1 Overall structure of the survey paper

Section 4 discusses the predictions made in many fields of social networks by using the homophily effect. Section 5 introduces a comparative study of the social network analysis, conducted by measuring homophily in multiple applications, the different types of network models constructed in each of the proposed models. Section 6 includes the different types of datasets used to validate these proposed, homophilous models. Section 7 discusses the state-of-art methods used for detecting homophily in social networks, the limitations of these approaches, and directions for future research and concludes the survey. Fig. 1, shows the overall structure of the paper.

2 Methodology

The research was conducted in two steps:

1. A keyword-based search was used for collecting conference papers, journals, and other types of articles from the web scientific indexing service: Google Scholar[1]. The following query was used for the keyword-based search: [*"social media"*] *AND*

[1] https://scholar.google.ca/

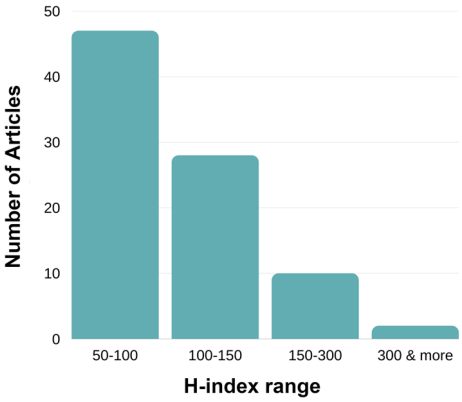

Fig. 2 The number of journals' H-index from a range of 50 and above

[*"Homophily" OR "degree distribution"*]. In this way, we were able to exclude articles that mentioned only social media and were not related to the homophily domain.

2. We have mainly focused on the articles which satisfies one of the three conditions: a) the venue where the article has been published that has an h-index of 50 or above, b) the articles belonging from Q1 or Q2 journals c) the articles having a minimum of 100 citations. We have used this approach as it is important to not only find the appropriate papers based on keywords but also to extract papers from top venues. The papers selected were from 2015 onwards to examine the methodologies utilized in recent years in greater depth. However, if any articles have major contributions, such as introducing novel algorithms or approaches used in measuring the degree of homophily then, these articles are considered for this survey because homophily is not a recent concept and the impact of these papers is more important than the year of publication.

In the initial stage, 347 articles were identified. The h-index, number of citations, and year of publication were used as features. Using these features, we were able to filter out 108 relevant research articles out of 347 papers. The h-index metric for each of the journals where the articles were published, was collected from the SCImago Journal Rank (SJR) website[2]. Recent surveys have also reported adopting a similar approach [37, 49, 64]. Figure 2, shows the h-index of the articles cited in the survey. We can see that most of the articles' h-index is from 50-100. H-index is a journal-level metric that is used to evaluate the impact factor and citations of the publications, obtained from the SJR website. Table 3 of Appendix A shows the articles selected for this survey with the venue, number of citations, quartile, h-index, as well as the year of publication.

An article from a high h-index venue, with a high number of citations, shows that the paper is reliable and trustworthy for the academic community. Figure 3 shows the total percentage of journals, conferences, and other types of articles such as book chapters, workshop papers that have been cited in this paper. It can be seen from Fig. 3 that most of the articles selected for this survey are from journals. Furthermore, in Fig. 4 we can see that majority of the articles are taken from Q1 journals. However, for the conference and workshop articles, the information about belonging to a certain quartile or h-index metrics was missing since these metrics are journal-level metrics. For such cases, we have only focused on the remaining metrics. Besides, the papers were selected from 2015 onwards so that the approaches

[2]https:/www.scimagojr.com

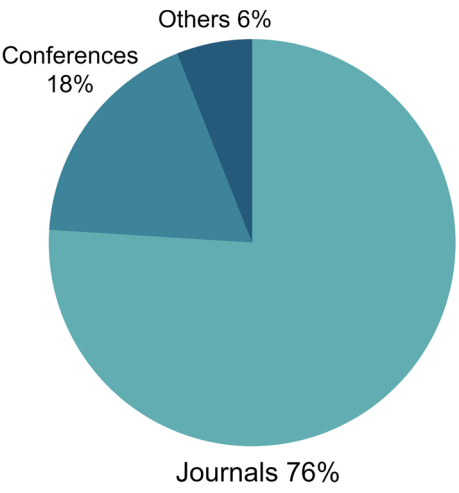

Fig. 3 Percentages of Journals, Conferences, and other types of articles cited in the survey paper

used in recent years could be studied more exhaustively [61, 121]. Figure 5 shows that most of the articles have been selected from 2015 onwards.

A word cloud, as shown in Fig. 6, was generated from the abstracts of the papers selected to get a visualization of the most important word in the field of homophily [48]. We implemented a simple python code to form the word cloud. The abstracts are pre-processed by converting the text to lower case, removing the punctuation, and commonly used English stop words, available in the NLTK library. Then the word cloud is built using a word cloud library. The importance of each word is represented by the font size and colour. The darker the colour of the word, the more significant the word is. The larger the font size of the word, the more frequently the word has occurred in the abstracts. Figure 6 shows that the homophily word has the largest font which depicts that most of the papers were about

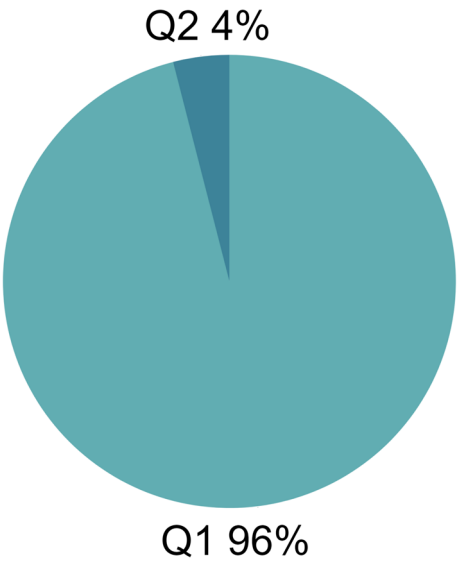

Fig. 4 Proportion of Q1 and Q2 Journals referenced in the survey paper

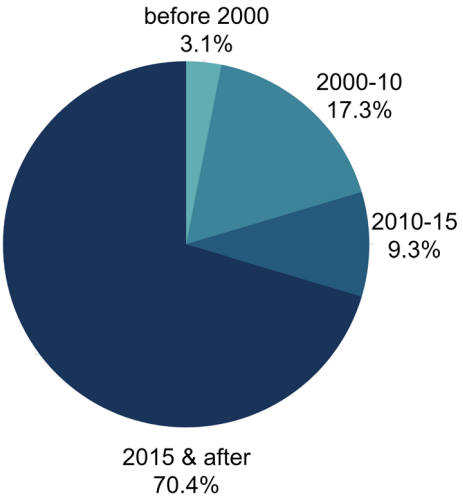

Fig. 5 Percentages of the Year of Publication of each of the articles

homophily-related issues. Moreover, the word Twitter and social networks have the second-largest font. The colour of the font is darker which shows most of the papers discussed social media. Moreover, Fig. 6 shows that words like social, graph, networks, echo, and homophily are used quite frequently. Interesting, the word *echo* is used quite frequently as well, as it is either used to raise their voices, or talk about politics or debate about a particular issue.

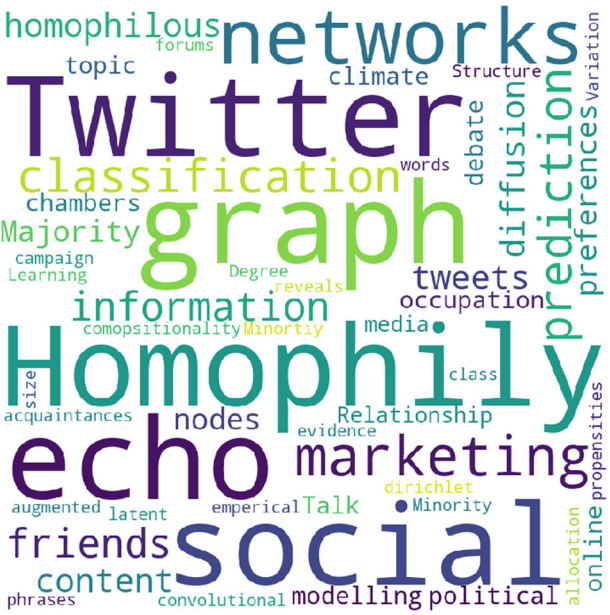

Fig. 6 Visual representation of the important words from the abstracts

3 Role of Homophily in social media

The Internet can connect people, with all kinds of interests, all around the world. As a result, it can be assumed that when social ties are formed between individuals in social media, homophily will be less likely to appear among users. However, homophily has rather shown to increase in social media platforms [12, 57, 119, 120]. For instance, the earliest research has highlighted that, when people used Microsoft Instant Messaging, users were more interested to converse with others belonging to equivalent ages, locations, and native language [119]. Moreover, the more individuals communicated with one another, the more related their online searches were [19, 20, 100]. Similarly, increased homophily was also detected among groups of Facebook friends, where they had similar thoughts on ideology or political orientation [63, 69]. The effect of homophily has been studied in various domains to observe if homophily has any effect on social networks or not. For example, Twitter is a popular micro-blogging platform that has been considered an effective tool for studying the interactions between social media users [58, 71, 94]. Homophilic studies have been conducted at an exhaustive rate in the domain of politics, marketing, and sociology as well [39, 45, 47, 112]. Insightful information has been extracted from these approaches. The following subsections discuss in detail the approaches that have been proposed in these domains.

3.1 Politics

In the field of politics, homophilic studies range from analyzing the users participating in political debates to observing the network of politically engaged users on a variety of political activities [31, 39, 42, 111, 117]. On a politically oriented website named "Essembly", users were observed to form positive and negative ties with people having parallel thoughts and different reasoning on an ideology respectively [117]. Furthermore, when smaller networks were studied more in-depth, several characteristics such as gender, age, and level of education proved to be strong predictors with network structural characteristics [26]. Moreover, these characteristics were used for investigating the existence and strength of positive ties among individuals [35, 97].

Homophily principle was also used to study the flow of political information among the majority and minority groups on the Twitter platform [39, 42, 59]. Recent studies have shown that the majority of larger groups received political information more quickly than smaller groups. Both groups were exposed to similar political information and it was observed that the flow of information was faster among the larger group. Substantial evidence of homophily was detected when users following a specific political party were more likely to connect with other users following the same party. The flow of information through the social media network was faster among the majority groups since they had more network connections and so they received more information at a faster rate. To sum up, political information is considered extremely influential information. Therefore, increasing exposure to such information among the like-minded majority (large) users can further increase political divergence among the users.

