Self-Similarity of Twitter Users

Masoud Fatemi
Linnaeus University
Växjö/Kalmar, Sweden
University of Eastern Finland
Kuopio/Joensuu, Finland
masoud.fatemi@uef.fi

Kostiantyn Kucher
Linnaeus University
Växjö/Kalmar, Sweden
kostiantyn.kucher@lnu.se
Linköping University
Norrköping, Sweden

Mikko Laitinen
Linnaeus University
Växjö/Kalmar, Sweden
mikko.laitinen@lnu.se
University of Eastern Finland
Kuopio/Joensuu, Finland

Pasi Fränti
University of Eastern Finland
Kuopio/Joensuu, Finland
pasi.franti@uef.fi

2021 Swedish Workshop on Data Science (SweDS) | 978-1-6654-1830-0/21/$31.00 ©2021 IEEE | DOI: 10.1109/SweDS53855.2021.9638288

***Abstract*—Earlier studies have established that the (perceived) similarity of users is highly subjective and reflects more on how people respect/admire others rather than their characteristics or behavioral similarities. We study this phenomenon among Twitter users, and while confirm that it is indeed the case, we further explore the components of similarity by investigating it using data from three categories (interactions between egos and alters, profile-based activity history, and linguistic content in the messages). We use interactions as estimation for admiration and observe that it has more impact and a higher correlation to the perceived similarity than other objective measures, including similarity based on user profiles and their use of hashtags.**

***Index Terms*—Social network analysis; Ego network; User similarity; Users interactions; Activity history.**

I. Introduction

We investigate user similarity in social media. The broad framework is such that *social media* has opened up to be a big and rich source of data [1], [2]. Analyzing this massive data with manual, computational, or interactive methods can lead to novel insights that can be employed in a variety of applications and fields in fundamental research [2], [3]. The notions of social networks, together with social media and *social network analysis* (SNA), offer powerful models and approaches for understanding social structures [4]. Social network analysis embraces a wide range of applications and can be used in different domains from internet applications of location-aware recommendation [5] to understanding behavior patterns of large numbers of individuals in social sciences and the humanities, where the dynamics of interactional behavior can substantially broaden the evidence based on fixed social categories. The underlying idea of SNA is to establish user similarities to identify the most similar individuals through various interactional and social factors [6], [7].

Previous literature identifies three categories of individual similarity that in some cases may lead two individuals who were initially unacquainted to establish a connection and initiate interaction in a social network [8]. First, *self-view* similarity is a *dyadic* method that indicates how similar two individuals are according to self-ratings. The second category is the *perceived similarity*. Opposite to the self-view, the perceived similarity is an *idiosyncratic* mode that quantifies the similarity between two individuals based on a specific trait according to their perceptions. The last one is *peer-view* similarity, a *group* approach, in which peer views are used to quantify two individuals' similarity on a specific trait [8].

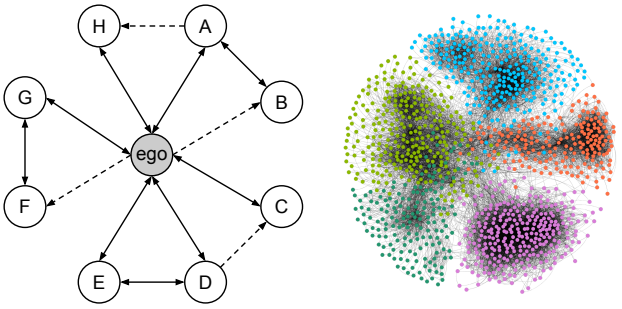
As discussed in Section I-C, we adopt the perceived-similarity approach. We ask Twitter users to provide a list of accounts that are similar to themselves and then compare this list to large user-generated data of interactions. The objective is to identify which interaction category most effectively predicts similarity.

A. Twitter Ego Networks

A literature review from computer science and social anthropology reveals that *ego networks* are the cornerstones in studying social networks [9], [10]. They are the primary structural characteristics of individual networks [9]. Indeed, the concept of an ego network is essential when identifying key features of human behavior. Depending on the application of interest or methods in analyzing ego networks, a range of definitions appear in past literature [11]. As illustrated in Fig. 1, we define an ego network to consist of a single individual or an account (*ego*) and the other accounts directly connected to the ego (*alters*) and the links between alters [12].

As primary material in the empirical part, we use mutual interaction data and user profile information obtained from Twitter. It is a micro-blogging and social network application that enables users to share text (up to 280 characters excluding URLs, mentions, and hashtags), photos, videos, and voice messages [13]. A social network on Twitter incorporates an *ego node*, *those followed*, and *those who follow*. The ego is an account (a node) that has a direct connection to all of the other accounts inside the network. Those followed or *friends* are accounts that an individual (ego) is following, while followers are the ones who follow the ego node. Twitter users can generate content and maintain interaction with their social networks and other accounts via *tweets*, *mentions (replies)*, and *retweets* that users post. Mentions or replies refer to the response of other users to someone's tweet. When retweeting, it is possible to add text or other modalities to the original tweet (retweet with a quotation).

