Journal Title
XX(X):1–16
©The Author(s) 2020
Reprints and permission:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/ToBeAssigned
www.sagepub.com/
SAGE

Influence and performance of user similarity metrics in followee prediction

Antonela Tommasel[1] and Daniela Godoy[1]

Abstract

Followee recommendation is a problem rapidly gaining importance in Twitter as well as in other micro-blogging communities. Hence, understanding how users select whom to follow becomes crucial for designing accurate and personalised recommendation strategies. This work aims at shedding some light on how homophily drives the formation of user relationships by studying the influence of diverse recommendation factors on tie formation. The selected recommendation factors were studied considering multiple alternatives for assessing them in terms of user similarity. A data analysis comparing the similarity amongst Twitter users and their followees, regarding two commonly-used followee recommendation factors (topology and content) was performed in the context of a followee recommendation task. This study is amongst the firsts to analyse the effect of the different criteria for followee recommendation in micro-blogging communities, and the importance of thoroughly analysing the different aspects of user relationships to define the concept of user similarity. The study showed how the choice of the different factors and assessment alternatives affects followee recommendation. It also verified the existence of certain patterns regarding friends and random users' similarities, which can condition the adequacy of the available similarity metrics.

Keywords

Followee prediction, micro-blogging communities, similarity metrics

Introduction

The rapid growth and exponential usage of social digital media increased the popularity of micro-blogging platforms, characterised by linked social entities, which have become an important part of the daily life of millions of users around the world. A representative example is *Twitter*, in which social entities are subscribed users and links between them are following relationships, not necessarily reciprocal. Generally, these relationships are driven by the phenomenon of homophily which establishes that people tend to strengthen their connection to other similar individuals [1]. In social networking sites, users follow other users with no need for that relation reciprocated or even accepted. In fact, most users tend to avoid the disturbance from uninteresting users, thus they may not follow their followees back [2].

Homophily has been extensively studied in sociology literature [1, 3] by conducting surveys on human subjects. Traditionally, homophily has been analysed in terms of user similarity, which in turn has been used to explain concepts such as community development, segregation and mobility. However, this raises two concerns. First, how to define and quantify the concept of similarity given the broad spectrum of alternatives. Second, due to the nature of the sociological studies and the experimental evaluation, their conclusions might not be extensible to the online world and, particularly, social media. Relationships in social media might be formed based on the same basis than in the real-world. However, given the differences between those two environments, it might be difficult to determine whether the same factors governing relationships in the real world also influence relationships in social media [4]. For example, in the case of face-to-face relationships, those with others with similar interests or opinions can be promoted by the exposure to socio-demographically similar people in places such as schools, universities, workplaces or even neighbourhoods [5]. However, in online environments users usually know others only through their profiles.

The differences between the real and online environments pose the question regarding which are the primary factors that drive homophily in online social networks (OSNs), and how they affect the formation of new ties. As socio-demographic information is rarely present on social media data, it is necessary to focus on the role of users' interests and behaviour as homophily drivers. Understanding what fosters the formation of social relations in OSNs becomes crucial for accurately assessing user similarity and hence defining precise and personalised strategies to be applied in recommendation systems. Moreover, the exponential grow of online activity hinders the ability of users to find relevant and reliable information, which creates a potential overload and prevents timely access to items of interest. This has increased the demand for recommender systems, which act as information filtering systems handling the problem of information overload that users normally encounter by providing them with personalised recommendations. A recurrent topic in recommender systems research is the

[1] ISISTAN Research Institute (CONICET-UNCPBA)

Corresponding author:
ISISTAN Research Institute (CONICET-UNCPBA) Campus Universitario, UNICEN University, Tandil, Bs. As., Argentina

generation of metrics to accurately assess the similarity between users or items [6].

A review of a wide range of works applying the concepts of homophily and user similarity for recommendation tasks [7–12], amongst others, has shown that the reliance on such concepts was not justified neither by a conceptual analysis of the involved similarity aspects nor users' context. Moreover, most works were based on similarity metrics stemming from other research areas such as geometry, biology or economics.

Motivated by the explicit differences between the real and online worlds and the lack of a systematic study of homophily in OSNs, this study aims at understanding how people connect in micro-blogging platforms by analysing the importance of different behavioural aspects and interests of users for adequately characterising social ties. Specifically, this study is founded on the following research questions. First, whether homophily principles drive the formation of social ties, i.e. whether users establishing social ties share similar characteristics. Second, which factors drive the formation of social ties, i.e. how to effectively measure user similarity. Third, considering that similarity can be based on distinct aspects of users' interests and behaviour, how the different aspects contribute to strengthen the homophily amongst friends. Fourth, whether user similarity is restricted to friends, i.e. how similarity amongst friends compares to similarity with other random social network users. To answer these questions, the concept of user similarity was explored in terms of a statistically analysis of diverse traditionally used similarity metrics, not only by assessing the relation between users and their friends, but also the relation between a user and the rest of the online community.

The rest of this work is organised as follows. The Literature Review Section presents related research. The Hypotheses and Research Model Section describes the defined hypothesis and the research model based on two data dimensions. The Research Method Section provides a description of the study methodology by describing the collected dataset and how the hypotheses were tested. The Data Analysis and Findings Section analyses the followee preferences of the studied users regarding the different recommendation factors, across the diverse similarity metrics. Then, the Discussion and Implications Section discusses the findings and some practical implications. Finally, the Conclusions Section summarises the conclusions drawn from this study.

Literature Review

The homophily principle has been extensively studied in the context of real-world data by conducting numerous surveys with human subjects. For example, McPherson et al. [1] studied how the similarity between individuals in terms of socio-demographic characteristics (e.g. geographical and locality factors) can foster the development of social ties, but neglected the relevance of users' interests. Selfhout et al. [13] leveraged on the reinforcement-affect theory to state that similarities in terms of feelings, views and opinions can trigger implicit responses that increase people's attraction. The effect of the Big Five personality dimensions on the formation of dyads is studied by [14] and [15], suggesting that each dimension has an important and differentiated role in friendship selection. Both studies focused on demonstrating the existence of similarities between individuals in a dyad, but did not explore the similarity with outsiders. The mentioned studies have in common that all of them were conducted in a physical world scenario by surveying groups of human subjects [4]. Often, subjects belonged to specific geographical locations, with similar socio-demographic characteristics. Thereby, ties were subjected to social influence, which inherently favoured the conclusions of the studies, hindering their applicability to the online domain.

The advent of OSNs has offered new strategies to evaluate the homophily theories on a much wider scale. For example, Singla and Richardson [7] applied data mining techniques to the study of a MSN Messenger network and discovered that people chatting together share personal characteristics, such as demographic data and queries to search engines (which were regarded as users' interests). Findings also showed that people who do not necessarily chat together but have common friends also tend to share some similar characteristics. Gilbert and Karahalios [8] defined variables regarding user demographics and interactions to predict tie strength on *Facebook*. Tommasel et al. [16] studied the impact of personality in the friendship selection process in *Twitter* verifying the hypotheses presented in [14, 15]. Tang et al. [17] defined user similarity in terms of gender and geographic location as a driver for retweeting behaviour. [18] also showed that homophily plays an important role in determining with whom to connect, as users predominantly choose to follow and interact with others from the same national identity.

The described studies have mainly focused on the existence of coincidences amongst demographic information that, in OSNs might be either unavailable or untrustworthy. It was even argued that this way of assessing homophily can put minority groups at a disadvantage by restricting their ability to establish links with a majority group or to access novel information [19]. Conversely, one of the strongest factors for evaluating homophily in the virtual world, although often neglected in physical world studies, is the matching interests of individuals. Several approaches have been proposed in the literature to recommend users worth following defining user similarity in terms of users' interests [20], network topology [5, 21–24], personality [9] and popularity [25, 26], geographical location [27, 28], the content users publish [10], or even emotions [29]. These works assumed the existence of homophily and only studied the performance of the selected similarity metrics in relation to the precision of recommendations, without analysing the adequacy of the metrics for measuring similarity, i.e. whether such metrics could accurately represent user similarity, or whether according to those metrics homophily also existed amongst strangers.

In OSNs, several metadata elements have been used for quantifying homophily. [30] studied the presence of homophily in three systems that combine social tagging with OSNs (*Flickr*, *Last.fm* and *aNobii*). The analysis suggested that users with similar interests are more likely to be friends, and therefore topical similarity among users based solely on their annotation metadata should be predictive of social links. Xu and Zhou [31] showed the homophily effects through hashtags, where users engaging with certain hashtags have higher chance of forming ties. Two patterns of homophily through hashtags were identified in this work. On the one hand, hashtag homophily can be established between two users sharing the same hashtags as the are more likely to form ties. On the other hand, a pattern where homophily alienates users who do not share the same hashtags, which have a lower likelihood of forming ties. [32] observed the effect of homophily in individuals' willingness to participate in collective actions in *Facebook* (e.g. protests). [33] carried out an in-depth investigation on the role of semantic homophily in a network of *Twitter* mentions. A temporal analysis of communication reveals that links persisting over several months present stable properties, such as semantic (content similarity) and status (social influence) similarity between source and receiver, which are not observed in short-lived links.

Finally, Bisgin et al. [4] aimed at exploring the principle of homophily based solely on topic similarity over the used tags. The study considered three social networks *BlogCatalog*, *Last.fm* and *LiveJournal*. At a dyadic level, their results showed that people sharing a social tie often do not share interests. At a community level, the authors found that people did not only have similar interests with other members of the same communities, but also to the whole population, suggesting that homophily also existed with outsiders. According to the authors, this implied that communities evolve based on the tie density of groups of users that do not have distinctive interests. Moreover, studies over a random rewired version of the dataset suggested that ties were not driven by homophily. In the overall, results seemed to contradict the assumption that homophily fosters the formation of social ties. This study raised several concerns regarding whether conventional theories established based on real-world observations hold when analysing OSNs. However, the study lacked of a conceptual study of how to assess similarity, as it implicitly assumed that users' interests are only expressed using tags.