Furthermore, social networks of users exchanging views about global warming on Twitter were examined. The users' attitudes towards global warming were classified based on their message content [46, 111]. The social networks were categorized by opinion-based homophily and the users were manually labelled as "skeptic" and "activist", based on their message content. Results have shown that users generally communicate only with other similar-minded users, in communities that are influenced by a common view. Moreover, the messages of like-minded users have shown to be a positive sentiment in most of the cases,

whereas, messages from skeptics and activists held a more negative comment. Overall, discussions of climate change in social media often take place in the polarising "echo chambers" where political issues are discussed, and also in "open forums" and mixed-opinion communities [35].

3.2 Sociology

Homophily, meaning "love of sameness", is considered to be a sociological theory that like-minded people will be inclined towards each other and will have a tendency to act in a similar way [60]. This behaviour of individuals has been studied on social media platforms as well. Social media generally consists of majority and minority groups, where the majority group is considered to have stronger connections with one another in its group and also tends to have higher network communications [7, 39]. Compared to the majority group, the minority group not only has fewer members in its group but is also deprived of receiving information quickly [50]. Thus, the relationship between majority and minority groups in social media is studied in depth to observe how the groups and the size of the groups are formed and the groups react to one another in social networks [41, 50].

To study the influence of homophily between the minority groups, the levels of homophily were calculated by combining the centrality measures with the preferential attachment network [4, 50, 89]. The model focused on multiple ranges of homophily and density of populations by capturing the degree distributions and ranked the minority groups in empirical social networks of scientific collaboration and dating contacts [50]. Experimental results have shown that as the volume of the minority group decreased, the heterophilic interactions were greater than the homophilic interactions. However, multiple assumptions were made. For instance, all the members of the minority groups were considered to be equally active and behave in a similar pattern and the group size differences were omitted. These factors can cause a biased estimate in the ranking of the groups. A major drawback was also faced when validating the proposed model such as finding adequate numbers of large-scale data representing the minority groups, since, remote and hard-to-reach minority groups are often absent from the social network datasets [99].

3.3 Marketing

Comparative analysis has also been conducted on homophily and social influence effects on product purchasing [66]. This analysis examined problems related to whether a company should target customers based on homophily or the social influence effect. If a company relies on the homophily principle then they target the existing customer's friends directly as they tend to purchase similar products. Whereas, if the firm emphasizes the social influence of the existing customers then they only target the existing customers and rely on them to promote to their social circles. Therefore, for cost-effective marketing strategies, it is extremely important to separate these two effects. However, such approaches are challenging since both phenomenons end up producing similar outcomes. As a result, a product choice model has been designed via the hierarchical Bayesian model which was implemented with a dataset that consists of both communication and product purchase information over three months provided by Asian Telecom Company [66]. A strong homophily effect was detected on the choice of products. When one of the factors was ignored in the Bayesian model, it resulted in an overestimation of the other factor and this shows that social influence and homophily effects are highly connected. Ignoring any of the factors leads to biased estimates in the Bayesian model. Furthermore, as network structures are

versatile [28], it was difficult for the model to detect strong and weak ties in network structures. This is because some people in the network might have many friends with weak ties to one another, while others might have few friends but with extremely strong ties. As a result, the model can be further improved to inspect the strength and the impact of social ties concerning a customer's decision on product purchasing. This will help to identify customer preferences in such versatile networks. Thus, an improved model is required that can further differentiate the effects of homophily and social influence.

Gender homophily was also investigated at a global scale in the online book market such as amazon.com to address the research question of whether readers are more biased to reading books from the same authors or did diversity exist in book selections [13]. In online book markets, book sales lead to important ties among the books. These book ties create large book networks that model the collective information about the reading habits of the customer. The study was conducted with a large volume of a dataset that had records of the books sold to readers to investigate gender homophily on a large scale. Assortative measures, such as Newman's degree-based Assortative co-efficient metric was used to measure homophily between gender-based readers and authors [81, 82, 90]. Assortativity or Assortative Mixing is defined as the tendency of a network's node to connect to other nodes that behave in similar patterns [16]. Their findings have shown strong homophily in the book networks [13]. Particularly, readers that prefer reading female-authored books tend to buy other female books. While the male-authored book readers tend to buy fewer female-authored books, compared to male-authored books. Thus, these findings have revealed the essence of gender homophily. However, their study had some limitations such as finding the name of the main author as many books have non-English author names and it is difficult to figure out the gender from non-English author names.

Research has also been conducted to study the presence of homophilic patterns based on the usage of hashtags in a Twitter mention network based on a Cause-Related Marketing (CRM) campaign [113]. CRM is a mutually beneficial collaboration between a corporation and a nonprofit organization. It is designed to promote social responsibility in the public community. However, CRM is a risky and controversial issue since this campaigning varies from receiving skepticism to full support of the customers [5, 78]. Gillette's CRM campaign "The Best Men Can Be" was used, to test the hypothesis that whether homophily exists in such marketing campaigns or not [113]. The brand's goal was to address concerns based on gender inequality and bullying of men and encourage a better lifestyle for youngsters. The company, moreover, guaranteed to donate 1 million to NGOs fighting for gender inequality. When the campaign ad was released, the ad received positive feedback from some of the customers because of its positive message [52]. However, others felt that the ad was a bit offensive representing men as sexually harassing, and bullying. Thus, it received negative feedback from the rest of the customers [108]. Hence, two groups were formed in social media where one group was supporting the cause while the other group opposed its motive. As the brand is well-known and discussions on this campaign became a trending topic on Twitter. Thus, this CRM campaign was an ideal fit to analyze how the users communicated and reacted with other users on Gillette's ad. For CRM's marketing campaign, topic modelling was used for extracting information [87]. Topic modelling was conducted on 100,000 original tweets, profiling the topics related to the CRM campaign tweets [10]. Based on the users' engagement, the network of the CRM event's related hashtags was analyzed with the aid of Exponential Random Graph Models (ERGMs) [101]. ERGMs are statistical models that are commonly used for analyzing data regarding online social networks [21, 105]. Results generated from this model showed an increased tendency of homophily on the network of users. The degree of homophily inspected was based upon the

common views of the users. The results of the topic modelling on Twitter have revealed that users are highly dependent on established social networking platforms to discuss important issues [38, 52]. Furthermore, users tend to react more to the tweets of influential, popular users which enables the users to be more reactive during online discussions. Moreover, these users showed homophily in the usage of hashtags. Thus, ideological hashtags served as measures of homophily as these hashtags refer to a person's identity and thoughts [11]. One such example of an ideological hashtag is the usage of *#BlackLivesMatter* or *#AllLivesMatter*, which reflects the user's ideological position is based on social justice issues to a large extent. Therefore, hashtags not only express a user's self-identity but also helps similar users to identify and connect in a versatile community but with same ideology [108].

4 Using Homophily for predictions

Homophily concepts can be implemented for predicting certain features. For example, homophily principles have been applied in the link prediction area for studying the probability of one user to be connected with another user. On the other hand, homophily in Twitter has also been examined to predict the occupation of users based on the information of their followers and followings' IDs. The usage of the homophily concept in predicting certain features is discussed below.

4.1 Predicting occupations

To predict the occupation of Twitter users, the homophily principle was used to conduct social network analysis on the biographical content of the user's follower/following community [86]. Occupation prediction is considered a multi-classification problem since the model is specialized in predicting multiple occupational classes. Furthermore, the results concluded that a user's follower/following community provides insightful information for identifying the occupational group of each of the users. The model was designed with Graph Convolutional Network (GCN), which has enhanced the model to work efficiently by training on only a small fraction of data. GCN is a recently proposed graph-based neural network learning model, which specializes in learning graph-structured network data [55]. Thus, by using the homophily principle and GCN, a better result was achieved for predicting occupation class with an accuracy of 61%. Similar work was done to predict the occupational class of Twitter users, where the dataset contained the historical tweets of the users [92]; however, the accuracy of only 50% was achieved.

4.2 Predicting links

The homophily principle was used as motivation in various works of link prediction, where link prediction is calculated based on the similarity between two entities. As a result, it can be used to predict future possible links in social networking platforms [23]. For example, researchers have proposed a model to investigate the associated links of the document's topic distribution between people discussing related topics. This study shows how topic distribution is mainly affected by the distribution of the topics of its nearest neighbours [54, 113, 115].

Particularly, a joint model was proposed in which link structure has been applied to define clusters. Here, each of the clusters was allocated with its segregated Dirichlet before

topic distribution. Using such priors has shown to be very helpful as in previous works only document priors were applied [73, 123]. Discriminative and max-margin approaches [123, 124] have been used for designing the contextual documents and generating good link predictions. Moreover, lexical terms have been used in the decision function to improve the strength of the prediction [83].

In summary, users in social media are not only comfortable at expressing their self-identities but also tend to connect, with similar interests, in a versatile community. Several studies have been carried out to study the homophilic patterns especially among the majority and minority groups. Studies have deduced that the majority and larger groups receive information faster than smaller, minority groups and information reaches like-minded individuals more quickly [39, 44, 68]. Whereas, the minority groups are deprived of receiving information instantaneously as such groups have fewer members in their groups hence have fewer network connections [50]. On the other hand, a strong homophily effect was detected on customers having similar product tastes and based on the attraction of users having a common view [66, 111, 113, 118].

5 Comparative study of related works for homophily detection

In this section, comparative analysis has been performed on various approaches proposed to detect and calculate the level of homophily. Besides, the network and language models defined by each of the states of art methods are also explained in detail.

5.1 Measuring the degree of Homophily

Multiple approaches have been proposed for measuring the level of homophily with the aid of topic modelling and network modelling. Recently, the degree of homophily has also been calculated by combining both textual as well as network features by using a highly efficient neural network model that has outperformed the existing traditional methods [39, 50, 86, 98, 113].

5.1.1 Topic modeling

Topic modelling is considered as an unsupervised machine learning technique that does not need the aid of humans to determine the topic of a set of documents [9]. It can automatically detect the main theme of given paragraphs or documents by clustering groups of words or similar expressions that best portray the set of documents or paragraphs. The topic model proposed based on the homophily concept specializes in detecting high-quality topics to test the hypothesis that whether people talking about similar topics are connected or not [115]. For example, the Latent Dirichlet Model (LDA), is a topic model, that maps the documents to the topics based on the distribution of words [10]. LDA model was modified with the concepts of homophily to not only detect the high-quality topics but also predict whether the people having similar posts in social media are connected or not [10].