Twitter ego networks are directed graphs in terms of friends and followers. Fig. 1 demonstrates two sample ego networks. Fig. 1(a) represents a dummy ego network with 9 accounts and 22 links. Fig. 1(b) illustrates five real and very large ego networks with interconnections (1,220 nodes and 19,139



(a) A dummy ego net (9 accounts and 22 edges).

(b) Five real ego networks from Twitter and their interconnections (1,220 nodes and 19,139 edges).

Fig. 1. Two examples of social networks.

edges). While a Twitter network consists of both friends and followers, we are more interested in friends networks than the whole or follower networks. As argued in [14], a following (friend) tie is, from a social and interactional perspective, a slightly stronger than a follower tie. The reason is that to become a friend with someone on Twitter, opposite to having a follower, users make some efforts (find and follow) [14].

B. User Similarity on Twitter

The majority of past methods of similarity analysis and those determining similarity profiles are based on either individual ego networks or contents that users post on Twitter. Since the objective is to focus on perceived similarity, we consider both the user-generated data of interactions and linguistic content as well as activity history. Earlier studies suggest that the (perceived) similarity of users in social media is extremely subjective, and each user might have his/her own interpretation of the phenomenon [6], [15]. In other words, instead of considering similarity based on specific traits or behavioral characteristics, previous studies of perceived similarity aim at measuring the extent to which users admire others rather than observing what interactional components actually contribute to similarity [16].

Evaluating literature on similarity analysis, and especially those detecting similar users in networks, suggests that measures for similarity analysis among social network users have been extensively focused on in the fields of information retrieval and graph theory [17], [18]. Determining similar users can be accomplished utilizing either data generated by users in networks or benefiting from models and techniques in the graph theory, such as centrality analysis, sub-graph isomorphism, and community detection [19].

In [20], Zhang et al. utilized textual data generated by Twitter users to identify communities within the networks. The authors applied this idea assuming that users who reside in the same community can be considered similar to each other. However, considering just one data modality without considering the accounts that generated the data, such as troll accounts, might yield unreliable results. The authors in [21], via characterizing the Twitter friends and followers concepts as out-degree and in-degree, defined a graph structure to analyze user behavior under the category of graph analytic techniques. However, the authors in [21] mainly concentrated on tweeting patterns on Twitter rather than detecting similar users.

Dib et al. in [22] proposed a model to detect similar users for followee recommendations. Their model utilizes lexical and semantic analysis to extract features from the content posted on profiles. Later, using a topology-based candidate search that was made for the user of interest, the authors developed a network to stream tweets. Applying a semantic analysis to the tweets and calculating the similarities was the next step in their user recommendations [22]. In 2020, Sridhar and Sanagavarapu in [23] proposed an account recommender model, in which the idea was to construct a social interaction network based on the similarity of tweet content. They extracted features via a semantic analysis and applied a hypernym feature engineering method to improve the quality of the features. Later, the authors utilized the k-nearest neighbors model to evaluate the similarity of tweets to be used when recommending accounts to be followed [23].

Orlandi et al. in [24] focused on user profiling techniques. These techniques are mainly used for expressing knowledge of users and their interests to provide personalised profile recommendations, and in [24] the authors proposed a method to automatically create user profiles by utilizing semantic techniques. TSim [25], which was proposed in 2018 by AlMahmoud and Al-Khalifa, is another model for identifying and investigating similarity of Twitter users based on their social interactions. TSim considers both friends and followers, while we are inclined to believe that a friends network is a stronger network than a friends plus followers network as it reduces the likelihood of including strangers or bot/troll accounts which aim at superficially conflating the network size [14].

There is an abundance of research on user similarity on social networks [26]–[28]. However, there is a lack of knowledge of which factors, such as user interactions, can affect the user similarity problem in network studies. Additionally, it is still unclear how these factors influence similarity, and whether the best factor's predictive accuracy is good enough to be employed in practice. This paper **aims** to analyze and evaluate the users perceived similarity problem in Twitter networks with respect to three features. These features are obtained from user-generated data, their activity history, and mutual interaction that users have with their social networks. To provide a comprehensive investigation of the perceived similarity problem from different perspectives, this paper not only examines the content created by Twitter users plus their interactions within the networks, but also encodes the patterns upon which they generate content and interact with others.