Despite the evidence that similarity fosters the attraction between individuals, the explanation of such effect continues to be the subject of debate [3]. For example, existing models are unable to explain why attraction occurs more in laboratory than in field studies, or the lack of attraction even in the presence of similarity regarding the negative traits. Additionally, it has been questioned why similarity regarding peripheral factors does not lead to less attraction than similarity on important factors [3]. Similar concerns have been expressed regarding online social relations [4]. This brings into question how to effectively model similarity, and which is the effect of such perceived similarity. However, the studies over OSNs data have merely relied on the phenomenon of homophily by applying similarity metrics without studying their pertinence and relevance to the task to be performed. Moreover, to the best of our knowledge no previous study has explicitly analysed the characteristics of the missing relations in social networks, i.e. how similarity behaves amongst strangers.

Hypotheses and Research Model

Motivated by the observed differences between the real and online worlds, this work proposes a systematic and novel study of homophily in OSNs aiming at discovering how homophily is reflected on the established online social relations, i.e. how traditionally used similarity metrics capture the essence of homophily. To determine the strength of homophily, ties are analysed from a wider point of view by not only assessing the characteristics of friendships, but also, how people relate to strangers in terms of their similarity.

Founded on previous sociology and psychology research that established the existence of homophily on real-world friendship relations [1, 14, 34], and to answer the motivational research questions, this study centres on the existence of homophily in OSNs in the context of a friend prediction task. According to the findings in [1], homophily can be expressed in diverse manners. For example, geography, race, religion, age, and even belief were shown to influence the formation of social ties by fostering interaction and attraction between individuals. As regards OSNs, Thelwall [35] also established the existence of socio-demographical (e.g. religion, age, country and ethnicity) homophily in *MySpace*. Interestingly, Verbrugge [36] found that the factors driving homophily might change according to the characteristics of the analysed group of individuals. For example, social ties amongst adults in some cities in Germany were more structured by work occupation than those in USA. Additionally, in Taiwan, relations complied with the normatives and social values governing daily life [37].

In the context of OSNs, users' interests and behaviour are traditionally analysed in terms of topological or content-based factors. Although a systematic study on the effect of each possible factor has not been performed in the literature, the results of followee recommendation suggest variations in the precision of recommendations according to the selected factor. For example, Armentano et al. [10] reported better precision results for content-based factors than for topological ones. On the other hand, Hannon et al. [38] reported that the combination of topology with content-based information achieved worse results than topology. Hence, each factor might not be equally important to every individual.

This study is guided by two hypotheses, which do not only refer to the criteria under analysis, but also to the intrinsic characteristics of users and social media sites that might influence their preferences. For example, users' behaviour (in terms of number of friends or the level of posting participation) might alter their friendship preferences. To verify the defined hypotheses, it is necessary not only to study the similarity between users across diverse metrics, but also how such similarity between friends compares to the similarity with other random OSN users.

As previously stated, in real-world studies it was found that although social ties are effectively driven by homophily regardless of the different geographic locations, the specific characteristics of both the environment in which the interactions occur and the involved individuals might have an effect over the factors leading to homophily. For instance, regarding socio-demographic factors, gender homophily was shown to be lower on Anglosajon societies when compared with African American and Hispanics ones [1]. Additionally, Cuevas et al. [28] claimed that location and language dictated the degree of geographic homophily. For example, users in countries with languages different than English (such as Brazil) exhibited a higher level of geographic homophily in their relations than users in English speaking countries (such as UK or Canada), who tended to relate with users in other countries. Following this notion, it could be inferred that the same situation applies to OSNs, i.e. the environmental characteristics of the OSN under analysis, which encourage certain types of activities, might condition the factors driving the formation of social ties. In this regard, the first hypothesis states that:

H1. *The characteristics of the social network under analysis influence the overall importance or relevance of the diverse factors.*

Specifically, in an information centric network, i.e. a OSN that is guided by the desire of consuming information (e.g. *Twitter*), the content similarity between users will be higher than the topological one. Conversely, on a friendship based network (e.g. *Facebook*), relationships will be driven by topological factors. In this context, Armentano et al. [10] and Hannon et al. [38] reported contradicting results whether content-based or topology factors achieved the highest precision in *Twitter* recommendations. On the other hand, in *Facebook* similar interests or socio-demographic characteristics achieved worse precision than recommendations based on topological factors [39]. Recently, [40] argued that structural diversity of common neighbourhoods had a positive influence in networks such as *LinkedIn* or *BlogCatalog* (i.e. content-oriented networks), whilst a negative influence in networks such as *Facebook* and *Friendster* (i.e. social oriented networks). As the level of user participation (measured as the number of followees, tweets or interactions, amongst other possibilities) might also impact on the characteristics of selected followees, this hypothesis aims at verifying whether content-based relations have greater relevance in information-oriented networks, and whether such impact is related to user participation in the social network.

As exposed in previous works, the factors driving homophily are not unique, and might inter-relate with interesting effects. In real-world studies, the combination of several factors was shown to make social relations less likely than the individual factors would have suggested [41]. Similarly, in OSNs, friendships might attend, possibly simultaneously, to several reasons. For example, individuals might choose to follow some individuals because they share mutual friends, others because they are celebrities, or others because they publish interesting information, amongst other possible explanations. Besides having multiple and diverse factors, there are multiple alternatives for assessing each of them. For example, topological similarity can be measured by considering neighbour-based, path-based or random walk-based metrics [42]. Similarly, content-based similarity can be computed by diverse metrics based on the actual content people post, the used tags, the comments people leave on others' content, or even the writing style. For example, Armentano et al. [10] and Hannon et al. [38] analysed content homophily based on the pre-processed content of tweets, whereas Chechev and Georgiev [20] considered the hashtags and links in tweets, obtaining contradicting results. In this context, the second hypothesis states that:

H2. *The diverse criteria for characterising users' interests and behaviour and their associated similarity metrics target different aspects of user relationships and, consequently, each combination of factor and similarity metric leads to differences in the quality of recommended followees.*

This hypothesis deepens on the concept of user similarity aiming at exposing that choosing the relevant recommendation factors is not sufficient for guaranteeing high quality recommendations. In this context, it explores the importance of adequately defining the concept of user similarity in the context of the followee recommendation problem.

Research Method

The influence of homophily in the formation of new social ties in the microblogging community was studied by analysing the characteristics and effects of diverse user similarity definitions. To that end, two of the most commonly used similarity factors in followee and friendship prediction were modelled. First, topological factors on which most of the traditional link prediction algorithms rely on. Second, content-based factors, which reflect the interest of users regarding the information they share and consume.

Twitter was the social networking site chosen for assessing the impact of the followee recommendation factors and similarity metrics. The rationale behind this decision is that it is embedded in everyday social and communicative interactions around the world, and its role as a public, global and real-time communications provides a glimpse on contemporary society as such [43]. *Twitter*'s easiness of use has converted it in a media for sharing news or reports about events of the everyday life through politics or emergencies. This is completed by the possibility to access to its data, in comparison to the data of other social networking sites. Almost 90% of the user *Twitter* accounts are public, implying the richness of the information that can be obtained from such network. The *Twitter* dataset was created by crawling a set of $3,453$ target users who frequently tweeted about multiple topics. Approximately a half of the target users were originally included in [44], comprising politicians, musicians, environmentalists and other users. The originally crawled users were chosen based not only on their topology, but also considering user context, such as their activities, location and shared information, to improve the representativeness of the selected sample regarding the social and information diffusion processes of the full graph. The remaining target users were selected from their followee set to increase user diversity, as they were chosen regardless of their popularity or posting activity.

To guarantee both meaningful content-based profiles and an extensive topological network, several restrictions were imposed on users to be selected. First, users must have more than 10 followees and more than 10 published tweets regardless whether the tweets were originally posted by the user or they are retweets. Second, the user account must had been listed as English, and the first set of retrieved tweets must had also been written in English. For determining tweets' language, the first 200 downloaded tweets of each user (or less, depending on the total number of published tweets) were analysed using TextCat[*].

For all target users and their followees, user account information, tweets, favourite tweets, followees and followers were retrieved from *Twitter*, through the *Twitter* API[†]. Table 1 summarises user statistics. For average values, the standard deviation is shown between parentheses, exposing that the number of tweets, followees and followers are distributed over a great range of values. In the analysed dataset, 25% of the target users have fewer than 36 followees, and 50% of the users fewer than 125. This implies that the dataset covers a wide spectrum of users, ranging from users only seeking information (i.e. users with a few followees) to celebrities (i.e. users with many followees). For each target user, a set of randomly selected non-followed users was also collected to analyse the correspondence between the similarities between users and their followees, and the similarities with other strangers, or users who might not have been of interest to target users. In all cases, for each target user a number of non-followed users equal to the double of the number of actual followees was selected, provided that the similarities between such users and the target ones had a random distribution. Randomness was analysed with the Wald-Wolfowitz test for continuous data as defined in [45]. As no order exists between the events, i.e. the target user similarity with each of the newly selected users, randomness was tested against both trends and first order negative serial correlation. In the former case, the similarity distribution was tested against itself at different times.