Generally, the most frequent words of each of the documents are aggregated into K and words clusters by using the k means algorithm [67]. Thus, for any word token $w_{d,n}$, for the word token belonging to a cluster k, any other token $z_{d,n}$ being a neighbor of $w_{d,n}$ will also belong to cluster k. The d represents the document and n represents the number

of documents. Therefore, in order to find the topic k's major words, skip-gram transition probability [72] is calculated for each $w_{k,i}$ word as in (1).

$$S_{k,i} = \sum_{j=1, j\neq i}^{N_k} p(w_{k,j}|w_{k,i}) \tag{1}$$

where, N_k indicates the number of words in topic k, words with the highest probabilities are used as the designated topic words for each of the documents in the sample. Regression is used to compute the topic distribution between d and d', for predicting the link between the two documents, which is dependent on the similarity of their topic patterns. Therefore, the regression value is defined in (2).

$$R_{d,d'} = \eta^T (\overline{z_d} \circ \overline{z_{d'}}) + \tau^T (\overline{w_d} \circ \overline{w_{d'}}) \tag{2}$$

where, $\overline{z_d} = \frac{1}{N_d} \sum_n z_{d,n}$. Similarly, $\overline{w_d} = \frac{1}{N_d} \sum_n w_{d,n}$; $\circ$ denotes the Hadamard product [43]; η and τ are the assigned topic based weight vectors and document link predictions.

In some studies, the perplexity metric was used as a measurement for evaluating the model's topic modelling performance [22]. Perplexity is a measurement based on the quality of a probability model predicting a sample. Results have shown that the proposed model outperformed the traditional LDA model for topic modelling. Furthermore, for validating the model in terms of its performance for predicting document link prediction, the Predictive Link Rank (PLR) metric was used. PLR outputs the average rank of a document with the documents to which it has been linked. High training performance was achieved that showed user interactions can contribute to better link prediction. However, the testing performance score was much lower than the training performance score. This shows that the model has over-fitted since the model could not perform well with the testing dataset. Even though the new model outperformed the traditional LDA method, for the document link predicting task the overfitting issue was not resolved.

LDA-based topic modelling focuses on topics co-occurring frequently. However, the main drawbacks of LDA based approach are the need of specifying the "appropriate" number of topics that the LDA has to predict [10]. Statistical indices have been proposed to address this issue [2, 14] which include differentiating each pair of the topics by the difference of each pair of topics or their distance. Although these methods can approximately calculate the number of topics in a given corpus, proper gold standards or benchmarks still do not exist. Therefore, human interpretation is still required for rendering the topics into the unsupervised topic modelling method [17]. Most importantly, the performance of these methods is not analyzed in documents, such as tweets, which consist of only sentences with a few words. Thus, for measuring the homophily of users using similar hashtags in Twitter posts, the co-occurrence of the topic was calculated by using Mimno et al.'s approach [74, 113]. Mimno et al. states that for any document, the leading words in a topic profile are likely in the same document. The coherence in Mimno et al.'s approach is defined as:

$$C(t; v^{(t)}) = \sum_{m=2}^{M} \sum_{l=1}^{m-1} \log \frac{D(v_m^{(t)}, v_l^{(t)}) + 1}{D(v_l^{(t)})} \tag{3}$$

In (3), $D(v)$ denotes the frequency the document, word v and $V^{(t)}$, topic t has a list of M words. The topic model can predict better when the result has an output that is close to

zero. [93]. To evaluate the quality of the predicted topics, the distance between each of the topics was calculated using Jensen-Shannon divergence [93, 104].

5.1.2 Exponential random graph modeling

Exponential random graph models (ERGMS) were used to evaluate whether the networks of users engaged in conversations influence the users' responses in online discussions [113]. In ERGMs, nodes' connections at the same degree are considered to be an indicator of users conversing frequently [76]. These graph models are statistical models that are composed of network structures [65]. ERGMs have been used in multiple domains of social networks, such as social media settings and to study communication between the users for modelling the presence of any ties that may exist between the network's local and structural factors [36, 95, 96].

Generally, the ERGM model is designed by aggregated tweets and hashtags which are posted by the users [113]. The top hashtags are categorized as either conceptual or ideological markers based on the definitions provided by Blevins et al. [11]. The ideological hashtag refers to the identity or identification, perspectives. On the other hand, conceptual markers are considered as personal thoughts on particular events. Then, the users mentioned in the network as @user and the classified hashtags are fed into statistical modelling as nodal attributes. The mentioned network is modelled because it helps to visualize a discussion that starts by actively exchanging information with one another rather than relying on what others have said.

5.1.3 Network modeling

Network modelling is a flexible way of representing a group of connected objects. The objects are represented as nodes or vertices and the connection between the nodes is represented by edges. Network modelling is generally used to visualize the various types of networks. These models are constructed from the social media data consisting of users and how the users are connected. Many researchers have claimed that the flow of information in social media is dependent highly on the majority groups where more users are connected. Moreover, the majority of groups have a larger social circle compared to the minority groups. The reason behind having a larger social circle is due to homophily [39, 50].

The homophily among majority groups was studied between a network of politically engaged Twitter users [24, 39]. Due to the shortage of measuring the political orientation and ideology of Twitter users directly, the researchers emphasized that the users following politicians from the two major parties- *Conservatives* and *Liberals* of the House of Representatives, in the 2012 general election. To analyze the degree of homophily, the individuals are divided into two groups - conservatives *(C)* and liberals *(L)*, depending on which political party each group supports such that: $t \in C, L$. The group sizes are normalized and symbolized as wt which is the weight of the tweet, where, $w_C + w_L = 1$. Conservatives were randomly selected as the majority group and liberals were considered as the minority group. Therefore, $w_C \geq 0.5$.

When two individuals belonging to the same group are randomly selected, then the probability of the two individuals communicating with each other is denoted by π_s. The probability of the two individuals connecting belonging to different groups is represented by π_d. Moreover, it is also logical to assume that individuals belonging to the same group have a higher tendency to interact with each other compared to two individuals of two different groups communicating with one another. As a result, $\pi_s > \pi_d$. Thus, an individual

belonging to a group t will be having similar $\pi_s w_t$ interactions, and $\pi_d(1 - w_t)$ different interactions. Therefore, in this study, the Homophily principle has been evaluated as such:

$$H_t = \frac{\pi_s w_t}{\pi_s w_t + \pi_d(1 - w_t)} \tag{4}$$

In (4), the greater the value of $\pi_s w_t$, the higher the degree of homophily. As a result, if conservatives are more prominent and links are formed with similar types, then conservatives would tend to be more homophilous. Similarly, liberals would be more heterophyllous. Furthermore, the time taken to reach the information to the majority and minority groups was also taken into account. So it was considered that each user produces information with probability of ε at time $\tau = 0$. The Bass model or the Bass Diffusion Model was used to generating the proposed model [8]. The Bass model uses a differential equation that describes the procedure of new products getting adopted in a population. In this case, the product is the political tweets. If the interaction occurs between two users, then the user exposed to information transfers the information to the unexposed user with a probability of q. Hence, by following the Bass model, the rate of information diffusion is defined as such:

$$F_t^\tau = F_t^{\tau-1} + (1 - F_t^{\tau-1}) f_t^\tau \tag{5}$$

In (5), F_t^τ is defined as the fraction of group t receiving information at time τ which is then connected to the fraction transmitted information at time $\tau - 1$. F_t^τ is the chances of group t getting information at time τ if not being exposed to information at time $\tau - 1$. Therefore, f_t^τ is defined as:

$$f_t^\tau = q w_t \pi_s F_t^{\tau-1} + q(1 - w_t)\pi_d F_{-t}^{\tau-1} - q^2 w_t(1 - w_t)\pi_s \pi_d F_t^{\tau-1} F_{-t}^{\tau-1} \tag{6}$$

The symbol $-t$ in (6) refers to the other group. The first term denotes the likelihood of receiving the information from the individual belonging to the same group. The second term denotes the likelihood of receiving information from a different group. The third term refers to the likelihood of both groups receiving information. In conclusion, if biased interactions are present, $(\pi_s > \pi_d)$, the majority group member will receive information faster than the minority group $(F_C^\tau > F_L^\tau)$ for every τ times. Based on receiving a higher probability score for the majority group, it is deduced that homophily is directly proportional to the rate of flow of information among the users. Moreover, the diffusion of information is relatively uniform with groups having a higher number of connections based on having similar political orientations. Thus, larger groups are exposed to information at a faster rate. Therefore, a close relation of homophily and diffusion of information is shown in this approach.

A similar approach was also used to measure the level of homophily between the minority groups [50] which is shown in (4). The homophily was used as an additional parameter in the famous model of preferential attachment proposed by Barabási and Albert [4]. Preferential attachment means that a node is more likely to receive new links if it has a higher number of connections. Thus, such nodes are more powerful since they can tightly hold links with one another. The growth of complex evolving networks was calculated using the fitness model which is based on the Barabási–Albert model [50]. In this model, nodes with different types of characteristics can grasp links at different rates. Hence, the fitness is calculated by the degree distribution of each of the nodes. This is how the model can predict a node's growth. In the preferential Attachment model, at each step, a new node that just approached, its degree, and group attachment is calculated for the possibility of the node to

be attached to the pre-existing nodes. The chances of node j to be connected to node i is defined as:

$$\Pi_i = \frac{h_{ij}k_i}{\sum_l h_{lj}k_l} \tag{7}$$

In (7), k_i generally, represents the degree of node i and h_{ij} is the similarity between nodes i and j. The similarity between each of the nodes is built, based on the nodes' attachment when the network was generated. If by any chance, the new node is not confronting individuals from the same network, it can stay deserted till the node confronts a newly approached node that is coming from the same network.

On the other hand, when the online debate on climate change was studied, the degree of homophily among the individuals was measured on the number of times the edges were connecting users on homogeneous/heterogeneous views [111]. The high frequency of edges between the homogeneous users, and similarly, the low frequency of edges between the heterogeneous users were considered as the measure of homophily. The probability of picking node i as the root or focus node for a given edge were denoted by:

$$P_{source}(i) = \frac{k_{out}(i)}{\sum_{j\in a,s} k_{out}(j)} \tag{8}$$

$$P_{target}(i) = \frac{k_{in}(i)}{\sum_{j\in a,s} k_{in}(j)} \tag{9}$$

In (8) and (9), k_{out} is node out-degree and k_{in} is node in-degree. This mathematical technique generates nodes' networks with homogeneous degree distributions.