We define an individual's ego network as all the links that he or she directly establishes with alters [29], and we focus on detecting the most similar alters to an individual. The underlying idea comes from social sciences and assumes that people tend to build communities that supply a meaningful framework in their quotidian life [30]. Another key assumption is that people

attempt to maintain interaction with people they appreciate or respect more, and the hypothesis is that they consider the same people most similar to themselves. Utilizing user interaction in networks, profile information, and the use of textual material (hashtags), we aim at answering the following **questions**: First, to what extent can user-generated data on Twitter be employed for investigating the similarity between users? Second, what user-generated data most effectively predicts similarity?

To fulfill the research **objectives**, we first designed an online survey to collect ground-truth data from Twitter users. Next, we streamed user-generated data through the Twitter API for those accounts that had been mentioned in the survey. We then developed a quantitative model to analyze the similarity of Twitter users employing these data. Finally, via evaluating the results, we try to answer the research questions.

C. Data

We collected our data directly from Twitter via connecting to the Twitter API using Python in two phases. First, considering the concept of perceived similarity, defined as quantifying the similarity between two individuals based on perception, we prepared an online survey and advertised it, and asked Twitter users to list the usernames of the 10 accounts most similar to themselves[1]. The respondents could freely decide on the similarity criteria. Second, we retrieved all the available data from the networks of those who filled in the survey. Depending on the number of accounts in a network and the amount of data, the collection time varied considerably. Table I summarizes the collected data statistics. In total, we collected 16,816,460 tweets, retweets, mentions, and quotations (up to 3,200 most recent items) from 14 ego networks and 8,744 accounts. The difference between the number of friends ('Friends') and the size of the retrieved networks ('Networks') is because of the private accounts whose data cannot be accessed by anyone outside the network.

The rest of this paper is organized as follows. Section II introduces our approaches for investigating the similarity of Twitter users. Section III evaluates our measures with real data that we streamed and collected directly from Twitter, and Section IV concludes the paper.

II. Detecting Similar Users in Social Network

We extract and analyze user similarity via three approaches and then compare the result with the ground truth data that we collected from the similarity survey. We utilize activity history, user-generated data, and the interaction that the ego nodes had with their social networks to identify the most similar accounts to themselves. Then, we compare these results with the list of most similar individuals from the survey. Fig. 2 presents an overview of our approach and the three computational perspectives employed to extract and detect similar users.

[1]The survey is available at: http://cs.uef.fi/~fatemi/usersimilarity

TABLE I
Twitter Data Statistics

#	Gender	Friends	Networks	Tweets	Retweets	Mentions	Quotations
1	Male	167	165	83,490	1,529	161,059	20,060
2	Male	118	110	38,333	429	65,316	6,680
3	Female	305	258	104,298	1,993	212,724	19,628
4	Female	197	191	103,392	1,683	232,960	27,130
5	Female	319	298	165,319	1,987	322,984	45,349
6	Female	3,856	3,790	2,265,328	32,910	4,999,656	840,223
7	Female	987	905	291,908	6,746	1,059,173	184,749
8	Male	542	515	191,743	3,884	515,107	90,254
9	Female	453	381	155,659	3,146	402,642	90,538
10	Male	468	457	407,942	5,001	460,663	55,790
11	Male	1013	881	528,451	11,949	943,800	70,143
12	Male	236	203	195,213	3,456	220,382	25,919
13	Male	261	260	205,232	2,482	280,609	33,548
14	Male	333	330	212,531	2,338	368,144	32,858
	8 M / 6 F	**9,255**	**8,744**	**4,948,839**	**79,533**	**10,245,219**	**1,542,869**

Note: We anonymized all the accounts, due to the privacy preservation.

A. Interaction-Based Similarity

Our first approach is tied to the interactions that users establish and maintain in their social networks. As pointed out, a Twitter interaction occurs and may continue once an account holder interacts (e.g. replies to a tweet or retweets content) with content from another account. The underlying idea is that if two users have more interactions than other accounts, then the probability is high that they are more similar in some traits than others, viz. they might have similar interests, or be interested in similar topics. Note that this does not mean that the two nodes would have to behave similarly, since they might, for instance, have opposing political views, in which case similarity consists of shared interest in politics. It goes without saying that similarity can consist of anything, ranging from personality traits to individual views or social factors.