Tweets' terms were filtered according to two text processing strategies to build the content-based profiles. The first one ($FULL$) considered tweets' full-text, whereas the second one ($PROC$) applied lexical and syntactical pre-processing steps to tweets. The pre-processing included removing all non-English tweets, keeping only nouns and verbs, and applying the Porter Stemmer algorithm [46] to reduce the syntactic variations of terms and to improve the probability of finding similarities between profiles.

As previously mentioned, this study is based on two factors that are commonly used in followee and friendship prediction: topological (Section Topological Factors) and content-based factors (Section Content-based Factors).

Topological Factors

Most link prediction algorithms are based on topological features. Generally, these algorithms consider user's neighbourhood or topological paths for computing user similarity. Table 2 presents the neighbourhood metrics and local similarity indexes based on topological features [47] that were included and analysed in this study. The first three metrics correspond to neighbourhood metrics, whereas the rest correspond to local similarity indexes.

Content-based Factors

Micro-blogging platforms have become a popular communication tool amongst Internet users. Millions of users share opinions, details of their personal life or discuss with other users through millions of messages posted daily, converting these platforms into both informational and social networks [48]. In most sites, users establish social relations by choosing friends and subscribing to the content they publish. Hence, content arises as an important factor for recommending who to follow, as users are likely to become friends with whom they share content preferences. Users' interests can be defined considering profiles based on the content they publish, or the content they read or consider interesting. Whereas the first alternative assesses users' interests regarding the information they create and publish, the second one analyses users' interests in terms of the information they consume, i.e. the information they marked as interesting. These profiles will be referred as *publishing profile* and *reading profile* respectively.

In *Twitter*, content is represented by the tweets users write. The set of tweets t for a user u_j can be denoted as:

$$tweets(u_j) = \{t_i, ..., t_n\} \tag{1}$$

The *publishing profile* of a user considers all published tweets, assuming that users post about things that are interesting to them and want others to read. Formally, the *publishing profile* of user u_j can be defined as:

$$pub-profile(u_j) = tweets(u_j) \tag{2}$$

The goal of building a *reading profile* is to accurately capture users' interests regarding the information they consume. In *Twitter*, if a user likes to read tweets regarding a certain topic, he/she is expected to follow users tweeting on those topics. However, followees could tweet about multiple topics, which might not be all of interest to users. Thus, it is important to identify the specific tweets that users considered interesting. *Twitter* provides two mechanisms for expressing interest and engagement on other users' tweets. First, analogously as when bookmarking Web sites, tweets can be marked as favourites. Second, tweets can be retweeted, i.e they are reposted or forwarded to other *Twitter* users. When users retweet, such tweet is visible to their followers, meaning that the original tweet is shared with more people. Hence, favourited and retweeted tweets are key mechanisms for information diffusion, conveying the information users are actually interested in consuming [49].

This leads to two alternatives for creating the *reading profile* of a user u_j. First, a reading profile containing only the favourited tweets ($tweets_{Fav}$), as Equation 3 shows. Second, a reading profile containing only the tweets that the user has retweeted ($tweets_{RT}$), as Equation 4 proposes.

$$\begin{aligned} read-profile_{Fav}(u_j) = tweets_{Fav}(u_k) \\ \forall k \in followees\,(u_j) \end{aligned} \tag{3}$$

[*] http://odur.let.rug.nl/vannoord/TextCat/

[†] https://api.twitter.com

Table 1. Data collection general statistics

Total number of target users	3,453
Total number of tweets	3,227,782
Average number of tweets per user	935.86 (± 1,200.21)
Total number of followee relations	1,650,208
Average number of followee relations per user	478.46 (± 2,440.53)
Total number of follower relations	23,626,904
Average number of follower relations per user	6,850.36 (± 187,662.64)

Table 2. Assessing Topological Similarity

Neighbourhood Metric	*Common Neighbours*	Measures the overlap of the ego-centric networks of two users, including both outgoing and incoming links. This metric is an adaptation of the Jaccard similarity measure.	$\frac{\|\Gamma(x)\cap\Gamma(y)\|}{\|\Gamma(x)\cup\Gamma(y)\|}$
Neighbourhood Metric	*Common Followees*	Measures to what extend two users follow the same set of users. If two users follow the same users, they are likely to have similar interests and thus, be interested in the same type of information.	$\frac{\|\Gamma_{out}(x)\cap\Gamma_{out}(y)\|}{\left\|\Gamma_{out}(x)\cup\Gamma_{out}(y)\right\|}$
Neighbourhood Metric	*Common Followers*	Measures to what extend two users are followed by the same people, and thus share the same audience.	$\frac{\|\Gamma_{in}(x)\cap\Gamma_{in}(y)\|}{\left\|\Gamma_{in}(x)\cup\Gamma_{in}(y)\right\|}$
Similarity Index	*Salton*	Computes the distance between the neighbourhood of each user represented as a binary vector.	$\frac{\|\Gamma(x)\cap\Gamma(y)\|}{\sqrt{\|\Gamma(x)\|\times\|\Gamma(y)\|}}$
Similarity Index	*Sørensen*	Measures the number of shared neighbours, and penalises it with the sum of the neighbourhoods sizes.	$\frac{2\|\Gamma(x)\cap\Gamma(y)\|}{\|\Gamma(x)\|+\|\Gamma(y)\|}$
Similarity Index	*Hub Promoted Index (HPI)*	Measures the number of shared neighbours and penalise it by the minimum neighbourhood size.	$\frac{\|\Gamma(x)\cap\Gamma(y)\|}{min\{\|\Gamma(x)\|,\|\Gamma(y)\|\}}$
Similarity Index	*Hub Depressed Index (HDI)*	Measures the number of shared neighbours and penalise it by the maximum neighbourhood size.	$\frac{\|\Gamma(x)\cap\Gamma(y)\|}{max\{\|\Gamma(x)\|,\|\Gamma(y)\|\}}$
Similarity Index	*Leicht-Holme-Newman Index (LHNI)*	Measures the number of shared neighbours and penalises it by the product of the neighbourhood sizes.	$\frac{\|\Gamma(x)\cap\Gamma(y)\|}{\|\Gamma(x)\|\times\|\Gamma(y)\|}$

where x and x denote the nodes for which the similarity score is computed, $\Gamma(x)$ denotes the set of neighbours of x, and $|\Gamma(x)|$ denotes the degree of post x.

$$read - profile_{RT}(u_j) = tweets_{RT}(u_k) \quad \forall k \in followees\,(u_j) \tag{4}$$

In turn, both alternatives can be combined as:

$$read - profile_{Fav-RT}(u_j) = tweets_{Fav}(u_k) \cup tweets_{RT}(u_k) \quad \forall k \in followees\,(u_j) \tag{5}$$

comprising all the favourited and retweeted tweets of user u_j, that were posted by any of their k followees.

User profiles are represented following the traditional vector space model [50], in which each vector dimension corresponds to an individual term appearing in the considered set of tweets weighted by its frequency of appearance. Note that weighting strategies requiring knowledge of the full tweet collection, such as TF-IDF cannot be applied. As profiles are intended to be used in real-time settings, posts would be constantly arriving, leading to two implications. First, there is no fixed available document corpus on which base the IDF computation. Second, if the data collection is considered to expand every time new tweets are known, the TF-IDF score of each feature has to be periodically computed, resulting in an inefficient approach. Note that, not only the statistics of terms in the newly arriving tweet would be computed, but also, the IDF statistics of the other terms should also be updated. Thus, although some information regarding the overall terms' relevance might be lost, in highly dynamic environments it is preferable to use more efficient weighting schemes, such as term frequency.

Once profiles are built, the similarity between them can be computed using the cosine similarity metric [50]. For followee recommendation, the profile of target users should be matched to those of the potential followees. For example, the $read\text{-}profile_{RT}$ of target users could be matched with the $pub\text{-}profile$ of potential followees, which would be denoted as $read_{RT}\text{-}pub$. On the other hand, the same profile for the target user and the potential followees could be also matched. For example, $read_{RT}$ denotes the matching of the $read\text{-}profile_{RT}$ profiles.

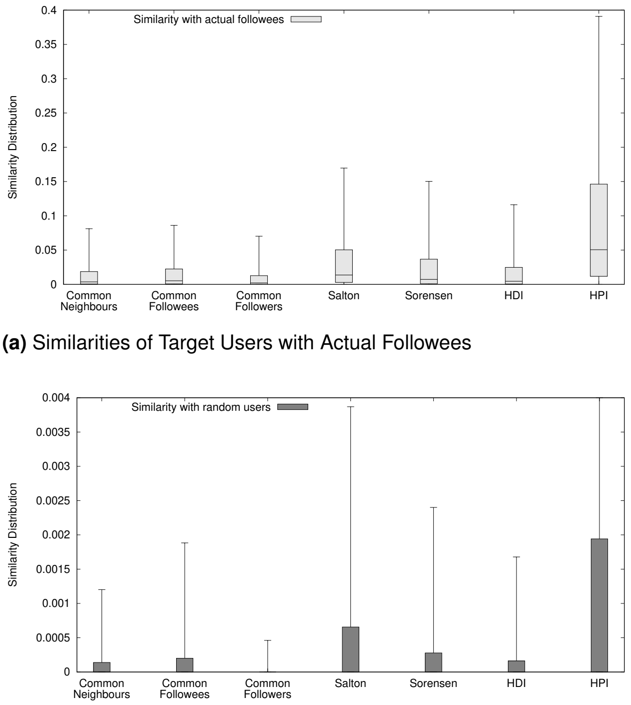
Data Analysis and Findings

The following sections describe the data analysis performed to study the influence of the different factors and similarity definitions in followee selection. The analysis focuses on understanding how user similarity conditions friendship, whether similarity patterns exist between friends, and how such similarities compare to the similarity with other randomly non-followed users. To that end, two hypotheses were defined. The first one aims at determining whether content-based relations have greater relevance in information-oriented networks, and whether such impact is related to user participation in the social network. In this context, for each factor (i.e. topology and content), the overall followee similarity distribution is presented and compared to the overall similarity distribution with randomly non-followed users. Outliers can be defined as observations that lie at an abnormal distance from the other values in the distribution, i.e. they are dissimilar to the majority of the remaining data points. In this case, outliers represent similarities between users and their followees that significantly differ from the remaining similarities. In this context, such similarities could be removed as they might not represent the characteristics of the majority of the users, forcing a skewing of the data distributions towards either the low or high values. Outliers were detected following Tukey's method [51], which is applicable to both normal or skewed data as it does not make any assumption regarding the data distribution.