Another type of metric that is used to measure the quantity of homophily includes Newman's global assortativity coefficient [81]. Assortativity or Assortative Mixing is defined as the propensity of a network's node to connect to other nodes that behave in similar patterns [88]. Newman's global assortativity coefficient is a vertex degree metric used to identify whether vertices in the graph tend to mix with homogeneous or heterogeneous vertices. The global assortativity coefficient is measured in terms of the discrete characteristics for each vertex. The coefficient gives values within the range [-1,1] [13]. The coefficient is zero for a given network if the vertices connect randomly. A high positive value indicates that the high degree vertices link preferentially with other high degree vertices and the other way around for the low degree vertices. The assortative metric was used to measure the degree of gender homophily in online book networks [13]. This coefficient is mainly applied for evaluating whether customers from the same gender tend to buy books written by the same gender author or are the books bought randomly by chance.

$$r^k = \frac{\sum_i e_{ii} - \sum_i a_i^2}{1 - \sum_i a_i^2} \tag{10}$$

Equation 10 shows how Newman's degree-based assortativity has been defined for calculating homophily for the given situation. For a given book graph G^k, the assortativity coefficient r is denoted by i, j, where $i, j \in \{$*male first author, female first author, book collections*$\}$. Out of all the edges in G^k, the portion of edges that link two books from any of the gender categories i and j is depicted as e_{ij}. The fraction of occurrences of same-gender edges in G^k is denoted by the term $\sum_i e_{ii}$. The a_i term represents the portion of edges in which one of the ends is a book from category i gender. For a randomly connected graph $\sum_i a_i^2$ term shows the portion of same-gender occurring edges. The normalized differences between the real and random fraction of the same occurring gender edges in a graph G^k are denoted by r^k.

5.1.4 Combining network and textual features

Recently, a new neural network model has been proposed known as the Graph Convolutional Network (GCN) [55]. GCN has been used for extracting the textual and the network features for identifying the homophily connection between the Twitter users and their followers/followings list [86]. GCN has not only enabled the model to achieve high performance but also the model was successfully trained with only a fraction of data. GCN graph-based neural network model *f(X,A)* with layer-wise propagation rules is defined as such:

$$\widehat{A} = D^{-1/2}(A + \lambda I)D^{-1/2} \tag{11}$$

$$X^{l+1} = \sigma(\widehat{A}X^l W^l + b^l) \tag{12}$$

In (11) and (12), X denotes the matrix of the features for each of the nodes(users). X^0 is the initial feature with a input size of *(nodes * features)* and A is the adjacency matrix of the dimensions *(nodes * nodes)*. D represents the degree matrix of $A + \lambda I$, where λ is the hyperparameter that controls the weight of the node among its neighbourhood. W^l, b^l are the trainable weights and bias for the l^th layer. In each GCN layer, nodes accumulate their closest neighbours' features by linearly converting the representation using weight W and bias b respectively. σ represents the activation function used in the GCN model for optimizing the performance. Then, the number of GCN layers determines the path of the node from its closest neighbours' features. The inputs of the adjacency matrix $\widehat{A}$ are all the network IDs (target users and their followers and following list IDs) which is a feature matrix of the biographical descriptions of each of the target users' followers/following lists. Pan et al. claimed that accuracy of 61% was obtained, which outperformed the results of the existing methods [86]. Thus, GCN was able to extract the rich network and textual information to learn the homophily connection between the users. However, the model was trained with only a small fraction of data, and thus if a larger amount of data would be used the model would achieve a better result for predicting the occupation of target users.

In summary, various types of models have been proposed for measuring the degree of homophily. Mainly network modelling and topic modelling have been used to find out the strength of a relationship of a user with the user's nearest neighbours, and to also investigate, how many users are involved with one another. Network modelling focused on the network features which involved calculating the number of connections each user has and the strength of the connection of the users with one another. On the other hand, topic modelling focused on clustering the main topics of each of the users, based on their textual tweets and how similar the usage of topics are with one another. Recently, the degree of homophily has also been evaluated by incorporating textual and network features with a highly efficient neural network model that has outperformed the existing traditional methods for multi-classification problems.

6 Datasets

Multiple datasets have been used for measuring the degree of homophily in multiple fields, such as link prediction and flow of information among majority and minority groups [39, 86]. For example, the dataset provided by Asian Telecom Company was used to analyze

if the homophily effect can influence product purchasing [66]. This section discusses the various datasets used to study the homophily effect.

6.1 Twitter-based datasets

Currently, there are no standard datasets that have been used to study the effect of homophily. Rather in most cases, datasets were generated with the aid of Twitter Search API [84]. Relevant words or hashtags have been used to find tweets and hashtags among the users who have posted such tweets. For example, to search for tweets about climate change, hashtags such as *#climatechange* and *#globalwarming* are used to find such specific tweets about climate change along with the user information [111]. Based on this search, an extensive network is generated from these users. Different approaches such as network modelling and topic modelling are used to measure the degree of homophily. In Table 1 the dataset index numbers - 1, 2, 4, 5, 6, 10 show examples of some of the datasets that were generated to study the effect of homophily by using this approach, as the mode of data collection was very much alike. Hence, the description of the following datasets gives an overview of how networks are generated by using the Twitter Search API.

6.1.1 Political information flow dataset

The Twitter dataset for analyzing the flow of political information among majority and minority groups was constructed by targeting the politically engaged users [39]. These users were following at least one account of a candidate running for the 2012 US elections. As a result, over 2.2 million users' data were collected from which 90 million network links were approximately identified. Users following more accounts of Republican political candidates than accounts connected with Democrats candidates were categorized as conservatives. Similarly, the users following more Democratic accounts were categorized as liberals. To measure the level of communication among the groups of supporters, approximately 500,000 retweets of the candidates' tweets, and tweets that mention candidates were also collected and analyzed. The flow of political information among the groups of Twitter users was measured based on whether or not the users received a candidate tweet or mention through these networks. Moreover, the rate of information flowing through the political network was taken into account by measuring the time taken for these retweets to diffuse across the networks.

6.1.2 Climate change dataset

Twitter API was used to collect tweets between January 2013 and May 2013 that consisted of the trending hashtags on global warming such as- *#globalwarming*, *#climatechange*, *#agw (an acronym for "anthropogenic global warming")*, *#climate*, and *#climaterealists* [111]. Moreover, followers of each of the users posting such tweets were also identified. Hence, 590,608 distinct tweets from 179,180 distinct users were used to generate the dataset. Hashtags were mainly used to search tweets as Twitter users commonly use hashtags to pinpoint a specific occasion. This enables users to search and participate in relevant discussions. Mean Sorensen similarity also known as F1 score, [103] was calculated for each of the hashtags. Sorensen's similarity score is within a range of 0 (no overlap) to 1 (identical). Greater values showed greater constancy among a major population of active users.

Table 1 Details of some of the datasets used to validate the presence of Homophily in the digital environment

	Dataset	Source	Size	Public Private	Year	Advantage	Additional Information
1	Occupation Dataset	Twitter	34,630 unique users and 586,303 edges	Public	2019	Achieved the highest accuracy at predicting the occupation of 5000 target users	Predicted the occupation of target users based on the biographical content of the target users' social circles [86].
2	Political Information Flow Dataset	Twitter	2.2 million Twitter users with 90 million network links	Private	2018	Time taken for information to flow among the majority of voters could be efficiently calculated	Information flows faster among users with higher number of connection compare to users with less number of connections [39].
3	Majority-Minority Dataset	Generate artificial undirected network	The undirected network consisted of 5000 nodes and averaged over 20 simulations	Private	2018	Calculates homophily by combining the centrality measures with the preferential attachment network	Captures the degree distributions and ranks of the majority and minority in empirical social networks [50].
4	Climate Change Dataset	Twitter	590,608 distinct tweets from 179,180 distinct users were used	Private	2015	Hashtag analysis was conducted using mean Sorensen similarity [103]	Sorensen similarity could successfully detect homophily as greater values showed a greater constancy among a major population of active users [111].
5	Emotions and Political Talk Dataset	Twitter	70 datasets were collected based on 10 controversial topics, each dataset has tweets of 1500 users	Private	2016	Identifying the emotions based on political conversation by using k-mean clustering	Their findings suggested that oppositional tone were associated more with negative emotion clusters, while supportive clusters overlapped more often with positive emotion clusters [42].
6	Debate Dataset	Twitter	Collection of 900,000 tweets	Private	2019	Multi-classification problem of detecting arguments between users by training the model with labels - against, favor or none	Achieved f1 score of 0.60 by using Linear SVM method for predicting any argument occurring between users [59].

Table 1 (continued)

	Dataset	Source	Size	Public Private	Year	Advantage	Additional Information
7	Online Books Dataset	Amazon	778,005 British and 1,461,206 American books data including book's ISBN, name of the authors, and literary genre metadata	Public	2019	Homophily is measured by Newman's global assortativity coefficient [81]	Strong gender homophily was found where the coefficients by gender is around 0.47 [13].
8	Asian Telecom Dataset	Asian Telephone company	300 million phone call histories of the company's approx 3.7 million customers	Private	2015	Hierarchical Bayesian model was developed with the communication information	Strong homophily effect was detected on product purchasing [66].
9	Weibo Users' Dataset	Sina Weibo website	Posts of 2000 users	Private	2015	Homophily is calculated by predicting links between the users' posts	The model could obtain better link prediction scores between the users by calculating the rate of similarity of each of the tweets [115].
10	CRM Campaigning Dataset	Twitter	100,000 posts from 75,302 unique twitter users	Private	2020	performed topic modeling on original tweets by using exponential random graph models (ERGM)	Strong homophily was detected among users using certain hashtags [113].

Table 2 Table shows the major groups (left column) and classified jobs with multiple sub-major groups (middle column) by Standard Occupation Classification. The right-most column represents the number of main users [86]

Occupational Class	Standard Occupation Classification	Users
1	Managers, Directors, Senior Officials	461
2	Professional Occ.	1,611
3	Associate Profession, Technical Occ.	926
4	Administrative Secretarial Occ.	162
5	Skilled Trades Occ.	768
6	Caring, Leisure, Other Service Occ.	259
7	Sales and Customer Service Occ.	58
8	Process, Plant, Machine Operatives	188
9	Elementary Occ.	124

6.1.3 Occupational twitter dataset

While most of the datasets were generated by using relevant keywords/hashtags, the occupation dataset was generated by using the biographical content of each of the target users. The Occupation dataset of Table 1 shows the details of the dataset. Occupational Twitter Dataset has public access and maps 5,191 Twitter users to 9 major occupational classes [92]. The dataset consists of User IDs and the historical tweets of each of the users. The Occupational prediction problem is considered a multi-classification problem as the model focuses on predicting multiple occupational classes. The occupations of each of the users were manually labelled with the aid of Standard Occupation Classification (SOC) from the UK[3] which is shown in detail in Table 2.