Fig. 3(a) demonstrates how we extracted the interactions from an ego node (John Doe) and the nodes in his network. Fig. 3(a) illustrates that this Twitter user has 165 alters (friends) and 3,221 messages (tweets + retweets + mentions + quotations). To extract the list of most frequent interactions, bottom table in Fig. 3(a), we decomposed the ego node messages and computed how many interactions this ego had had with each of his friends. After that, we ranked the

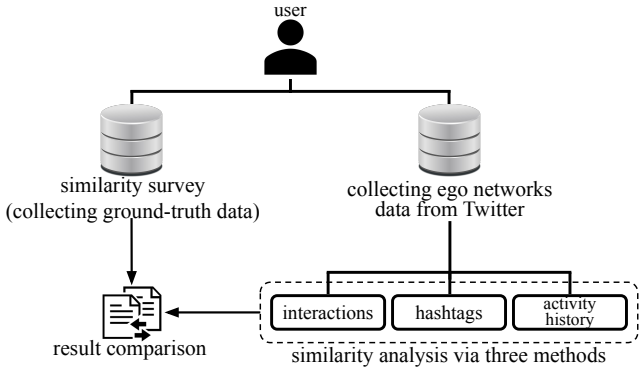


Fig. 2. Overview of the proposed model.

accounts based on frequency and selected the top 10 ones. This procedure results in two lists of Twitter users, the first one provided by the user (John Doe) using the online survey, and the second one is the result of extracting his recent social interactions in the application. From the mathematical point of view, these lists are considered two sets of distinct entities, and the similarity analysis can be accomplished by calculating the set similarity.

Based on our empirical results, we argue below in Section III-A that interaction-based similarity seems to be a superior method for measuring similarity over the other methods of the profile activity history or hashtags.

B. Profile-Based Similarity

The second approach quantifies Twitter users' activity history. Table II shows that we extract a set of activity-based features (seven features), and then utilize them to create user profiles for each account and thus all the nodes in the network.

The first calculated feature is *Age*, and it equals to the number of days that an account has been active until our data collection. *Tweet* indicates the total number of tweets (including retweets and replies) that an account has published. As the third feature, we applied the idea of [31] to compute the *Reputation* for each account and giving insight of the credibility of a user. Based on the formula in Table II, the reputation value for verified Twitter accounts, such as celebrities and politicians, is close to zero, since there is a drastic difference between the number of friends and followers for these verified accounts. *Favorite* indicates the total number of times an individual likes others' tweets. *Tweet rate* is the fifth feature and indicating the average number of tweets that an account publishes per day. The last two features are related to the hashtags that users integrate into their messages. The *Hashtags* category represents the total number of unique hashtags (types) that an account has used so far, and *Hashtag density* indicates the number of hashtags (tokens) per tweet of an account (taking all the hashtag tokens into account instead of considering the unique types).

Fig. 3(b) indicates profiles that we built from John Doe's ego network using the features that we extracted from his activity history. After building profiles, to scale the extracted features and to make them comparable, we apply a min-max normalization and transform the values into the [0, 1] interval.

Finally, we calculate the distance between the built profiles by calculating *Euclidean distance* [32] between the respective profile vectors consisting of the 7 dimensions (features) listed in Table II.

We assume that accounts with the same activity patterns ought to be more similar to each other than those with differing activities. Consequently, we rank one's friends based on the distance that was calculated using the activity profiles. The lower the distance between two profiles, the more similar these profiles are considered to be.

C. Hashtag-Based Similarity

The third method to extract and analyze Twitter user similarity involves hashtags. They are utilized as tags or topics for tweets, and users attach them to tweets to showcase the topics discussed. In detail, social media users take advantage of hashtags as labels to indicate succinctly what is being written, and they always begin with the '#' sign. Tables III and IV visualize two dummy hashtag sets with their frequencies. The idea is that if two individuals regularly use similar sets of hashtags and share a substantial number of hashtags, these individuals are probably more similar to each other than those whose hashtag similarity is lower. That is, if two people are similar in some specific traits, they will probably care, chat, and write about similar topics [20].

The first step in this part consists of extracting all the hashtags. For the 14 ego nodes and their alters, we collect the hashtags and their frequencies. Next, using Formula 1 we calculate the similarity between each ego node and its alters and rank them.

$$Sim(A,B) = \frac{\sum\{min(n_a, n_b) | n \in (A \cap B)\}}{total\ number\ of\ hashtags} \quad (1)$$

Here, A and B are two tag sets (tags and their frequencies such as Tables III and IV) that belong to two users, and n_a and n_b are the frequencies of a specific hashtag in A and B, respectively. For instance, for the two hashtag sets in Tables III and IV, using Formula 1, the similarity value will be: $Sim(A,B) = (1+7)/(16+17) \simeq 0.24$.

Fig. 4 presents a more comprehensive illustration of the ground-truth data (Survey) that a Twitter user, such as John Doe, provided and three lists that we extracted from his network content according to the analysis of interactions, hashtags, and activity profiles. This comparison shows that there are 6, 3, and 0 similar users between the Survey data and the Interactions, Hashtags, and Profiles respectively. Applying *Jaccard similarity coefficient* (*JSC*) calculations [32], [33], the similarity values for John Doe's ego network, which take into account interactions, hashtags, and activity history are 33.3, 17.6, and 0. For John Doe's ego network, the interaction category turns out to be the best way to identify similar users among the nodes in his network.