For followee recommnendation, a pool of potential followees to be recommended was built for each target user by including the actual followees and the set of randomly selected users. Then, potential followees were ranked according to the chosen similarity metric and selected by the recommendation algorithm. The quality of recommendations was evaluated by analysing whether the actual followees were recommended, i.e. whether the selected factor and similarity metrics were enough for adequately identifying users who were already deemed as interesting. Particularly, it was evaluated by selecting the top-N recommended followees and computing the overall precision defined as the percentage of relevant recommendations (the number of actual followees that were discovered) regarding the total recommendations. For all experimental evaluations, N was set to 5, 10, 15 and 25 positions of the ranked recommendation list, and for each list of length N, the reported precision corresponds to the aggregated precision for all target user.

In all cases requiring the analysis of the significance of the observed differences, statistical tests were used based on [45]. Sample normality was evaluated by analysing their skewness, kurtosis, and performing both the Shapiro and the Anderson-Darling tests.

The similarity metrics achieving the highest recommendation precision were analysed to determine whether the level of user participation has an effect on the characteristics of selected followees. To that end, target users were grouped into four equal parts delimited by the first quartile, median and third quartile, according to their number of followees or published tweets. In this context, considering all users sorted in ascending order according to either the number of followees or the number of published tweets, the median represents the value that separates the 50% of the higher values from the 50% of the lower ones, i.e. represents the value in the middle of the distribution. Then, the first quartile represents the median of the first half of the data distribution, marking the point at which 25% of the values (either the number of followees or the number of published tweets) are lower than the first quartile and the remaining 75% are higher. Similarly, the third quartile represents the median of the second half of the data marking the point at which 75% of the values are lower and the remaining 25% is higher. User grouping was based on quartiles because they are not based on the supposition of a symmetric distribution of data and not influenced by data outliers. Thereby, the interquartile range is an adequate and robust statistic when data is skew (as the mean and average values in Table 1 show), or when the data characteristics are not known in advance [51].



(a) Similarities of Target Users with Actual Followees

(b) Similarities of Target Users with Random Users

Figure 1. Similarity Distribution for the Topology Factor

The second hypothesis explores the concept of user similarity and how it influences friendship. Considering the distribution patterns in both friend and randomly selected users, user similarity's effectiveness was studied across diverse metrics in a followee recommendation task.

Topological Factors

Figure 1 presents the similarity distribution for each topological metric described in the Topological Factors Section, for both the actual followees and the randomly selected non-followed users. The similarity distributions of LHNI are not included in as they were at least two magnitude orders smaller than the other chosen metrics. For each metric, the randomness of its score distribution was tested using the Wald-Wolfowitz test. For all target users, results showed that followees were not chosen at random, i.e. their similarities did not correspond to a random distribution. As Figure 1 shows, the similarity distributions are higher for the similarity indexes than for the neighbourhood metrics. However, for both metric types, similarities were lower than 0.4, resulting in low to moderate topological similarities in general.

Besides analysing the non-randomness of distributions, the statistical difference between the similarity distributions of the actual followees and the randomly selected users was analysed. The Mann-Whitney test for unrelated samples was used, setting the confidence value (p-value) to 0.01 and defining the null and the alternative hypotheses. The null hypothesis stated that no difference existed amongst the similarity distributions, i.e. the similarity distribution of the actual followees and the random users where equal. Conversely, the alternative hypothesis stated that there was a non-incidental difference between both distributions. For each metric, more than the 95% of the target users showed significant differences in the distributions. In other words, there was approximately 5% of target users, who despite not choosing friends at random, did not show a significant pattern of similarities with the followees that allowed distinguishing them from randomly selected users. This shows that there are some users who do not seem to engage with followees according to their topological similarity. Additionally, as Figure 1 depicts, the one sided statistical test showed that similarity distribution with the actual followees was higher than that of the randomly selected users.

Additionally, it was tested whether the metrics measured different aspects of user similarity, i.e. whether their results were unrelated. The Wilcoxon test for related samples was used, setting the p-value to 0.01, and defining the null and the alternative hypotheses. The null hypothesis stated that no difference existed amongst the similarity metrics, whereas the alternative one stated that each metric had a distinctive score, different from the other metrics. Cliffs's Delta was used to quantify the effect size between the compared similarities. Table 3 summarises the observed effect sizes, where an empty cell means that the observed differences were not statistically significant, while in the other cases there exists a significant statistical difference with a confidence of 0.01. As it can be observed, the null hypothesis was rejected for most pairs of metrics with a few exceptions. In this regard, no difference was shown between Salton and Sørensen, *HDI*, *LHNI* and *Common Neighbours*, and between *HDI* and *HPI*. In other cases, even though there existed a significant difference, the effect was negligible. That was the case of *Common Neighbours*, *Common Followees* and *Common Followers*, and *Common Followees* and *HDI*. As the Table shows, there is a statistically significant difference between the scores observed for the similarity indexes and neighbourhood metrics. These differences can be explained in terms of how metrics are defined. As Table 2 shows, when comparing *Common Neighbours* with the similarity indexes, they only differ in the denominators. Moreover, the denominators of most of the similarity indexes are always lower than those of the neighbourhood metric, as the degree of union of the set of neighbours is presumably going to be higher than the degree of either set, or to the half of the sum of both degrees (as in *Sørensen*). The only exception to this is situation *LHNI*, in which the product of both degrees should be higher than the union of the degrees, yielding a higher denominator and thus lower similarity scores than *Common Neighbours*.

Figure 2 shows the precision results of using the presented metrics in the context of a recommendation task. As shown, precision ranged between 0.97 and 1.0, which could imply that all metrics could accurately distinguish between the actual followees and the random non-followed users. This distinction could be due to the different similarity distributions shown by the followees and the non-followees. As Figure 1 showed, non-followees had a significantly lower similarity distribution, causing that when sorting by similarity, those users were mostly ranked at the bottom of the ranking, and thus were not selected for recommendation.

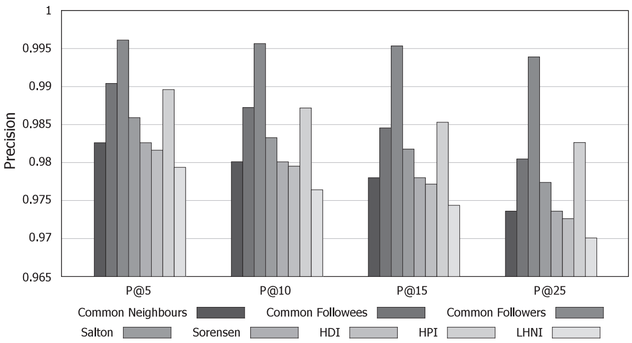
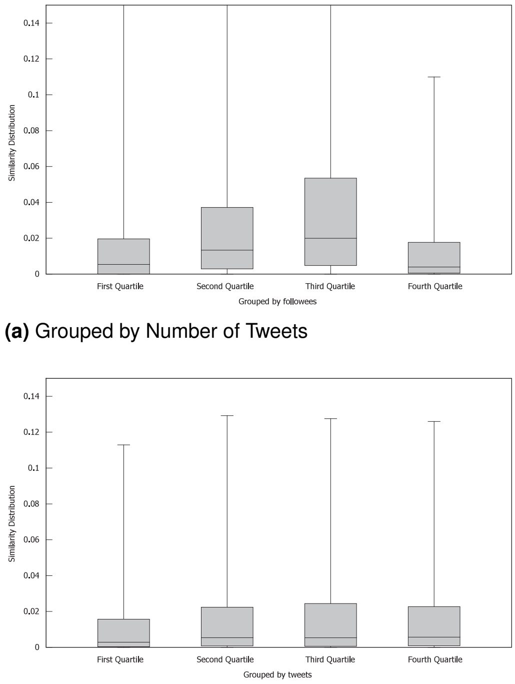


Figure 2. Comparison of Precision Results for the Topology Factor

A statistical analysis of the observed differences based on the Wilcoxon test with a confidence of 0.01 was performed for each top-N. The analysis showed that although differences in almost all cases were statistically significant, their effect size was negligible. This implies that regardless of their similarity distribution, in most cases metrics behaved alike during recommendation. A few exceptions were found as the length of the ranking increased. When considering the top-25, the differences between precision results showed small to medium effect sizes, for example when comparing *Common Followers* with some of the similarity indexes. In this context, it could be inferred that as the number of users to recommend, and thus the number of potential mistakes increases, the selection of the similarity metric to use becomes relevant for achieving the best possible recommendation results. On the other hand, those pairs of metrics that did not show statistically significant differences for their similarities, did not show statistically significant differences regarding their recommendation results.