The dataset was initially used to predict the occupation of the main Twitter user based on their historical tweets [92]. Later, the dataset was further extended, to analyze deeper into the network information. The biographical descriptions of the following and followers' IDs for each of the main user ID [86] were added. Biographical descriptions were extracted from the 160-a character-long summary that a user writes about themselves in their profile. Thus, the extended dataset had the followers and the following information for about 4,557 for main users. Due to account suspension and protected tweets, the remaining users' accounts could not be reached. Table 2 shows the occupational class distribution of the users' occupational dataset. The biographical information of the main users was not taken into account since the main users' occupations were manually annotated in the dataset.

To construct the social network, each follower/following relationship is considered as an undirected edge for predicting the occupational classes of the main users [86]. In this social network, the Twitter users are considered as being connected via common followers and following (follow) IDs. The following IDs which only connect a few of the main IDs are considered to be weak since the flow of information between the main user IDs will be less than these follow IDs. As a result, the network was further refined by keeping only the following IDs that have more than 10 connections to the main user IDs.After performing the refining step, an unweighted graph was constructed in which all the main IDs were connected and the refined graphs consisted of 34,630 unique users including the

[3] https://www.ons.gov.uk/

4,557 main users. In the network, only 2,550 main user IDs have at least one direct connection with another main user ID. Thus, when constructing the network model the main user IDs often shared common follow IDs, which enabled the researchers to extract rich network information.

6.2 Non-twitter based datasets

Datasets for validating homophilous models were generated from other types of social media platforms as well, such as the Weibo Sina website, which is a Chinese microblogging site [102, 115]. Furthermore, synthetic networks were also developed by using an artificial undirected network. However, the artificial dataset was not thoroughly described in-depth [50]. Dataset indices 3, 7 8, and 9 from Table 1, shows the datasets generated from other sources. The following paragraphs give an overview of how these datasets were developed for analyzing the effects of homophily.

6.2.1 Asian telecom dataset

For studying the effects of homophily for product purchasing [66], purchases of Caller Ring-Back Tones (CRBT) data were provided by an Asian mobile network. This network data was mainly used for predicting consumers' product choice decisions and purchase timings. The dataset consisted of three months of detailed 300 million phone call histories of the company's approximately 3.7 million customers. The call attributes were the caller or callee phone numbers and the duration of the phone conversation. The CRBT product was bought by 750,000 customers. The pattern of the CRBT purchasing data was explored to find out the main driving forces of the customers to buy the product. The communication between friends was tracked and analyzed using this dataset. It was observed that, when friends of the customer get exposed to the ringtone by calling them, they tend to purchase the same product. Moreover, CRBT is a cheap and economical product that has been purchased by more than 750,000 customers so the researchers claimed that the dataset of phone call histories provides a convenient platform for studying communications on this product.

6.2.2 Online books dataset

Datasets from amazon.com and amazon.co.uk were used to study the effects of homophily [13]. The study was carried out on online book sales by Amazon on the English language book markets. The data was collected from over 3 million books, which included 778,005 British and 1,461,206 American books. Book's ISBN, name of the authors, and literary genre metadata were analyzed. The unique ISBN for each of the books was fetched from Amazon's website. In most of the cases, different editions with individual ISBNs were released for the same book title across the two markets (amazon.com and amazon.co.uk).

6.2.3 Weibo users dataset

To validate the proposed topic model for link prediction [115], Data was extracted from the Sina Weibo[4]. The dataset contains about 2,000 verified users, in which each user is represented by a single document. The link information between the pairs of users was also

[4] https://www.weibo.com

collected when both the users' posts were present in the dataset. The link information refers to three types of interactions which include mentioning, retweeting, and following in the Weibo website.

Generally, Twitter Search API is used to generate the desired network data for validating the proposed homophilous models. Users and their textual tweets are mainly used as features to develop the dataset. Social media platforms have other features such as images and videos which are posted by users. However, such features have not yet been included in the datasets for evaluating homophily. Furthermore, the size of the datasets varies a lot, ranging from 2,000 to 75,000 of users' posts. Yet, no benchmark has been set to have a minimum standard size of the dataset to validate any of the proposed models. Other than using Twitter Search API, microblogs and artificial data such as Weibo Sina and the artificial undirected network have also been used to generate the network data.

7 Conclusion & future directions

The homophily principle in the domain of social network analysis is an important concept that has been studied broadly. It has been used to examine the behaviour of users on social media platforms. Generally, in social media, users tend to connect with others where they have similar interests. Several studies have been carried out to study the homophilic patterns especially among majority and minority groups. These studies have deduced that majority and larger groups receive information faster than smaller, minority groups. The information reaches like-minded individuals quicker. Besides, multiple types of models have been proposed for measuring the degree of homophily. Network modelling and topic modelling have been mainly used for analyzing the strength of a relationship of a user with the user's nearest neighbours. Network modelling was conducted on network features and topic modelling was conducted on users' textual tweets respectively. Recently, the effect of homophily has also been studied by combining textual and network features with a highly efficient neural network model. However, the content of the interactions occurring between users in social media is still inadequately understood [111]. Furthermore, the content of social media ranges from texts to videos, which needs different types of analysis. Most of the studies focused either on textual posts, hashtags tweeted by users, mentions of users, or users' network connections. However, there is a research gap as comprehensive studies have not been conducted, concerning images and videos posted by users. Whether these contents of social media have any effect on the degree of homophily in online platforms is not fully understood as of yet. Therefore, state-of-art methods should also focus on measuring the level of homophily by using these features.

In this paper, we presented a survey on the usage of the Homophily principle for computing social network analysis. This thorough survey will enable researchers to explore new methods to measure the degree of homophily. In summary, multiple methodologies have been proposed over the years to measure the level of homophily by using different types of modelling approaches. This includes topic modelling, network modelling, ERGMs, where each of the models has its own merits and drawbacks. These models are validated by either using synthetic data or by using real-life data from social networking sites such as Twitter. However, the range of data used by each of the proposed models varied to a great extent ranging from posts from only 2,000 users to 75,000 users' posts. Therefore, for better comparability of different features and methods, we argue for a benchmark dataset for homophily detection.

Appendix A: Table of References

Table 3 The information of all the references selected for the survey

Title	Venue	Citations	Quartile	H-index	Year
Friendship as a social process: A substantive and methodological analysis [60]	Freedom and control in modern society	3069	–	–	1954
Birds of a feather: Homophily in social networks [70]	Annual review of sociology	15216	–	151	2001
Why does everybody hate me? balance,status, and homophily: The triumvirate of signed tie formation [116]	Social Networks	43	Q1	85	2015
How social ties transcend class boundaries ? Network variability as tool for exploring occupational homophily [15]	Social Networks	-	Q1	85	2020
Quantifying segregation in an integrated urban physical-social space [114]	Journal of the Royal Society Interface	-	Q1	114	2019
Trustworthy health-related tweets on social media in Saudi Arabia: tweet metadata analysis [1]	Journal of medical Internet research	1	Q1	116	2019
For better or for worse? A systematic review of the evidence on social media use and depression among lesbian, gay, and bisexual minorities [29]	Journal of medical Internet research	6	Q1	116	2019

Table 3 (continued)

Title	Venue	Citations	Quartile	H-index	Year
Social media and human need satisfaction: Implications for social media marketing [125]	Business Horizons	194	Q1	67	2015
Will they come and will they stay? Online social networks and news consumption on external websites [68]	Business Horizons	194	Q1	67	2015
Political homophily in social relationships: Evidence from online dating behavior [44]	The Journal of Politics	100	-	50	2017
Homophily of music listening in online social networks of China [122]	Social Networks	5	Q1	85	2018
Latent homophily or social influence? An empirical analysis of purchase within a social network [66]	Management Science	85	Q1	221	2016
Twitter Homophily: Network Based Prediction of User's Occupation [86]	Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics	-	-	51	2019

Table 3 (continued)

Title	Venue	Citations	Quartile	H-index	Year
Homophily, group size, and the diffusion of political information in social networks: Evidence from Twitter [39]	Journal of public economics	178	Q1	124	2016
Analyze users' online shopping behavior using interconnected online interest-product network [40]	WCNC	1	-	80	2018
Race, school integration, and friendship segregation in America [75]	American journal of Sociology	1330	Q1	160	2001
Shared contexts, shared background, shared values–Homophily in Finnish parliament members' social networks on Twitter [56]	Telematics & Informatics	3	Q1	52	2019
Building the community: Endogenous network formation, homophily and pro social sorting among therapeutic community residents [110]	Drug and Alcohol Dependence	-	Q1	151	2020

Table 3 (continued)

Title	Venue	Citations	Quartile	H-index	Year
Purity homophily in social networks [25]	Journal of Experimental Psychology: General	83	Q1	138	2016
Structural transition in social networks: The role of homophily [79]	Scientific reports	1	Q1	149	2019
Information diffusion and opinion change during the gezi park protests: Homophily or social influence? [27]	Database: The Journal of biographical logical Databases and Curation	88	-	65	2016
Trip distribution modeling with Twitter data [91]	Computers, Environment and Urban Systems	2	Q1	74	2019
Good Games, bad host? Using big data to measure public attention and imagery of the Olympic Games [51]	Cities	5	Q1	74	2019
Using Facebook for Qualitative Research: A Brief Primer [32]	Journal of medical Internet research	-	Q1	116	2019
Sensitivity analysis for contagion effects in social networks [109]	Sociological Methods & Researchs	124	Q1	65	2011
Birds of a schedule flock together: Social networks, peer influence, and digital activity cycles [62]	Computers in Human Behavior	3	Q1	137	2018

Table 3 (continued)

Title	Venue	Citations	Quartile	H-index	Year
Social media engagement: What motivates user participation and consumption on YouTube? [53]	Computers in Human Behavior	254	Q1	137	2017
Social media and web presence for patients and professionals: evolving trends and implications for practice [6]	PM&R	31	-	53	2017
The media inequality: Comparing the initial human-human and human-AI social interactions [77]	Computers in Human Behavior	59	Q1	137	2017
Sentiment analysis on Twitter: A text mining approach to the Syrian refugee crisis [80]	Telematics and Informatics	80	Q1	52	2018
Network analysis reveals open forums and echo chambers in social media discussions of climate change [111]	Global environmental change	201	Q1	147	2015
The ambivalence of cultural homophily: Field positions, semantic similarities, and social network ties in creative collectives [7]	Poetics	4	Q1	54	2019

Table 3 (continued)