III. Empirical Results

A. Detecting Similarity of Twitter Users

Table V shows the results of our methods for Twitter user similarity analysis. We extracted the similar account lists based on the social interactions, hashtags, and activity history for all the 14 ego networks shown in Table I using the procedures

TABLE II
ACTIVITY-BASED FEATURES

Features	Description
Age (days)	$age = present\ day - created\ day$
Tweet	the total number of tweets
Reputation	$reputation = \frac{friends}{friends+followers}$
Favorite	the total number of likes
Tweet rate	$tweet\ rate = \frac{tweets}{age}$
Hashtags	the total number of unique hashtags
Hashtag density	the total number of hashtags per tweet

	Friends	Tweets	Retweets	Mentions	Quotations
John Doe	165	1,206	9	1,898	108

	Account	Retweets	Mentions	Quotations	Total
1	Candi Lile	4	20	5	29
2	Lawan Moad	0	13	3	16
3	Mona Meder	1	10	4	15
⋮					
164	Tad Pugh	2	8	2	12
165	Jose Sly	1	4	3	8

(a)

	Accounts	Age (days)	Tweets	Favorites	Reputations	Tweets rate	Hashtags	Hashtags densities
1	John	3,334	5,269	44,031	0.63	1.58	127	0.035
2	Candi	3,381	2,256	3,098	0.31	0.67	194	0.138
3	Tad	3,128	18,944	27,090	0.50	6.07	26	0.002
⋮								
166	Jose	1,477	6,335	95,194	0.77	4.29	19	0.003

(b)

Fig. 3. Examples of (a) extracted interactions from an ego network, and (b) how profiles are built using the introduced activity-based features.

TABLE III
A: JOHN'S HASHTAGS

Hashtags	Frequencies
SoundCloud	1
Covid	11
Mdpi	1
HR1044	3
Total	**16**

TABLE IV
B: CANDI'S HASHTAGS

Hashtags	Frequencies
Worldcup	3
Covid	7
twitter	1
FIVB	4
SoundCloud	2
Total	**17**

introduced in Sections II-A, II-B, and II-C. Next, we sorted each list into a descending order and selected the top 10 accounts for the final stage. Lastly, applying *JSC* [32], [33], we calculated the similarity values for social interactions, hashtags, and activity history between the ground-truth lists that the ego nodes had provided and the extracted lists.

As Table V demonstrates, the interaction-based similarity measurement has the highest accuracy on average (19.2%), much higher than the hashtag-based (7.7%) and the activity-based (1.9%) similarity analyses. The calculated values for different approaches suggest that the analysis of ego node interactions with their friends turns out to be the most effective way to locate the similar users in a network and outperforms the other methods that are used here.

B. Effect of Network Size

The authors in [34] discussed the idea that social network size is an important aspect in network and that size plays an important, yet understudies, role in various fields, including social media technological design, sociology, and so on. In addition, it has been argued that larger social networks (in terms of the number of nodes in the network) might be more beneficial than smaller ones, because size brings in the potential of having more nodes that can carry more information and thus increase the diversity of social contacts [34]. In this regard, we conducted an additional investigation to evaluate the effect of ego network size on the user similarity problem. Using *Pearson correlation analysis* [32], [35], we calculated the linear correlation between the results of our three approaches and the network sizes. The correlation values of the network sizes and similarity categories (interaction, hashtags, and activity) are -0.65, -0.22, and -0.23, respectively. There is a linear correlation between the three approaches and the network size values, and the negative coefficient values suggest that the accuracy of the methods decreases when increasing the number of nodes in networks. In other words, we can find users that are similar to the ego more effectively in smaller networks than in larger ones.

Fig. 5 visualizes the correlation analysis results. The 14 ego networks were sorted based on the size of the networks ('Networks' column in Table I), and then we plotted (Fig. 5) the network sizes against the similarity values for each network (see Table V). As we mentioned earlier, the interaction-based

Survey	Interactions	Hashtags	Profiles
1. Tad Pugh	1. Dion Marcella	1. Lawana Moad	1. Gavin Miedema
2. Mona Madere	2. Candi Lile	2. Nellie Branson	2. Savannah Craver
3. Candi Lile	3. Tad Pugh	3. Carley Samson	3. Darren Brobst
4. Lawana Moad	4. Rolf Weishaupt	4. Shawna Cifaldi	4. Afton Merchant
5. Quentin Warthen	5. Kathleen Narciso	5. Mona Madere	5. Beth Brummer
6. Kathleen Narciso	6. Jose Sly	6. Dean Izzard	6. Chrissy Haan
7. Carley Samson	7. Les Brumett	7. Kena Mccraney	7. Lucina Calvillo
8. Maryeta Webster	8. Laronda Mayers	8. Sharyl Amaker	8. Lucile Rosas
9. Jose Sly	9. Stephnie Edward	9. Twanda Heyman	9. Inocencia Hess
10. Stephnie Edward	10. Doglas Vanauken	10. Jose Sly	10. Darline Brazzell

Fig. 4. Example results of similar accounts extraction.