Finally, as regards how user behaviour (expressed as the level of user participation on the social network) affects followee selection of followees, Figure 3 presents the *Common Followees* similarity distribution for the target users divided according to the statistical distribution of their number of followees or number of published tweets. Each group in the Figure represents a quartile. As previously mentioned, quartiles divide a distribution into four equal parts, in which the first quartile represents the value separating the 25% lower and 75% of higher data values, the median separates the lower and higher halves and the third quartile the 75% lower and 25% of higher data values. This metric was chosen for illustrating the effect of user behaviour as it is one of the most commonly used topological metrics in the literature. As observed, the number of shared tweets does not significantly affect the topological similarity distribution, implying that the publishing activity is not strictly related to the topological characteristics of

Table 3. Analysis of differences between topological similarities



(a) Grouped by Number of Tweets

(b) Grouped by Number of Followees

Figure 3. Influence of User Behaviour in Topology-based Similarity

the chosen followees. This could be due to the fact that as users publish more, they might be more interested in befriending users according to the shared content regardless of the topological similarity. Additionally, this interrelation between content aspects and topological similarities could indicate the existence of different sub-groups of followees, which are selected according to different criteria.

On the other hand, when grouping users according to their number of followees, at first, as the number of followees increases, the similarity distribution also increases. This could be caused by different phenomena. Even though Twitter is mostly a content-based network, at first, when users create their account, in most cases, they are recommended users that can be found in their email contacts (provided they granted the access to it), or contacts in other social media site (for example, *Instagram* includes in contact suggestions *Facebook* friends). In this case, it is expected that topological similarity will increase as more followees are added as they mostly belong to the same network and are likely to be connected. As a next step, recommendations will include followees-of-followees, which in case of being accepted will increase similarity even more with the already selected followees. Nonetheless, it might occur that after users add everyone in their close topology network, they might start expressing interest for the shared content instead of the topology, which would be accompanied by higher content and lower topological similarity.

Content-related Factors

To analyse the content-based related factors, the different user profiling strategies based on users' interests were combined. Figure 4 presents the similarity distribution for each combination of the profiling strategies described in the Content-based Factors Section for both the actual followees and the randomly selected non-followed users. As depicted, content-based similarities spanned over a higher range of values than the topological ones. Whilst topological similarities spanned between 0 and 0.4, content-based similarities spanned up to 0.95. This could be related to the content-based nature of *Twitter*, in which the content users share is a stronger motivation than proximity for following others. This is also fostered by the echo chamber and filter bubbles phenomena [52] that states that users tend to relate to others confirming their narratives and holding similar beliefs, which manifests through stronger content-based similarities. In this sense, similarly to the expanded topological recommendations, as users start befriending others sharing a particular content, they are likely to be recommended users sharing similar content. In addition, this implies that topological and content-based similarities do not share the same space, and thus cannot be directly combined into a single ranking for recommendation. For each metric, it was tested whether their distribution was random using the Wald-Wolfowitz test. In all cases, results showed that similarities did not have a random distribution, i.e. followees were not chosen at random.

As Figure 4 shows, regardless of the selected profiling strategy, the full-text of tweets lead to higher similarities between target users and their followees. As similarity decreases with the reduction of syntactic variations of words imposed by the $PROC$ processing strategy, these results could be explained by the existence of tweets sharing non-meaningful words, instead of being related by their relevant content. Interestingly, the highest similarity distributions were obtained for those alternatives including $read_{RT-FULL}$, although those profiling strategies are also the ones that present two of the highest dispersions, with a 75% of the followees having similarities ranging between 0.4 and 1. Contrarily, the lowest similarity distributions were obtained for $read_{Favs}$ independently from whether the tweets' content was processed. It is worth noting that the distribution of similarities regarding pub were higher than for several profiling strategies based only on the reading profile, and combining both reading and publishing profiles. Particularly, pub showed a higher similarity distribution than $read_{Favs}$, $read_{Favs} - pub$, $read_{RT-PROC} - pub_{PROC}$ and $read_{RT-Favs-PROC} - pub_{PROC}$. This situation could be related to users not narrowing their interests to one interest or activity, and selecting followees according to such different interests. First, it shows that users tend to share content, in opposition to only being lurkers that consume information. Second, it shows that the own content they share differs from the content that they choose to share or save. For example, the share and published content could belong to different topics. The observed differences between the $read_{RT}$ and $read_{Favs}$ shows that users make distinctions between the content they want to include in their profile and shared with others (the retweets), and the content they simply express interest in (the favourites). The highest similarity

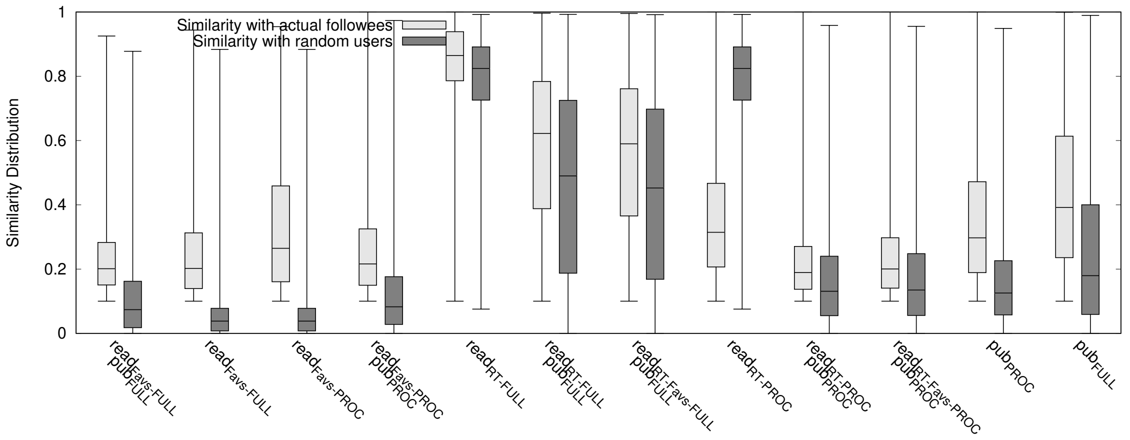


Figure 4. Similarity Distributions of Target Users with Followees and Random Users for the Content-based Factor

distributions were observed for a $read - profile$ implying that the interests of users and their friends are not only similar, but that they befriend people posting on the same topics. These results contrast with those in [10], which reported that users and their followees do not publish similar content.

As for the topological factors, the differences between the similarity distributions of the actual followees and the randomly selected users were analysed with the Mann-Whitney test for unrelated samples. Results varied according to the analysed profiling strategy. As regards $read_{RT-PROC}$ only a 9% of users showed no statistically significant differences between the similarity distributions. Conversely, for $read_{RT-FULL}$ more than a 81% of target users did not show statistically significant differences. For the remaining metrics, approximately half of the users did not show differences. Consequently, the similarity distribution of the actual followees resulted statistically similar to that of the randomly selected users. This is in line with [53, 54] that show that the presence of homophily regarding content-based or topical interest is independent of the topological relations between the users, i.e. like-mined users are not necessarily always connected to each other. Intuitively, given the large number of topics and content possibilities, it is easier to find a random user with similar content-based interests than a user with a similar topological structure.

Similarly to the topology factor, for each similarity metric it was tested whether differences existed between the similarity scores obtained for each actual followee by means of the Wilcoxon test for related samples. The observed effect sizes are summarised in Table 4. Results showed that there was a statistically significant difference between the results of each pair of profiles. Nonetheless, despite being statistically different, such differences were, in some cases, negligible. On the other cases, the compared profiling alternatives were unrelated, which shows that they effectively analyse different and independent aspects of user similarity. Particularly, $read_{RT-FULL}$was shown to have large differences with every other profile, followed by $read_{RT-FULL} - pub_{FULL}$. These differences could be caused by the diverse interests of users which, as previously mentioned, manifest in the different content-related activities.

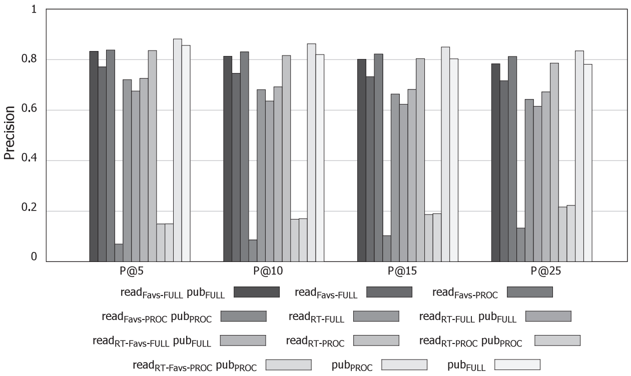


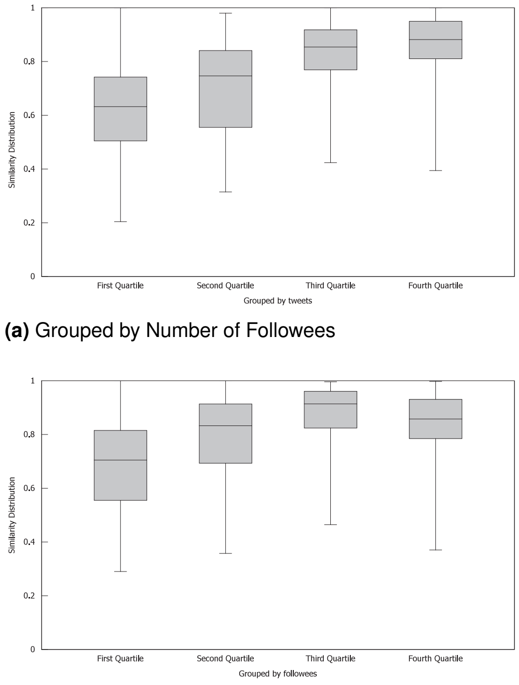
Figure 5. Comparison of Precision Results for the Content-based Factor

Figure 5 depicts the precision results for all the combinations of profiling strategies which, as can be observed, are lower than those obtained for the topology metrics. Note that having high similarity distributions does not necessarily translates into high precision results. For example, the highest quality recommendations were not obtained for $read_{RT}$, which showed the highest similarities between target users and their actual followees. This particular case can be explained based on the observed distributions of similarities amongst followees and the randomly selected users. As both distributions are similar, it is difficult to correctly identify the actual followees. Statistical analysis of the observed differences based on the Wilcoxon test with a confidence of 0.01 was performed for each top-N. Table 5 summarises the effect sizes for the top-10. As it can be observed, differences were statistically significant for all but one pair. The majority of differences showed a medium or large effect. Moreover, even when the effect size of the similarity differences were negligible, some of such differences resulted in large effects over the precision differences.