Title	Venue	Citations	Quartile	H-index	Year
A survey on deep learning in medical image analysis	Medical image analysis [64]	2991	Q1	113	2017
Deep learning in agriculture: A survey [49]	Computers and electronics in agriculture	434	Q1	96	2018
Graph embedding techniques, applications, and performance: A survey [37]	Knowledge-Based Systems	521	Q1	94	2018
Deep learning based recommender system: A survey and new perspectives [121]	ACM Computing Surveys	437	Q1	132	2019
Homophily, structure, and content augmented network representation learning [119]	2016 IEEE 16th international conference on data mining (ICDM)	55	Q1	100	2016
The social structure of political echo chambers: Variation in ideological homophily in online networks [12]	Political Psychology	150	Q1	80	2019
Equality of opportunity in classification: A causal approach [120]	Advances in Neural Information Processing Systems	13	-	54	2018
Yes, there is a correlation: -from social networks to personal behavior on the web [100]	Proceedings of the 17th international conference on World Wide Web	344	-	64	2008

Table 3 (continued)

Title	Venue	Citations	Quartile	H-index	Year
The old boy (and girl) network: Social network formation on university campuses [69]	Journal of public economics	443	Q1	123	2008
YouTube vloggers' popularity and influence: The roles of homophily, emotional attachment, and expertise [58]	Journal of Retailing and Consumer Services	-	Q1	65	2020
Dynamic social balance and convergent appraisals via homophily and influence mechanisms [71]	Automatica	1	Q1	239	2019
Dominant frames in legacy and social media coverage of the IPCC Fifth Assessment Report [85]	Nature Climate Change	151	Q1	136	2015
Twitter users change word usage according to conversation-partner social identity [106]	Social Networks	47	Q1	85	2015
Stance polarity in political debates: A diachronic perspective of network homophily and conversations on Twitter [59]	Data & Knowledge Engineering	1	Q2	79	2019
Valence-based homophily on Twitter: Network analysis of emotions and political talk in the 2012 presidential election [42]	New media & society	63	Q1	87	2016
Hashtag homophily in twitter network: Examining a controversial cause-related marketing campaign [113]	Computers in Human Behavior	-	Q1	137	2020

Table 3 (continued)

Title	Venue	Citations	Quartile	H-index	Year
Polarized frames on "climate change" and "global warming" across countries and states: Evidence from Twitter big data [46]	Global Environmental Change	132	Q1	147	2015
Homophily influences ranking of minorities in social networks [50]	Scientific reports	20	Q1	149	2018
Status seeking and perceived similarity: a consideration of homophily in the social servicescape [41]	International Journal of Hospitality Management	29	Q1	93	2017
Emergence of scaling in random networks [4]	Science	36019	Q1	1058	1999
Spectral centrality measures in complex networks [89]	Physical Review E	148	Q1	190	2008
The influence of cause-related marketing on consumer choice: does one good turn deserve another? [5]	Journal of the academy of marketing Science	1383	Q1	148	2000

Table 3 (continued)

Title	Venue	Citations	Quartile	H-index	Year
Latent semantic indexing: A probabilistic analysis [87]	Journal of Computer and System Sciences	1280	Q2	81	2000
Latent dirichlet allocation [10]	Journal of machine Learning research	31189	Q1	173	2003
Exponential random graph model parameter estimation for very large directed networks [105]	PloS one	4	Q1	268	2020
Measuring the impact of spammers on e-mail and Twitter networks [21]	International Journal of Information Management	10	Q1	91	2019
Friend or frenemy? Experiential homophily and educational track attrition among premedical students [38]	Social Science & Medicine	1	Q1	213	2018
A belief-based theory of homophily [52]	Games and Economic Behaviory	5	Q1	84	2019
Tweeting for social justice in# Ferguson: Affective discourse in Twitter hashtags [11]	new media & society	3	Q1	87	2019
An adaptive temporal-causal network model for social networks based on the homophily and more-becomes-more principle [108]	Neurocomputing	7	Q1	110	2019
Semi-supervised classification with graph convolutional networks [55]	arXiv	3196	-	-	2016

Table 3 (continued)

Title	Venue	Citations	Quartile	H-index	Year
An analysis of the user occupational class through Twitter content [92]	Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)	134	-	51	2015
A simple model of homophily in social networks [23]	European Economic Review	66	Q1	116	2016
Birds of a feather linked together: A discriminative topic model using link-based priors [115]	Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing	5	-	88	2015
Effect of homophily on network formations [54]	Communications in Nonlinear Science and Numerical Simulation	21	Q1	96	2017
MedLDA: maximum margin supervised topic models [123]	Journal of Machine Learning Research	443	Q1	173	2012
Topic models conditioned on arbitrary features with dirichlet-multinomial regression [73]	arXiv	389	-	-	2012

Table 3 (continued)

Title	Venue	Citations	Quartile	H-index	Year
Gibbs max-margin topic models with data augmentation [124]	The Journal of Machine Learning Research	75	Q1	173	2014
Lexical and hierarchical topic regression [83]	Advances in neural information processing systems	61	-	54	2013
Distributed representations of words and phrases and their compositionality [72]	Advances in neural information processing systems	18726	-	54	2013
A density-based method for adaptive LDA model selection [14]	Neurocomputing	276	Q1	110	2009
Reading tea leaves: How humans interpret topic models [17]	Advances in neural information processing systems	1668	-	54	2009
Optimizing semantic coherence in topic models [74]	Proceedings of the conference on empirical methods in natural language processing	918	-	88	2011
The effect of calorie posting regulation on consumer opinion: A flexible latent Dirichlet allocation model with informative priors [93]	Marketing Science	34	Q1	113	2017
Specification of exponential-family random graph models: terms and computational aspects [76]	Journal of statistical software	302	Q1	115	2008

Table 3 (continued)

Title	Venue	Citations	Quartile	H-index	Year
Exponential random graph models for social networks: Theory, methods, and applications [65]	Cambridge University Press	705	-	-	2013
Opening the black box of link formation: Social factors underlying the structure of the web [36]	Social Networks	77	Q1	85	2009
An introduction to exponential random graph (p*) models for social networks [95]	Social Networks	1677	Q1	85	2007
Assortative mixing in networks [82]	Physical review letters	4986	Q1	602	2002
Gender homophily in online book networks [13]	Information sciences	4	Q1	169	2019
Clustering and preferential attachment in growing networks [81]	Physical review E	1822	Q1	76	2001
Multiscale mixing patterns in networks [88]	Proceedings of the National Academy of Sciences	32	Q1	737	2018
Assortativity of suicide-related posting on social media. [16]	American Psychologist	3	Q1	219	2020
Reconsidering power in multi stakeholder relationship management [96]	Management Communication Quarterly	12	Q1	55	2018

Table 3 (continued)

Title	Venue	Citations	Quartile	H-index	Year
Why the Bass model fits without decision variables [8]	Marketing Science	1044	Q1	113	1994
Tweetmotif: Exploratory search and topic summarization for twitter [84]	Fourth International AAAI Conference on Weblogs and Social Media, 2010	411	-	60	2010
A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons [103]	Journal of Machine Learning Research	2871	-	-	1948
On finding the natural number of topics with latent dirichlet allocation: Some observations [2]	Pacific-Asia Conference on knowledge discovery and data mining	470	-	-	2010
Social cognitive theory of mass communication [3]	Media Effects	5922	-	-	2009
Stability of topic modeling via matrix factorization [9]	Expert Systems with Applications	66	Q1	207	2018
Cross-lingual language model pretraining [22]	Advances in Neural Information Processing Systems	635	-	284	2019
Metal-phenolic networks as a versatile platform to engineer nanomaterials and biointerfaces [28]	Nano Today	230	Q1	143	2017

Table 3 (continued)

Title	Venue	Citations	Quartile	H-index	Year
Viewer aggression and homophily, identification, and parasocial relationships with television characters [30]	Journal of Broadcasting & Electronic Media	444	Q1	68	2003
Bones, body parts, and sex appeal: An analysis of #thinspiration images on popular social media [34]	Body Image	223	Q1	83	2015
Predicting tie strength with social media [35]	Proceedings of the SIGCHI conference on human factors in computing systems	1826	-	-	2009
The hadamard product [43]	Proc. Symp. Appl. Math	358	-	-	1990
Development of word cloud generator software based on python [48]	Procedia engineering	30	-	74	2017
Platform structures, homing preferences, and homophilous propensities in online social networks [57]	Journal of Management Information Systems	20	Q1	144	2017
5G Internet of Things: A survey [61]	Journal of Industrial Information Integration	757	Q1	24	2018
Some methods for classification and analysis of multivariate observations [67]	Proceedings of the fifth Berkeley symposium on mathematical statistics and probability	31175	-	-	1967
Brand activism: Does courting controversy help or hurt a brand? [78]	International Journal of Research in Marketing	13	Q1	102	2020

Table 3 (continued)

Title	Venue	Citations	Quartile	H-index	Year
Approaches to recruiting 'hard-to-reach'populations into research: a review of the literature [99]	Health Promotion Perspectives	330	-	-	2011
Markov chain Monte Carlo estimation of exponential random graph models [101]	Journal of Social Structure	903	-	-	2002
Probabilistic topic models [104]	Handbook of latent semantic analysis	2089	-	-	2007
Exploiting homophily effect for trust prediction [107]	Proceedings of the sixth ACM international conference on Web search and data mining	266	-	-	2018
Analysis of political discourse on twitter in the context of the 2016 US presidential elections [117]	Government Information Quarterly	109	Q1	103	2017
Gratifications of using Facebook, Twitter, Instagram, or Snapchat to follow brands: The moderating effect of social comparison, trust, tie strength, and network homophily on brand identification, brand engagement, brand commitment, and membership intention [90]	Telematics and Informatics	25	Q1	66	2017
Semantic homophily in online communication: evidence from Twitter [98]	Online Social Networks and Media	20	Q1	-	2017

Table 3 (continued)

Title	Venue	Citations	Quartile	H-index	Year
JUUL: spreading online and offline [19]	Handbook of latent semantic analysis	59	Q1	161	2018
Public diplomacy networks: China's public diplomacy communication practices in twitter during Two Sessions [47]	Public Relations Review	7	Q1	82	2020
A network analysis of official Twitter accounts during the West Virginia water crisis [33]	Computers in Human Behavior	42	Q1	178	2016
The effects of offline events on online connective actions: an examination of# BoycottNFL using social network analysis [20]	Computers in Human Behavior	5	Q1	178	2021
Strategies to find audience segments on Twitter for e-cigarette education campaigns [18]	Addictive Behaviors	10	Q1	127	2019
Strategies to find audience segments on Twitter for e-cigarette education campaigns [112]	Information Processing and Management	2	Q1	101	2020

Table 3 (continued)

Title	Venue	Citations	Quartile	H-index	Year
"THE RUSSIANS ARE HACKING MY BRAIN!" investigating Russia's internet research agency twitter tactics during the 2016 United States presidential campaign [33]	Computers in Human Behavior	37	Q1	178	2016
Gender, rank, and social networks on an enterprise social media platform [26]	Social Networks	21	Q1	98	2020
Modeling interurban mentioning relationships in the US Twitter network using geo-hashtags [24]	Computers, Environment and Urban Systems	-	Q1	92	2021
Can we vote with our tweet? On the perennial difficulty of election forecasting with social media [45]	International Journal of Forecasting	79	Q1	96	2015
Not all emotions are created equal: Expressive behavior of the networked public on China's social media site [102]	Computers in Human Behavior	40	Q1	178	2016
Structural diversity effect on hashtag adoption in Twitter [118]	Physica A: Statistical Mechanics and its Applications	16	Q2	166	2018

Acknowledgments We would like to thank the reviewers for their helpful comments on our work. This work is supported by the Natural Sciences and Engineering Research Council of Canada (NSERC).