TABLE V
SIMILARITIES CALCULATED VIA THREE PROPOSED METHODS FOR THE 14 COLLECTED EGO NETWORKS

#	Gender	Interactions (%)	Hashtags (%)	Profiles (%)
1	Male	33.3	17.6	0.0
2	Male	11.1	5.3	11.1
3	Female	11.1	17.6	5.3
4	Female	25.0	11.1	0.0
5	Female	25.0	5.3	0.0
6	Female	0.0	5.3	0.0
7	Female	11.1	5.3	0.0
8	Male	17.6	11.1	5.3
9	Female	17.6	0.0	0.0
10	Male	25.0	0.0	0.0
11	Male	5.3	0.0	0.0
12	Male	33.3	11.1	0.0
13	Male	25.0	17.6	0.0
14	Male	25.0	0.0	5.3
avg.		19.2	7.7	1.9

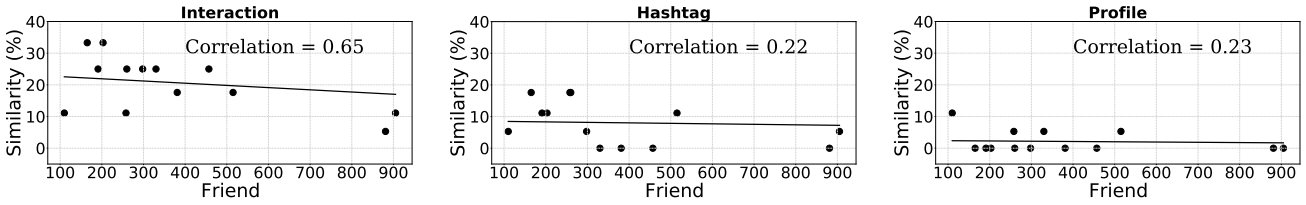


Fig. 5. Correlation analysis between the size of ego networks (number of friends) and the accuracy of three proposed approaches for measuring user similarity. The correlation values of the network sizes and similarity categories (interaction, hashtags, and activity) are -0.65, -0.22, and -0.23, respectively.

similarity has the highest average value for calculating the user similarity, which is superior when compared with the other two approaches; its absolute correlation value is also the largest among the three approaches ($|\rho| = 0.65$). In other words, when compared with the hashtag-based and the activity-based similarities, the interaction-based similarity decreases more substantially when the ego network size increases.

C. Male Ego Networks vs. Female Ego Networks

Out of the 14 ego networks that we collected from Twitter, eight identify as males (male ego nodes), and the rest as females (female ego nodes). We compared the average similarity values calculated in Table V for the male against female ego networks. As shown in Fig. 6, the male ego networks result in higher values than females for all the current methods. Moreover, adding this additional category does not change the overall result. For both the male and female egos, using the interaction-based similarity measure between the ego and the alters results in the highest accuracy.

IV. Conclusion

We have focused on user similarity in social networks and particularly on Twitter and have utilized a set of particular methods to measure similarity. The empirical part analyzed how effectively these methods can be used to measure Twitter user similarity. The proposed methods aimed at investigating the extent to which various factors that are directly observable in user-generated data, such as users interactions, can affect the user similarity problem in a directed graph network. We also assessed the impact of these factors as possible predictors to measure which one is the most influential when it comes to the user similarity problem.

We employed actual user interactions, hashtags in user posts, and individual activity history as three approaches to extract features and to measure user similarity between the ego node and the alters. The results indicate that utilizing user interactions has the most impact on the similarity problem and the highest accuracy for predicting similar users. Based on our observations, we propose that user-generated data on Twitter, and especially deep network interaction data, can be employed to identify the most similar users and user groups. This information can be further utilized in social network analyses in other fields, such as sociolinguistics that focuses on how language variation and change is embedded in the social structures in which it is used [30], [36], [37]. What is more, we investigated that the size of ego networks can slightly affect the accuracy of the similarity problem in networks. In more detail, there is a negative linear correlation between the proposed method and the size of ego networks. That is, by increasing the network size, we witness decreasing accuracy in locating similar users. Additionally, we examined the effect of gender (male egos vs. female egos) on the accuracy of identifying similar users. The observations suggest that the accuracy is higher for each of the three proposed methods when the ego node is male, and thus, further improvements concerning female accounts are required in the future.