Finally, Figure 6 presents the similarity distribution for $read_{RT} - pub_{FULL}$. Similarly as in Figure 3, target users are grouped in quartiles according to the statistical distribution of their number of followees or the number of the published tweets. Such profiling was chosen for illustrating how user behaviour can influence the characteristics of

Table 4. Analysis of differences between content-based similarities

Table 5. Analysis of precision differences between content-based similarities



(a) Grouped by Number of Followees

(b) Grouped by Number of Tweets

Figure 6. Influence of User Behaviour in Content-based Similarity

the selected followees as it was the strategy showing the highest similarity distribution in combination with the smallest interquartile range (the 50% of followee similarities ranged between 0.8 and 0.9). As depicted, user behaviour had a greater impact on the content-related factor than on the topology one. When grouping users according to their followee number, the tendency of similarity distributions is similar to that of the topology factor. For those users in the first three quartiles, similarities tended to increase. However, for those users in the fourth quartile, the similarities were lower than those of the third one. These results are in agreement with [55] which stated that homophily is non-monotonic, as it does not grow perpetually. Instead, beyond a point, the increased social relations do not guarantee increased similarity. For users with few followees (i.e. in the first and second quartiles), similarities spanned over a large range than those of the other quartiles. Then, as the number of followees increased, the range of similarities became smaller. This is also in partial agreement with Dey et al. [55], which indicated that as the number of followees increased, the median similarity started to decrease, but, unlike the results in [55], in this case, the variability of similarities decreased, showing more content-cohesive selection of friends. On the contrary, as the number of published tweets increased, the similarities with the actual followees also increased. Similarly to the previous case, the range over which the similarities spanned was higher for those users in the first and second quartiles. This situation exposes that as users tend to be more involved in content-sharing, they choose followees that are aligned with their content interests.

Discussion and Implications

In this Section, an analysis of each of the proposed hypotheses that guided the study in relation with the obtained results is presented. Finally, the implications and possible applications of the performed data analysis are stated.

Homophily According to Diverse Recommendation Factors

The first hypothesis aimed at verifying whether in information centric networks user relations are guided by information needs. It also aimed at verifying whether the user participation level (defined as the number of established social relations or actually published tweets) had an impact on the characteristics of selected followees.

As the data analysis showed, content-based similarities spanned over a greater range of values than the topological ones. Particularly, $read_{RT-FULL}$ similarities are very high, hinting the existence of homophily not only between target users and their followees, but also between the target users and the followees of their followees. The implications of these high $read_{RT-FULL}$ similarities are two-fold. First, it means that both target users and their followees relate to similar kind of users, revealing the existence of shared characteristics between target users and their followees, which leads to homophily. However, this also implies that users share similar characteristics with strangers, and thus that homophily is not only restricted to friendship relations. Interestingly, the other pure reading profile (i.e. $read_{Favs}$) showed contrasting results, as similarities with followees were statistically higher that those with strangers. These results imply that users are more selective regarding the content they save in comparison to the content they retweet, i.e. the content they want to easily find, and the content they want others to immediately see. On the other hand, as regards the topological factor, similarities were restricted to a smaller and lower range, and showed strong statistical differences between the followee and random populations.

Whilst it might be desirable that the selected followees have unique characteristics which would help to distinguish them from other users, if similarity distributions are different, metrics will not be reliable for this task. On the other hand, finding that random users have similar characteristics to the actually selected followees could generate noise, hindering the search of actual interesting users. This is evidenced by the precision results obtained. The similarity between the friend and random populations had an effect on the quality of recommendations expressed by their precision, which could also explain the diversity of results obtained by the studies aiming at recommending followees in *Twitter* (for example [10, 38]).

The effect of the described phenomenon is noticeable on the diverse content-based profiling strategies and can be explained by the content-guided nature of *Twitter*. As shown, user activity had a greater impact on the content-related factors than on the topological one. When considering topology, followee selection was not affected by the number of posted tweets. However, the higher the number of followees, the higher the similarities with the target user. In this regard, having more followees increased the number of users with whom the neighbourhoods was shared. However, for those users having the highest followee numbers, similarities had a similar distribution to that of the users on the first quartile. These results suggest that not all followees are chosen by their topological similarity, as they tended to share few topological ties with their followees. If that would be the case, target users with the highest followee number should also share the highest similarities with their followees.

As regards the content-related factor, results showed that as the number of published tweets increased, the spanning range of similarities shrank, implying that highly participative users tend to choose followees sharing the same interests with lower dispersion. Similarly to the previous case, the similarity spanning range was wider for those users in the first two quartiles, meaning that users who mostly read content do not focus over a unique topic, instead they choose to follow users posting information covering a great range of topics, which are probably not worthy of retweeting. These results agree with those in [53] and [55], which stated that as the number of followees increased, content-similarity in dyads also tended to first increase. The analysis allows inferring that the motivation for choosing followees are not unique nor static, and might change according to users' activities. Moreover, it can be inferred that followee selection might not respond to a unique factor. These results are in agreement with those in [36], which concurred on the existence of different motivations for starting friendships according to environmental characteristics. These motivations for forming new ties might also be related to the characteristics of the social network analysis. For example, García-Martín and García-Sánchez [56] found differences in the motivations for using either *Facebook* and *Twitter*, which effected the type of people with whom users interacted. According to the study, young Spanish people used more *Facebook* than *Twitter* for social purposes involving friends and relatives, whilst they used more *Twitter* than *Facebook* for communicating with strangers.

According to [57] and [58] the relevant aspects of the social environment can be regarded as foci around which individuals organise their social relations. Hence, people connecting around a particular focus of activity tend to present similar behaviour regarding such activity. In the context of *Twitter*, the focus or activity would be sharing content. As a result, it is expected that users in a dyad share similar posting behaviour, which could translate into high content-based similarity. At the same time, people associated with the same focus may vary widely on traits that are not core to the activities of the focus [58, 59]. In this regard, as the focus of *Twitter* is not the establishment of social relations, it is expected that the structural similarity would be lower that the content-based ones. In agreement to [60], exhibiting lower similarities does not necessarily imply that similarity is not an important predictor of the quality of online friendships, as exposed by the precision results obtained.

Although there is no consensus regarding why homophily seems to occur between strangers, this phenomenon could be explain in terms of the foci around which the relations occur (i.e. the content driven nature of *Twitter*) [58]. As all users are motivated to share content, it is expected that content-similarity might be higher that structural similarity amongst strangers. Moreover, results are also in agreement with those in [61], which stated that the number of people with whom traits are shared, i.e. similar people, can influence the homophily towards strangers. Particularly, the authors showed that when users relate with a more exclusive group (i.e. a small group in which not necessarily all users are explicitly related), homophily amongst users is higher, whereas when users relate with more inclusive groups (i.e. big groups), homophily tends to decrease. In this context, the exclusiveness of groups can be measured in terms of number of followees. The analysis showed that, although the same tendencies are observed, as the number of followees increased the median of the similarity distribution with strangers was lower than that for the actual followees, implying a differentiation between users and their surrounding context.

The previous results allowed to validate the hypothesis that the characteristics of the context on which social relations are developed influence the exhibited homophily. These results agree with those found for real-world social relations, in that not every factor yields the same degree of homophily. Particularly, the characteristics of the social networks influence user behaviour, which in turn affects the characteristics of the selected followees. As a result, it can be stated that in an information centric network, social ties are guided by the desire of consuming and sharing information. Also, followee selection was showed to be affected by user behaviour, meaning that interests are not static and that followee selection can be motivated not only by the context of the social network but also by users' behaviour and interests. More importantly, these results allowed to verify the existence of relationship patterns found for real-world relations in an online environment, showing a consistency between offline and online behaviour.

Deciding on the User Similarity Metric

The second hypothesis aimed at demonstrating that identifying the most relevant predictive factor (e.g. content or topology) is not sufficient for guaranteeing high quality recommendations. To this aim, the differences amongst the diverse alternatives for measuring the similarity between users were explored.

For both recommendation factors, the spanning range of similarities was not directly related to the quality of recommendations. In both cases, the fact that the similarities amongst target users and their actual followees was low for a metric, did not imply that such metric would achieve low precision results, as is the case of *LHNI* and $read_{Favs}$. This is related with the findings in [60], which expressed that low similarity does not reduce the importance of the similarity as a predictor of relationship quality.

It is also interesting to analyse the elements that each metric takes into account. For instance, *LHNI* penalises user similarity if any of the neighbourhoods sizes is big, leading to extremely low values if either user has many followees, whilst *HPI* penalises similarity using the minimum neighbourhood size. Although those metrics presented the lowest and highest similarity distributions respectively, they can be misleading in analysing *Twitter* topological similarity.

The statistical analysis of the dependence amongst the similarity metrics showed that several topology metrics were statistically dependent. Although the metrics assess different aspects of social relations, they are intrinsically related, which leads to similar score distributions and even recommendation quality. Conversely, no statistical relationship was found amongst the similarities based on the diverse content-based profiling strategies, implying that each of them assesses diverse aspects of user interests.