References

1. Albalawi Y, Nikolov NS, Buckley J (2019) Trustworthy health-related tweets on social media in saudi arabia: tweet metadata analysis. Journal of medical Internet research 21(10):e14731
2. Arun R, Suresh V, Madhavan CV, Murthy MN (2010) On finding the natural number of topics with latent dirichlet allocation: Some observations. In: Pacific-asia conference on knowledge discovery and data mining, pp. 391–402. Springer
3. Bandura A (2009) Social cognitive theory of mass communication. In: Media effects, pp. 110–140. Routledge
4. Barabási A. L., Albert R (1999) Emergence of scaling in random networks. Science 286(5439):509–512
5. Barone MJ, Miyazaki AD, Taylor KA (2000) The influence of cause-related marketing on consumer choice: does one good turn deserve another? Journal of the academy of marketing Science 28(2):248–262
6. Barreto JE, Whitehair CL (2017) Social media and web presence for patients and professionals: evolving trends and implications for practice. PM&R 9(5):S98–S105
7. Basov N (2019) The ambivalence of cultural homophily: Field positions, semantic similarities, and social network ties in creative collectives Poetics
8. Bass FM, Krishnan TV, Jain DC (1994) Why the bass model fits without decision variables. Marketing science 13(3):203–223
9. Belford M, Mac Namee B, Greene D (2018) Stability of topic modeling via matrix factorization. Expert Syst Appl 91:159–169
10. Blei DM, Ng AY, Jordan MI (2003) Latent dirichlet allocation. Journal of machine Learning research 3(Jan):993–1022
11. Blevins JL, Lee JJ, McCabe EE, Edgerton E (2019) Tweeting for social justice in# ferguson:, Affective discourse in twitter hashtags. New Media & Society 21(7):1636–1653
12. Boutyline A, Willer R (2017) The social structure of political echo chambers: Variation in ideological homophily in online networks. Political Psychology 38(3):551–569
13. Bucur D (2019) Gender homophily in online book networks. Information sciences 481:229–243
14. Cao J, Xia T, Li J, Zhang Y, Tang S (2009) A density-based method for adaptive lda model selection. Neurocomputing 72(7-9):1775–1781
15. Cepić D., Tonković Ž (2020) How social ties transcend class boundaries? network variability as tool for exploring occupational homophily. Soc Networks 62:33–42
16. Cero I, Witte TK (2020) Assortativity of suicide-related posting on social media. Am Psychol 75(3):365
17. Chang J, Gerrish S, Wang C, Boyd-Graber JL, Blei DM (2009) Reading tea leaves: How humans interpret topic models. In: Advances in neural information processing systems, pp 288–296
18. Chu KH, Allem JP, Unger JB, Cruz TB, Akbarpour M, Kirkpatrick MG (2019) Strategies to find audience segments on twitter for e-cigarette education campaigns. Addictive behaviors 91:222–226
19. Chu KH, Colditz JB, Primack BA, Shensa A, Allem JP, Miller E, Unger JB, Cruz TB (2018) Juul: spreading online and offline. J Adolesc Health 63(5):582–586
20. Chung TLD, Johnson O, Hall-Phillips A, Kim K (2021) The effects of offline events on online connective actions: an examination of# boycottnfl using social network analysis. Comput Hum Behav 115:106623
21. Colladon AF, Gloor PA (2019) Measuring the impact of spammers on e-mail and twitter networks. Int J Inf Manag 48:254–262
22. Conneau A, Lample G (2019) Cross-lingual language model pretraining. In: Advances in neural information processing systems, pp. 7057–7067
23. Currarini S, Matheson J, Vega-Redondo F (2016) A simple model of homophily in social networks. Eur Econ Rev 90:18–39
24. Cvetojevic S, Hochmair HH (2021) Modeling interurban mentioning relationships in the us twitter network using geo-hashtags. Comput Environ Urban Syst 87:101621
25. Dehghani M, Johnson K, Hoover J, Sagi E, Garten J, Parmar NJ, Vaisey S, Iliev R, Graham J (2016) Purity homophily in social networks. J Exp Psychol Gen 145(3):366
26. Di Tommaso G, Gatti M, Iannotta M, Mehra A, Stilo G, Velardi P (2020) Gender, rank, and social networks on an enterprise social media platform. Soc Networks 62:58–67

27. Dincelli E, Hong Y, DePaula N (2016) Information diffusion and opinion change during the gezi park protests: Homophily or social influence? Proceedings of the Association for Information Science and Technology 53(1):1–5
28. Ejima H, Richardson JJ, Caruso F (2017) Metal-phenolic networks as a versatile platform to engineer nanomaterials and biointerfaces. Nano Today 12:136–148
29. Escobar-Viera CG, Whitfield DL, Wessel CB, Shensa A, Sidani JE, Brown AL, Chandler CJ, Hoffman BL, Marshal MP, Primack BA (2018) For better or for worse? a systematic review of the evidence on social media use and depression among lesbian, gay, and bisexual minorities. JMIR mental health 5(3):e10496
30. Eyal K, Rubin AM (2003) Viewer aggression and homophily, identification, and parasocial relationships with television characters. Journal of Broadcasting & Electronic Media 47(1):77–98
31. Fincham K (2019) Exploring political journalism homophily on twitter: a comparative analysis of us and uk elections in 2016 and 2017. Media and Communication 7(1):213–224
32. Franz D, Marsh HE, Chen JI, Teo AR (2019) Using facebook for qualitative research: a brief primer. Journal of medical Internet research 21(8):e13544
33. Getchell MC, Sellnow TL (2016) A network analysis of official twitter accounts during the west virginia water crisis. Comput Hum Behav 54:597–606
34. Ghaznavi J, Taylor LD (2015) Bones, body parts, and sex appeal: an analysis of# thinspiration images on popular social media. Body image 14:54–61
35. Gilbert E, Karahalios K (2009) Predicting tie strength with social media. In: Proceedings of the SIGCHI conference on human factors in computing systems, pp. 211–220
36. Gonzalez-Bailon S (2009) Opening the black box of link formation: Social factors underlying the structure of the web. Soc Networks 31(4):271–280
37. Goyal P, Ferrara E (2018) Graph embedding techniques, applications, and performance: a survey. Knowl-Based Syst 151:78–94
38. Grace MK (2018) Friend or frenemy? experiential homophily and educational track attrition among premedical students. Social Science & Medicine 212:33–42
39. Halberstam Y, Knight B (2016) Homophily, group size, and the diffusion of political information in social networks: Evidence from twitter. Journal of public economics 143:73–88
40. Han S, Qiao Y, Zhang Y, Lin W, Yang J (2018) Analyze users' online shopping behavior using interconnected online interest-product network. In: 2018 IEEE Wireless communications and networking conference (WCNC), pp. 1–6. IEEE
41. Hanks L, Line N, Yang W (2017) Status seeking and perceived similarity: a consideration of homophily in the social servicescape. Int J Hosp Manag 60:123–132
42. Himelboim I, Sweetser KD, Tinkham SF, Cameron K, Danelo M, West K (2016) Valence-based homophily on twitter: Network analysis of emotions and political talk in the 2012 presidential election. New media & society 18(7):1382–1400
43. Horn RA (1990) The hadamard product. In: Proc. Symp. Appl. math, vol. 40, pp. 87–169
44. Huber GA, Malhotra N (2017) Political homophily in social relationships: Evidence from online dating behavior. The Journal of Politics 79(1):269–283
45. Huberty M (2015) Can we vote with our tweet? on the perennial difficulty of election forecasting with social media. Int J Forecast 31(3):992–1007
46. Jang SM, Hart PS (2015) Polarized frames on "climate change" and "global warming" across countries and states: Evidence from twitter big data. Glob Environ Chang 32:11–17
47. Jia R, Li W (2020) Public diplomacy networks: China's public diplomacy communication practices in twitter during two sessions. Public Relations Review 46(1):101818
48. Jin Y (2017) Development of word cloud generator software based on python. Procedia engineering 174:788–792
49. Kamilaris A, Prenafeta-Boldú FX (2018) Deep learning in agriculture: a survey. Computers and electronics in agriculture 147:70–90
50. Karimi F, Génois M, Wagner C, Singer P, Strohmaier M (2018) Homophily influences ranking of minorities in social networks. Scientific reports 8(1):1–12
51. Kassens-Noor E, Vertalka J, Wilson M (2019) Good games, bad host? using big data to measure public attention and imagery of the olympic games. Cities 90:229–236
52. Kets W, Sandroni A (2019) A belief-based theory of homophily. Games and Economic Behavior 115:410–435
53. Khan ML (2017) Social media engagement: What motivates user participation and consumption on youtube? Comput Hum Behav 66:236–247
54. Kim K, Altmann J (2017) Effect of homophily on network formation. Commun Nonlinear Sci Numer Simul 44:482–494