Our plans for the future work include improvements of our approaches to increase the accuracy of the methods and to account for particularly challenging cases and scenarios discussed below, for instance, regarding the effect of network size and account holder's gender. In addition, combining our computational approaches with interactive visual analyses of social networks [38] and social media [39] to facilitate research in sociolinguistics is also part of our plans [40].

Acknowledgment

This research acknowledges the funding from the Center for Data Intensive Sciences and Application (DISA) at Linnaeus University. DISA has enabled this multidisciplinary work by bringing together people from various fields.

References

[1] C. Lin, Y. He, R. Everson, and S. Ruger, "Weakly supervised joint sentiment-topic detection from text," *IEEE Transactions on Knowledge and Data Engineering*, vol. 24, no. 6, pp. 1134–1145, Jun. 2012.

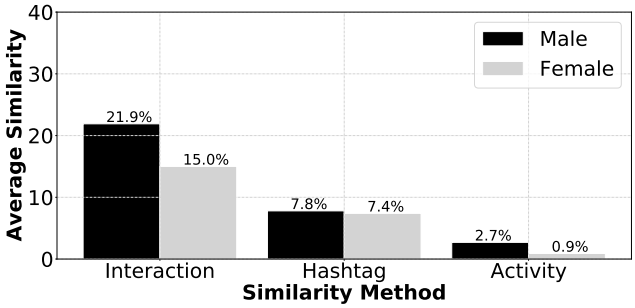


Fig. 6. A comparison of male vs. female accounts w.r.t similarity accuracy.

[2] K. Kucher, R. M. Martins, C. Paradis, and A. Kerren, "StanceVis Prime: Visual analysis of sentiment and stance in social media texts," *Journal of Visualization*, vol. 23, no. 6, pp. 1015–1034, Dec. 2020.

[3] M. Fatemi and M. Safayani, "Joint sentiment/topic modeling on text data using a boosted restricted Boltzmann machine," *Multimedia Tools and Applications*, vol. 78, no. 15, pp. 20 637–20 653, Aug. 2019.

[4] J. Scott, "Social network analysis," *Sociology*, vol. 22, no. 1, pp. 109–127, Feb. 1988.

[5] P. Fränti, K. Waga, and C. Khurana, "Can social network be used for location-aware recommendation?" in *Proceedings of the International Conference on Web Information Systems and Technologies — Volume 1: WEBIST*, ser. WEBIST '15, INSTICC. SciTePress, 2015, pp. 558–565.

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

[7] M. E. Newman and M. Girvan, "Finding and evaluating community structure in networks," *Physical Review E*, vol. 69, no. 2, p. 026113, Feb. 2004.

[8] M. van Zalk and J. Denissen, "Idiosyncratic versus social consensus approaches to personality: Self-view, perceived, and peer-view similarity," *Journal of Personality and Social Psychology*, vol. 109, no. 1, pp. 121–141, Jul. 2015.

[9] R. Dunbar, *Grooming, Gossip, and the Evolution of Language*. Harvard University Press, 1996.

[10] V. Arnaboldi, A. Passarella, M. Conti, and R. I. Dunbar, "Chapter 5: Evolutionary dynamics in Twitter ego networks," in *Online Social Networks*, ser. Computer Science Reviews and Trends. Elsevier, 2015, pp. 75–92.

[11] V. Arnaboldi, M. Conti, M. La Gala, A. Passarella, and F. Pezzoni, "Ego network structure in online social networks and its impact on information diffusion," *Computer Communications*, vol. 76, pp. 26–41, Feb. 2016.

[12] R. Burt, "Structural holes versus network closure as social capital," in *Social Capital: Theory and Research*. Routledge, 2001.

[13] I. Eleta and J. Golbeck, "Multilingual use of Twitter: Social networks at the language frontier," *Computers in Human Behavior*, vol. 41, pp. 424–432, 2014.

[14] M. Laitinen, J. Lundberg, M. Levin, and A. Lakaw, "Revisiting weak ties: Using present-day social media data in variationist studies," in *Exploring Future Paths for Historical Sociolinguistics*. John Benjamins Publishing Company, 2017, pp. 303–325.

[15] J. Weng, E.-P. Lim, J. Jiang, and Q. He, "Twitterrank: finding topic-sensitive influential twitterers," in *Proceedings of the third ACM international conference on Web search and data mining*, 2010, pp. 261–270.

[16] R. M. Montoya, R. S. Horton, and J. Kirchner, "Is actual similarity necessary for attraction? a meta-analysis of actual and perceived similarity," *Journal of Social and Personal Relationships*, vol. 25, no. 6, pp. 889–922, 2008.

[17] A. Goel, A. Sharma, D. Wang, and Z. Yin, "Discovering similar users on Twitter," in *Proceedings of the Workshop on Mining and Learning with Graphs*, ser. MLG '13, 2013.