Precision results for the topology metrics apparently implied that all metrics were capable of accurately distinguishing between the actual followees and the random set of users. However, as the random population had lower median results than the actual followees, they would be never discovered by the recommendation algorithm, as ranking users according to their similarity would place the actual followees in the first positions (as they are more similar to target users), leaving the random users at the end of the ranking, thus obtaining high precision results. Hence, it could be inferred that only assessing the precision of recommendations could be misleading for understanding the followee phenomenon.

As regards the content-related factors, recommendation quality was lower than that of the topological metrics. This could be due to the resemblance of the similarity distributions between the actual followees and the randomly selected users, which hinders the possibilities of finding the actual followees as they are all similar. Interestingly, the lowest precision results were obtained in most cases when considering the processed tweets. However, this implies that reducing the syntactic variations of words and only keeping verbs and nouns results in lower similarity values, implying that the higher similarity scores could be due to tweets sharing non-meaningful words and stopwords, instead of being actually content related. The most accurate recommendations were obtained when considering $pub-profile$, followed by $read_{RT-PROC}$, meaning that the published content might be more important for identifying similar followees than the content users have explicitly showed interest. Additionally, recommendation quality based on $read_{RT-FULL}$ was not amongst the best performing profiles, even when it had the highest similarity distribution. Note that $read_{RT-PROC}$ and $read_{RT-FULL}$ had the minimum and maximum statistical coincidence between the similarity distributions of actual followees and non-followed users, respectively. As a result, the selection of the similarity metric should be conditioned by the similarity distribution of not only the actual followees, but also by that of a random population, as the latter could affect recommendation performance.

These results validated the hypothesis that the concept of user similarity has to be carefully analysed as metrics could be biased, and hence not being useful for accurately assessing the relationship between target users and their followees. Moreover, results showed how choosing the wrong metric could affect the recommendation task by hindering the accurate search of potential followees.

Implications

The main goal of this study was to shed some light on the relative importance of different aspects of users' online behaviour, such as social relationships and published content, in the accurate prediction of followees. The findings of this study allowed to verify each of the defined hypotheses, and established the correspondences between the studies over real-world relations and online social networks [36, 57, 58, 60, 61]. The study also allowed to verify the importance of considering the characteristics of the environment in relation to the characteristics of strangers and the similarities towards them to effectively assess the factors guiding the friend selection. Consequently, the performed data analysis showed the existence of patterns between the level of user activity on the micro-blogging site and the characteristics of selected followees.

The findings indicate that tie formation is not a simple process. Instead, it is related to the intrinsic nature and interests of users, and at the same time is conditioned by the environment in which social ties arise. Although ties are built based on common interests, those interests might not be evident or easily distinguished amongst all possible factors. The strength of this study is the performed analysis of the homophilic friendship formation on two levels. First, analysing the factors driving the homophilic relations in connection with the environment and user behaviour. Second, the specific measurement of homophily, i.e. the impact of adequately choosing how to measure user similarity. In turn, this allows to discover with whom users would want to become friends and with whom they actually become friends, which sheds light on the underlying processes.

Several contributions arise from this study. First, the study broadens the analysis of the homophily effect to the context of OSNs, showing that many processes originally described for real-world relationships also hold in online networking sites. The obtained results allow to examine the real-world friendship theories and enrich them. Although numerous studies have been based on the concept of homophily, none of them performed a systematic analysis of such phenomenon and the factors driving it. Second, guidelines for choosing which factors to include in the recommendation system can be derived, as well as how to measure such factors. This is also relevant in terms of the considerations needed to effectively evaluate the performed recommendations. Third, the study regarding the similarity metrics could help to refine existing recommendation algorithms by allowing to adequately measure and weight user similarity. Fourth, as user behaviour was shown to condition the characteristics of selected users by showing that preferences might respond to a combination of the diverse factors (as expressed by Block and Grund [41]), the findings of this study could be used for designing recommendation strategies that combine and adapt the importance of recommendation factors to each users' characteristics. Fifth, the performed analysis allowed to infer that user interests are dynamic and change over time as users share more content and follow more users, implying that the selection of recommendation factors should be also dynamic to cope with the changing behaviour of users. As a result, the findings could be the cornerstone for understanding how users select their followees and thus designing strategies for improving the performance of followee recommendation systems. The implications are not only useful in the context of friend recommendation but also for product recommendation, within friendship networks. Companies can use this findings to design efficient marketing strategies for social media.

Conclusions

Given the exponential number of active users in micro-blogging communities, a careful analysis of the criteria to guide the accurate selection and recommendation of potential followees is crucial. The findings indicate that tie formation is not only related to the intrinsic nature and interests of users, but also conditioned by the surrounding environment. Although ties are built based on common interests, such interests might not be easily distinguished amongst all possible factors. The strength of this study is that it analysed the process of homophilic friendship formation on two levels. First, the factors driving the homophilic ties in relation with the environment and user behaviour. Second, the specific measurement of homophily, i.e. the impact of adequately choosing how to measure user similarity. In turn, this allows to discover with whom users would want to become friends, and with whom they actually become friends.

The performed analysis allowed to verify the proposed hypotheses, and hence answer the research questions guiding the study. The first question focused on whether the formation of social ties was influenced by user similarity. Evidence of similarity between users and their friends was found confirming the existence of homophily amongst them. Additionally, users and their friends were shown to present different similarity patterns according to the diverse factors under analysis, which have distinctive effects over followee selection. This agrees with social theories defined for real-world friendships related to the traits driving tie formation, answering the second research question. Moreover, the study showed a relationship between the characteristics of social networks, and the behaviour and manifestation of user interests when selecting followees, hinting the answer to the third question referring to whether all aspects contribute to strengthen friend homophily. In this regard, the study stated the importance of analysing the level of users' activity and participation for assessing the similarity with other users, and how the definition of user similarity affects the quality of the potentially recommended followees. These findings demonstrate the importance of OSN's characteristics and users' behaviour for performing the best recommendations. Finally, the study shed light on the relationship between users and strangers, and the reasons fostering the similarity coincidences. These results answered the fourth question highlighting the importance of considering the environmental characteristics in terms of strangers and the similarities towards them to effectively assess the factors guiding friend selection.

This work presents some limitations. First, recommendation factors were individually considered. However, users might base their decision of choosing a followee on several and distinctive reasons. As a result, not every followee is relevant according all factors, implying that the importance of each factor varies according to each user's interests and behaviour, as hinted by the performed data analysis. Future works should analyse how to combine the multiple factors. Second, evaluation was only performed on an offline setting in which only positive examples are available (i.e. the actual user followees). In this context, the lack of an explicit relation between two users can be considered as an implicit indication that they are not interested in each other. However, such absence could be due to the fact that users have not yet discovered each other. In such case, even though the recommendation would be still be counted as an incorrect one by precision and hit-rate metrics, it could be appropriate and valuable. The same situation applies for the analysis of the similarity distributions of actual and random non-followed users. Hence, it would be interesting to test the hypotheses in an online environment with explicit feedback from users.

Finally, this study raises interesting questions for future research. First, which is the combined effect of the recommendation factors, i.e. whether combining the factors allows to find other patterns of social ties. Second, whether the findings hold in other similar environments, i.e. whether users in different content-driven social networking sites share the same behavioural tendencies. This study focuses only on one social networking site disregarding the possibility of users having multiple profiles in diverse sites. Hence, it would be interesting to study how users behave in different types of networks. For example, whether users having both *Twitter* and *Facebook* accounts maintain their behaviour across the different networking sites, or they are influenced by the environmental characteristics.

Acknowledgements

This work has been partially funded by ANPCyT, through project PICT-2018-03323.

References

[1] McPherson M, Smith-Lovin L and Cook JM (2001) Birds of a feather: Homophily in social networks. *Annual Review of Sociology* 27(1): 415–444.

[2] Liang H and Fu Kw (2017) Information overload, similarity, and redundancy: Unsubscribing information sources on Twitter. *Journal of Computer-Mediated Communication* 22(1): 1–17.

[3] Montoya RM and Horton RS (2013) A meta-analytic investigation of the processes underlying the similarity-attraction effect. *Journal of Social and Personal Relationships* 30(1): 64–94.

[4] Bisgin H, Agarwal N and Xu X (2012) A study of homophily on social media. *World Wide Web* 15(2): 213–232.

[5] Golder SA and Yardi S (2010) Structural predictors of tie formation in Twitter: Transitivity and mutuality. In: *Proceedings of the 2010 IEEE Second International Conference on Social Computing (SOCIALCOM '10)*. Minneapolis, MN, USA, pp. 88–95.

[6] Bobadilla J, Ortega F, Hernando A and Gutiérrez A (2013) Recommender systems survey. *Knowledge-Based Systems* 46: 109–132.

[7] Singla P and Richardson M (2008) Yes, there is a correlation - From social networks to personal behavior on the Web. In: *Proceeding of the 17th International Conference on World Wide Web (WWW '08)*. Beijing, China, pp. 655–664.

[8] Gilbert E and Karahalios K (2009) Predicting tie strength with social media. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '09)*. pp. 211–220.

[9] Tommasel A, Corbellini A, Godoy D and Schiaffino SN (2016) Personality-aware followee recommendation algorithms: An empirical analysis. *Engineering Applications of Artificial Intelligence* 51: 24–36.

[10] Armentano M, Godoy D and Amandi A (2013) Followee recommendation in Twitter based on text analysis of micro-blogging activity. *Information Systems* 38(8): 1116–1127.