55. Kipf TN, Welling M (2016) Semi-supervised classification with graph convolutional networks. arXiv:1609.02907
56. Koiranen I, Koivula A, Keipi T, Saarinen A (2019) Shared contexts, shared background, shared values–homophily in finnish parliament members' social networks on twitter. Telematics Inform 36:117–131
57. Kwon HE, Oh W, Kim T (2017) Platform structures, homing preferences, and homophilous propensities in online social networks. J Manag Inf Syst 34(3):768–802
58. Ladhari R, Massa E, Skandrani H (2020) Youtube vloggers' popularity and influence: the roles of homophily, emotional attachment, and expertise. J Retail Consum Serv 54:102027
59. Lai M, Tambuscio M, Patti V, Ruffo G, Rosso P (2019) Stance polarity in political debates: a diachronic perspective of network homophily and conversations on twitter. Data & Knowledge Engineering 124:101738
60. Lazarsfeld PF, Merton RK et al (1954) Friendship as a social process: a substantive and methodological analysis. Freedom and control in modern society 18(1):18–66
61. Li S, Da Xu L, Zhao S (2018) 5g internet of things: a survey. Journal of Industrial Information Integration 10:1–9
62. Liang H, Shen F (2018) Birds of a schedule flock together: Social networks, peer influence, and digital activity cycles. Comput Hum Behav 82:167–176
63. Linvill DL, Boatwright BC, Grant WJ, Warren PL (2019) "the russians are hacking my brain!" investigating russia's internet research agency twitter tactics during the 2016 United States presidential campaign. Comput Hum Behav 99:292–300
64. Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, Van Der Laak JA, Van Ginneken B, Sánchez C. I. (2017) A survey on deep learning in medical image analysis. Medical image analysis 42:60–88
65. Lusher D, Koskinen J, Robins G (2013) Exponential random graph models for social networks: Theory, methods, and applications Cambridge University Press
66. Ma L, Krishnan R, Montgomery AL (2015) Latent homophily or social influence? an empirical analysis of purchase within a social network. Manag Sci 61(2):454–473
67. MacQueen J et al (1967) Some methods for classification and analysis of multivariate observations. In: Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, vol. 1, pp 281–297. Oakland, CA, USA
68. Mahmood A, Sismeiro C (2017) Will they come and will they stay? online social networks and news consumption on external websites. J Interact Mark 37:117–132
69. Mayer A, Puller SL (2008) The old boy (and girl) network: Social network formation on university campuses. Journal of public economics 92(1-2):329–347
70. McPherson M, Smith-Lovin L, Cook JM (2001) Birds of a feather: Homophily in social networks. Annual review of sociology 27(1):415–444
71. Mei W, Cisneros-Velarde P, Chen G, Friedkin NE, Bullo F (2019) Dynamic social balance and convergent appraisals via homophily and influence mechanisms. Automatica 110:108580
72. Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J (2013) Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems, pp. 3111–3119
73. Mimno D, McCallum A (2012) Topic models conditioned on arbitrary features with dirichlet-multinomial regression. arXiv:1206.3278
74. Mimno D, Wallach HM, Talley E, Leenders M, McCallum A (2011) Optimizing semantic coherence in topic models. In: Proceedings of the conference on empirical methods in natural language processing, pp. 262–272. Association for Computational Linguistics
75. Moody J (2001) Race, school integration, and friendship segregation in america. American journal of Sociology 107(3):679–716
76. Morris M, Handcock MS, Hunter DR (2008) Specification of exponential-family random graph models: terms and computational aspects. Journal of statistical software 24(4):1548
77. Mou Y, Xu K (2017) The media inequality: Comparing the initial human-human and human-ai social interactions. Comput Hum Behav 72:432–440
78. Mukherjee S, Althuizen N (2020) Brand activism: Does courting controversy help or hurt a brand? International Journal of Research in Marketing
79. Murase Y, Jo HH, Török J, Kertész J, Kaski K (2019) Structural transition in social networks: the role of homophily. Scientific reports 9(1):1–8
80. Nazan Ö, Ayvaz S (2018) Sentiment analysis on twitter: a text mining approach to the syrian refugee crisis. Telematics Inform 314:136–147
81. Newman ME (2001) Clustering and preferential attachment in growing networks. Physical review E 64(2):025102

82. Newman ME (2002) Assortative mixing in networks. Physical review letters 89(20):208701
83. Nguyen VA, Ying JL, Resnik P (2019) Lexical and hierarchical topic regression. In: Advances in neural information processing systems, pp. 1106–1114
84. O'Connor B, Krieger M, Ahn D (2010) Tweetmotif: Exploratory search and topic summarization for twitter. In: Fourth international AAAI conference on weblogs and social media
85. O'Neill S, Williams HT, Kurz T, Wiersma B, Boykoff M (2015) Dominant frames in legacy and social media coverage of the ipcc fifth assessment report. Nat Clim Chang 5(4):380–385
86. Pan J, Bhardwaj R, Lu W, Chieu HL, Pan X, Puay NY (2019) Twitter homophily: Network based prediction of user's occupation. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 2633–2638
87. Papadimitriou CH, Raghavan P, Tamaki H, Vempala S (2000) Latent semantic indexing: a probabilistic analysis. J Comput Syst Sci 61(2):217–235
88. Peel L, Delvenne JC, Lambiotte R (2018) Multiscale mixing patterns in networks. Proceedings of the National Academy of Sciences 115(16):4057–4062
89. Perra N, Fortunato S (2008) Spectral centrality measures in complex networks. Physical Review E 78(3):036107
90. Phua J, Jin SV, Kim JJ (2017) Gratifications of using facebook, twitter, instagram, or snapchat to follow brands: the moderating effect of social comparison, trust, tie strength, and network homophily on brand identification, brand engagement, brand commitment, and membership intention. Telematics Inform 34(1):412–424
91. Pourebrahim N, Sultana S, Niakanlahiji A, Thill JC (2019) Trip distribution modeling with twitter data. Comput Environ Urban Syst 77:101354
92. Preoţiuc-Pietro D., Lampos V, Aletras N (2015) An analysis of the user occupational class through twitter content. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pp. 1754–1764
93. Puranam D, Narayan V, Kadiyali V (2017) The effect of calorie posting regulation on consumer opinion: a flexible latent dirichlet allocation model with informative priors. Mark Sci 36(5):726–746
94. Qudar MMA, Mago V (2020) Tweetbert:, A pretrained language representation model for twitter text analysis. arXiv:2010.11091
95. Robins G, Pattison P, Kalish Y, Lusher D (2007) An introduction to exponential random graph (p*) models for social networks. Social networks 29(2):173–191
96. Saffer AJ, Yang A, Taylor M (2018) Reconsidering power in multistakeholder relationship management. Manag Commun Q 32(1):121–139
97. Sandhu M, Vinson CD, Mago VK, Giabbanelli PJ (2019) From associations to sarcasm: Mining the shift of opinions regarding the supreme court on twitter. Online Social Networks and Media 14:100054
98. Šćepanović S, Mishkovski I, Gonçalves B, Nguyen TH, Hui P (2017) Semantic homophily in online communication: evidence from twitter. Online Social Networks and Media 2:1–18
99. Shaghaghi A, Bhopal RS, Sheikh A (2011) Approaches to recruiting 'hard-to-reach'populations into research: a review of the literature. Health promotion perspectives 1(2):86
100. Singla P, Richardson M (2008) Yes, there is a correlation: -from social networks to personal behavior on the web. In: Proceedings of the 17th international conference on World Wide Web, pp. 655–664
101. Snijders TA (2002) Markov chain monte carlo estimation of exponential random graph models. Journal of Social Structure 3(2):1–40
102. Song Y, Dai XY, Wang J (2016) Not all emotions are created equal: Expressive behavior of the networked public on china's social media site. Comput Hum Behav 60:525–533
103. Sørensen T, Sørensen T, Sørensen T, SORENSEN T, Sorensen T, Sorensen T, Biering-sørensen T (1948) A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on danish commons
104. Steyvers M, Griffiths T (2007) Probabilistic topic models. Handbook of latent semantic analysis 427(7):424–440
105. Stivala A, Robins G, Lomi A (2020) Exponential random graph model parameter estimation for very large directed networks. Plos one 15(1) e0227804
106. Tamburrini N, Cinnirella M, Jansen VA, Bryden J (2015) Twitter users change word usage according to conversation-partner social identity. Soc Networks 40:84–89
107. Tang J, Gao H, Hu X, Liu H (2013) Exploiting homophily effect for trust prediction. In: Proceedings of the sixth ACM international conference on Web search and data mining, pp. 53–62
108. van den Beukel S, Goos SH, Treur J (2019) An adaptive temporal-causal network model for social networks based on the homophily and more-becomes-more principle. Neurocomputing 338:361–371

109. VanderWeele TJ (2017) Sensitivity analysis for contagion effects in social networks. Sociological Methods & Research 54(13):3058–3070
110. Warren K, Campbell B, Cranmer S, De Leon G, Doogan N, Weiler M, Doherty F (2020) Building the community: Endogenous network formation, homophily and prosocial sorting among therapeutic community residents. Drug Alcohol Depend 207:107773
111. Williams Hywel TPEA (2015) Network analysis reveals open forums and echo chambers in social media discussions of climate change. Global environmental change 32:126–138
112. Xiong J, Feng X, Tang Z (2020) Understanding user-to-user interaction on government microblogs: An exponential random graph model with the homophily and emotional effect. Information Processing & Management 57(4):102229
113. Xu S, Zhou A (2020) Hashtag homophily in twitter network: Examining a controversial cause-related marketing campaign. Comput Hum Behav 102:87–96
114. Xu Y, Belyi A, Santi P, Ratti C (2019) Quantifying segregation in an integrated urban physical-social space. Journal of the Royal Society Interface 16(160):20190536
115. Yang W, Boyd-Graber J, Resnik P (2015) Birds of a feather linked together: a discriminative topic model using link-based priors. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pp. 261–266
116. Yap J, Harrigan N (2015) Why does everybody hate me? balance, status, and homophily: the triumvirate of signed tie formation. Soc Networks 40:103–122
117. Yaqub U, Chun SA, Atluri V, Vaidya J (2017) Analysis of political discourse on twitter in the context of the 2016 us presidential elections. Gov Inf Q 34(4):613–626
118. Zhang A, Zheng M, Pang B (2018) Structural diversity effect on hashtag adoption in twitter. Physica A:, Statistical Mechanics and its Applications 493:267–275
119. Zhang D, Yin J, Zhu X, Zhang C (2016) Homophily, structure, and content augmented network representation learning. In: 2016 IEEE 16Th international conference on data mining (ICDM), pp. 609–618. IEEE
120. Zhang J, Bareinboim E (2018) Equality of opportunity in classification: a causal approach. In: Advances in neural information processing systems, pp. 3671–3681
121. Zhang S, Yao L, Sun A, Tay Y (2019) Deep learning based recommender system: a survey and new perspectives. ACM Computing Surveys (CSUR) 52(1):1–38
122. Zhou Z, Xu K, Zhao J (2018) Homophily of music listening in online social networks of china. Soc Networks 55:160–169
123. Zhu J, Ahmed A, Xing EP (2012) Medlda: maximum margin supervised topic models. J Mach Learn Res 13(Aug):2237–2278
124. Zhu J, Chen N, Perkins H, Zhang B (2014) Gibbs max-margin topic models with data augmentation. The Journal of Machine Learning Research 15(1):1073–1110
125. Zhu YQ, Chen HG (2015) Social media and human need satisfaction: Implications for social media marketing. Business horizons 58(3):335–345