[18] V. Kuikka, "Terrorist network analyzed with an influence spreading model," in *Complex Networks IX*. Springer, 2018, pp. 185–197.

[19] W. M. Campbell, C. K. Dagli, and C. J. Weinstein, "Social network analysis with content and graphs," *Lincoln Laboratory Journal*, vol. 20, no. 1, pp. 61–81, Jan. 2013.

[20] Y. Zhang, Y. Wu, and Q. Yang, "Community discovery in Twitter based on user interests," *Journal of Computational Information Systems*, vol. 8, no. 3, Feb. 2012.

[21] Q. Yan, L. Wu, and L. Zheng, "Social network based microblog user behavior analysis," *Physica A: Statistical Mechanics and its Applications*, vol. 392, no. 7, pp. 1712–1723, 2013.

[22] B. Dib, F. Kalloubi, E. H. Nfaoui, and A. Boulaalam, "Semantic-based followee recommendations on Twitter network," *Procedia Computer Science*, vol. 127, pp. 505–510, 2018.

[23] S. Sridhar and S. Sanagavarapu, "Twitter social networking graph using hypernym based semantic similarity detection," in *Proceedings of the International Conference on Smart Electronics and Communication*, ser. ICOSEC '20. IEEE, 2020, pp. 28–35.

[24] F. Orlandi, J. Breslin, and A. Passant, "Aggregated, interoperable and multi-domain user profiles for the social web," in *Proceedings of the International Conference on Semantic Systems*, ser. I-SEMANTICS '12. ACM, 2012, pp. 41–48.

[25] H. AlMahmoud and S. Al-Khalifa, "TSim: A system for discovering similar users on Twitter," *Journal of Big Data*, vol. 5, no. 39, Oct. 2018.

[26] C. G. Akcora, B. Carminati, and E. Ferrari, "Network and profile based measures for user similarities on social networks," in *Proceedings of the IEEE International Conference on Information Reuse & Integration*, ser. IRI '11. IEEE, 2011, pp. 292–298.

[27] V. Kuikka, "Influence spreading model used to community detection in social networks," in *International Conference on Complex Networks and their Applications*. Springer, 2017, pp. 202–215.

[28] A. Tommasel and D. Godoy, "Influence and performance of user similarity metrics in followee prediction," *Journal of Information Science*, 2020.

[29] M. Laitinen, M. Fatemi, and J. Lundberg, "Size matters: Digital social networks and language change," *Frontiers in Artificial Intelligence*, vol. 3, p. 46, Jul. 2020.

[30] L. Milroy and C. Llamas, "Social networks," in *The Handbook of Language Variation and Change*, 2nd ed. John Wiley & Sons, Ltd, 2013, ch. 19, pp. 407–427.

[31] J. Lundberg, J. Nordqvist, and M. Laitinen, "Towards a language independent Twitter bot detector," in *Proceedings of the Conference of the Association of Digital Humanities in the Nordic Countries*, ser. DHN '19, vol. 2364. CEUR Workshop Proceedings, 2019, pp. 308–319.

[32] J. Scott and P. J. Carrington, Eds., *The SAGE Handbook of Social Network Analysis*. SAGE Publications, 2011.

[33] V. Thada and V. Jaglan, "Comparison of Jaccard, Dice, cosine similarity coefficient to find best fitness value for web retrieved documents using genetic algorithm," *International Journal of Innovations in Engineering and Technology*, vol. 2, no. 4, pp. 202–205, Aug. 2013.

[34] H. Rainie and B. Wellman, *Networked: The New Social Operating System*. MIT Press, 2012.

[35] R. Taylor, "Interpretation of the correlation coefficient: A basic review," *Journal of Diagnostic Medical Sonography*, vol. 6, no. 1, pp. 35–39, Jan. 1990.

[36] W. Labov, *Principles of Linguistic Change, Volume 2: Social Factors*. Wiley, 2001.

[37] Z. Fagyal, S. Swarup, A. M. Escobar, L. Gasser, and K. Lakkaraju, "Centers and peripheries: Network roles in language change," *Lingua*, vol. 120, no. 8, pp. 2061–2079, Aug. 2010.

[38] J. Du, Y. Xian, and J. Yang, "A survey on social network visualization," in *Proceedings of the International Symposium on Social Science*, ser. ISSS '15. Atlantis Press, Aug. 2015, pp. 419–423.

[39] S. Chen, L. Lin, and X. Yuan, "Social media visual analytics," *Computer Graphics Forum*, vol. 36, no. 3, pp. 563–587, Jun. 2017.

[40] K. Kucher, M. Fatemi, and M. Laitinen, "Towards visual sociolinguistic network analysis," in *Proceedings of the International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP '21) — Volume 3: IVAPP*, ser. IVAPP '21. SciTePress, 2021, pp. 248–255.