[11] Chen R, Hua Q, Wang B, Zheng M, Guan W, Ji X, Gao Q and Kong X (2019) A novel social recommendation method fusing user's social status and homophily based on matrix factorization techniques. *IEEE Access* 7: 18783–18798.

[12] Fabbri F, Bonchi F, Boratto L and Castillo C (2020) The effect of homophily on disparate visibility of minorities in people recommender systems. In: *Proceedings of the International AAAI Conference on Web and Social Media*. pp. 165–175.

[13] Selfhout M, Denissen J, Branje S and Meeus W (2009) In the eye of the beholder: Perceived, actual, and peer-rated similarity in personality, communication, and friendship intensity during the acquaintanceship process. *Journal of Personality and Social Psychology* 96(6): 1152–1165.

[14] Cuperman R and Ickes W (2009) Big Five predictors of behavior and perceptions in initial dyadic interactions: Personality similarity helps extraverts and introverts, but hurts "disagreeables". *Journal of personality and social psychology* 97(4): 667.

[15] Selfhout M, Burk W, Branje S, Denissen J, van Aken M and Meeus W (2010) Emerging late adolescent friendship networks and Big Five personality traits: A social network approach. *Journal of Personality* 78(2): 509–538.

[16] Tommasel A, Corbellini A, Godoy D and Schiaffino S (2015) Exploring the role of personality traits in followee recommendation. *Online Information Review* 39(6): 812–830.

[17] Tang X, Miao Q, Quan Y, Tang J and Deng K (2015) Predicting individual retweet behavior by user similarity. *Knowledge-Based Systems* 89(C): 681–688.

[18] Zubiaga A, Wang B, Liakata M and Procter R (2019) Political homophily in independence movements: Analyzing and classifying social media users by national identity. *IEEE Intelligent Systems* 34(6): 34–42.

[19] Karimi F, Génois M, Wagner C, Singer P and Strohmaier M (2018) Homophily influences ranking of minorities in social networks. *Scientific Reports* 8.

[20] Chechev M and Georgiev P (2012) A multi-view content-based user recommendation scheme for following users in Twitter. In: *Proceedings of the 4th International Conference on Social Informatics (SocInfo'12), LNCS*, volume 7710. pp. 434–447.

[21] Yin D, Hong L and Davison BD (2011) Structural link analysis and prediction in microblogs. In: *Proceedings of the 20th ACM International Conference on Information and Knowledge Management (CIKM '11)*. Glasgow, UK, pp. 1163–1168.

[22] Valverde-Rebaza JC and de Lopes AA (2012) Structural link prediction using community information on Twitter. In: *Proceedings of the 4th International Conference onComputational Aspects of Social Networks (CASoN 2012)*. pp. 132–137.

[23] De A, Bhattacharya S, Sarkar S, Ganguly N and Chakrabarti S (2016) Discriminative link prediction using local, community, and global signals. *IEEE Transactions on Knowledge & Data Engineering* 28(8): 2057–2070.

[24] Kim K and Altmann J (2017) Effect of homophily on network formation. *Communications in Nonlinear Science and Numerical Simulation* 44: 482–494.

[25] Liang B, Liu Y, Zhang M, Ma S, Ru L and Zhang K (2014) Searching for people to follow in social networks. *Expert Systems with Applications* 41(16): 7455–7465.

[26] Liu Y, Li L, Wang H, Sun C, Chen X, He J and Jiang Y (2018) The competition of homophily and popularity in growing and evolving social networks. *Scientific Reports* 8.

[27] Takhteyev Y, Gruzd A and Wellman B (2011) Geography of Twitter networks. *Social Networks* .

[28] Cuevas R, Gonzalez R, Cuevas A and Guerrero C (2014) Understanding the locality effect in Twitter: Measurement and analysis. *Personal and Ubiquitous Computing* 18(2): 397–411.

[29] Feltoni Gurini D, Gasparetti F, Micarelli A and Sansonetti G (2015) Enhancing social recommendation with sentiment communities. In: *Proceedings of the 16th International Conference on Web Information Systems Engineering (WISE 2015)*. Miami, FL, USA, pp. 308–315.

[30] Aiello LM, Barrat A, Schifanella R, Cattuto C, Markines B and Menczer F (2012) Friendship prediction and homophily in social media. *ACM Transactions on the Web* 6(2).

[31] Xu S and Zhou A (2020) Hashtag homophily in twitter network: Examining a controversial cause-related marketing campaign. *Computers in Human Behavior* 102: 87–96.

[32] Korkmaz G, Kuhlman CJ, Goldstein JD and Vega-Redondo F (2020) A computational study of homophily and diffusion of common knowledge on social networks based on a model of Facebook. *Social Network Analysis and Mining* 10(1).

[33] Šćepanović S, Mishkovski I, Gonçalves B, Nguyen TH and Hui P (2017) Semantic homophily in online communication: Evidence from Twitter. *Online Social Networks and Media* 2: 1–18.

[34] McPherson JM and Smith-Lovin L (1987) Homophily in voluntary organizations: Status distance and the composition of face-to-face groups. *American Sociological Review* : 370–379.

[35] Thelwall M (2009) Homophily in MySpace. *Journal of the American Society for Information Science and Technology* 60(2): 219–231.

[36] Verbrugge LM (1977) The structure of adult friendship choices. *Social Forces* 56(2): 576.

[37] Chen YF (2014) See you on Facebook: exploring influences on Facebook continuous usage. *Behaviour & Information*

Technology 33(11): 1208–1218.

[38] Hannon J, Bennett M and Smyth B (2010) Recommending Twitter users to follow using content and collaborative filtering approaches. In: *Proceedings of the 4th ACM Conference on Recommender Systems (RecSys'10)*. Barcelona, Spain, pp. 199–206.

[39] Naruchitparames J, Güneş MH and Louis SJ (2011) Friend recommendations in social networks using genetic algorithms and network topology. In: *IEEE Congress on Evolutionary Computation*. pp. 2207–2214.

[40] Dong Y, Johnson RA, Xu J and Chawla NV (2016) Structural diversity and homophily: A study across more than one hundred large-scale networks. *CoRR* abs/1602.07048.

[41] Block P and Grund T (2014) Multidimensional homophily in friendship networks. *Network Science* 2(02): 189–212.

[42] Liben-Nowell D and Kleinberg J (2003) The link prediction problem for social networks. In: *Proceedings of the 12th International Conference on Information and Knowledge Management (CIKM '03)*. New Orleans, LA, USA, pp. 556–559.

[43] Weller K, Bruns A, Burgess J and Mahrt M (2013) *Twitter and Society*. Switzerland: Peter Lang International Academic Publishers.

[44] Choudhury MD, Lin YR, Sundaram H, Candan KS, Xie L and Kelliher A (2010) How does the data sampling strategy impact the discovery of information diffusion in social media? In: *Proceedings of the 4th International AAAI Conference on Weblogs and Social Media (ICWSM'10)*.

[45] Corder GW and Foreman DI (2009) *Nonparametric Statistics for Non-Statisticians: A Step-by-Step Approach*. New Jersey: Wiley.

[46] Porter M (1997) *Readings in information retrieval*, chapter An algorithm for suffix stripping. Morgan Kaufmann Publishers Inc., pp. 313–316.

[47] Lü L and Zhou T (2011) Link prediction in complex networks: A survey. *Physica A: Statistical Mechanics and its Applications* 390(6): 1150–1170.

[48] Romero DM and Kleinberg JM (2010) The directed closure process in hybrid social-information networks, with an analysis of link formation on Twitter. In: *Proceedings of the 4th International Conference on Weblogs and Social Media (ICWSM 2010)*. Washington, DC, USA.

[49] Liu Z, Liu L and Li H (2012) Determinants of information retweeting in microblogging. *Internet Research* 22(4): 443–466.

[50] Salton G and McGill MJ (1983) *Introduction to modern information retrieval*. New York: McGraw-Hill.

[51] Tukey JW (1977) *Exploratory Data Analysis*. Addison-Wesley.

[52] Gillani N, Yuan A, Saveski M, Vosoughi S and Roy D (2018) Me, my echo chamber, and I: Introspection on social media polarization. In: *Proceedings of the 2018 World Wide Web Conference (WWW '18)*. Lyon, France, p. 823–831.

[53] Choudhury MD (2011) Tie formation on Twitter: Homophily and structure of egocentric networks. In: *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing*. pp. 465–470.

[54] Fani H, Bagheri E and Du W (2017) Temporally like-minded user community identification through neural embeddings. In: *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management (CIKM '17)*. Singapore, Singapore, p. 577–586.

[55] Dey K, Shrivastava R, Kaushik S and Garg K (2019) Assessing topical homophily on Twitter. In: *Complex Networks and Their Applications VII*. pp. 367–376.

[56] García-Martín J and García-Sánchez JN (2015) Use of Facebook, Tuenti, Twitter and MySpace among young Spanish people. *Behaviour & Information Technology* 34(7): 685–703.

[57] Feld SL (1981) The focused organization of social ties. *American Journal of Sociology* 86(5): 1015–1035.

[58] Feld S and Grofman B (2009) Homophily and the focused organization of ties. *The Oxford Handbook of Analytical Sociology* : 521–543.

[59] Feld SL (1982) Social structural determinants of similarity among associates. *American Sociological Review* 47(6): 797–801.

[60] Antheunis ML, Valkenburg PM and Peter J (2015) The quality of online, offline, and mixed-mode friendships among users of a social networking site. *Cyberpsychology: Journal of Psychosocial Research on Cyberspace* 6(3).

[61] Launay J and Dunbar RI (2015) Does implied community size predict likeability of a similar stranger? *Evolution and Human Behavior* 36(1): 32–37.